# Learning Objective

The purpose of this exercise is to introduce you to the process of preparing raw quantitative data for analysis. Raw data is messy and irregular. The first step in any data analysis is to examine the nature of the data itself, label all the data elements correctly, and screen for any data entry errors, missing cases, or other irregularities. This investment of time up front will save you an inordinate amount of later on in your analysis, and also enhance the accuracy and quality of your subsequent findings.

# Directions

Download a copy of the dataset, the questionnaire/code book that identifies the range of possible responses to each question, and these exercise directions. Submit a completed Excel workbook, showing all of the processes and analyses described below, with your answers to questions 1 to 6 on a separate worksheet in the workbook. Read the through the directions carefully so that you understand the goal of each step.

## Sample Data

You will be working with a subset of dataset from the **Pew Research Centre Core Trends**research project, Jan. 3-10, 2018**.** The research focused on attitudes towards the internet and social media use and whether Americans think it’s good or bad for society.

# Using Excel

I expect everyone to have basic skills in Microsoft Excel. If you are not familiar with Excel, work through the [MS tutorials](https://support.office.com/en-us/article/excel-for-windows-video-training-9bc05390-e94c-46af-a5b3-d7c22f6990bb?ui=en-US&rs=en-US&ad=US). Having said that, please do not spend hours if you are stuck – take a break, ask a friend, or email me or your TA for assistance.

**Please note that full step-by-step instructions are NOT given**, students might be using different versions and different platforms for Excel.

**Please do not use the online version of excel as it does not support the Analysis ToolPak and in.**

**BEFORE YOU BEGIN**, you will need the [**Analysis ToolPak**](https://support.office.com/en-us/article/use-the-analysis-toolpak-to-perform-complex-data-analysis-6c67ccf0-f4a9-487c-8dec-bdb5a2cefab6)**,** which is a Microsoft Office Excel add-in. First, check the Data tab. If you don’t see ‘Data Analysis …’ at the right-hand end of the ribbon (or at the bottom of the Tools menu) you will need to install it.

1. Click the **File tab**, and then click **Options**.
2. Click **Add-Ins**, and then in the **Manage** box, select **Excel Add-ins**.
3. Click **Go**.
4. In the **Add-Ins available** box, select the **Analysis ToolPak** check box, and then click **OK** (you don’t need the VBA version).
5. After you load the Analysis ToolPak, the **Data Analysis** command is available on the **Data** tab in the **Analysis** group at the right-hand end of the ribbon.

# Spreadsheet Set-up and Data Protection

Our first task is to ensure that the data is protected so that we don’t change any of the values by mistake. In order to do this:

* First rename the existing data sheet – call it **Raw Data**.
* Create a copy of the **Raw Data** sheet. Rename the copy **Working Sheet**.
* Return to the Raw Data sheet and protect it from any changes (click on the Format button). Double check that you are unable to make any changes on this sheet!
* Save your workbook.
* **Remember, keep saving your work regularly**!

# Creating Informative Formatting & Coding

Change to the Working Sheet. Set up this worksheet so that it is easy to understand and so that other people using it can easily identify what each column of data. In order to do this:

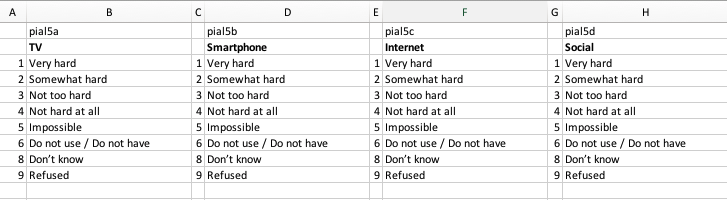
* Review the questionnaire summary document to understand which question relates to which column of values.
* Insert a row at the top and type in a meaningful column name for each column. Read the questionnaire to help you think of a good label for each question. For example, pial5a could be called “TV” and pial5b could be called “Smartphone.”
* Click in cell A3 and use the Freeze Panes function to ensure that your column headings remain visible as you scroll down the page.

