



LAB INSTRUCTION FOR BASIC COMPUTER SKILL AND DATA MANAGEMENT IN FIELD EPIDEMIOLOGY USING MS EXCEL

DR. PAISIN LEKCHAROEN

Lab Instruction for basic computer skill and data management in field epidemiology using MS Excel

Contents

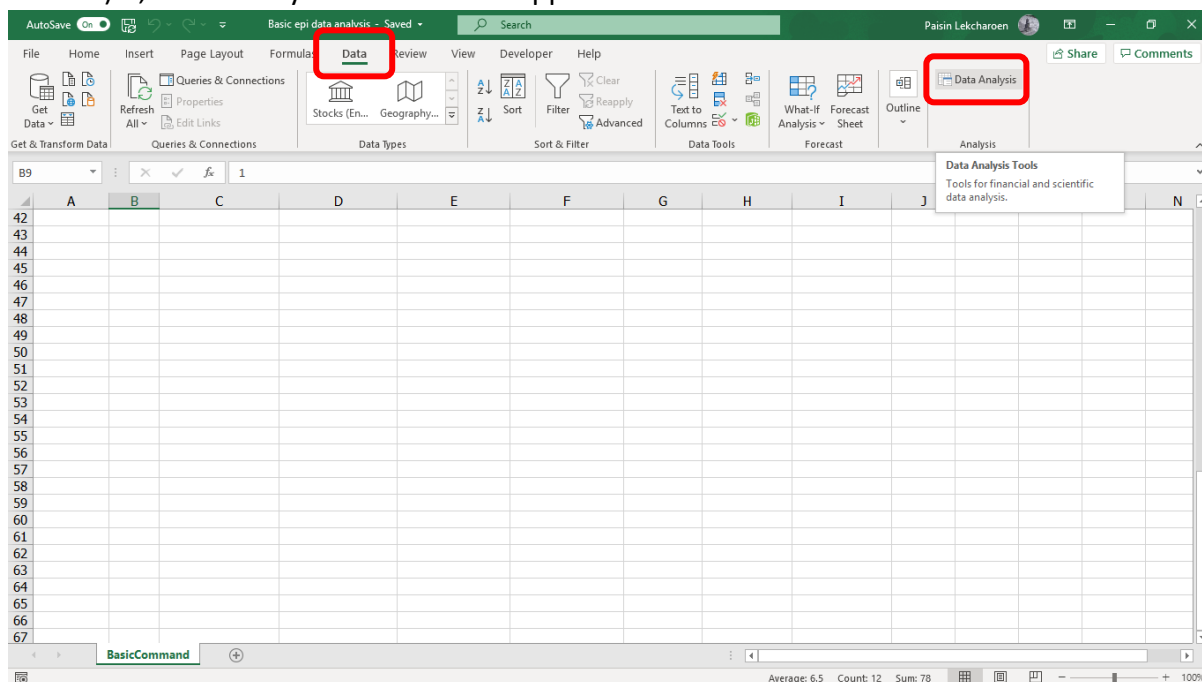
Pre-requirement	3
Activate 'Analysis ToolPak'	4
Review basic calculation function of Excel	6
Question 1 Which month has the highest prevalence of disease X and disease Y?	7
Question 2 What is the difference between disease X and disease Y prevalence per 100,000 population this year?	7
Question 3 Which city has the highest disease prevalence and what is the prevalence per 1,000 population?	7
Question 4 Which city has the lowest disease prevalence in female?	7
Question 5 How much the highest difference in disease prevalence between gender?	7
Part 1 Data entry and validation when entering data.	8
Part 2 Data conversion.....	24
Exercise 2.1 Find a season-year (Syr) of birth for participants.	25
Exercise 2.2 Calculate BMI for participants	35
Exercise 2.3 Define consumption and exercise behaviors, and case condition of all participants ..	36
Exercise 2.4 Vlookup function.....	37
Part 3 Data cleaning and recoding.....	39
Exercise 3.1 Filter	39
Exercise 3.2 Data Validation	46
Exercise 3.3 Pivot Table	48
Exercise 3.4 Look for and eliminate duplicates.	58
Exercise 3.5 Recode.	62
Part 4 Basic data analysis using Pivot Table.....	66
Exercise 4.1 Find a proportion of PhD graduate among participants.....	66
Question 6 What is a proportion of participants who was born in wet season?	69
Exercise 4.2 Find a central tendency and dispersal of variables.....	69
Question 7 Which gender has higher average weight?	75
Question 8 How much different among average height of male and female?	75
Question 9 Which gender has lower variation of BMI?.....	76
Exercise 4.3 Find average BMI among different regularities of activities.	76
Exercise 4.4 How to find odds ratio.	82
Question 10 what is the study design for this study?.....	82



Question 11 What is the odds ratio for this table?.....	83
Question 12 What is your interpretation?.....	83
Question 13 What is the odds of having regular sleep (often/always) and having.....	85
Question 14 Do participants who have overweight and normal BMI have different eating regularity?.....	85
Question 15 Please interpret the likelihood of regularity of talking with friend and being overweight?	85
Part 5 Basic data analysis.....	85
Exercise 5.1 Descriptive statistics	86
Question 16 What is the average weight of the participants?	89
Question 17 What is a standard variation of height of participants?.....	89
Question 18 What direction is the skewness of BMI?	90
Exercise 5.2 Comparison of means in 2 independent groups.....	90
Question 19 Which gender has higher average height?.....	92
Question 20 Is BMI of participants who have more regularity (often and always) in physical exercise less than of those who have not?	99
Exercise 5.3 Comparison of an average BMI among different age groups and different birth season.	99
Question 21 Are there any differences of BMI among participants who has different age and birth season?.....	99

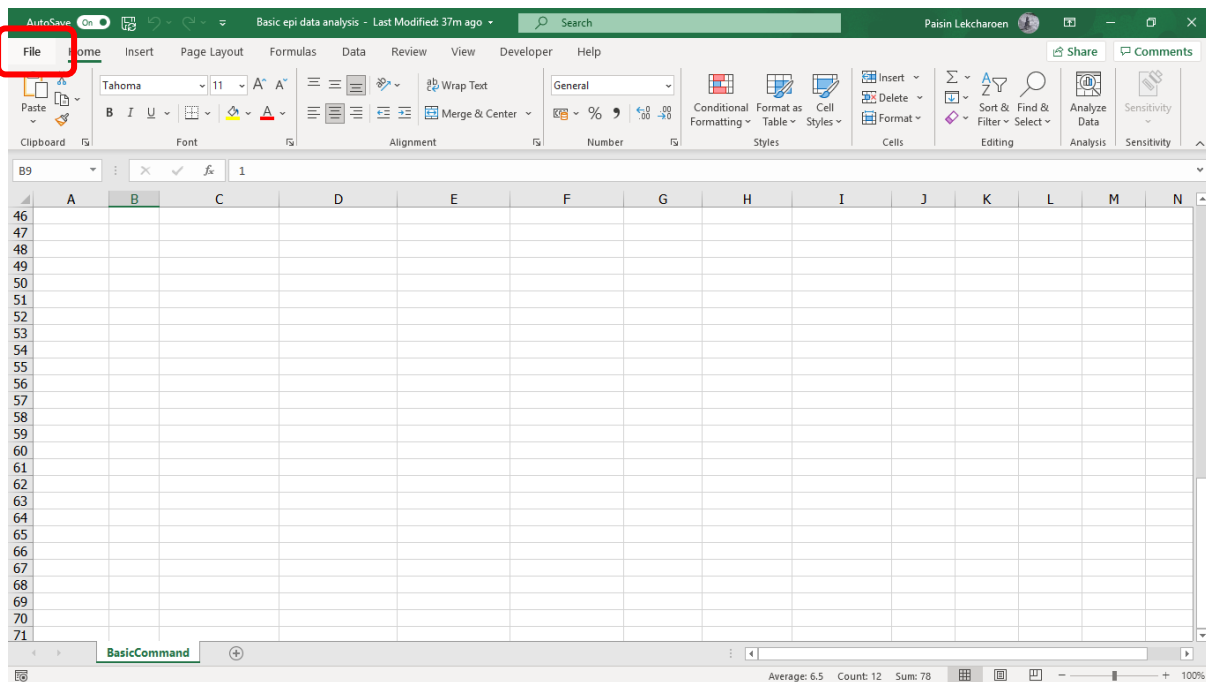
Pre-requirement

1. Download MS Excel workbook 'Basic epi data analysis'.
2. Ensure you already have 'Analysis Toolpak' add-in installed then the 'Data Analysis' tool would appear on the MS Excel's toolbar. If not, download it from <https://support.microsoft.com/th-th/office/%e0%b9%82%e0%b8%ab%e0%b8%a5%e0%b8%94-analysis-toolpak-%e0%b9%83%e0%b8%99-excel-6a63e598-cd6d-42e3-9317-6b40ba1a66b4?ui=th-th&rs=th-th&ad=th>
3. Install (if you just download it) or activate (if you already have this package in your add-ins) it, 'Data Analysis' tools should appear on the 'Data' ribbon.

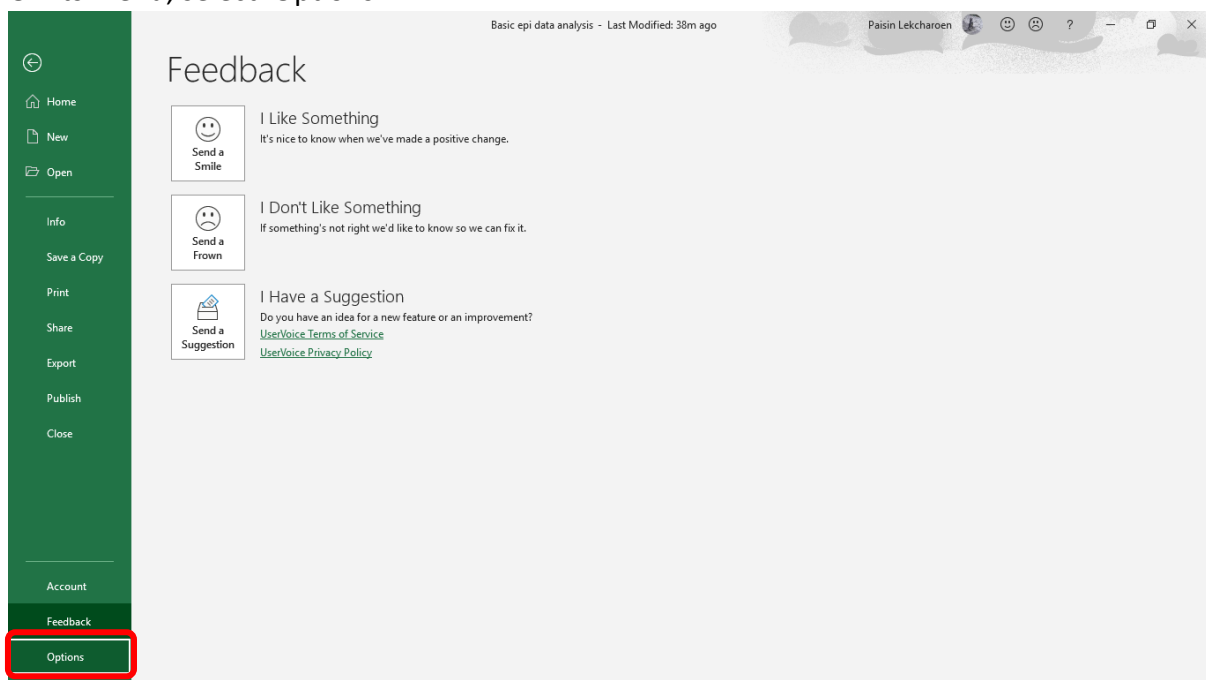


Activate 'Analysis ToolPak'

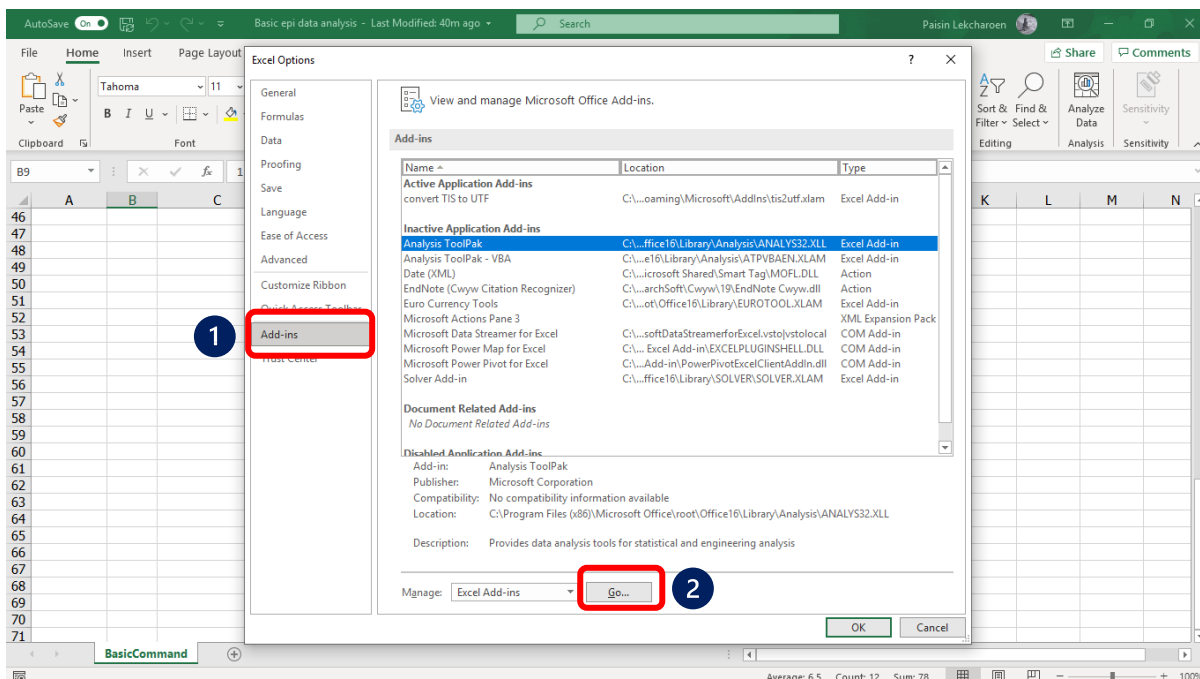
1. Click 'File' ribbon on the toolbar.



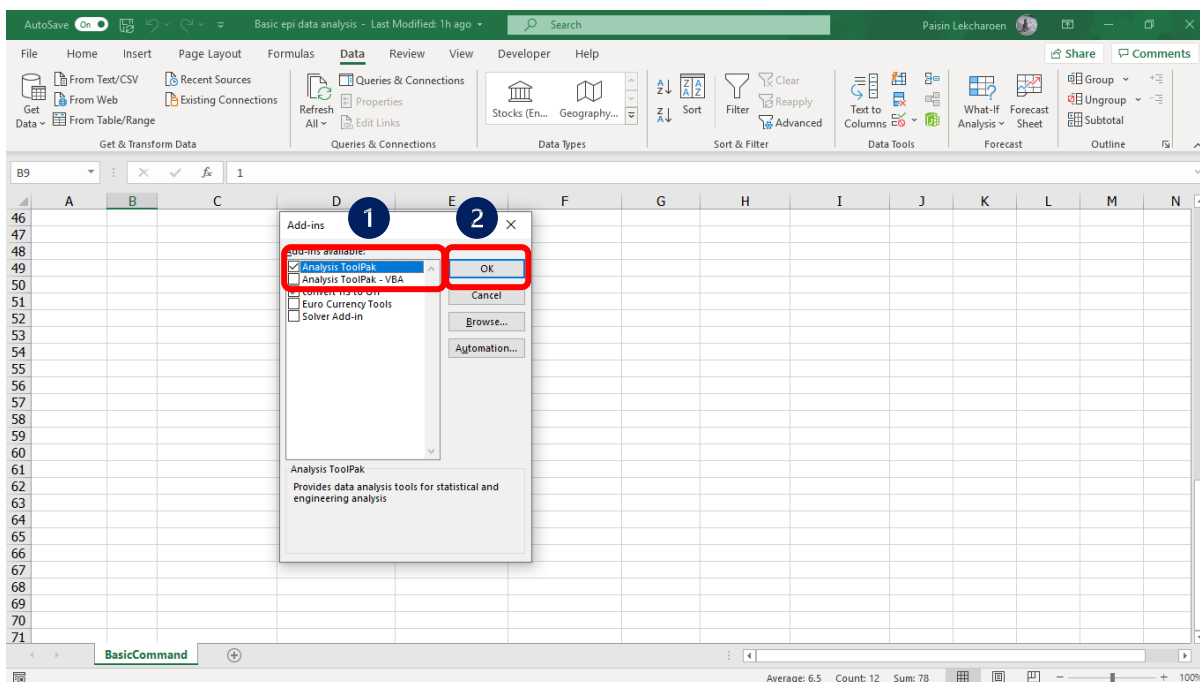
2. On its menu, select 'Options'.



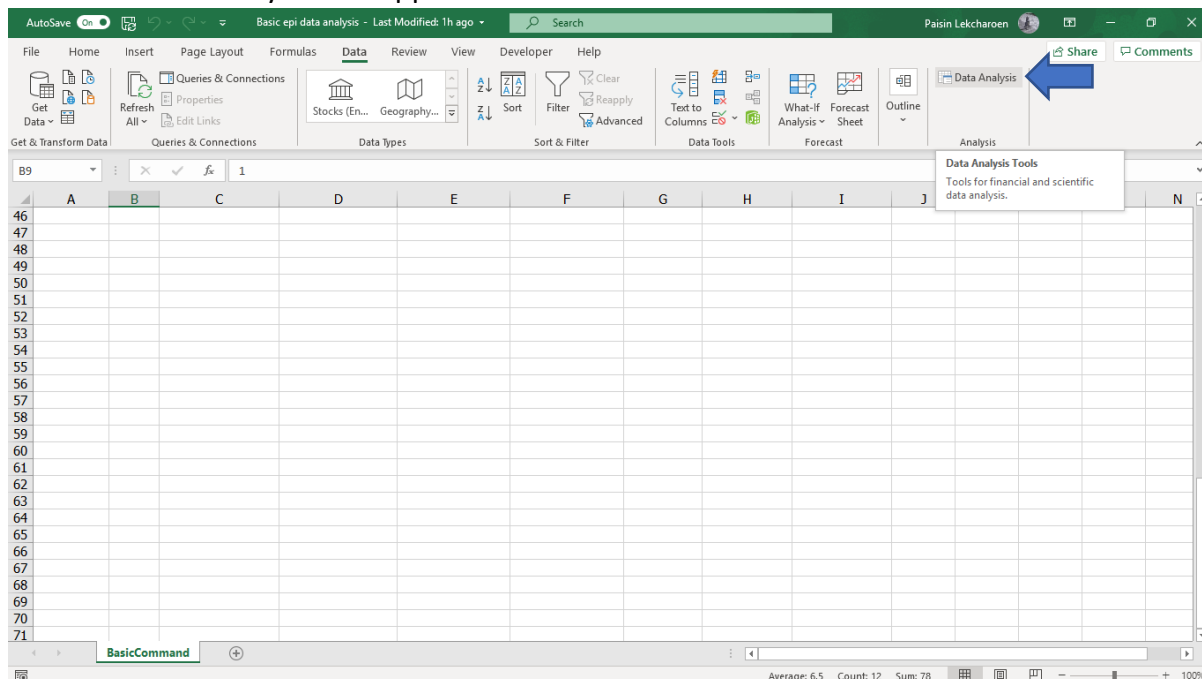
3. The 'Excel Options' window will appear. The 'Analysis ToolPak' will be showed in the 'Inactivated Application Add-ins' (as highlighted in blue). Then click on 'Go' to open the 'Add-ins' toolbox.



4. Check on a box 'Analysis ToolPak' (as highlighted in blue). Then click 'OK' to activate this add-in.



5. Then the 'Data Analysis' will appear in the 'Data' ribbon.



Review basic calculation function of Excel

Go to workbook 'Basic epi data analysis', then choose worksheet 'BasicCommand'. Calculate prevalence of disease for Table 1 and 2. Try to answer these questions.

Questions for Table 1

Question 1 Which month has the highest prevalence of disease X and disease Y?

Answer: Click or tap here to enter text.

Question 2 What is the difference between disease X and disease Y prevalence per 100,000 population this year?

Answer: Click or tap here to enter text.

Questions for Table 2

City	MalePop	DisX_male	FemalePop	DisX_female	Prev_total	Prev_male	Prev_female	DifferentPrev
A	2000	23	3563	11	11			
B	2300	2	2341	13	13			
C	3400	10	2345	12	12			
D	1353	24	1444	22	22			
E	1256	35	2322	43	43			
F	6780	12	7800	15	15			
G	1900	25	2341	23	23			
H	3455	33	2800	31	31			
I	1255	34	1677	12	12			
J	5460	45	5560	34	34			

Question 3 Which city has the highest disease prevalence and what is the prevalence per 1,000 population?

Answer: Click or tap here to enter text.

Question 4 Which city has the lowest disease prevalence in female?

Answer: Click or tap here to enter text.

Question 5 How much the highest difference in disease prevalence between gender?

Answer: Click or tap here to enter text.

Part 1 Data entry and validation when entering data.

Functions in use:

- Format Cells
- Data Validation
- Define Name
- Allow Edit Range
- Protect Sheet

1. Familiarize with the questionnaire (download from ...)

Questionnaire ID

Questionnaire for R-FETPV trainee, 2021

Part I Personal information

Q1. Given name - Last name ឈ្មោះត្រកូលនិងឈ្មោះ Q2. Age អាយុ

Q3. Gender: Male Female Other Q4. Weight ទម្ងន់

Q5. Height មមាត់

Q6. Nationality: Bhutanese Burmese Cambodian Chinese Filipino Indonesian Laotian Malaysian Nepalese Thai Vietnamese Other

Q7. Education: Bachelor Master PhD Other

Q8. Affiliation: Government Non-government Academic Other

Q9. How long have you been in recent position? ឆ្នាំ

Part II R-FETPV training

Q10. Do you have previous experience in field epidemiology training? Yes No

Q11. What species are studied in your animal health situation analysis? (choose all that applicable) Aquatic Poultry Ruminant Swine Horse Dog and cat Wildlife Other

Q12. What types of health problem related to your animal health situation analysis? Non-infectious Infectious >> (Virus Bacteria Protozoa Helminth Multiple) Other

Q13. How often is your activity when you free from this training in a week?

Going to cinema	<input type="checkbox"/> Never <input type="checkbox"/> Rarely <input type="checkbox"/> Sometimes <input type="checkbox"/> Often <input type="checkbox"/> Always
Watching TV	<input type="checkbox"/> Never <input type="checkbox"/> Rarely <input type="checkbox"/> Sometimes <input type="checkbox"/> Often <input type="checkbox"/> Always
Watching YouTube	<input type="checkbox"/> Never <input type="checkbox"/> Rarely <input type="checkbox"/> Sometimes <input type="checkbox"/> Often <input type="checkbox"/> Always
Physical exercise	<input type="checkbox"/> Never <input type="checkbox"/> Rarely <input type="checkbox"/> Sometimes <input type="checkbox"/> Often <input type="checkbox"/> Always
Shopping	<input type="checkbox"/> Never <input type="checkbox"/> Rarely <input type="checkbox"/> Sometimes <input type="checkbox"/> Often <input type="checkbox"/> Always
Go walking/sightseeing	<input type="checkbox"/> Never <input type="checkbox"/> Rarely <input type="checkbox"/> Sometimes <input type="checkbox"/> Often <input type="checkbox"/> Always
Eating	<input type="checkbox"/> Never <input type="checkbox"/> Rarely <input type="checkbox"/> Sometimes <input type="checkbox"/> Often <input type="checkbox"/> Always
Sleeping	<input type="checkbox"/> Never <input type="checkbox"/> Rarely <input type="checkbox"/> Sometimes <input type="checkbox"/> Often <input type="checkbox"/> Always
Reading magazine/comics	<input type="checkbox"/> Never <input type="checkbox"/> Rarely <input type="checkbox"/> Sometimes <input type="checkbox"/> Often <input type="checkbox"/> Always
Cooking	<input type="checkbox"/> Never <input type="checkbox"/> Rarely <input type="checkbox"/> Sometimes <input type="checkbox"/> Often <input type="checkbox"/> Always
Meditation	<input type="checkbox"/> Never <input type="checkbox"/> Rarely <input type="checkbox"/> Sometimes <input type="checkbox"/> Often <input type="checkbox"/> Always
Talking with friends	<input type="checkbox"/> Never <input type="checkbox"/> Rarely <input type="checkbox"/> Sometimes <input type="checkbox"/> Often <input type="checkbox"/> Always
Chaffing via social media	<input type="checkbox"/> Never <input type="checkbox"/> Rarely <input type="checkbox"/> Sometimes <input type="checkbox"/> Often <input type="checkbox"/> Always

Questionnaire ID

Part III Health information

Q14. Do you have any of these problems after the course started?

Yes >> (choose all that applicable) Sleepless Diarrhea Fever [Cough](#)

Sore throat Vomiting Upper respiratory tract infection Physical injury Other

No

Q15. When did you initially observe the above symptoms? (DDMMYY)

Q16. What do you do when you or your family get influenza-like infection?

Rest at home Visit pharmacy Visit clinic/hospital Other

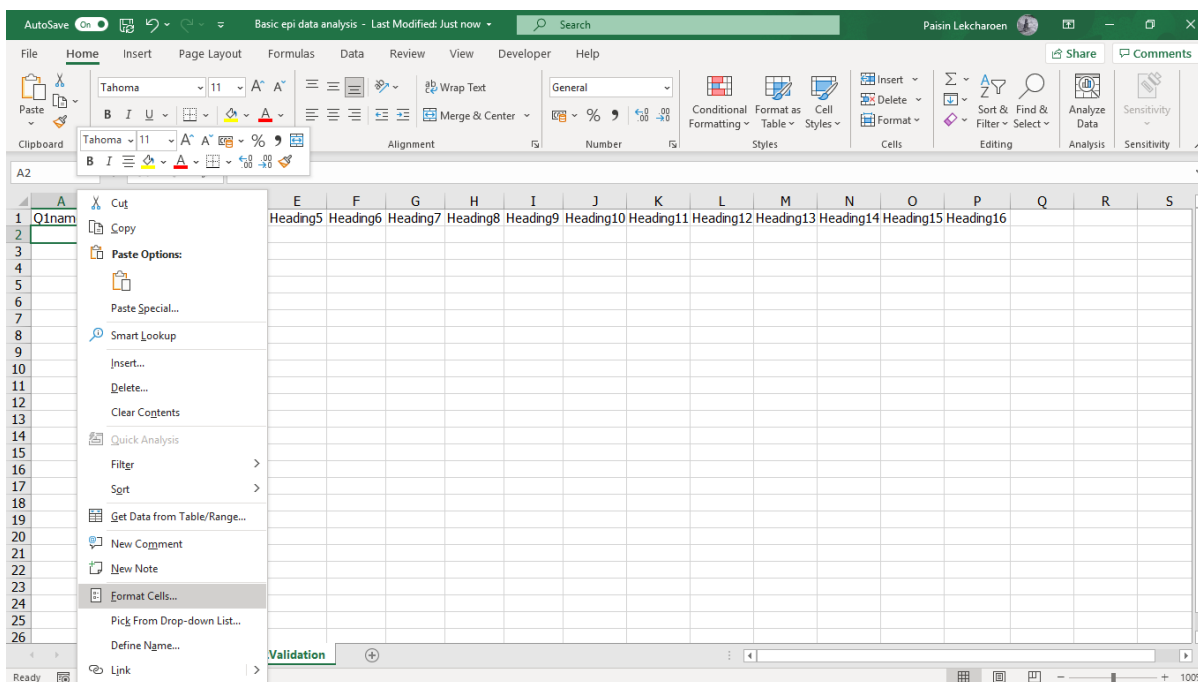
Interviewer

2. Go to 'Prt1Entry&Validation' sheet in 'Basic epi data analysis' workbook. You may see that the first row contains 16 headings: Heading1 to Heading16, correspond to 16 questions in the questionnaire.

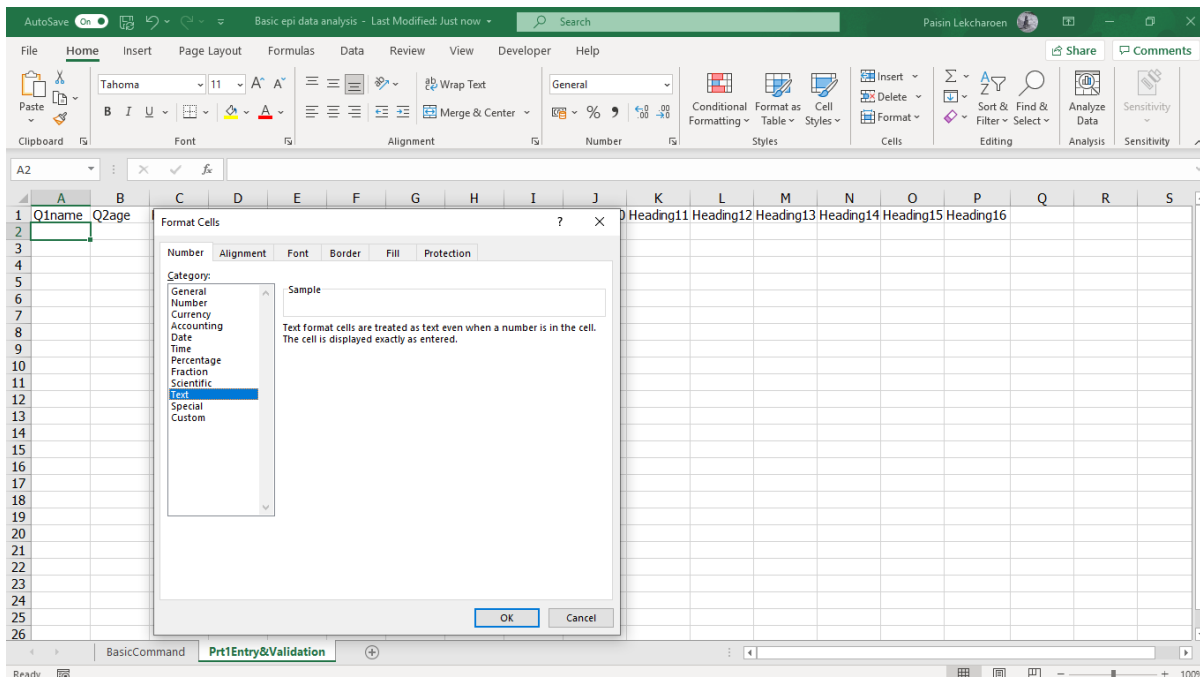
The screenshot shows the Microsoft Excel interface. The title bar reads 'Basic epi data analysis - Last Modified: 2m ago'. The ribbon is set to 'Data'. The active sheet is 'Prt1Entry&Validation'. The first row of the worksheet contains 16 columns, each with a heading from 'Heading1' to 'Heading16'. The rest of the worksheet is empty.



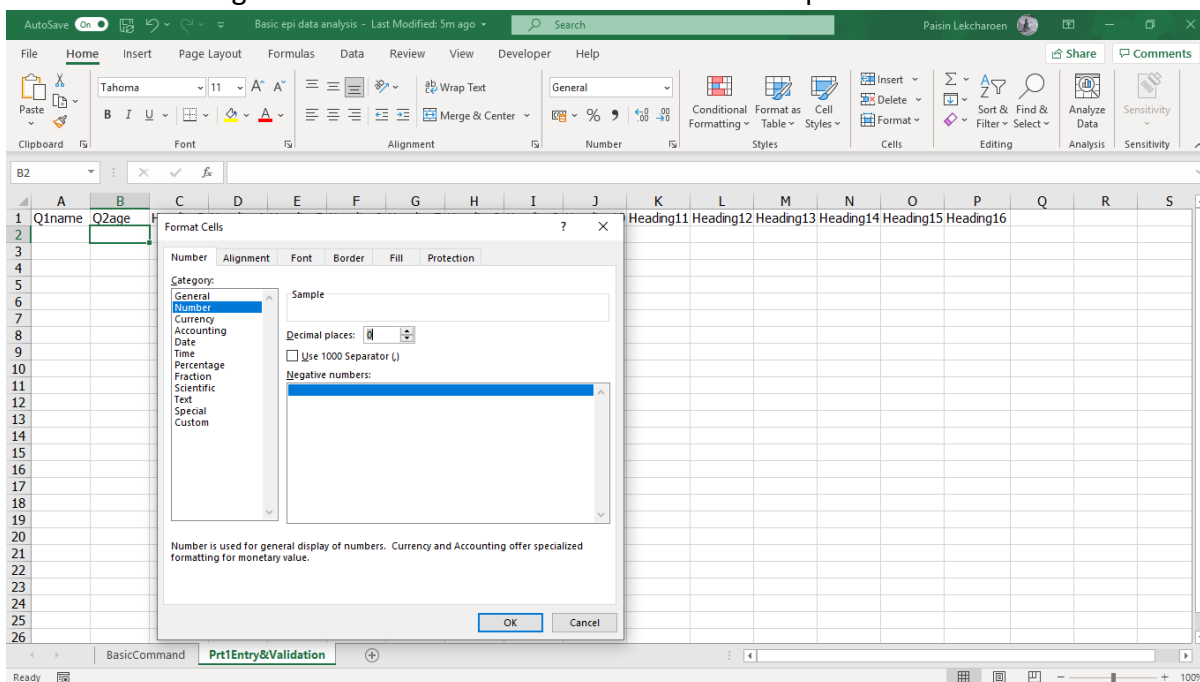
- Please create a Data Entry Form using this sheet appropriately.
 - Change the heading name correspond to the question.
 - e.g., Heading1 → Q1name
 - The heading should be short, uses limited characters, represents the corresponding question or data that it may contain, be able for tracking back to the original data, and avoids space in the heading.
 - Be compatible to the Software using in analytical process.
 - *Some questions, particularly the questions that can have multiple answers, can contain more than one column in the data entry form. Adjust them appropriately. Consider about how you are going to analyze them.
 - Each row represents only one element/subject. The unit of the subject can be individual animal, farm, or any unit as defined in the methodology.
 - Each subject should be indexed (that may derive from a subject ID or a questionnaire ID)
3. Use appropriate data format by using 'Format Cells' to define correct cell format for each column (or question). The cell formats can be text, number, date, etc.
- 3.1. Right click on the cell(s) that you want to define its/their property.
- 3.2. Choose 'Format Cells'



3.3. Select appropriate cell format from the 'Category' box in the 'Number' ribbon in the 'Format Cells' window. For example, the suitable cell format for 'Q1name' is text. Select 'Text' then click OK.



3.4. The cell format for 'Q2age' (upon your heading change) is number. So, select 'Number'. Age contains no decimal so indicate 'Decimal places' to '0'.

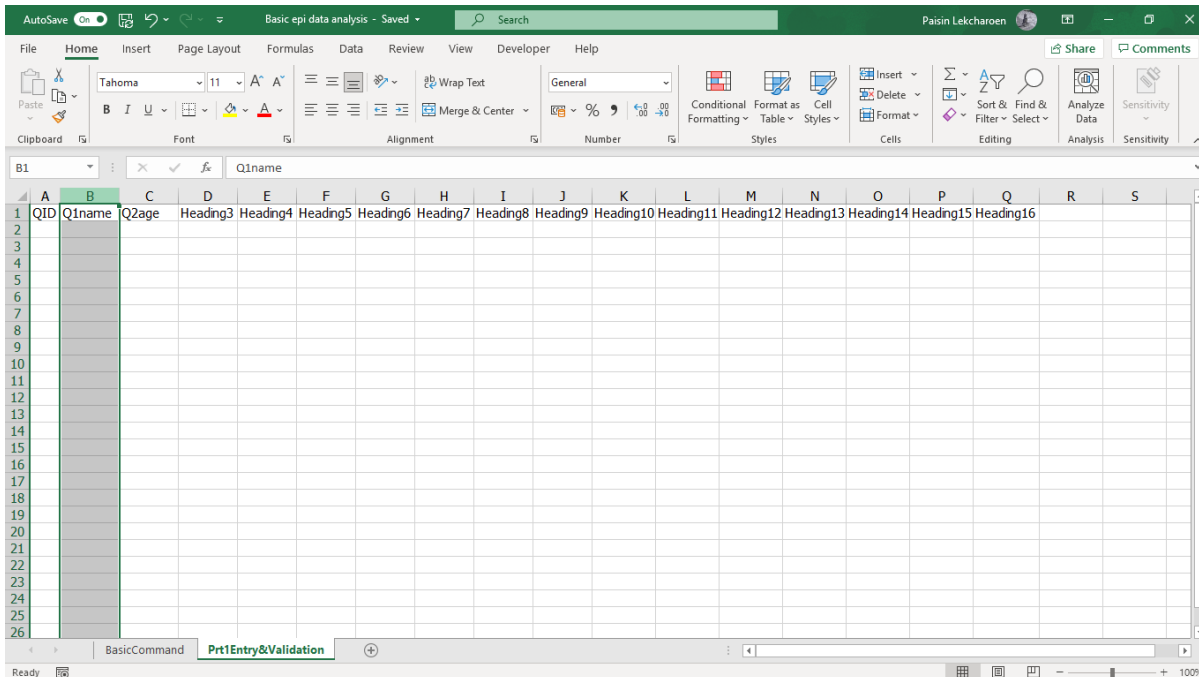


3.5. Continue to the remaining columns.

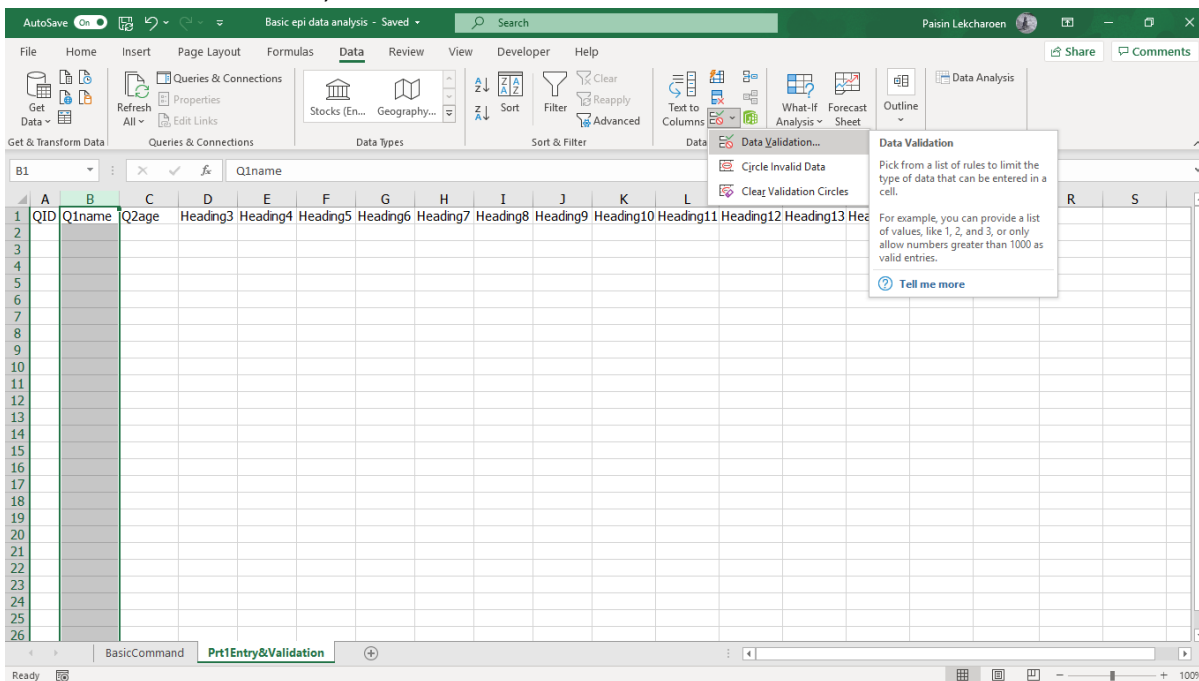


4. Mask the cells to limit possible values (answers) for each column. This can help us avoiding wrong inputs or typos during data entry.

