Huiting Wu

Professor Name: Taejoon Kim

Stat 495: Data Analysis with SAS/Python

28 September 2023

STAT 495 – Project 1: Cleaning and Inspecting Data

1. Data Cleaning:

   (a) There are 342 companies are included in the Stocks.csv dataset.

   (b) Missing values for each variable:

   - Name: 0
   - stock market ticker symbol (Symbol): 0
   - Sector: 0
   - stock price (Price): 1
   - dividend as a percentage of stock price (Dividend): 0
   - Price to-earnings ratio (PE): 1
   - earnings per share (EPS): 1
   - The lowest price over the last 52 weeks(52 week low): 2
   - The highest price over the last 52 weeks(52 week high): 1
   - Market Cap: 0
   - EBITDA: 0

2. Data Inspection:

(a) There are 11 different sectors are in included in the dataset. The number of companies for each sector of the dataset:

- Consumer Discretionary: 55
- Consumer Staples: 25
- Energy: 6
- Financials: 59
- Health Care: 34
- Industrials: 51
- Information Technology: 38
- Materials: 20
- Real Estate: 27
- Telecommunication Services: 3
- Utilities: 23

(b) The average of the difference between the 52-week-high stock price and the 52-week-low stock price for each sector:

- Telecommunication Services: 9.936667
- Utilities: 11.761338
- Consumer Staples: 18.865200
- Energy: 19.945000
- Consumer Discretionary: 21.274196
- Information Technology: 24.308947
- Real Estate: 26.793333

- Financials: 27.044976
- Health Care: 29.963824
- Materials: 30.874000
- Industrials: 31.816667

The sector with highest average difference is Industrials. The sector with the lowest average difference is Telecommunication Services.

(c) The standard deviation of earnings-per-share for each sector:

- Telecommunication Services: 1.026174
- Utilities: 1.433574
- Energy: 1.593675
- Consumer Staples: 2.022614
- Consumer Discretionary: 2.129783
- Information Technology: 2.427637
- Real Estate: 2.585953
- Health Care: 2.800441
- Materials: 2.826266
- Industrials: 3.171595
- Financials: 3.553237

The sector with the highest standard deviation is Financials. The sector with the lowest standard deviation is Telecommunication Services.

(d) The five companies with the largest dividends:

- Iron Mountain Incorporated: 6.05

- Mattel Inc.: 5.97
- ONEOK: 5.44
- Seagate Technology: 5.15
- Weltower Inc.: 5.00

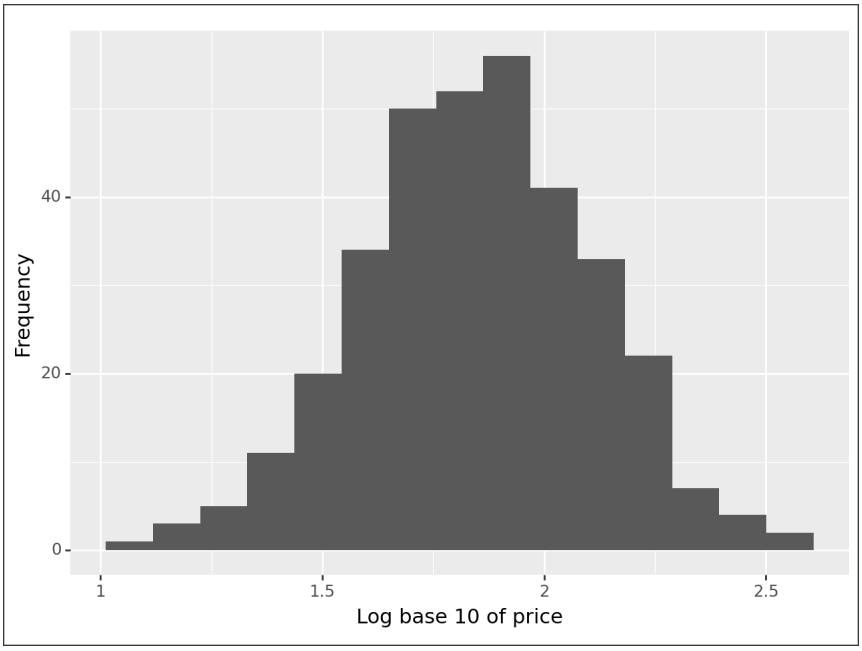
(e) The five companies with the smallest earnings per share:

