Written Assignment 03

AUTHOR
Owen Senowitz

PUBLISHED
September 21, 2024

Assignment Goal

The goal of this assignment is to demonstrate your understanding of exploratory data analysis. You may use any programming language of your choice. **R** and **Python** are the two popular languages for the **programmatic** data exploration.

Assignment Specification

The first step is to identify a **relevant** open-source dataset. The dataset should have at least 500 instances and contain a mix of at least ten **continuous** and **categorical** variables. The next step is to perform various exploratory data analysis tasks discussed in the class. The final step is to summarize your findings. You may use Quarto/RMarkdown or Jupyter/Python to respond to this assignment. Use this document as a template to prepare your response.

Some open-source data sources for this assignment are the following. This list is not exhaustive and you are not required to select a datset from this list. If you come across other repositories of open source datasets, please post links on MS Teams for the benefit of other students.

1. Integrated Postsecondary Education Data System (IPEDS) is a system of 12 interrelated survey components conducted annually that gathers data from every college, university, and technical and vocational institution that participates in the federal student financial aid programs. IPEDS Website
2. U.S. Securities and Exchange Commission (SEC) Financial Statement Data Sets
3. United Nations (UN) Datasets
4. World Bank Open Data
5. The Library of Congress Datasets
6. NASDAQ Historical Datasets
7. The World Factbook and Guide to Country Comparisons

1 The Data Quality Report

Document the data quality report in two separate tables, one for the continuous features and another for the categorical features. Use the table format discussed in the class.

1.1 Continuous Features

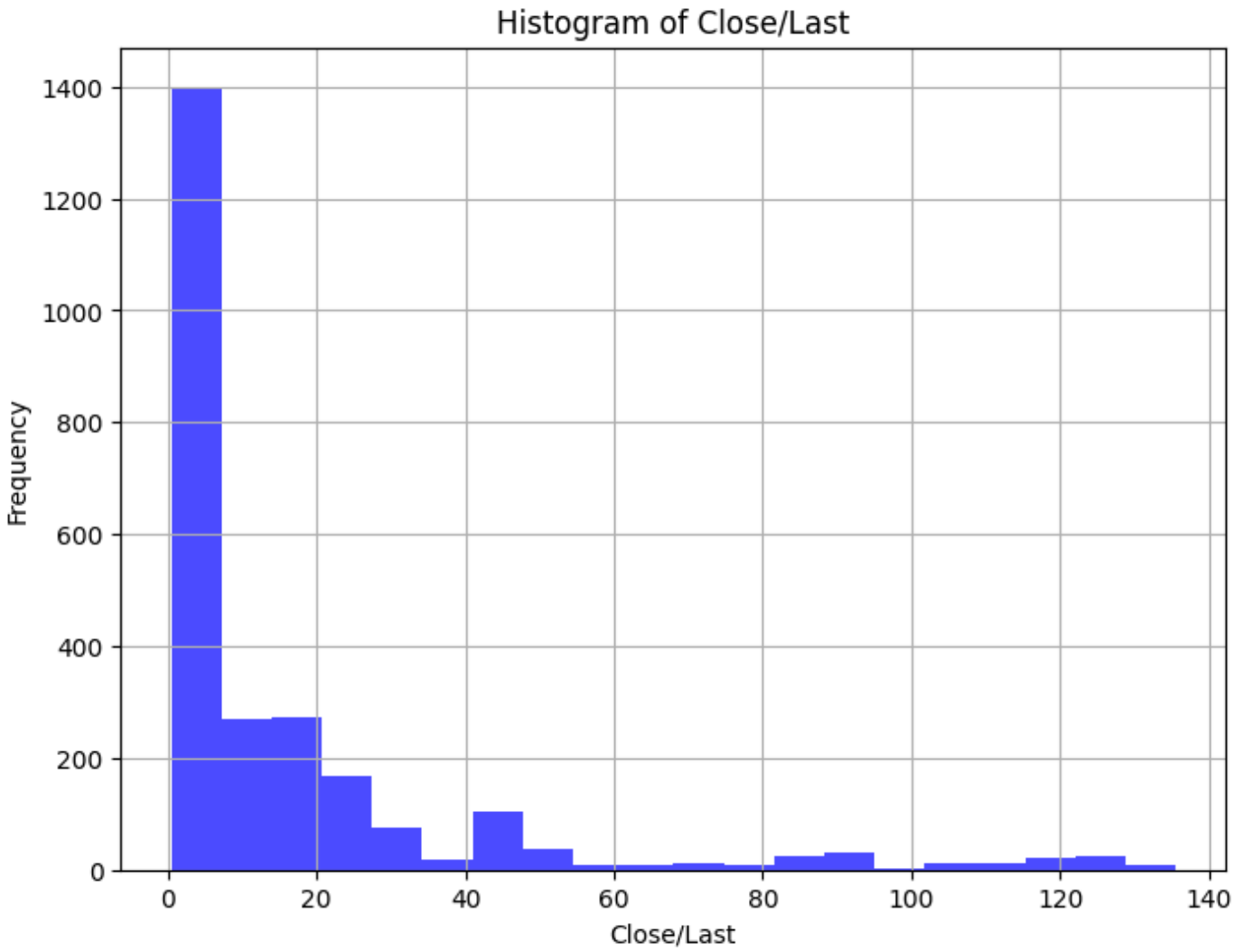
Feature	Mean	Median	Standard Deviation	Minimum	Maximum	Missing Values
Close/Last	17.03	6.1883	25.84	0.4196	135.58	0
Volume	467196589.63	415380400.0	252617500.70	45645120	3688131600	0
Open	17.03	6.1948	25.86	0.4233	139.8	0
High	17.35	6.2695	26.37	0.4325	140.76	0
Low	16.68	6.0972	25.27	0.4193	132.42	0