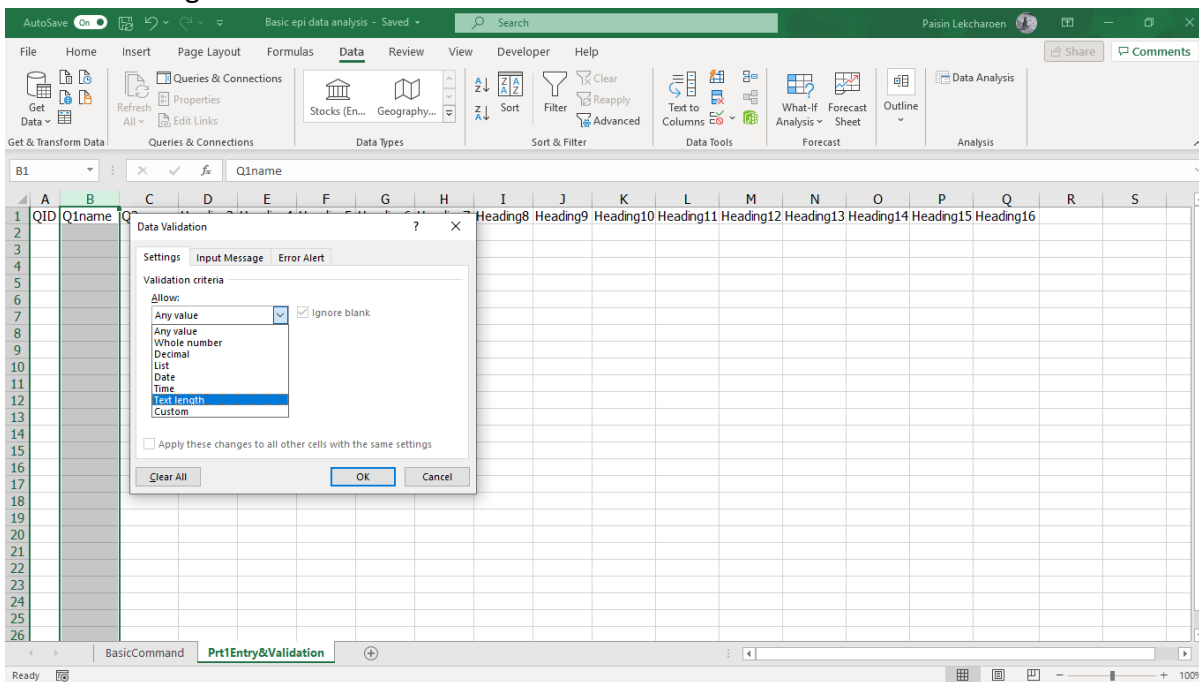
4.1. E.g., for Q1name, texts are allowed. Highlight the column B 'Q1name' (after inserting and indexing column A with the questionnaire ID - 'QID').



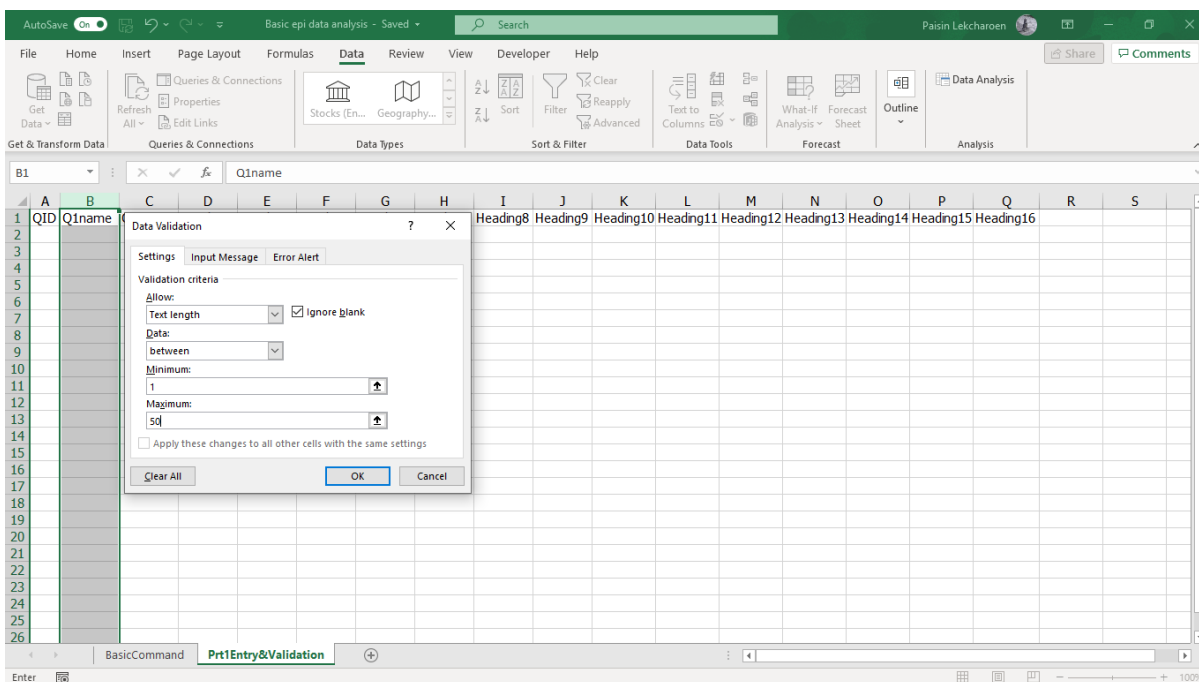
4.2. On the Data ribbon, select 'Data Validation' function.



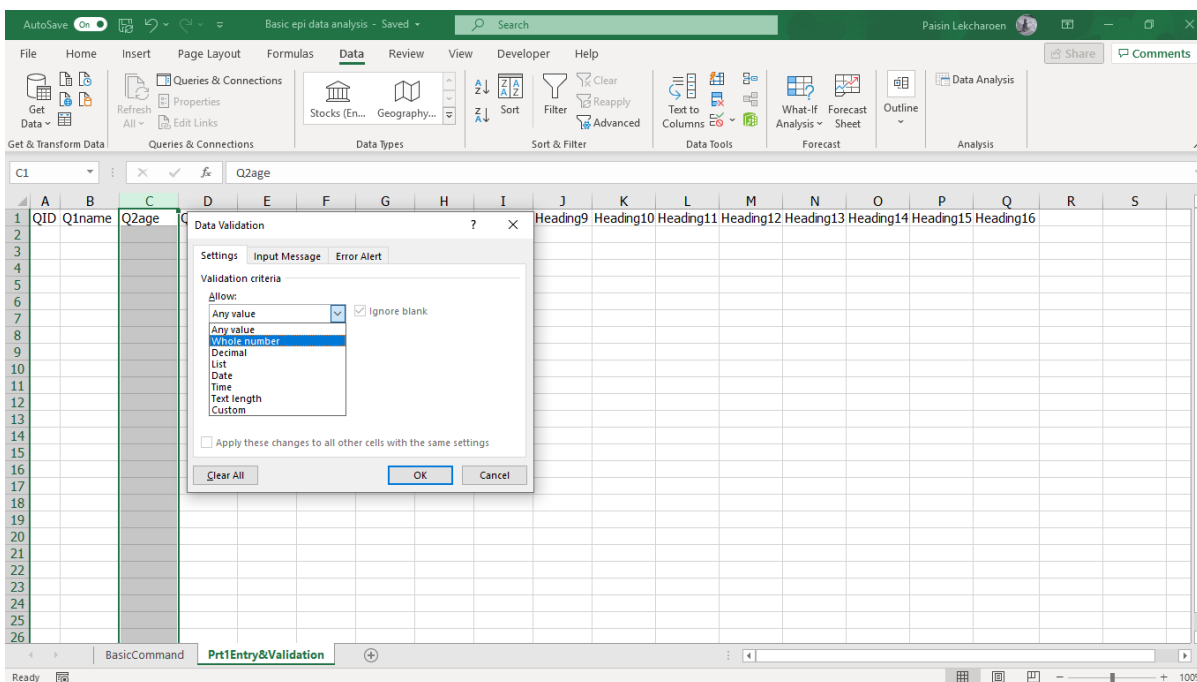
4.3. In the 'Data Validation' window, as the value allowed for this question is text, select 'Text length' from the 'Allow:' box.



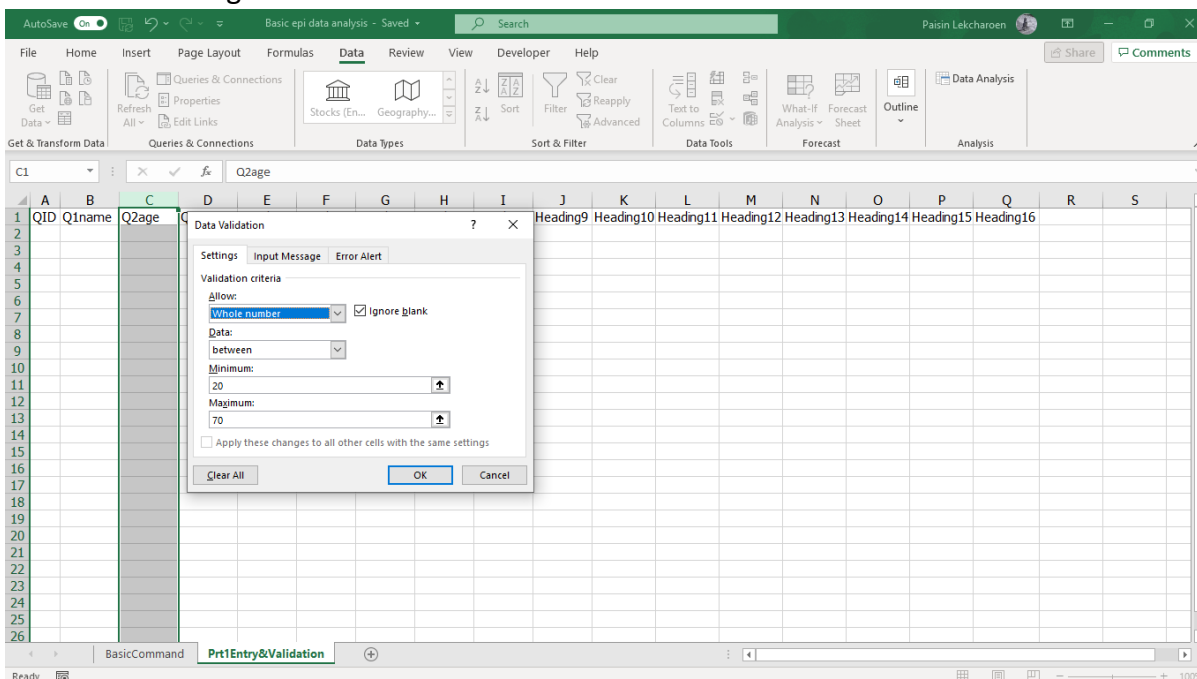
4.4. In the 'Data:' box, select 'Between'. Then indicate minimum and maximum length as you preferred. In this example, the minimum length is 1 and the maximum length is 50. Then click 'OK'.



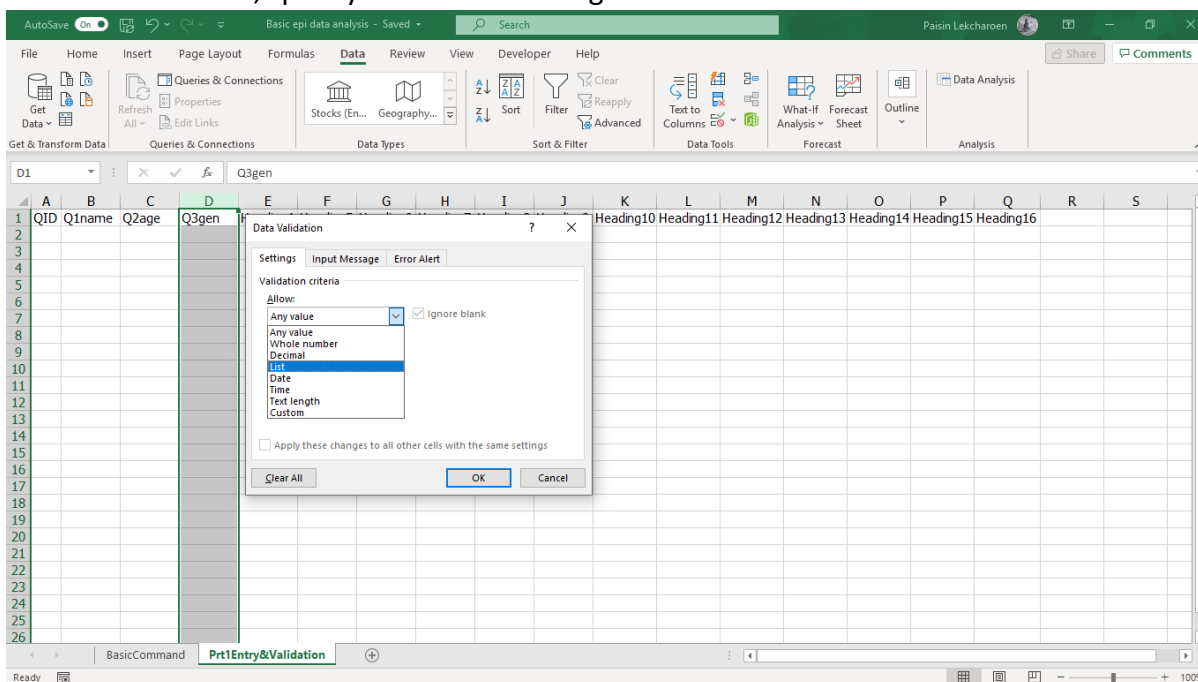
4.5. For 'Q2age' which only a whole number is allowed, select 'Whole number' from the 'Allow:' box in the 'Data Validation' window.



4.6. Select 'Between' for Data. As we expected that the participants are people who have already been in employment, so the age should be in between age at graduation and retirement. For this example, we indicate 20 as minimum and 70 as maximum age. Then click 'OK'.

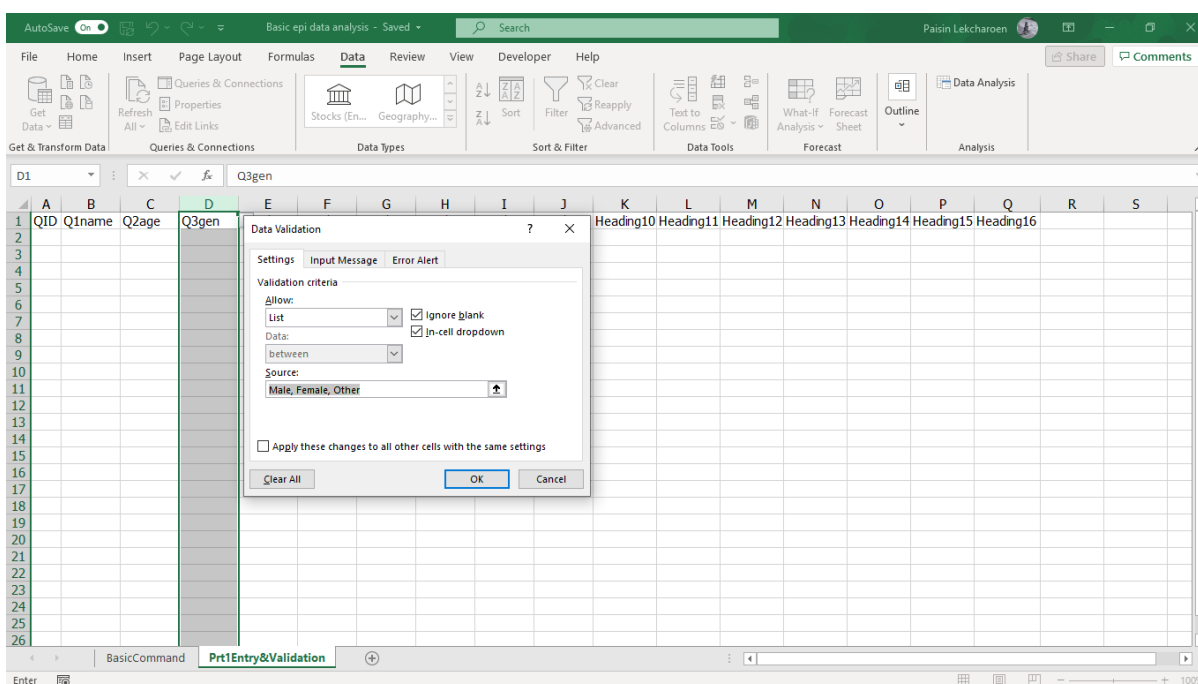


4.7. The 'Q3gen' representing a gender of participant allows three values: Male, Female, and Other. So, specify these values using 'List' from 'Allow:' box.

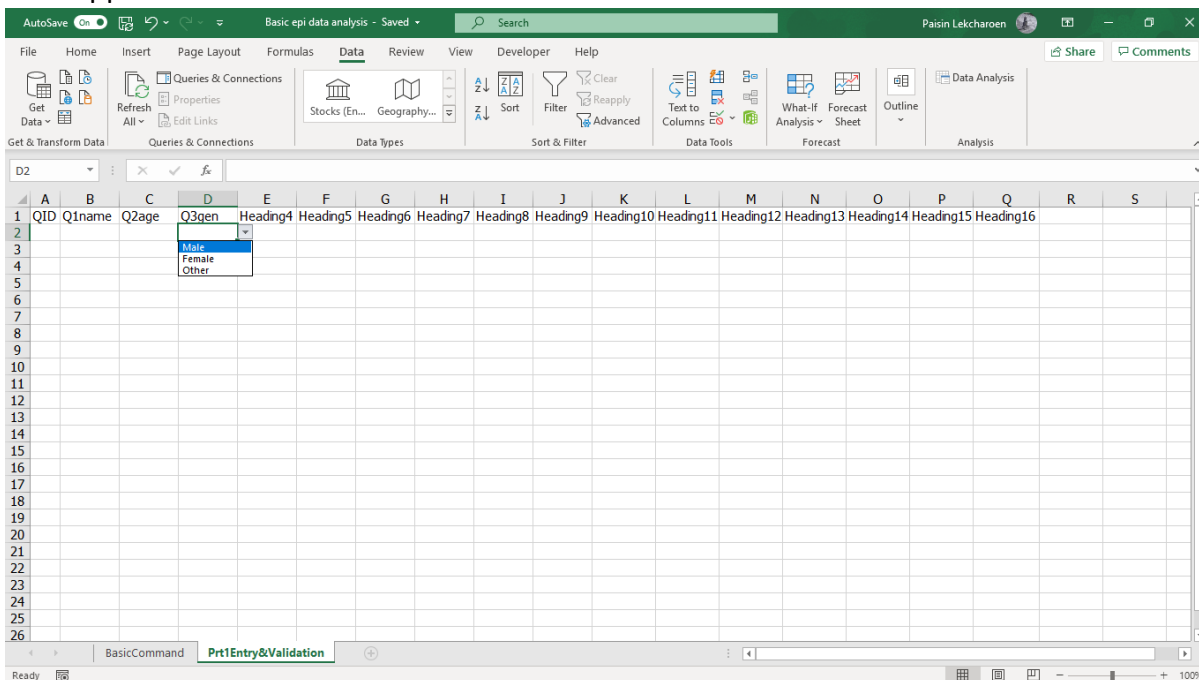


4.8. When selecting 'List', the 'Source:' box is activated. You can put values by two method:

- 1) Put values you use as an alternative in the question one by one, separated by comma (,), e.g., 'Male, Female, Other' (as shown in the figure below).
- 2) Put name of defined list (that contains a list of value Male, Female, and Other that you create elsewhere in different worksheet in this workbook (practice provided later).



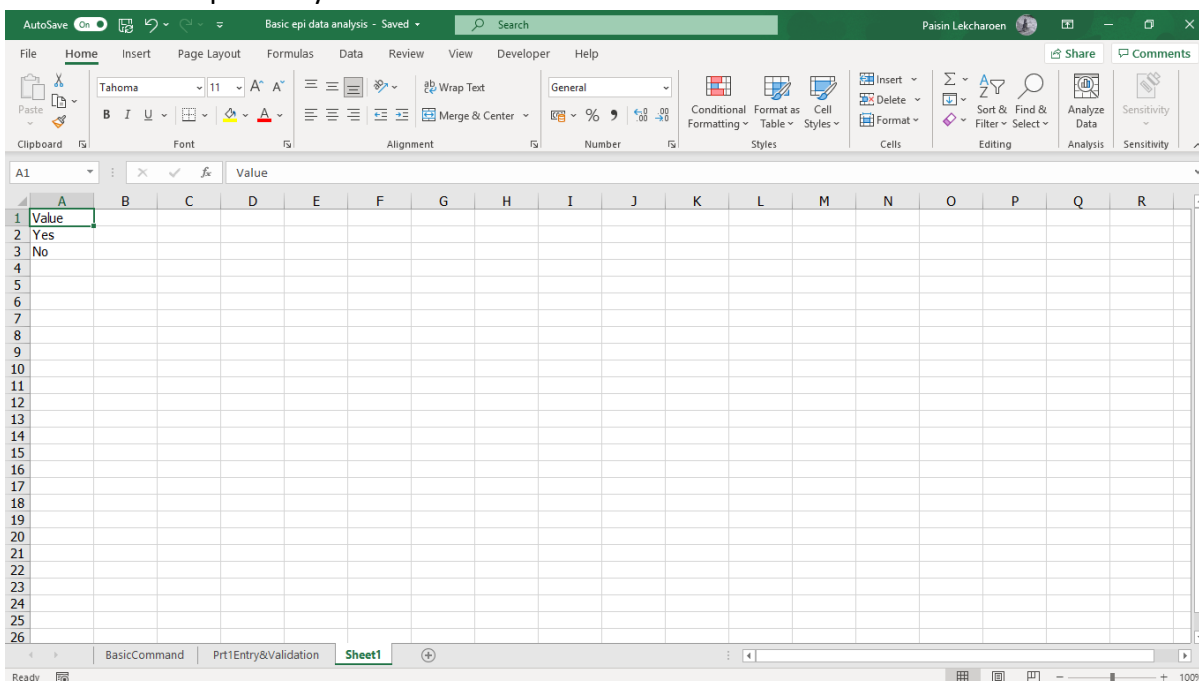
4.9. So, when you want to enter data for this question, a drop-down list of values will appear.



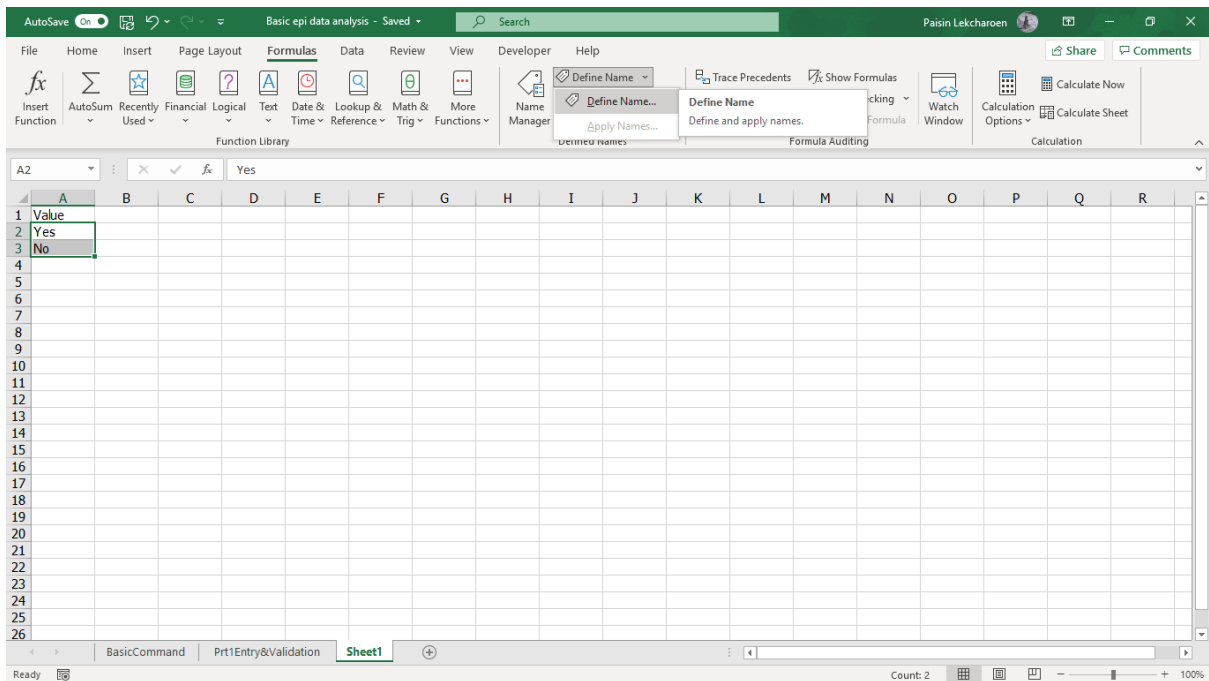
4.10. For some questions which have a same list of answer, you can define a 'List' of values and apply this list to those questions.

For example, 'Q10epiexp' (if the participant has epidemiological training experience) has 'Yes' and 'No' as the alternatives as same as those of 'Q11aq' (if the participant study about aquatic animal or not?).

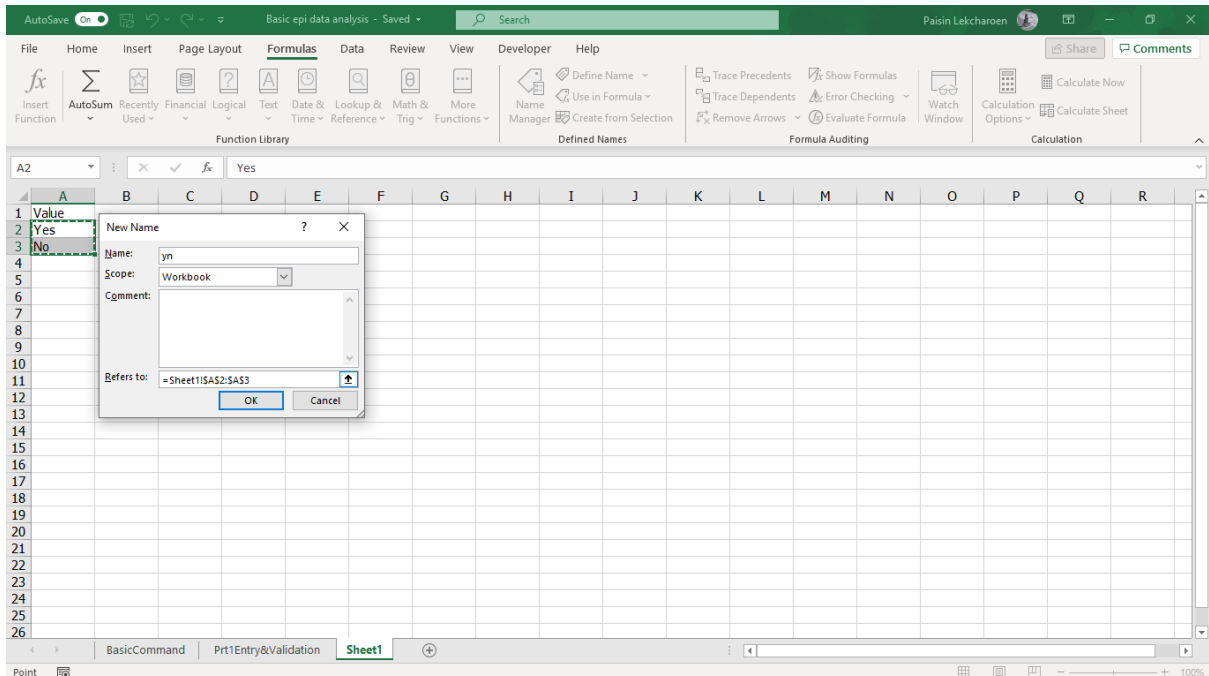
- Insert new worksheet.
- In the new worksheet, put a name for the first column (column A) as 'Value' in the first row. Then put 'Yes' and 'No' in the second and third row, respectively.



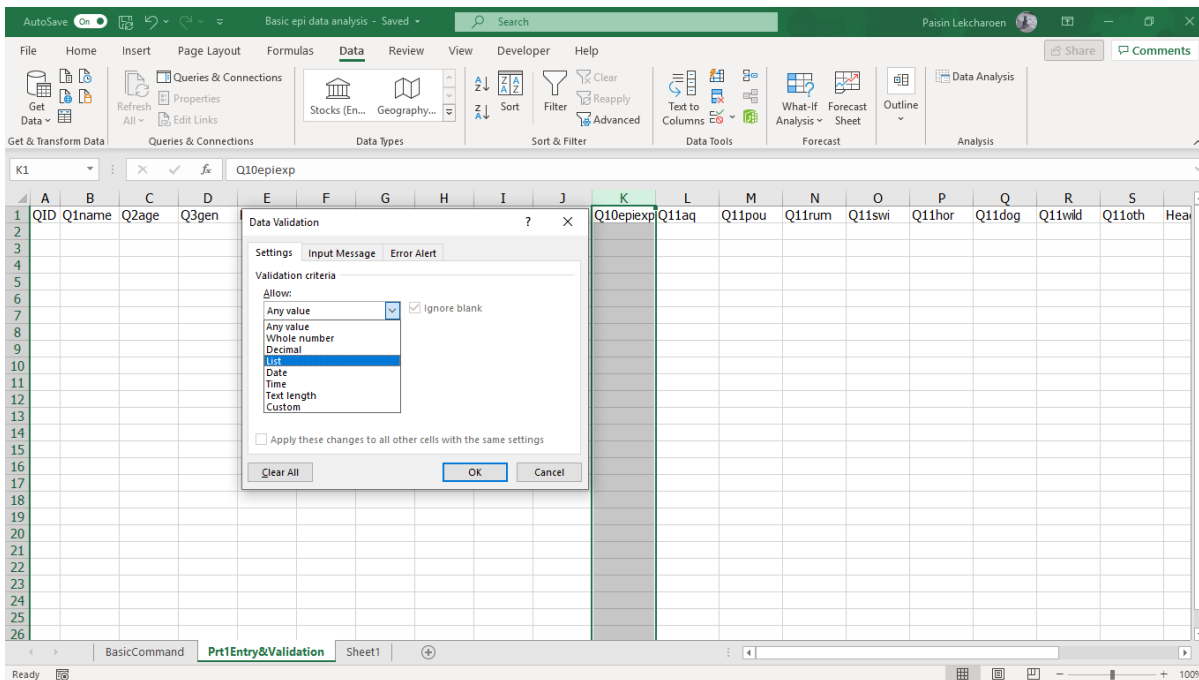
- Highlight across Yes and No. Then, click on 'Formulas' ribbon, select 'Define Name' function.



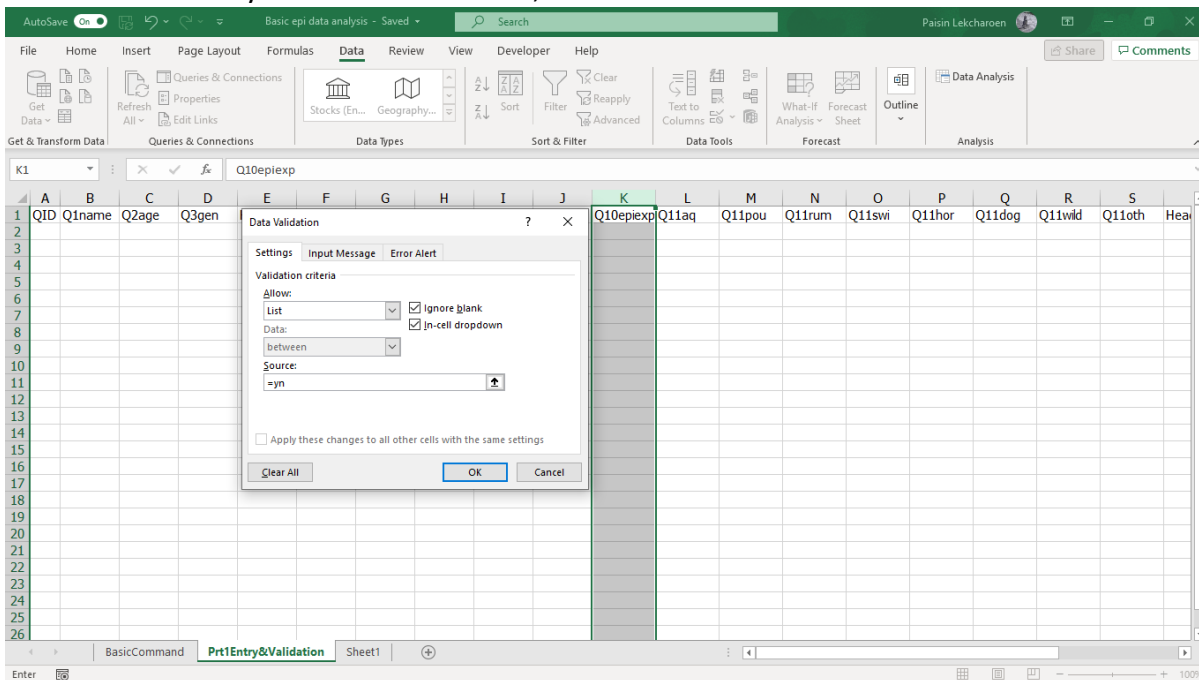
- A 'New Name' window appears. Write 'yn' (stands for a list containing Yes and No values) in the 'Name' box. By the way in the 'Refers to' box, it is shown as '=Sheet1!\$A\$2:\$A\$3' while a moving rectangle appear across the cells A2 and A3 which contain Yes and No values. Click 'OK'.



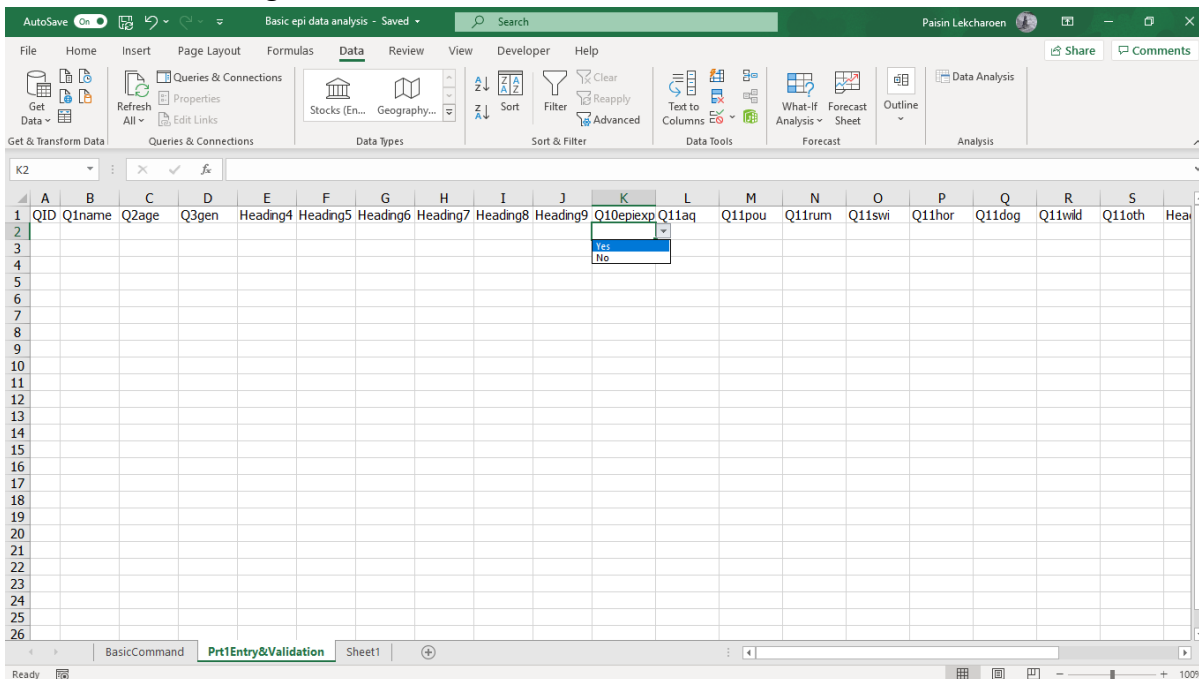
- Go back to the working sheet (Prt1Entry&Validation). Apply the 'yn' list to 'Q10epiexp'. Click 'Data' ribbon and select 'Data Validation' function. Then choose 'List' from the 'Allow' box.



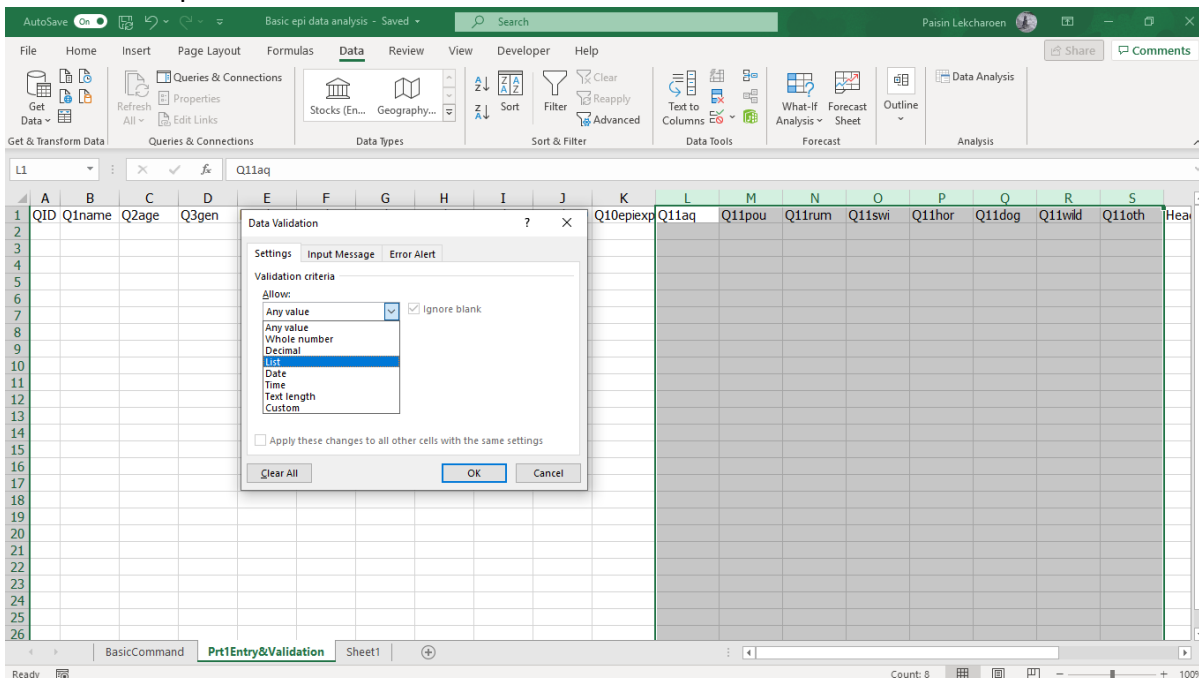
- Put '=yn' in the 'Source' box, then click 'OK'.

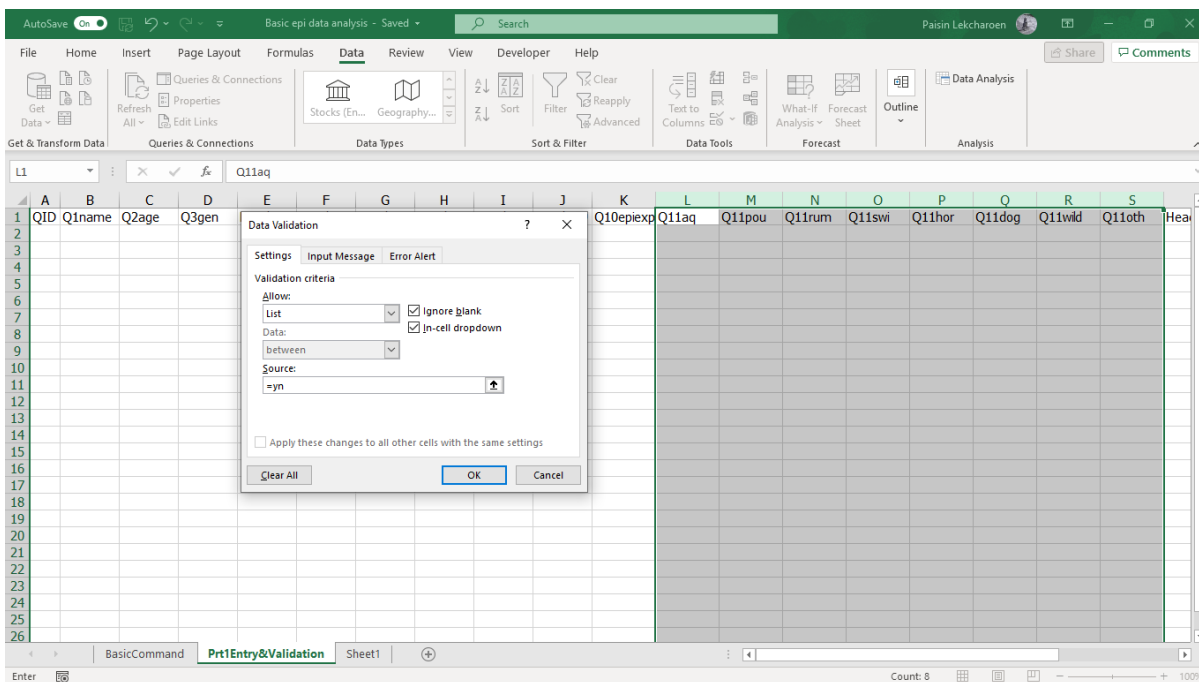


- Now you can choose only ‘Yes’ or ‘No’ from the drop-down list when entering data.