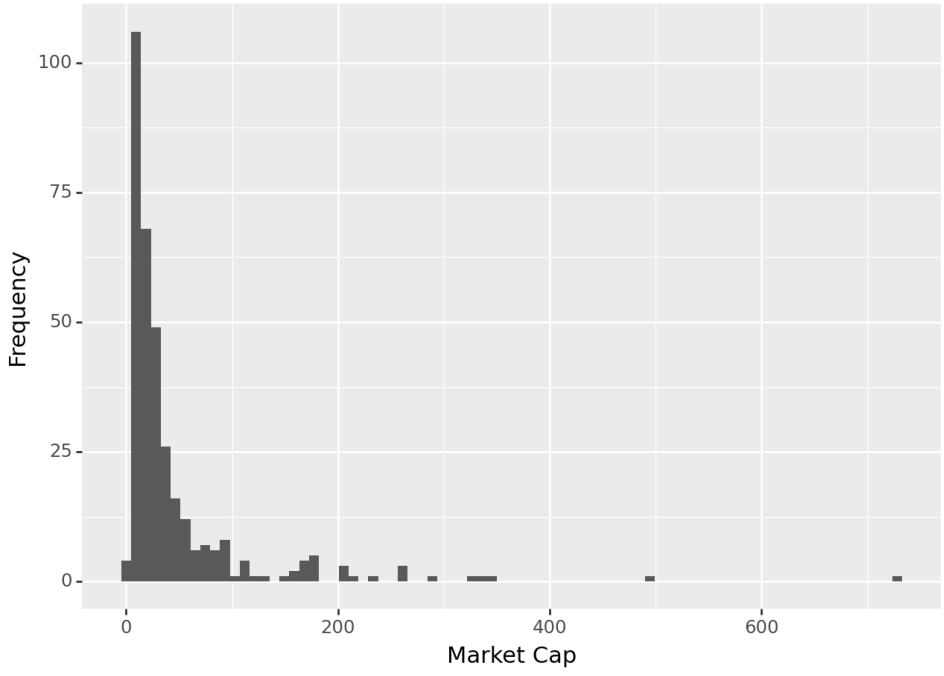
- Kinder Morgan: 0.25
- Leucadia National Crop.: 0.34
- Microchip Technology: 0.41
- Iron Mountain Incorporated: 0.42
- Western Union Co: 0.51

(f) The companies that do not pay a dividend are:

- E*Trade
- Express Scripts
- Lumen Technologies
- Mylan N.V.
- United Continental Holdings
- Varian Medical Systems

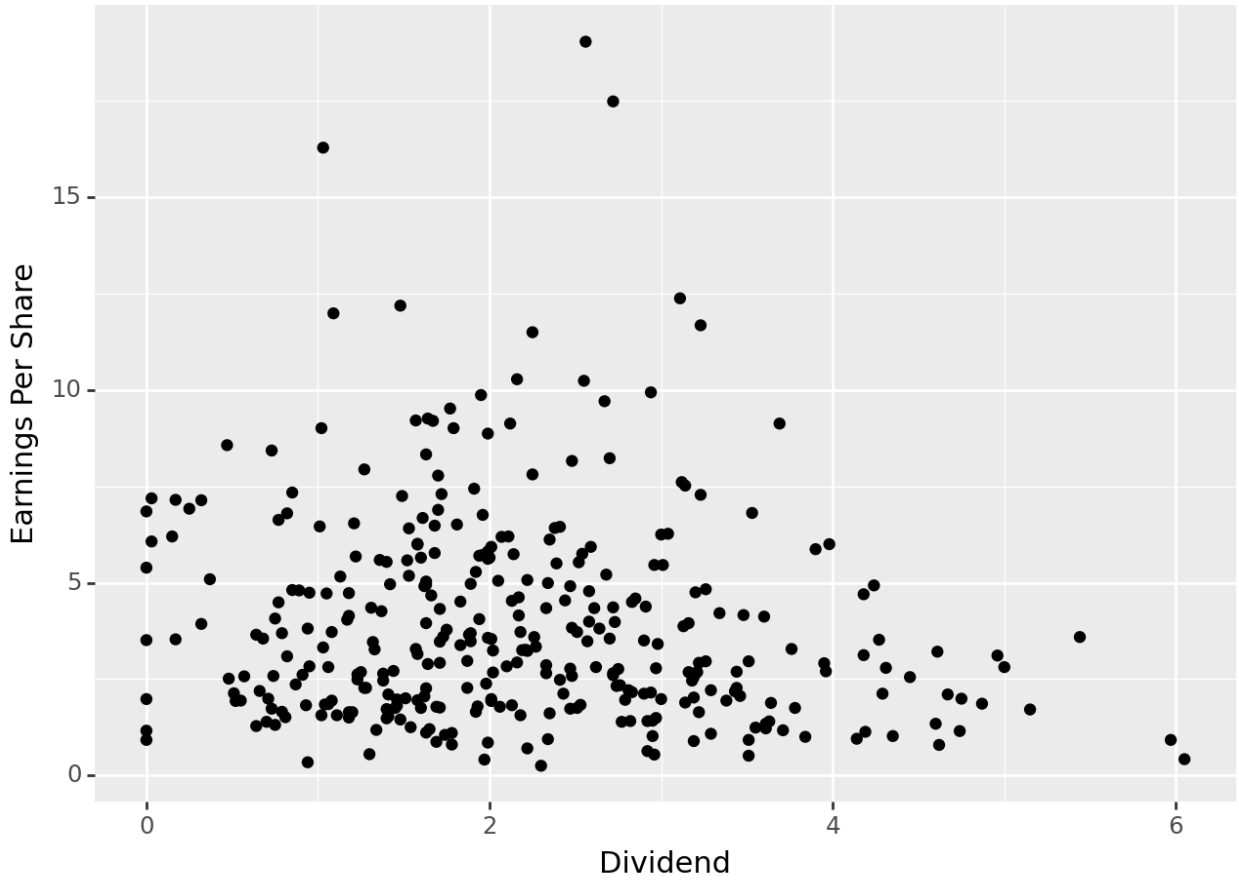
(g)