1.2 Categorical Feature

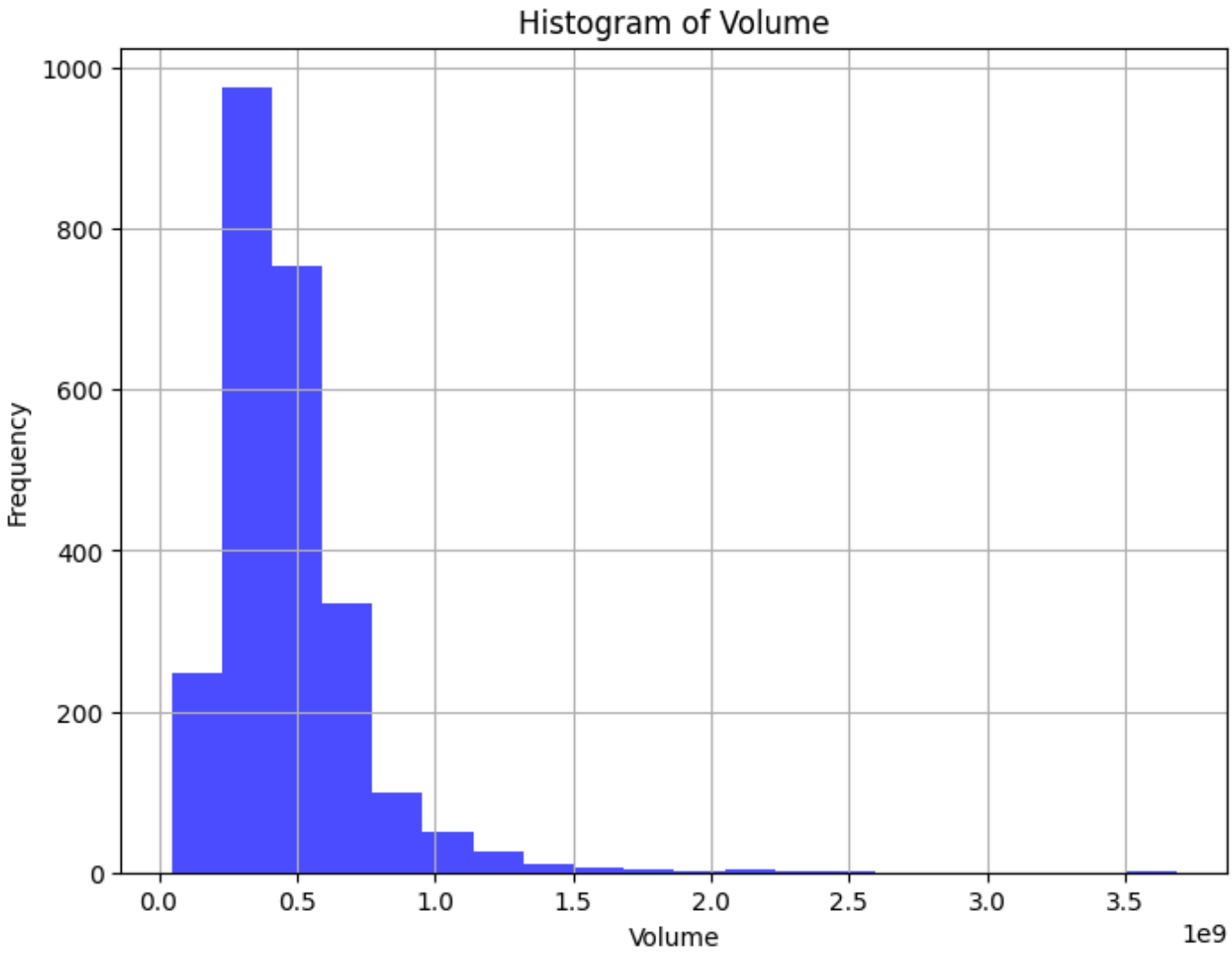
Feature	Unique Values	Missing Values	Most Frequent Value	Most Frequent Value Count
Date	2517	0	01/02/2015	1

2 Histograms of Continuous Features

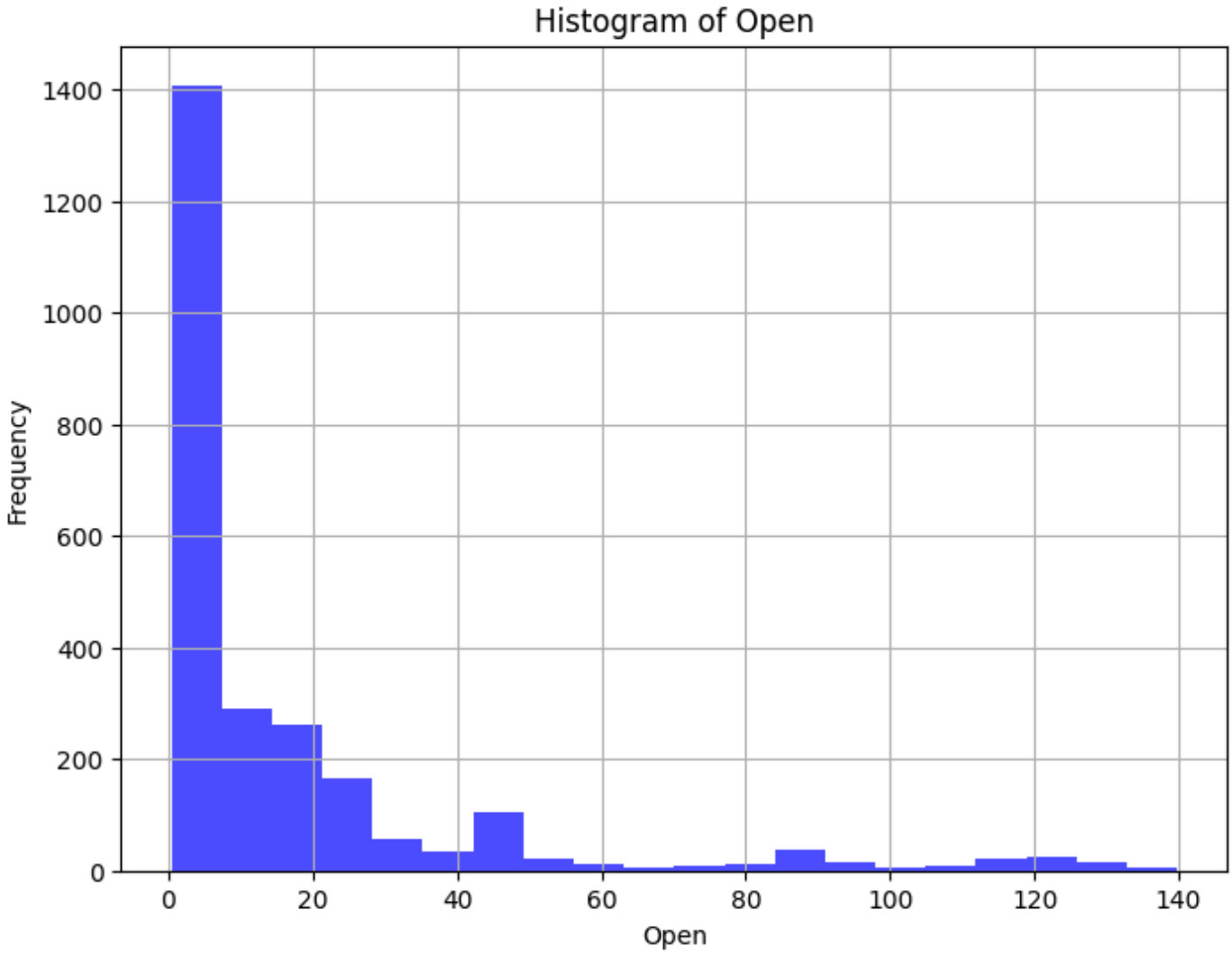
Create a **histogram** for each continuous feature. What probability distributions the histograms reveal? For example, uniform, normal (unimodal), unimodal (skewed right), unimodal (skewed left), exponential, and multimodal.



