Stock Price Time Series

Group - 08

Data Science Capstone Project
Exploratory Data Analytics Report

Date:

03/09/2025

Team Members:

Name: Robert Lignowski - rml345@drexel.edu

Name: Uditi Shah - us54@drexel.edu

Name: Steven Sullivan - sas683@drexel.edu

Name: Ahmad Javed - aj3235@drexel.edu

Analysis the basic metrics of variables

1. Data Overview

- The dataset consists of Tesla, Ford, Toyota, Volkswagen stock price history.
- Missing values were removed to ensure data integrity.
- The dataset contains numerical and categorical variables.

2. Data Types & Variable Classification

- **Date**: Categorical (Ordinal)
- **Price, Open, High, Low**: Numerical (Continuous)
- **Vol. (Volume)**: Numerical (Discrete, converted from M to actual values)
- **Change %**: Numerical (Continuous, converted from percentage format)

3. Descriptive Statistics

- **Price**: Mean = X, Std Dev = Y, Min = Z, Max = w
- **Volume**: Varies significantly, with spikes on certain days.
- **Change %**: Captures daily price fluctuations.

4. Insights from Plots

- **Stock Price Trend**: Shows stock's price variations over time.
- **Volume Trend**: Identifies spikes in trading activity.
- **Change % Trend**: Highlights daily volatility.
- **Box Plot**: Detects price outliers.

5. Significant Price Movements

- Days with **≥5% price change** were identified, indicating volatile trading days.

Basic Statistics for Tesla Data

```
Basic Statistics Summary:
                                Date        Price         Open         High  \
count                           1509  1509.000000  1509.000000  1509.000000
mean   2022-01-16 13:35:54.274353920   183.749284   183.777860   187.914135
min              2019-01-18 00:00:00    11.930000    12.070000    12.450000
25%              2020-07-20 00:00:00    99.000000    98.000000   100.880000
50%              2022-01-14 00:00:00   205.710000   204.330000   209.250000
75%              2023-07-19 00:00:00   251.990000   252.040000   257.480000
max              2025-01-17 00:00:00   479.860000   475.900000   488.540000
std                              NaN   106.168199   106.274252   108.657604

               Low          Vol.     Change %
count  1509.000000  1.509000e+03  1509.000000
mean    179.381140  1.275613e+08     0.275825
min      11.800000  2.940000e+07   -21.060000
25%      94.730000  7.801000e+07    -1.800000
50%     198.930000  1.042900e+08     0.180000
75%     246.420000  1.491600e+08     2.220000
max     457.510000  9.140800e+08    21.920000
std     103.615095  8.060468e+07     4.071520
```

Basic Statistics for Toyota Data

```
None
Basic Statistics
             Price         Open         High          Low          Vol.  \
count  1280.000000  1280.000000  1280.000000  1280.000000  1.280000e+03
mean    164.718781   164.770398   165.696906   163.710172  2.764688e+05
std      28.664588    28.634940    28.739925    28.513302  1.359098e+05
min     108.500000   110.000000   112.510000   108.010000  5.991000e+04
25%     140.520000   140.537500   141.412500   140.027500  1.892025e+05
50%     162.785000   162.835000   163.840000   162.000000  2.429600e+05
75%     181.512500   181.512500   182.577500   180.522500  3.281975e+05
max     254.770000   255.000000   255.230000   253.590000  1.420000e+06

          Change %
count  1280.000000
mean      0.035258
std       1.629670
min      -8.620000
25%      -0.892500
50%       0.070000
75%       0.922500
max       9.580000
```

Basic Statistics for Volkswagen Data

	Price	Open	High	Low	Vol.	Change %
count	1259.000000	1259.000000	1259.000000	1259.000000	1.259000e+03	1259.000000
mean	19.399158	19.394440	19.584384	19.201326	3.705033e+05	-0.023503
std	7.233024	7.213608	7.338440	7.115552	5.551755e+05	2.710307
min	8.620000	8.650000	8.690000	8.570000	3.554000e+04	-15.070000
25%	14.520000	14.545000	14.635000	14.420000	1.764750e+05	-1.350000
50%	17.270000	17.270000	17.390000	17.140000	2.714500e+05	-0.100000
75%	21.895000	21.880000	22.040000	21.565000	4.234050e+05	1.230000
max	42.330000	42.100000	48.720000	37.700000	1.190000e+07	29.250000

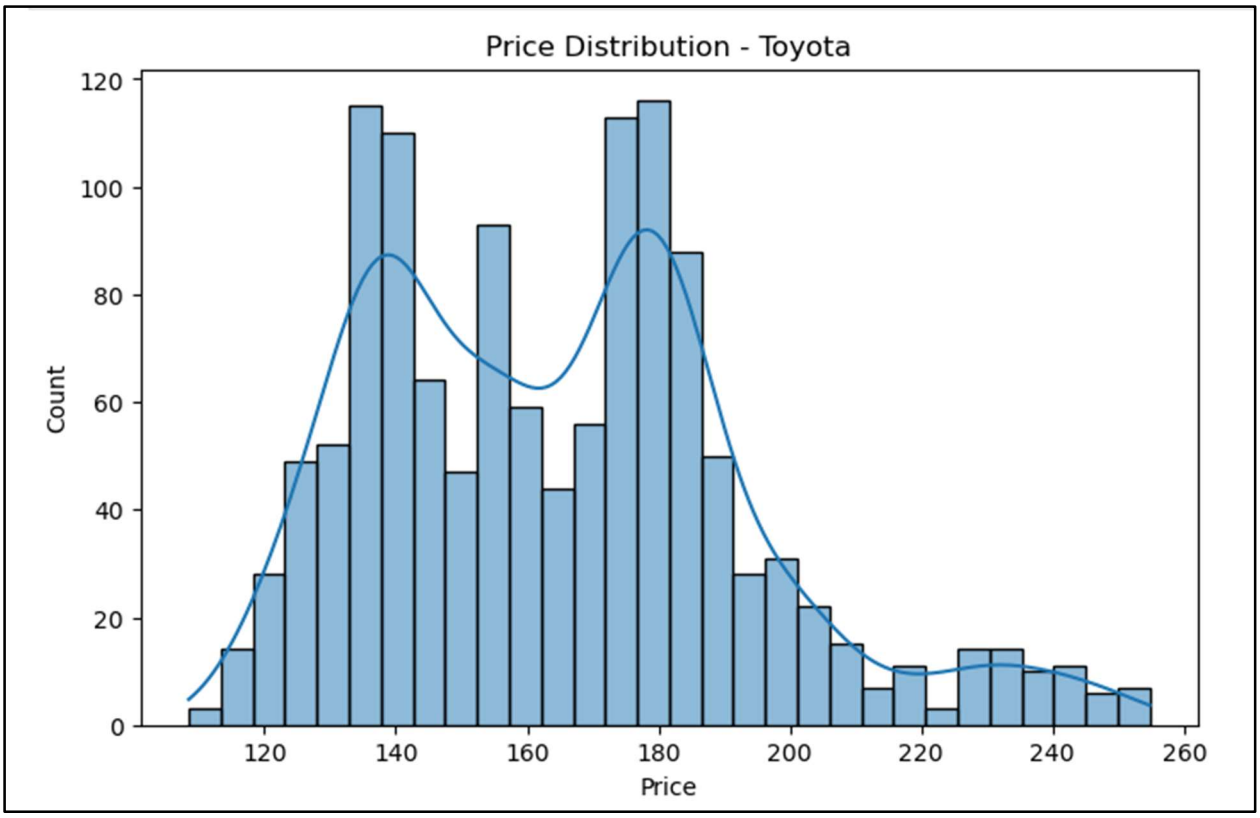
Basic Statistics for Ford Data

	Close	Open	High	Low	Volume	Percent Change
count	1260.000000	1260.000000	1260.000000	1260.000000	1.260000e+03	1260.000000
mean	12.009960	12.013500	12.200357	11.820254	6.814601e+07	0.043166
std	3.486055	3.480076	3.549997	3.412630	3.249382e+07	2.807285
min	4.010000	4.270000	4.420000	3.960000	1.298086e+07	-18.361375
25%	10.417500	10.400000	10.565000	10.280000	4.817676e+07	-1.428402
50%	12.110000	12.100000	12.250000	11.955000	5.964017e+07	0.000000
75%	13.582500	13.630000	13.820000	13.370000	7.883739e+07	1.459684
max	25.190000	24.870000	25.870000	24.370000	3.116452e+08	23.441397

