Introduction

The world today revolves around data with every institution and organization trying to include a data analyst/scientist in their team. Analyzing a set of data to analyze it and gain some information from them might prove to be a challenging task however, that is what most of the future trends depend upon. As such, in the very said field, cryptocurrency is an up-and-coming field in today's market. The concept of crypto currency started from back in 2009 when the first open source decentralized crypto known by the name of "bitcoin" was released. Following that, many other cryptocurrencies were developed.

The project is based on analyzing the historical data of the identified cryptocurrencies, while also tallying its popularity on the social platform twitter. This analyzing of the dataset is to determine which of the selected cryptocurrency would be suitable for better investment.

Problem Statement

The main aim of the project was to analyze the historical dataset of the cryptocurrencies to analyze the trends emerging from the data and determine which among those would be a better choice of investment. To achieve this aim, a list of objectives was set out to accomplish the main criteria for the project:

◊ Preprocessing of the initial dataset using SAS and its various functionalities to generate a master table.

◊ Creating required variables in the dataset by summarizing and recoding columns to adjust and calculate the required variables.

◊ Using different visualization techniques such as bar charts and donut charts to analyze the trends from the historical data.

◊ Using test such as non-parametric tests as well as linear models and linear regressions to analyze the mean and variance.

◊ Using ARIMA model to forecast and make predictions for the dataset.

Datasets

The datasets for the master tables were scraped from the site:
https://www.investing.com/crypto/
The website holds the data of all the cryptocurrencies from when it began trading in the market. For the project, we have two master tables.

One master table is for all the historical data of the cryptocurrencies that holds the cryptocurrency by week, month, and year. We recoded the columns to find the change %, which is the difference in the high and low price of the crypto currency based on the date. We also calculated the percentage of growth for the said crypto. All of this was done using functionalities such as TRANWRD, SUBSTRING, string replace, CASE expression, Joins, recoding columns, and computing columns. One of the hardest parts of the dataset was to alter the format of the date. The date conversions were challenging and was converted to weak, year and month format successfully.

The second master table holds the tweets for the selected crypto currencies. The table holds the columns for the name of the cryptocurrency, the date, and its popularity. The popularity holds the number of times the name of the cryptocurrency popped up in the tweets for the date. For this, functions such as COUNT, DISTINCT, recoding columns and computing columns were required.

Data Scraping

The popularity of the cryptocurrency was determined by how many tweets each of the coins had each day. A specific timeframe, i.e., a more recent timeframe from 2017-2020, was selected due to cryptocurrencies seeing an increase in both popularity and transactions during this period.

To gather the data, a snscrape tool was used to scrape the Twitter data of the selected coins. To set that up we needed the right python setup in our system (Python 3.8 or higher). The development version was used instead of the released version as it supersedes the need to access Twitter URLs using the **--jsonl** argument in JSON format. Using the command line interface (Anaconda PowerShell) to execute the script. Since historical data was required. the arguments

--since and **"until:"** were used to define the time range and **--max-results # 10000** was used to set the cap of the results to ten thousand records. This resulted in the original dataset of the tweets scraped for analysis which was later cleaned.

Data Cleaning

The part of data cleaning includes vigorous methods of recoding and summarizing the formats of the dataset as per the requirement for the project. For instance, one of the most crucial yet important process for data cleaning included the date modification for the dataset. The provided initial date format was not the one that was viable for the project. We used the PUT and INPUT function to properly convert the datatype of the date/year for the dataset.

The validation to ensure that the data was clean was done by making sure that there were no mandatory constraints that were mission. Also, unique combinations were used to ensure that the data formats were changes appropriately regarding the format of the data. This included constraints such as datatype and range to be properly validated. The functions INPUT, PUT and formats were used to ensure that the number, dates were within the range and the constraints on the datatype with numeric and character were properly segregated.

When merging the entirety of the collected dataset for the cryptocurrency to create a master table, we came across the challenge for missing data. For the popularity of the cryptocurrency in their beginning phase, the data is missing, and we did not change the rows for the missing value because then it would clash with the accuracy of the forecasting for the cryptocurrency since it is a time-series forecasting.