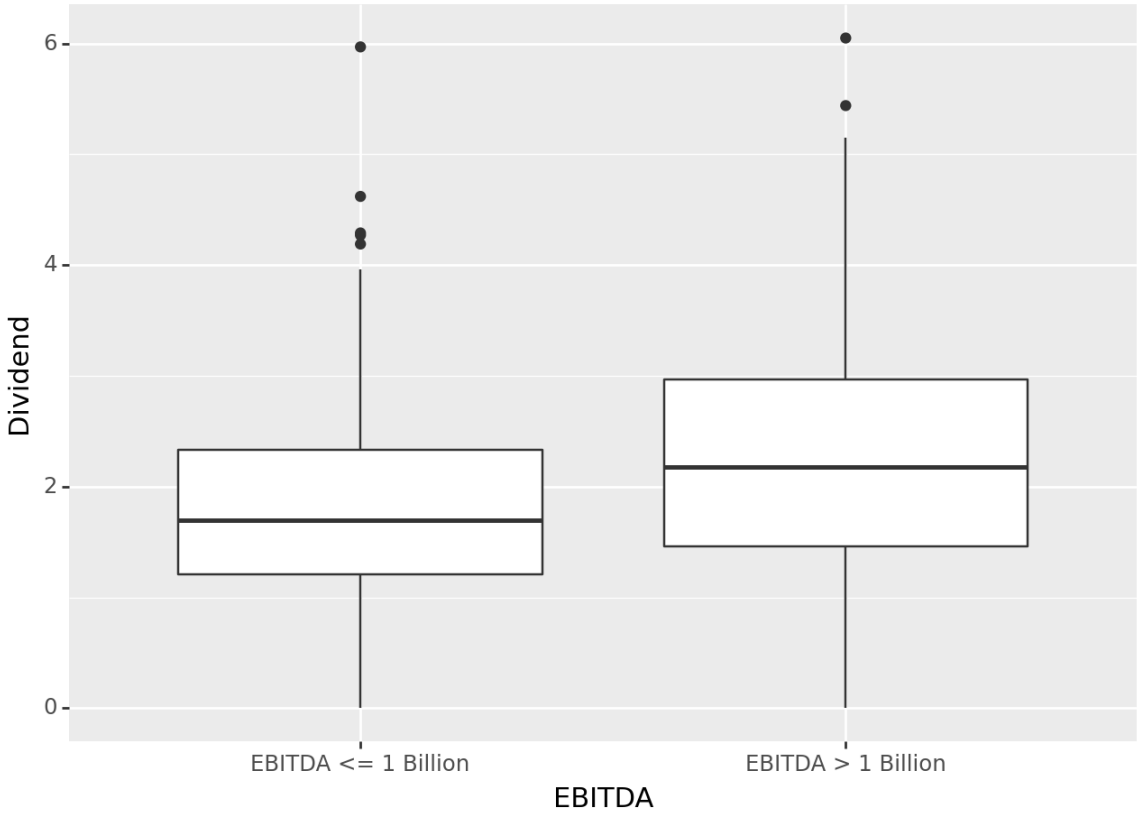
The shape of frist graph is bell shape. The shape of the second graph is right-skewed.

(h)



The scatterplot is with a weak negative correlation. It may need transformation for the variables to make a linear relationship.

(i)



- This graph shows the median of dividends of companies with an EBITDA greater than 1 billion is higher than the companies with EBITDA less than or equal to 1 billion.
- The companies with an EBITDA greater than 1 billion has larger maximum dividends than the companies with EBITDA less than or equal to 1 billion.
- The companies with an EBITDA greater than 1 billion has a wider IQR than the companies with EBITDA less than or equal to 1 billion.
- There are 4 outliers of the companies with EBITDA less than or equal to 1 billion.
- There are 2 outliers of the companies with an EBITDA greater than 1 billion.