

Predicting direction of Ethereum using Logistic Regression

A project presented for the partial requirement for the completion of NMST731 module

Stephen Moorcroft 201802156

Palesa Eunice Baisitse 201902507

Ryan Isaacs 202001716

Department of Mathematical Sciences, Sol Plaatje University,

Kimberley South Africa

© Sol Plaatje University, April 2022.

1. **Introduction**

Ethereum is currently the second biggest crypto currency in the world, and because of its popularity it is intriguing to investors. The challenge is therefore trying to predict the exchange rates of Ethereum. Hence in this project we are going to predict the direction of Ethereum.

The data collected contains 3 variables namely, Direction, Price and Volume. We chose to use a logistic regression model because the target variable is binary. Since the dependant variable “Direction” can either take a value of 0 or 1,

The aim of the project will be to implement Logistic regression on this data to classify the binary outcome of Direction, based on its Price and Volume. Also we will study what effects closing price and volume have on the direction of Ethereum.

1. **Methodology**

The data we are using has a response variable that has only two outcomes where the direction of ethereum can either go up or down because the data satisfies all the assumptions of logistic regression we chose this as our model.

The model takes the structure:

Further simplification we see:

Models the probability of success as a function of the predictor variables.

So the Logistic regression model has a random component which states that the distribution of Y should be Binomial (n, π) where π is the probability of success. The systematic component contains all predictor variables that are linear in the parameters ( for our case it β1 is the price and β2 is the volume. The link function of logistic regression is logit or (logit(π))

Assumptions of Logistic Regression model:

* The response variable is a binary or dichotomous variable (yes/no)
* Logistic regression does not require a linear relationship between the dependent and independent variables.
* There are no influential outliers in the continuous predictors
* There is no high multi-collinearity among the predictors

1. **Descriptive analysis**

From our descriptive analysis we are more familiar type of data we have and also the structure of the data. We know our target variable to be Direction and the predictors to be Price and Volume. The data set is well balanced and that will benefit the model.

In our Price column we see that the range is about 1379.33, the variance seen in the boxplot is relatively small compare to that of the volume. We see that on average there is around 39 091 197 trade that are made.

