World Stock Prices Analysis

Indhu Sri Krishnaraj
MS in Artificial Intelligence

1. Introduction

Empirical Analysis is the process of testing a hypothesis using empirical evidence which can be any information or data. The dataset that will be used in this project for analysis is a World Stock Prices dataset. This report contains the analysis of the performed empirical analysis on the variables using various approaches. In this project, various tests are performed using population data and sample extracted from that data. The tests that will be discussed here are Mean Test, Proportion Test, Two Sample Test, Correlation Test, Regression Models and Multi-Regression Models. All of these operations are performed in this final phase of the project.

2. Dataset

To accomplish this, we have selected a world stock dataset. This dataset contains 10 independent variables and one dependent variable. The variables are described below:

Open: The opening price of the stock, indicating initial market sentiment and investor confidence.

High: The highest price reached by the stock during the trading day, reflecting peak market demand.

Low: The lowest price reached by the stock during the trading day, indicating potential support levels.

Close: The closing price of the stock, reflecting market sentiment at the end of the trading day.

Volume: The trading volume, representing the total number of shares traded on that date, providing insights into market liquidity and investor activity.

Dividends: Information about dividends paid on that date, enabling dividend yield analysis and income estimation.

Stock Splits: Details about any stock splits that occurred, affecting the stock's price and market capitalization.

Brand_Name: The name of the brand or company associated with the stock, facilitating company-specific analysis and comparison.

Ticker: The ticker symbol for the stock, allowing for easy identification and tracking.

Industry_Tag: The industry category or sector to which the brand belongs, aiding in sector-specific analysis and comparison.

Country: The country where the brand is headquartered or primarily operates, providing geographical context and enabling regional analysis.

The sample dataset is shown in below figures:

Open	High	Low	Close	Volume
4.840000	4.910000	4.630000	4.670000	7441900.0
397.049988	397.989990	386.119995	386.299988	3866600.0
564.349976	569.219971	562.659973	563.830017	1311500.0
138.550003	139.369995	135.199997	135.289993	46263700.0
179.259995	179.699997	175.399994	175.490005	58436200.0

Dividends	Stock Splits	Brand_Name	Ticker	Industry_Tag	Country
0.0	0.0	37.0	46.0	9.0	6.0
0.0	0.0	32.0	37.0	6.0	6.0
0.0	0.0	15.0	12.0	20.0	6.0
0.0	0.0	4.0	6.0	5.0	6.0
0.0	0.0	8.0	0.0	22.0	6.0

3. One Sample Test

A sample of 100 rows is randomly selected from the population data.

3.1 Test for Mean

To perform test for mean, a few variables are selected. These are Close, Open and Volume, each of which will be tested below. The test statistic for mean test is given below by equation (1). z denotes the test statistic; x denotes the sample mean while X denotes the guess mean. S is the population Standard Deviation and n is the sample size (100 in this case).

$$z = \frac{x - X}{\frac{S}{\sqrt{n}}} \tag{1}$$

Close

```
Sample Mean: 71.4655403316021
Population Standard Deviation: 117.19096308450717
Guess Mean: 70
CI = [-1.96, 1.96]
z = 0.12505574602585087
```

Claim

H0: The population mean for Close variable is 70

H1: The population mean for Close variable is not 70

Hypothesis Testing

The test statistic z = 0.12

CI

The confidence interval CI = [-1.96, 1.96]

Interpretation

- -1.96 <= 0.12 <= 1.96; Hence, the test statistic z is in the range of CI.
- Therefore, null hypothesis is accepted.
- The population mean for Close variable is 70.

Open

```
Sample Mean: 71.37208537431603
Population Standard Deviation: 117.20315835421648
Guess Mean: 70
CI = [-1.96, 1.96]
z = 0.11706897609100712
```

Claim

H0: The population mean for Open variable is 70

H1: The population mean for Open variable is not 70

Hypothesis Testing

The test statistic z = 0.11

CI

The confidence interval CI = [-1.96, 1.96]

Interpretation

- -1.96 <= 0.11 <= 1.96; Hence, the test statistic z is in the range of CI.
- Therefore, null hypothesis is accepted.
- The population mean for Open variable is 70.

Volume

```
Sample Mean: 20519497.93
Population Standard Deviation: 88643605.16094734
Guess Mean: 20000000
CI = [-1.96, 1.96]
z = 0.058605234867959634
```

Claim

H0: The population mean for Volume variable is 20000000

H1: The population mean for Volume variable is not 20000000

Hypothesis Testing

The test statistic z = 0.05

CI

The confidence interval CI = [-1.96, 1.96]

Interpretation

- -1.96 <= 0.05 <= 1.96; Hence, the test statistic z is in the range of CI.
- Therefore, null hypothesis is accepted.
- The population mean for Volume variable is 20000000.

