**Homework 2**

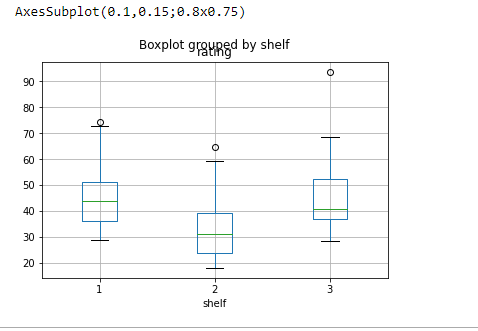
Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Tyler DiNapoli-Chiappelli\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

There are five questions (15 points) in this assignment. The minimum increment is 0.5 point. Solve them and fill the answers in the blank space.

1. **Breakfast Cereals.** Find the dataset HW2\_Cereals.csv on Blackboard. The table below describes the variables in the dataset. Write a Python code to explore and summarize the data as follows.

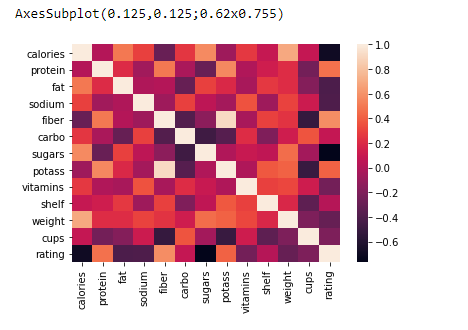


a. Use appropriate graphs to detect if *shelf* is relevant to *rating*. Attach that graph in the space below. Based on your judgment, are they relevant? Explain your answer. (2 points)

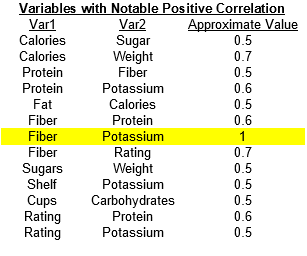


Yes, based upon this graph, it does seem as if the shelf a particular box of cereal is placed upon does in fact have some relationship with its consumer rating. As we look across these data groups, we can see that the mean, median, and IQR of the “rating” variable changes rather substantially as the shelf changes. Granted, it is worth noting that there may be some self-selection bias in this relationship, as grocery stores may purposefully place certain products on certain shelves based off of consumer reports to begin with. So, it not clear in which “direction” this observed relationship goes.

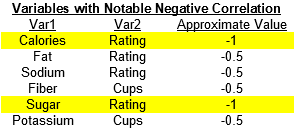
b. Attach the heat map of the correlation matrix in the space below. Which pair of variables is most strongly positively correlated? Which is most strongly negatively correlated? (2 points)



The variables that are most strongly positively correlated, other than a variable correlated with itself, are as follows in this table:



The variables that are most strongly negatively correlated, other than a variable correlated with itself, are as follows in this table:



For both tables, please note that the variables highlighted in yellow are too perfect to be true. Rather than a strong correlation, I suspect there is something amiss with how the data was collected.

Submit your Python notebook file with the filename [DM2020] HW2\_Q1\_YOURFULLNAME. ipynb (Hereafter YOURFULLNAME refers to your first and last name).

2. Several new airports have just opened in major cities, opening the market for potential new routes (a route refers to a pair of airports). Hereafter we refer to those major cities as M cities. A major airline (not Southwest) has a goal to predict average ticket fare on these potential new routes. The analytics team has found the following dataset in the company’s data warehouse (available on Blackboard with the filename “HW2\_Airfares.csv”). They consist of the variables listed in the following table. Note that the set of M cities does not contain any city included in the CSV file.