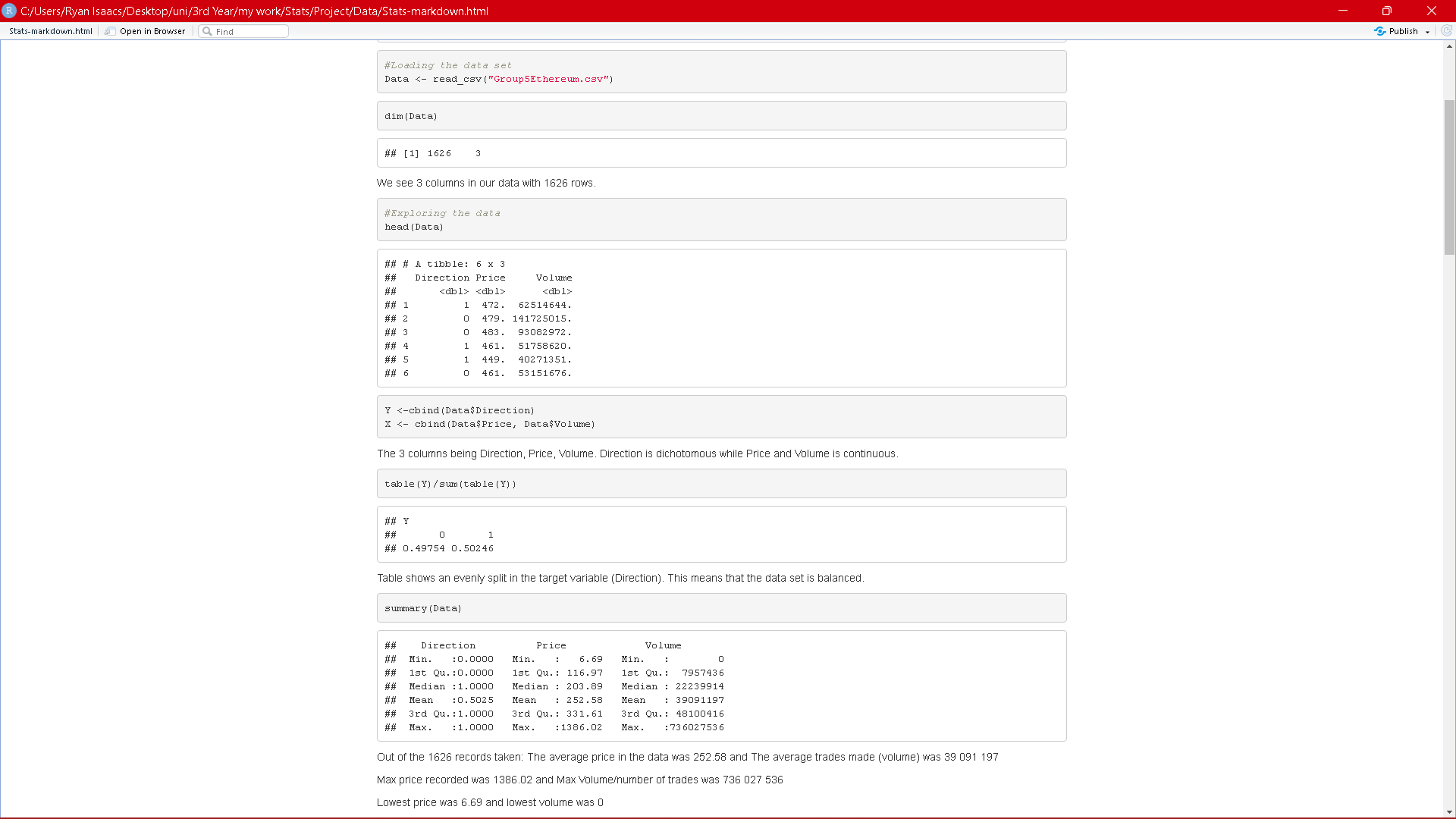


Figure 1: Exploratory data analysis

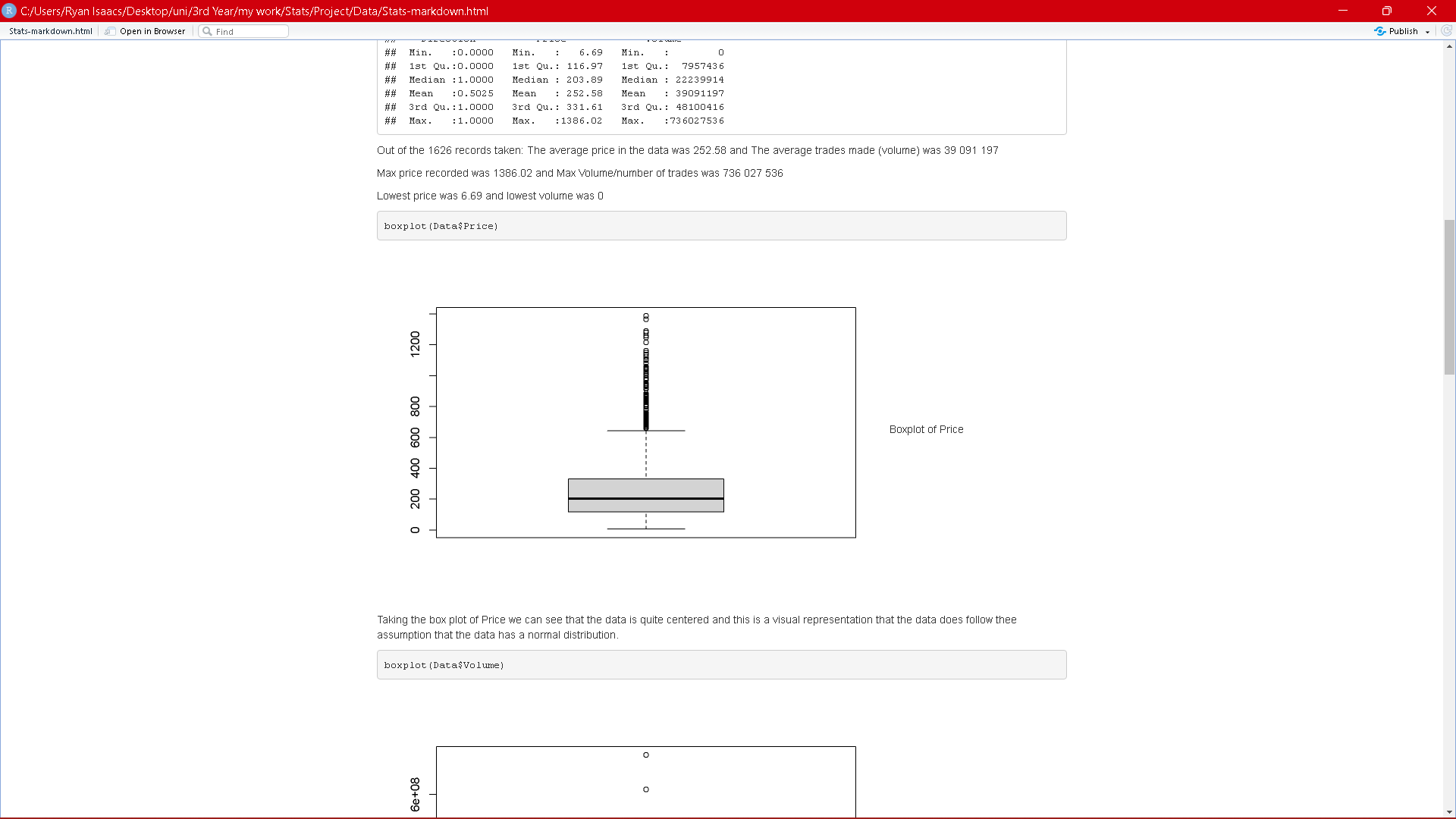


Figure 2: Price Boxplot

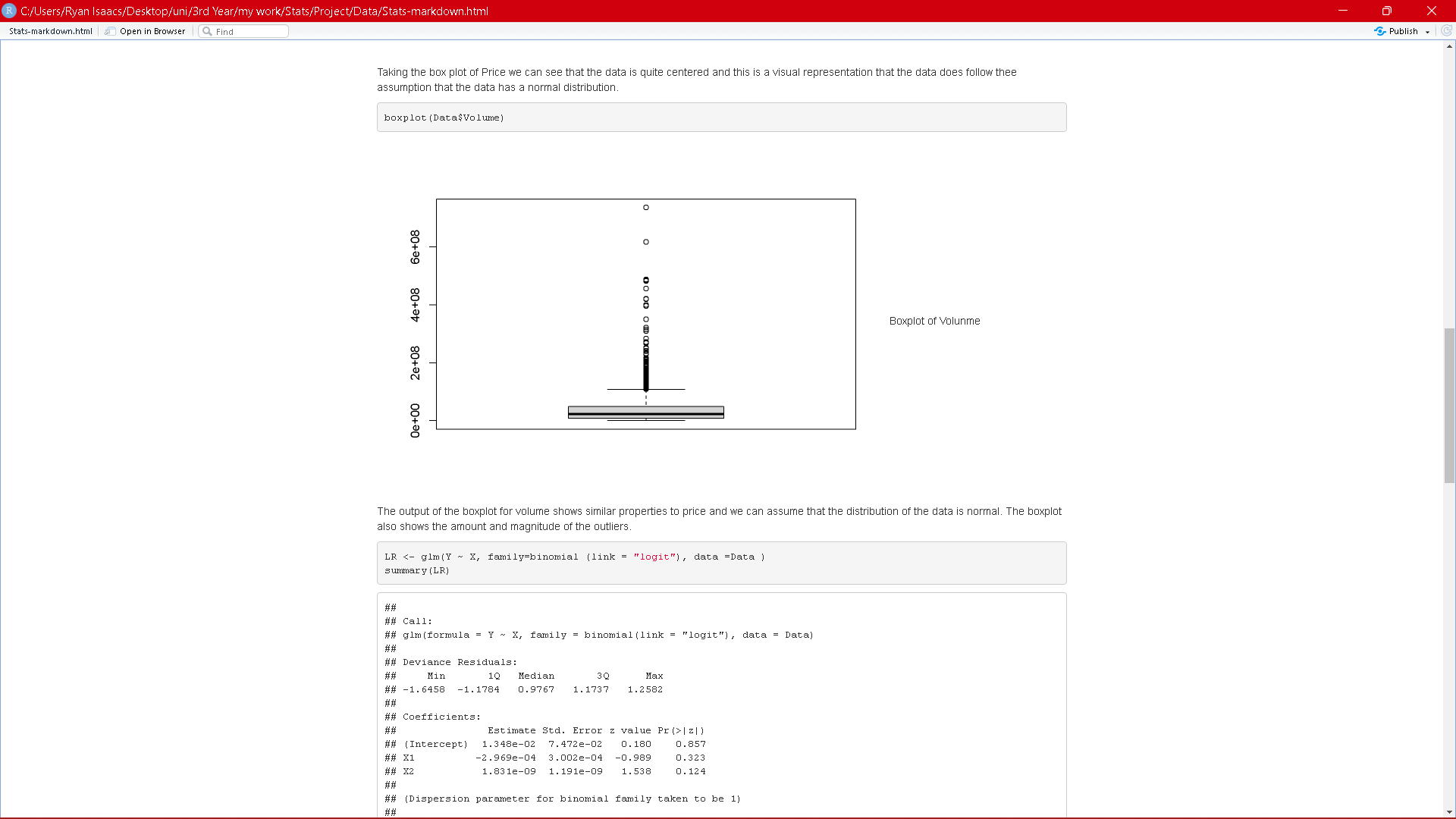


Figure 3:Volume Boxplot

1. **Test for model assumptions:**

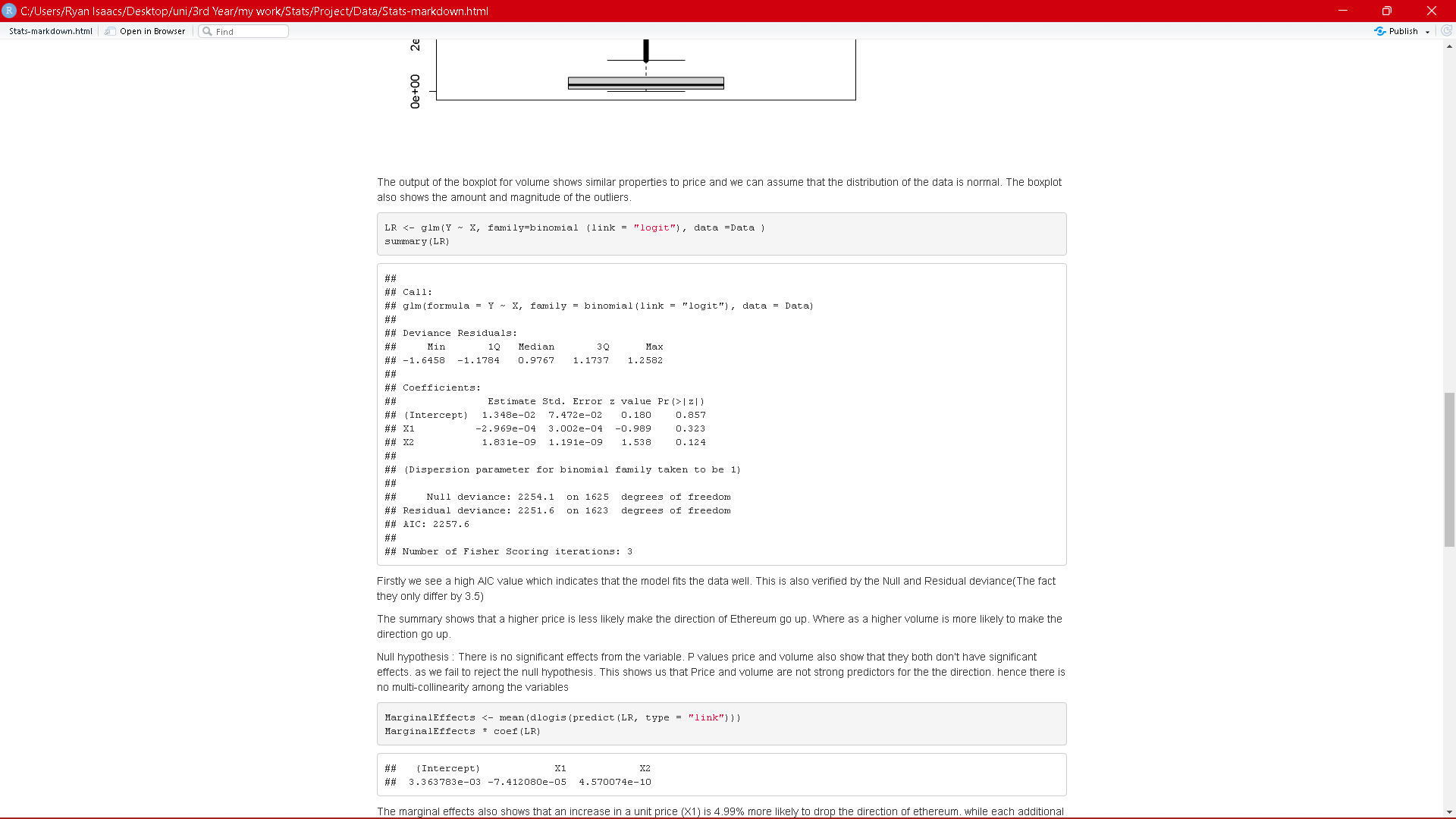


Figure 4: Logistic Regression Summary

|  |  |
| --- | --- |
| **Assumption** | **Validation** |
| The response variable is a binary or dichotomous variable (yes/no) | In figure 1 we see that our response variable “Direction” is indeed binary/dichotomous. |
| Logistic regression does not require a linear relationship between the dependent and independent variables. | This assumption is already satisfied because there is no need for linear relationship. |
| There are no influential outliers in the continuous predictors | When we did the outlier treatment, we saw only a 2% increase in the models accuracy and concluded that the outliers do not greatly influence the model. |
| There is no high multi-collinearity among the predictors | Figure 5 shows the test for multi-collinearity, in which we concluded that there is no significant correlation between the predictor variables |
| Normality in data | Boxplots in figure 2&3 indicate a normal distribution for the Price and Volume variables |

1. **Results, interpretations and discussions**

In this section we will interpret the results of the Logistic Regression model and discuss these results and also our findings.

In Figure 4 we can find a large amount of information of the Logistic Regression model. Firstly we see a high AIC value which indicates that the model fits the data well. This is also verified by the Null and Residual deviance (The fact they only differ by 3.5).The summary shows that a higher price is less likely make the direction of Ethereum go up. Whereas a higher volume is more likely to make the direction go up.

Null hypothesis: There is no significant effects from the variable.

P values price and volume also show that they both don't have significant effects as we fail to reject the null hypothesis. This shows us that Price and volume are not strong predictors for the direction, hence there is no multi-collinearity among the variables.