3.2 Test for Proportion

To perform test for proportion, the same variables are selected. These are Close, Open and Volume, each of which will be tested below. The test statistic for proportion test is given below by equation (2). z denotes the test statistic; p denotes the sample proportion while P denotes the population proportion. n is the sample size (100 in this case).

$$z = \frac{p - P}{\sqrt{\frac{P(1-P)}{n}}} \tag{2}$$

Close

```
Sample proportion: 0.27
Population proportion: 0.2751248422715753
CI = [-1.96, inf]
z = -0.1147582698873859
```

Claim

H0: The proportion of stocks having Close value more than 65 is 0.27

H1: The proportion of stocks having Close value more than 65 is less than 0.27

Hypothesis Testing

The test statistic z = -0.11

CI

The confidence interval CI = [-1.96, inf]

Interpretation

- -1.96 <= -0.11 <= inf; Hence, the test statistic z is in the range of CI.
- Therefore, null hypothesis is accepted.
- The proportion of stocks having Close value more than 65 is 0.27.

Open

```
Sample proportion: 0.26
Population proportion: 0.27508552187107915
CI = [-1.96, inf]
z = -0.33781823245707715
```

Claim

H0: The proportion of stocks having Open value more than 65 is 0.27

H1: The proportion of stocks having Open value more than 65 is less than 0.27

Hypothesis Testing

The test statistic z = -0.33

CI

The confidence interval CI = [-1.96, inf]

Interpretation

- -1.96 <= -0.33 <= inf; Hence, the test statistic z is in the range of CI.
- Therefore, null hypothesis is accepted.
- The proportion of stocks having Open value more than 65 is 0.27

Volume

```
Sample proportion: 0.21
Population proportion: 0.1593405611378609
CI = [-1.96, inf]
z = 1.384162748474168
```

Claim

H0: The proportion of stocks having Volume more than 23491025 is 0.16

H1: The proportion of stocks having Volume more than 23491025 is less than 0.16

Hypothesis Testing

The test statistic z = 1.38

CI

The confidence interval CI = [-1.96, inf]

Interpretation

- -1.96 <= 1.38 <= inf; Hence, the test statistic z is in the range of CI.
- Therefore, null hypothesis is accepted.
- The proportion of stocks having Volume more than 23491025 is 0.16

4. Two Sample Test

Another sample of 200 rows is randomly selected from the population data.

4.1 Two Sample Test for Mean

To perform two sample test for mean, a few variables are selected. These are High, Low and Dividends, each of which will be tested below. The test statistic for two sample mean test is given below by equation (3). z denotes the test statistic; x_1 denotes the mean of first sample and x_2 denotes the mean of second sample. s_1 denotes the standard deviation of the first sample and s_2 denotes the standard deviation of the second sample. n_1 is the sample size of first sample (100 in this case) and n_2 is the sample size of second sample (200 in this case).

$$z = \frac{x_1 - x_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{3}$$

High

```
First population mean: 72.03416178851722
Second population mean: 57.07661898186036
First population standard deviation: 162.7423622793537
Second population standard deviation: 99.54332564296732
CI = [-1.96, 1.96]
z = 0.8435723709840648
```

Claim

H0: The mean of High is equal for both samples

H1: The mean of High is not equal for both samples

Hypothesis Testing

The test statistic z = 0.84

CI

The confidence interval CI = [-1.96, 1.96]

Interpretation

- -1.96 <= 0.84 <= 1.96; Hence, the test statistic z is in the range of CI.
- Therefore, null hypothesis is accepted.
- The mean of High is equal for both populations

Low

```
First population mean: 70.63835885211907
Second population mean: 55.587255360525326
First population standard deviation: 159.55236796897083
Second population standard deviation: 95.78267311262937
CI = [-1.96, 1.96]
z = 0.8683376059845432
```

Claim

H0: The mean of Low is equal for both samples

H1: The mean of Low is not equal for both samples

Hypothesis Testing

The test statistic z = 0.86

CI

The confidence interval CI = [-1.96, 1.96]

Interpretation

- -1.96 <= 0.86 <= 1.96; Hence, the test statistic z is in the range of CI.
- Therefore, null hypothesis is accepted.
- The mean of Low is equal for both populations

Dividends

```
First population mean: 0.0
Second population mean: 0.0011375
First population standard deviation: 0.0
Second population standard deviation: 0.01225718305452606
CI = [-1.96, 1.96]
z = -1.312428736719717
```

Claim

H0: The mean of Dividends is equal for both samples

H1: The mean of Dividends is not equal for both samples

Hypothesis Testing

The test statistic z = -1.31

CI

The confidence interval CI = [-1.96, 1.96]

Interpretation

- -1.96 <= -1.31 <= 1.96; Hence, the test statistic z is in the range of CI.
- Therefore, null hypothesis is accepted.
- The mean of Dividends is equal for both populations

