Assignment #1: Excel Data Prep Exercise

Learning goals of this week:

Pivot table / Unpivot (reshape) / Linking tables / Excel data prep functions

By going through this exercise, you will be essentially taking the baby steps of storytelling with data (without visualization yet).

You should use your virtual machine on **mycloud.gatech.edu (COB-LAB)** to complete this exercise, or at least the -reshape- and WWW import part, because those two features are not available or hard to use in the Mac version of Excel.

I recommend that you log in to Canvas and download a copy of this file on the same virtual desktop where you will be working on this exercise.

Once you log in to the virtual machine, use this link to download the data
http://bit.ly/bdpvcardata

(BDPV car data)

> *Note: This car dataset is publicly available and the source will be provided later. You can search for it yourself if you are curious but finding it online won't help you finish this assignment.*

Use the "car data.txt" file. This dataset shows the average gas efficiency (MPG, miles per gallon) of cars produced by each manufacturer in each year (across different models they produce). Note that there might be missing values.

For this exercise, you need to submit an Excel file titled
"**ExcelPrep_1_YourFirstNameInitial_YourLastName.xlsx**"

Please note: If you need help with this question, please be sure to watch the data prep videos. If you understand the demos in the videos, you should have no problem with this exercise. Ask the instructor if you need any further clarifications about the video demos.

1. (0%) (**not graded but make sure you try this**) Decide on a text editor (not rich text editors such as Word or Pages that allow you to format words) that you want to use for the semester. It doesn't matter which tool you use (here is a good complete list https://en.wikipedia.org/wiki/List_of_text_editors) but when you work with data, you cannot open them in Word or Pages. I recommend Sublime Text, https://www.sublimetext.com/ , which is shareware for all major OS platforms, but no limit on time so you can at least use it for the semester as an evaluation. Use the text editor of your choice to open the dataset for this exercise, "car data.txt" in the text editor, and take a look.

2. (10%) **Import** this file into Excel.

Tip: what's the delimiter for this file? Is it comma? Or tabs? You should be able to see this in the pure text editors. Try open a new window in the pure text editors, and type in a "space", vs. a "tab" (by pressing the "tab" key on your keyboard. You will be able to see the difference).

The import creates a worksheet in your Excel. **Save this file** as ExcelPrep_1_YourFirstNameInitial_YourLastName.xlsx.

3. (30%) (**unpivot/reshape – you should use mycloud.gatech.edu – COB-LAB**) Use a new worksheet to reshape this file into the "long" format. Each row should represent a Make-Year combination, e.g., one row could have BMW in the first column, 1992 in the second column, and 19.67 in the third column.

   Rename the columns to "make" "year" "avgMPG", respectively.
   Rename this new worksheet "Long0"

   Note: if you know other ways to do unpivot, you don't necessarily need to use the method described in the lecture video.

4. (10%) Copy and paste **only rows without missing values for avgMPG** into a new worksheet. Name this new worksheet "noBlanks"

   Tip: use the filter – the downward arrow in the avgMPG cell. Filter out the blanks, then copy/paste. However make sure you are only copying/pasting the visible rows. Here are the instructions: https://support.microsoft.com/en-us/office/copy-visible-cells-only-6e3a1f01-2884-4332-b262-8b814412847e

5. (20%) (**Pivot Table**) Create a table with two columns based on the "noBlanks" worksheet. The first column is year, and the second column is the average MPG of cars (only those that we have the data for). Each row should represent a year (i.e., the unique identifier of each row, or the "primary key" of this table, should be the year). Rename the column headers to "Year" and "avgCarMPG", respectively. Format second column to make sure it only shows two decimal digits (e.g. 20.22). Rename this worksheet as "Year-MPG".

   What do the numbers in the second column represent? Pick any number in that column (e.g., B10, but can be others) and add a comment to explain that number (select that cell, go to INSERT—COMMENT.

   Without creating visualizations, do you see any trend? Add your thoughts as a "comment" for the last row of this table.

6. (10%) (**Import data from WWW**) In the virtual machine, fortunately, they have a more recent version of Excel than the one I used for the demo. Use the *Data* → *From Web* button to import data from this URL: https://www.usinflationcalculator.com/gasoline-

prices-adjusted-for-inflation/. On the next screen, click on "Connect". Then, select "Table 0" on the left panel. Then, click "**Load**". Rename this worksheet "GasPrice".

7. (20%) (**Linking**) Return to the "Year-MPG" worksheet. For each year, create a new column, and use the -lookup- function (do not just copy/paste, you must use the lookup function) to extract the "Gas Prices Adjusted for Inflation (In 2020 Dollars)" value from the "GasPrice" worksheet. See demo if you have any questions.

   Tip #1: One pleasantly surprising difference you will see is that, in this new version of Excel, you don't even need to worry about absolute references (A3 rather than A3, as we showed in the video)!

   Tip #2: Some of you (depending on your computer and version of Excel) may see an error in the Lookup function. Here is the reason. It is possible that, since you imported the data from two different places (one from the CSV, the other from the web), even though both have a "Year" column, they might contain different formats. For example,

   "1990"

   (with the double quotes, as text qualifiers") will show up as 1990 on your screen, but internally Excel recognizes it as having the double quotes surrounding the number. In that case, if you are looking for

   1990

   In a column that stores years as

   "1990"

   The lookup function will generate an error message because it can't find 1990 (no double quotes). The problem arose because of a difference between the CSV import and the WWW import. Depending on the version of Excel, there are slight differences. In particular, the "Year" was stored as a text in the CSV import, and it was stored as "Numbers" in the WWW import. In this case, when you do LOOKUP() it is looking for 1990 (no double quotes) in a column that has "1990" (with double quotes, but not visible in excel).

   To address this, you can use Excel's "text to column" function, under Data // Data Tools. No need to change the default options. It'll be able to take those things out. Your lookup function should work after that. In the most recent version of Excel, if your column has something like "1990" (with double quotes), Excel will give you a little warning message in the form of a green triangle in the upper left corner of the cells, saying "numbers stored as texts", and you can "convert to numbers" that way --- which would give you the same result. Then, refresh the Pivot table. It should work.

More generally, if you are not sure whether a column is stored as text or number, you can use ISTEXT() or ISNUMBER() functions, even if they appear like numbers.

Save what you did, and submit the **ExcelPrep_1_YourFirstNameInitial_YourLastName.xlsx** file.

That wraps up the exercise. Keep a copy of this file. A few things for you to think about (your answers to the following questions are optional and not graded, but important):

This whole data prep process is what you would have needed to do to test one idea: does gas prices relate to car MPG in some way? When we talk about visualizations you can refer back to this to see if you see any remarkable pattern between these two variables.

- Are these data sufficient?
- What else would you add to this "picture" when thinking about their relationships?
- If you don't like what you found, can you think of OTHER variables that might relate to the average MPG in some way.
- Or, perhaps MPG did not turn out to be that interesting for you – are there other variables in the data that you'd look into?
- Or, are there other car-related datasets out there for you to look at?