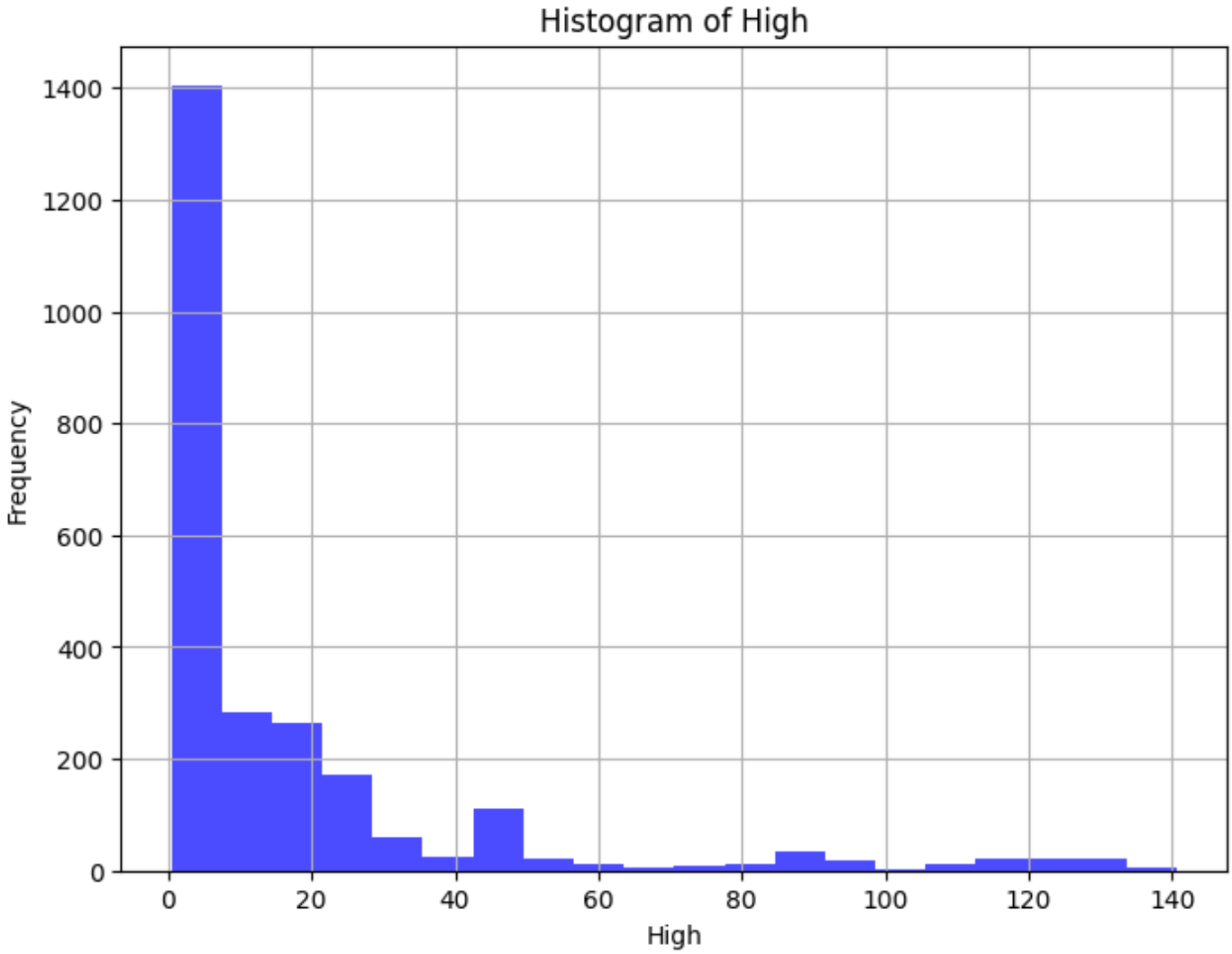
Skewed Right (Unimodal)