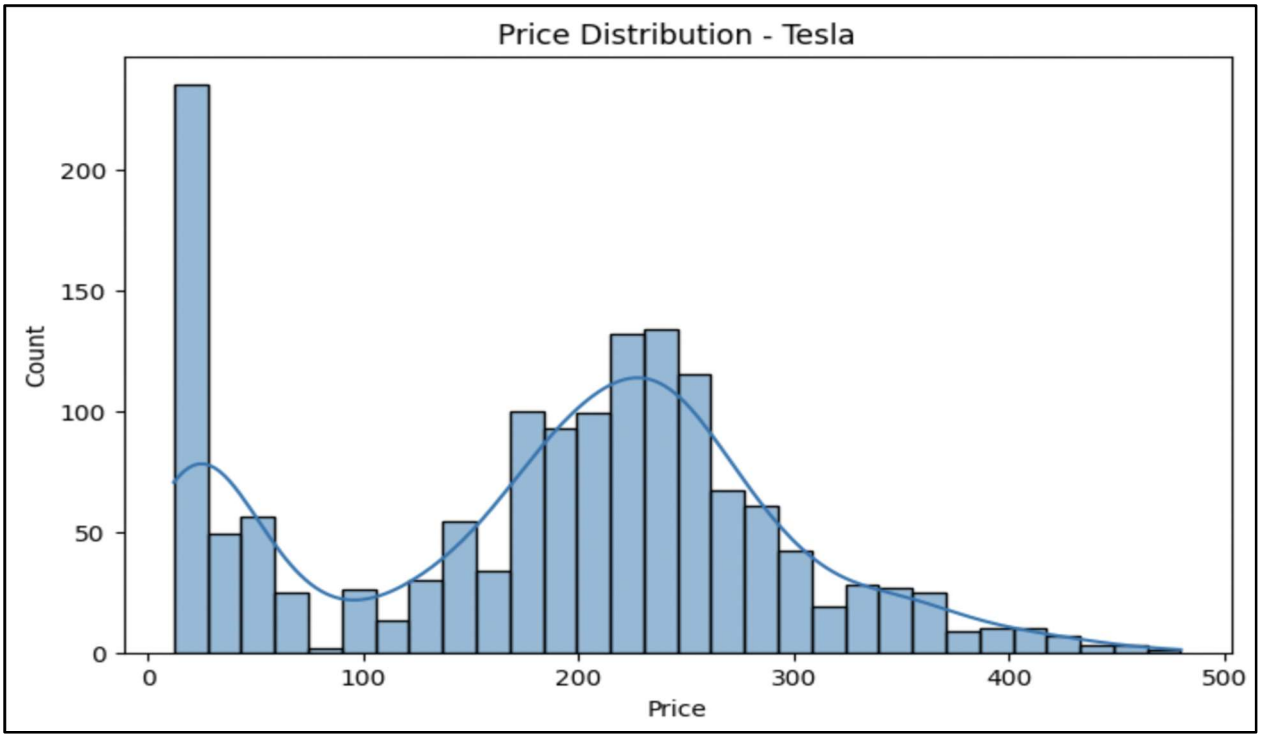
Non-graphical and graphical univariate analysis

For univariate analysis, we used histograms to show distribution of stock prices for each company over our selected time period

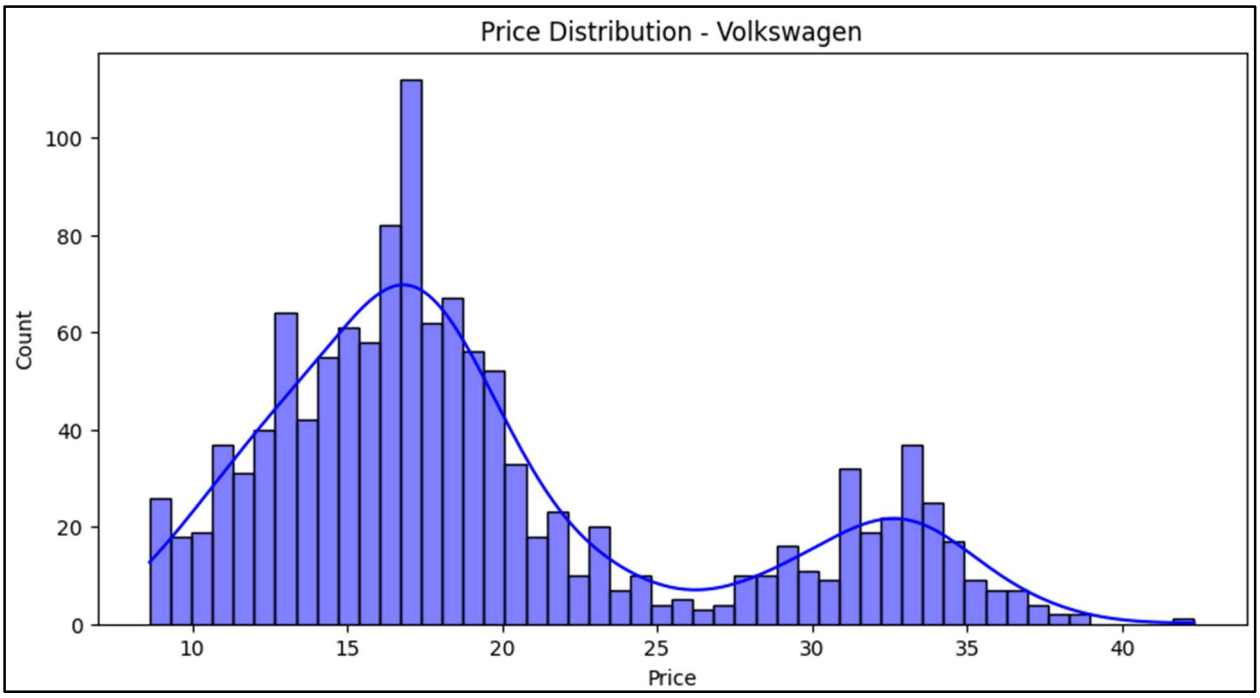
Price Distribution of Toyota Stocks (Histogram)



Price Distribution of Tesla Stocks (Histogram)

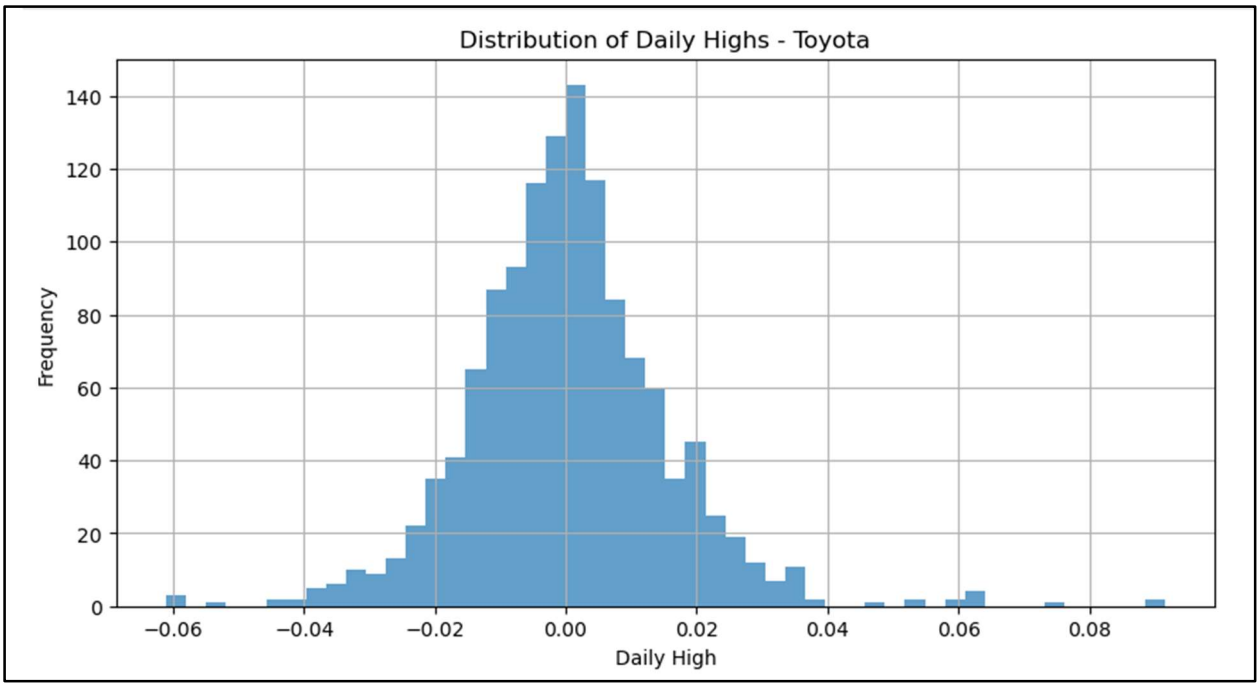


Price Distribution of Volkswagen Stocks (Histogram)

