Bitcoin Price Prediction (lightweight CSV) Data Analysis

By: Ron Brody

Link to Kaggle Case and Data: https://www.kaggle.com/team-ai/bitcoin-price-prediction

Situation/Objectives:

Using the bitcoin price history data from 4/28/2013 to 6/31/2017, I want to predict the open prices from 7/1/2017-7/7/2017. I also want to see which variables have the biggest impact on what the open price will be. The training data set includes 1556 rows while the test data set includes only 7 rows.

Variables:

	Date	Open	High	Low	Close	Volume	Market Cap
1	31-Jul-17	2763.24	2889.62	2720.61	2875.34	860575000	45535800000

Above me is a record in the training data set

Date- date bitcoin price was announced

Open- Price that bitcoin opened on that day

High- Highest price for bitcoin on that day

Low- Lowest price for bitcoin on that day

Close – Price that bitcoin closed on for that day

Volume – Volume of bitcoin on that day

Market Cap- Market cap of bitcoin on that day

Data Cleaning:

```
FALSE  TRUE
10649   243
```

I first created a table to determine how many null values are in the bitcoin training dataset and found there to be 243 null values out of the 10,892 values in the table. I then created cross table tables for each variables to determine which columns my null values are coming from.

```
FALSE  TRUE
10649   243
```

```
FALSE  TRUE
 1313   243
```

After examining a few different variables I found that all the missing values came from the 'Volume' column of my training data set.

```
bitcoin.median   45301400
```

I then created a formula to calculate the median of the volume column. Using this formula, I found the median to be 45,301,400.

I then filled all the missing values in the Volume column with this median of 45,301,400 and now have completed filling in the null values.

Model

```
Call:
lm(formula = Open ~ High + Low + Close + Volume + `Market Cap`,
    data = bitcoin)

Coefficients:
 (Intercept)          High           Low         Close        Volume
   4.507e+00     8.502e-01     4.895e-01    -5.042e-01    -5.634e-08
`Market Cap`
   1.078e-08
```

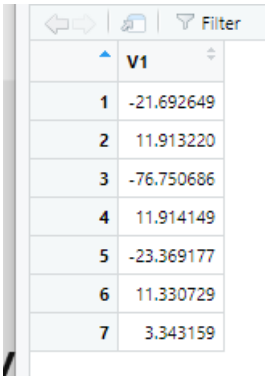
I then created a linear regression model to predict the Open price for bitcoin in the first week of August 2017. I used every variable to predict the open price except for date because it is an ID variable and it not measurable.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.507e+00  8.684e-01    5.19 2.38e-07 ***
High          8.502e-01  1.662e-02   51.15  < 2e-16 ***
Low           4.895e-01  2.377e-02   20.59  < 2e-16 ***
Close        -5.042e-01  2.449e-02  -20.59  < 2e-16 ***
Volume       -5.634e-08  3.815e-09  -14.77  < 2e-16 ***
`Market Cap`  1.078e-08  6.019e-10   17.91  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As you can see from the summary of the step model, all of the variables have an equally strong relationship with the open price of the bitcoin with P-values very close to 0. The high price and

close price both have the biggest impact on the open price with their 8.502 and -5.042 estimates.

Performance Metrics



	V1
1	-21.692649
2	11.913220
3	-76.750686
4	11.914149
5	-23.369177
6	11.330729
7	3.343159

I then used the linear regression model to predict what the open price of the first days of August will be and I found that I was off by a specific amount of dollars on each day which is shown in the dataset above

rmse_linear	32.3613329564721

This RMSE calculation shows I was off by 32.36 dollars on average for my predictions of the open price

mape_linear	0.00786926364961084

This MAPE calculation shows I was off by .7% on average on each prediction that I made for the open price on those 7 days as well