Module 4 Lab

Part 1

For this part of the lab we'll work through the mbta.xlsx data. The Massachusetts Bay Transportation Authority ("MBTA") manages America's oldest subway, as well as Greater Boston's commuter rail, ferry, and bus systems. It's your first day on the job as the T's data analyst and you've been tasked with analyzing average ridership through time. Complete the following data cleaning tasks.

1. Import the data. Note, you'll need to skip the first row and have the first column be used as the index. Check out the `read_excel()` docs to figure out how to do this. The resulting DataFrame should contain 11 rows and 59 columns and look similar to the below DataFrame.

	mode	2007-01	2007-02	2007-03	2007-04	2007-05	2007-06	2007-07	2007-08
1	All Modes by Qtr	NaN	NaN	1187.653	NaN	NaN	1245.959	NaN	NaN
2	Boat	4.000	3.600	40.000	4.300	4.900	5.800	6.521	6.572
3	Bus	335.819	338.675	339.867	352.162	354.367	350.543	357.519	355.479
4	Commuter Rail	142.200	138.500	137.700	139.500	139.000	143.000	142.391	142.364
5	Heavy Rail	435.294	448.271	458.583	472.201	474.579	477.032	471.735	461.605
6	Light Rail	227.231	240.262	241.444	255.557	248.262	246.108	243.286	234.907
7	Pct Chg / Yr	0.020	-0.040	0.114	-0.002	0.049	0.096	-0.037	0.004
8	Private Bus	4.772	4.417	4.574	4.542	4.768	4.722	3.936	3.946
9	RIDE	4.900	5.000	5.500	5.400	5.400	5.600	5.253	5.308
10	Trackless Trolley	12.757	12.913	13.057	13.444	13.479	13.323	13.311	13.142
11	TOTAL	1166.974	1191.639	1204.725	1247.105	1244.755	1246.129	1243.952	1223.323

11 rows × 59 columns

2. How many missing values are in each column? How many missing values are there in total.

3. It appears that the data are organized with observations stored as columns rather than as rows. You can fix that. First, though, you can address the missing data. All of the NA values are stored in the first row. This row really belongs in a different data frame; it is a quarterly average of weekday MBTA ridership. Since this data set tracks monthly average ridership, you'll remove that row. Similarly, the 7th row (Pct Chg / Yr) and the 11th row (TOTAL) are not really observations as much as they are analysis. Go ahead and remove the 7th and 11th rows as well. After removing the first, seventh, and eleventh rows, what are the dimensions of the new DataFrame?

4. In this data, variables are stored in rows instead of columns. The different modes of transportation (commuter rail, bus, subway, boat, ...) are variables, providing information about each month's average ridership. The months themselves are observations (Currently, months are listed as variable names; rather, they should be in their own column). You can tell which is which because as you go through time, the month changes, but the modes of transport offered by the T do not. As is customary, you want to represent variables in columns rather than rows.

1. Pivot the rows and columns of the mbta data so that all columns are variables of the data. This should result in 3 columns - mode, date, and number of riders in thousands (`thou_riders`).
2. What are the new dimensions of this data?

5. Your data set is already looking much better! Your boss saw what a great job you're doing and now wants you to do an analysis of the T's ridership during certain months across all years. Your data set has months in it, so that analysis will be a piece of cake. There's only one small problem: if you want to look at ridership on the T during every January (for example), the month and year are together in the same column, which makes it a little tricky. You'll need to separate the month column into distinct month and year columns to make life easier.

1. Split the month column of mbta at the dash and create a new month column with only the month and a year column with only the year.
2. View the head of this new mbta data set.

6. Every month, average weekday commuter boat ridership was around 4,000. Then, one month it jumped to 40,000 without warning? Unless the Olympics were happening in Boston that month (they weren't), this value is certainly an error. You can assume that whoever was entering the data that month accidentally typed 40 instead of 4.

1. Locate the row and column of the incorrect value.
2. Replace the incorrect value with 4.

7. Congrats, your data is now clean and ready for analysis.

1. Compute the mean ridership per mode.
2. Compute the mean ridership per mode for the month of January.
3. Which year had the greatest ridership for the boat mode?
4. On average, which month experiences the greatest number of passengers on the Heavy Rail mode?

Part 2

For this module we will be using the **completejourney_py** data sets that you saw in the lesson 4b reading:

1. Using the transactions and demographics data, how many of the 1,469,307 transactions do we have demographic information for?

2. Using the transactions and demographics data, compute the total `sales_value` by age category to identify which age group generates the most sales.

3. Identify all different products that contain “pizza” in their `product_type` description. Which of these products produces the greatest amount of total sales (compute total sales by product ID and product type)?

4. Identify all products that are categorized (`product_category`) as “pizza” but are considered a “snack” or “appetizer” (via `product_type`). Which of these products (`product_id`) have the most number of sales (measured by `quantity`)?

5. Identify all products that contain “peanut butter” in their product_type. How many unique products does this result in? For these products, compute the total `sales_value` by month based on the `transaction_timestamp` . Which month produces the most sales value for these products?