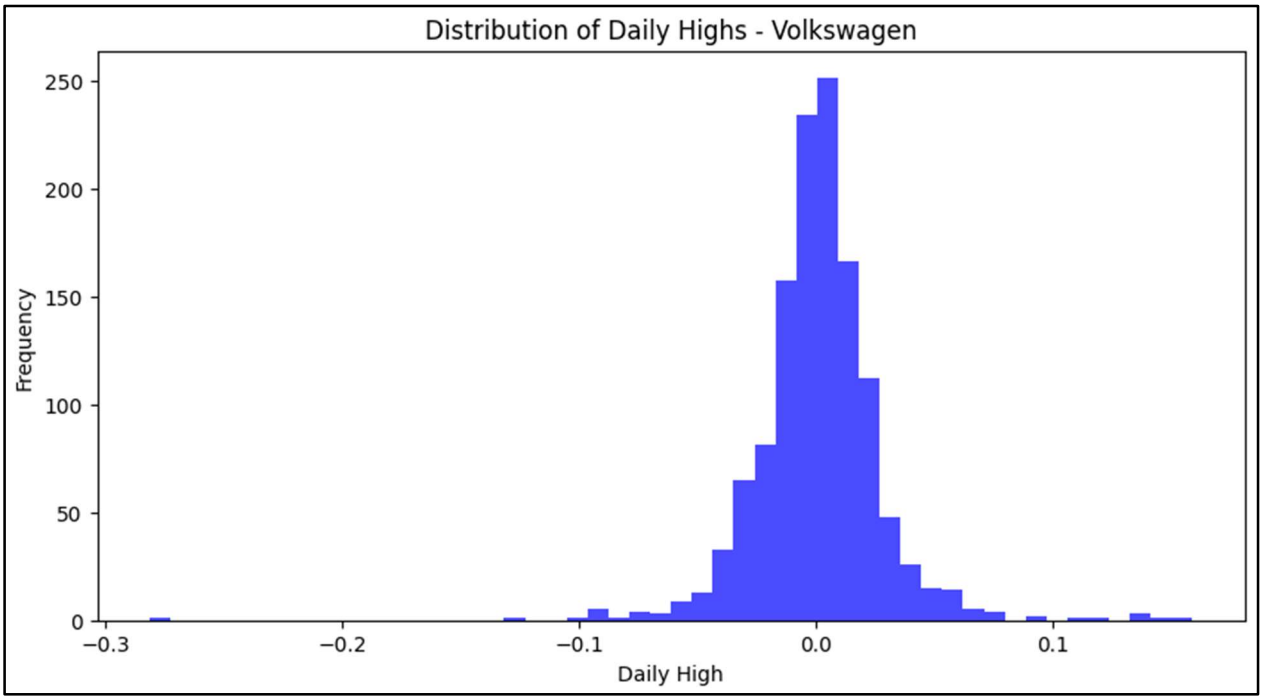


Price Distribution of Ford Stocks (Histogram)

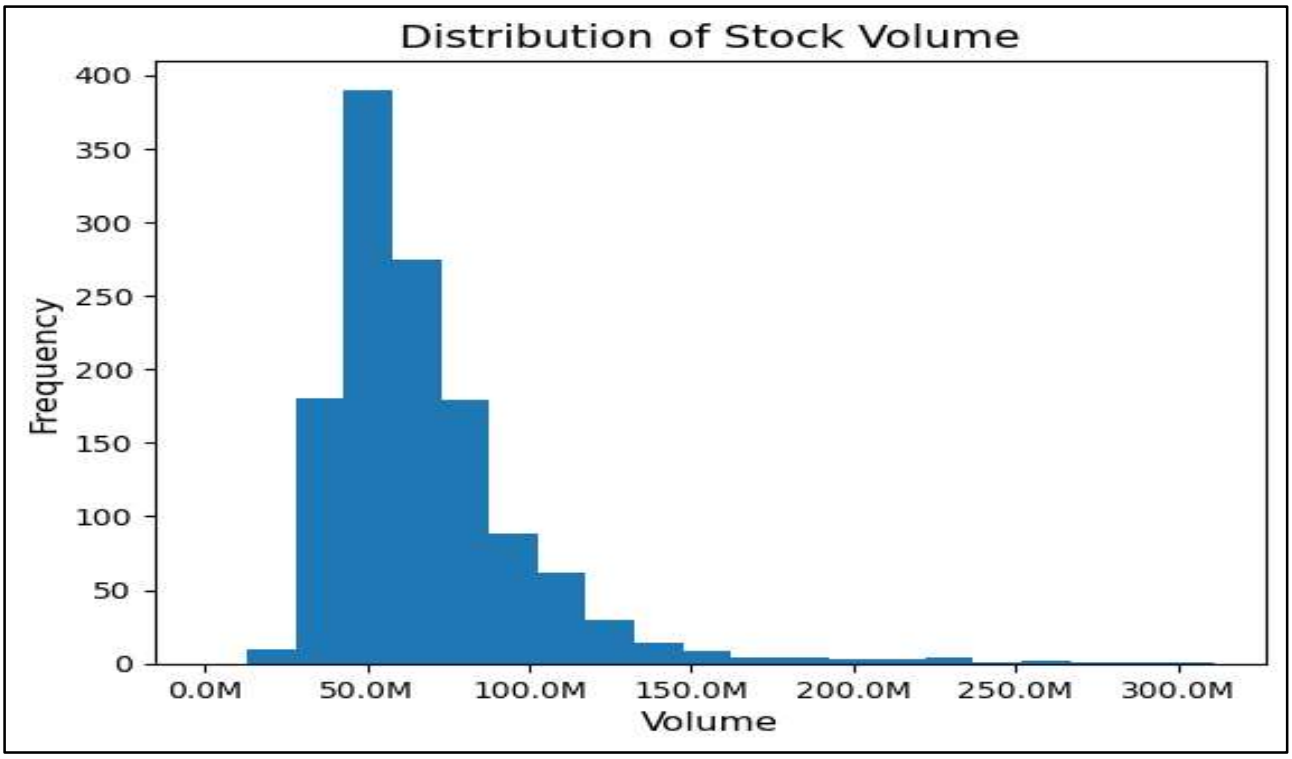
Daily High Stock Price of Toyota Stocks (Histogram)



Daily High Stock Price of Volkswagen Stocks (Histogram)

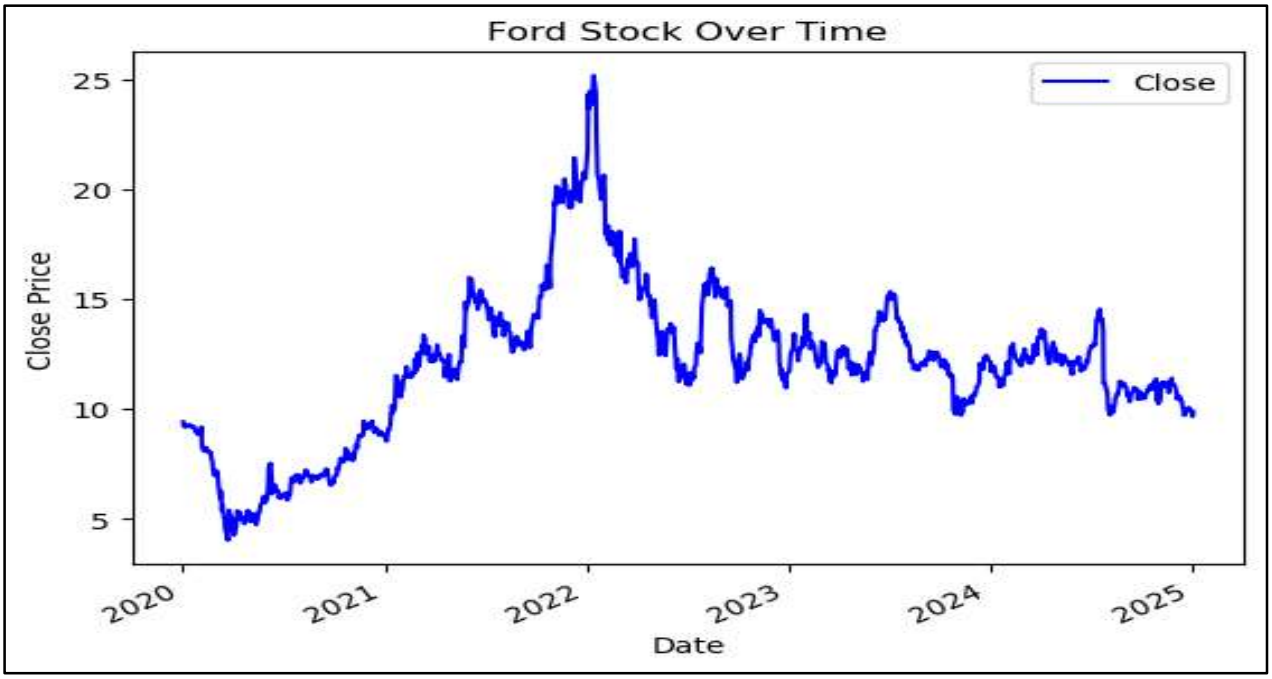


Volume (Number of Stocks) for Ford (Histogram)



