**Homework 2**

Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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)

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)

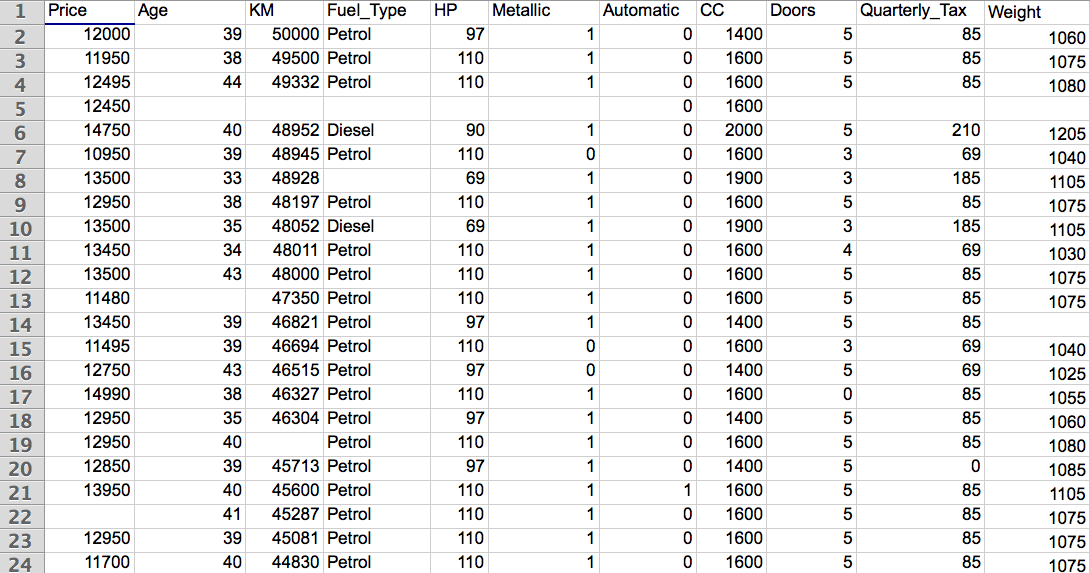
Submit your Python notebook file with the filename [DM2020] HW2\_Q1\_YOURFULLNAME. ipynb (Hereafter YOURFULLNAME refers to your first and last name).

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).

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)**



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)**

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 |