## Creating a coding Sheet

While headings on the worksheet help, we still don’t know what the numbers in each column mean. It will be helpful when we are reporting our analyses to be able to label our findings correctly.

Add a new worksheet, call it **Coding Sheet**, and enter descriptions for each set of nominal and ordinal indicators for each variable. Make sure that the codes are in the first column and the descriptions in the second column for each set of codes.

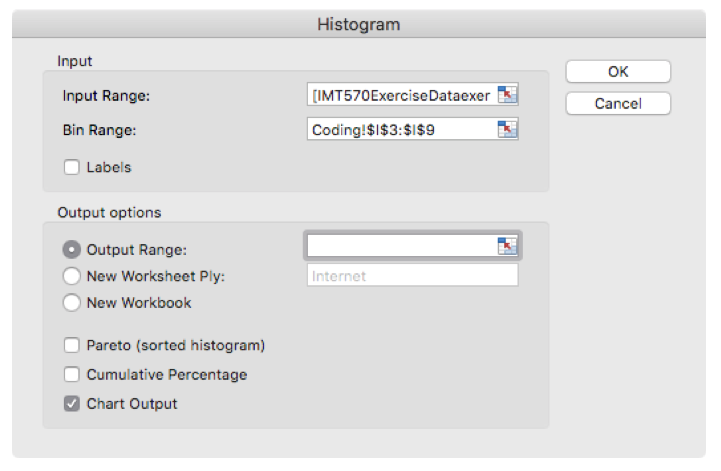
Do not create a set of codes for the variable AGE. This is a ratio variable (sometimes called a continuous variable). Here’s an example.



# Creating Histograms

Create a frequency table summarizing the frequencies of each value for the first variable (TV). Eventually, you will do this for all the nominal or ordinal variables in the dataset. This table will provide an opportunity to inspect the data.

* Create a new sheet called **Histograms** and enter a heading in cell A1 called **TV**.
* On the Data tab, click on ‘Data Analysis’ and select Histogram.
* For Input Range, click on the sheet icon to the right of the box. The box will minimize and you can move to the **Working Sheet** and select all the numeric values in the TV column. Click on the sheet icon to return to the histogram window.
* For Bin Range, click on the sheet icon to the right of the box. Move to the Coding Sheet and select the cells holding the values for TV (not the labels, just the numbers). Click on the sheet icon to return to the histogram window.
* Now you need to specify where to put the output. For Output Range, select a cell below your TV label on the Histogram sheet.
* Finally, select Chart output at the bottom of the Histogram Window to also display a chart of your results, and click OK.



You should see a table with headings Bin and Frequency, and a histogram chart.

Repeat these steps with every nominal or ordinal variable in the data set. This should include all of them except for respid and AGE (which is a ratio variable).

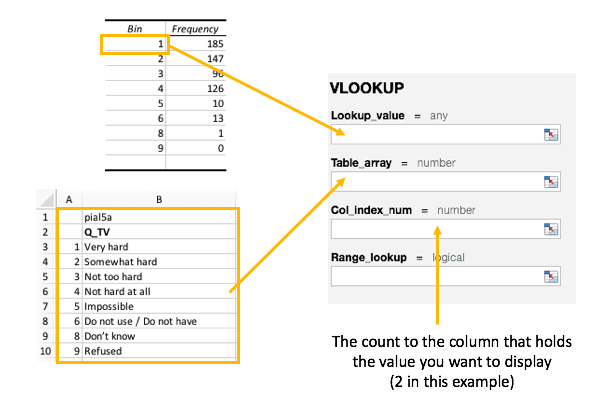
# Formatting the Histogram Table

Unfortunately, the table is not very informative because it has the extra category ‘More’ which is a default category for values that fall outside your Bin range (for some variables with missing cases, some cases will fall into the More category). Also, the Bin labels are numeric.