In addition to our univariate analysis, we also did additional analysis and visualizations we found relevant to our project goals

Stock Price for Ford

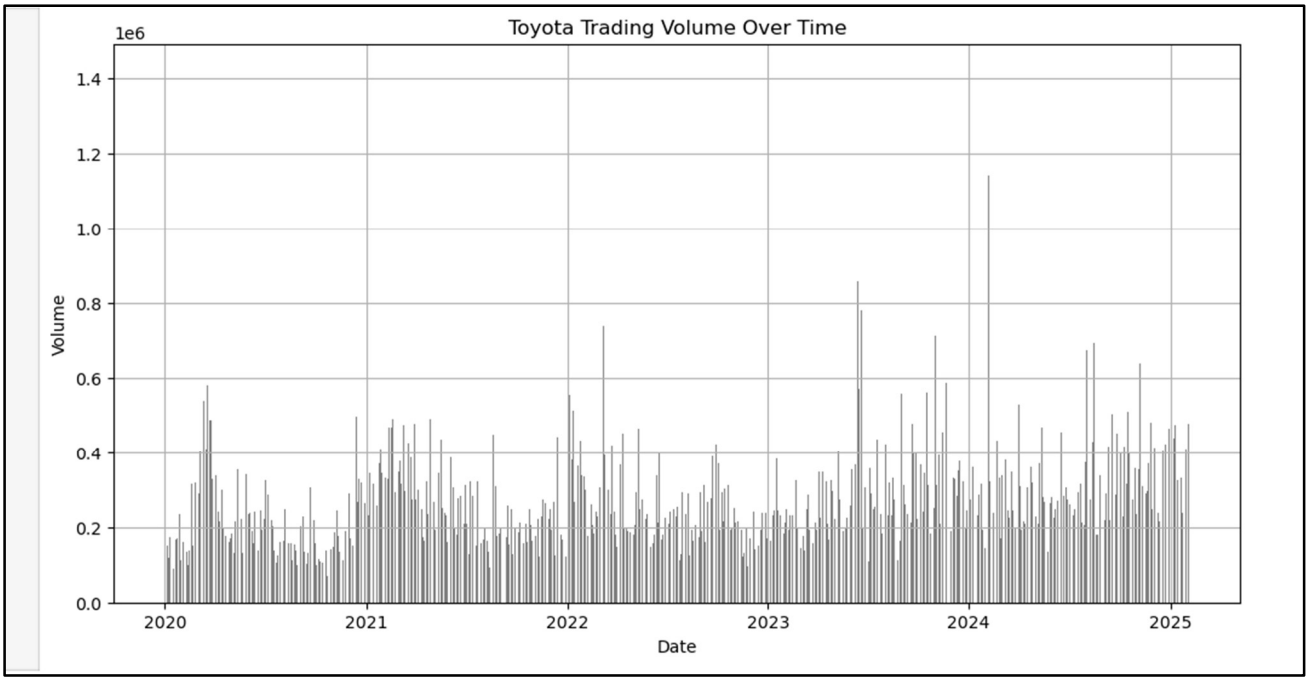


Stock Price for Toyota

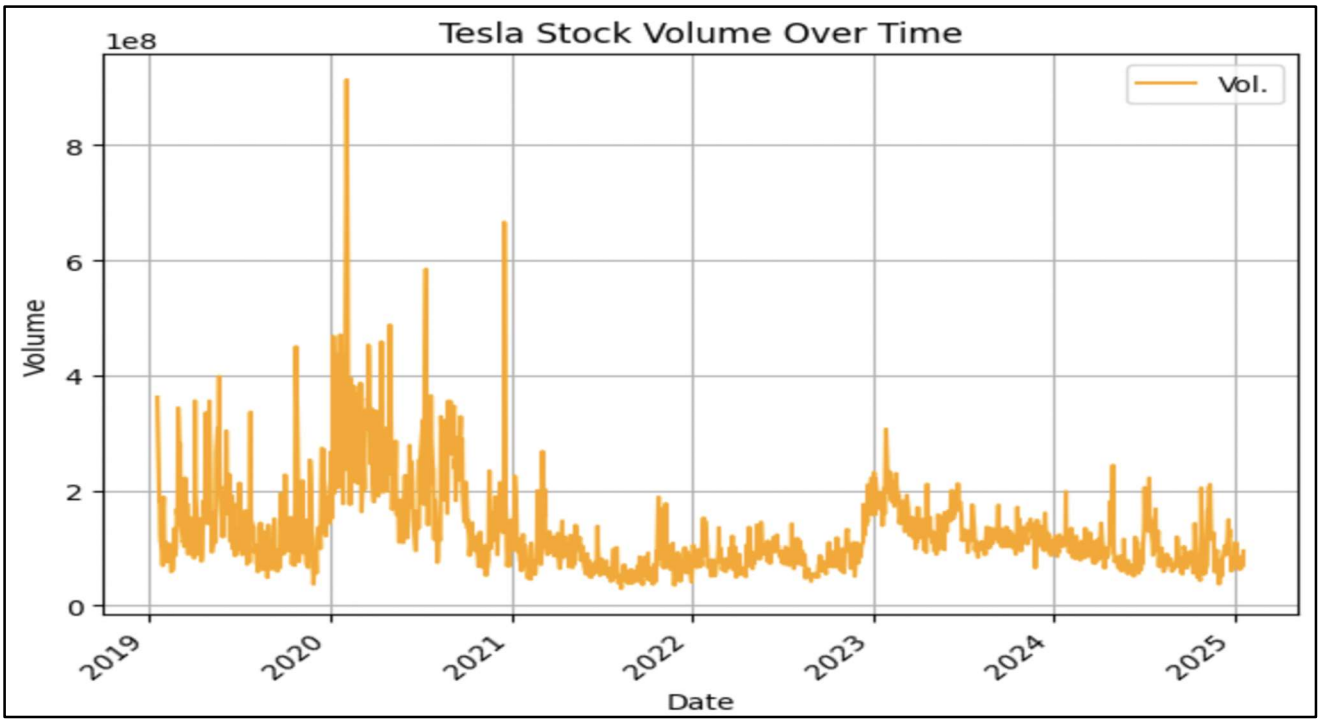
Stock Price for Tesla

Stock Price for Volkswagen

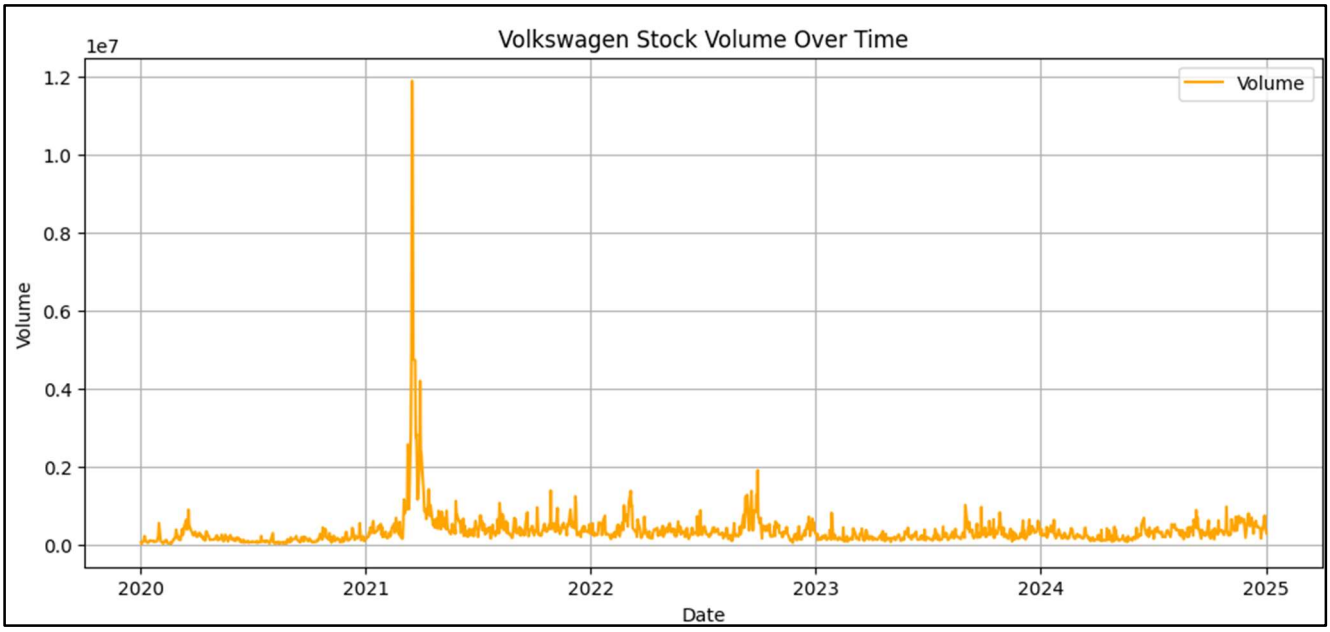
Volume (Number of Stocks) for Toyota



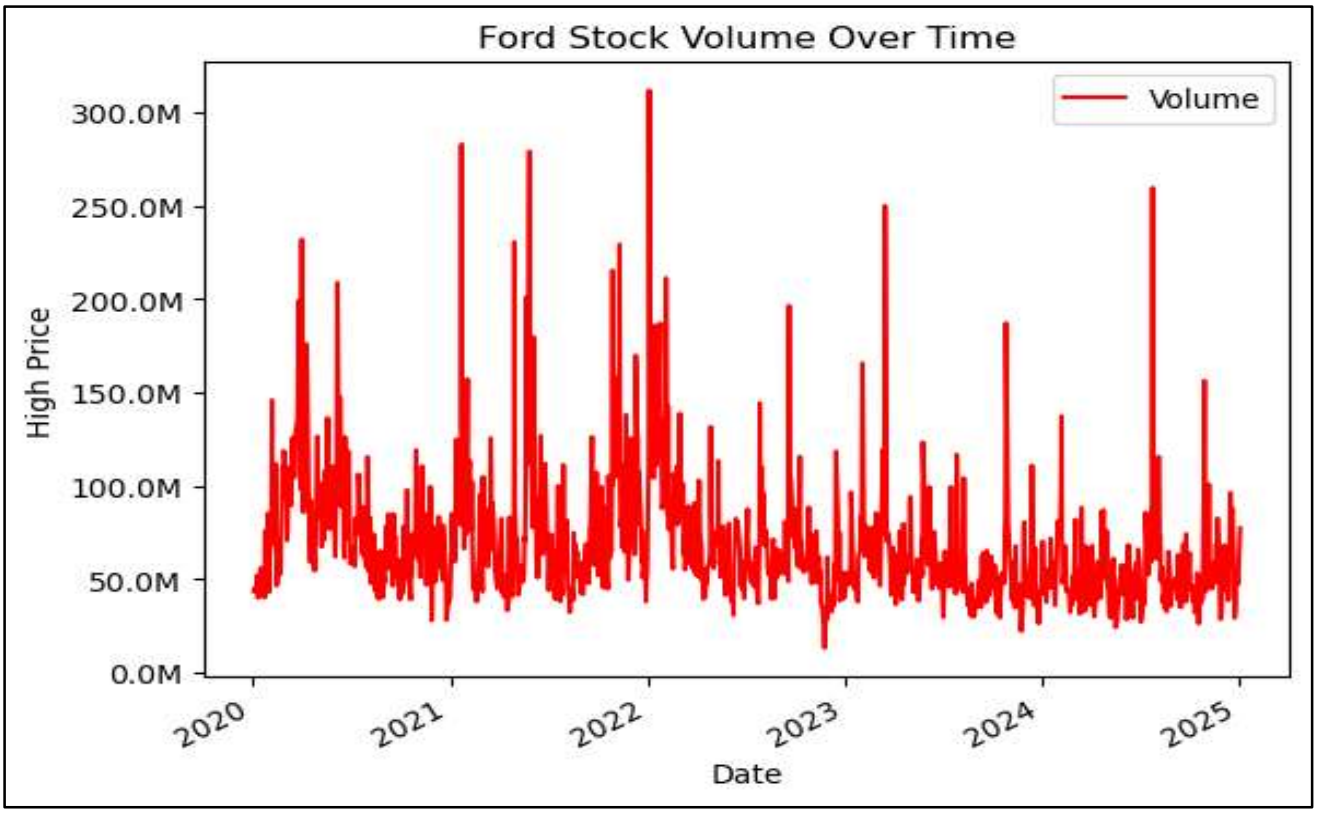
Volume (Number of Stocks) for Tesla



Volume (Number of Stocks) for Volkswagen



Volume (Number of Stocks) for Ford



Missing value analysis and outlier analysis