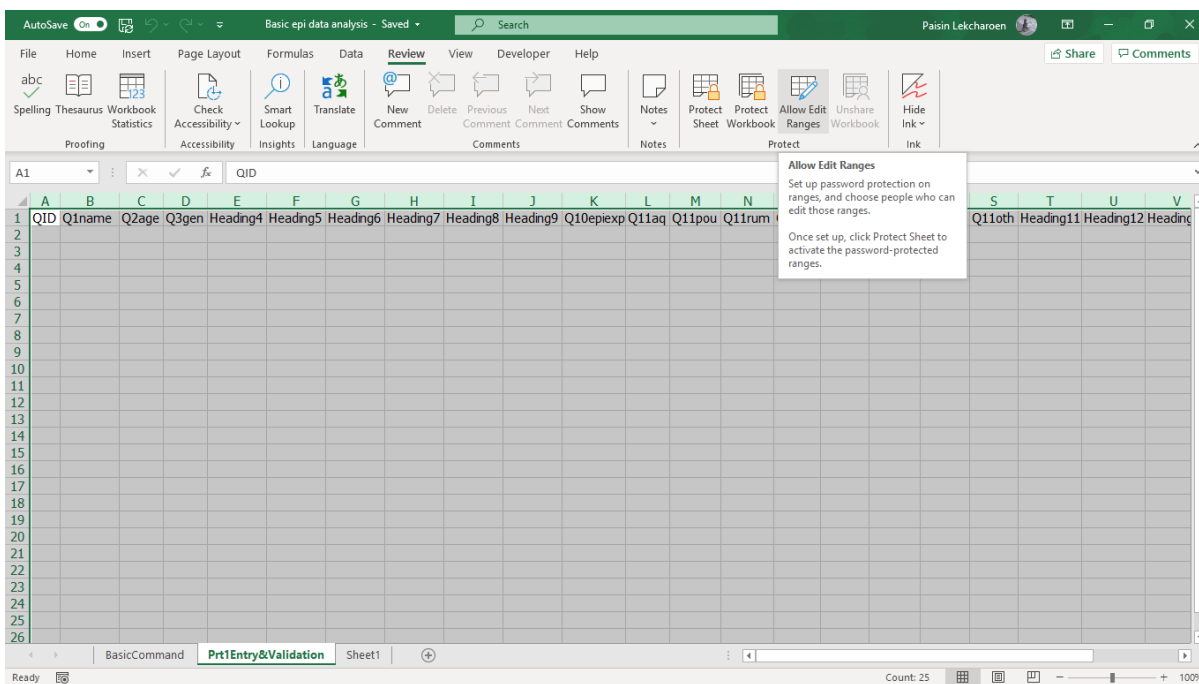
- You can also apply the list to multiple questions which have the same answer alternatives at once. ‘Q11aq’ to ‘Q11oth’ are derived from question number 11. All these questions need ‘Yes’ or ‘No’ answer if it is not missing. So, highlight all of these questions. Then apply the list ‘yn’ for these questions.



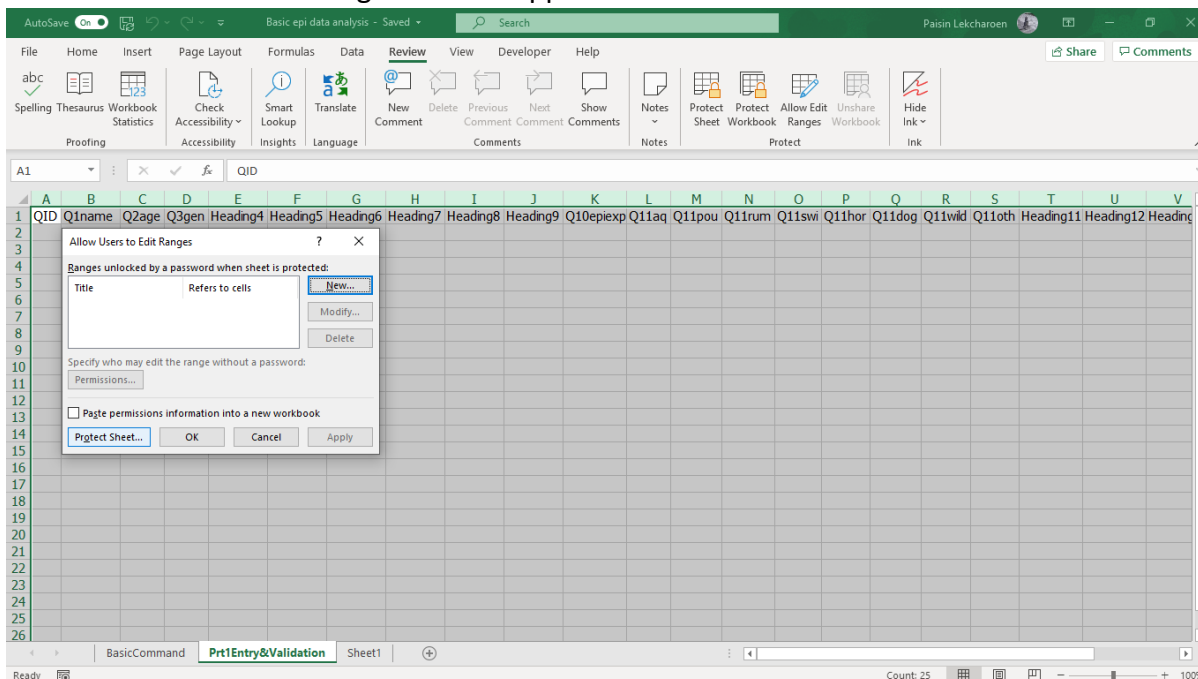


4.11. Continue to all columns as appropriate.

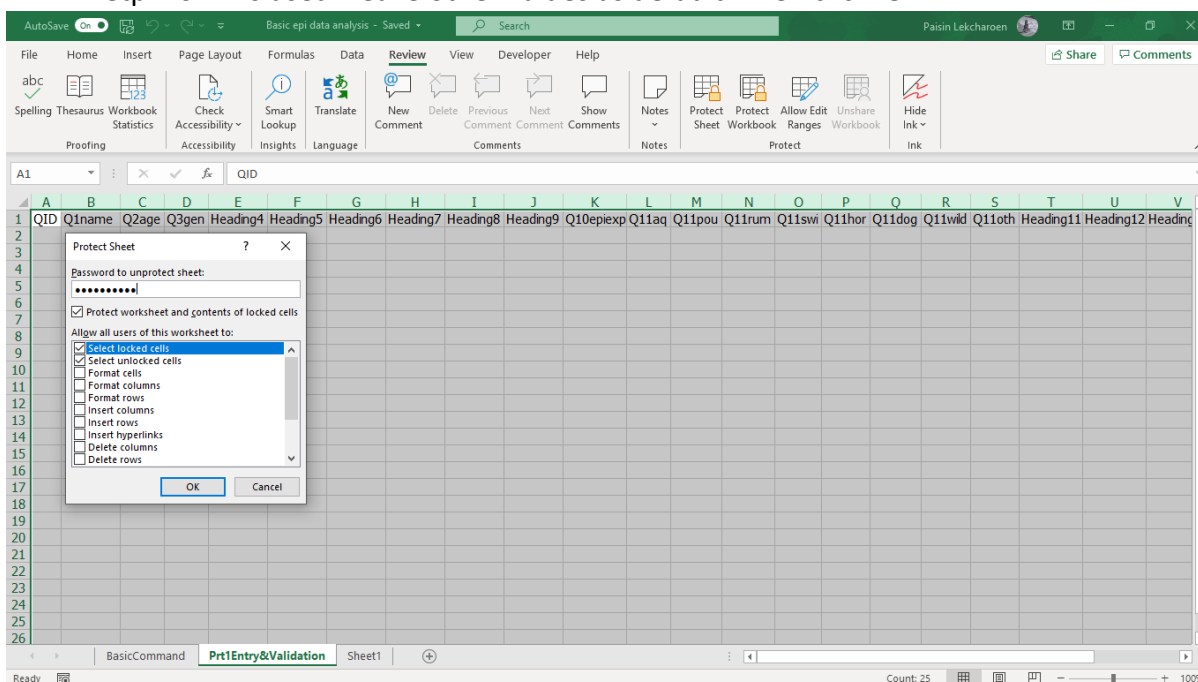
4.12. Use 'Allow Edit Ranges' function on range of cells that are going to be used by highlighting across all columns and rows you preferred. This function can be selected from 'Review' ribbon.



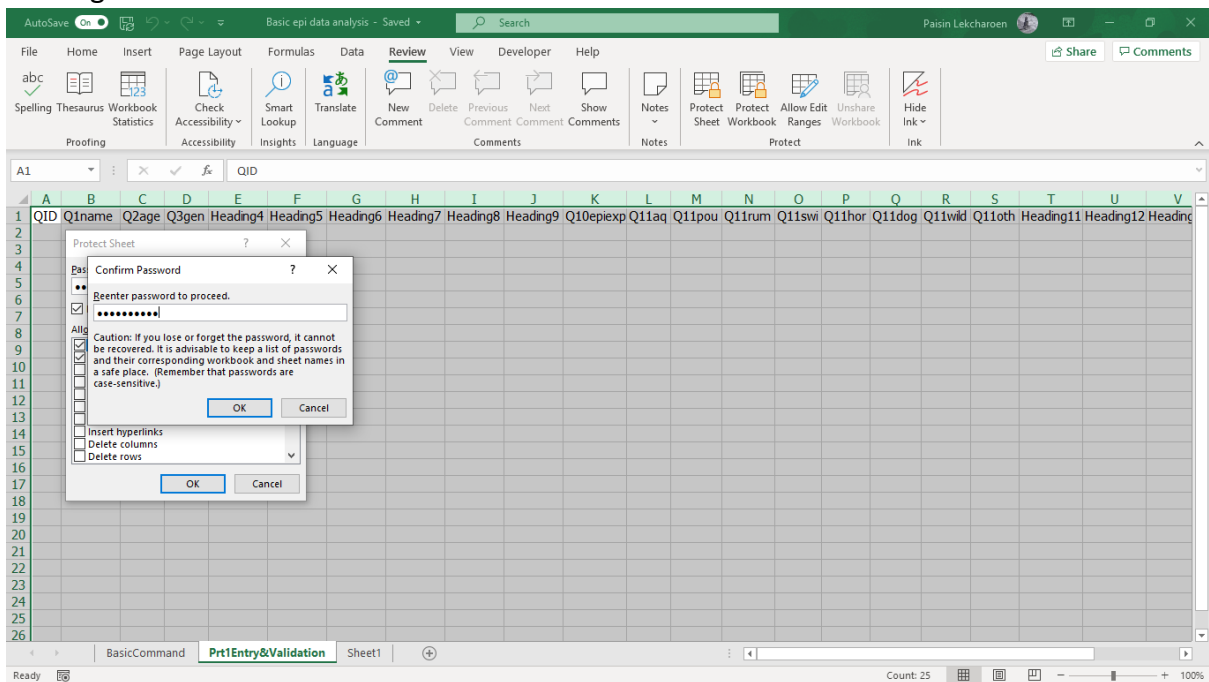
4.13. 'Allow Edit Range' window appears. Click 'Protect Sheet'.



4.14. Put a password in 'Password to unprotect sheet' box. In this case, 'rfetpv2021' is used. Leave other values as default. Then click 'OK'.



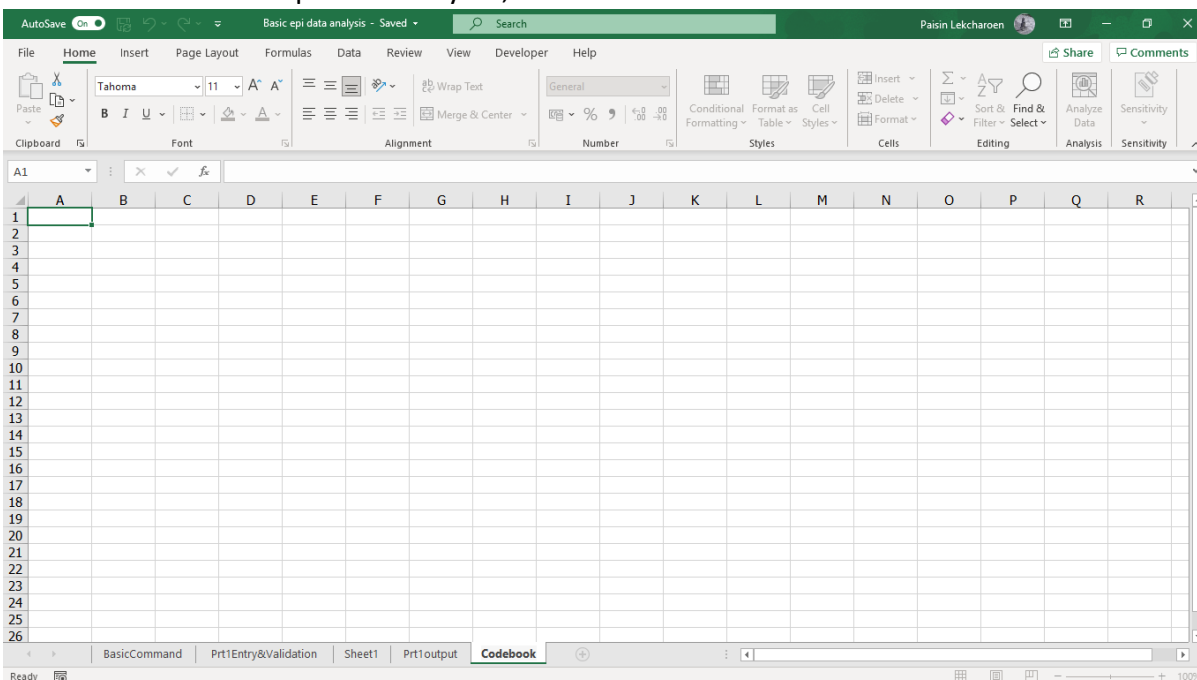
- 4.15. The program will ask you to re-confirm the password. Put the same password again and click 'OK'.



- 4.16. So, the worksheet is protected. If anyone is going to make change in this worksheet. It will be asked to unprotect the sheet by providing correct password. This is useful when you want to send a data entry form to your colleagues or when data entry is completed.
- 4.17. Now unprotect the worksheet to allow data entry. Then, enter data record from the example questionnaires.
- 4.18. An example of Exercise part 1 output is provided. See worksheet 'Prt1output'. Look around about how to head each column, how to arrange each question in a suitable format for their relevant answering, and how to define list of answering values. Compare with one you have already created.

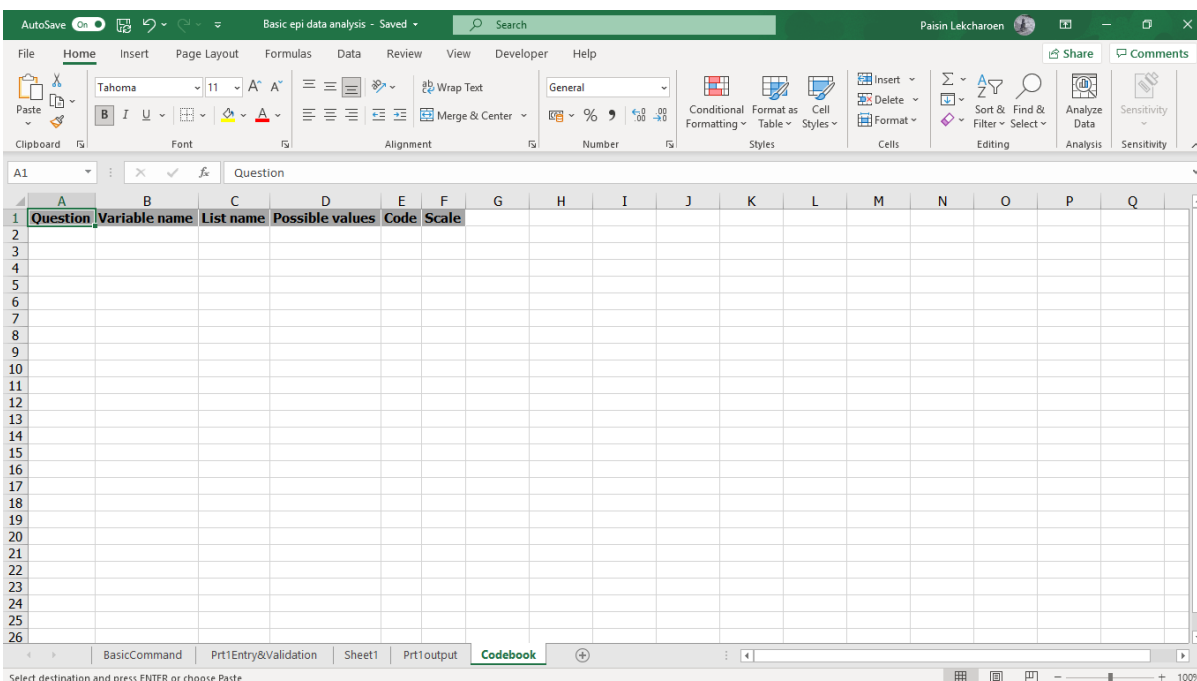
5. Generate a codebook sheet for data entry reference from the relevant questions in the questionnaire. It helps you easily keep track with the questions and answers, and allow easily interpretation of the database.

5.1. From the 'Basic epi data analysis', insert new sheet and name 'Codebook'.



5.2. Six columns will be used in the Codebook sheet including:

- 1) Question
- 2) Variable name
- 3) List name
- 4) Possible values
- 5) Code
- 6) Scale



5.3. Put information in each column according to the question in the questionnaire. Examples are shown for questions number 1 to 6.

Question	Variable name	List name	Possible values	Code	Scale
1	Q1name	Name and given name	Text		Nominal
2	Q2age	Age (years)	18-65		Ratio
			No answer	999	
3	Q3gen	Gender	Male	1	Nominal
4			Female	2	
5			Other	3	
6			No answer	9	
7	4 Q4wt	Weight (kg)	Number as indicated (>30)		Ratio
8			No answer	999	
9	5 Q5ht	Height (cm)	Number as indicated		Ratio
10			No answer	999	
11	6 Q6nat	Nationality	Bhutanese	1	Nominal
12			Burmese	2	
13			Cambodian	3	
14			Chinese	4	
15			Filipino	5	
16			Indonesian	6	
17			Laotian	7	
18			Malaysian	8	
19			Nepalese	9	
20			Thai	10	
21			Vietnamese	11	
22			Other	12	
23			No answer	99	

5.4. Continue to all questions. Make sure that you assign appropriate code and correct scale for all questions.

5.5. An example of a complete codebook relevant to 'Prt1output' is provided as 'CBKoutput' worksheet.



Part 2 Data conversion

A dataset from a questionnaire survey is provided in worksheet 'Prt2conversion'. **Thirty-two trainees** participate in this survey and provide answer of the 16-question questionnaire. Fifty variables, deriving from 16 questions, are obtained including questionnaire ID (QID). However, some variables are not acquired during the survey and need some conversions from the original dataset.

In addition, a worksheet 'CBKcplt' for a complete codebook is also provided. Use this worksheet for practicing 'Define Name' function to create lists of values for operating in 'Vlookup' function.

Functions in use:

- Text to Columns
- If
- And
- Left
- Right
- Vlookup
- Concatenate

Additional columns:

- ✓ DoB : Date of birth of the participants

At the end of this part, you should have additional column including:

- ✓ A column representing 'Date' of birth
- ✓ A column representing 'Month' of birth
- ✓ A column representing 'Year' of birth
- ✓ A column representing 'Season' of birth
- ✓ A column representing abbreviated season
- ✓ A column representing abbreviated year
- ✓ A column representing a season-year group
- ✓ A column representing body mass index
- ✓ Two columns representing body mass index interpretation
- ✓ A column indicating consumption behavior
- ✓ A column indicating exercise behavior
- ✓ A column indicating case status

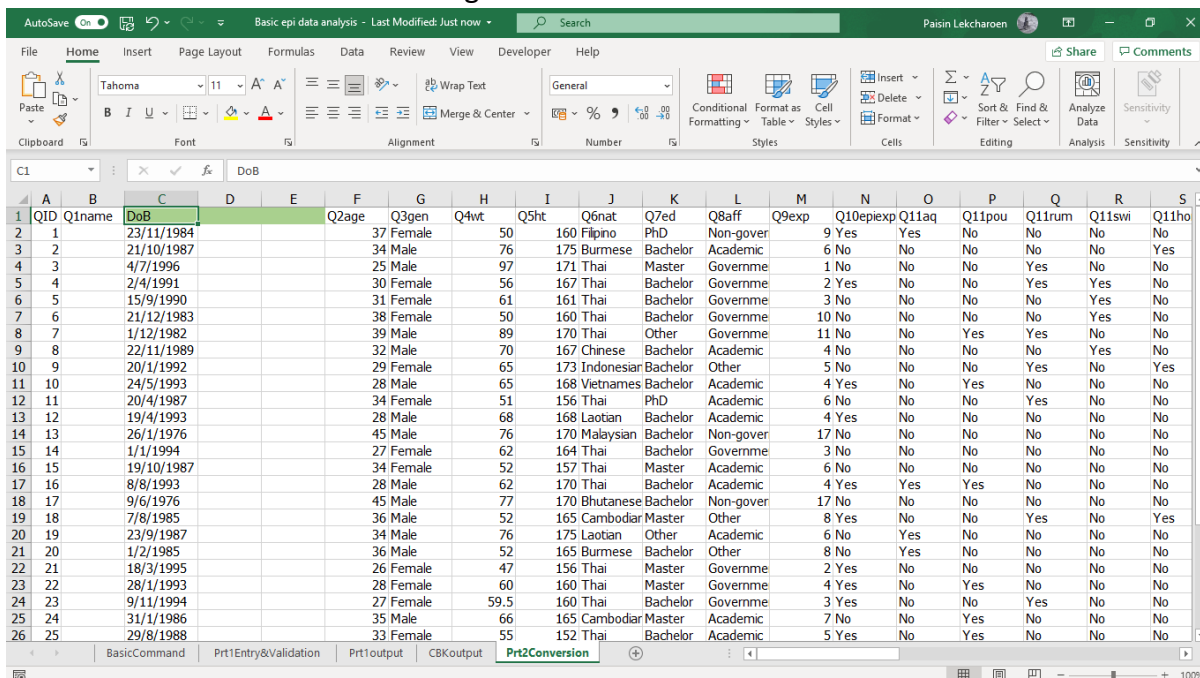
Exercise 2.1 Find a season-year (Syr) of birth for participants.

Table A. A season-year group (Syr) values.

Year of birth	Era	Season	
		Wet	Dry
1970-1979	1970s	W70	D70
1980-1989	1980s	W80	D80
1990-1999	1990s	W90	D90

Given a wet season is between May and October and a dry season is between November and April.

1. Use 'Text to columns' function to separate a column 'DoB' into three columns of 'Day', 'Month', and 'Year'.
 - 1.1. Insert 2 new columns to the right of 'DoB'.



1.2. Highlight column 'DoB'. Then, from the 'Data' ribbon, select 'Text to Columns' function.

The screenshot shows the Microsoft Excel interface with the 'Text to Columns' dialog box open. The 'DoB' column in the spreadsheet is highlighted in green. The dialog box is titled 'Text to Columns' and contains the following text:

Split a single column of text into multiple columns.

For example, you can separate a column of full names into separate first and last name columns.

You can choose how to split it up: fixed width or split at each comma, period, or other character.

The dialog box has three main sections: 'Original data type', 'Choose the file type that best describes your data', and 'Preview of selected data'. The 'Delimited' radio button is selected under 'Choose the file type that best describes your data'. The 'Preview of selected data' section shows a list of dates: 23/11/1984, 21/10/1987, 4/7/1996, 2/4/1991, 15/9/1990, 21/12/1983, 1/12/1982, 22/11/1989, 20/1/1992, 24/5/1993, 20/4/1987, 19/4/1993, 26/1/1976, 1/1/1994, 19/10/1987, 8/8/1993, 9/6/1976, 7/8/1985, 23/9/1987, 1/2/1985, 18/3/1995, 28/1/1993, 9/11/1994, 31/1/1986, 29/8/1988.

1.3. Step 1, choose 'Delimited' as an original data type, then click 'Next'.

The screenshot shows the Microsoft Excel interface with the 'Convert Text to Columns Wizard - Step 1 of 3' dialog box open. The 'Delimited' radio button is selected under 'Choose the file type that best describes your data'. The 'Next >' button is highlighted. The background shows the same spreadsheet as in the previous screenshot.

The dialog box contains the following text:

Convert Text to Columns Wizard - Step 1 of 3

The Text Wizard has determined that your data is Delimited.

If this is correct, choose Next, or choose the data type that best describes your data.

Original data type

Choose the file type that best describes your data:

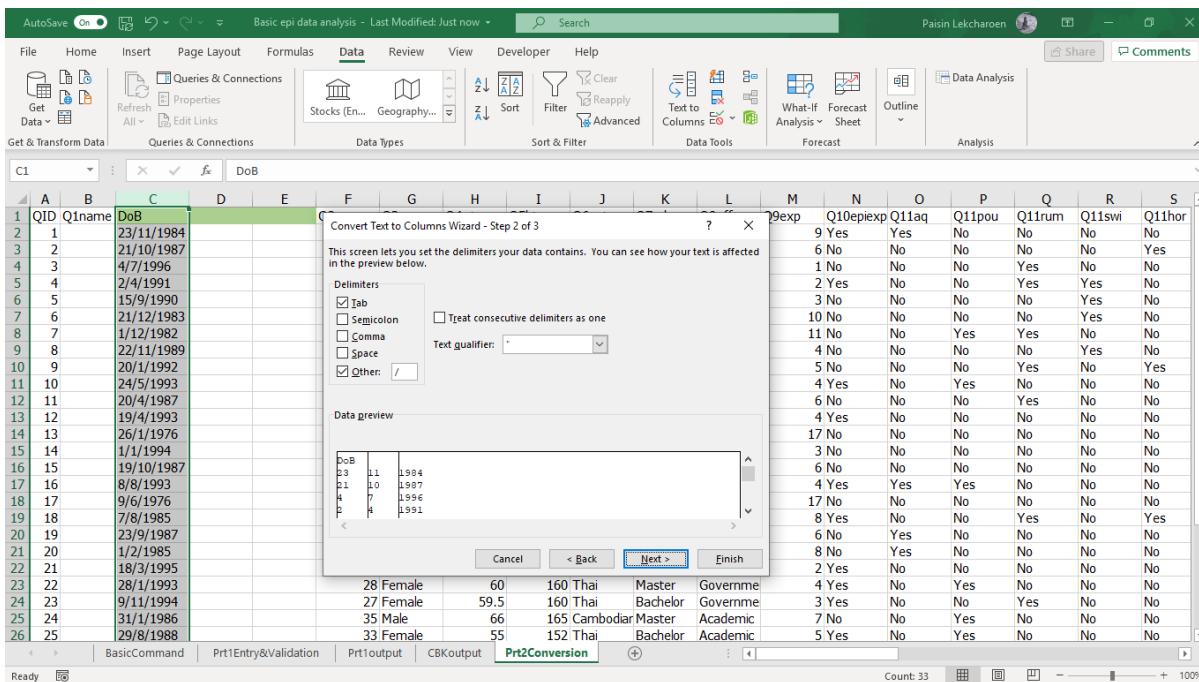
Delimited - Characters such as commas or tabs separate each field.

Fixed width - Fields are aligned in columns with spaces between each field.

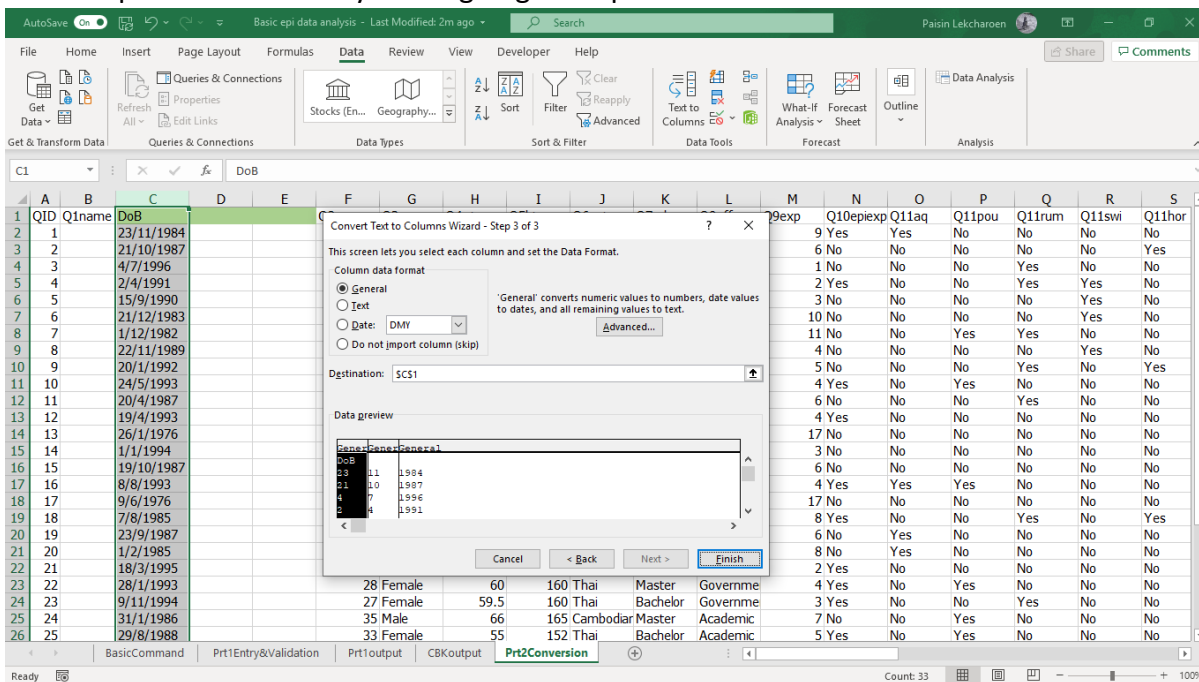
Preview of selected data:

1 DoB
2 23/11/1984
3 21/10/1987
4 4/7/1996
5 2/4/1991
6 15/9/1990
7 21/12/1983
8 1/12/1982
9 22/11/1989
10 20/1/1992
11 24/5/1993
12 20/4/1987
13 19/4/1993
14 26/1/1976
15 1/1/1994
16 19/10/1987
17 8/8/1993
18 9/6/1976
19 7/8/1985
20 23/9/1987
21 1/2/1985
22 18/3/1995
23 28/1/1993
24 9/11/1994
25 31/1/1986
26 29/8/1988

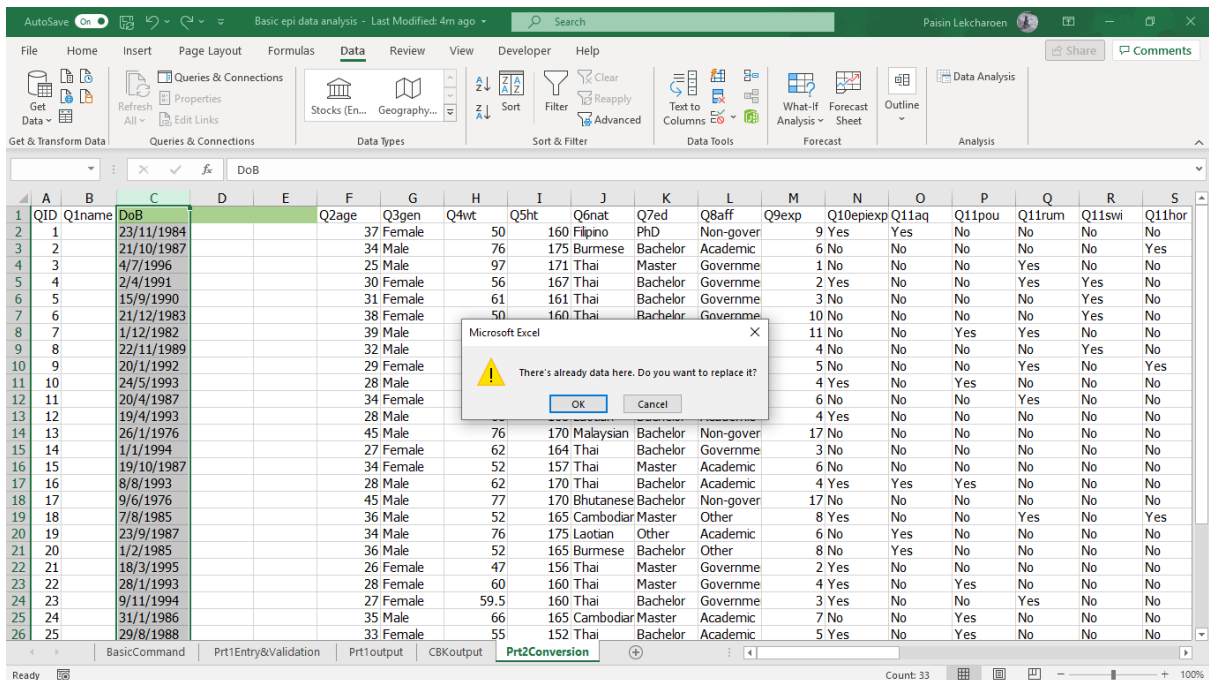
1.4. Step 2, check on 'Other' as a Delimiters. Put '/' in the space following 'Other'. Then click 'Next'.



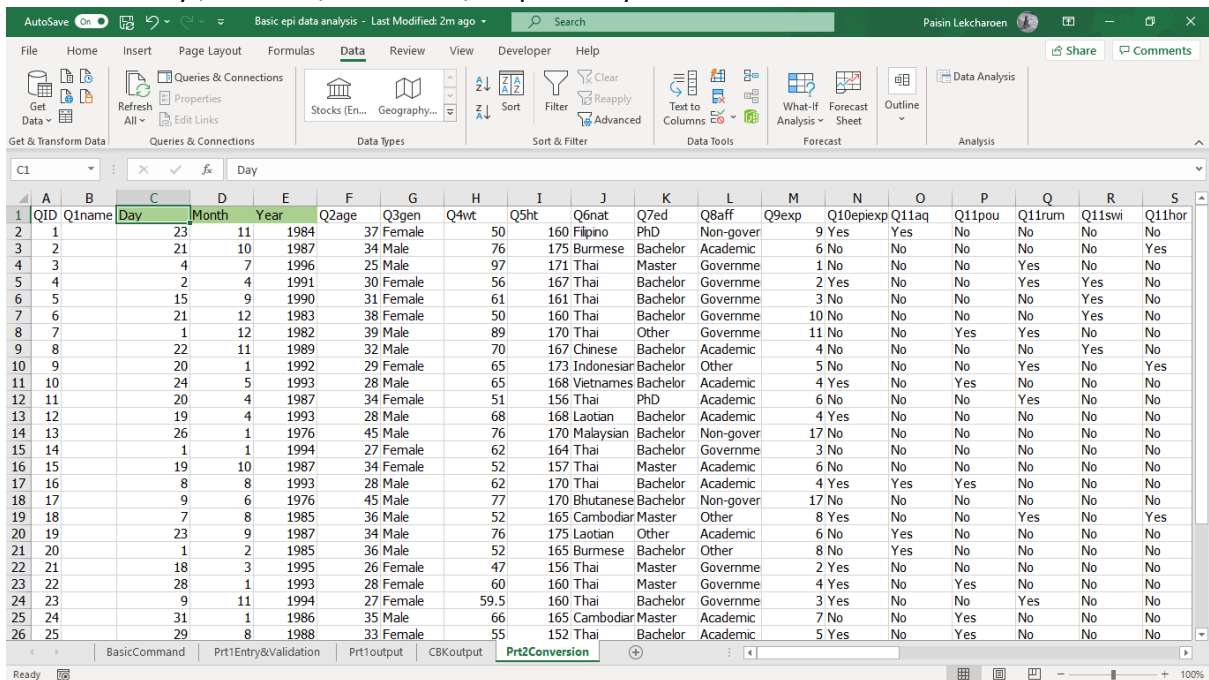
1.5. Step 3, no change is required in this step. Look through a Data preview box for the output variables that you are going to acquire. Click 'Finish'.



1.6. Windows will ask you if you want to replace existing data or not because it is going to replace data in 'DoB' column. Click 'OK'. Be aware that you provide enough space for your output otherwise your original data in the next column will be replaced and lost.



1.7. Now, you have 3 columns (1 replaced DoB and 2 new columns). Change the heading into 'Day', 'Month', and 'Year', respectively.



1.8. Insert a new column between 'Month' and 'Year'. Give its name 'Season'.

The screenshot shows an Excel spreadsheet with the following columns: QID, Q1name, Day, Month, Season, Year, Q2age, Q3gen, Q4wt, Q5ht, Q6nat, Q7ed, Q8aff, Q9exp, Q10epiexp, Q11aq, Q11pou, Q11rum, Q11swi. The 'Season' column is currently empty. The data rows contain various demographic and educational information.

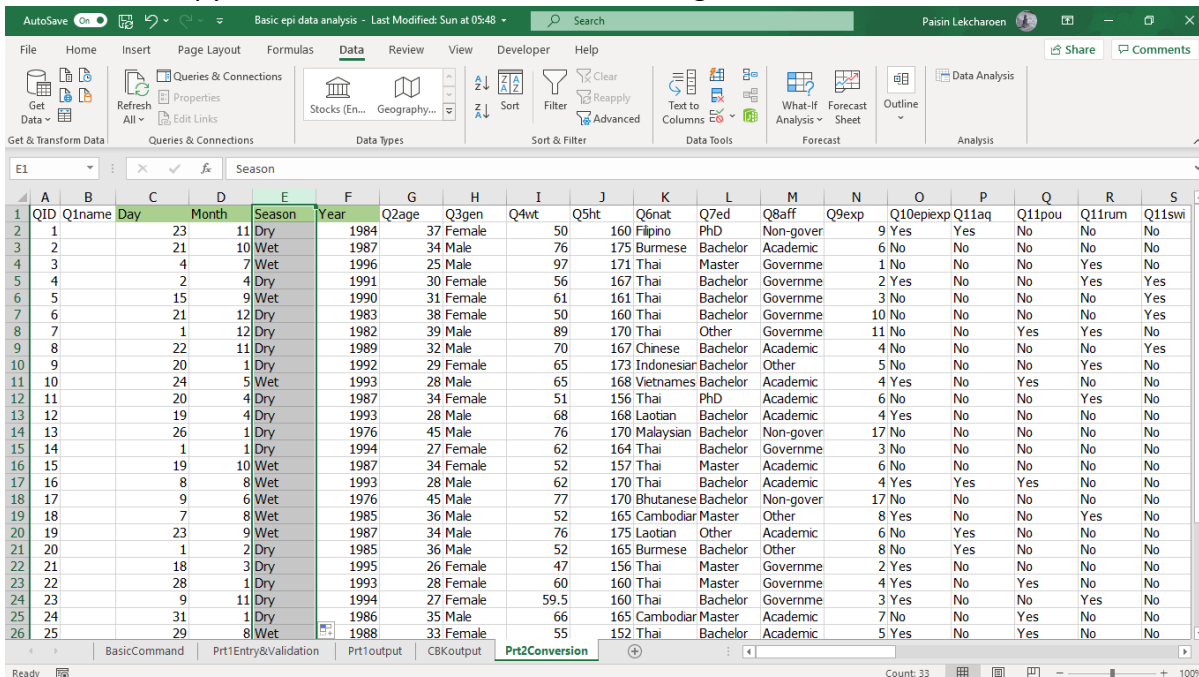
1.9. In cell E2, use 'If' function to convert 'Month' of birth into 'Season' of birth. As given above, Month 5 (May) to 10 (October) is 'Wet' season otherwise 'Dry' season. So, provide a condition for 'If' function using 'And' function so that a value between 5 to 10 is met. If the condition is true, return a value 'Wet', otherwise 'Dry'.

