**HW2**

**Name: Yiyoung Kim**

**Date: 2024.01.31**

The analysis presented herein pertains to data extracted from the Naver Finance Market Capitalization Ranking page, a prominent financial portal providing comprehensive information on the South Korean stock market.

The Naver Finance Market Capitalization Ranking page is a valuable resource for investors seeking insights into the market dynamics of South Korean stocks. The webpage provides a detailed list of stocks ranked by market capitalization, offering a snapshot of the relative size and importance of each listed company within the market.

The primary focus of this analysis revolves around the exploration and understanding of various financial metrics associated with the listed stocks, with an emphasis on market capitalization. Market capitalization is a fundamental variable that encapsulates the collective valuation of a company in the stock market. By delving into additional variables such as trading volume, foreign ownership ratio, return on equity (ROE), and other financial indicators, this analysis aims to uncover patterns, correlations, and trends within the South Korean stock market.

The chosen website serves as a rich repository of data, allowing us to perform exploratory data analysis (EDA) and gain valuable insights into the dynamics of the market.

* What is the purpose of the analysis?

The goal is to explore patterns, relationships, and outliers in the South Korean stock market dataset. Insights into market capitalization, stock fluctuations, foreign ownership ratios, and other metrics aim to inform investment decisions.

* What is the structure of the dataset?

The dataset is sizable, containing information on South Korean stocks. It includes both categorical and numerical variables such as stock names, prices, market capitalization, and financial ratios.

* What are the key features or variables in the dataset

Crucial variables include stock names, current prices, market capitalization, trading volumes, fluctuation rates, foreign ownership ratios, ROE, and operating rates.

- Are there missing values in the dataset?  
Yes, the dataset contains missing values in three variables. Specifically, 'OperatingRate' has 674 missing values, 'ForeignRate' has 450 missing values, and 'ROE' has 1144 missing values. To ensure data integrity for subsequent analyses, an appropriate strategy, such as imputation or removal, will be implemented to address these missing values.

* What are the summary statistics of numerical variables

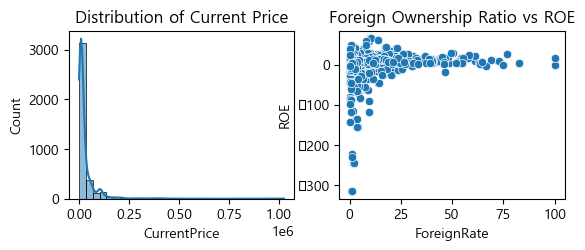
Mean Price: 28933.646153846155

Median Price: 11682.5

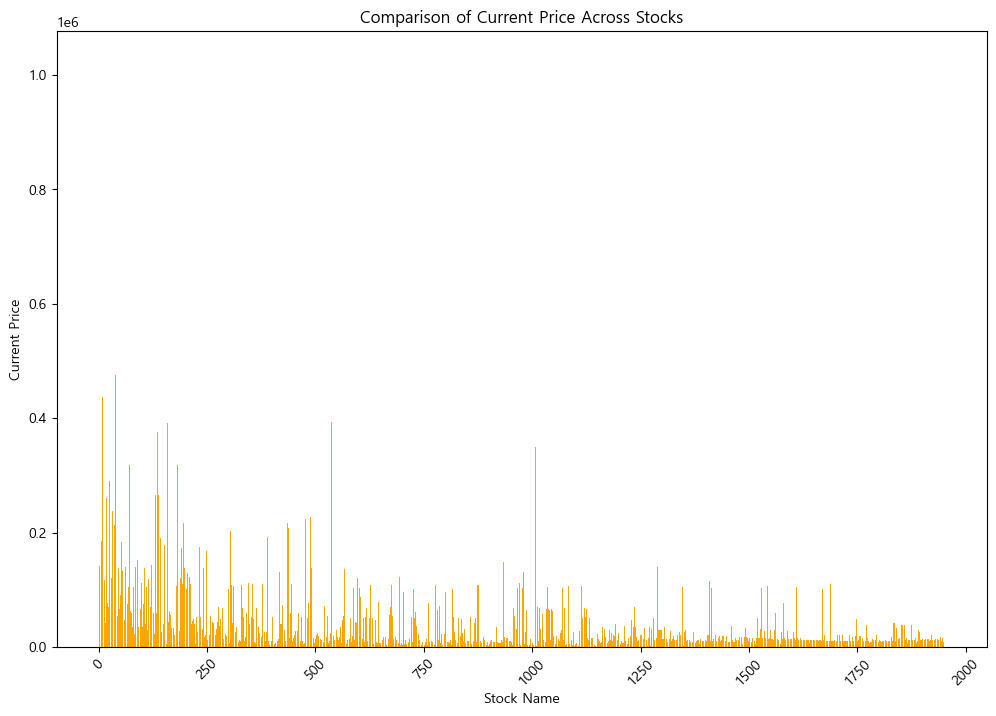
Standard Deviation of Price: 59013.582147297195

* How are categorical variables distributed?