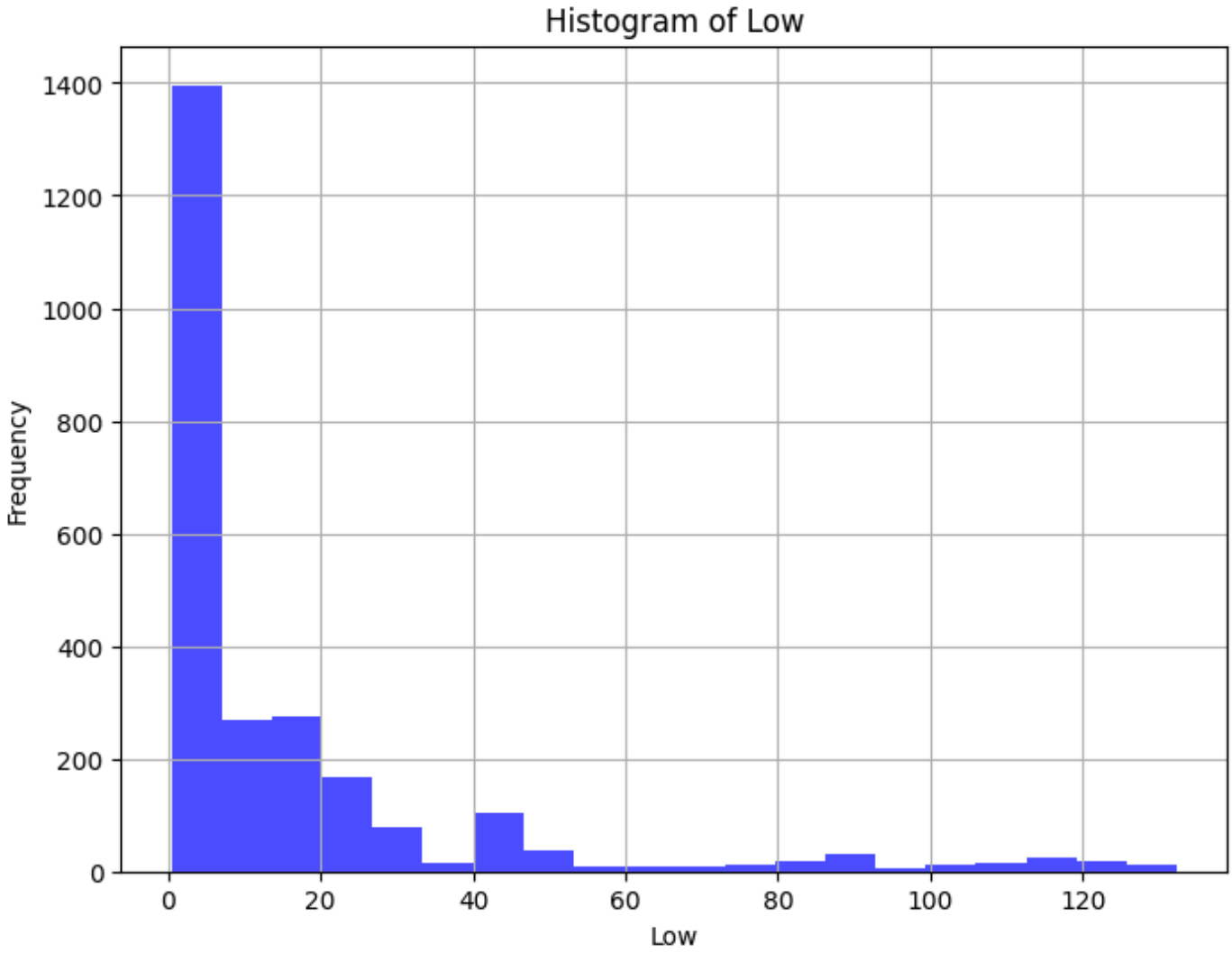
Skewed Right (Unimodal)



Skewed Right (Unimodal)



Skewed Right (Unimodal)



Skewed Right (Unimodal)

3 Identification of Data Quality Issues

Consider the missing values, irregular cardinality problems, and outliers. Summarize the **data quality issues** using a three-column table. The first column is the feature name, the second column is the associated data quality issue, and the third column describes potential handling strategies.

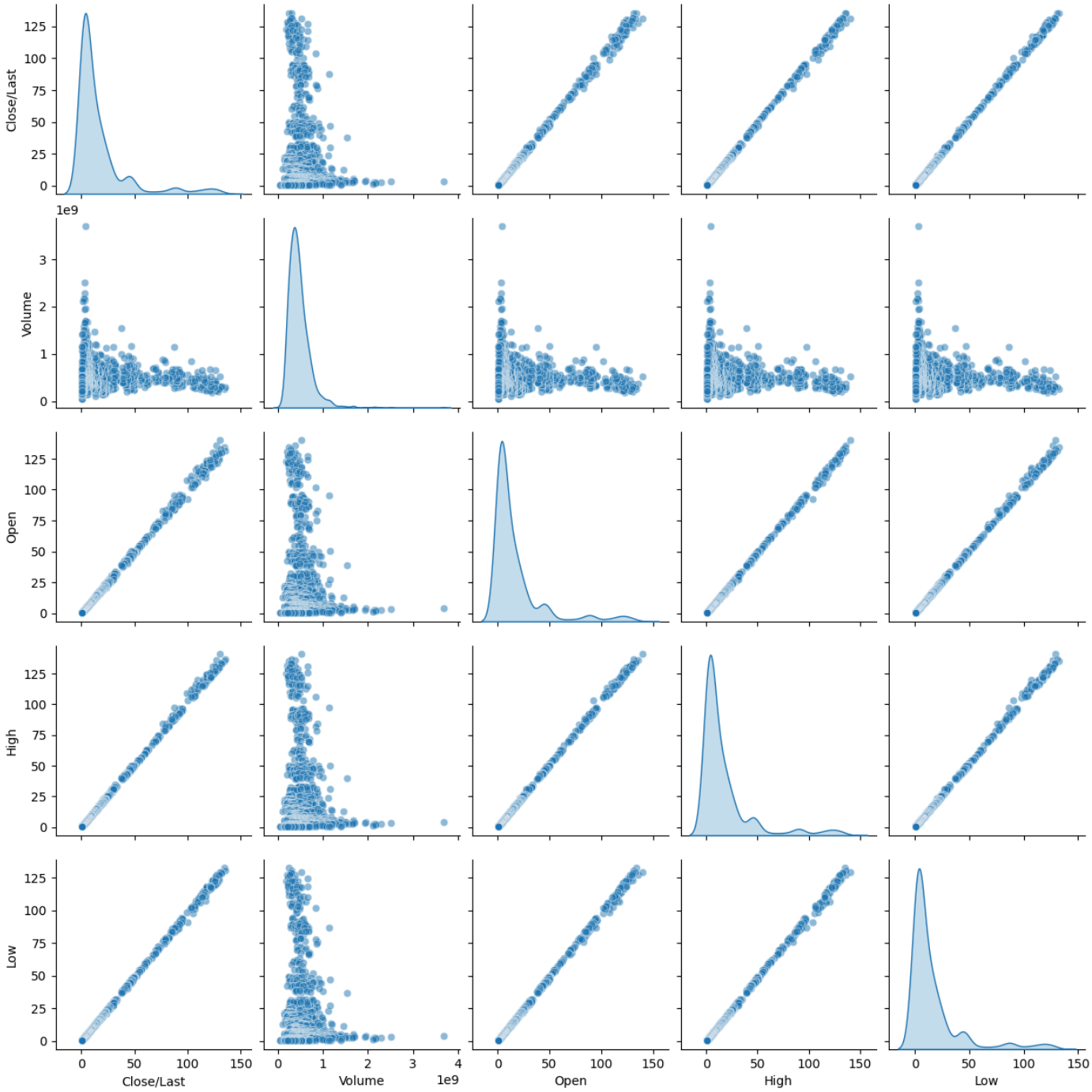
Since this is stock market data there are no data quality issues because it is highly regulated and handled by reliable sources.