4.2 Two Sample Proportion Test

To perform two sample test for proportion, the same variables are selected. These are High, Low and Dividends, each of which will be tested below. The test statistic for two sample mean test is given below by equation (3). z denotes the test statistic; p_1 denotes the proportion of first sample and p_2 denotes the proportion of second sample. P denotes the population Proportion. n_1 is the sample size of first sample (100 in this case) and n_2 is the sample size of second sample (200 in this case).

$$z = \frac{p_1 - p_2}{\sqrt{P(1-P)(\frac{1}{n_1} + \frac{1}{n_2})}} \quad (4)$$

High

```
First population proportion: 0.21
Second population proportion: 0.225
Standard population proportion: 0.25406340593309096
CI = [-1.96, 1.96]
z = -0.28133490298439556
```

Claim

H0: The proportion of stocks having High value more than 72 is equal for both samples

H1: The proportion of stocks having High value more than 72 is not equal for both samples

Hypothesis Testing

The test statistic z = -0.28

CI

The confidence interval CI = [-1.96, 1.96]

Interpretation

- -1.96 <= -0.28 <= 1.96; Hence, the test statistic z is in the range of CI.
- Therefore, null hypothesis is accepted.
- The proportion of stocks having High value more than 72 is equal for both samples

Low

```
First population proportion: 0.31
Second population proportion: 0.24
Standard population proportion: 0.29417378902102925
CI = [-1.96, 1.96]
z = 1.2543003179409375
```

Claim

H0: The proportion of stocks having Low value more than 60 is equal for both samples

H1: The proportion of stocks having Low value more than 60 is not equal for both samples

Hypothesis Testing

The test statistic z = 1.25

CI

The confidence interval CI = [-1.96, 1.96]

Interpretation

- -1.96 <= 1.25 <= 1.96; Hence, the test statistic z is in the range of CI.
- Therefore, null hypothesis is accepted.
- The proportion of stocks having Low value more than 60 is equal for both samples

Dividends

```
First population proportion: 0.0
Second population proportion: 0.01
Standard population proportion: 0.008897134257720204
CI = [-1.96, 1.96]
z = -0.869500692985797
```

Claim

H0: The proportion of stocks having Dividends value more than 0.0005 is equal for both samples

H1: The proportion of stocks having Dividends value more than 0.0005 is not equal for both samples

Hypothesis Testing

The test statistic z = -0.86

CI

The confidence interval CI = [-1.96, 1.96]

Interpretation

- -1.96 <= -0.86 <= 1.96; Hence, the test statistic z is in the range of CI.
- Therefore, null hypothesis is accepted.
- The proportion of stocks having Dividends value more than 0.0005 is equal for both samples

5. Correlation Test

To perform correlation test, six variables are selected. These are Country, Ticker, Industry Tag, Brand Name, High and Volume, each of which will be tested below. The test statistic for correlation test is given below by equation (5). z denotes the test statistic; c denotes the sample correlation and n is the sample size (100 in this case).

$$z = \frac{c}{\sqrt{\frac{(1-c^2)}{n-2}}} \tag{5}$$

Country

```
Sample correlation coefficient: 0.11542220809843841
CI = [-1.661, 1.661]
z = 1.1503096331909108
```

Claim

H0: Country and Close Values are not correlated

H1: There is some degree of correlation between Country and Close Values

Hypothesis Testing

The test statistic z = 1.15

CI

The confidence interval CI = [-1.661, 1.661]

Interpretation

- -1.661 <= 1.15 <= 1.661; Hence, the test statistic z is in the range of CI.
- Therefore, null hypothesis is accepted.
- Country and Close Values are not correlated

Ticker

```
Sample correlation coefficient: -0.19195876618931568
CI = [-1.661, 1.661]
z = -1.936304301122827
```

Claim

H0: Ticker and Close Values are not correlated

H1: There is some degree of correlation between Ticker and Close Values

Hypothesis Testing

The test statistic z = -1.93

CI

The confidence interval CI = [-1.661, 1.661]

Interpretation

- -1.93 < -1.661; Hence, the test statistic z is not in the range of CI.
- Therefore, null hypothesis is rejected.
- There is some degree of correlation between Ticker and Close Values

Industry Tag

```
Sample correlation coefficient: 0.07915896745362107
CI = [-1.661, 1.661]
z = 0.7861005768354207
```

Claim

H0: Industry Tag and Close Values are not correlated

H1: There is some degree of correlation between Industry Tag and Close Values

Hypothesis Testing

The test statistic z = 0.78

CI

The confidence interval CI = [-1.661, 1.661]

Interpretation

- -1.661 <= 0.78 <= 1.661; Hence, the test statistic z is in the range of CI.
- Therefore, null hypothesis is accepted.
- Industry Tag and Close Values are not correlated.