|  |  |
| --- | --- |
| **Table DESCRIPTION OF VARIABLES FOR AIRFARE EXAMPLE** | |
| S\_CODE | Starting airport's code |
| S\_CITY | Starting city |
| E\_CODE | Ending airport's code |
| E\_CITY | Ending city |
| COUPON | Average number of coupons (a one-coupon flight is a nonstop flight, a two-coupon flight is a one-stop flight, etc.) for that route |
| VACATION | Whether (Yes) or not (No) a vacation route |
| SW | Whether (Yes) or not (No) Southwest Airlines serves that route |
| S\_INCOME | Starting city's average personal income |
| E\_INCOME | Ending city's average personal income |
| S\_POP | Starting city's population |
| E\_POP | Ending city's population |
| SLOT | Whether or not either endpoint airport is slot controlled (this is a measure of airport congestion) |
| GATE | Whether or not either endpoint airport has gate constraints (this is another measure of airport congestion) |
| DISTANCE | Distance between two endpoint airports in miles |
| PAX | Number of passengers on that route during period of data collection |
| FARE | Average fare on that route |

Do you think the analytics team can utilize this data set to help the airline company to achieve its goal? If yes, explain why you think it is helpful and then further discuss which variables and how many observations the analytics team should include in the analysis. Justify your choices. If no, explain why you do not think it is helpful in detail. (2 points).

This data set could be helpful for the company to achieve its forecasting goal, as long as the cities in “M” are in the United States. This dataset has several useful attributes. First, these data are already for major US cities, such as Boston and New York City, so any made comparisons are with similar categories of markets. Second, this data set contains data such as city population, income, “vacation destination”, and distances. These variables could be used to estimate a demand function for the locations in this data, which could potentially be used as a proxy for demand functions in the M cities. Finally, and critically, these data show the average fare on a particular route. This is of notable importance because these data reflect the net effect of competition from other carriers on the final ticket prices for these routes. Assuming the company can figure out which other carriers they will be competing with on the routes in M, they can determine whether to what degree the demand function they estimate using this data will apply to the business environment for M.

When estimating their model, the analytics team should include the variable FARE as the dependent variable and the variables COUPON, VACATION, S\_INCOME, E\_INCOME, S\_POP, E\_POP, DISTANCE, and PAX as the independent variables. Here is a breakdown of the rationale:

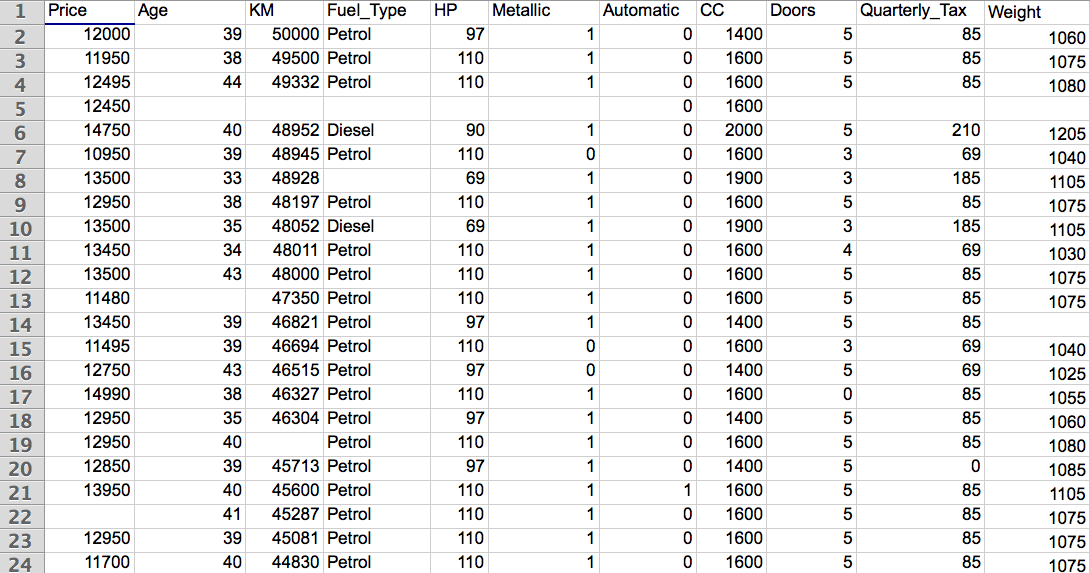
**FARE:** the company’s explicitly stated goal is to estimate average ticket prices for M. FARE is quite literally the average ticket price for this data set.