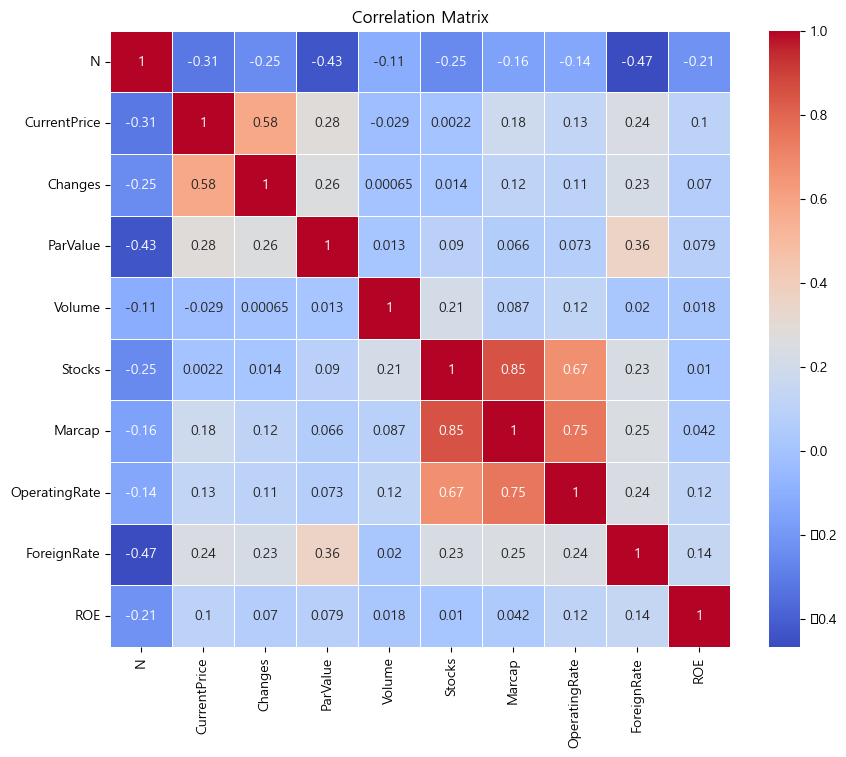
The visualization results indicate that the distribution of 'CurrentPrice' is left-skewed, with a notable concentration around the x-coordinate 0.0. The histogram, plotted with 30 bins and a kernel density estimate (kde), illustrates a prevalence of lower current prices.

The visualization results indicate that in the scatter plot of 'ForeignRate' against 'ROE,' the majority of data points are concentrated in the upper-left region. There is a notable trend with points clustering predominantly in the upper part of the plot. This suggests a potential positive relationship between the foreign ownership ratio ('ForeignRate') and the return on equity ('ROE'), with higher 'ROE' values associated with higher 'ForeignRate' values.

- Are there correlations between variables?

In the scatter plot depicting 'ForeignRate' against 'ROE,' the data points are relatively evenly distributed across the plot. However, a discernible trend is observed, characterized by a downward-sloping diagonal in the bottom right. This suggests a potential negative correlation between the foreign ownership ratio ('ForeignRate') and the return on equity ('ROE'), with lower 'ROE' values associated with higher 'ForeignRate' values

The vertical left and horizontal top sections of the matrix exhibit a blue (negative) distribution, indicating potential negative correlations. In the central area, there is a slight reddish (positive) trend, particularly in the diagonal descending from the middle-right section.

Specifically, the variables 'stock,' 'marcap,' and 'operating rate' demonstrate a notable positive correlation, as indicated by the reddish coloration. This suggests that these variables tend to move in tandem, with an increase in one associated with an increase in the others. The negative correlations observed in the top and left sections suggest potential inverse relationships between certain pairs of variables.

* What visualizations can help in understanding the data better?

Utilizing visualizations such as histograms, box plots, scatter plots, and other relevant charts will enhance our understanding of the dataset's characteristics and reveal patterns that may not be immediately apparent.