**STAT 495 – Project 1: Cleaning and Inspecting Data**

***(due: 10/05/2023 11:59 pm)***

Answer all questions fully and include all requested answers and output onto a single .pdf, .doc, or .docx file. All written work for the report will need to be **typed**, and all requested plots need to be generated through Python and included in your report. All plots need to be correctly labeled and easy to read/understand. In addition to the typed report, upload the .ipynb file along with comments indicating which code corresponds to which question and final cleaned dataset as a .csv file. **In summary, upload three files on Canvas: (1) a report with brief answers (.pdf, .doc, or .docx) (2) code (.ipynb) (3) cleaned dataset (.csv)**

Data for completing this project can be found at Assignments Project 1 tab on Canvas. Note, the dataset may be a bit inaccurate: most data were collected several years ago, though slight updates have been made to some observations in the dataset.

**Cleaning and Inspecting data:**

The file Stocks.csv contains data on the closing stock price on a single day for a large number of companies. For each company, the dataset provides the name (Name), stock market ticker symbol (Symbol), sector (Sector), stock price (Price), dividend as a percentage of stock price (Dividend), price to-earnings ratio (PE), earnings per share (EPS), lowest and highest share price over the last 52 weeks (52 week low and 52 week high), market capitalization of the company’s shares in billions of dollars (Market Cap), and earnings before interest, taxes, depreciation, and amortization in billions of dollars (EBITDA).

**1. Data Cleaning:**

1. (5pts) How many companies are included in the Stocks.csv dataset?
2. (5pts) How many missing values are there for each variable?
3. (5pts) Clean up the missing value(s) in each variable by replacing with its average.
4. (5pts) Remove all companies with an earnings-per-share greater than 20 from this dataset.
5. (5pts) Add a column to this dataset obtained by taking the log base 10 of the price.

(g) (5pts) Add a column to this dataset obtained by taking the difference between the 52-week-high and the 52-week-low of the stock price.

(h) (10pts) Create a variable that is 1 if the company has an EBITDA greater than1 billion and is 0 otherwise. Add this variable as a column in the dataset.

(i) (5pts) Replace the 2nd entry of the Name variable with your full name.

(j) (10pts) Save this cleaned dataset as a **Cleaned\_Stocks.csv** file. (Hint: use **pd.to\_csv()** function)

**2. Data Inspection: For this section, use the cleaned dataset from Part 1.**

1. (5pts) How many different sectors are included in the dataset and how many companies of each sector of the dataset are there?
2. (5pts) For each sector, give the average the difference between the 52-week-high stock price and the 52-week-low stock price for companies within that sector (the variable constructed in part 1g). Which sector has the highest average difference? Which sector has the lowest average difference? (Hint: You can reduce your workload by using the aggregate() function)
3. (5pts) For each sector, give the standard deviation of the earnings-per-share for companies within that sector. Which sector has the highest standard deviation? Which sector has the lowest standard deviation?
4. (5pts) Give the five companies with the largest dividends.
5. (5pts) Give the five companies with the smallest earnings per share.
6. (5pts) List all companies that do not pay a dividend (that is, Dividend = 0).
7. (5pts) Make two histograms: one of log base 10 of price and one of market cap. Describe the shape of both histograms.
8. (5pts) Make a scatterplot where dividend is on the x-axis and the earnings per share is on the y-axis. Describe the shape of this plot.
9. (5pts) Make side-by-side boxplots comparing the distribution of the dividends between companies with an EBITDA greater than 1 billion to those with an EBITDA less than or equal to 1 billion. What does this plot show? (Hint: It may be useful to use the variable created in part 1h) to make this graph)