Feature	Data Quality Issue	Handling Strategy
Close/Last	No issues detected	No action required
Volume	No issues detected	No action required
Open	No issues detected	No action required
High	No issues detected	No action required

Feature	Data Quality Issue	Handling Strategy
Low	No issues detected	No action required
Date	No issues detected	No action required

4 Scatterplot Matrix

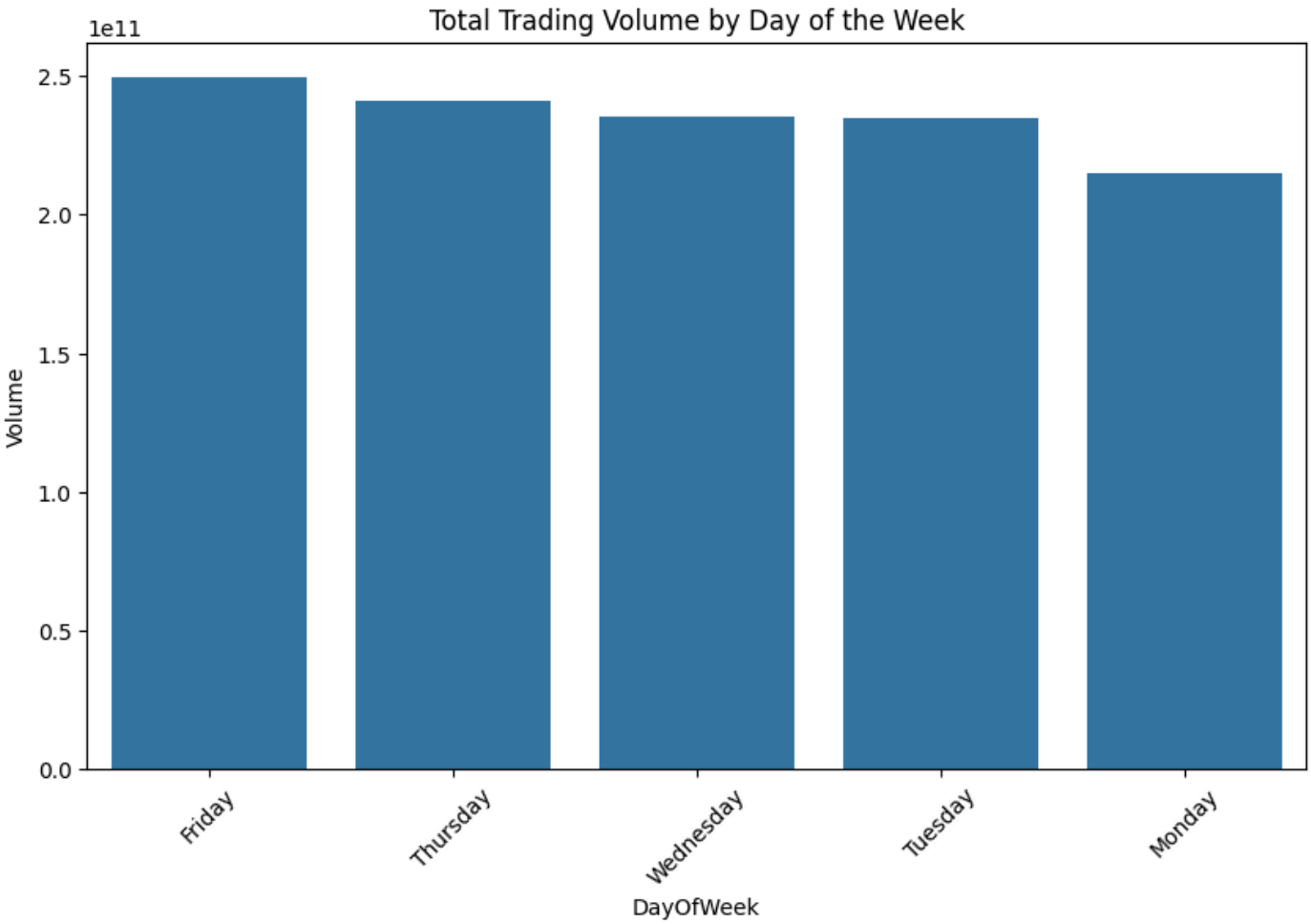
Construct the **scatterplot matrix** for the continuous features and comment on what you observed.