The dataset was taken from a reliable data source, the consistency and uniformity of the dataset were an integral part. The cleaning of the dataset included standardizing the computed columns with proper formats, correcting any syntax errors, cleaning irrelevant data, and computing relevant columns.

Summarized Columns

This section includes the columns that were recoded, summarized, and computed to adjust it to the dataset as per the variables required for the project.

For the master table that held the tweets there were two computed columns: the first columns were the name of the crypto where the dataset for the cryptos were segregated by their names. The second column was the column popularity where the popularity for the cryptocurrencies were calculated using different SAS functions. The count for the popularity depended on the appearance of the name of the cryptocurrencies in the tweets for the given date.

For the second master table of the dataset for the cryptocurrencies, variables, and columns as change%, change growth rate, popularity and the date columns were recoded, summarized, and computed.

Analytics | Visualizations

For the project, most of the SAS functionalities taught throughout the semester were used to ensure that the columns and their format were corresponding to the dataset. Similarly, functionalities such as, INPUT/PUT function were used to change the format from numeric to character and vice-versa. There was also the use of append function to append the mini tables of the tweets to a master table. The next step involved the visualizations of the final dataset using the knowledge from the module. The tools used for these were visualizations provided in SAS such as the bar chart and the donut pie chart.

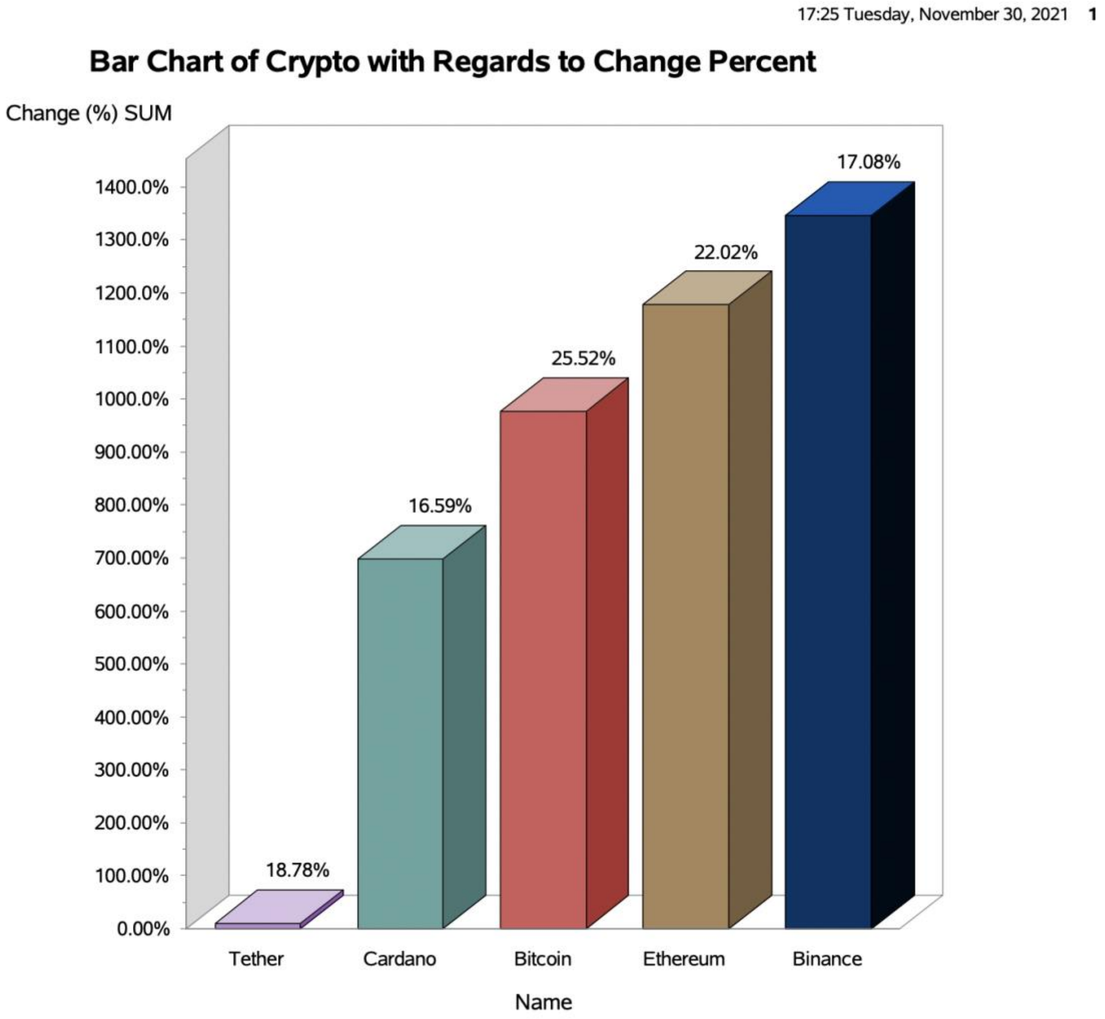

Fig 1: Bar chart to show the change percent of the cryptocurrencies

The image in *figure 1,* shows us the results from the bar chart, which has been visualized for each of the cryptocurrency regarding their change%. The change% is the rate of difference of the price for the cryptocurrency. The cryptocurrencies have been plotted against the variable change% and arranged in a descending format from their height. We can see that the change % for the cryptocurrencies are: 18.78% for Tether, 16.59% for Cardano, 25.52% for Bitcoin, 22.02% for Ethereum and 17.08% for Binance.

The next visualization for the dataset was made from the Donut Chart as shown in the *figure 2*. The Donut chart shows the growth percent for each of the cryptocurrency and its impact. As we can see from the visualization that the highest change was in the cryptocurrency Binance and the least in Tether.