Four each of our chosen company data, we were able to have full data as shown by the image below

Missing Values for Tesla

```
Missing values per column:
 Date        0
Price        0
Open         0
High         0
Low          0
Vol.         0
Change %     0
dtype: int64
```

Missing Values for Ford

```
Missing Values in Dataset:

Date              0
Close             0
Open              0
High              0
Low               0
Volume            0
Percent Change    0
```

Missing Values for Volkswagen

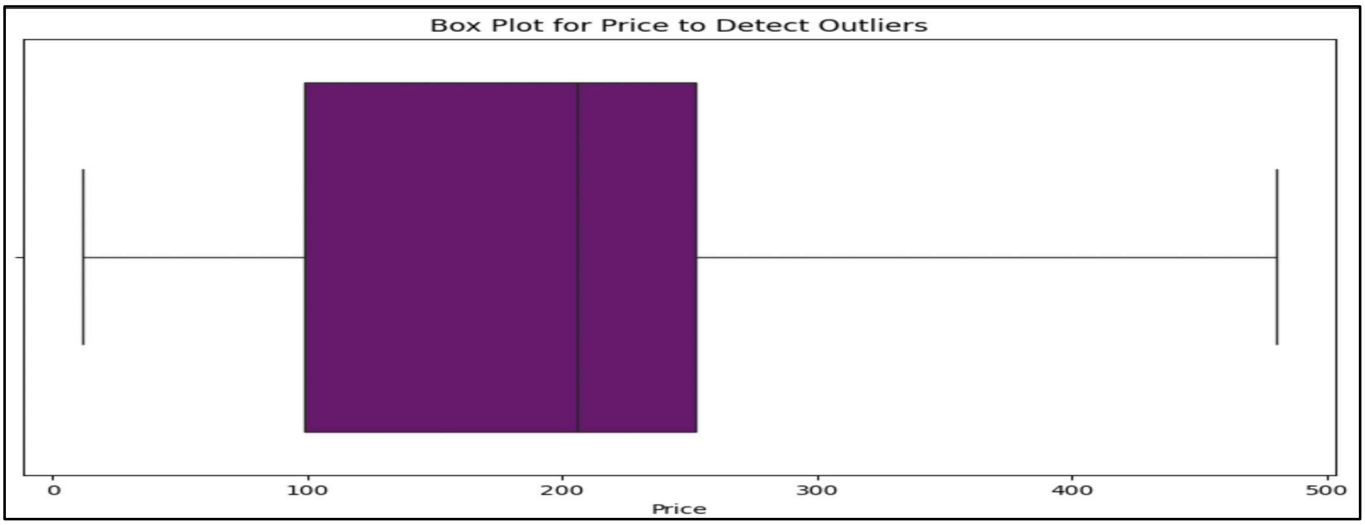
```
Missing Values Per Column:
Price        0
Open         0
High         0
Low          0
Vol.         0
Change %     0
dtype: int64
```