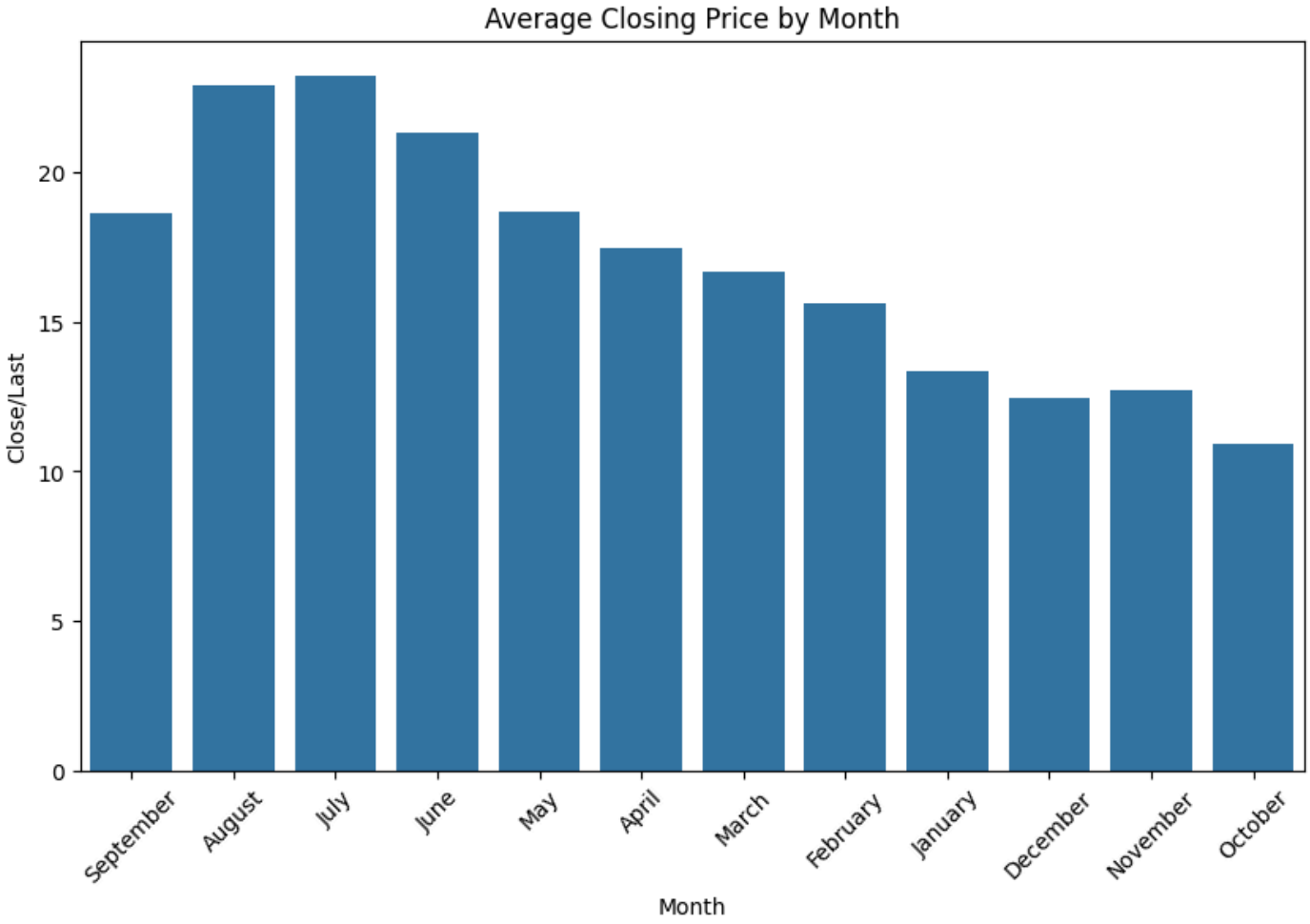


The scatterplot matrix shows a strong positive correlation between the price-related features (Close/Last, Open, High, Low), indicating that they move together in a linear fashion. Volume, on the other hand, does not exhibit any clear relationship with the price features, as seen by the dispersed points in the scatterplots. The distribution of Volume is right-skewed, indicating that most trading days have lower volumes with a few days showing very high volumes.