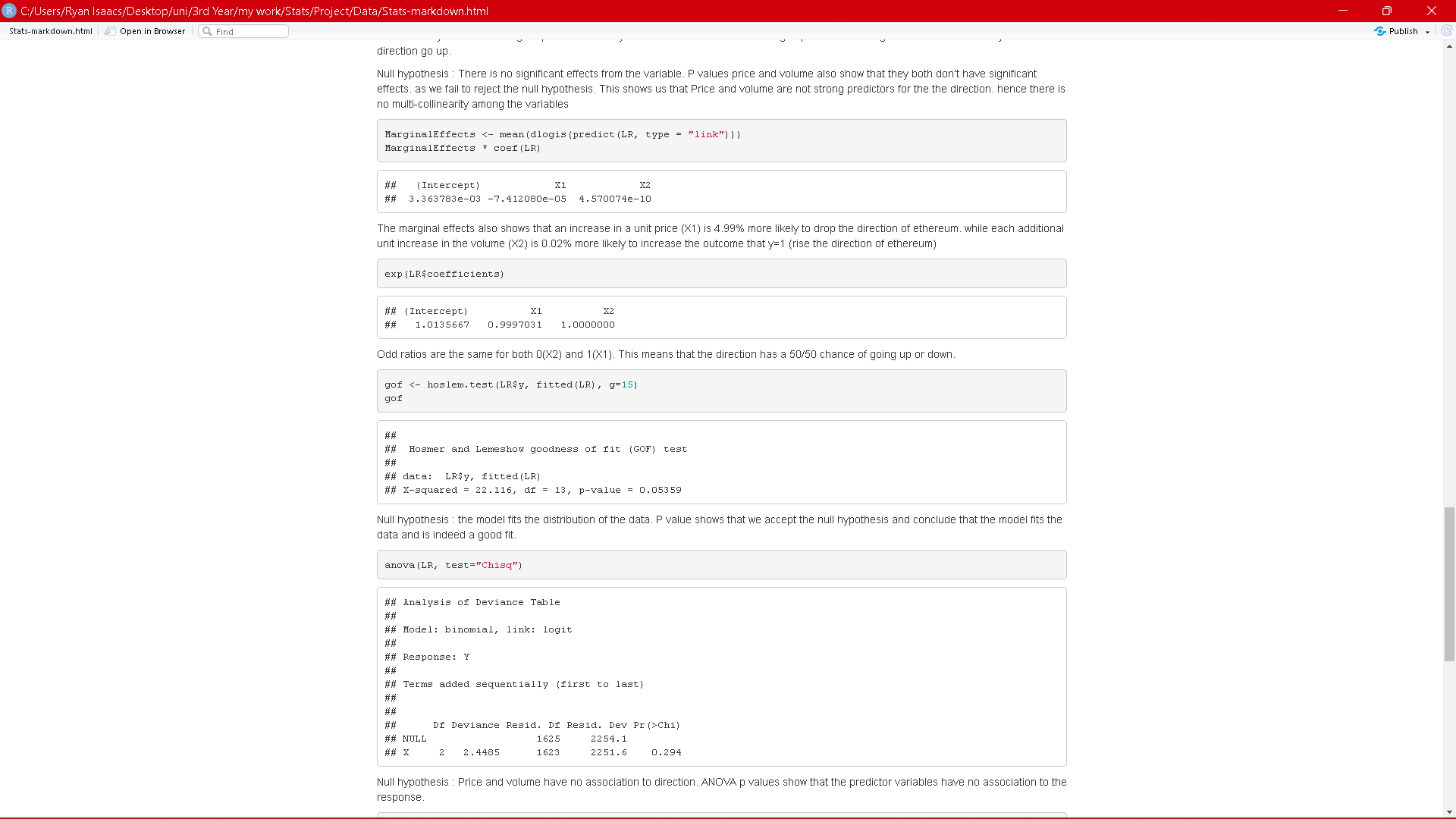


Figure 5: Relevant model tests

Figure 5 shows the marginal effects that an increase in a unit price (X1) is 4.99% more likely to drop the direction of Ethereum while each additional unit increase in the volume (X2) is 0.02% more likely to rise the direction of Ethereum (increase the outcome that y=1).

Odd ratios are the same for both 0(X2) and 1(X1). This means that the direction has a 50/50 chance of going up or down.

For the goodness of fit test, null hypothesis: the model fits the data.

P value shows that we accept the null hypothesis and conclude that the model fits the data and is indeed a good fit. This can be backed up by the residual deviance where the difference is small indicating a good fit.

Thee ANOVA, null hypothesis: Price and volume have no association to direction.

ANOVA p values show that the predictor variables have no association to the response.