Brand Name

```
Sample correlation coefficient: -0.17773624483250583
CI = (-1.661, 1.661)
z = -1.7879667746076198
```

Claim

H0: Brand Name and Close Values are not correlated

H1: There is some degree of correlation between Brand Name and Close Values

Hypothesis Testing

The test statistic z = -1.78

CI

The confidence interval CI = [-1.661, 1.661]

Interpretation

- -1.78 < -1.661; Hence, the test statistic z is not in the range of CI.
- Therefore, null hypothesis is rejected.
- There is some degree of correlation between Brand Name and Close Values.

High

```
Sample correlation coefficient: 0.999993778667428
CI = [-1.661, 1.661]
z = 2806.4305843225975
```

Claim

H0: High and Close Values are not correlated

H1: There is some degree of correlation between High and Close Values

Hypothesis Testing

The test statistic z = 2806

CI

The confidence interval CI = [-1.661, 1.661]

Interpretation

- 2806 > 1.661; Hence, the test statistic z is not in the range of CI.
- Therefore, null hypothesis is rejected.
- There is some degree of correlation between High and Close Values

Volume

```
Sample correlation coefficient: -0.1357136491468841
CI = [-1.661, 1.661]
z = -1.35604254727644
```

Claim

H0: Volume and Close Values are not correlated

H1: There is some degree of correlation between Volume and Close Values

Hypothesis Testing

The test statistic z = -1.35

CI

The confidence interval CI = [-1.661, 1.661]

Interpretation

- -1.661 <= -1.35 <= 1.661; Hence, the test statistic z is in the range of CI.
- Therefore, null hypothesis is accepted.
- Volume and Close Values are not correlated

6. Regression

A regression model is built using one independent and one dependent variable. We build this model for all independent variables with the dependent variable 'Close'. The regression equation is given by equation (6). y denotes the value of the dependent variable, m is a slope, x denotes the value of independent variable and c is the intercept.

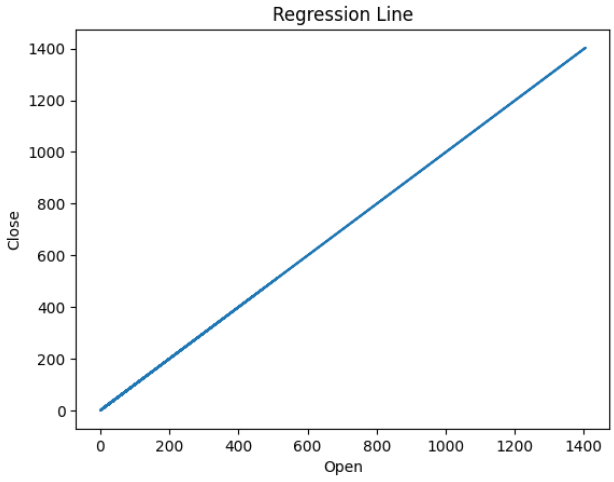
$$y = mx + c \tag{6}$$

Open

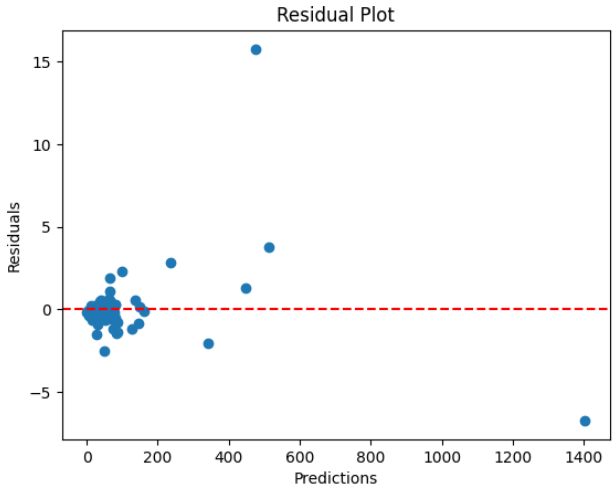
Regression Equation

```
y = 0.9989932755243812*x +
0.16530698250834064
```

Regression Line



Residual Plot



Interpretation

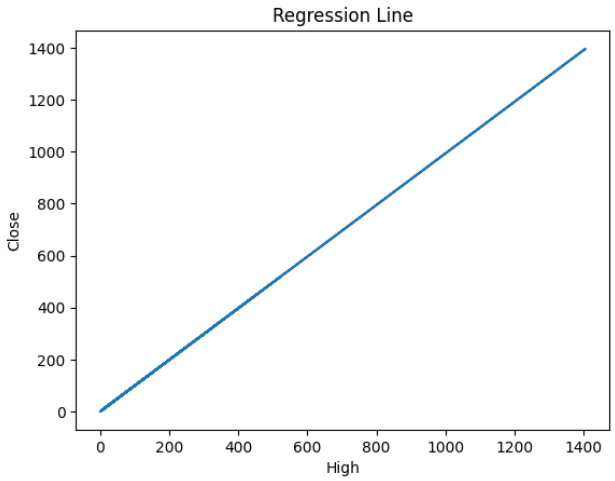
The points are scattered randomly in the residual plot, hence it is a good model to predict the target variable.

