The aspirations of this project are to build a machine learning model that will predict the price of Bitcoin based on Tweets mentioning Bitcoin. We being this project by downloading to CSV file from Kaggle. We were able to find two CVS including one with historical pricing and another collecting tweets with hashtag Bitcoin. Unfortunately, this is where we ran into our first problem. Our tweets data was extremely huge, over 3 million lines in a CVS, the file was over 1.5 GB.

Although the file took several minutes to download, we were not deterred from using the data, yet. We had no problems dropping and changing the names of the columns. We changed the data type of an columns with dates or numbers, the originally came in as string. After further inspecting our tweets data, it was clear that we only needed to use a few columns as most of the data was irrelevant.

One of the biggest road blocks was trying to merge both data sets. The one column both had in common was the Date column. After several test attempts, we realized the data type had to be datatime64. At this time, we also found out that our data only overlapped for a few months, from 02/05/2021 to 07/06/2021. Because of this we were left with very few data points. From this point we decided to do the machine learning based only on the pricing data. We would use both data sets to visualize data in our dashboard.

The first machine learning model we used was linear regression on the pricing data. With the time left we were only able to achieve test scores of 66%. Given more time we could have run more test to achieve a better test score.

We were also able to use machine learning models, ARMA, AIRMA and SARIMA. We had better success with these models. Autoregressing moving average (ARMA) model predicted there would be a downward trend. Autoregressive integrated moving average (ARIMA) model's prediction was un upper trend. Finally, seasonal autoregressive integrated moving average (SARIMA) model's prediction was a convincing exponential graph. Our first thoughts were to buy more bitcoin! These results were based on the complete data starting from 2013. From 2013 to 2017 the pricing remained relatively low. It wasn’t until mid-2017 that there started to be volatility in pricing.

In our last tests we ran the data starting from 2017. This changed the machine learning predicted. Previously we had an exponential growth. If we only test the volatility the prediction was an exponential downtrend. Our thoughts then changed to, we need more data!