Missing Values for Toyota

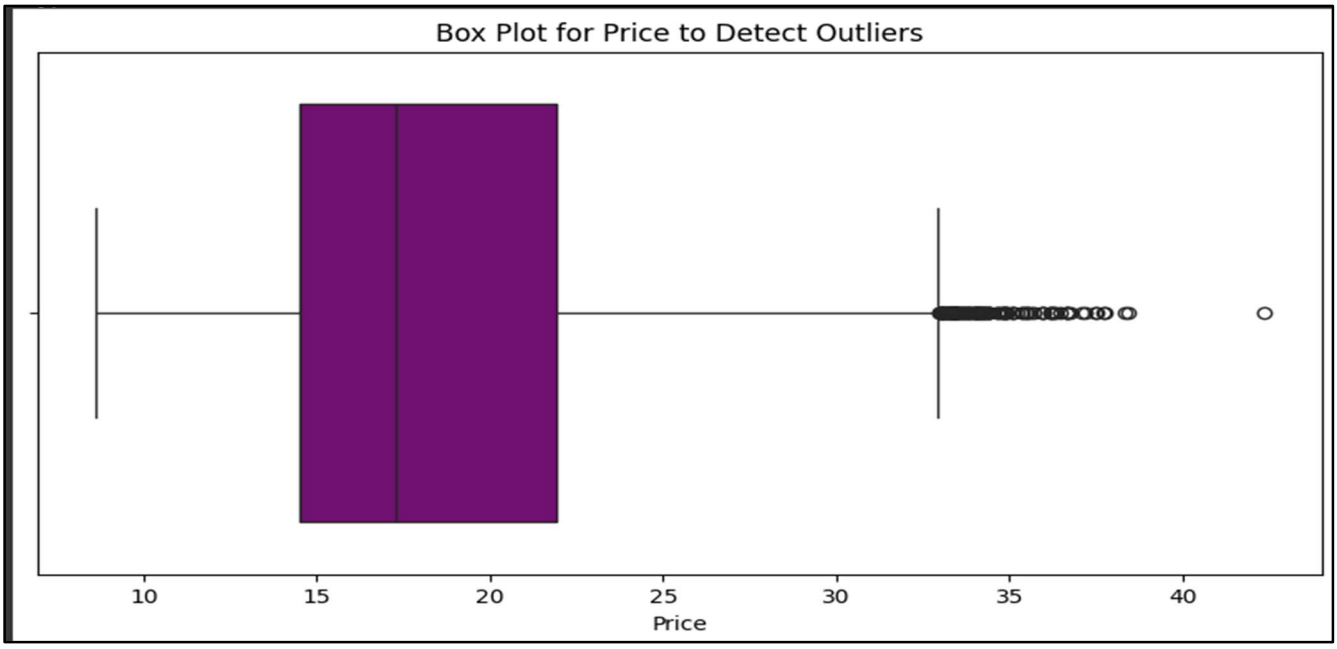
```
Date          0
Price         0
Open          0
High          0
Low           0
Vol.          0
Change %      0
```

For outlier analysis, we used a box and whisker plot to show stock price outlier for each of our companies