High

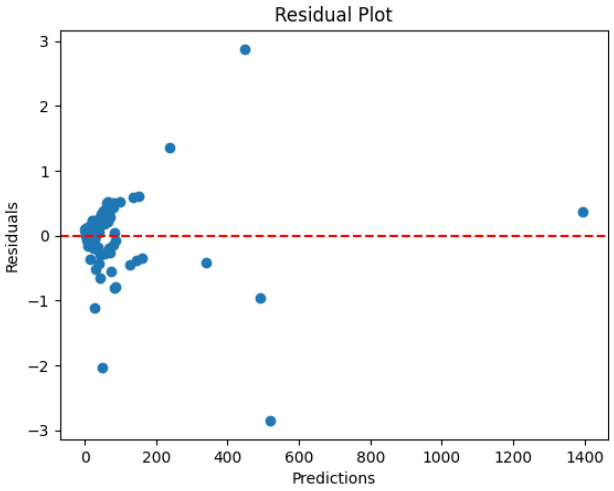
Regression Equation

```
y = 0.9934888973370033*x + -
0.09959963426717877
```

Regression Line



Residual Plot



Interpretation

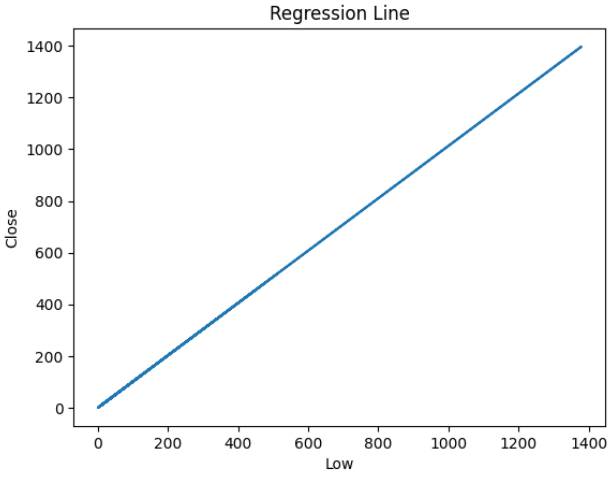
The points are scattered randomly in the residual plot, hence it is a good model to predict the target variable.

Low

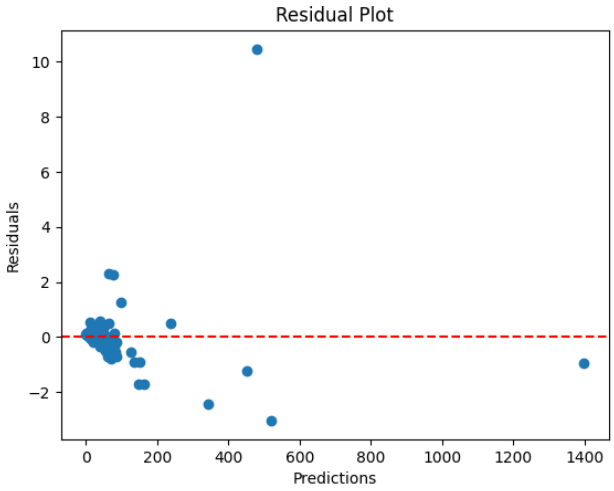
Regression Equation

```
y = 1.0133284335026373*x + -
0.11431718921286915
```

Regression Line



Residual Plot



Interpretation

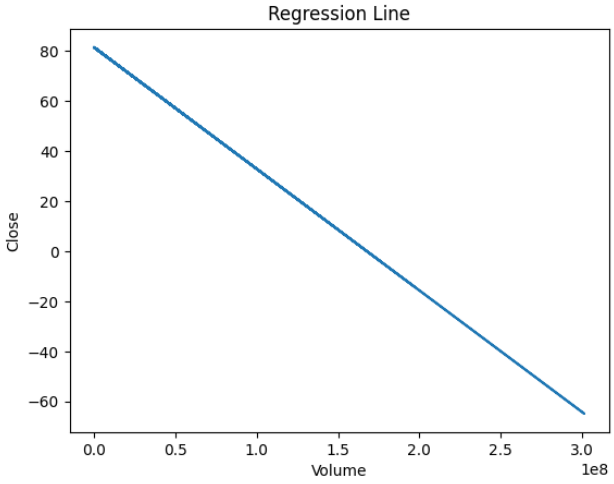
The points are scattered randomly in the residual plot, hence it is a good model to predict the target variable.