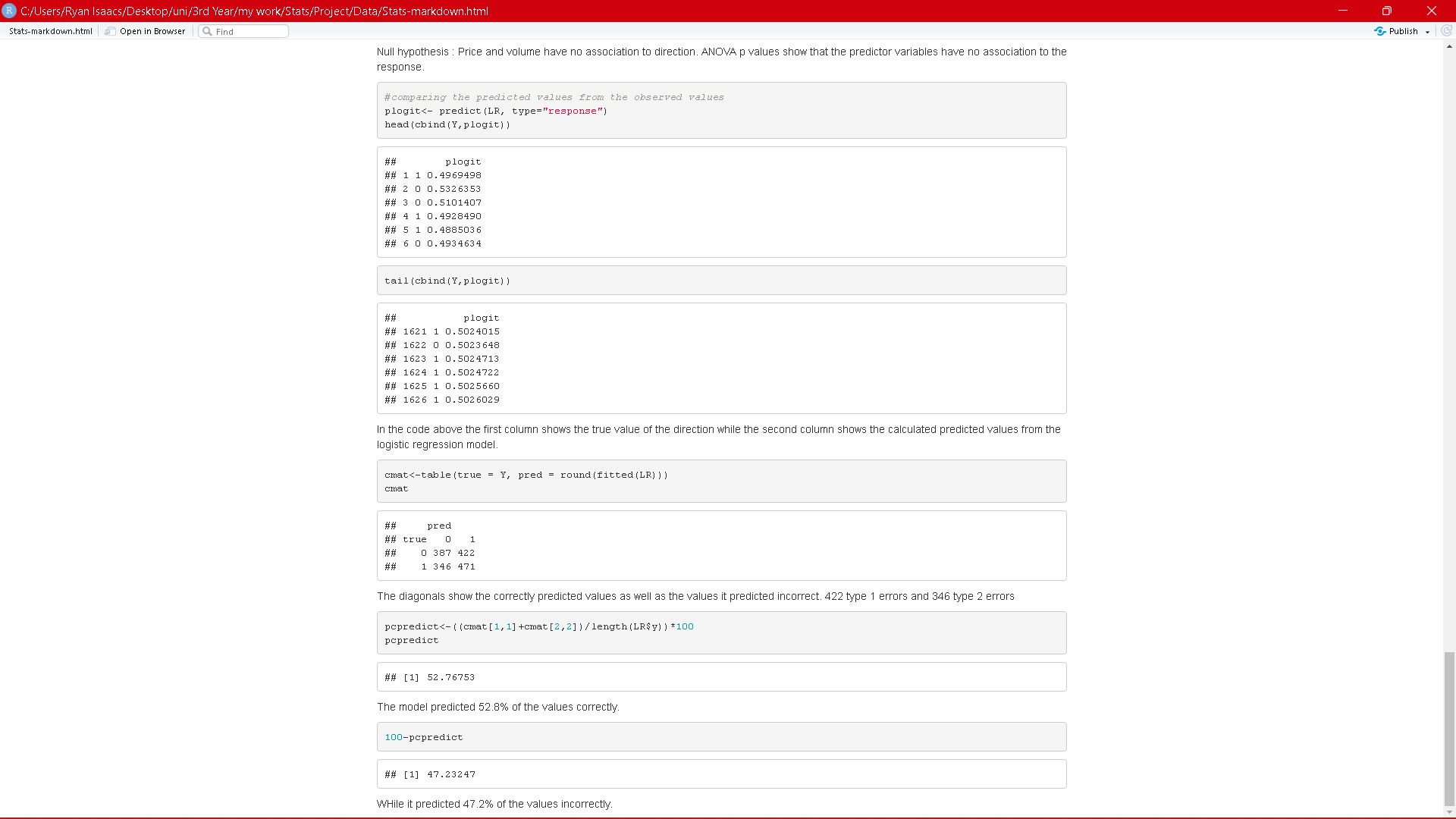


Figure 6; Prediction results

The beginning of figure 6 shows the head and tail of the predictions made by the model with the actual Direction. To better understand this we create a confusion matrix to see the correct predictions and also the type 1 & 2 errors. From this we can calculate the accuracy score of the model. We see that the model correctly predicts 52.76% of the data and has a misfit percentage of 47.23%

1. **Conclusions**

In conclusion we aimed to predict the Direction of Ethereum base on its price and the volume (Number of trades made). The problem was a binary classification problem because the Direction could only take the values of 0 or 1(Go up or down), and therefore we decided to use the Binary Logistic Regression model.

We obtained results that show Ethereum exchange rate is very difficult to predict based on only price and volume. Not to say that these variables do not have an effect on the direction but the effect they have is just not significant enough to make very accurate predictions. Our model managed to have an accuracy above 50%, which is not the best but we think that there is many external factor that play a big role in the exchange rate of Ethereum.

Examples being the current war between Ukraine and Russia, also billionaires like Elon Musk and Donald Trump who are able to influence not only the market but people who invested in Ethereum.

Our conclusion is that Price and Volume is just not sufficient enough to predict the direction of Ethereum based on the reasons we have discussed.

**References**

online.stat.psu.edu. (n.d.). *12.1 - Logistic Regression | STAT 462*. [online] Available at: <https://online.stat.psu.edu/stat462/node/207/>.

Statistics Solutions. (2019). *Multicollinearity - Statistics Solutions*. [online] Available at: https://www.statisticssolutions.com/multicollinearity/.