5 Visualizing Pairs of Categorical Features

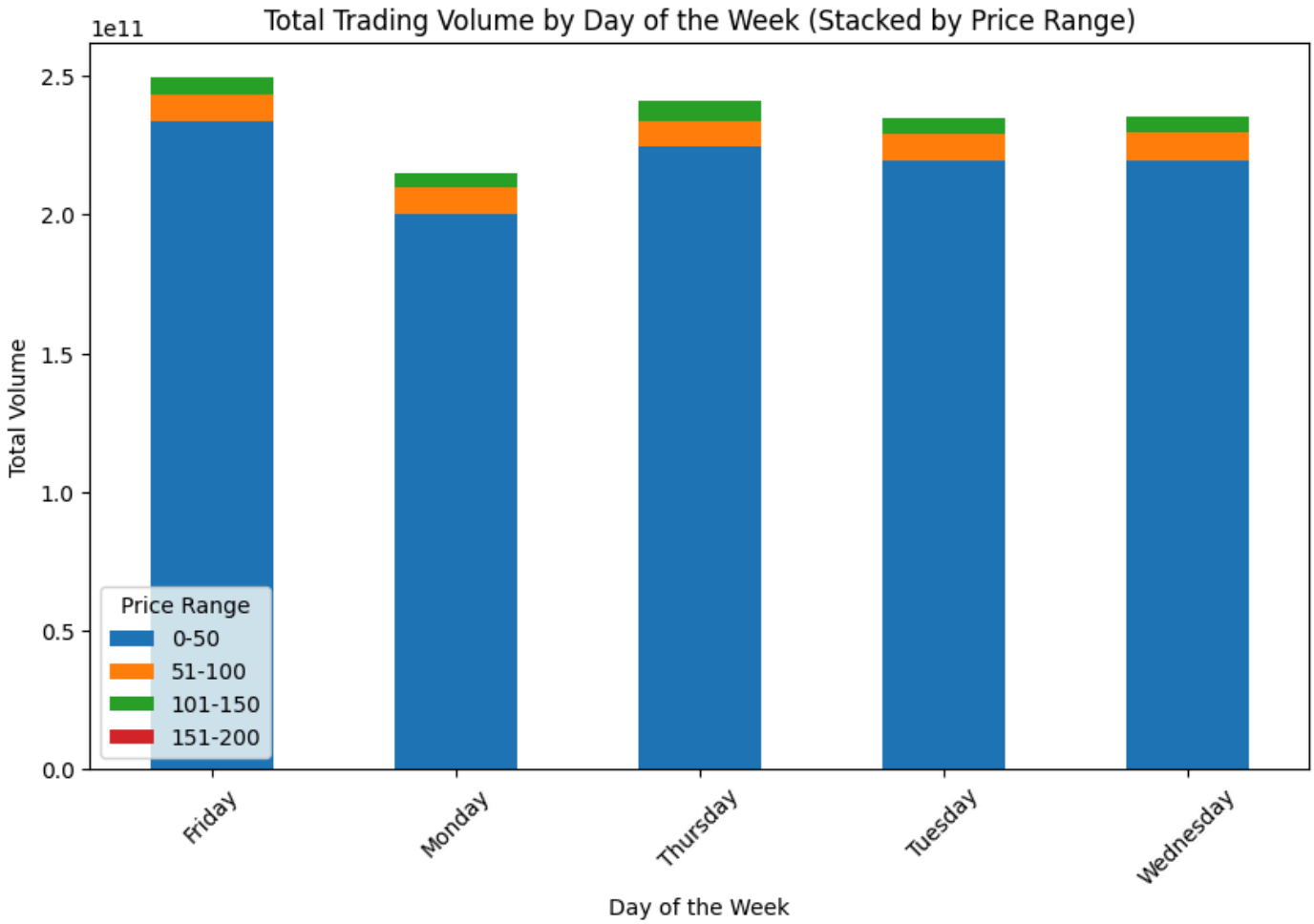
Use multiple **barplot visualizations**.


Total Trading Volume by Day of the Week
1e11
2.5
2.0
1.5
1.0
0.5
0.0
Volume
Friday
Thursday
Wednesday
Tuesday
Monday
DayOfWeek



6 Visualizing Relationship Between a Catergorical and Continuous Feature

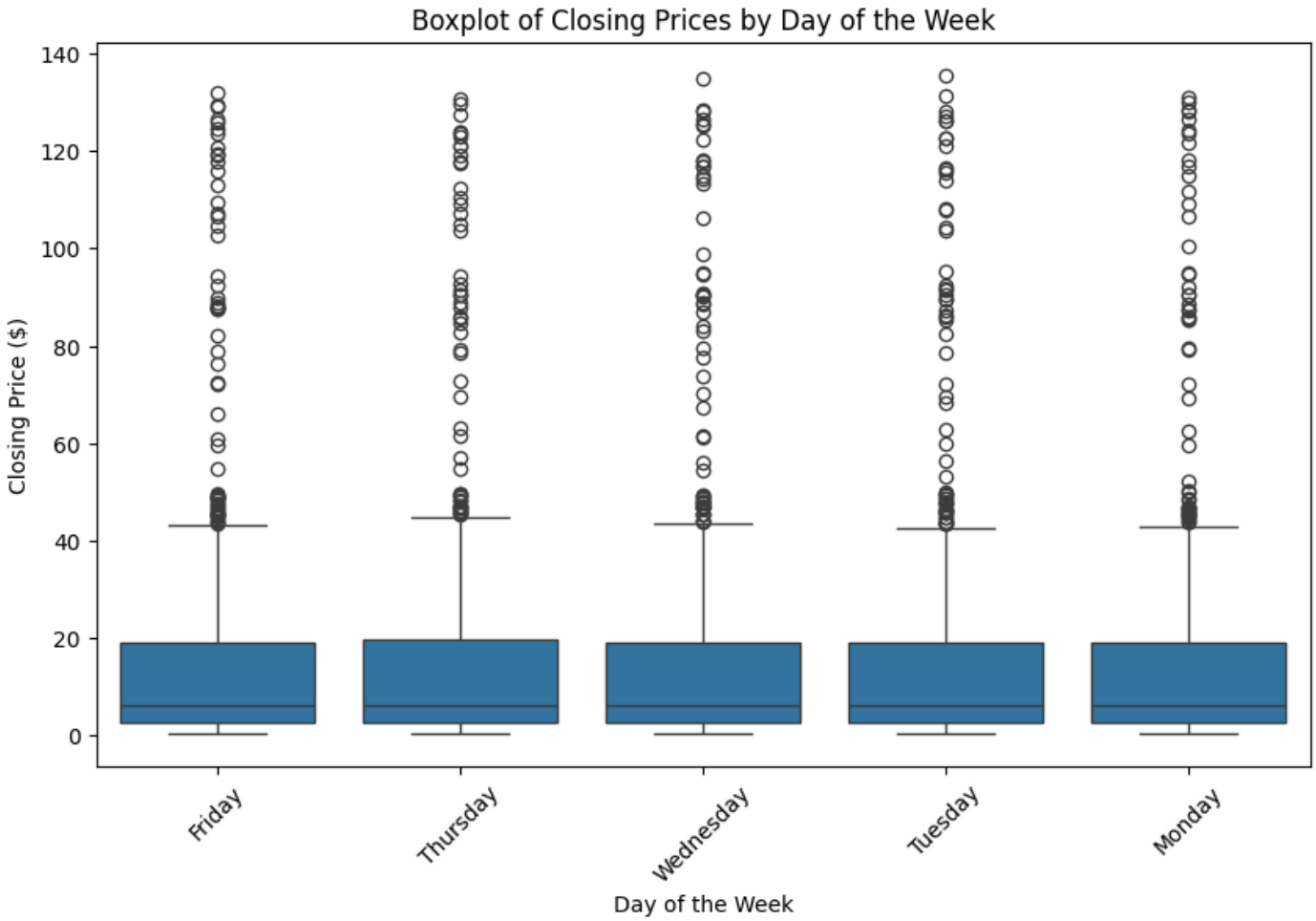
For a subset of the categorical and continuous features, perform **stacked barplot visualizations**. Comment on what you observed.