Volume

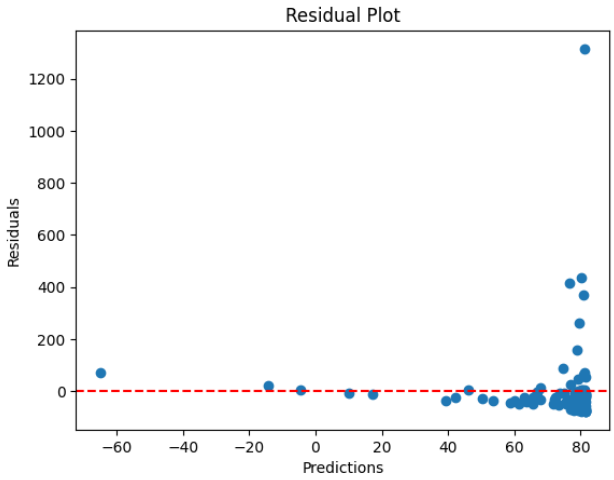
Regression Equation

```
y = -4.854398451838654e-07*x +
81.42652222999195
```

Regression Line



Residual Plot



Interpretation

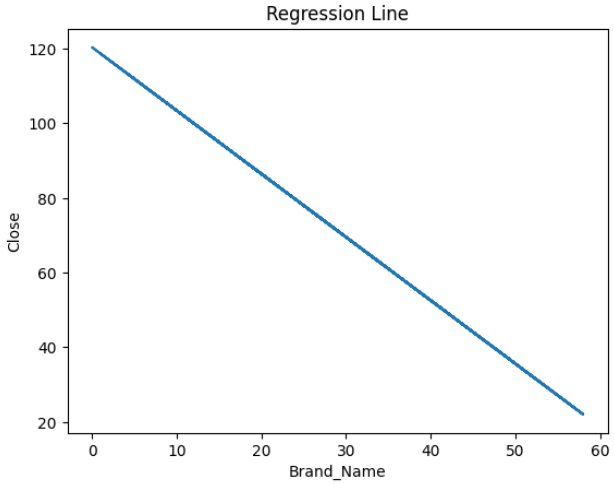
The points are not scattered randomly along the axis in the residual plot. Most of the points are below the axis, hence it is not a good model to predict the target variable.

Brand Name

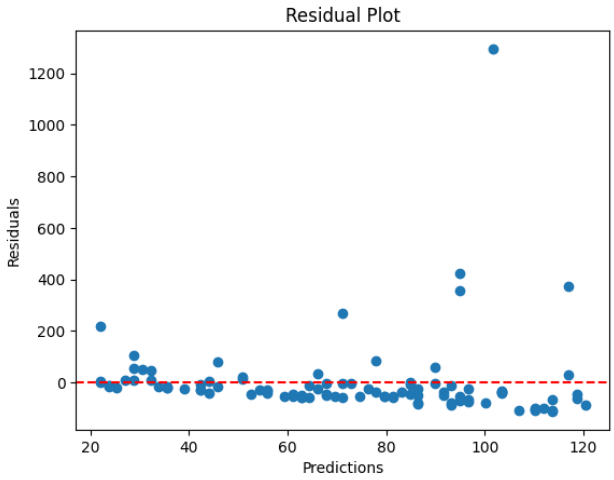
Regression Equation

```
y = -1.6958847237407841*x +
120.39181461152373
```

Regression Line



Residual Plot



Interpretation

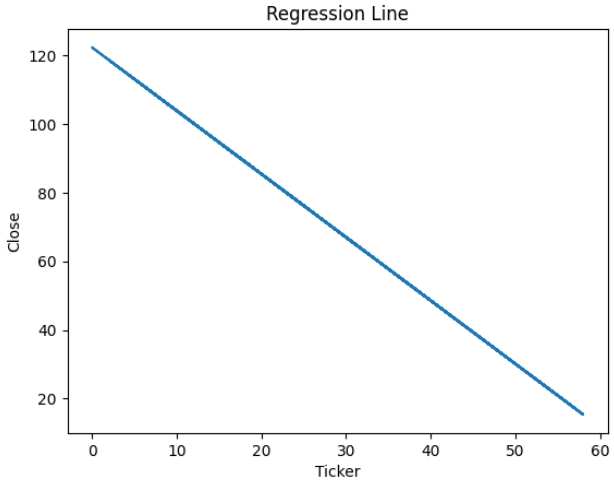
The points are not scattered randomly along the axis in the residual plot. Most of the points are below the axis, hence it is not a good model to predict the target variable.