- “=IF (logical test, value if true, value if false)”
- “= IF (AND(D2>4, D2<11), “Wet”, “Dry”)”

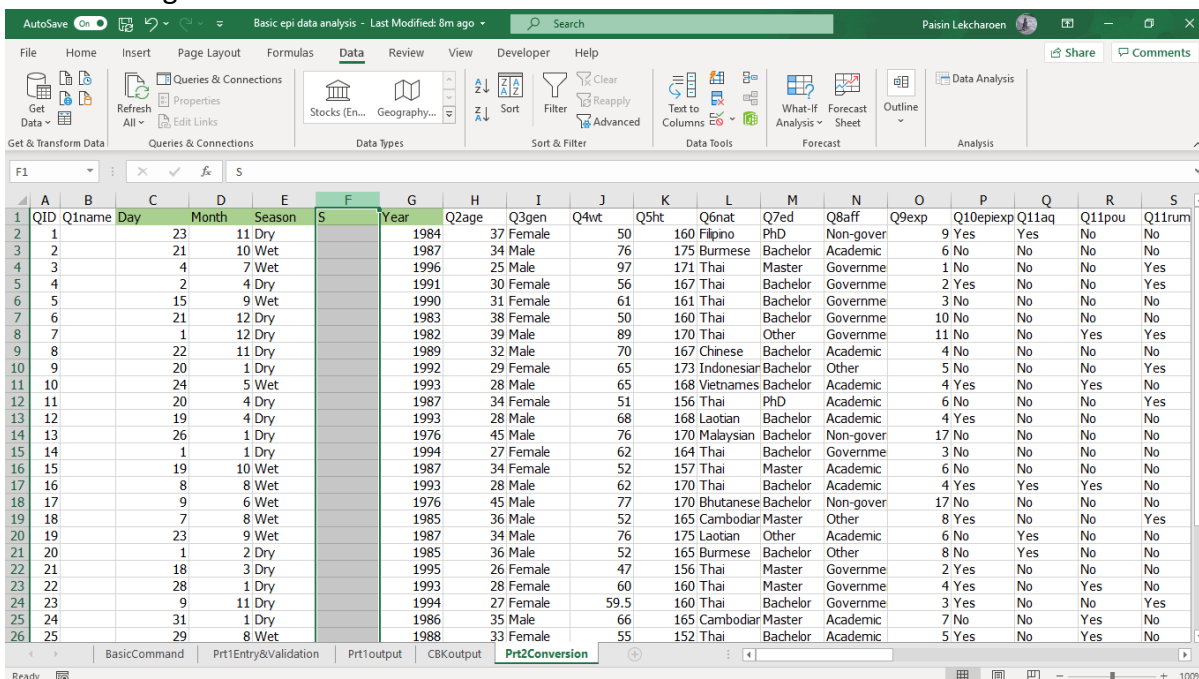
The screenshot shows the same Excel spreadsheet as in 1.8, but now with the formula `=IF(AND(D2>4,D2<11),"Wet","Dry")` entered in cell E2. The formula bar at the top displays this formula. The 'Season' column now contains 'Wet' for rows where the month is between 5 and 10, and 'Dry' otherwise.



1.10. Copy the command for all cells in the range.

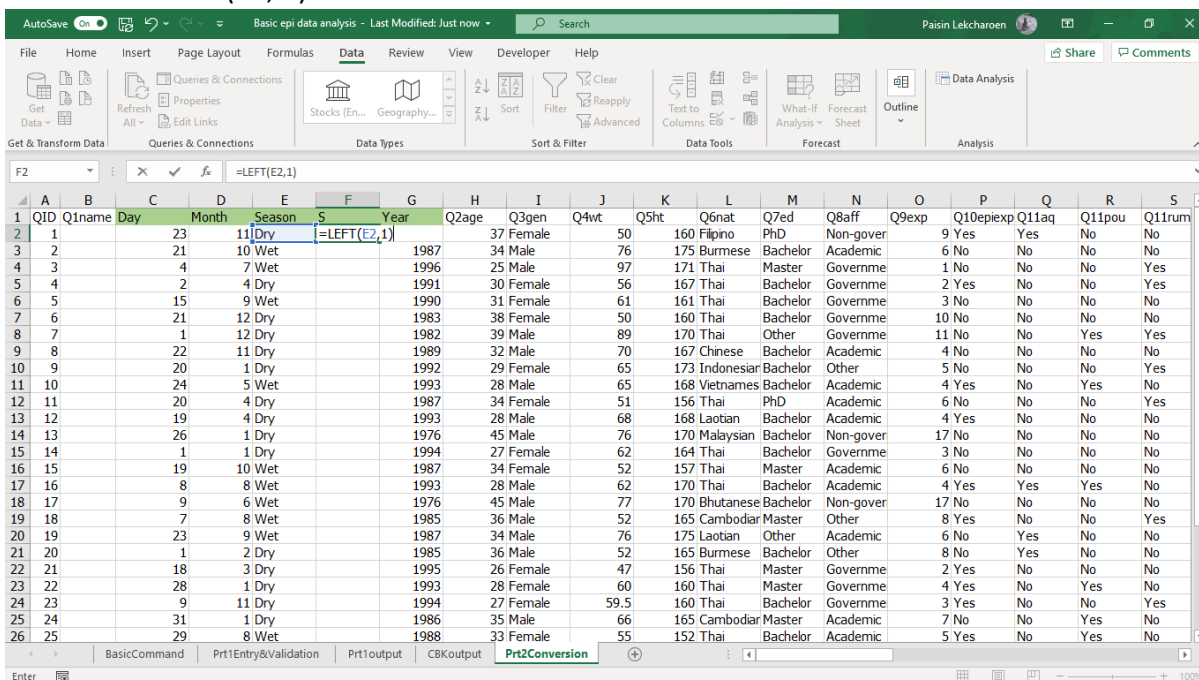


1.11. In a particular circumstance, you may like to shorten a value to a specified number of a text string, in order to make a code or an index, or abbreviate a name. At this point in the example, you would like to have only one character of 'Season': W for wet and D for Dry. Firstly, insert a column between 'Season' and 'Year'. Give a heading name 'S'.

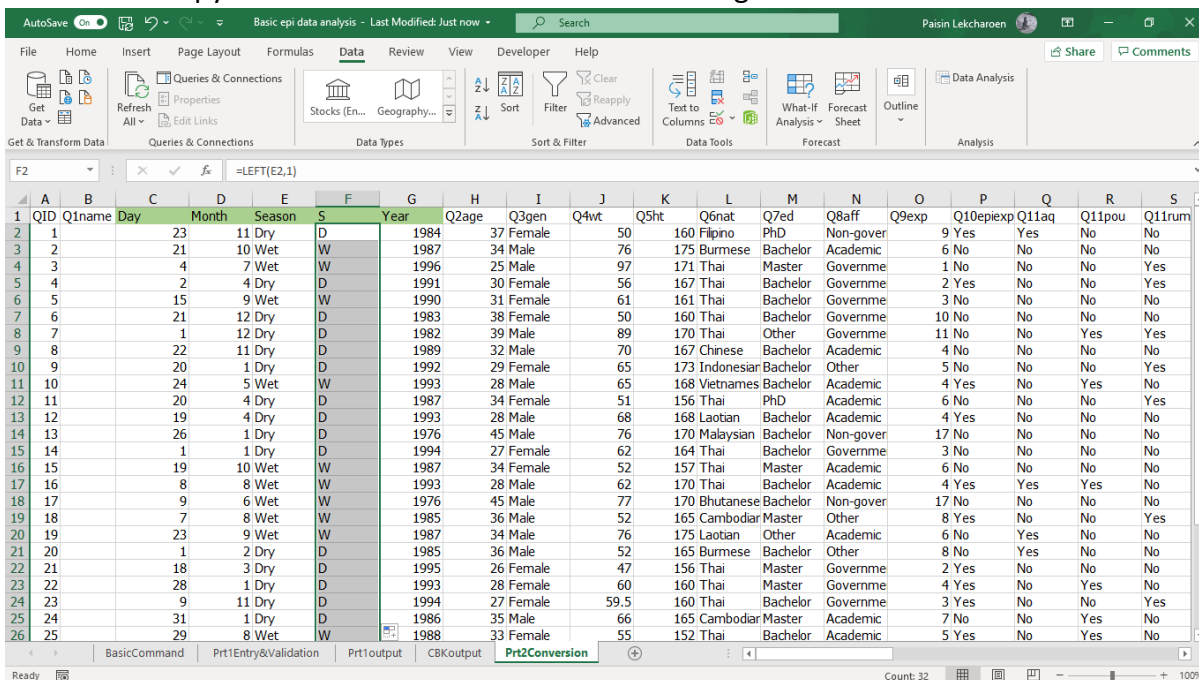


1.12. In cell F2, use function 'Left' to specify the first character from the value in cell E2.

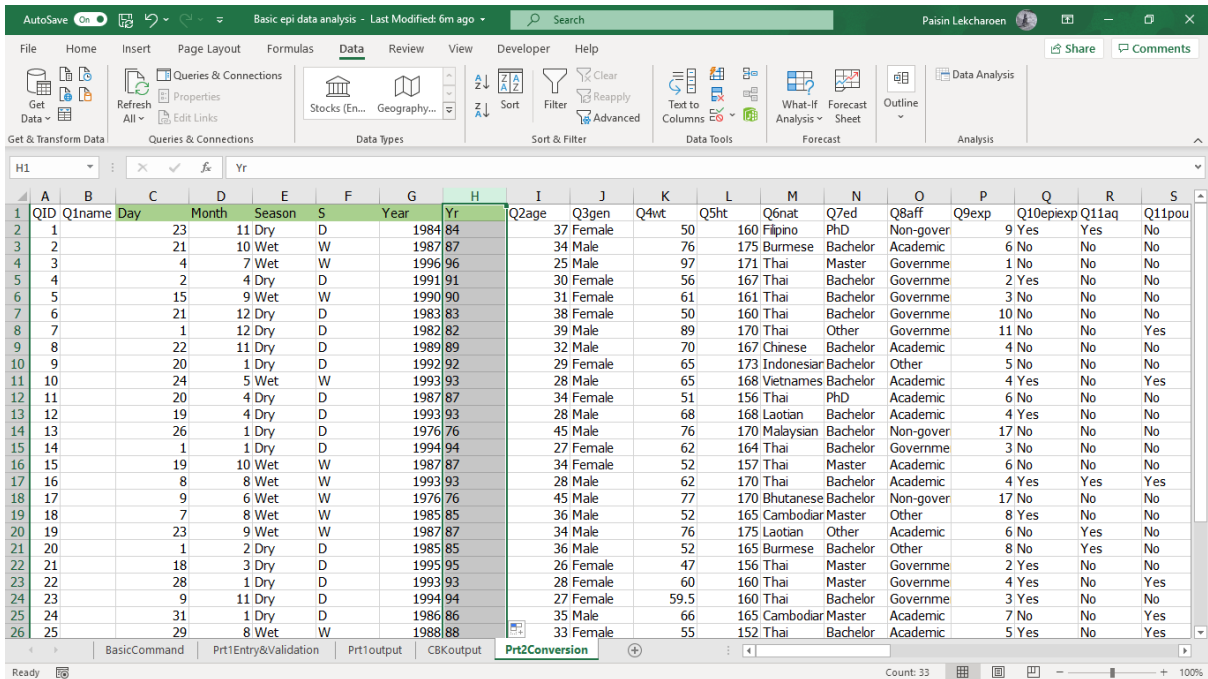
- “= LEFT (Text, Number of character)”
- “= LEFT (E2, 1)”



1.13. Copy the command for all records in the range.

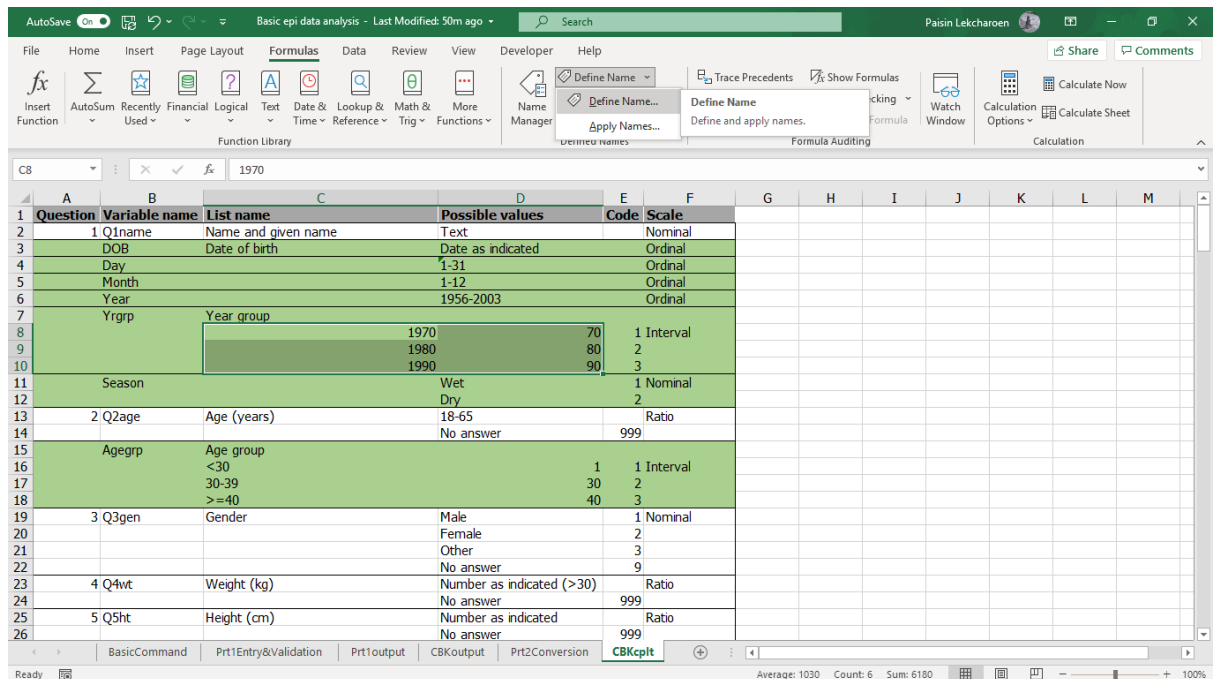


1.14. Now, you can specify 2 characters from the end of a text string in the column 'Year' and put it in the newly inserted column 'Yr' using a function 'Right'.



1.15. As in table A, year groups are divided into 3 eras: 1970s (70), 1980s (80), and 1990s (90). Insert a new column 'YrGr' for a Year Group variable.

- Define a list name 'yrgrp' using values in worksheet 'CBKcplt'. Highlight values in columns 'List name' and 'Possible values' of variable 'Yrgrp'.



- Use 'Define Name' function to create a list and name it 'yrgrp'. Click 'OK'.

The screenshot shows the 'Define Name' dialog box in Excel. The 'Name' field is set to 'yrgrp'. The 'Refers to' field is set to '=CBKcpt!\$C\$8:\$D\$10'. The background table has the following data:

Question	Variable name	List name	Possible values	Code	Scale
1	Yrgrp	Year group	1970 1980 1990	70 80 90	1 Interval 2 3
2	Q2age	Age (years)			1 Nominal 2 999
3	Agegrp	Age group	<30 30-39 >=40		1 Interval 2 3
4	Q3gen	Gender			1 Nominal 2 3 9
5	Q4wt	Weight (kg)			Ratio No answer 999
6	Q5ht	Height (cm)			Ratio No answer 999
7	Q6nat	Nationality	Bhutanese Burmese Cambodian Chinese Filipino		1 Nominal 2 3 4 5

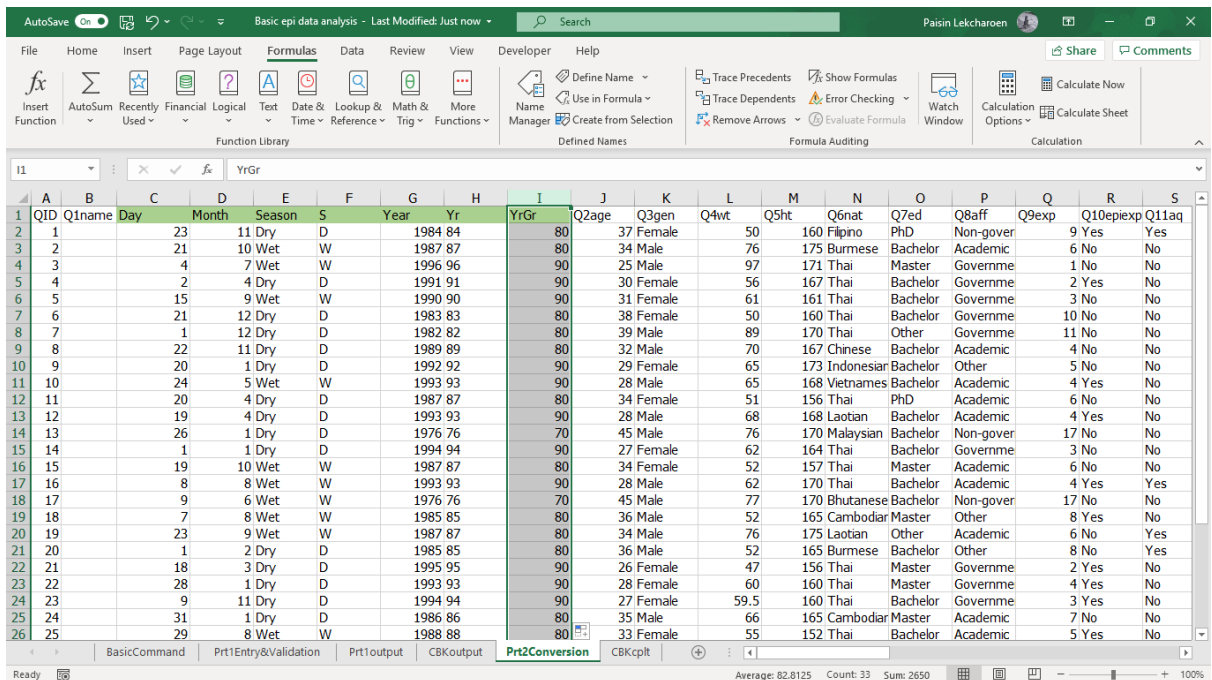
- Use 'Vlookup' function to look for a 'Yr' value in the table yrgrp and return a 'YrGr' value. Firstly, insert a new column and name it 'YrGr'. Then put a command:
 - “= VLOOKUP (look up value, table array, column index number, range look up)
 - “= VLOOKUP (G2, yrgrp, 2)”, so you are going to look for a value 1984 (as in cell G2) in the table 'yrgrp' (that you have already created) then return a value in column 2 in the same row of that table.

The screenshot shows the 'VLOOKUP' formula being applied in cell I2. The formula bar shows '=VLOOKUP(G2, yrgrp, 2)'. The background table is a detailed dataset with columns: QID, Q1name, Day, Month, Season, S, Year, Yr, YrGr, Q2age, Q3gen, Q4wt, Q5ht, Q6nat, Q7ed, Q8aff, Q9exp, Q10epiexp, Q11aq.

QID	Q1name	Day	Month	Season	S	Year	Yr	YrGr	Q2age	Q3gen	Q4wt	Q5ht	Q6nat	Q7ed	Q8aff	Q9exp	Q10epiexp	Q11aq	
2	1	23	11	Dry	D	1984	84	=VLOOKUP(G2, yrgrp, 2)			50	160	160	Phil	Non-gover	9	Yes	Yes	
3	2	21	10	Wet	W	1987	87							Bachelor	Academic	6	No	No	
4	3	4	7	Wet	W	1996	96			25	Male	97	171	Thai	Master	Government	1	No	No
5	4	2	4	Dry	D	1991	91			30	Female	56	167	Thai	Bachelor	Government	2	Yes	No
6	5	15	9	Wet	W	1990	90			31	Female	61	161	Thai	Bachelor	Government	3	No	No
7	6	21	12	Dry	D	1983	83			38	Female	50	160	Thai	Bachelor	Government	10	No	No
8	7	1	12	Dry	D	1982	82			39	Male	89	170	Thai	Other	Government	11	No	No
9	8	22	11	Dry	D	1989	89			32	Male	70	167	Chinese	Bachelor	Academic	4	No	No
10	9	20	1	Dry	D	1992	92			29	Female	65	173	Indonesian	Bachelor	Other	5	No	No
11	10	24	5	Wet	W	1993	93			28	Male	65	168	Vietnamese	Bachelor	Academic	4	Yes	No
12	11	20	4	Dry	D	1987	87			34	Female	51	156	Thai	PhD	Academic	6	No	No
13	12	19	4	Dry	D	1993	93			28	Male	68	168	Laotian	Bachelor	Academic	4	Yes	No
14	13	26	1	Dry	D	1976	76			45	Male	76	170	Malaysian	Bachelor	Non-gover	17	No	No
15	14	1	1	Dry	D	1994	94			27	Female	62	164	Thai	Bachelor	Government	3	No	No
16	15	19	10	Wet	W	1987	87			34	Female	52	157	Thai	Master	Academic	6	No	No
17	16	8	8	Wet	W	1993	93			28	Male	62	170	Thai	Bachelor	Academic	4	Yes	Yes
18	17	9	6	Wet	W	1976	76			45	Male	77	170	Bhutanese	Bachelor	Non-gover	17	No	No
19	18	7	8	Wet	W	1985	85			36	Male	52	165	Cambodian	Master	Other	8	Yes	No
20	19	23	9	Wet	W	1987	87			34	Male	76	175	Laotian	Other	Academic	6	No	Yes
21	20	1	2	Dry	D	1985	85			36	Male	52	165	Burmese	Bachelor	Other	8	No	Yes
22	21	18	3	Dry	D	1995	95			26	Female	47	156	Thai	Master	Government	2	Yes	No
23	22	28	1	Dry	D	1993	93			28	Female	60	160	Thai	Master	Government	4	Yes	No
24	23	9	11	Dry	D	1994	94			27	Female	59.5	160	Thai	Bachelor	Government	3	Yes	No
25	24	31	1	Dry	D	1986	86			35	Male	66	165	Cambodian	Master	Academic	7	No	No
26	25	29	8	Wet	W	1988	88			33	Female	55	152	Thai	Bachelor	Academic	5	Yes	No

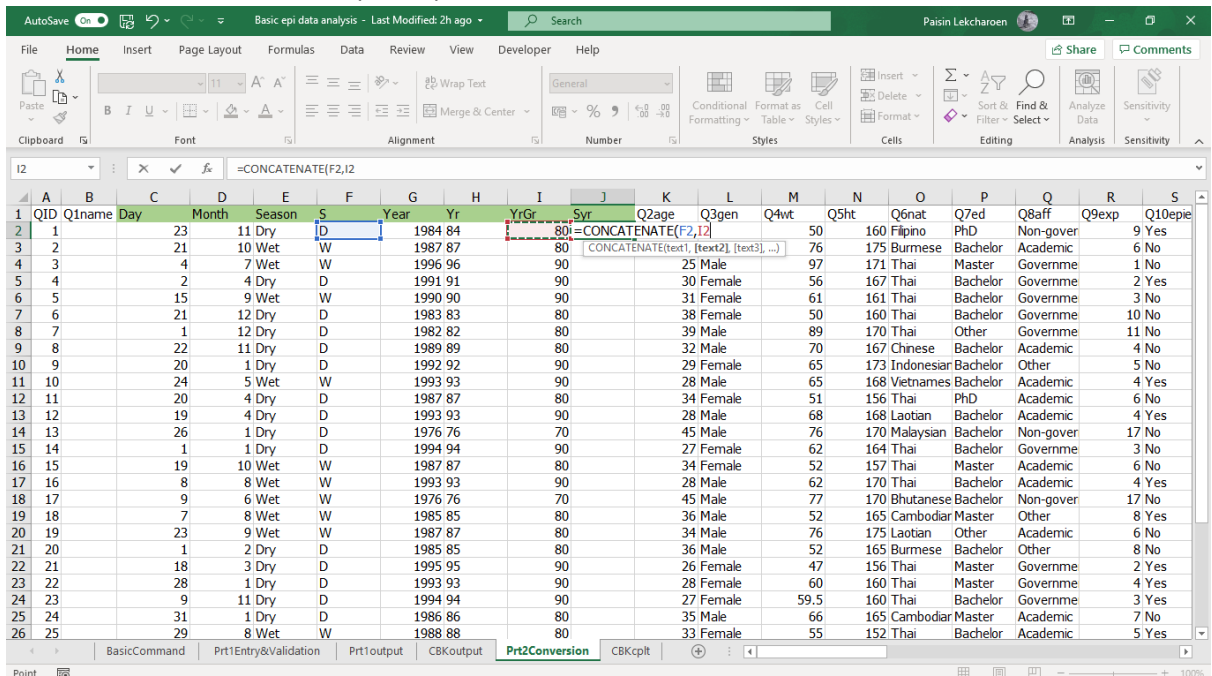


- Copy the command for all records.



1.16. Insert a new column 'Syr' for season-year variable. Use 'Concatenate' function to combine text strings from 'S' and 'YrGr' columns. Put a command in cell J2:

- “= CONCATENATE (Text1, Text2, ...)
- “= CONCATENATE (F2, I2)”



1.17. Copy the command to all records.



Exercise 2.2 Calculate BMI for participants.

Calculate body mass index (BMI) and classify BMI condition for all participants.

$$BMI = \frac{\text{Body weight (kg)}}{\text{Height (m)}^2}$$

Table B. BMI Class

BMI	Class
< 18.50	Underweight
18.50 – 22.90	Normal
23.0 – 24.90	Overweight
> 25.00	Obesity

1. Calculate BMI using the above equation in new column 'BMI'.
2. In order to classify BMI class, you can manipulate it by 2 methods:
 - 2.1. Use 'Vlookup' function (thus, you need to create and define name of 'a table of listed variables' in the codebook worksheet first), or
 - 2.2. Use 'If' function to indicate a proper condition.

Try to operate it by both methods and compare the results. Put the output from 2.1 in a new column 'BMI_con1' and those from 2.2 in another new column 'BMI_con2'.

Exercise 2.3 Define consumption and exercise behaviors, and case condition of all participants.

You observe that consumption and exercising behavior could associate with BMI condition of participants. Eating habit and exercising regularity are provided in 'Q13eat' and 'Q13exer', respectively. However, they were collected in ordinal 'Likert' scale. You would like to convert those data into 2 categories which are 'Normal' and 'Eat' groups for consumption behavior, and 'Exercise' and 'Not exercise' for exercising behavior.

- The participants who often or always eat are considered as having 'Eat' behavior, the rest are 'Normal'.
- The participants who often or always have physical exercise are considered as having 'Exercise' behavior and the rest are 'Not exercise'.

Then, you would like to define a case status using this condition:

- The participants who have 'Overweight' or 'Obesity' BMI class are considered as 'Case', the rest are 'Not case'.

Use 'If', 'Or' or 'And' function to manage data conversion.

Place outputs for consumption behavior, exercising behavior, and case status in new columns 'Eat', 'Exe', and 'Case', respectively.

	AV	AW	AX	AY	AZ	BA	BB	BC	BD	BE	BF	BG	BH	BI	BJ	BK	BL	BM
1	Q14splss	Q14djar	Q14fev	Q14coug	Q14soto	Q14vom	Q14uri	Q14rnj	Q14oth	Q15onset	Q16getsick	BMI	BMI_con1	BMI_con2	Eat	Exe	Case	
2											Other							
3	Yes	No	No	No	No	No	No	No	Yes		Visit pharmacy							
4											Rest at home							
5											Other							
6	No	No	No	No	No	No	No	No	No	44298	Visit clinic/hospital							
7											Visit pharmacy							
8	No	No	No	Yes	No	No	No	No	No	44300	Other							
9											Visit pharmacy							
10											Visit clinic/hospital							
11											Other							
12	No	No	No	No	No	No	No	No	Yes	44310	Visit clinic/hospital							
13											Visit clinic/hospital							
14	No	No	No	No	No	No	No	Yes	No	44309	Visit clinic/hospital							
15											Rest at home							
16											Visit clinic/hospital							
17											Visit clinic/hospital							
18											Visit clinic/hospital							
19											Rest at home							
20											Visit clinic/hospital							
21											Rest at home							
22											Visit pharmacy							
23											Other							
24											Rest at home							
25	No	No	Yes	Yes	Yes	No	No	No	No	44302	Other							
26											Rest at home							

*** Answers of command for 'BMI', 'BMI_con1', 'BMI_con2', 'Eat', 'Exe', and 'Case' will be sent to those who complete all exercise in the following week.



Exercise 2.4 Vlookup function

In some circumstances, needed data are separated and stored in two different databases. See worksheet 'Vlookup'. There are 2 databases. The first table comprises population data by district. The second table reports number of cases and deaths by district.

You would like to combine these 2 databases into a combined table.

1. The first 3 columns of combined data table are duplicated from health data table. Last 3 columns are empty and need to be retrieved from population data table.

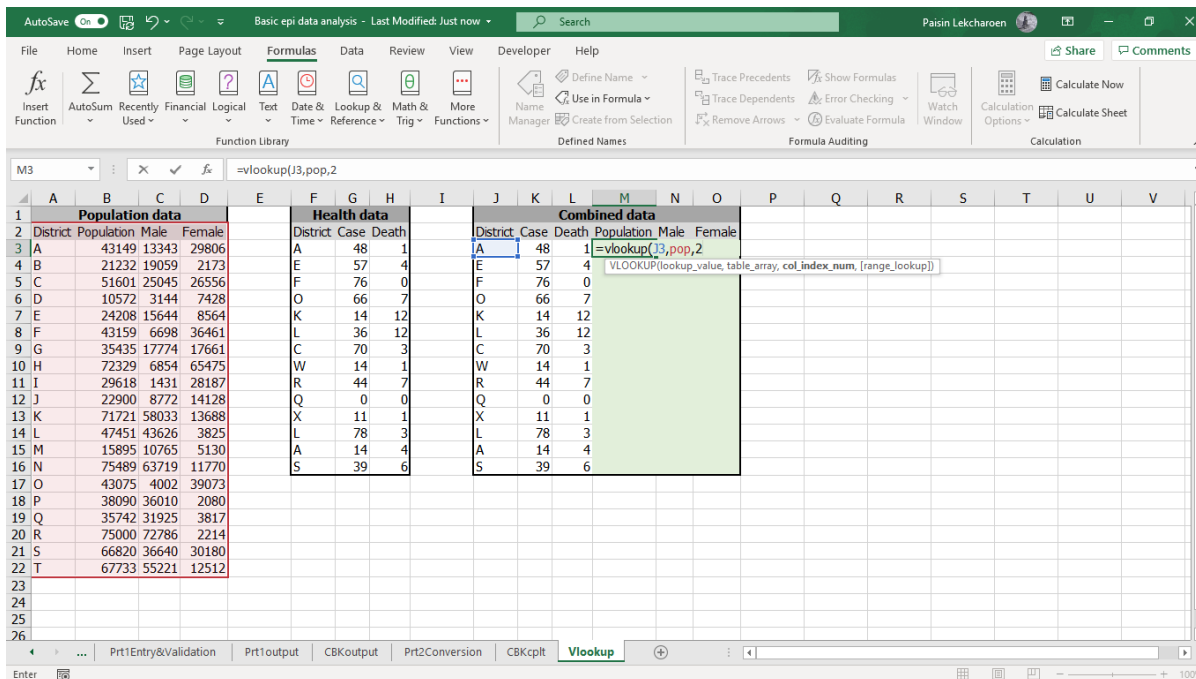
Population data				Health data			Combined data					
District	Population	Male	Female	District	Case	Death	District	Case	Death	Population	Male	Female
A	43149	13343	29806	A	48	1	A	48	1			
B	21232	19059	2173	E	57	4	E	57	4			
C	51601	25045	26556	F	76	0	F	76	0			
D	10572	3144	7428	O	66	7	O	66	7			
E	24208	15644	8564	K	14	12	K	14	12			
F	43159	6698	36461	L	36	12	L	36	12			
G	35435	17774	17661	C	70	3	C	70	3			
H	72329	6854	65475	W	14	1	W	14	1			
I	29618	1431	28187	R	44	7	R	44	7			
J	22900	8772	14128	Q	0	0	Q	0	0			
K	71721	58033	13688	X	11	1	X	11	1			
L	47451	43626	3825	L	78	3	L	78	3			
M	15895	10765	5130	A	14	4	A	14	4			
N	75489	63719	11770	S	39	6	S	39	6			
O	43075	4002	39073									
P	38090	36010	2080									
Q	35742	31925	3817									
R	75000	72786	2214									
S	66820	36640	30180									
T	67733	55221	12512									

2. Use 'Define Name' to create a list of value table of population data. Give the table name 'pop'.

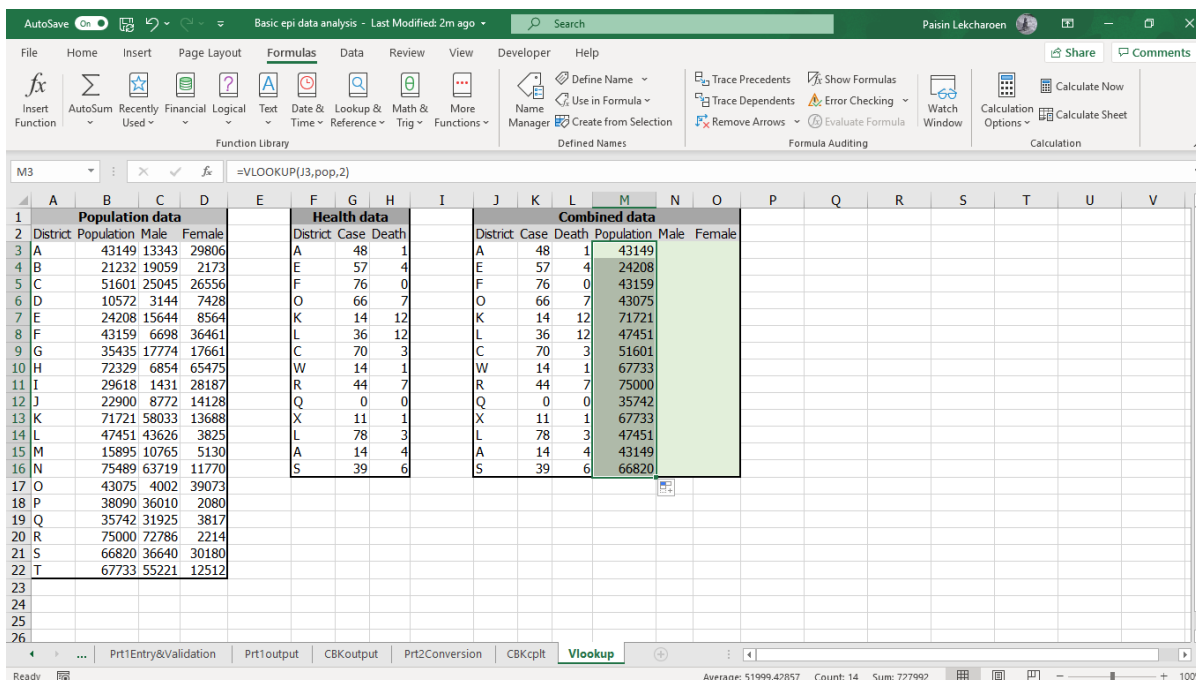
Population data				Health data			Combined data					
District	Population	Male	Female	District	Case	Death	District	Case	Death	Population	Male	Female
A	43149	13343	29806	A	48	1	A	48	1			
B	21232	19059	2173	E	57	4	E	57	4			
C	51601	25045	26556	F	76	0	F	76	0			
D	10572	3144	7428	O	66	7	O	66	7			
E	24208	15644	8564	K	14	12	K	14	12			
F	43159	6698	36461	L	36	12	L	36	12			
G	35435	17774	17661	C	70	3	C	70	3			
H	72329	6854	65475	W	14	1	W	14	1			
I	29618	1431	28187	R	44	7	R	44	7			
J	22900	8772	14128	Q	0	0	Q	0	0			
K	71721	58033	13688	X	11	1	X	11	1			
L	47451	43626	3825	L	78	3	L	78	3			
M	15895	10765	5130	A	14	4	A	14	4			
N	75489	63719	11770	S	39	6	S	39	6			
O	43075	4002	39073									
P	38090	36010	2080									
Q	35742	31925	3817									
R	75000	72786	2214									
S	66820	36640	30180									
T	67733	55221	12512									

- These 2 databases can be linked via 'District' variable. So, in cell M3 that you need population data, use 'Vlookup' function to look for the 'District' value in its own table (J3) and targeted defined table (population table 'pop', column 1). Ask it to return the value from column 2. Thus:

➤ “= VLOOKUP (J3, pop, 2)”



Then copy the command for all records.



- Look for 'Male' and 'Female' values and put them in the combined data table.



Part 3 Data cleaning and recoding

After a data entry procedure, the database should be cleaned before operating data analysis. Some variables may be recoded into a suitable value that is compatible with the software being used in analytical processes.

During a data entry process, questionnaires (of different questionnaire ID) and data entry form prepared in Part 1 may be distributed to many team members in order to reduce time spent at this step. In the steps of entering and combining data, several problems and erroneous data entering may occur.

Now you receive a combined database from your team member as you see in worksheet 'Prt3clean'. This database contains several errors. Try to find and correct those errors!

Functions in use:

- Filter
- Data Validation
- Circle Invalid Data
- Pivot Table
- Conditional Formatting
- Remove Duplicates
- Define Name
- Vlookup
- Find and Replace*

Exercise 3.1 Filter

1. Use 'Filter' function to find erroneous data entry. Go to 'Data' ribbon and select 'Filter' function. Make sure the active cell is any cells on the first row.

The screenshot shows the Microsoft Excel interface with the 'Data' ribbon selected. The 'Filter' button is highlighted, and its dropdown menu is open. The menu options include 'Turn on filtering for the selected cells.' and 'Then, click the arrow in the column header to narrow down the data.' The spreadsheet below shows a table with columns for QID, QIname, Day, Month, Season, Year, S, Yr, Yrgrp, Sjr, Q2age, Q3gen, Q4wt, Q5i, Q9exp, Q10epiexp, Q11aq, Q11pou, Q11rum, Q11swi, Q11hor, and Q11dog. The table contains data for various questionnaires and their characteristics.

Q1	Q1nan	D	Mon	Seasi	Ye	Yrg	Syr	Q2ai	Q3ge	Q4	Q5	Q6nat	Q7ed	Q8aff	Q9ei	Q10epie	Q11	Q11p	Q11ru	Q11s	Q11h	Q11d	
1	1	23	11	Dry	19						50	160	Filpno	PHD	Non-government	9	Yes	Yes	No	No	No	No	No
2	1	21	10	Wet	19						76	175	Burmese	Bachelor	Academic	6	No	No	No	No	No	Yes	Yes
3	2	4	7	Wet	19						97	171	Thai	Master	Government	1	No	No	No	Yes	No	No	No
4	3	4	4	Dry	19						56	167	Thai	Bachelor	Government	2	Yes	No	No	Yes	Yes	No	No
5	4	2	4	Dry	19						61	161	Thai	Bachelor	Government	3	No	No	No	No	Yes	No	No
6	5	15	9	Wet	19						50	160	Thai	Bachelor	Government	10	No	No	No	No	Yes	No	No
7	6	21	12	Dry	19						170	89	Thai	Other	Government	31	No	No	Yes	Yes	No	No	No
8	7	1	13	Dry	19						170	89	Thai	Other	Government	31	No	No	Yes	Yes	No	No	No
9	7	1	13	Dry	19						70	167	Chinese	Bachelor	Academic	4	No	No	No	No	Yes	No	No
10	8	22	11	Dry	19						65	173	Indonesian	Bachelor	Other	5	No	No	No	No	Yes	No	Yes
11	9	20	1	Dry	19						65	168	Vietnamese	Bachelor	Academic	4	Yes	No	Yes	No	No	No	No
12	10	24	5	Wet	19						51	156	Thai	PhD	Academic	6	No	No	No	Yes	No	No	No
13	11	20	4	Dry	19						68	168	Laotian	Bachelor	Academic	4	Yes	No	No	No	No	No	Yes
14	12	19	4	Dry	19						76	170	Malaysian	Bachelor	Non-government	17	No	No	No	No	No	No	No
15	13	26	1	Dry	19						76	170	Malaysian	Bachelor	Non-government	17	No	No	No	No	No	No	No
16	13	26	1	Dry	19						62	164	Thai	Bachelor	Government	3	No	No	No	No	No	No	No
17	14	1	1	Dry	19						52	157	Thai	Master	Academic	6	No	No	No	No	No	No	Yes
18	15	19	10	Wet	19						62	170	Thailand	Bachelor	Academic	4	Yes	Yes	Yes	No	No	No	No
19	16	8	8	Wet	19						77	170	Bhutanese	Bachelor	Non-government	17	No	No	No	No	No	No	No
20	17	9	6	Wet	19						52	165	Cambodian	Master	Other	8	Yes	No	No	Yes	No	Yes	No
21	18	7	8	Wet	19						76	175	Laotian	Other	Academic	6	No	Yes	No	No	No	No	No
22	19	23	9	Wet	19						76	175	Laotian	Other	Academic	6	No	Yes	No	No	No	No	No
23	19	23	9	Wet	19						52	165	Burmese	Bachelor	Other	8	No	Yes	No	No	No	No	No
24	20	31	2	Dry	19						17	156	Thai	Master	Government	2	Yes	No	No	No	No	No	No
25	21	18	3	Dry	19	1995	D	95	90	D90	26	Female											
26	22	18	1	Dry	19	1993	D	93	90	D90	28	Female											

- 3.2.2. Season is Wet or Dry.
- 3.2.3. Year or Working Experience should be relevant with Age.
- 3.2.4. Following answer for a sub-question or that emerged when you separate a multiple-answer question into multiple single questions, should not break a condition of the proxy main question. For example, if you say 'No' for a question asking whether you are sick or not, you should have 'No' for all following questions asking about clinical symptoms.

AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB	BC	BD	BE	BF	BG	BH	BI	
1	Q13cook	Q13medi	Q13tak	Q13chat	Q14sick	Q14spl	Q14di	Q14fr	Q14col	Q14so	Q14vo	Q14i	Q14d	Q14o	Q15ons	Q16getsick	BMI	BMI_cond	BMI_cond
2	Never	Often	Often	Often	No										Other	19.531	Normal	Normal	
4	Rarely	Rarely	Often	Sometimes	No										Rest at home	33.173	Obesity	Obesity	
5	Never	Never	Rarely	Always	No										Other	20.08	Normal	Normal	
7	Rarely	Sometimes	Sometimes	Sometimes	No										Visit pharmacy	19.531	Normal	Normal	
10	Often	Sometimes	Often	Often	No										Visit pharmacy	25.1	Obesity	Obesity	
11	Never	Never	Rarely	Sometimes	No										Visit clinic/hospital	21.718	Normal	Normal	
12	Rarely	Sometimes	Sometimes	Always	No										Other	23.03	Overweight	Overweight	
14	Never	Sometimes	Sometimes	Sometimes	No										Visit clinic/hospital	24.093	Overweight	Overweight	
17	Sometimes	Rarely	Rarely	Sometimes	No										Rest at home	23.052	Overweight	Overweight	
18	Often	Rarely	Sometimes	Often	No										Visit clinic/hospital	21.096	Normal	Normal	
20	Rarely	Sometimes	Sometimes	Never	No										Visit clinic/hospital	21.453	Normal	Normal	
21	Rarely	Rarely	Sometimes	Sometimes	No										Visit clinic/hospital	26.644	Obesity	Obesity	
22	Rarely	Sometimes	Often	Sometimes	No										Rest at home	19.1	Normal	Normal	
23	Never	Never	Rarely	Sometimes	No										Visit clinic/hospital	24.816	Overweight	Overweight	
24	Never	Never	Rarely	Sometimes	No										Visit clinic/hospital	24.816	Overweight	Overweight	
24	Rarely	Sometimes	Often	Sometimes	No										Rest at home	19.1	Normal	Normal	
25	Sometimes	Sometimes	Always	Always	No										Visit pharmacy	6.9855	Underweight	Underweight	
26	Rarely	Rarely	Sometimes	Sometimes	No										Other	8.8757	Underweight	Underweight	
27	Never	Never	Sometimes	Rarely	No										Rest at home	23.242	Overweight	Overweight	
29	Never	Never	Sometimes	Rarely	No										Rest at home	23.805	Overweight	Overweight	
30	Never	Never	Sometimes	Rarely	No										Rest at home	23.805	Overweight	Overweight	
32	Never	Sometimes	Sometimes	Often	No										Other	26.235	Obesity	Obesity	
33	Rarely	Never	Sometimes	Always	No										Other	24.533	Overweight	Overweight	
35	Sometimes	Sometimes	Often	Often	No								44314		Visit clinic/hospital	30.071	Obesity	Obesity	
36	Sometimes	Sometimes	Often	Often	No										Visit clinic/hospital	30.071	Obesity	Obesity	

Refer to the codebook or questionnaire. Check for any compatible values.



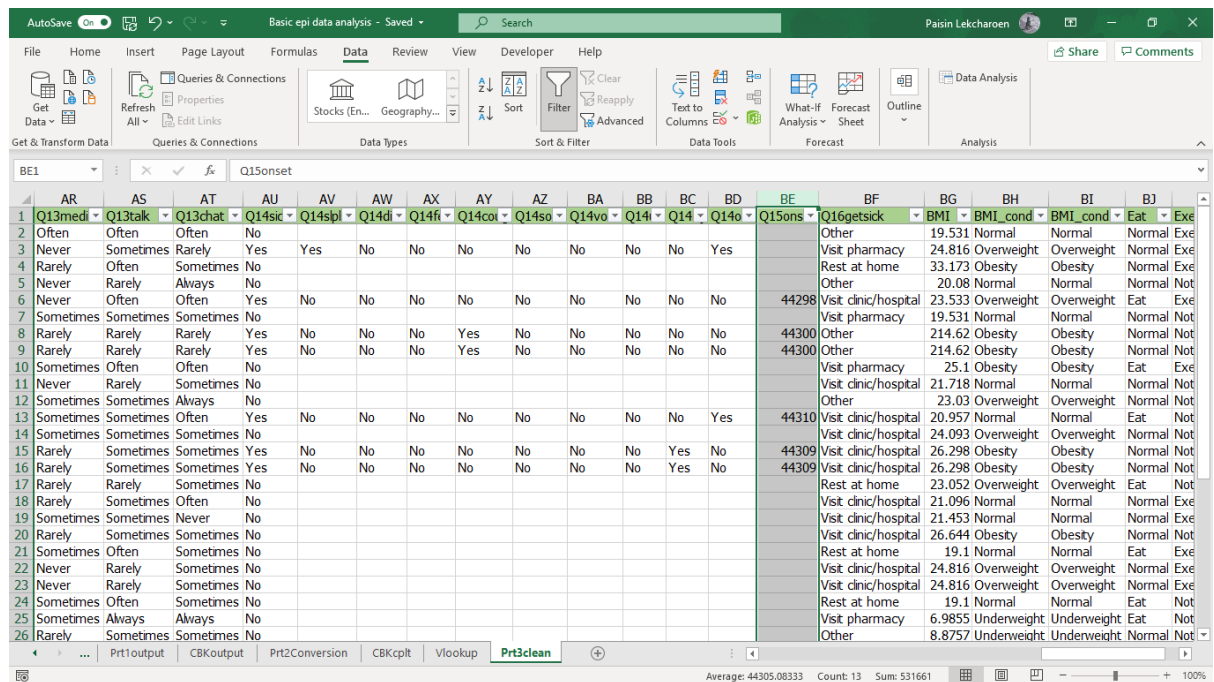
3.3. Missing or blank value.

The screenshot shows an Excel spreadsheet with columns labeled AR through BJ. The data in these columns consists of categorical responses like 'Never', 'Rarely', 'Sometimes', 'Often', and 'Always'. A filter dropdown is open for column BE, displaying a list of values: 44300, 44301, 44302, 44303, 44307, 44308, 44309, 44310, and (Blanks). The spreadsheet also shows a ribbon with 'Data' and 'Filter' options, and a status bar at the bottom indicating '12 of 37 records found'.

3.4. Any typos for example, one may input 'Thai', 'Thailand', or 'Thialand' for the question about nationality.

The screenshot shows an Excel spreadsheet with columns labeled Q1 through Q11d. The data includes demographic information such as age, sex, and nationality. A filter dropdown is open for column Q6nat, displaying a list of nationalities: Cambodian, Chinese, Filipino, Indonesian, Laotian, Malaysian, Thai, Thailand, and Vietnamese. The spreadsheet also shows a ribbon with 'Data' and 'Filter' options, and a status bar at the bottom indicating '12 of 37 records found'.

3.5. Wrong data format. For example, if you have not set a data format at the first place, e.g., 'Onset' should be allowed for 'Date' format, it will return a 'General' type of data format as a number or text string.

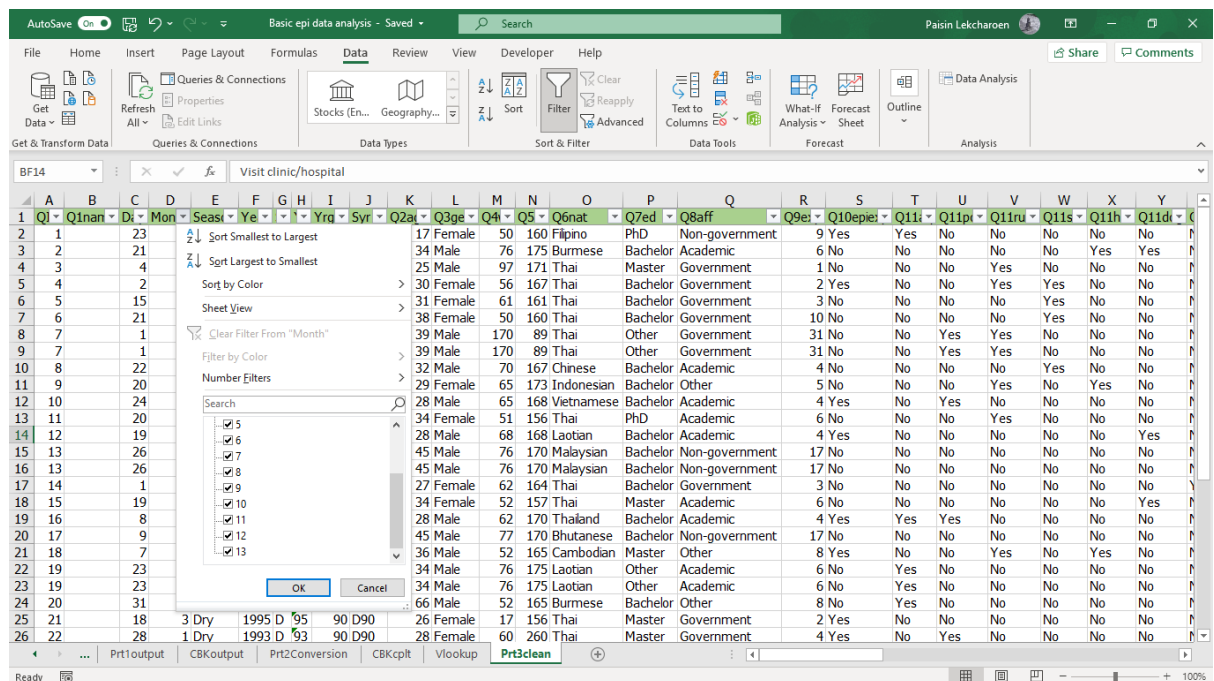


3.6. Highlight any erroneous data/cells/cell ranges for further recheck and correction.

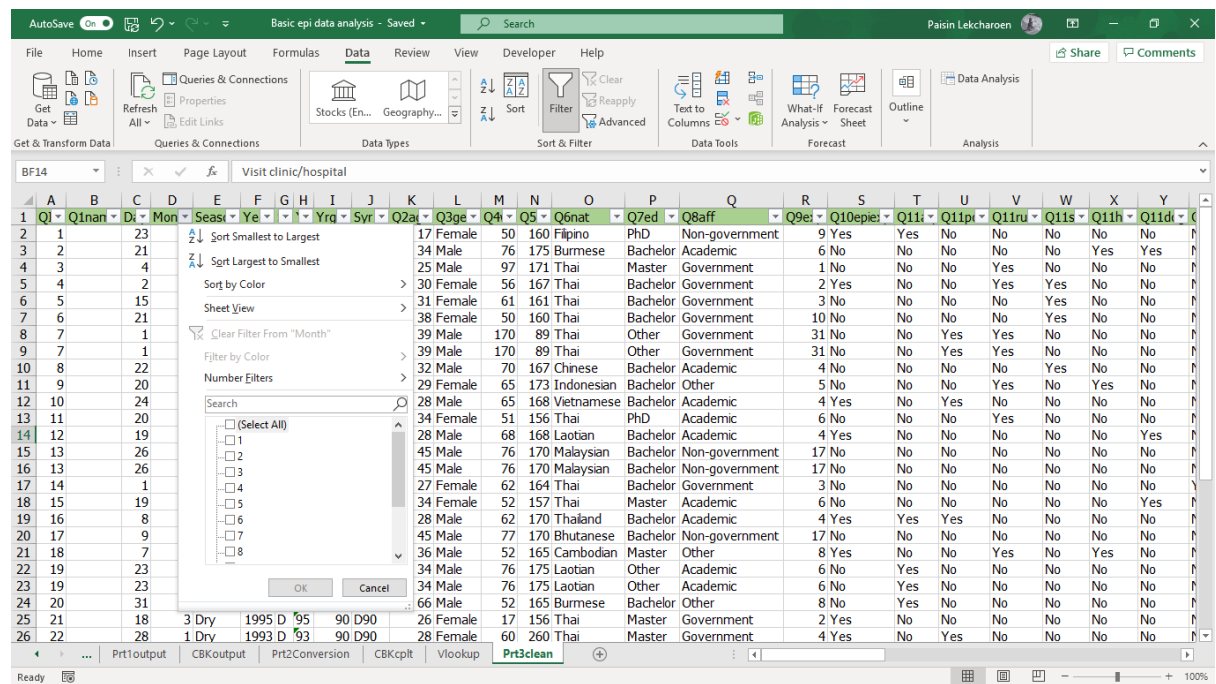
4. If you find any irrelevant non-missing data, return to the original questionnaire to correct them.

For example, in 'Month', there are records containing value '13', which is incorrect.

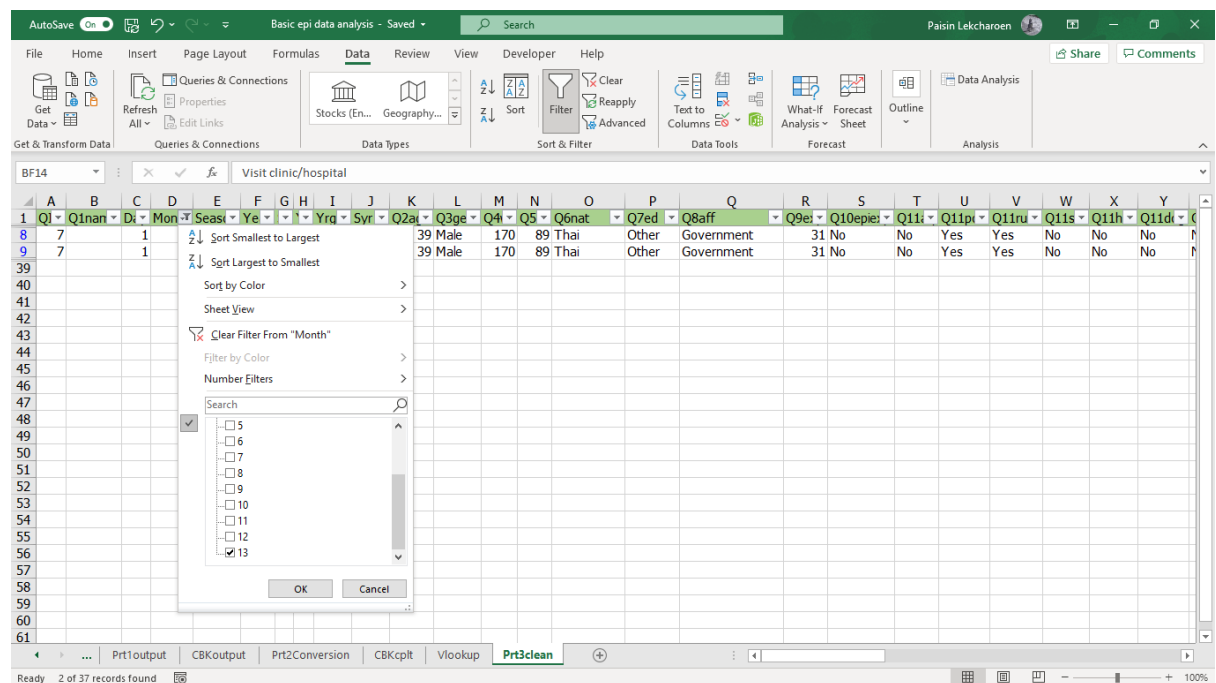
4.1. Click on drop-down. Look through the list of values.



4.2. Uncheck a box in front of '(Select All)'.



4.3. Then check only a box in front of '13'.



4.4. Only records containing value '13'. See QID of these records. Find correct data from the original data in the questionnaire and correct it.

The screenshot shows an Excel spreadsheet with the following data visible:

Q1	Q1nat	D	Mon	Seas	Ye	D	Yrg	Syr	Q2a	Q3ge	Q4	Q5	Q6nat	Q7ed	Q8aff	Q9e	Q10epie	Q11	Q11p	Q11ru	Q11s	Q11h	Q11d
7		13	Dry		1982	D	82	80	D80	39	Male	170	89	Thai	Other	Government	31	No	No	Yes	Yes	No	No
9		13	Dry		1982	D	82	80	D80	39	Male	170	89	Thai	Other	Government	31	No	No	Yes	Yes	No	No

The status bar at the bottom indicates: Ready 2 of 37 records found. Average: 13 Count: 3 Sum: 26. The task pane shows 'Prt3clean' selected.



Exercise 3.2 Data Validation

You can use 'Data Validation' function to validate your database. It is easy if you already have indicated a data format and a range of possible values before you enter data (as in Part 1). However, you can still use this function after a data entering process.

1. For example, a column 'Day' can have a range of 1 to 31. So, select 'Data Validation' from 'Data' ribbon after highlight all the column 'Day'.

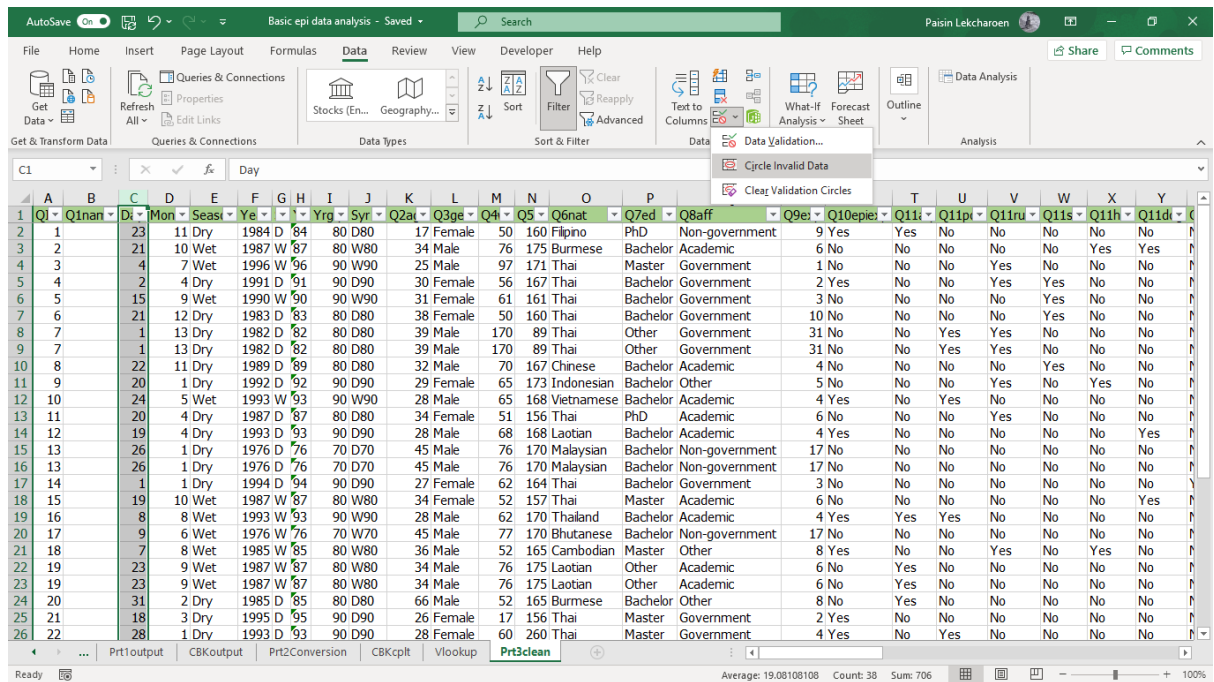
The screenshot shows the Microsoft Excel interface with the 'Data' ribbon selected. The 'Data Validation' dialog box is open, showing the 'Settings' tab. The 'Allow' dropdown is set to 'Whole number', and the 'Ignore blank' checkbox is checked. The 'Data' dropdown is set to 'between'. The 'Minimum' value is 1 and the 'Maximum' value is 31. The background spreadsheet shows a table with columns for Date (Day, Month, Year), Season, and various demographic and educational data.

2. Allow 'Whole Number' as a valid data format. Indicate 1 as minimum and 31 as maximum value. Then click 'OK'.

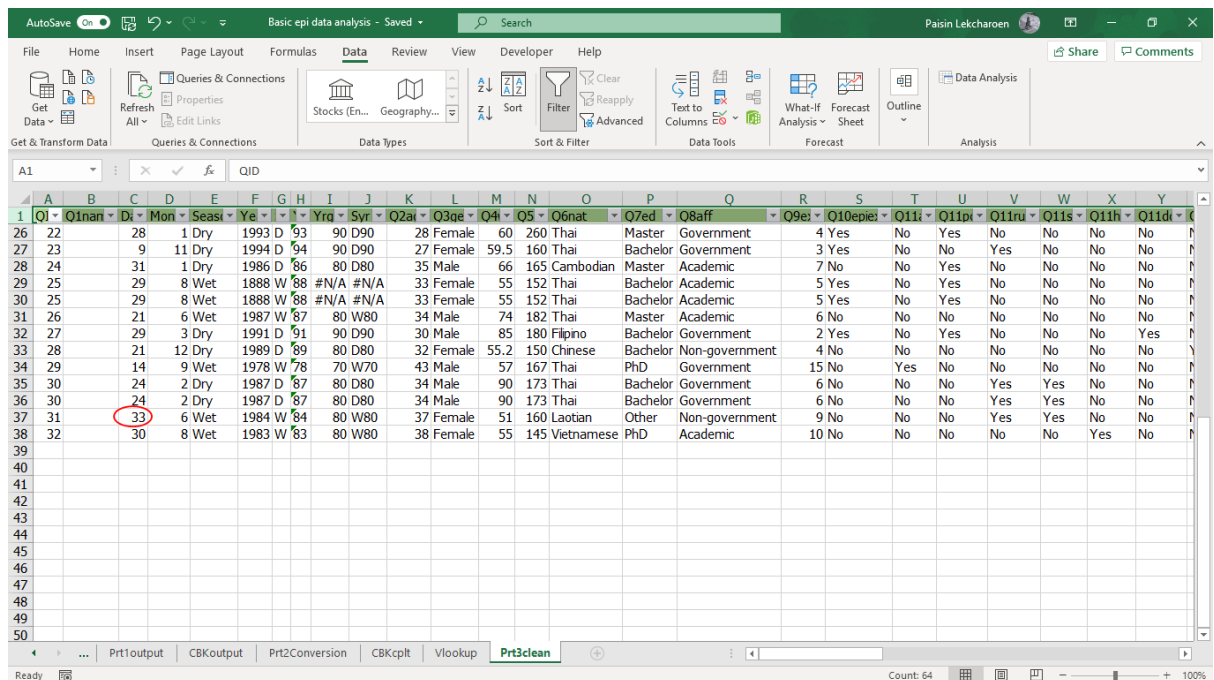
The screenshot shows the Microsoft Excel interface with the 'Data Validation' dialog box open, showing the 'Settings' tab. The 'Allow' dropdown is set to 'Whole number', and the 'Ignore blank' checkbox is checked. The 'Data' dropdown is set to 'between'. The 'Minimum' value is 1 and the 'Maximum' value is 31. The background spreadsheet shows the same table as in the previous screenshot.



3. Now, select 'Circle Invalid Data'.



4. It will show a red circle on cell(s) containing invalid data. So, you can recheck and correct it.



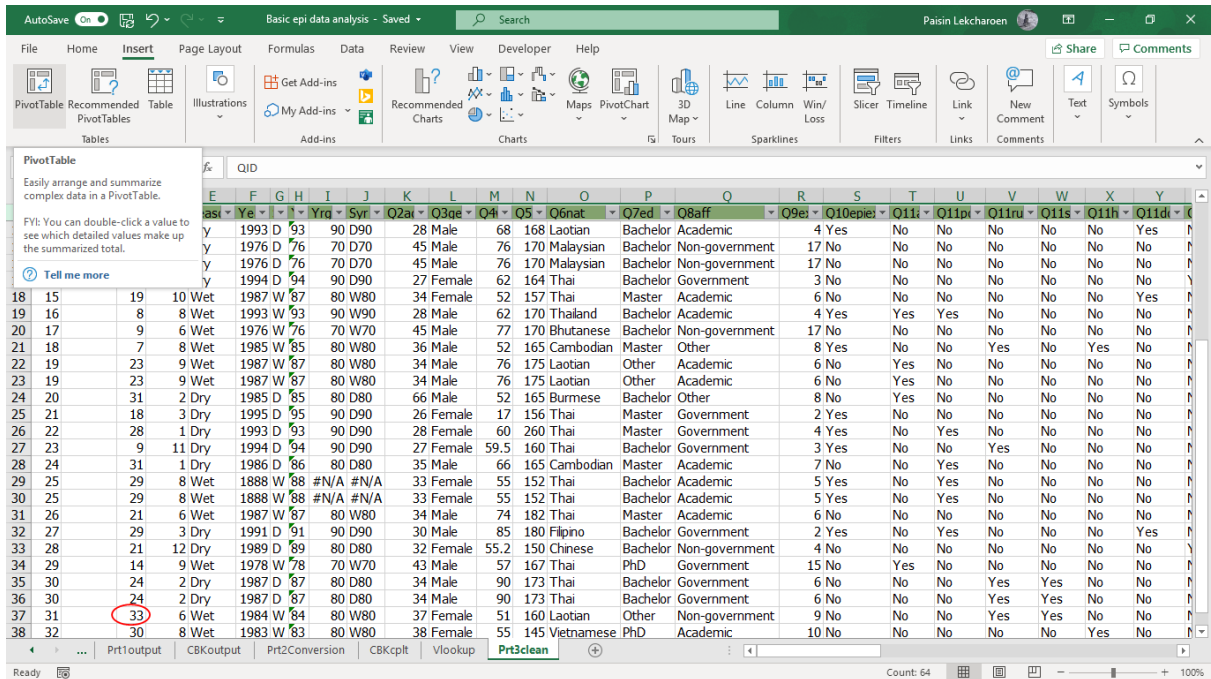
5. Apply this function to other columns and see if it circles the same cells as you have highlighted from previous exercise.



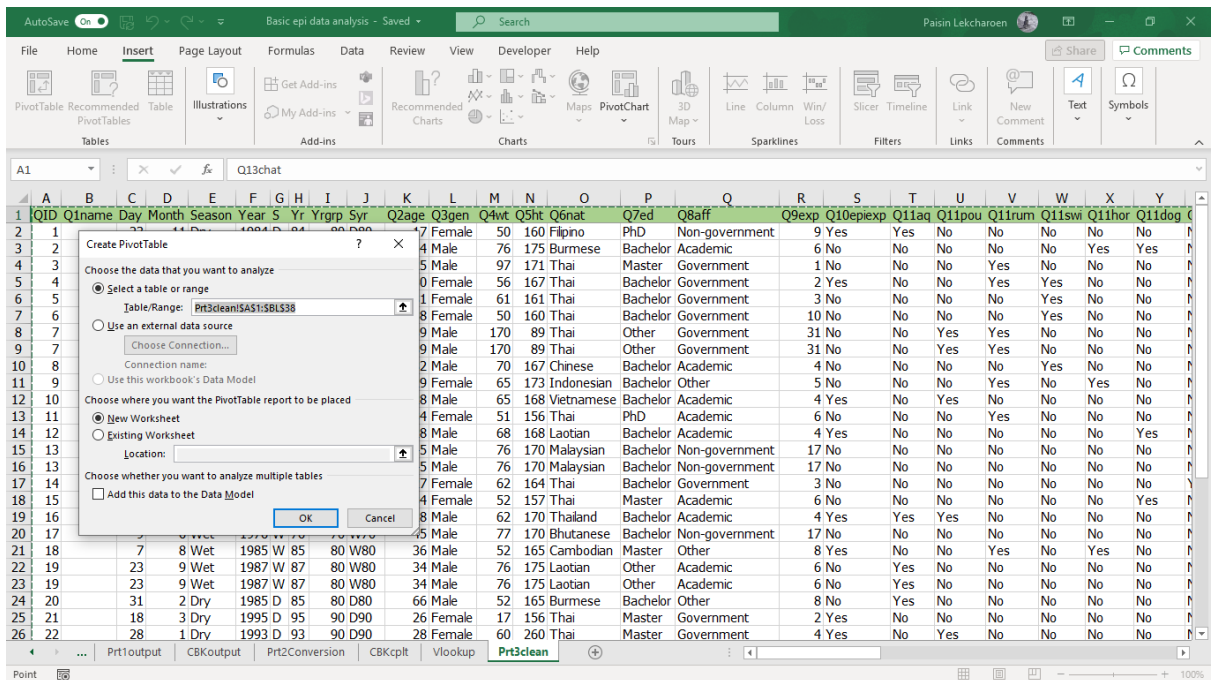
Exercise 3.3 Pivot Table

You can also use a pivot table to validate your database.

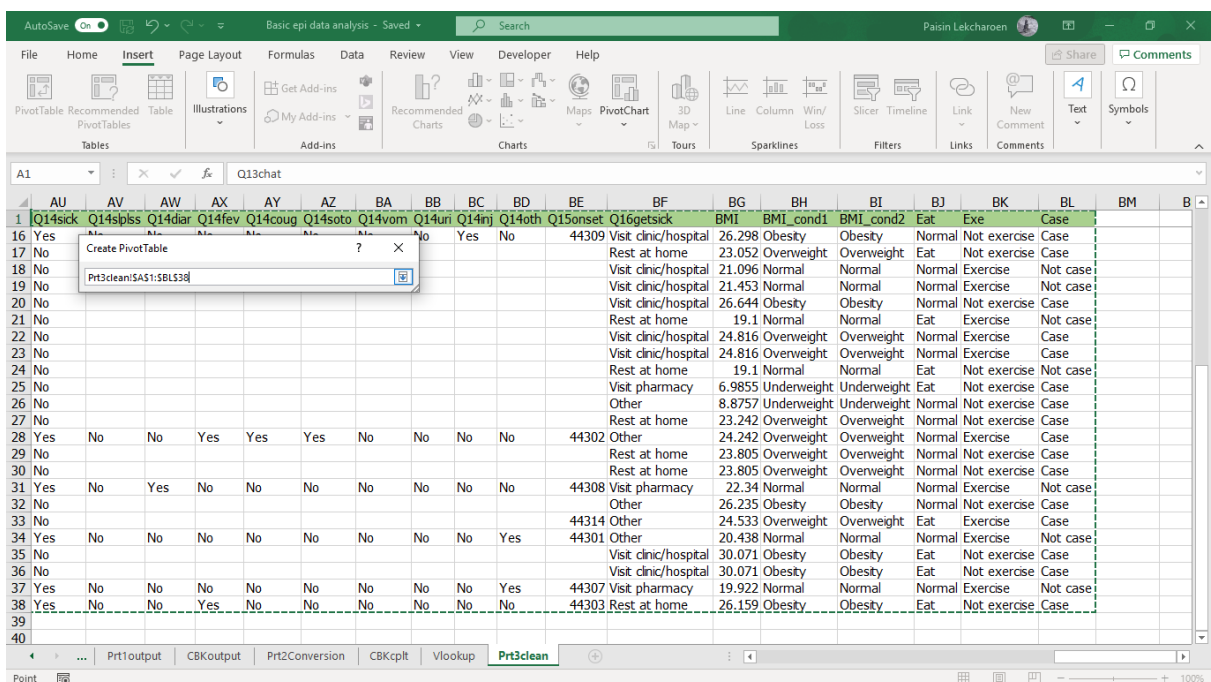
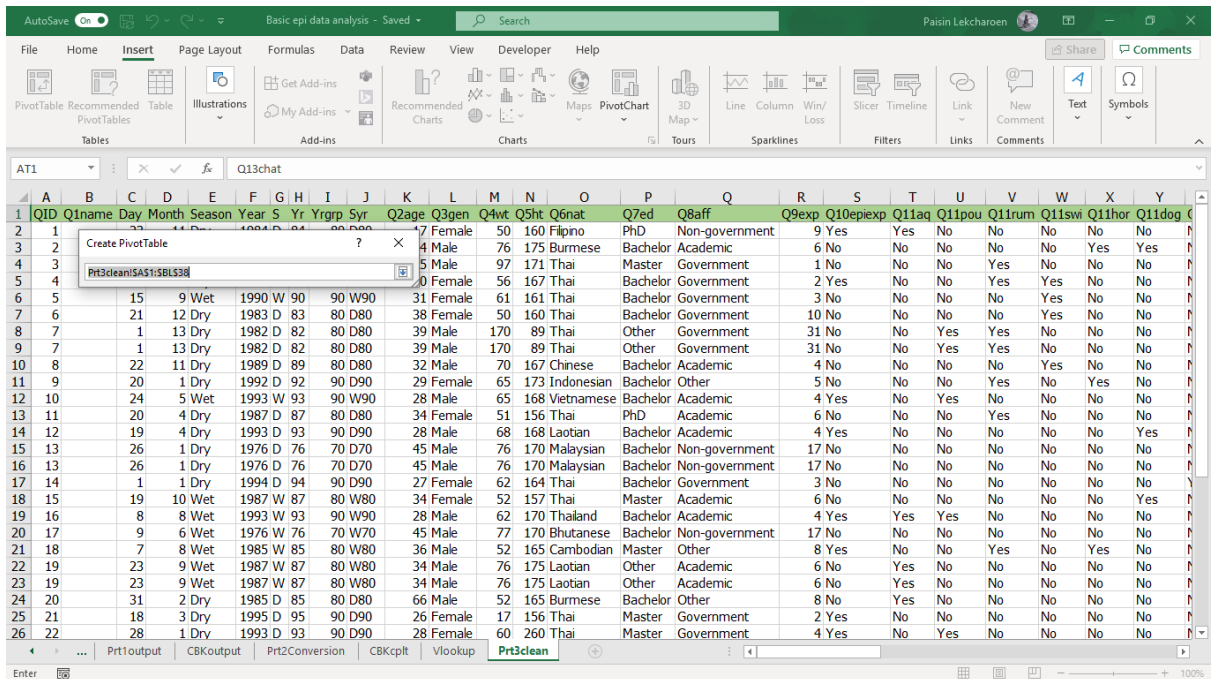
1. On 'Insert' ribbon, select 'Pivot Table'.



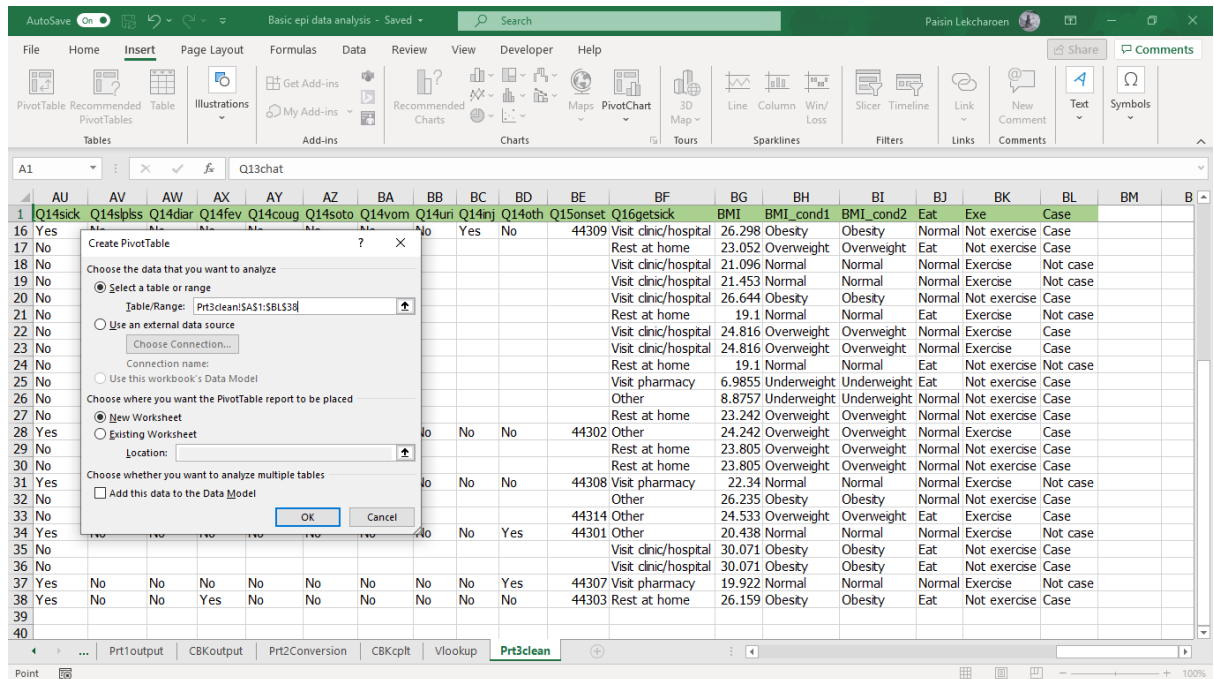
2. It will automatically select a table or range of cell you want to further your analysis. However, if it does not select data range properly, you can define it by click on the upward arrow at the end of 'Table/Range' box.



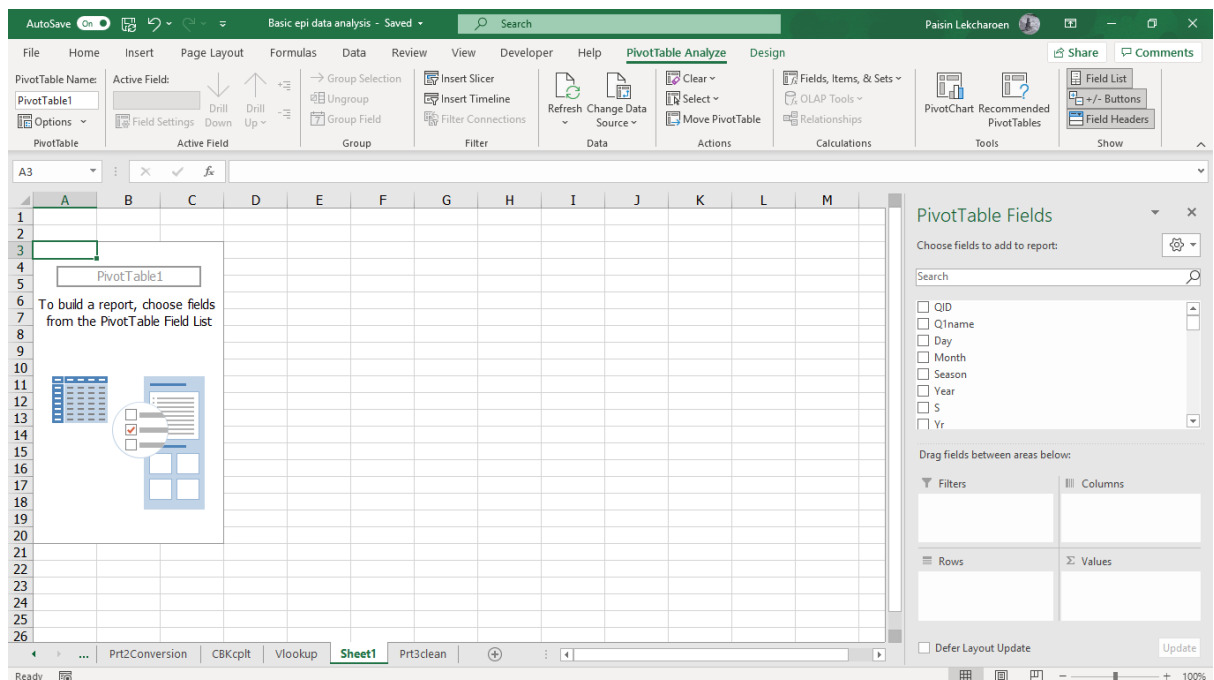
3. A 'Create PivotTable' window appear. Drag cursor to cover all cells that you want to use. The data in the box will be moving upon your dragging. Then click on downward arrow.



4. Then, select where you want to place the PivotTable. In this case, 'New Worksheet' is preferred.



5. A new worksheet (in this case, sheet1) which contains PivotTable space, will present beside (to the left, normally) your Prt3clean worksheet. You may name this worksheet accordingly (in this case, it names 'PVT'. PivotTable fields show all variables from your original table.



6. Drag 'QID' variable in variable list in the PivotTable fields and drop it in ' Σ Values' box. The PivotTable will show a value of '622' and the heading indicates 'Sum of QID'. This indicates that it sums up all values from the variable 'QID' which is not our purpose.