The stacked barplot shows total trading volume by day of the week, segmented by different price ranges. Friday exhibits the highest total trading volume, with the majority falling into the lower price range (0-50). All other days have relatively similar trading volumes, with Monday showing a slightly lower total volume compared to the other days. The higher price ranges (e.g., 101-150) contribute less to the overall volume compared to the lower price ranges across all days.

7 Boxplot Visualizations

For a subset of the categorical and continuous features, perform **boxplot visualizations**. Comment on what you observed.



The boxplot shows the distribution of closing prices for each day of the week. Across all days, the median closing price remains fairly consistent, with most prices falling within the same range. However, there is a significant number of outliers (prices above $40), which is typical in stock market data due to price volatility. These outliers are present across all days, indicating that extreme price movements are not isolated to a specific day of the week. The interquartile range (IQR) is also similar for each day, indicating that the general spread of stock prices remains stable throughout the week.

8 Covariance Matrix

For the continuous features, construct the **covariance matrix**. Comment on what you observed.

	Close/Last	Volume	Open	High	Low
Close/Last	6.676405e+02	-2.717345e+08	6.678506e+02	6.811328e+02	6.529092e+02
Volume	-2.717345e+08	6.381560e+16	-2.675637e+08	-2.585401e+08	-2.800415e+08
Open	6.678506e+02	-2.675637e+08	6.688159e+02	6.818036e+02	6.534324e+02
High	6.811328e+02	-2.585401e+08	6.818036e+02	6.952984e+02	6.662445e+02
Low	6.529092e+02	-2.800415e+08	6.534324e+02	6.662445e+02	6.387599e+02

9 Correlation Matrix

For the continuous features, construct the **correlation matrix**. Comment on what you observed.

	Close/Last	Volume	Open	High	Low
Close/Last	1.000000	-0.041630	0.999435	0.999712	0.999799
Volume	-0.041630	1.000000	-0.040955	-0.038813	-0.043862
Open	0.999435	-0.040955	1.000000	0.999817	0.999720
High	0.999712	-0.038813	0.999817	1.000000	0.999722
Low	0.999799	-0.043862	0.999720	0.999722	1.000000

10 Range Normalization

List the continuous features that require **range normalization**. What is the rationale for your selection? Perform the range normalization and show the values before and after the normalization.

Before Normalization:

Close/Last	Volume	Open	High	Low
116.00	382462400	117.06	118.6181	115.3901
117.87	293506400	117.35	119.6600	117.2500
113.37	310318900	115.89	117.7000	113.2200
115.59	231925900	118.17	118.8000	114.8300
116.78	248772300	116.79	118.1800	114.3600

After Normalization:

Close/Last	Volume	Open	High	Low
0.855135	0.092469	0.836845	0.842213	0.870986
0.868970	0.068047	0.838926	0.849637	0.885076
0.835677	0.072663	0.828451	0.835670	0.854546
0.852102	0.051141	0.844809	0.843509	0.866743
0.860906	0.055766	0.834908	0.839091	0.863183

11 Binning

Do you see the need for converting a subset of the continuous features into categorical features? Select two such continuous features and convert the first into a categorical feature using the **equal-width binning*** and the second using **equal-frequency binning**. Show the feature values after the equal-width and equal-frequency binning.

Original and Equal-width Binned 'Close/Last':

Close/Last	Close/Last_Binned_Equal_Width
116.00	Very High
117.87	Very High
113.37	Very High
115.59	Very High
116.78	Very High

Original and Equal-frequency Binned 'Volume':

Volume	Volume_Binned_Equal_Frequency
382462400	Medium
293506400	Low
310318900	Medium
231925900	Low
248772300	Low

12 Undersampling

Do you see a need for undersampling? **Undersampling** is used to reduce the instances from the majority class so that the final dataset is balanced. For example, a binary classification problem has a target/outcome variable that takes two values, say, *approved* and *denied*. In the dataset, if 70% of the instances have the *approved* value for the target variable, the dataset is *imbalanced*. Ideally, the dataset should have approximately equal number of instances for each the values the target variable takes. This article illustrates the undersampling.

PriceChange Decrease 1373 Increase 1144 Name: count, dtype: int64

Since the difference between the two classes is not drastic. The imbalance ratio is about 55% to 45% which is relatively mild compared to highly imbalanced datasets where one class dominates the other example 80% vs 20%

13 Oversampling

Oversampling arises when we have too few instances from a class (called the minority class) relative to other classes. To boost the participation of the minority class in the (training) dataset, more observations from the minority class are generated usually by replicating the samples from the minority class.

In your dataset, do you see the need for oversampling? If so, which features require oversampling?

While there is an imbalance, it's not severe. In scenarios where the imbalance is small like here, you could handle it by adjusting class weights in your model similar to the strategy for avoiding undersampling.

14 Summary

Summarize the findings you have discovered through the exploratory data analysis. The summary should about a page and should serve as an executive report for non-technical people.

In this analysis of stock market data, we observed that there is a mild class imbalance between price increases and decreases, with decreases slightly outnumbering increases. The price-related features show strong positive correlations, while trading volume has little correlation with price movements. Given the regulated nature of stock data, no significant data quality issues were detected, and strategies like adjusting class weights may be more appropriate than oversampling or undersampling to handle the small imbalance.