Stock Price Outliers for Tesla

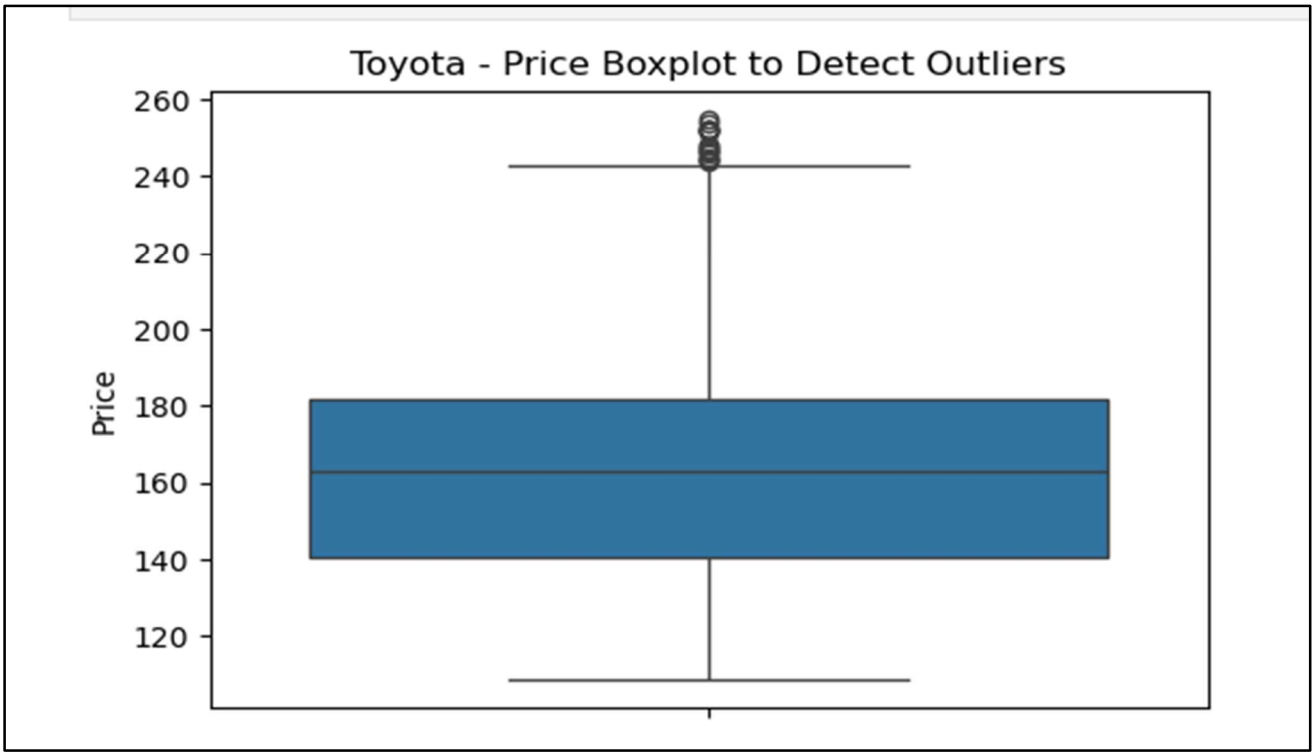


Stock Price Outliers for Volkswagen



Stock Price Outliers for Ford

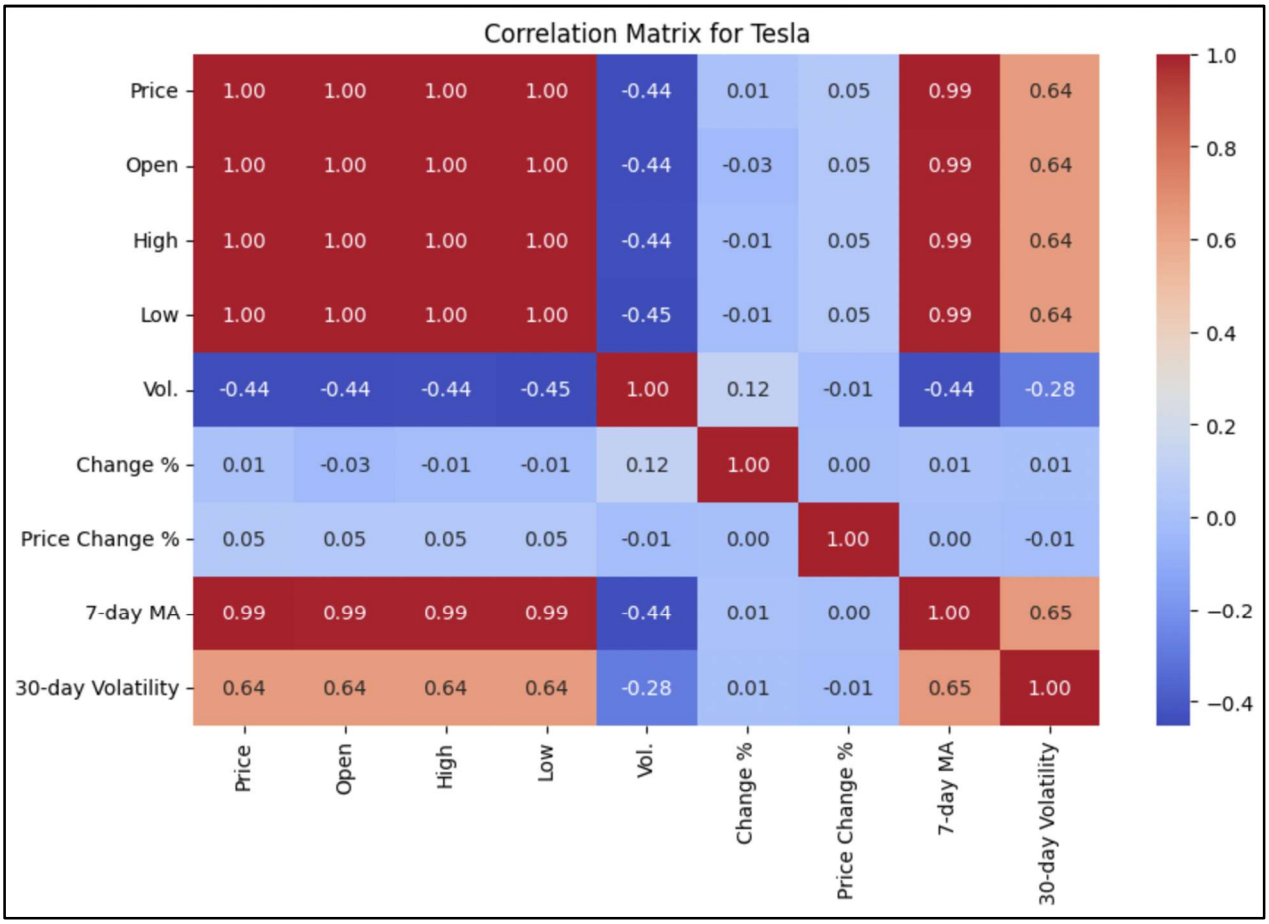
Stock Price Outliers for Toyota



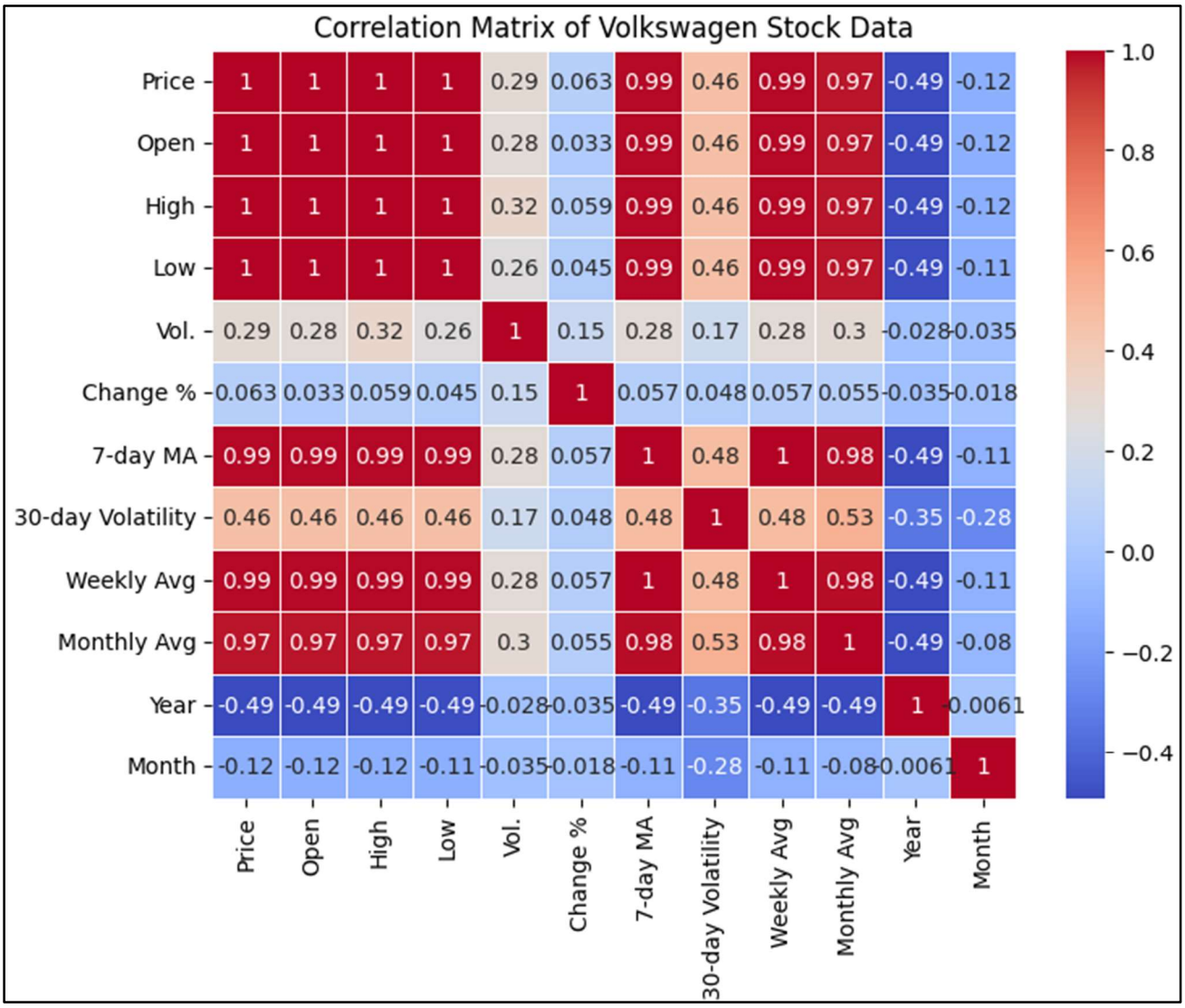
Feature engineering and analysis

For our feature engineering, we used a correlation matrix to show the relationships between each variable used in our analysis for each car company