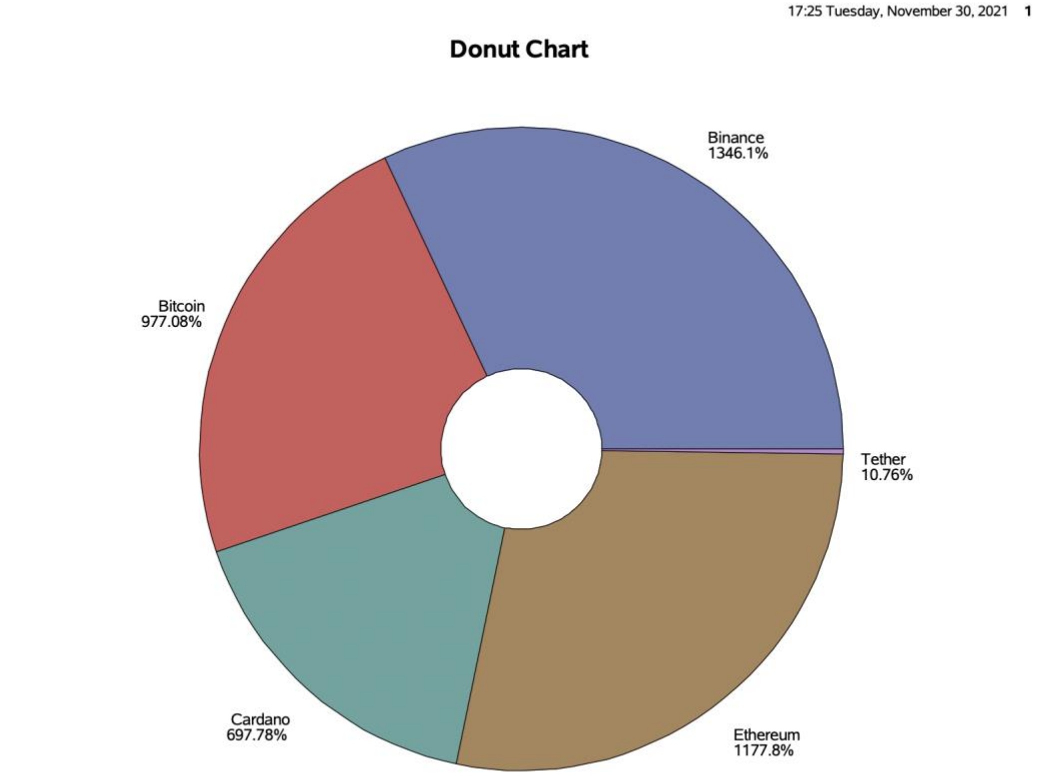

Fig: Pie chart to show the growth of the cryptocurrencies

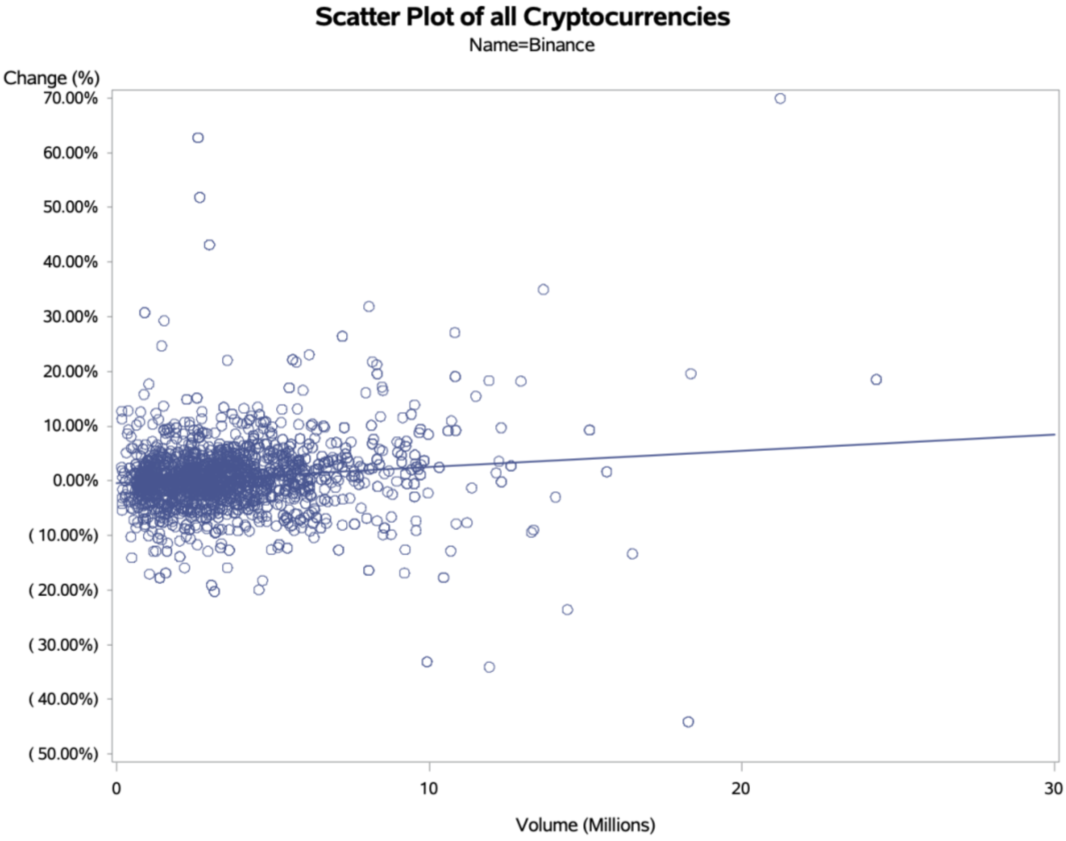

Fig 3: Scatter plot to show the change% of Binance in its volume presented in millions