The screenshot shows the Microsoft Excel interface with the PivotTable Analyze ribbon selected. The PivotTable in the worksheet has a single cell containing 'Sum of QID' with a value of 622. The PivotTable Fields task pane on the right shows the 'QID' field checked and placed in the 'Values' area. The 'Sum of QID' dropdown is visible in the Values area.

7. Click on the arrow at the end of 'Sum of QID' item in ' Σ Values' box. Select 'Value Field Settings'.

The screenshot shows the same Excel interface as above, but with the 'Value Field Settings' dialog box open for the 'Sum of QID' field. The dialog box is positioned over the PivotTable Fields task pane. The 'Sum of QID' dropdown is highlighted, and the 'Value Field Settings...' option is selected in the context menu.

8. Select 'Count' instead of 'Sum' and click 'OK'.

The screenshot shows the Microsoft Excel interface with a PivotTable. The PivotTable Name is 'PivotTable1' and the Active Field is 'Sum of QID'. The value in cell A4 is 622. A 'Value Field Settings' dialog box is open, showing the 'Source Name' as 'QID' and the 'Custom Name' as 'Count of QID'. The 'Summarize Values By' section is set to 'Summarize value field by', and the 'Sum' option is selected in the list. The 'Number Format' button is visible at the bottom of the dialog box.

9. The value indicates 37 for the 'Count of QID'. Do you think this is a correct value? No, you have collected data from 32 participants as indicated in Part 2. So, it should have something wrong. There may be some duplicates or incorrect QID.

The screenshot shows the Microsoft Excel interface with a PivotTable. The PivotTable Name is 'PivotTable1' and the Active Field is 'Count of QID'. The value in cell A4 is 37. The 'PivotTable Fields' task pane on the right shows 'QID' selected in the 'Values' area, and 'Count of QID' is displayed in the 'Σ Values' section.

10. Now, drag and drop 'QID' variable in the 'Row' box. Each row will show each value for QID (which are 1 to 32). The Count of QID column shows a count of each value in the row. Each count should be 1; however, the count for '7', '13', '19', and some others is 2, suggesting duplicate records.

The screenshot shows the Excel interface with a PivotTable. The PivotTable Fields task pane on the right shows 'QID' in the Rows area and 'Count of QID' in the Values area. The PivotTable data is as follows:

QID	Count of QID
1	1
2	1
3	1
4	1
5	1
6	1
7	2
8	1
9	1
10	1
11	1
12	1
13	2
14	1
15	1
16	1
17	1
18	1
19	2
20	1
21	1
22	1
23	1
24	1
25	1
26	1

11. Move to check 'Day' variable. Clear the value in the PivotTable fields by uncheck a box in front of QID variable. Then, drag and drop 'Day' into the 'Row' box. It will show Day number in each row. You may notice that '33' is an invalid value for Day.

The screenshot shows the Excel interface with a PivotTable. The PivotTable Fields task pane on the right shows 'Day' in the Rows area and 'Count of Day' in the Values area. The PivotTable data is as follows:

Day	Count of Day
1	1
2	1
3	1
4	1
5	1
6	1
7	1
8	1
9	1
10	1
11	1
12	1
13	1
14	1
15	1
16	1
17	1
18	1
19	1
20	1
21	1
22	1
23	1
24	1
25	1
26	1
27	1
28	1
29	1
30	1
31	1
32	1
33	1
Grand Total	33

12. Drag and drop QID variable under Day variable in the Row box. You may see that the QID that contains a value of Day 33 is QID 31. This helps you to identify which questionnaire has incorrect value or which record may have invalid data entering.

The screenshot shows an Excel PivotTable with the following data:

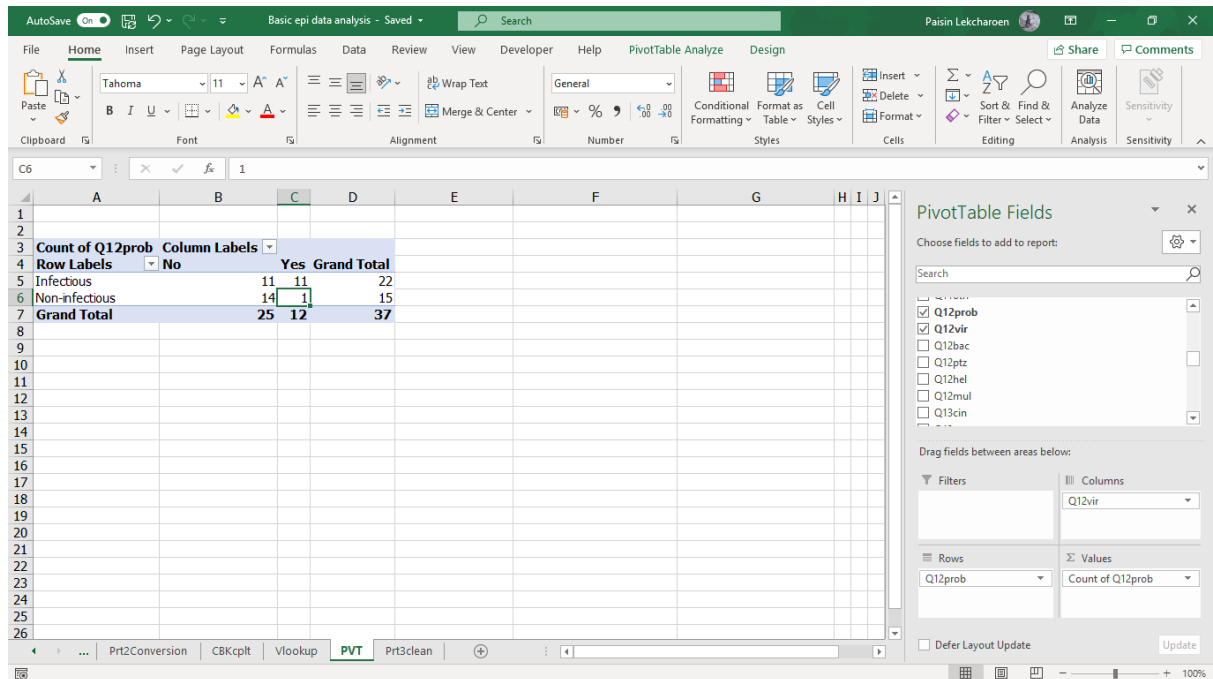
Day	QID
28	
22	
8	
23	
1	
19	
24	
10	
30	
26	
13	
28	
22	
29	
25	
27	
30	
32	
31	
20	
24	
33	31
Grand Total	

13. PivotTable can help to identify invalid condition. For example, put Q12prob in both Row and Values boxes. Set a value as a count. It will count a number of infectious and non-infectious records.

The screenshot shows an Excel PivotTable with the following data:

Row Labels	Count of Q12prob
Infectious	22
Non-infectious	15
Grand Total	37

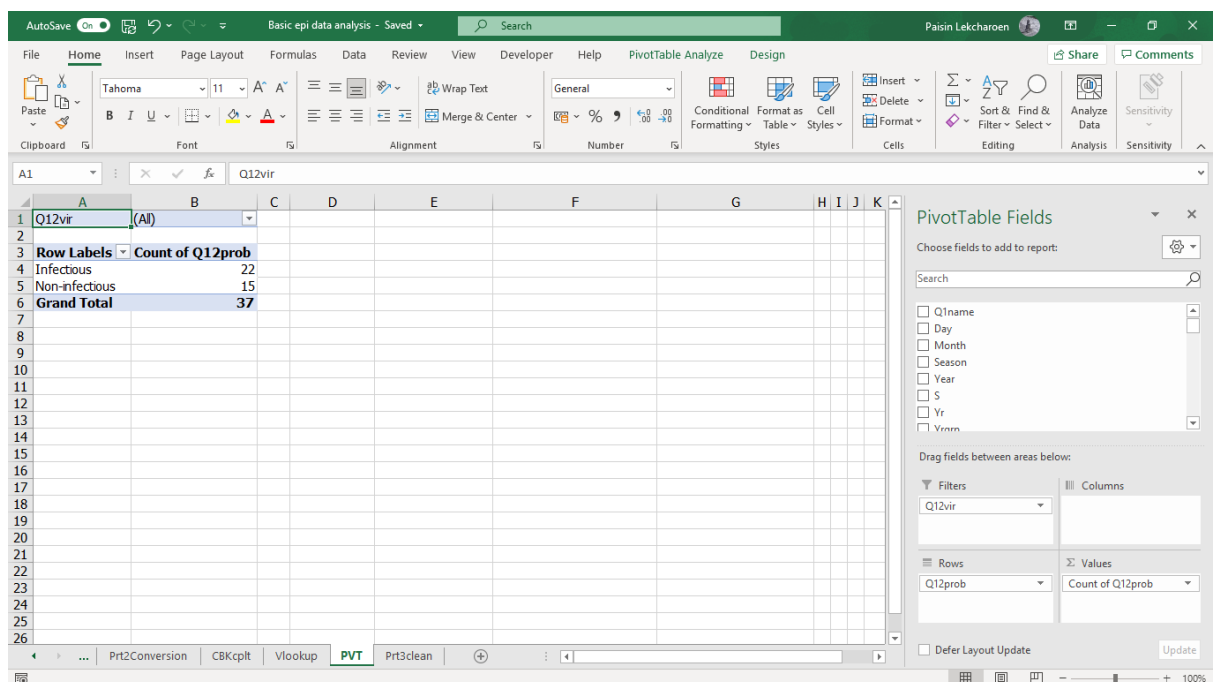
14. Let you check that how many 'Infectious' records are studies about viruses. Drag and drop Q12vir in Column box. You may see that there are 11 records in Infectious row answering Yes and 11 records answering No which is possible (there are other infectious problems such as bacteria and protozoa!). However, there is one record in Non-infectious row answering Yes which is not possible. Then you can go back and look for the correct value.



The screenshot shows the Microsoft Excel interface with a PivotTable. The PivotTable Fields task pane on the right has 'Q12vir' in the Columns area and 'Q12prob' in the Rows area. The PivotTable data is as follows:

Count of Q12prob	Column Labels		
Row Labels	No	Yes	Grand Total
Infectious	11	11	22
Non-infectious	14	1	15
Grand Total	25	12	37

15. Now, if you want to know which record has incorrect condition, move Q12vir from Column box to Filter box. The filter for possible value of Q12vir, which are Yes and No, will appear above the table. Now it shows all values in the table.



The screenshot shows the Microsoft Excel interface with the same PivotTable. The PivotTable Fields task pane on the right has 'Q12vir' in the Filters area and 'Q12prob' in the Rows area. The PivotTable data is as follows:

Count of Q12prob	
Infectious	22
Non-infectious	15
Grand Total	37

16. Click the arrowhead at the end of the filter of Q12vir. Select only Yes and click OK.

The screenshot shows the Microsoft Excel interface with a PivotTable. The PivotTable is located in the range A1:K26. The PivotTable Fields task pane is open on the right, showing the following configuration:

- Filters:** Q12vir
- Columns:** (Empty)
- Rows:** Q12prob
- Values:** Count of Q12prob

The PivotTable data is as follows:

Q12vir	Count of Q12prob
Yes	11
No	1
Grand Total	12

17. Again, it sums up a number of records answering Yes for each row.

The screenshot shows the Microsoft Excel interface with the PivotTable summarized. The PivotTable is located in the range A1:K26. The PivotTable Fields task pane is open on the right, showing the following configuration:

- Filters:** Q12vir
- Columns:** (Empty)
- Rows:** Q12prob
- Values:** Count of Q12prob

The PivotTable data is as follows:

Q12vir	Count of Q12prob
Yes	11
No	1
Grand Total	12

18. Then, drag and drop QID variable under Q12prob in the Row box. So, you can identify which record may contain incorrect value.

The screenshot shows an Excel PivotTable with the following data:

Q12vir	Count of Q12prob
Infectious	11
2	1
4	1
5	1
15	1
23	1
25	2
27	1
30	2
31	1
Non-infectious	1
6	1
Grand Total	12

19. Play around with other variables. Then correct all errors you can identify.

Exercise 3.4 Look for and eliminate duplicates.

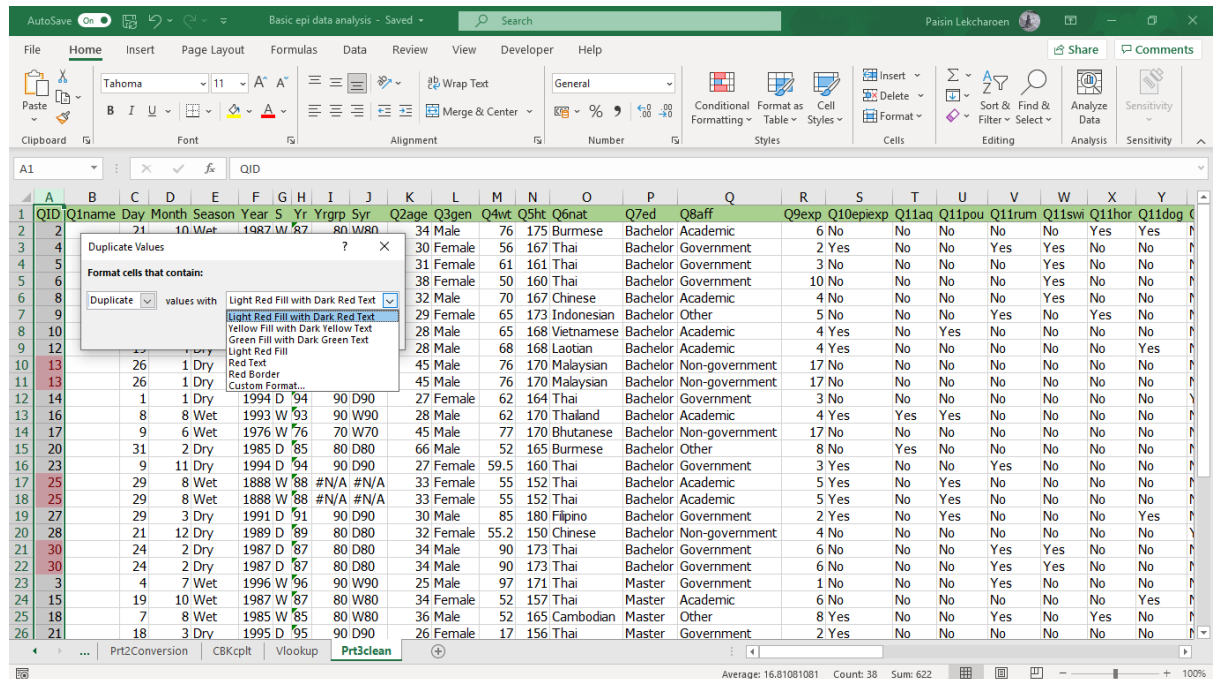
In the previous exercise, you may notice that there are some duplicates. You have to clean up those records before furthering your analysis.

1. Identify duplicate QID using 'Conditional Formatting' function.

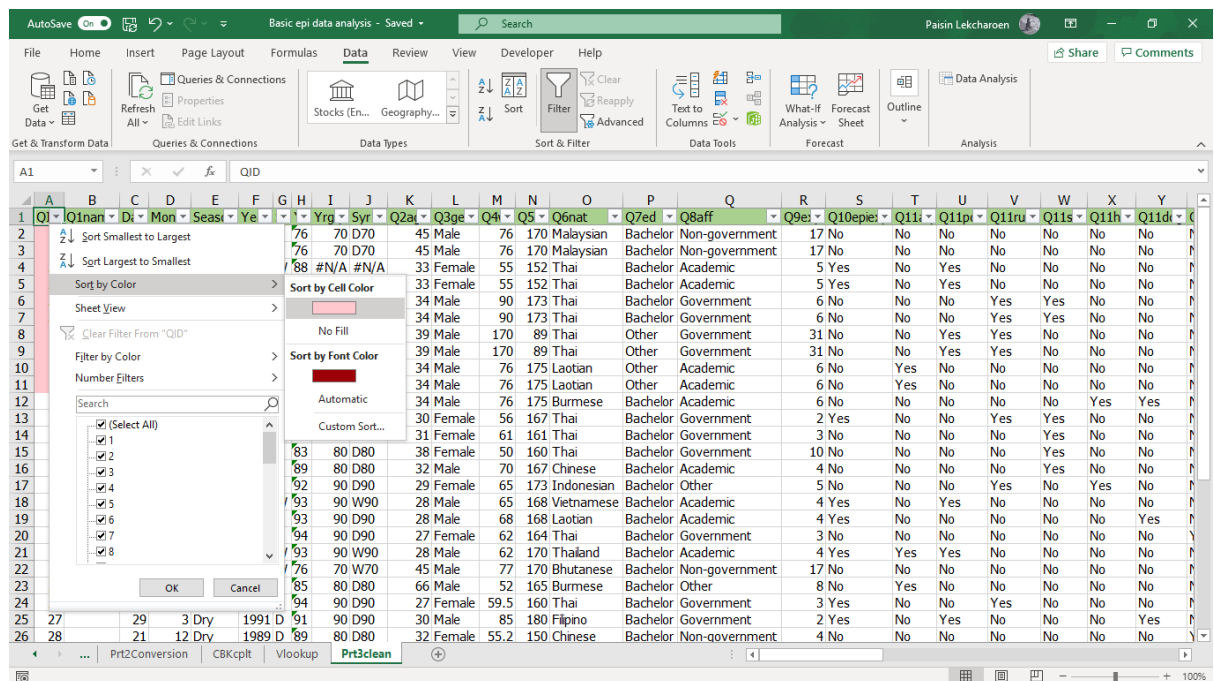
1.1. Highlight QID column, then select 'Conditional Formatting' function from 'Home' ribbon.

1.2. Select 'Highlight Cells Rules', then select 'Duplicate Values'.

1.3. A 'Duplicate Values' appears. Choose 'Duplicate' and select cell highlight as you would like.



1.4. The duplicate values will be highlighted. You can sort the highlighted values so it is easy to delete.



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
QID	Q1name	Day	Month	Season	Year	S	Yr	Yrgrp	Svr	Q2age	Q3gen	Q4vst	Q5ht	Q6nat	Q7ed	Q8aff	Q9es	Q10epie	Q11i	Q11p	Q11ru	Q11s	Q11h	Q11d	Q11c
13		26	1	Dry	1976	D	76	70	D70	45	Male	76	170	Malaysian	Bachelor	Non-government	17	No	No	No	No	No	No	No	No
13		26	1	Dry	1976	D	76	70	D70	45	Male	76	170	Malaysian	Bachelor	Non-government	17	No	No	No	No	No	No	No	No
25		29	8	Wet	1888	W	88	#N/A	#N/A	33	Female	55	152	Thai	Bachelor	Academic	5	Yes	No	Yes	No	No	No	No	No
25		29	8	Wet	1888	W	88	#N/A	#N/A	33	Female	55	152	Thai	Bachelor	Academic	5	Yes	No	Yes	No	No	No	No	No
30		24	2	Dry	1987	D	87	80	D80	34	Male	90	173	Thai	Bachelor	Government	6	No	No	No	Yes	Yes	No	No	No
30		24	2	Dry	1987	D	87	80	D80	34	Male	90	173	Thai	Bachelor	Government	6	No	No	No	Yes	Yes	No	No	No
7		1	13	Dry	1982	D	82	80	D80	39	Male	170	89	Thai	Other	Government	31	No	No	Yes	Yes	No	No	No	No
7		1	13	Dry	1982	D	82	80	D80	39	Male	170	89	Thai	Other	Government	31	No	No	Yes	Yes	No	No	No	No
19		23	9	Wet	1987	W	87	80	W80	34	Male	76	175	Laotian	Other	Academic	6	No	Yes	No	No	No	No	No	No
19		23	9	Wet	1987	W	87	80	W80	34	Male	76	175	Laotian	Other	Academic	6	No	Yes	No	No	No	No	No	No
2		21	10	Wet	1987	W	87	80	W80	34	Male	76	175	Burmese	Bachelor	Academic	6	No	No	No	No	No	Yes	Yes	No
4		2	4	Dry	1991	D	91	90	D90	30	Female	56	167	Thai	Bachelor	Government	2	Yes	No	No	Yes	No	Yes	No	No
5		15	9	Wet	1990	W	90	90	W90	31	Female	61	161	Thai	Bachelor	Government	3	No	No	No	No	Yes	No	No	No
6		21	12	Dry	1983	D	83	80	D80	38	Female	50	160	Thai	Bachelor	Government	10	No	No	No	No	Yes	No	No	No
8		22	11	Dry	1989	D	89	80	D80	32	Male	70	167	Chinese	Bachelor	Academic	4	No	No	No	No	Yes	No	No	No
9		20	1	Dry	1992	D	92	90	D90	29	Female	65	173	Indonesian	Bachelor	Other	5	No	No	No	Yes	No	Yes	No	No
10		24	5	Wet	1993	W	93	90	W90	28	Male	65	168	Vietnamese	Bachelor	Academic	4	Yes	No	Yes	No	No	No	No	No
12		19	4	Dry	1993	D	93	90	D90	28	Male	68	168	Laotian	Bachelor	Academic	4	Yes	No	No	No	No	No	No	Yes
14		1	1	Dry	1994	D	94	90	D90	27	Female	62	164	Thai	Bachelor	Government	3	No	No	No	No	No	No	No	No
16		8	8	Wet	1993	W	93	90	W90	28	Male	62	170	Thailand	Bachelor	Academic	4	Yes	Yes	Yes	No	No	No	No	No
17		9	6	Wet	1976	W	76	70	W70	45	Male	77	170	Bhutanese	Bachelor	Non-government	17	No	No	No	No	No	No	No	No
20		31	2	Dry	1985	D	85	80	D80	66	Male	52	165	Burmese	Bachelor	Other	8	No	Yes	No	No	No	No	No	No
23		9	11	Dry	1994	D	94	90	D90	27	Female	59.5	160	Thai	Bachelor	Government	3	Yes	No	No	Yes	No	No	No	No
25		29	3	Dry	1991	D	91	90	D90	30	Male	85	180	Pinpo	Bachelor	Government	2	Yes	No	Yes	No	No	No	No	Yes
28		21	12	Dry	1989	D	89	80	D80	32	Female	55.2	150	Chinese	Bachelor	Non-government	4	No	No	No	No	No	No	No	No

2. After you identify any duplicate records, you can delete them manually or using 'Remove Duplicates' function.

2.1. On the 'Data' ribbon, select 'Remove Duplicates' function.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
QID	Q1name	Day	Month	Season	Year	S	Yr	Yrgrp	Svr	Q2age	Q3gen	Q4vst	Q5ht	Q6nat	Q7ed	Q8aff	Q9es	Q10epie	Q11i	Q11p	Q11ru	Q11s	Q11h	Q11d	Q11c
13		26	1	Dry	1976	D	76	70	D70	45	Male	76	170	Malaysian	Bachelor	Non-government	17	No	No	No	No	No	No	No	No
13		26	1	Dry	1976	D	76	70	D70	45	Male	76	170	Malaysian	Bachelor	Non-government	17	No	No	No	No	No	No	No	No
25		29	8	Wet	1888	W	88	#N/A	#N/A	33	Female	55	152	Thai	Bachelor	Academic	5	Yes	No	Yes	No	No	No	No	No
25		29	8	Wet	1888	W	88	#N/A	#N/A	33	Female	55	152	Thai	Bachelor	Academic	5	Yes	No	Yes	No	No	No	No	No
30		24	2	Dry	1987	D	87	80	D80	34	Male	90	173	Thai	Bachelor	Government	6	No	No	No	Yes	Yes	No	No	No
30		24	2	Dry	1987	D	87	80	D80	34	Male	90	173	Thai	Bachelor	Government	6	No	No	No	Yes	Yes	No	No	No
7		1	13	Dry	1982	D	82	80	D80	39	Male	170	89	Thai	Other	Government	31	No	No	Yes	Yes	No	No	No	No
7		1	13	Dry	1982	D	82	80	D80	39	Male	170	89	Thai	Other	Government	31	No	No	Yes	Yes	No	No	No	No
19		23	9	Wet	1987	W	87	80	W80	34	Male	76	175	Laotian	Other	Academic	6	No	Yes	No	No	No	No	No	No
19		23	9	Wet	1987	W	87	80	W80	34	Male	76	175	Laotian	Other	Academic	6	No	Yes	No	No	No	No	No	No
2		21	10	Wet	1987	W	87	80	W80	34	Male	76	175	Burmese	Bachelor	Academic	6	No	No	No	No	No	Yes	Yes	No
4		2	4	Dry	1991	D	91	90	D90	30	Female	56	167	Thai	Bachelor	Government	2	Yes	No	No	Yes	No	Yes	No	No
5		15	9	Wet	1990	W	90	90	W90	31	Female	61	161	Thai	Bachelor	Government	3	No	No	No	No	Yes	No	No	No
6		21	12	Dry	1983	D	83	80	D80	38	Female	50	160	Thai	Bachelor	Government	10	No	No	No	No	Yes	No	No	No
8		22	11	Dry	1989	D	89	80	D80	32	Male	70	167	Chinese	Bachelor	Academic	4	No	No	No	No	Yes	No	No	No
9		20	1	Dry	1992	D	92	90	D90	29	Female	65	173	Indonesian	Bachelor	Other	5	No	No	No	Yes	No	Yes	No	No
10		24	5	Wet	1993	W	93	90	W90	28	Male	65	168	Vietnamese	Bachelor	Academic	4	Yes	No	Yes	No	No	No	No	No
12		19	4	Dry	1993	D	93	90	D90	28	Male	68	168	Laotian	Bachelor	Academic	4	Yes	No	No	No	No	No	No	Yes
14		1	1	Dry	1994	D	94	90	D90	27	Female	62	164	Thai	Bachelor	Government	3	No	No	No	No	No	No	No	No
16		8	8	Wet	1993	W	93	90	W90	28	Male	62	170	Thailand	Bachelor	Academic	4	Yes	Yes	Yes	No	No	No	No	No
17		9	6	Wet	1976	W	76	70	W70	45	Male	77	170	Bhutanese	Bachelor	Non-government	17	No	No	No	No	No	No	No	No
20		31	2	Dry	1985	D	85	80	D80	66	Male	52	165	Burmese	Bachelor	Other	8	No	Yes	No	No	No	No	No	No
23		9	11	Dry	1994	D	94	90	D90	27	Female	59.5	160	Thai	Bachelor	Government	3	Yes	No	No	Yes	No	No	No	No
25		29	3	Dry	1991	D	91	90	D90	30	Male	85	180	Pinpo	Bachelor	Government	2	Yes	No	Yes	No	No	No	No	Yes
28		21	12	Dry	1989	D	89	80	D80	32	Female	55.2	150	Chinese	Bachelor	Non-government	4	No	No	No	No	No	No	No	No

2.2. A 'Remove Duplicates' window appears. Select all columns and click 'OK'.

The screenshot shows the 'Remove Duplicates' dialog box in Microsoft Excel. The dialog box has a title bar with a question mark and a close button. The main text says 'To delete duplicate values, select one or more columns that contain duplicates.' Below this, there are two buttons: 'Select All' (highlighted) and 'Unselect All'. There is also a checkbox labeled 'My data has headers' which is checked. A list box titled 'Columns' is open, showing a list of column headers: QID, Q1name, Day, Month, Season, Year, S, Yr, Yrgrp, Syr, Q2age, Q3gen, Q4wt, Q5ht, Q6nat, Q7ed, Q8aff, Q9exp, Q10epexp, Q11aq, Q11pou, Q11rum, Q11swi, Q11hor, and Q11dog. The 'QID' column is selected. The background spreadsheet shows a table with various data points, including names, ages, weights, heights, nationalities, and educational levels.

2.3. It will report a number of removed and remaining records.

The screenshot shows the same Excel spreadsheet as in the previous image, but with a message box open in the center. The message box has a blue information icon and contains the text: '5 duplicate values found and removed; 32 unique values remain.' There is an 'OK' button at the bottom of the message box. The spreadsheet data is partially visible behind the message box.

Hint: Variables containing erroneous data

Day, Month, Year, Yrgrp, Syr, Q2age, Q4wt, Q5ht, Q6nat, Q9exp, Q12vir, Q12bac, Q12mul, Q15onset



Exercise 3.5 Recode.

Many software are available for data management and analysis, e.g., MS Excel, EpiInfo, STATA, SAS, SPSS, and R. Each software requires different data types and formats. However, most software can manipulate data in numeric format. Therefore, it is more comfortable if you can change your text string data into numeric format by recoding them. Furthermore, some continuous data may be difficult to analyze for finding association. It is preferred to group or re-group these data.

1. Open 'Prt3recode' worksheet. Many variables contain text string as data. It will be better to use 'Vlookup' function to recode these values.

Before that, you need to generate table of values for each variable.

Let's start with Q2age. It is especially important to define age into age group. Insert a new column name 'Agegr'.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
QID	Q1name	Day	Month	Season	Year	S	Yr	Yrgrp	Syr	Q2age	Agegr	Q3gen	Q4wt	Q5ht	Q6nat	Q7ed	Q8aff	Q9exp	Agestart	Q10peixp	Q11aq	Q11pou	Q11rum	Q11swi
1	1	23	11	Dry	1984	D	84	80	D80	37		Female	50	160	Filipino	PhD	Non-government	9	28	Yes	Yes	No	No	No
2	2	21	10	Wet	1987	W	87	80	W80	34		Male	76	175	Burmese	Bachelor	Academic	6	28	No	No	No	No	Yes
3	3	4	7	Wet	1996	W	96	90	W90	25		Male	97	171	Thai	Master	Government	1	24	No	No	No	Yes	No
4	4	2	4	Dry	1991	D	91	90	D90	30		Female	56	167	Thai	Bachelor	Government	2	28	Yes	No	No	Yes	Yes
5	5	15	9	Wet	1990	W	90	90	W90	31		Female	61	161	Thai	Bachelor	Government	3	28	No	No	No	No	Yes
6	6	21	12	Dry	1983	D	83	80	D80	38		Female	50	160	Thai	Bachelor	Government	10	28	No	No	No	Yes	No
7	7	1	12	Dry	1982	D	82	80	D80	39		Male	89	170	Thai	Other	Government	11	28	No	No	Yes	Yes	No
8	8	22	11	Dry	1989	D	89	80	D80	32		Male	70	167	Chinese	Bachelor	Academic	4	28	No	No	No	No	Yes
9	9	20	1	Dry	1992	D	92	90	D90	29		Female	65	173	Indonesian	Bachelor	Other	5	24	No	No	No	Yes	No
10	10	24	5	Wet	1993	W	93	90	W90	28		Male	65	168	Vietnamese	Bachelor	Academic	4	24	Yes	No	Yes	No	No
11	11	20	4	Dry	1987	D	87	80	D80	34		Female	51	156	Thai	PhD	Academic	6	28	No	No	No	Yes	No
12	12	19	4	Dry	1993	D	93	90	D90	28		Male	68	168	Laotian	Bachelor	Academic	4	24	Yes	No	No	No	No
13	13	26	1	Dry	1976	D	76	70	D70	45		Male	76	170	Malaysian	Bachelor	Non-government	17	28	No	No	No	No	No
14	14	1	1	Dry	1994	D	94	90	D90	27		Female	62	164	Thai	Bachelor	Government	3	24	No	No	No	No	No
15	15	19	10	Wet	1987	W	87	80	W80	34		Female	52	157	Thai	Master	Academic	6	28	No	No	No	No	No
16	16	8	8	Wet	1993	W	93	90	W90	28		Male	62	170	Thai	Bachelor	Academic	4	24	Yes	Yes	Yes	No	No
17	17	9	6	Wet	1976	W	76	70	W70	45		Male	77	170	Bhutanese	Bachelor	Non-government	17	28	No	No	No	No	No
18	18	7	8	Wet	1985	W	85	80	W80	36		Male	52	165	Cambodian	Master	Other	8	28	Yes	No	No	Yes	No
19	19	23	9	Wet	1987	W	87	80	W80	34		Male	76	175	Laotian	Other	Academic	6	28	No	Yes	No	No	No
20	20	1	2	Dry	1985	D	85	80	D80	36		Male	52	165	Burmese	Bachelor	Other	8	28	No	Yes	No	No	No
21	21	18	3	Dry	1995	D	95	90	D90	26		Female	47	156	Thai	Master	Government	2	24	Yes	No	No	No	No
22	22	28	1	Dry	1993	D	93	90	D90	28		Female	60	160	Thai	Master	Government	4	24	Yes	No	Yes	No	No
23	23	9	11	Dry	1994	D	94	90	D90	27		Female	59.5	160	Thai	Bachelor	Government	3	24	Yes	No	No	Yes	No
24	24	31	1	Dry	1986	D	86	80	D80	35		Male	66	165	Cambodian	Master	Academic	7	28	No	No	Yes	No	No
25	25	29	8	Wet	1988	W	88	80	W80	33		Female	55	152	Thai	Bachelor	Academic	5	28	Yes	No	Yes	No	No

- Go to the codebook 'CBKcplt' worksheet. There is an information for 'Agegr' variable provided. Highlight 'Possible values' and 'Code' columns of 'Agegr'. Use 'Define Name' function to create a 'agegr' table. Click 'OK'.

The screenshot shows the Excel interface with the 'Formulas' tab active. The 'Agegr' variable is highlighted in green in the codebook. A 'New Name' dialog box is open, showing the name 'agegr', scope 'Workbook', and refers to '=CBKcplt!\$D\$16:\$E\$18'.

Question	Variable name	List name	Possible values	Code	Scale
1	Q1name	Name and given name	Text		Nominal
3	DOB	Date of birth	Date as indicated		Ordinal
4	Day		1-31		Ordinal
5	Month		1-12		Ordinal
6	Year		1956-2003		Ordinal
7	Yrgrp	Year group			
8			1970	70	1 Interval
9			1980	80	2
10			1990	90	3
11	Season		Wet		1 Nominal
12			Dry		2
13	2 Q2age	Age (years)	18-65		Ratio
14			No answer	999	
15	Agegrp	Age group			
16		<30		1	1 Interval
17		30-39		30	2
18		>=40		40	3
19	3 Q3gen	Gender	Male		1 Nominal
20			Female		2
21			Other		3
22			No answer		9
23	4 Q4wt	Weight (kg)	Number as indicated (>30)		Ratio
24			No answer	999	
25	5 Q5ht	Height (cm)	Number as indicated		Ratio
26			No answer	999	

- Go back to 'Prt3recode' worksheet. In cell L2, put a command as follow:

➤ '= VLOOKUP (K2, agegr, 2)'

The screenshot shows the Excel interface with the 'Formulas' tab active. The 'Prt3recode' worksheet is selected. Cell L2 contains the formula '=VLOOKUP(K2,agegr,2)'. The formula bar shows the formula being entered.

QID	Q1name	Day	Month	Season	Year	S	Yr	Yrgrp	Syr	Q2age	Agegr	Q3gen	Q4wt	Q5ht	Q6nat	Q7ed	Q8aff	Q9exp	Agestart	Q10peixp	Q11aq	Q11pou	Q11rum	Q11swi	Q
1	1	23	11	Dry	1984	D	84	80	D80	37	=VLOOKUP(K2,agegr,2)	Filipino	PhD		Non-government		9	28	Yes	Yes	No	No	No	No	N
2	2	21	10	Wet	1987	W	87	80	W80	34							6	28	No	No	No	No	No	Y	
3	3	4	7	Wet	1996	W	96	90	W90	25		Male	97	171	Thai	Master	Government	1	24	No	No	No	Yes	No	N
4	4	2	4	Dry	1991	D	91	90	D90	30		Female	56	167	Thai	Bachelor	Government	2	28	Yes	No	No	Yes	Yes	N
5	5	15	9	Wet	1990	W	90	90	W90	31		Female	61	161	Thai	Bachelor	Government	3	28	No	No	No	No	Yes	N
6	6	21	12	Dry	1983	D	83	80	D80	38		Female	50	160	Thai	Bachelor	Government	10	28	No	No	No	No	Yes	N
7	7	1	12	Dry	1982	D	82	80	D80	39		Male	89	170	Thai	Other	Government	11	28	No	No	Yes	Yes	No	N
8	8	22	11	Dry	1989	D	89	80	D80	32		Male	70	167	Chinese	Bachelor	Academic	4	28	No	No	No	No	Yes	N
9	9	20	1	Dry	1992	D	92	90	D90	29		Female	65	173	Indonesian	Bachelor	Other	5	24	No	No	No	Yes	No	Y
10	10	24	5	Wet	1993	W	93	90	W90	28		Male	65	168	Vietnamese	Bachelor	Academic	4	24	Yes	No	Yes	No	No	N
11	11	19	4	Dry	1987	D	87	80	D80	34		Female	51	156	Thai	PhD	Academic	6	28	No	No	No	Yes	No	N
12	12	19	4	Dry	1993	D	93	90	D90	28		Male	68	168	Laotian	Bachelor	Academic	4	24	Yes	No	No	No	No	N
13	13	26	1	Dry	1976	D	76	70	D70	45		Male	76	170	Malaysian	Bachelor	Non-government	17	28	No	No	No	No	No	N
14	14	1	1	Dry	1994	D	94	90	D90	27		Female	62	164	Thai	Bachelor	Government	3	24	No	No	No	No	No	N
15	15	19	10	Wet	1987	W	87	80	W80	34		Female	52	157	Thai	Master	Academic	6	28	No	No	No	No	No	N
16	16	8	8	Wet	1993	W	93	90	W90	28		Male	62	170	Thai	Bachelor	Academic	4	24	Yes	Yes	Yes	No	No	N
17	17	9	6	Wet	1976	W	76	70	W70	45		Male	77	170	Bhutanese	Bachelor	Non-government	17	28	No	No	No	No	No	N
18	18	7	8	Wet	1985	W	85	80	W80	36		Male	52	165	Cambodian	Master	Other	8	28	Yes	No	No	Yes	No	Y
19	19	23	9	Wet	1987	W	87	80	W80	34		Male	76	175	Laotian	Other	Academic	6	28	No	Yes	No	No	No	N
20	20	1	2	Dry	1985	D	85	80	D80	36		Male	52	165	Burmese	Bachelor	Other	8	28	No	Yes	No	No	No	N
21	21	18	3	Dry	1995	D	95	90	D90	26		Female	47	156	Thai	Master	Government	2	24	Yes	No	No	No	No	N
22	22	28	1	Dry	1993	D	93	90	D90	28		Female	60	160	Thai	Master	Government	4	24	Yes	No	Yes	No	No	N
23	23	9	11	Dry	1994	D	94	90	D90	27		Female	59.5	160	Thai	Bachelor	Government	3	24	Yes	No	No	Yes	No	N
24	24	31	1	Dry	1986	D	86	80	D80	35		Male	66	165	Cambodian	Master	Academic	7	28	No	No	Yes	No	No	N
25	25	29	8	Wet	1988	W	88	80	W80	33		Female	55	152	Thai	Bachelor	Academic	5	28	Yes	No	Yes	No	No	N



4. So, it will look for the same value as K2 in the agegr table and return a value in the second column in the same row of that table. Then copy the command for all records.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	
1	QID	Q1name	Day	Month	Season	Year	S	Yr	Yrgrp	Syr	Q2age	Agegr	Q3gen	Q4wt	Q5ht	Q6nat	Q7ed	Q8aff	Q9exp	Agestart	Q10epiexp	Q11aq	Q11pou	Q11rum	Q11swi	Q11
2	1		23	11	Dry	1984	D	84	80	D80	37	2	Female	50	160	Filipino	PhD	Non-government	9	28	Yes	Yes	No	No	No	Ni
3	2		21	10	Wet	1987	W	87	80	W80	34	2	Male	76	175	Burmese	Bachelor	Academic	6	28	No	No	No	No	No	Yi
4	3		4	7	Wet	1996	W	96	90	W90	25	1	Male	97	171	Thai	Master	Government	1	24	No	No	No	Yes	No	Ni
5	4		2	4	Dry	1991	D	91	90	D90	30	2	Female	56	167	Thai	Bachelor	Government	2	28	Yes	No	No	Yes	Yes	Ni
6	5		15	9	Wet	1990	W	90	90	W90	31	2	Female	61	161	Thai	Bachelor	Government	3	28	No	No	No	No	Yes	Ni
7	6		21	12	Dry	1983	D	83	80	D80	38	2	Female	50	160	Thai	Bachelor	Government	10	28	No	No	No	No	Yes	Ni
8	7		1	12	Dry	1982	D	82	80	D80	39	2	Male	89	170	Thai	Other	Government	11	28	No	No	Yes	Yes	No	Ni
9	8		22	11	Dry	1989	D	89	80	D80	32	2	Male	70	167	Chinese	Bachelor	Academic	4	28	No	No	No	No	Yes	Ni
10	9		20	1	Dry	1992	D	92	90	D90	29	1	Female	65	173	Indonesian	Bachelor	Other	5	24	No	No	No	Yes	No	Yi
11	10		24	5	Wet	1993	W	93	90	W90	28	1	Male	65	168	Vietnamese	Bachelor	Academic	4	24	Yes	No	Yes	No	No	Ni
12	11		20	4	Dry	1987	D	87	80	D80	34	2	Female	51	156	Thai	PhD	Academic	6	28	No	No	No	Yes	No	Ni
13	12		19	4	Dry	1993	D	93	90	D90	28	1	Male	68	168	Laotian	Bachelor	Academic	4	24	Yes	No	No	No	No	Ni
14	13		26	1	Dry	1982	D	82	70	D70	45	3	Male	76	170	Malaysian	Bachelor	Non-government	17	28	No	No	No	No	No	Ni
15	14		1	1	Dry	1994	D	94	90	D90	27	1	Female	62	164	Thai	Bachelor	Government	3	24	No	No	No	No	No	Ni
16	15		19	10	Wet	1987	W	87	80	W80	34	2	Female	52	157	Thai	Master	Academic	6	28	No	No	No	No	No	Ni
17	16		8	8	Wet	1993	W	93	90	W90	28	1	Male	62	170	Thai	Bachelor	Academic	4	24	Yes	Yes	Yes	No	No	Ni
18	17		9	6	Wet	1976	W	76	70	W70	45	3	Male	77	170	Bhutanese	Bachelor	Non-government	17	28	No	No	No	No	No	Ni
19	18		7	8	Wet	1985	W	85	80	W80	36	2	Male	52	165	Cambodian	Master	Other	8	28	Yes	No	No	Yes	No	Yi
20	19		23	9	Wet	1987	W	87	80	W80	34	2	Male	76	175	Laotian	Other	Academic	6	28	No	Yes	No	No	No	Ni
21	20		1	2	Dry	1985	D	85	80	D80	36	2	Male	52	165	Burmese	Bachelor	Other	8	28	No	Yes	No	No	No	Ni
22	21		18	3	Dry	1995	D	95	90	D90	26	1	Female	47	156	Thai	Master	Government	2	24	Yes	No	No	No	No	Ni
23	22		28	1	Dry	1993	D	93	90	D90	28	1	Female	60	160	Thai	Master	Government	4	24	Yes	No	Yes	No	No	Ni
24	23		9	11	Dry	1994	D	94	90	D90	27	1	Female	59.5	160	Thai	Bachelor	Government	3	24	Yes	No	No	Yes	No	Ni
25	24		31	1	Dry	1986	D	86	80	D80	35	2	Male	66	165	Cambodian	Master	Academic	7	28	No	No	Yes	No	No	Ni
26	25		29	8	Wet	1988	W	88	80	W80	33	2	Female	55	152	Thai	Bachelor	Academic	5	28	Yes	No	Yes	No	No	Ni

5. In the 'CBKcplt' worksheet, try to define tables necessary for using in recoding process, such as:

- ✓ Gender: male, female, other → 1, 2, 3
- ✓ Nationality: ..., ..., ... → 1, 2, 3, 4, 5, ...
- ✓ Yes, No → 1, 0
- ✓ Regularity: never, rarely, sometimes, often, always → 1, 2, 3, 4, 5
- ✓ Case, Not case → 1, 0

6. Now you can apply 'Vlookup' function for necessary variables to be recoded. Just change a referent table.

7. Another way to recode data is using 'Find and Replace'. For example, in order to replace 'Yes' and 'No' value with '1' and '0', respectively, select data range that you would like to recode the value. Then select 'Find & Select' function and choose 'Replace' function.