You can fix the ‘More’ problem easily, by deleting the row showing ‘More’ from the histogram table (don’t just delete the values in the cells – you need to actually delete the cells). Notice that the histogram changes a little to reflect this deletion. Now select the chart and modify the data range so that the chart only displays data for each of the bin ranges.

In order to add meaningful names next to the Bin labels, we will use a Vertical Lookup function. What we’re doing here is telling Excel to take the number in the Bin cells and look it up in your coding table. Find the description corresponding to that number and bring it back and report it next to the Bin value

* Insert a new column next to the bin column in you Histogram sheet and label it Response Label
* Click on the ***fx*** icon on the toolbar and find the VLOOKUP function and use it to obtain the correct description from the **Coding Sheet**.



If you got your lookup right, you should see the word ‘Very Hard’ appear in the cell next to the ‘1’ in the Bin column.

**Make sure** that you have used **absolute cell addresses** for the lookup table in your VLOOKUP function, so that the cell addresses of the lookup table don’t change if you copy the formula. To create an absolute address, insert $ before every cell letter and cell number. Example:

* Relative address: A2:B5
* Absolute address: $A$2:$B$5

# Formatting the Histogram Chart

The chart displayed by the histogram tool is hard to read, so once you have displayed the frequencies and chart, and added descriptive row labels, experiment with the chart to improve its appearance.

1. Click on the outside frame of the histogram chart, and you will see that a set of Chart Tools tabs appear on the ribbon.
2. On the Chart Design tab, click on ‘Select Data’ and Edit the horizontal axis labels to choose the cell range that will report descriptive labels.
3. Double click on the title in the chart and rename it appropriately. Do the same for the X axis.
4. Remove the series label from the chart
5. Now create a copy of the chart below the original and try to convert it to a pie chart, with descriptive labels instead of the bin numbers and percentages instead of the frequency counts. [HINT: Look on the Chart Design tab in the Type group for different chart types, and in the Chart Layout tab for ways to label the segments of the pie.]

Finally, repeat steps 1-3 to report the frequency table for every nominal and ordinal variable in the Pew data set (you may experiment with other chart forms as you wish, but are not required to do this for any variable other than the first one).

# Questions

Create a new sheet called **Answers** to record your answers to the following questions. Please note that some of the questions have multiple, inter-related parts. Be sure to answer the questions completely as you can.

For any given variable, a missing case is a subject where an observation has not been recorded. The data may be blank, or it may be that the subject refused to answer or indicated they did not know.

1. According to the frequency tables, which variable has the lowest number of missing cases? Which variable has the highest? What is the range of missing cases in the data set?
2. For some of the variables, indicators make a distinction between “Don’t Know” and “Refused to Answer. Under what conditions would it be important to make a distinction between these two responses? When would it not matter? Do you agree or disagree with the choices made by these researchers?

Identify the three variables with the largest number of missing cases. Sort the dataset by these three variables (largest, second largest, third largest – this will be a “custom sort”). When you are done, at the bottom of the sheet you will see cases with values of 8 or 9 for each of these variables, and possibly for others as well. This is a rough “eye-balling” of the data to identify subjects that might have a large number of missing cases.

1. How many missing cases would a subject need to have before you feel it would be necessary to remove the subject from further analysis? Do you see any cases you think should be removed? Why do you want to remove them?

Examine the frequencies of actual observations when answering questions 4-6 to get a sense of how responses to each question is distributed.

1. For the first 9 variables, are the responses roughly even across the indicators, or does one indicator have a substantially larger number of cases? How many different patterns do you see across these variables? Describe these distribution patterns in words.
2. For any variable distribution, the most common observation is called the Mode. Would the Mode be a good indicator to describe the center of the distribution for any of these 9 (nominal) variables? Why?
3. Look at the frequency distribution of the variables pertaining to level of education and income. In words, how would you characterize the group according this data? Are they highly educated or lowly educated? Are they low income, middle class, or wealthy? In other words, how would you interpret these numbers?