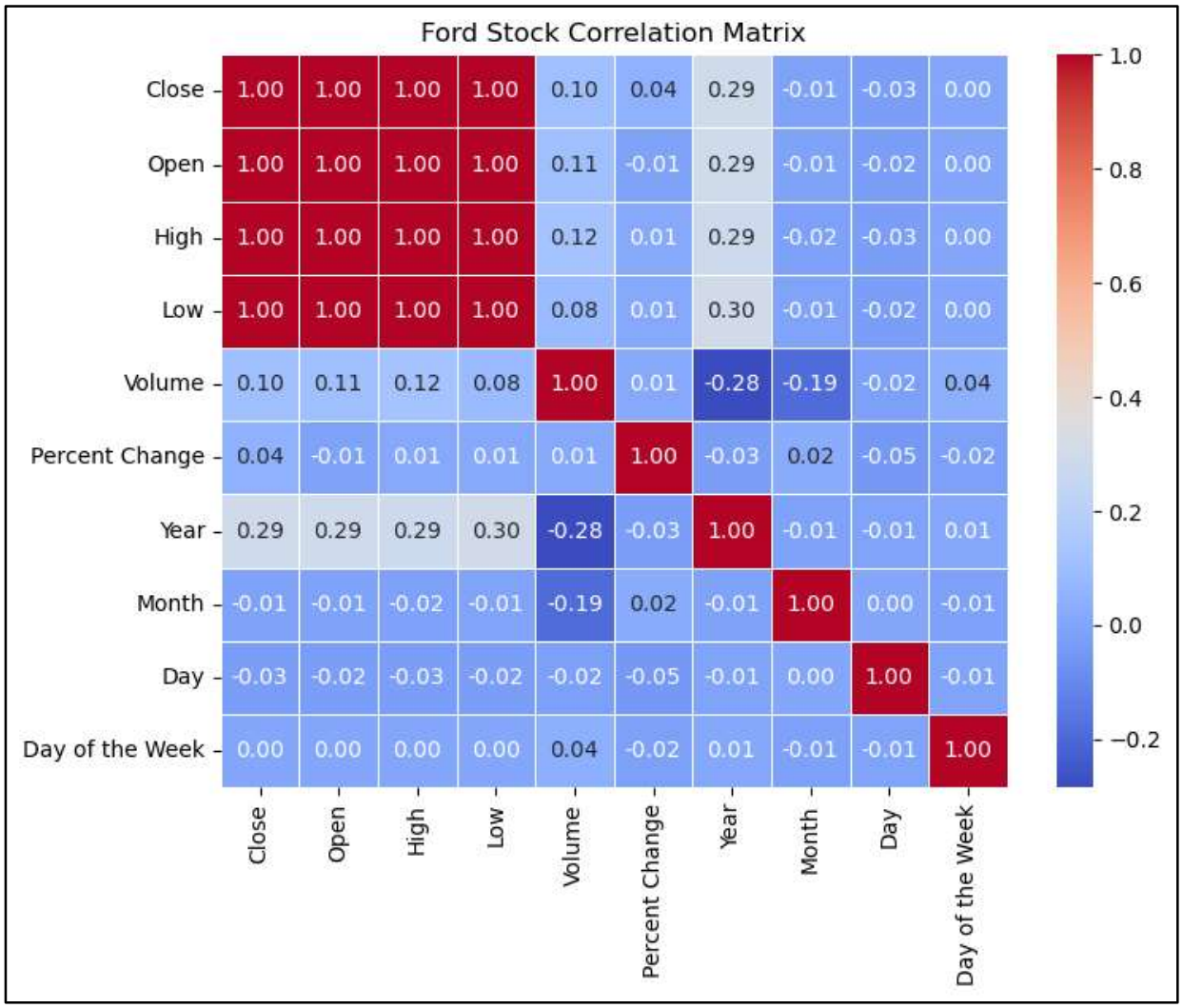
Correlation Matrix for Tesla Stock



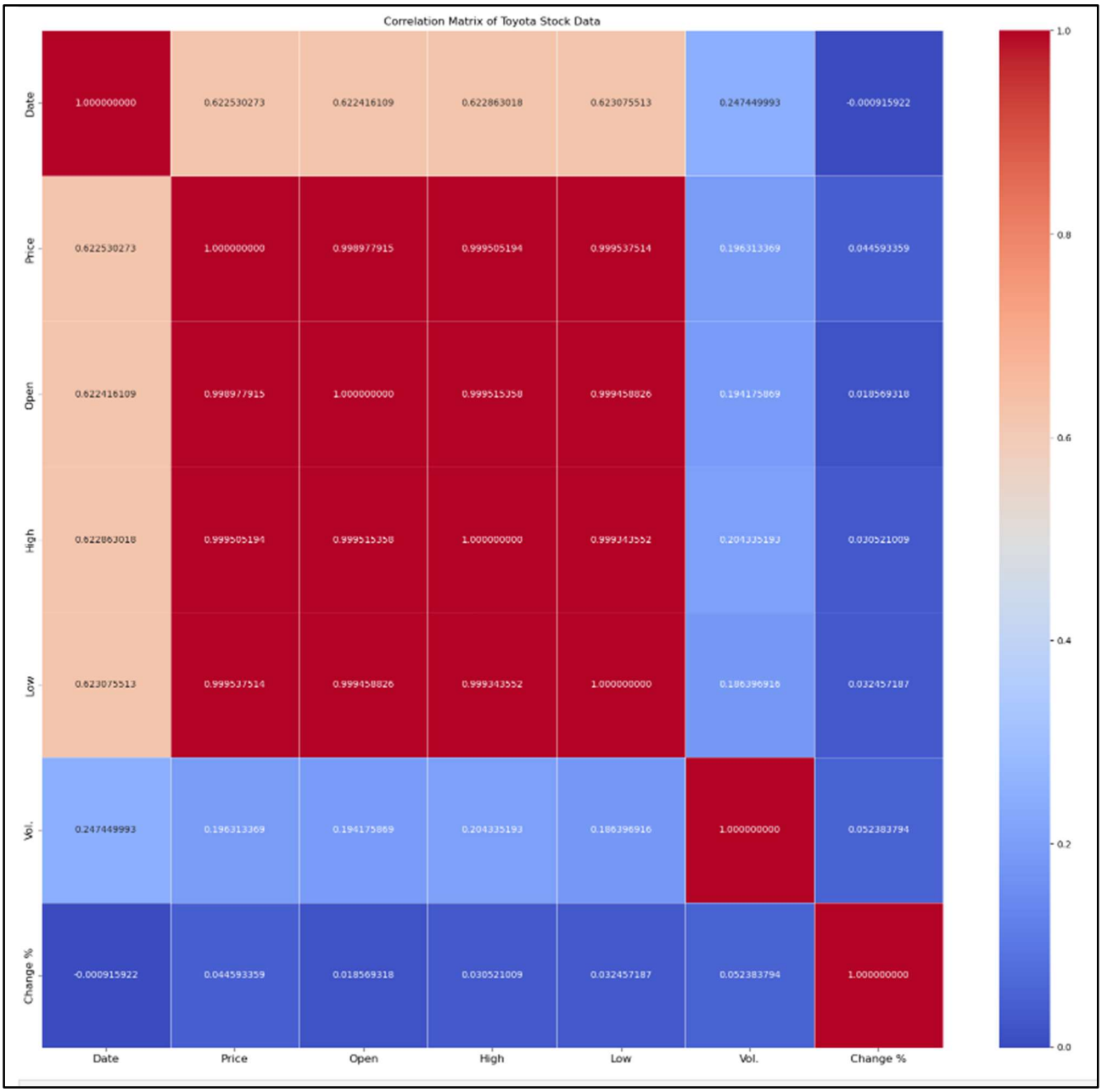
Correlation Matrix for Volkswagen Stock



Correlation Matrix for Ford Stock



Correlation Matrix for Toyota Stock



Explanation:

The correlation matrix helps identify relationships between numerical variables in the dataset. We first converted the 'Date' column to datetime format to ensure it's not included in calculations.

Then, we selected only numerical columns (Price, Open, High, Low, Volume, Change %) for correlation analysis. The correlation values range from -1 to 1:

- Positive values indicate a direct relationship (as one variable increases, the other also increases).
- Negative values indicate an inverse relationship (as one increases, the other decreases).
- Values close to 0 suggest no significant relationship.

This analysis helps in feature selection for predictive modeling, as highly correlated variables might be redundant.

Appendix

The following are examples of the code used for data exploration and graph creation:

Data Description:

```
stat_df = corr_df[['Date', 'Close', 'Open', 'High', 'Low', 'Volume', 'Percent Change']]
stat_df.describe()
```

Price Distribution:

```
ax = hist_df['Close'].hist(bins=20)
ax.set_title("Distribution of Closing Prices - Ford", fontsize=14)
ax.set_xlabel("Closing Price", fontsize=12)
ax.set_ylabel("Frequency", fontsize=12)
ax.grid(False)
```

Closing Price Line Graph:

```
hist_df.plot(x='Date', y='Close', kind='line',
        title='Ford Stock CLosing Price',
        xlabel='Date',
        ylabel='Close Price',
        color='blue')
plt.grid(False)
plt.show()
```

Average Ford Stock Line Graph:

```
month_df.set_index('Date', inplace=True)
month_df = month_df['Close'].resample('M').mean()
print(month_df.head())
```

```
Date
2020-01-31    9.130476
2020-02-29    7.973684
2020-03-31    5.522273
2020-04-30    4.921905
2020-05-31    5.257500
Freq: M, Name: Close, dtype: float64
```

```
month_df.plot(kind='line',
        title='Average Ford Stock Price per Month',
        xlabel='Date',
        ylabel='Average Close Price (Month)',
        color='green')
plt.grid(False)
plt.show()
```

Stock Volume Distribution:

```
ax = hist_df['Volume'].hist(bins=20)
ax = hist_df['Close'].hist(bins=20)
ax.set_title("Distribution of Stock Volume", fontsize=14)
ax.set_xlabel("Volume", fontsize=12)
ax.set_ylabel("Frequency", fontsize=12)
ax.grid(False)
ax.xaxis.set_major_formatter(mtick.FuncFormatter(lambda x, _: f'{x/1e6:.1f}M'))
```

Identifying Missing Values:

```
print("Missing Values in Dataset:\n")
print(stat_df.isnull().sum().to_string())
```

Box and Whisker Plot:

```
plt.figure(figsize=(8,4))
sns.boxplot(x=corr_df['Close'], color="lightblue", flierprops={'marker': 'o', 'markerfacecolor': 'red', 'markersize': 6})

plt.title("Distribution of Closing Prices", fontsize=14)
plt.xlabel("Closing Price", fontsize=12)

plt.show()
```

Correlation Matrix:

```
corr_df = hist_df.copy()
corr_df['Percent Change'] = corr_df['Close'].pct_change() * 100
corr_df['Percent Change'].fillna(0, inplace=True)
corr_df['Year'] = hist_df['Date'].dt.year
corr_df['Month'] = hist_df['Date'].dt.strftime('%B')
corr_df['Day'] = hist_df['Date'].dt.strftime('%d')
corr_df['Day of the Week'] = hist_df['Date'].dt.day_name()
```

```
corr_df['Day'] = corr_df['Day'].astype('int64')
corr_df['Month'] = pd.to_datetime(corr_df['Month'], format='%B').dt.month
days_mapping = {
    'Monday': 0, 'Tuesday': 1, 'Wednesday': 2, 'Thursday': 3,
    'Friday': 4, 'Saturday': 5, 'Sunday': 6
}
corr_df['Day of the Week'] = corr_df['Day of the Week'].map(days_mapping)
```

```
corr_matrix = corr_df.corr()
plt.figure(figsize=(8,6))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt=".2f", linewidths=0.5)
plt.title("Ford Stock Correlation Matrix")
plt.show()
```

Table of Contributions

The table below identifies contributors to various sections of this document.

	Section	Writing	Editing
1	**Analysis the basic metrics of variables**	**Uditi, Robert**	**Ahmad, Steven**
2	**Non-graphical and graphical univariate analysis**	**All Members**	**All Members**
3	**Missing value and outlier analysis**	**All Members**	**All Members**
4	**Feature engineering and analysis**	**All Members**	**All Members**
5	**Appendix**	**All Members**	**All Members**