The screenshot shows the Microsoft Excel interface with the 'Find and Replace' dialog box open. The dialog box is set to 'Replace' and the 'Find what' field contains 'Yes' and the 'Replace with' field contains '1'. The data table has columns for various variables like Q8aff, Q9exp, Q10epiexp, etc., with rows representing different categories like Non-government, Academic, and Government.

8. Put 'Yes' in the 'Find what' box and '1' in the 'Replace' box. Click 'Replace All'. It will report how many values have been replaced.

The screenshot shows the Microsoft Excel interface with the 'Find and Replace' dialog box open. The dialog box is set to 'Replace' and the 'Find what' field contains 'Yes' and the 'Replace with' field contains '1'. A message box is displayed in the center of the screen, stating 'All done. We made 51 replacements.'

9. Now you can replace 'No' with '0' within the same data range. You may extend this function for other variables that use the same answering system (Yes, No). By the way, you can use the similar recoded system (1, 0) for other variables such as Case/Not case, Eat/Normal, and Not exercise/Exercise.

Hint: Variables to be recoded: Q3gen, Q6nqt, Q7ed, Q8aff, Q9exp, Q10epiexp, Q11aq, Q11pou, Q11rum, Q11swi, Q11hor, Q11dog, Q11wild, Q11oth, Q12prob, Q12vir, Q12bac, Q12ptz, Q12hel, Q12mul, Q13cin, Q13tv, Q13yout, Q13exer, Q13shop, Q13walk, Q13eat, Q13sleep, Q13sleep, Q13read, Q13cook, Q13medi, Q13talk, Q13chat, Q14sick, Q14slplss, Q14diar, Q14fev, Q14coug, Q14soto, Q14vom, Q14uri, Q14inj, Q14oth, Q16getsick, BMI_con1, BMI_con2, Eat, Exe, and Case.

Part 4 Basic data analysis using Pivot Table.

'PivotTable' help you in pre-analytical process. In addition, this function is also useful in analytical steps. Go to 'Prt4pivot' worksheet' and find what you can do more with PivotTable function. Review how to start a Pivot Table from the previous exercise if needed.

Functions in use:

- PivotTable

Exercise 4.1 Find a proportion of PhD graduate among participants.

Try to complete this table with PivotTable.

Education level	Percentage
Bachelor	
Master	
PhD	
Other	

1. Insert PivotTable in a new worksheet. Make sure you have all variables needed for analysis.

2. Drag and drop Q7ed in Row and Value boxes.

The screenshot shows the Microsoft Excel interface with a PivotTable and the PivotTable Fields task pane. The PivotTable is located in the range A3:L8 and has the following data:

Row Labels	Count of Q7ed
Bachelor	18
Master	7
Other	3
PhD	4
Grand Total	32

The PivotTable Fields task pane on the right shows the following configuration:

- Choose fields to add to report:** Q7ed (checked)
- Drag fields between areas below:**
 - Filters:** (empty)
 - Columns:** (empty)
 - Rows:** Q7ed
 - Values:** Count of Q7ed

3. Change a value setting. In the 'Show Values As' box, select '% of Grand Total'.

The screenshot shows the same Microsoft Excel interface as above, but with the 'Value Field Settings' dialog box open for the 'Count of Q7ed' field. The 'Show Values As' dropdown is set to '% of Grand Total'.

The 'Value Field Settings' dialog box shows the following configuration:

- Field Name:** Count of Q7ed
- Show Values As:** % of Grand Total
- Options:**
 - Show Data As Percent of Grand Total
 - Show Data As Percent of Parent Row Total
 - Show Data As Percent of Total
 - Show Data As Percent of Grand Average
 - Show Data As Percent of Average Parent
 - Show Data As Percent of Average
 - Show Data As Percent of Average Parent
 - Show Data As Percent of Average

The screenshot shows an Excel PivotTable with the following data:

Row Labels	Count of Q7ed
Bachelor	18
Master	7
Other	4
PhD	3
Grand Total	32

The 'Value Field Settings' dialog box is open, showing the following configuration:

- Source Name: Q7ed
- Custom Name: Percentage of Q7ed
- Summarize Values By: Show Values As
- Show values as: % of Grand Total

The screenshot shows the same PivotTable after the 'Value Field Settings' dialog box has been applied. The data is now displayed as percentages:

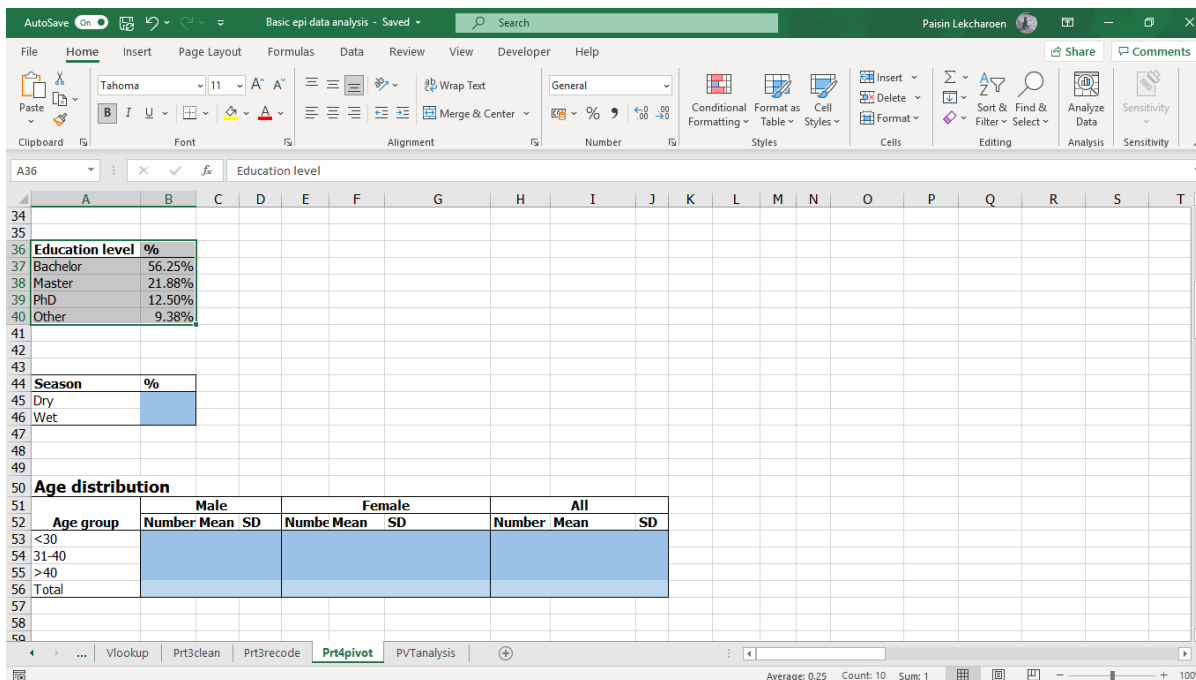
Row Labels	Percentage of Q7ed
Bachelor	56.25%
Master	21.88%
Other	9.38%
PhD	12.50%
Grand Total	100.00%

The PivotTable Fields task pane on the right shows the following configuration:

- Rows: Q7ed
- Values: Percentage of Q7ed



4. Put the output values in the table provided. Make sure to copy correct answer for each row.



Question 6 What is a proportion of participants who was born in wet season?

Answer: Click or tap here to enter text.

Exercise 4.2 Find a central tendency and dispersal of variables.

You would like to know average age of all participants and also among different genders.

Age group (year)	Male			Female			All		
	No	Mean	SD	No	Mean	SD	No	Mean	SD
<30									
31-40									
>40									
Total									



1. In the 'PVTanalysis', drag and drop Q2age into Row and Value boxes. The default value is Sum for continuous data.

The screenshot shows an Excel spreadsheet with a PivotTable. The PivotTable is located in the range B3:L21. The Row Labels are 'Q2age' and the Values are 'Sum of Q2age'. The data is as follows:

Q2age	Sum of Q2age
25	25
26	26
27	54
28	112
29	29
30	60
31	31
32	64
33	33
34	204
35	35
36	72
37	74
38	76
39	39
43	43
45	90
Grand Total	1067

The PivotTable Fields task pane on the right shows 'Q2age' selected for both Rows and Values. The default aggregation function is 'Sum of Q2age'.

2. You may want to see how many people contribute to each age year. Drag and drop Q2age into Value box again. Similarly, it shows the Sum of age and labels as 'Sum of Q2age2'. Set one of this value to Count.

The screenshot shows the same Excel spreadsheet as above, but with an additional column added to the PivotTable. The PivotTable is now in the range B3:L21. The Row Labels are 'Q2age', the first Values field is 'Sum of Q2age', and the second Values field is 'Sum of Q2age2'. The data is as follows:

Q2age	Sum of Q2age	Sum of Q2age2
25	25	25
26	26	26
27	54	54
28	112	112
29	29	29
30	60	60
31	31	31
32	64	64
33	33	33
34	204	204
35	35	35
36	72	72
37	74	74
38	76	76
39	39	39
43	43	43
45	90	90
Grand Total	1067	1067

The PivotTable Fields task pane on the right shows 'Q2age' selected for Rows. The first Values field is 'Sum of Q2age' and the second Values field is 'Sum of Q2age2'.

The screenshot shows an Excel spreadsheet with a PivotTable. The PivotTable is located in the range A3:K21. The PivotTable fields are:

Row Labels	Count of Q2age	Sum of Q2age2
25	1	25
26	1	26
27	2	54
28	4	112
29	1	29
30	2	60
31	1	31
32	2	64
33	1	33
34	6	204
35	1	35
36	2	72
37	2	74
38	2	76
39	1	39
43	1	43
45	2	90
Grand Total	32	1067

The PivotTable Fields task pane on the right shows the following configuration:

- Choose fields to add to report: Q2age (checked), Q3gen, Q4wt, Q5ht, Q6nat.
- Drag fields between areas below:
 - Filters: (empty)
 - Columns: Values
 - Rows: Q2age
 - Σ Values: Count of Q2age, Sum of Q2age2

3. And again. Drag and drop Q2age into Value box. But this time you may set the value to Average.

The screenshot shows the same Excel spreadsheet as above, but with an additional column in the PivotTable. The PivotTable fields are:

Row Labels	Count of Q2age	Sum of Q2age2	Average of Q2age
25	1	25	25
26	1	26	26
27	2	54	54
28	4	112	112
29	1	29	29
30	2	60	60
31	1	31	31
32	2	64	64
33	1	33	33
34	6	204	204
35	1	35	35
36	2	72	72
37	2	74	74
38	2	76	76
39	1	39	39
43	1	43	43
45	2	90	90
Grand Total	32	1067	1067

The PivotTable Fields task pane on the right shows the following configuration:

- Choose fields to add to report: Q2age (checked), Q3gen, Q4wt, Q5ht, Q6nat.
- Drag fields between areas below:
 - Filters: (empty)
 - Columns: Values
 - Rows: Q2age
 - Σ Values: Count of Q2age, Sum of Q2age2, Average of Q2age

The screenshot shows the 'Value Field Settings' dialog box in Microsoft Excel. The 'Source Name' is 'Q2age'. The 'Custom Name' is 'Average of Q2age'. Under 'Summarize value field by', the 'Average' option is selected. The background shows a PivotTable with the following data:

Row Labels	Count of Q2age	Sum of Q2age2	Sum of Q2age
25	1	25	25
26	1	26	26
27	2	54	54
28	4	112	112
29	1	29	29
30	2	60	60
31	1	31	31
32	2	64	64
33	1	33	33
34	6	204	204
35	1	35	35
36	2	72	72
37	2	74	74
38	2	76	76
39	1	39	39
43	1	43	43
45	2	90	90
Grand Total	32	1067	1067

The screenshot shows the same PivotTable after adding 'Average of Q2age' to the Values area. The 'Average of Q2age' column now displays the average of the 'Q2age' values for each row label. The 'Grand Total' for 'Average of Q2age' is 33.34375.

Row Labels	Count of Q2age	Sum of Q2age2	Average of Q2age
25	1	25	25
26	1	26	26
27	2	54	27
28	4	112	28
29	1	29	29
30	2	60	30
31	1	31	31
32	2	64	32
33	1	33	33
34	6	204	34
35	1	35	35
36	2	72	36
37	2	74	37
38	2	76	38
39	1	39	39
43	1	43	43
45	2	90	45
Grand Total	32	1067	33.34375



4. Now you can do similar way for SD.

The screenshot shows an Excel PivotTable with the following data:

Row Labels	Count of Q2age	Sum of Q2age2	Average of Q2age	StdDev of Q2age
25	1	25	25	#DIV/0!
26	1	26	26	#DIV/0!
27	2	54	27	0
28	4	112	28	0
29	1	29	29	#DIV/0!
30	2	60	30	0
31	1	31	31	#DIV/0!
32	2	64	32	0
33	1	33	33	#DIV/0!
34	6	204	34	0
35	1	35	35	#DIV/0!
36	2	72	36	0
37	2	74	37	0
38	2	76	38	0
39	1	39	39	#DIV/0!
43	1	43	43	#DIV/0!
45	2	90	45	0
Grand Total	32	1067	33.34375	5.252399605

5. You may see that it does not make any sense to get an average and a standard deviation for each age year because of the small sample. Let's try to group age into age groups. At the first value of the Row Labels column (A4), right click and select 'Group'.

The screenshot shows the same PivotTable as above, but with a context menu open over the first row label (25). The 'Group' option is highlighted. The PivotTable Fields task pane is visible on the right.

6. Group the age into 3 groups: ≤ 30 years, 31-40 years, and >40 years. Firstly, put 21 for the starting, 51 for the ending, and 10 for by. It will define 21-30 as the first group, 31-40 for the second group, and 41 to 50 for the last group.

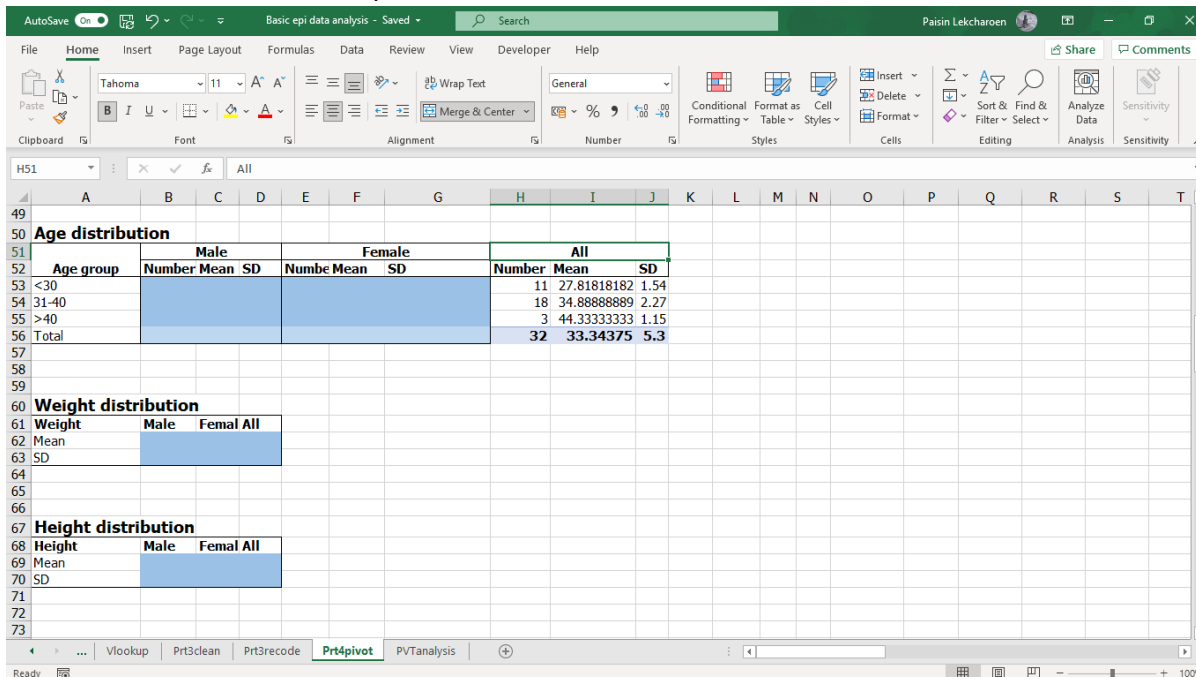
The screenshot shows an Excel PivotTable with the following data:

Row Labels	Count of Q2age	Sum of Q2age2	Average of Q2age	StdDev of Q2age
25	25	25	25	#DIV/0!
26	26	26	26	#DIV/0!
27	54	27	27	0
28	112	28	28	0
29	29	29	29	#DIV/0!
30	60	30	30	0
31	31	31	31	#DIV/0!
32	64	32	32	0
33	33	33	33	#DIV/0!
34	204	34	34	0
35	1	35	35	#DIV/0!
36	2	72	36	0
37	2	74	37	0
38	2	76	38	0
39	1	39	39	#DIV/0!
43	1	43	43	#DIV/0!
45	2	90	45	0
Grand Total	32	1067	33.34375	5.252399605

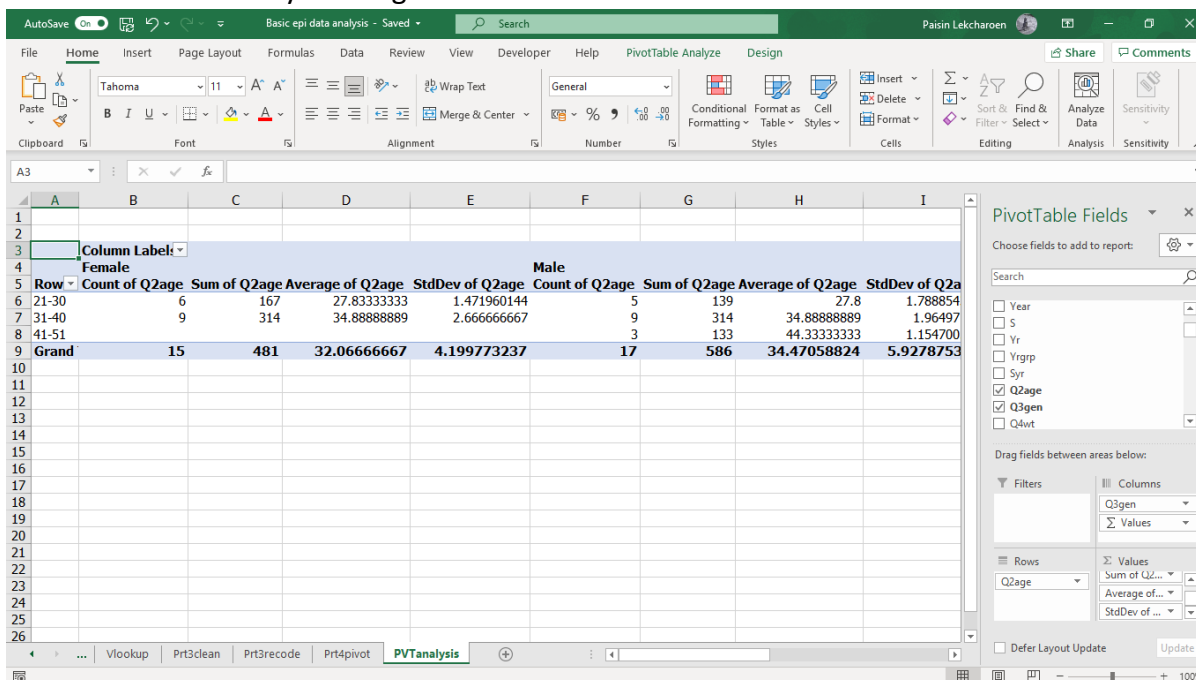
The screenshot shows the same Excel PivotTable after grouping. The data is summarized as follows:

Row Labels	Count of Q2age	Sum of Q2age2	Average of Q2age	StdDev of Q2age
21-30	11	306	27.81818182	1.53741223
31-40	18	628	34.88888889	2.272311311
41-51	3	133	44.33333333	1.154700538
Grand Total	32	1067	33.34375	5.252399605

7. Put the result into the table provided.



8. Now you would like to see those values of male and female. So, drag and drop Q3gen into Column box and you will get the results.



9. Find averages and standard deviations for weight, height, and BMI and answer the questions below.

Question 7 Which gender has higher average weight?

Answer: Click or tap here to enter text.

Question 8 How much different among average height of male and female?

Answer: Click or tap here to enter text.



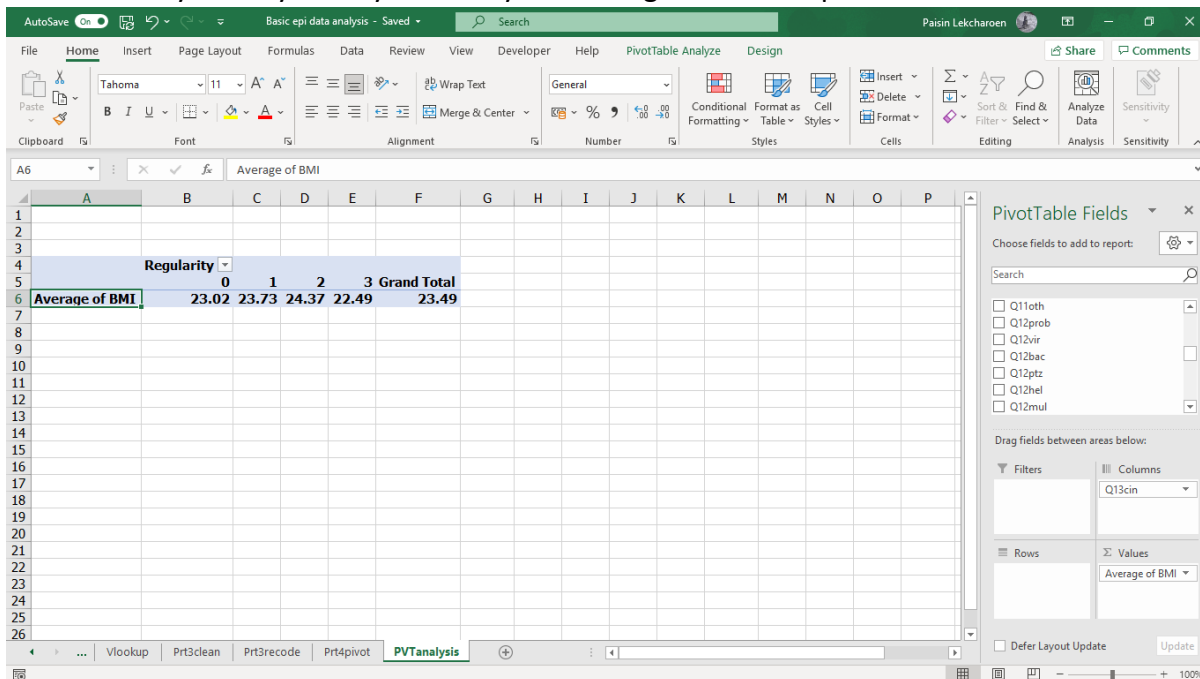
Question 9 Which gender has lower variation of BMI?

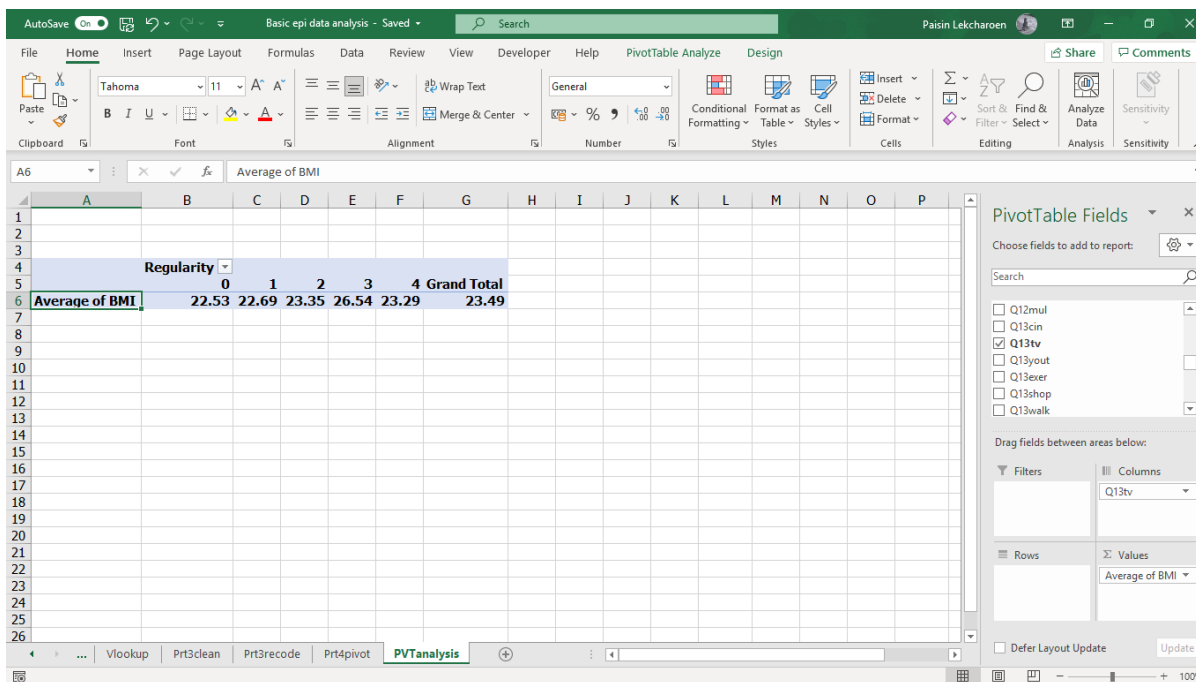
Answer: Click or tap here to enter text.

Exercise 4.3 Find average BMI among different regularities of activities.

Activity	Regularity					All
	Never	Rarely	Sometimes	Often	Always	
Go to cinema						
Watch TV						
Watch YouTube						
Physical exercise						
Go shopping						
Go walking						
Eat						
Sleep						
Read						
Cook						
Meditation						
Talk with friends						
Chat						

1. You can do it in the same way that you find these values in the previous exercise. Try to do it for every activity one by one. Fill your finding in the table provided.





2. Now you would like to compare score and BMI among those who have higher regularity of each activity. Again, you need to do it one by one for each activity.

Activity	Cumulative score	Average score	SD score	Average BMI
Go to cinema				
Watch TV				
Watch YouTube				
Physical exercise				
Go shopping				
Go walking				
Eat				
Sleep				
Read				
Cook				
Meditation				
Talk with friends				
Chat				

Values needed includes cumulative score, average score and SD, and average BMI.

Drop Q13cin into Filter box once and into Value box 3 times. Then drop BMI into Value box once.



AutoSave On Basic epi data analysis - Saved - Search Painsin Lekcharoen

File Home Insert Page Layout Formulas Data Review View Developer Help PivotTable Analyze Design

Get & Transform Data Queries & Connections Data Types Sort & Filter Data Tools Forecast

Sum of Q13cin

	A	B	C	D	E	F	G
1							
2	Q13cin	(All)					
3							
4	Sum of Q13cin	Sum of Q13cin2	Sum of Q13cin3	Sum of Q13cin4			
5	33	33	33	33			
6							
7							
8							
9							
10							
11							
12							
13							
14							
15							
16							
17							
18							
19							
20							
21							
22							
23							
24							
25							
26							

PivotTable Fields

Choose fields to add to report:

Search

Q12mul
 Q13cin
 Q13tv
 Q13yout

Drag fields between areas below:

Filters: Q13cin Columns: Σ Values

Rows: Σ Values
Sum of Q13cin
Sum of Q13cin2
Sum of Q13cin3
Sum of Q13cin4

Defer Layout Update Update

Vlookup Prt3clean Prt3recode Prt4pivot **PVTanalysis**

3. Set value for those in the Value box:

- Cumulative score of Q13cin.

AutoSave On Basic epi data analysis - Saved - Search Painsin Lekcharoen

File Home Insert Page Layout Formulas Data Review View Developer Help PivotTable Analyze Design

Get & Transform Data Queries & Connections Data Types Sort & Filter Data Tools Forecast

Sum of Q13cin

	A	B	C	D	E	F	G
1							
2	Q13cin	(All)					
3							
4	Sum of Q13cin	Sum of Q13cin2	Sum of Q13cin3	Sum of Q13cin4			
5	33	33	33	33			
6							
7							
8							
9							
10							
11							
12							
13							
14							
15							
16							
17							
18							
19							
20							
21							
22							
23							
24							
25							
26							

PivotTable Fields

Choose fields to add to report:

Search

Q12mul
 Q13cin
 Q13tv
 Q13yout

Drag fields between areas below:

Filters: Q13cin Columns: Σ Values

Rows: Σ Values
Sum of Q13cin
Sum of Q13cin2
Sum of Q13cin3
Sum of Q13cin4

Defer Layout Update Update

Vlookup Prt3clean Prt3recode Prt4pivot **PVTanalysis**

Value Field Settings

Source Name: Q13cin
Custom Name: Cumulative score of Q13cin

Summarize Values By Show Values As

Summarize value field by
Choose the type of calculation that you want to use to summarize data from the selected field

Sum
Count
Average
Max
Min
Product

Number Format OK Cancel

➤ Average score of Q13cin.

The screenshot shows an Excel PivotTable with the following data:

Q13cin	Sum of Q13cin2	Sum of Q13cin3	Sum of Q13cin4
(All)	33	33	33

The 'Value Field Settings' dialog box is configured as follows:

- Source Name: Q13cin
- Custom Name: Average of Q13cin
- Summarize Values By: Show Values As
- Summarize value field by: Average

➤ SD of Q13cin.

The screenshot shows an Excel PivotTable with the following data:

Q13cin	Average of Q13cin	Sum of Q13cin3	Sum of Q13cin4	
(All)	33	1.03125	33	33

The 'Value Field Settings' dialog box is configured as follows:

- Source Name: Q13cin
- Custom Name: StdDev of Q13cin3
- Summarize Values By: Show Values As
- Summarize value field by: StdDev

➤ Average BMI

The screenshot shows the Microsoft Excel interface with a PivotTable. The PivotTable is based on the 'BMI' source and has 'Q13cin' as the filter. The PivotTable fields are: Rows: Cumulative of Q13cin, Average of Q13cin, StdDev of Q13cin3, Sum of BMI; Columns: Values. The Value Field Settings dialog box is open, showing 'Source Name: BMI' and 'Custom Name: Average of BMI'. The 'Summarize Values By' section is set to 'Average'. The PivotTable data is as follows:

Cumulative of Q13cin	Average of Q13cin	StdDev of Q13cin3	Sum of BMI
33	1.03125	0.966682888	751.6552356

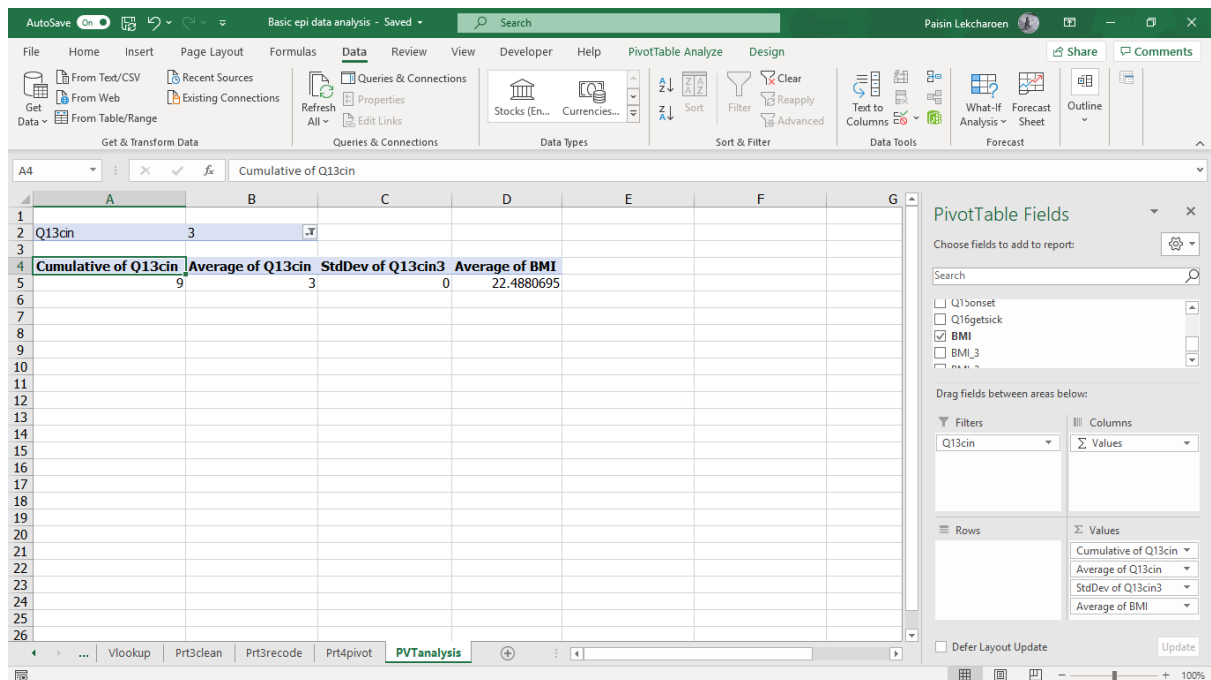
4. However, these values are of all regularity of this activity. But you need only those values of who has higher activity (often/always).

Click the arrowhead of the variable in the Filter box. Select only 3 (and 4) which represent Often (and Always) in regularity.

The screenshot shows the same Excel interface as above, but with the filter dropdown menu for 'Q13cin' open. The dropdown menu shows a search bar and a list of items: 0, 1, 2. The 'Average of BMI' value in the PivotTable is now 23.48922611. The PivotTable data is as follows:

StdDev of Q13cin3	Average of BMI
0.966682888	23.48922611

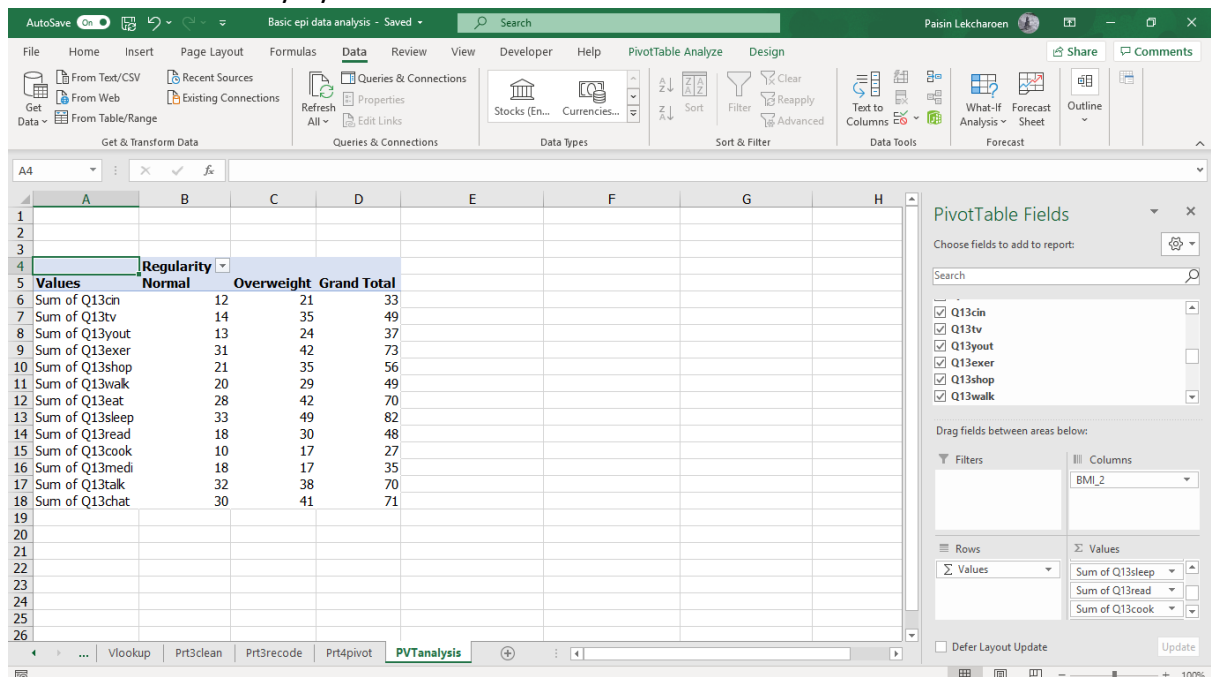
- You will get the results. Put your results into the table provided and do similar thing to the rest activities.



- Now you are going to see which are the top 5 activities of those who has normal and overweight condition.

Firstly, you have to compare cumulative score for each activity in those who has normal condition and also for those who has overweight.

Drop 'BMI_2' into Column box and all activities into Value box. You will have cumulative score for each activity by BMI conditions.



- Copy the results into another tale. Then sort the score for both BMI conditions to rank top 5 activities for each group. Use the results for the next exercise.



Exercise 4.4 How to find odds ratio.

From the previous exercise, you find which activities have higher rank for those who has overweight. So, you would like to quantify if there any association among activity and BMI condition.

The basic measures of frequency include risk ratio, odds ratio, prevalent ratio, and prevalent odds ratio upon study design.

Question 10 what is the study design for this study?

Answer: [Click or tap here to enter text.](#)

Odds ratio is a measure that can be used in any study designs. It compares the odds of different outcomes that are likely to be exposed with the exposure of interest. In this study, you are going to compare the odds of those who has overweight has exposed to certain regularity of activity with those who has normal condition.