**S\_INCOME, E\_INCOME, S\_POP, and E\_POP:** These are all economic variables which are likely to impact a demand function. It is worth noting here, that it is most prudent to run three different models. One which drops S\_INCOME and S\_POP, one which drops E\_INCOME and E\_POP, and one which includes all four. According to economic theory alone, it is not clear which model is “correct”, so all three should be created.

**COUPON, VACATION, DISTANCE, PAX:** While these are not explicitly economic variables, they likely will impact ticket prices. A dummy for VACATION will likely result in a different outcome when the dummy is on versus when it is off. And, COUPON, DISTANCE, and PAX will all reflect costs the airline will have to cover as part of its ticket pricing.

Regarding the number of observations in the current data set: given my proposed usage of 8 predictors, the data set’s 638 observations is more than enough to run a robust model, following the guidance of “observations = 10 x predictors”.

3. Suppose the sample dataset you retrieved from the IT department is the following. Identify the missing values in the table by labeling them. Write down your answers. For each missing value, write down how you would handle it. Note that you only need to write down the treatment you want to give and do not have to compute the specific value for imputation. You can put your answer in a batch mode if you believe a group of labeled missing values shall receive the same kind of treatment. (4 points) **(Need the class content on Sep. 28th)**



1: Price has one missing value in row 22. As this is the dependent variable, we must strike the entirety of row 22 from the data set.

2: Age has missing values in rows 5 and 13, KM has missing values in rows 5 and 19, HP has a missing value at row 5, Quarterly\_Tax has a missing value at row 5, and Weight has missing values at rows 5 and 14. These are all numerical independent variables whose missing values should be replaced with the median value of each variable.

3: Fuel\_Type has missing values in rows 5 and 8, Metallic has a missing value in row 5, and Doors has a missing value in row 5. As these are all independent categorical variables, each missing value should be replaced with the mode value in each respective column.

Here, it is worth mentioning that a human expert could assist in two ways. First, there are a few values which are impossible or highly unlikely; for example, a car having 0 doors in row 17 or a car having no quarterly tax in row 20. An individual with domain knowledge might be able to help fill these values. Similarly, rather than implementing the brute-force methods described above, a subject-matter expert might be able to advise on where naively imputing either the median or the mode is inappropriate.

4. Briefly discuss why we need to standardize numerical variables and code categorical variables in general. (2 points) **(Need the class content on Sep. 28th)**

Numerical values need to be standardized when the values occupy different ranges or scales. For example, if in the Used Car data, some subgroup of cars weighs between between 1000 kg and 1200 kg, and a different subgroup of cars weighs between 1500 kg and 2000kg, the large difference in the values of these subgroups will tend to distort the measurements our model will come up with. Therefore, in order to improve the fidelity of the model’s output, bringing all values in line with a standard scale is necessary.

Categorical value need to be coded as dummies, as opposed to natural numbers, for a similar reason. Coding as natural numbers such as 2, 3, or 4, will artificially alter a model’s numerical output because we as users would have already pre-assigned arbitrarily higher or lower values to certain observations. Thus, coding as 0 and 1 only is done to minimize model distortion.

5. Use Microsoft Excel to standardize the following two variables in the table. Next, find out the exchange rate between USD and Euro on the day you do this homework question. Use the rate to convert Income from USD to Euro. And then standardize the Income in Euro. Are the standardized values different between using Income in Euro and using Income in USD? Show all calculations. **(Need the class content on Sep. 28th)**

Submit your Excel spreadsheet with the filename [DM2020] HW2\_Q5\_YOURFULLNAME.xlsx (3 points)

|  |  |
| --- | --- |
| **Age** | **Income (USD $)** |
| 26 | 50,000 |
| 55 | 155,000 |
| 64 | 98,000 |
| 31 | 191,000 |
| 40 | 38,000 |
| 48 | 56,000 |

Please see submitted Excel sheet