Ticker

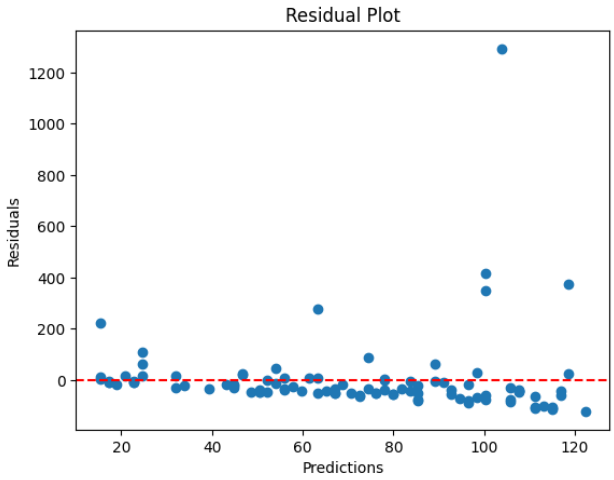
Regression Equation

```
y = -1.8441005722280117*x +
122.32583411365066
```

Regression Line



Residual Plot



Interpretation

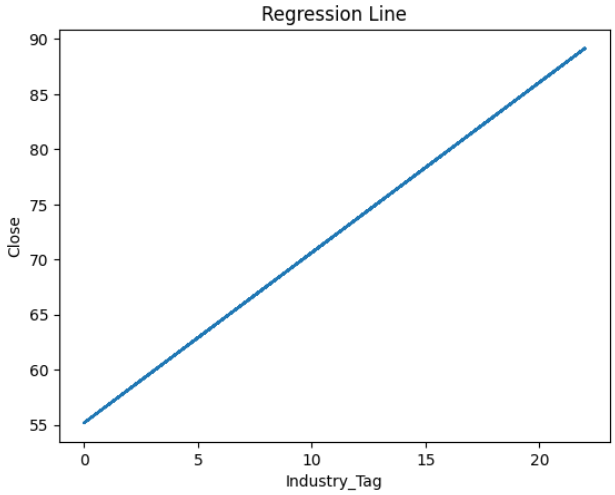
The points are not scattered randomly along the axis in the residual plot. Most of the points are below the axis, hence it is not a good model to predict the target variable.

Industry Tag

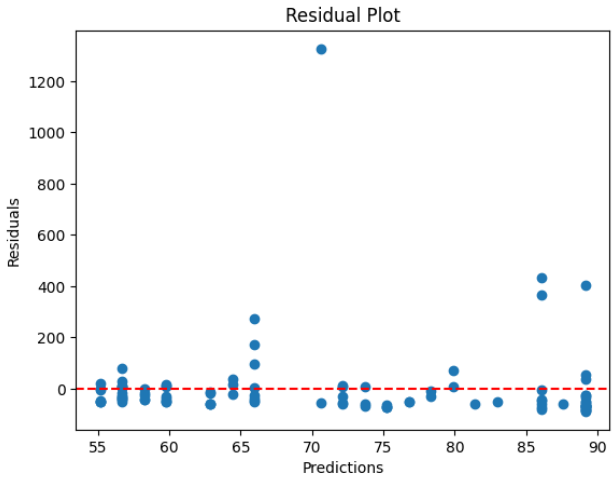
Regression Equation

```
y = 1.5442074681977835*x +
55.17415154211548
```

Regression Line



Residual Plot



Interpretation

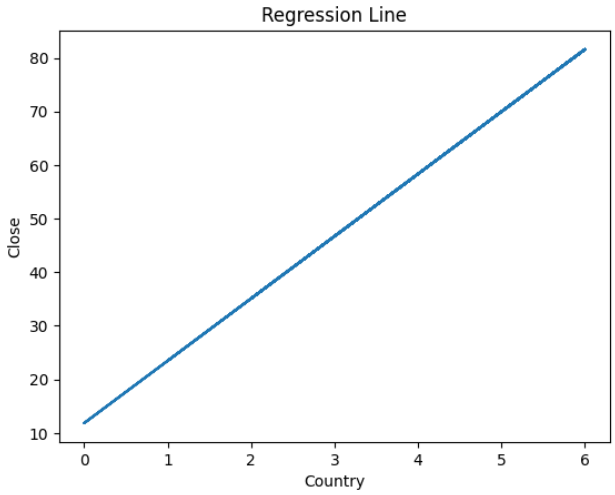
The points are scattered randomly in the residual plot, hence it is a good model to predict the target variable.

Country

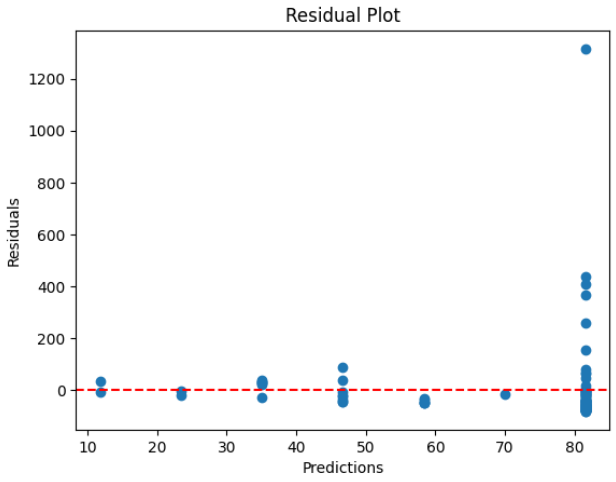
Regression Equation

```
y = 11.620890219410207*x +
11.850373506027744
```

Regression Line



Residual Plot



Interpretation

The points are scattered randomly in the residual plot, hence it is a good model to predict the target variable.