		BMI condition		
		Overweight	Normal	
Regularity of activity	High	a	b	a+b
	Low	c	d	c+d
		a+c	b+d	a+b+c+d

Odds of exposed among overweight = Proportion of overweight having high regularity/Proportion of overweight having low regularity

$$= [a/(a+c)]/[c/(a+c)]$$

$$= a/c$$

Odds of exposed among normal = Proportion of normal having high regularity/Proportion of normal having low regularity

$$= [b/(b+d)]/[d/(b+d)]$$

$$= b/d$$

Odds ratio = Odds of exposed among overweight/Odds of exposed among normal

$$= (a/c)/(b/d) = ad/bc$$

1. You have observed that sleeping activity is the same rank among those who has overweight and normal BMI. So, you would like to compare the next rank. You pick up exercise activity as the next exposure of interest.

Recode regularity into two groups: low and high. So, you can use it in the analysis more comfortably. In this case, exercising regularity has already been classified into two group under the variable 'Exe'.

Drop 'Exe' into Row box and 'BMI_2' into Column box. Then drop 'QID' into Value box and change it to a Count. So, you will get a 2x2 table.

The screenshot shows an Excel PivotTable with the following data:

Count of QID	Regularly	Overweight	Grand Total
Exercise	7	7	14
Not exercise	6	12	18
Grand Total	13	19	32

The PivotTable Fields task pane on the right shows the following configuration:

- Filters: (empty)
- Columns: BMI_2
- Rows: Exe
- Values: Count of QID

Question 11 What is the odds ratio for this table?

Answer: Click or tap here to enter text.

Question 12 What is your interpretation?

Answer: Click or tap here to enter text.

- Generally, we interpret the table above as those who has regular exercise associated with normal BMI. In causal relationship sense, we may expect to see the association among possible risk factor, in this case, less regular exercise, with problem, in this case, overweight. You can re-arrange the order of both values in both row and column.

The screenshot shows the same Excel PivotTable as above, but with a 'Sort A to Z' dialog box open for the 'Regularly' column. The dialog box has the following options:

- Sort A to Z
- Sort Z to A
- More Sort Options...
- Clear Filter From "BMI_2"
- Label Filters: (empty)
- Value Filters: (empty)
- Search: (empty)
- Checked items: (Select All), Normal, Overweight

The top screenshot shows a PivotTable with the following data:

Row Labels	Overweight	Normal	Grand Total
Count of QID	7	14	
Sort A to Z	6	18	
Sort Z to A	13	32	

The bottom screenshot shows the same PivotTable with 'Exe' as a filter:

Count of QID	Overweight	Normal	Grand Total
Not exercise	12	6	18
Exercise	7	7	14
Grand Total	19	13	32

It may have more sense to be interpret. However, the calculation remains the same.

3. Transfer the values to the 2x2 table provided in Prt4pivot worksheet. Other calculations for 95% confidence interval are also provided.

The screenshot shows an Excel spreadsheet with the following data:

2x2 Table between activity and BMI class		
Exercise	Overwe	Normal
136 Never, rarely, som	12	6
137 Often, Always	7	7
139 Odds	1.71	0.86
140 Odds ratio	2	
141 Ln(OR)	0.69315	
142 Upper 95% CI	8.39571	
143 Lower 95% CI	0.47643	
Sleep		
Overwe	Normal	
146 Often, Always		
147 Never, rarely, sometimes		
149 Odds	#DIV/0!	####
150 Odds ratio	#DIV/0!	
151 Ln(OR)	#DIV/0!	
152 Upper 95% CI	#DIV/0!	
153 Lower 95% CI	#DIV/0!	
Eat		
Overwe	Normal	
156 Often, Always		
157 Never, rarely, sometimes		

4. Find OR and 95%CI for other activities in the top 5 rank of those who has overweight. Complete all tables and try to answer the following questions.

Question 13 What is the odds of having regular sleep (often/always) and having overweight?

Answer: Click or tap here to enter text.

Question 14 Do participants who have overweight and normal BMI have different eating regularity?

Answer: Click or tap here to enter text.

Question 15 Please interpret the likelihood of regularity of talking with friend and being overweight?

Answer: Click or tap here to enter text.

Part 5 Basic data analysis

Excel has a power function for some statistical tests, e.g., 'Data Analysis' function.

Functions in use:

- Data Analysis
 - ✓ Descriptive Statistics
 - ✓ F-Test
 - ✓ t-Test
 - ✓ ANOVA



Exercise 5.1 Descriptive statistics

For some continuous data, descriptive statistics can be obtained by performing a basic data analysis using 'Data Analysis function.

- Go to 'Prt5dataana' worksheet. Find the measures of central tendency and variability of age, weight, height, and BMI of participants.

1.1. Go to 'Data' ribbon. Select 'Data Analysis' function

The screenshot shows the Microsoft Excel interface with the 'Data' ribbon selected. The 'Data Analysis' dropdown menu is open, displaying various analysis tools. The 'Data Analysis Tools' option is highlighted, which is the first step in performing a descriptive statistical analysis.

1.2. Select 'Descriptive Statistics' tool from 'Data Analysis' window

The screenshot shows the 'Data Analysis' dialog box open over the 'Prt5dataana' worksheet. The 'Descriptive Statistics' option is selected in the list of analysis tools. The dialog box also shows the 'Input Range' and 'Output Range' fields, which are essential for specifying the data to be analyzed and where the results should be placed.



1.3. Click the upward arrow to open window for specifying input range of data needed for analysis.

The screenshot shows the 'Descriptive Statistics' dialog box in Microsoft Excel. The 'Input Range' field is empty, and the 'Columns' radio button is selected. The dialog box is overlaid on a spreadsheet with columns labeled QID, Q1name, Day, Month, Season, Year, S, Yr, Yrgrp, Syr, Q2age, Q3gen, Q4wt, Q5ht, Q6nat, Q7ed, Q8aff, Q9exp, and Ag. The spreadsheet data includes columns for demographic and clinical information.

1.4. Select a range of data. For this exercise, choose a range of 'Q2age' first. You will see a dot borderline around the range you have selected and the range is shown in the descriptive statistics window. Then click on arrowhead at the end of the box.

The screenshot shows the 'Descriptive Statistics' dialog box in Microsoft Excel. The 'Input Range' field now contains '\$K\$1:\$K\$33', indicating that the 'Q2age' column has been selected. The dialog box is overlaid on the same spreadsheet as in the previous image. The spreadsheet data includes columns for demographic and clinical information.

1.5. In this case, data is grouped by column. You also include a heading in the data range, so check on 'Labels in First Row'.

The screenshot shows the 'Descriptive Statistics' dialog box in Microsoft Excel. The 'Input Range' is set to '\$K\$1:\$K\$33'. The 'Grouped By' option is set to 'Columns'. The 'Labels in First Row' checkbox is checked. The 'Output options' section shows 'New Worksheet By' selected. The 'Confidence Level for Means' is set to 95%.

1.6. Specify the output range somewhere in this sheet under the range of data.

The screenshot shows the 'Descriptive Statistics' dialog box in Microsoft Excel. The 'Input Range' is set to '\$A\$39:\$B\$55'. The 'Output Range' is set to '\$A\$39:\$B\$55'. The 'Labels in First Row' checkbox is checked. The 'Output options' section shows 'New Worksheet By' selected.

Descriptive summary for age, weight, height, and BMI of participants
 Function: Descriptive Analysis - Descriptive Statistics

1.7. Check on 'Summary Statistics' and specify confidence level equal to 95. Click 'OK'.

The screenshot shows the 'Descriptive Statistics' dialog box in Microsoft Excel. The 'Input Range' is set to '\$K\$1:\$K\$33'. The 'Output Range' is set to '\$A\$39:\$B\$50'. The 'Confidence Level for Mean' is set to 95%. The 'Summary statistics' checkbox is checked. The spreadsheet data includes columns for Age, Weight, Height, and BMI of participants.

The screenshot shows the results of a Descriptive Analysis in Microsoft Excel. The spreadsheet displays a table of statistics for age, weight, height, and BMI. The 'Q2age' is highlighted in the table.

Statistic	Value
Mean	33.34375
Standard Error	0.928502
Median	34
Mode	34
Standard Deviation	5.2524
Sample Variance	27.5877
Kurtosis	-0.07565
Skewness	0.550853
Range	20
Minimum	25
Maximum	45
Sum	1067
Count	32
Confidence Level	1.893692

1.8. Then, you will have measures of central tendency and dispersal of age. So, perform this calculation but for weight, height, and BMI of the participants.

Question 16 What is the average weight of the participants?

Answer: Click or tap here to enter text.

Question 17 What is a standard variation of height of participants?

Answer: Click or tap here to enter text.



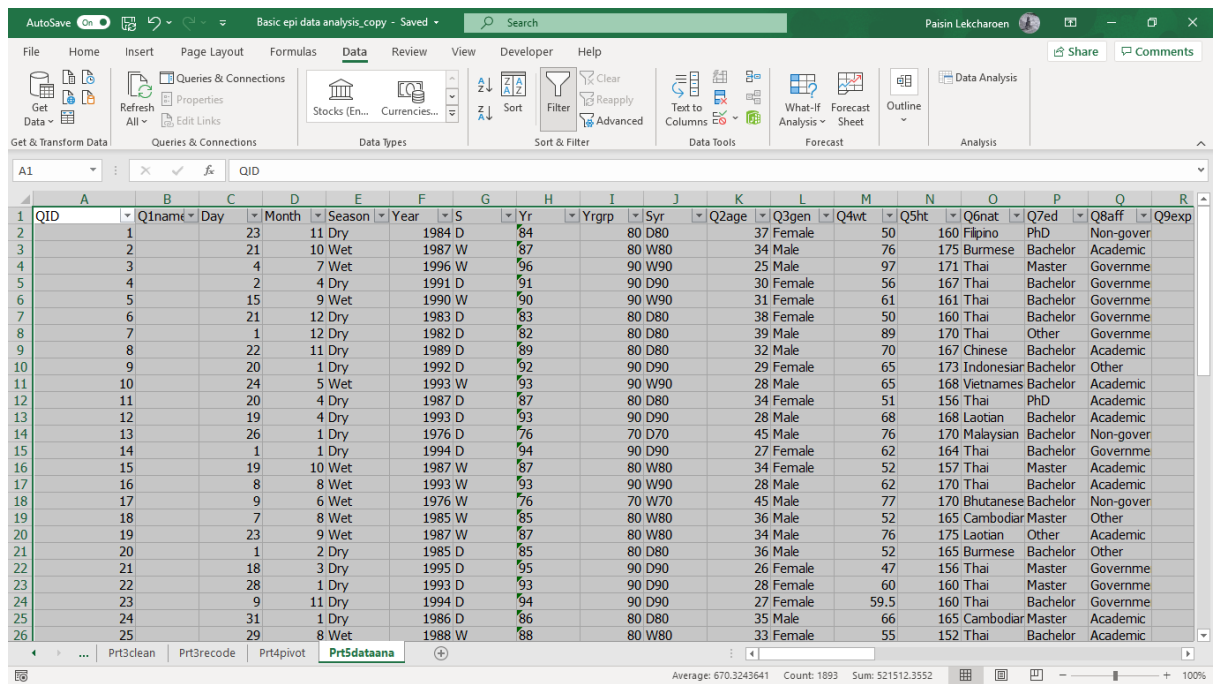
Question 18 What direction is the skewness of BMI?

Answer: Click or tap here to enter text.

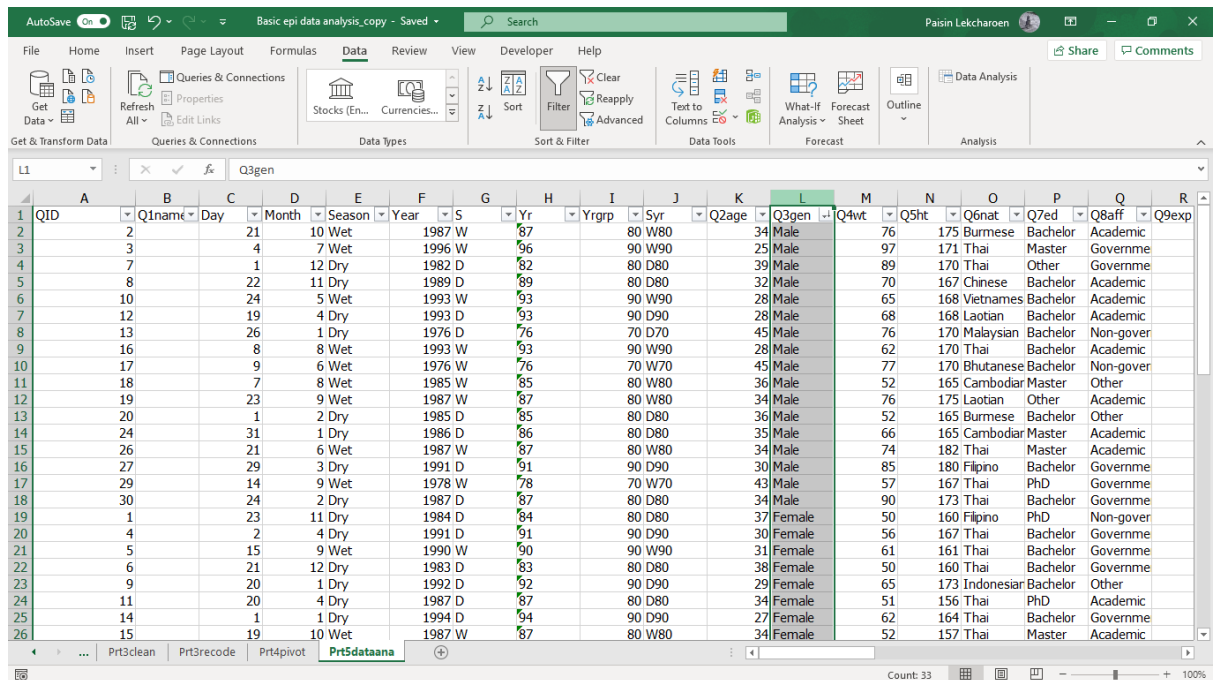
Exercise 5.2 Comparison of means in 2 independent groups.

In this exercise, you would like to compare some characteristics among participants. In this case, you will see whether male is higher than female or not, or significantly.

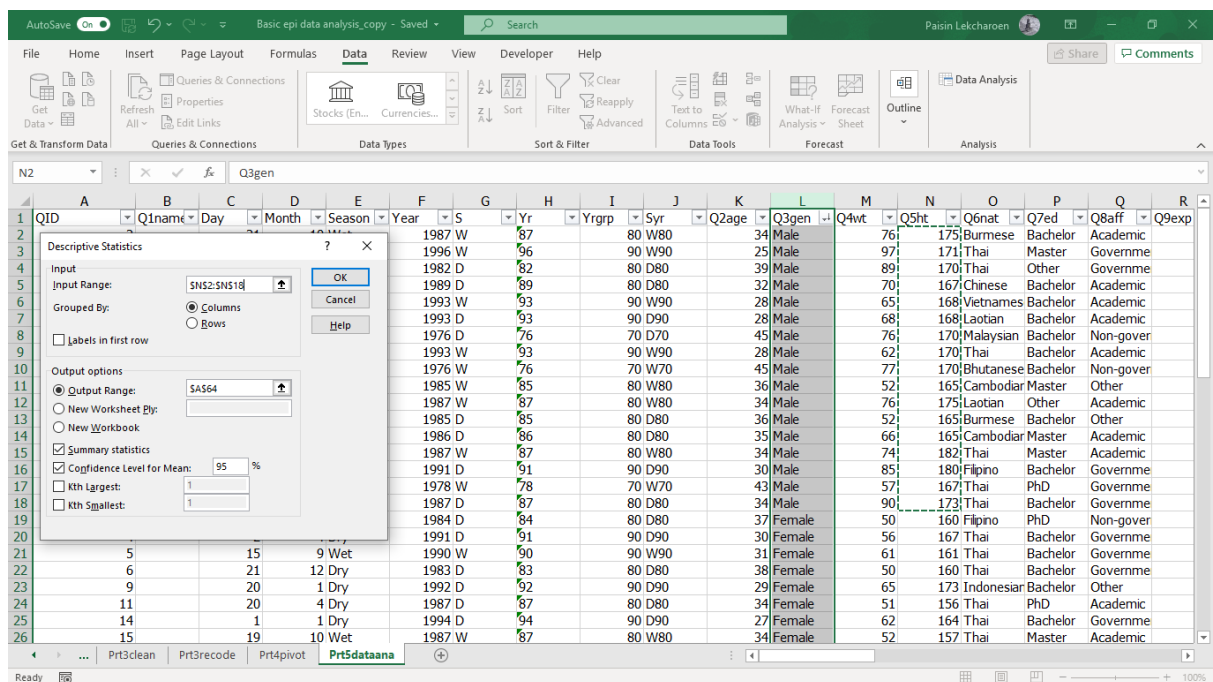
1. Perform ‘Descriptive Statistics’ function to see the average height and SD among male and female.
 - 1.1. Highlight the range of the original data. Go to ‘Data’ ribbon and select ‘Filter’.



1.2. Sort 'Q3gen' either 'A to Z' or 'Z to A'. In this case, 'Z to A' is used.



1.3. The input range comprises only 'Q5ht' only for 'Male' first. Change the heading of the output table to Male height.



	Male: height
61	Comparison of mean in 2 independent groups
62	Is male higher than female significantly?
63	
64	Male: height
65	
66	Mean 170.6471
67	Standard Error 1.209464
68	Median 170
69	Mode 170
70	Standard Deviation 4.986747
71	Sample Variance 24.86765
72	Kurtosis 0.54254
73	Skewness 1.019035
74	Range 17
75	Minimum 165
76	Maximum 182
77	Sum 2901
78	Count 17
79	Confidence Level 2.563949

1.4. Do the same thing for Female.

	Male: height	Female height
61	Comparison of mean in 2 independent groups	
62	Is male higher than female significantly?	
63		
64	Male: height	Female height
65		
66	Mean 170.6471	Mean 158.7333
67	Standard Error 1.209464	Standard Error 1.747016
68	Median 170	Median 160
69	Mode 170	Mode 160
70	Standard Deviation 4.986747	Standard Deviation 6.766162
71	Sample Variance 24.86765	Sample Variance 45.78095
72	Kurtosis 0.54254	Kurtosis 0.974286
73	Skewness 1.019035	Skewness 0.007724
74	Range 17	Range 28
75	Minimum 165	Minimum 145
76	Maximum 182	Maximum 173
77	Sum 2901	Sum 2381
78	Count 17	Count 15
79	Confidence Level 2.563949	Confidence Level 3.746976

1.5. So, you can compare all measures those of male and female.

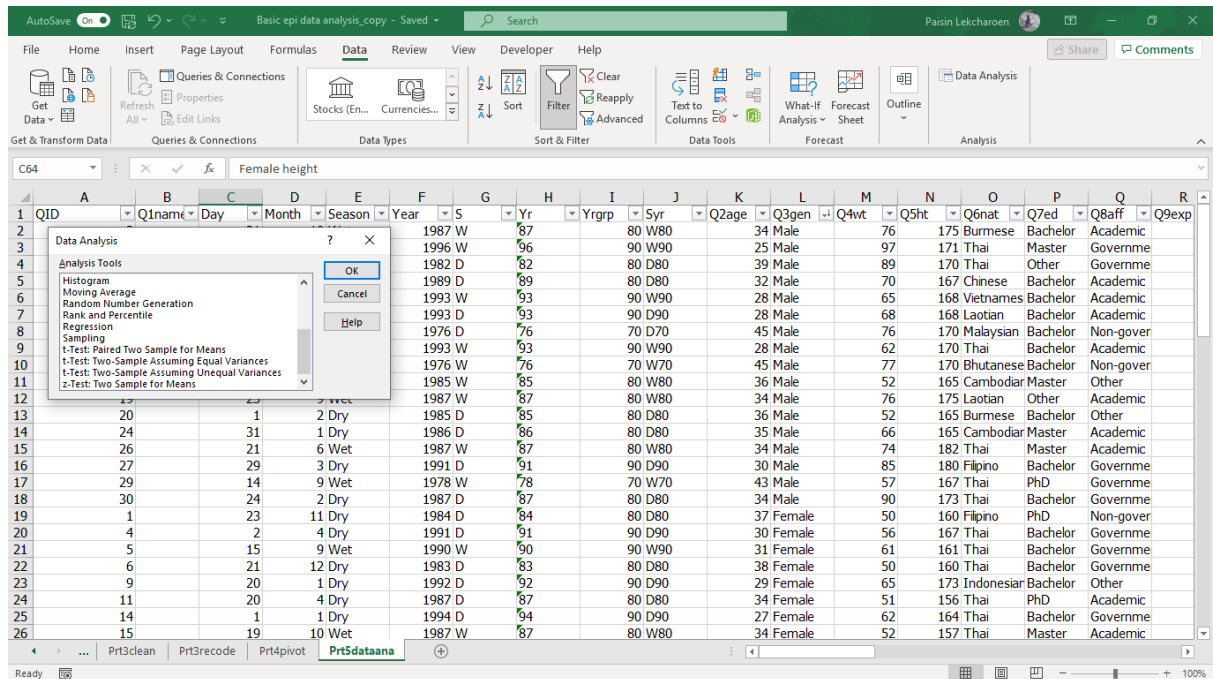
Question 19 Which gender has higher average height?

Answer: Click or tap here to enter text.

- It is shown that male has higher average height, but female has higher height variation. Are these findings exactly true? To compare the average height, in this case the mean, among 2 independent groups – male and female, the t-Test is appropriate and is going to be performed.



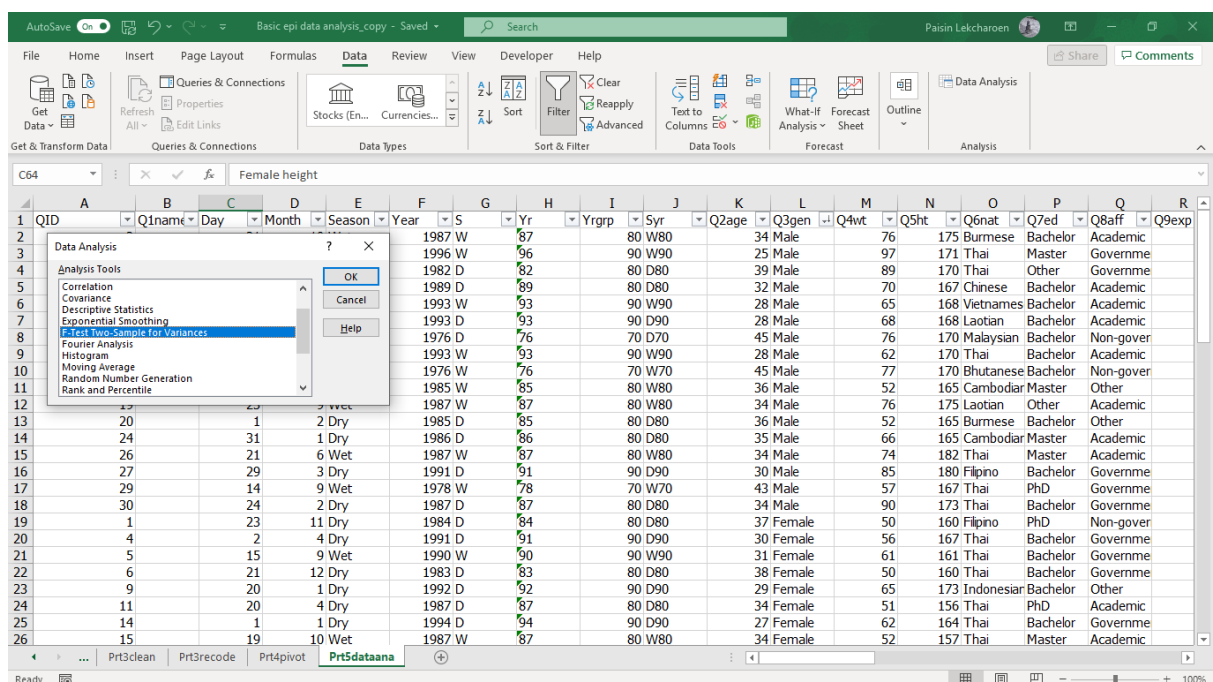
2.1. Go to 'Data Analysis' function again. In this time, you switch to use a function 't-Test'. But three different technics are provided.



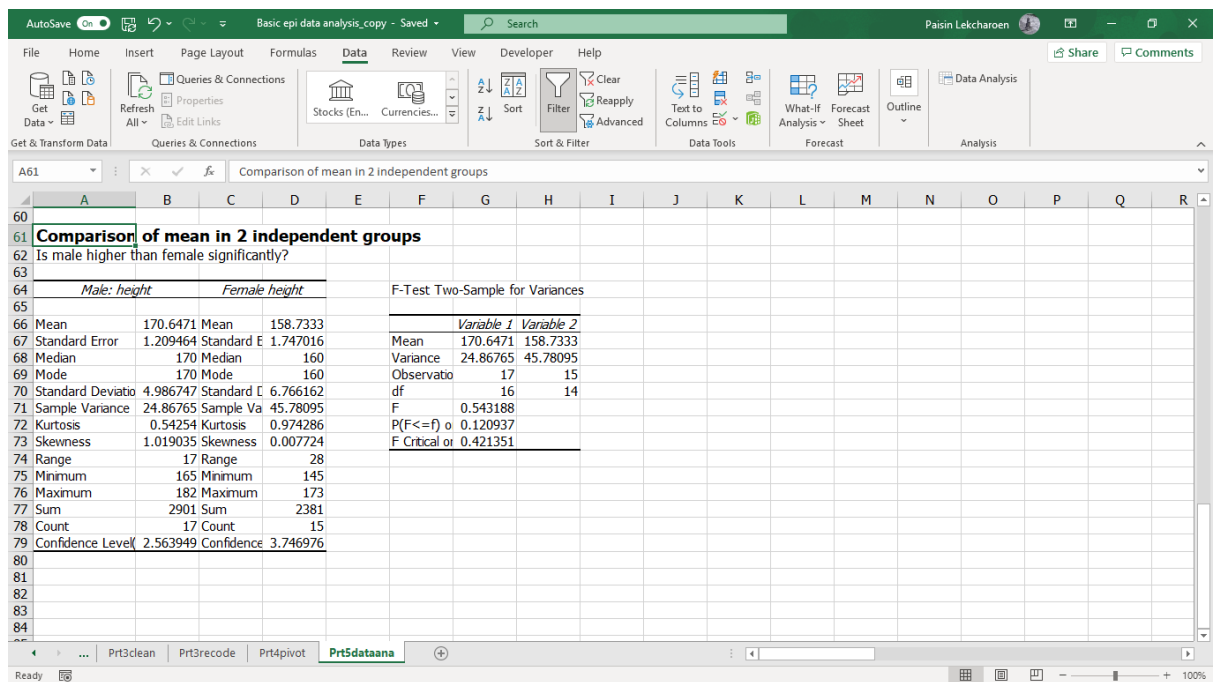
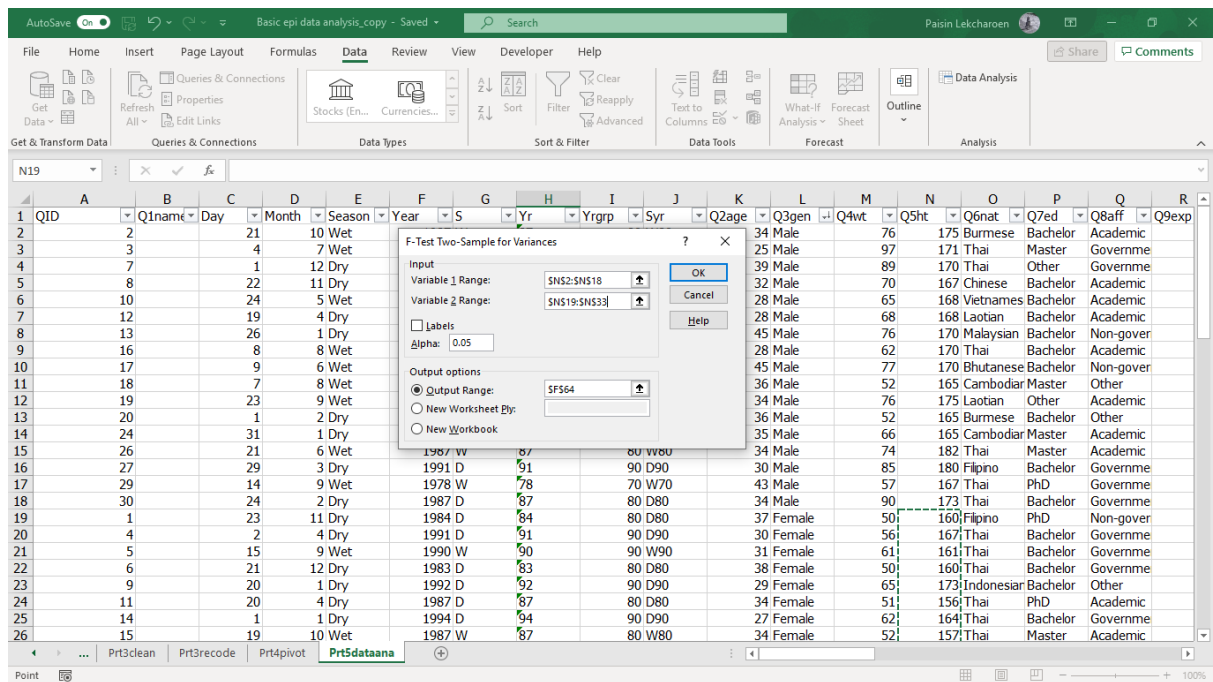
2.2. It is not a paired group so you cannot choose 't-Test: Paired Two Sample for Means'. Thus, your options are 't-Test: Two Sample Assuming Equal or Unequal Variances'. Even though you know the variances of the two groups, you still do not know whether it is exactly different.

3. Use F-Test to find out whether two sample variances are equal or unequal.

3.1. In 'Data Analysis' window, select to open 'F-Test Two-Sample for Variances' tools.



3.2. Put a range of male height in 'Variable 1 Range' and female height in 'Variable 2 Range'. Set 'Alpha' as 0.05. indicate 'Output Range' at the range of cells beside the results from descriptive statistics.



3.3. It calculates means and variances for variable 1 (male height) and variable 2 (female height). F is 0.543188 and F Critical one-tail is 0.42135. When F is less than 1, we will reject the null hypothesis (Two variances are equal) when $F < F$ Critical one-tail. So, we have not enough evidence to reject the null hypothesis.



4. From the previous step, it is suggested a 't-Test: Two Samples Assuming Equal Variance'.
- 4.1. In the 'Data Analysis' window, select 't-Test: Two-Sample Assuming Equal Variances'

The screenshot shows the Microsoft Excel interface with the 'Data Analysis' task pane open. The 't-Test: Two-Sample Assuming Equal Variances' option is selected in the list of analysis tools. The background spreadsheet contains data for a comparison of mean in 2 independent groups, with columns for QID, Q1name, Day, Month, Season, Year, S, Yr, Yrgrp, Syr, Q2age, Q3gen, Q4wt, Q5ht, Q6nat, Q7ed, Q8aff, and Q9exp.

- 4.2. Put male height in the variable 1 range input and female height in the variable 2 range input. No hypothesized mean difference needed. Set alpha as 0.05 and put the output range beside the previous result.

The screenshot shows the 't-Test: Two-Sample Assuming Equal Variances' dialog box in Microsoft Excel. The 'Variable 1 Range' is set to '\$N\$2:\$N\$18', the 'Variable 2 Range' is '\$O\$2:\$O\$33', and the 'Alpha' is set to 0.05. The 'Output Range' is set to '\$J\$64'. The dialog box also includes options for 'Labels', 'Hypothesized Mean Difference', and 'Output options' (Output Range, New Worksheet Ply, New Workbook).

Comparison of mean in 2 independent groups

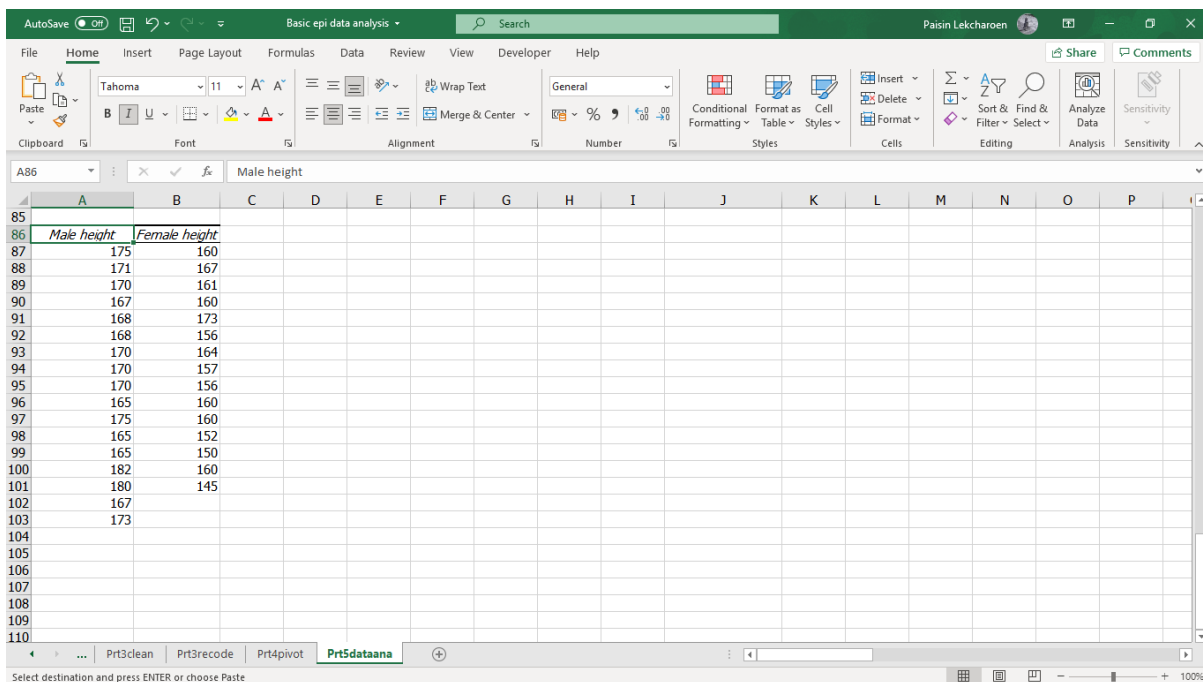
Is male higher than female significantly?

	Male: height	Female height	F-Test Two-Sample for Variances		t-Test: Two-Sample Assuming Equal Variances			
			Variable 1	Variable 2	Variable 1	Variable 2		
Mean	170.6471	158.7333	Mean	170.6471	158.7333	Mean	170.6471	158.7333
Standard Error	1.209464	1.747016	Variance	24.86765	45.78095	Variance	24.86765	45.78095
Median	170	160	Observations	17	15	Observations	17	15
Mode	170	160	df	16	14	Pooled Variance	34.62719	
Standard Deviation	4.986747	6.766162	F	0.543188		Hypothesized Mean	0	
Sample Variance	24.86765	45.78095	P(F<=f) o	0.120937		df	30	
Kurtosis	0.54254	0.974286	F Critical o	0.421351		t Stat	5.715235	
Skewness	1.019035	0.007724				P(T<=t) one-tail	1.55E-06	
Range	17	28				t Critical one-tail	1.697261	
Minimum	165	145				P(T<=t) two-tail	3.1E-06	
Maximum	182	173				t Critical two-tail	2.042272	
Sum	2901	2381						
Count	17	15						
Confidence Level	2.563949	3.746976						

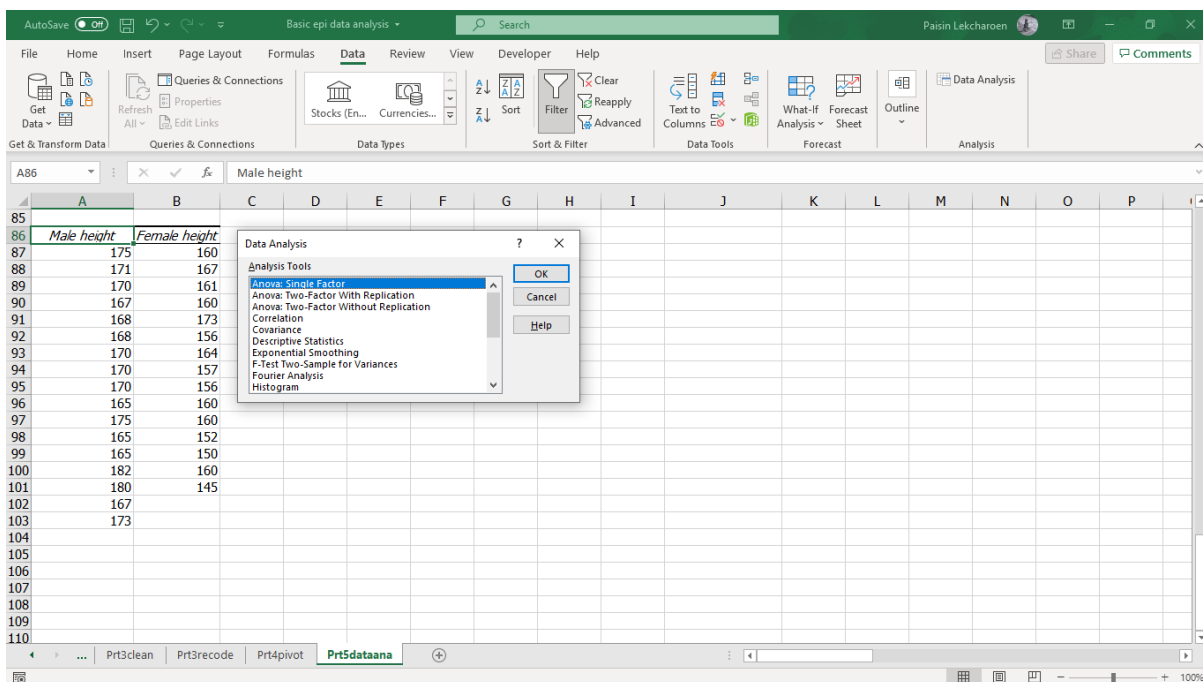
- 4.3. If you would like to know whether average male height is significantly higher than those of female or not, it is a one-tail test.
- 4.4. Your null hypothesis (H_0) is average male height is less than or equal to average female height and the alternative hypothesis (H_a) is average male height is greater than that of female.
- 4.5. The t stat (5.715235) is greater than the t Critical one-tail (1.697261). So, you have enough evidence to reject the null hypothesis. Therefore, you reject that average male height is less than or equal to average female height.
- 4.6. The P-value (one-tail) is 1.55×10^{-6} , that is less than 0.05. It also confirms that this difference is statistically significant.

5. You can also perform an ANOVA test.

5.1. Rearrange values of male and female height in a separated table.



5.2. Open 'Data Analysis' function, then choose 'ANOVA: Single Factor'.



- 5.3. Select a range of values covering all male and female height and put in the input range. Grouped by columns. Check on 'Labels in first row'. Set alpha as 0.05. Put the output range in the same sheet.

The screenshot shows the 'Anova: Single Factor' dialog box in Microsoft Excel. The dialog box is open over a data range in column D. The 'Input Range' is set to '\$4596:\$B\$103'. 'Grouped By' is set to 'Columns'. 'Labels in first row' is checked. 'Alpha' is set to 0.05. 'Output Range' is set to '\$D\$86'.

The screenshot shows the results of an ANOVA test in Microsoft Excel. The results are displayed in column C, starting from row 86. The results include a SUMMARY table and an ANOVA table.

Groups	Count	Sum	Average	Variance
Male height	17	2901	170.6471	24.86765
Female height	15	2381	158.7333	45.78095

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	1131.059	1	1131.059	32.66391	3.1E-06	4.170876786
Within Groups	1038.816	30	34.62719			
Total	2169.875	31				

- 5.4. You can see that the F value is greater than F Critical value. Therefore, you can reject the null hypothesis (average male height is equal to average female height).
- 5.5. The P-value shows similar thing as the t-Test.
6. Perform the similar test for BMI in those who do not have regular exercise and those who has.

Question 20 Is BMI of participants who have more regularity (often and always) in physical exercise less than of those who have not?

Answer: Click or tap here to enter text.

Exercise 5.3 Comparison of an average BMI among different age groups and different birth season.

This is a self-performing exercise. No example and lab direction provided.

Hint: Use 'Pivot Table' and 'ANOVA: Two-Factor without replication' tool.

Question 21 Are there any differences of BMI among participants who has different age and birth season?

Answer: Click or tap here to enter text.

END

Paisin Lekcharoen

Instructor

DVM, MVSc, FETP