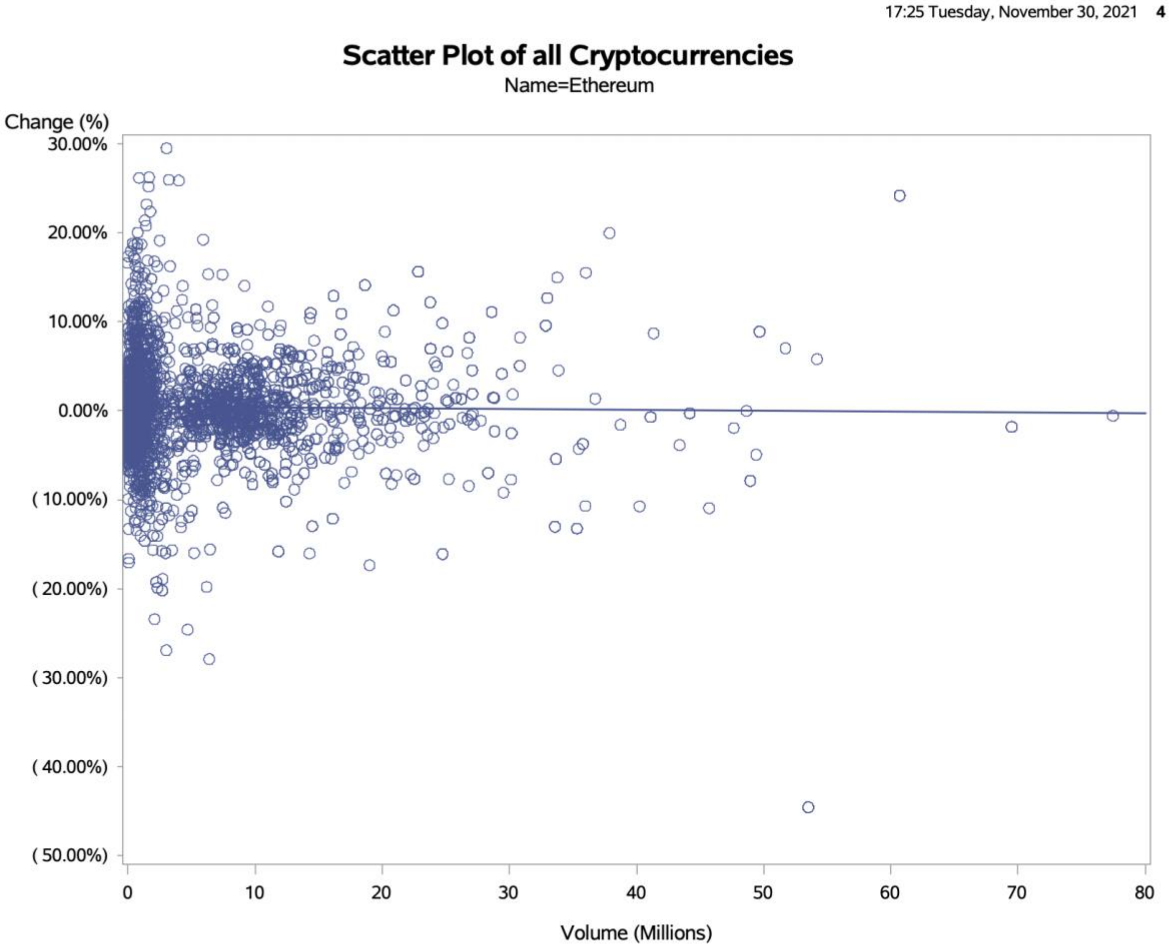

Fig 4: Scatter plot to show the change% of Ethereum in its volume presented in millions

The image in *figure 3, 4, 5, 6 and 7*, shows the use of the tool scatter plot to visualize the trend of the change percent in the selected cryptocurrency. For the report we have included the visualization for all the cryptos regarding change % and their count i.e., their volume in terms of millions.