7. Multi-Regression

A multi-regression model is built using more than one independent variables and one dependent variable. We build three models for the independent variables with the dependent variable 'Close'. The multi-regression equation is given by equation (7). y denotes the value of the dependent variable, while x_1, x_2, x_3, .. are the independent variables with a, b, c… as their co-efficient; ε is the intercept of the equation.

$$y = ax_1 + bx_2 + cx_3 + .. + \varepsilon \tag{7}$$

Model 1

```
# Model 1
x = np.array(sample[['Open', 'High',
'Low']])
y = np.array(sample['Close'])

regressor = LinearRegression()
regressor.fit(x, y)

y_pred = regressor.predict(x)

r2 = r2_score(y, y_pred)
adj_r2 = 1-(1-r2)*(n-1)/(n-p-1)

print(f'Regression Equation:
y={regressor.coef_[0]}*a +
{regressor.coef_[1]}*b +
{regressor.coef_[2]}*c +
{regressor.intercept_}')
print(f'Adjusted R2: {adj_r2}')
```

Variables

The independent variables in this model are Open, High and Low as these variables have the highest correlation with the dependent variable (from correlation test and regression).

Regression Equation

```
y=-0.017034891973385768*a +
0.8109676270075653*b +
0.2034552694554999*c + -
0.10776348739618413
```

Adjusted R2

The adjusted R2 value is 0.9999

Model 2

```
# Model 2
x = np.array(sample[['Volume',
'Dividends', 'Ticker']])
y = np.array(sample['Close'])

regressor = LinearRegression()
regressor.fit(x, y)

y_pred = regressor.predict(x)

r2 = r2_score(y, y_pred)
adj_r2 = 1-(1-r2)*(n-1)/(n-p-1)

print(f'Regression Equation:
y={regressor.coef_[0]}*a +
{regressor.coef_[1]}*b +
{regressor.coef_[2]}*c +
{regressor.intercept_}')
print(f'Adjusted R2: {adj_r2}')
```

Variables

The independent variables in this model are Volume, Dividends and Ticker.

Regression Equation

```
y=-5.543879226856478e-07*a + 0.0*b +
-1.983132250495282*c +
137.53608963222712
```

Adjusted R2

The adjusted R2 value is 0.0313

Model 3

```
# Model 3
x = np.array(sample[['Brand_Name',
'Industry_Tag', 'Country']])
y = np.array(sample['Close'])

regressor = LinearRegression()
regressor.fit(x, y)

y_pred = regressor.predict(x)

r2 = r2_score(y, y_pred)
adj_r2 = 1-(1-r2)*(n-1)/(n-p-1)

print(f'Regression Equation:
y={regressor.coef_[0]}*a +
{regressor.coef_[1]}*b +
{regressor.coef_[2]}*c +
{regressor.intercept_}')
print(f'Adjusted R2: {adj_r2}')
```

Variables

The independent variables in this model are Brand Name, Industry Tag and Country.

Regression Equation

```
y=-1.5736730847932796*a +
0.7959105657177795*b +
8.5809888968668176*c +
64.44867931863173
```

Adjusted R2

The adjusted R2 value is 0.0124

Interpretation

Among the three models, the best model is Model 1 with the variables High, Low and Open, because the Adjusted R2 is highest in this model in comparison with Models 2 and 3. A higher Adjusted R2 value explains that all the variables are statistically significant.

8. Milestones

The following milestones of this project have been reached. These are:

1. Dataset gathering.
2. EDA.
3. Hypothesis Testing.
4. Regression Model

Pending milestones:

Project Presentation (April 15): present our project to the class.

9. Group Responsibilities

Uday – Uday was responsible to carry out test for mean and proportion of Close variable, two sample test for High variable, correlation test for Country and Ticker variables, regression model for Open, Volume and Industry Tag variables and Model 1 of multi-regression model.

Indhu – Indhu was responsible to carry out test for mean and proportion of Open variable, two sample test for Low variable, correlation test for Industry Tag and Brand Name variables, regression model for High, Country and Ticker variables and Model 2 of multi-regression model.

Harika – Harika was responsible to carry out test for mean and proportion of Volume variable, two sample test for Dividends variable, correlation test for High and Volume variables, regression model for Low, and Brand Name variables and Model 3 of multi-regression model. It was also found that Dividends and Stock Splits do not contribute to the Close variable.

REFERENCES

[1] https://www.kaggle.com/datasets/nelgiriyewithana/world-stock-prices-daily-updating.

[2] https://www.researchgate.net/publication/356776423_STOCK_MARKET_ANALYSIS

[3] https://jfinswufe.springeropen.com/articles/10.1186/s40854-023-00548-5.

[4] https://www.jstor.org/stable/2351663

[5] https://www.sciencegate.app/keyword/3502137