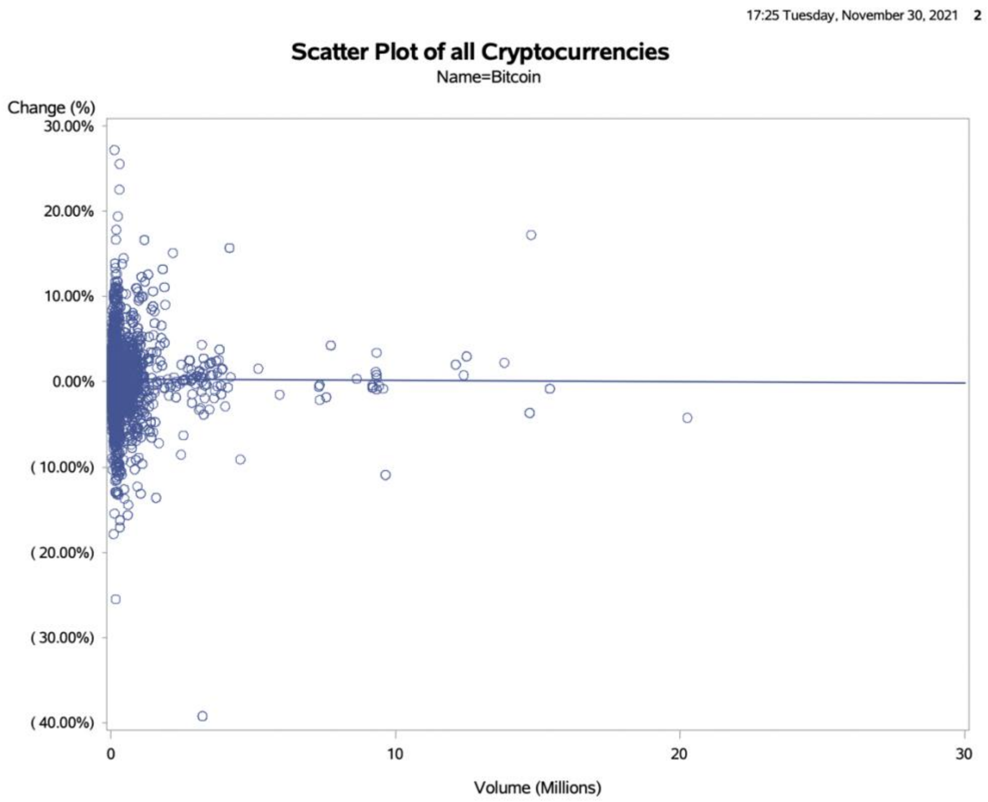

Fig 5: Scatter plot to show the change% of Bitcoin in its volume presented in millions

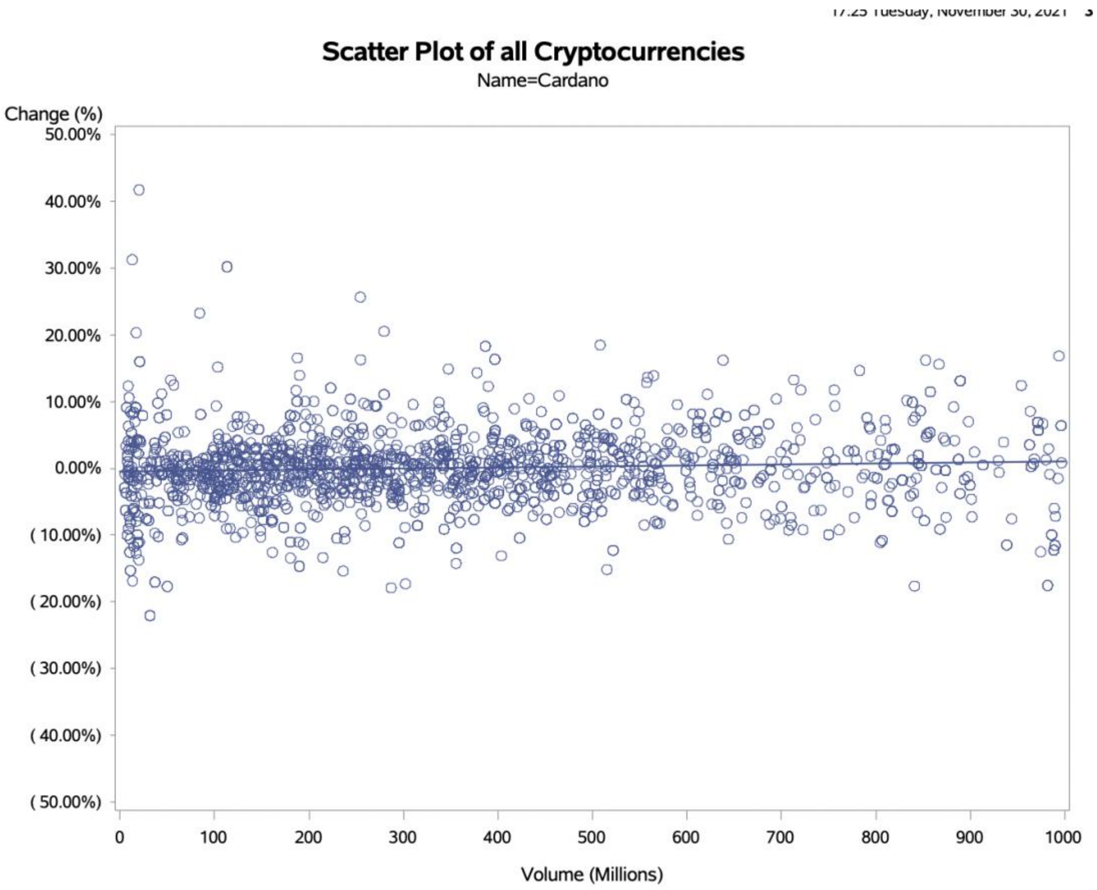

Fig 6: Scatter plot to show the change% of Cardano in its volume presented in millions

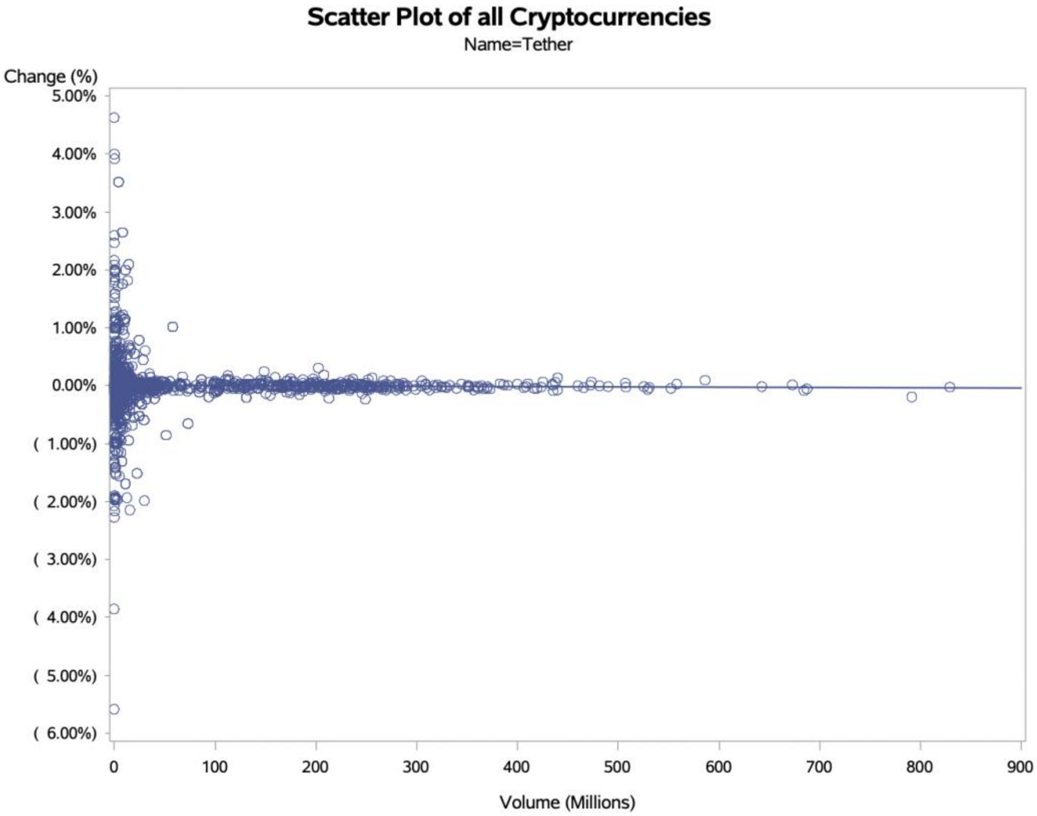

Fig 7: Scatter plot to show the change% of Tether in its volume presented in millions

Generalization

From the trends and the forecasts, we can see that the coins fluctuate at different rates. The explanation of each of the coin have been explained below:

- ◊ Binance: The most volatile coin (both long and short term) in the group but is a good investment if someone is looking to make a quick profit.

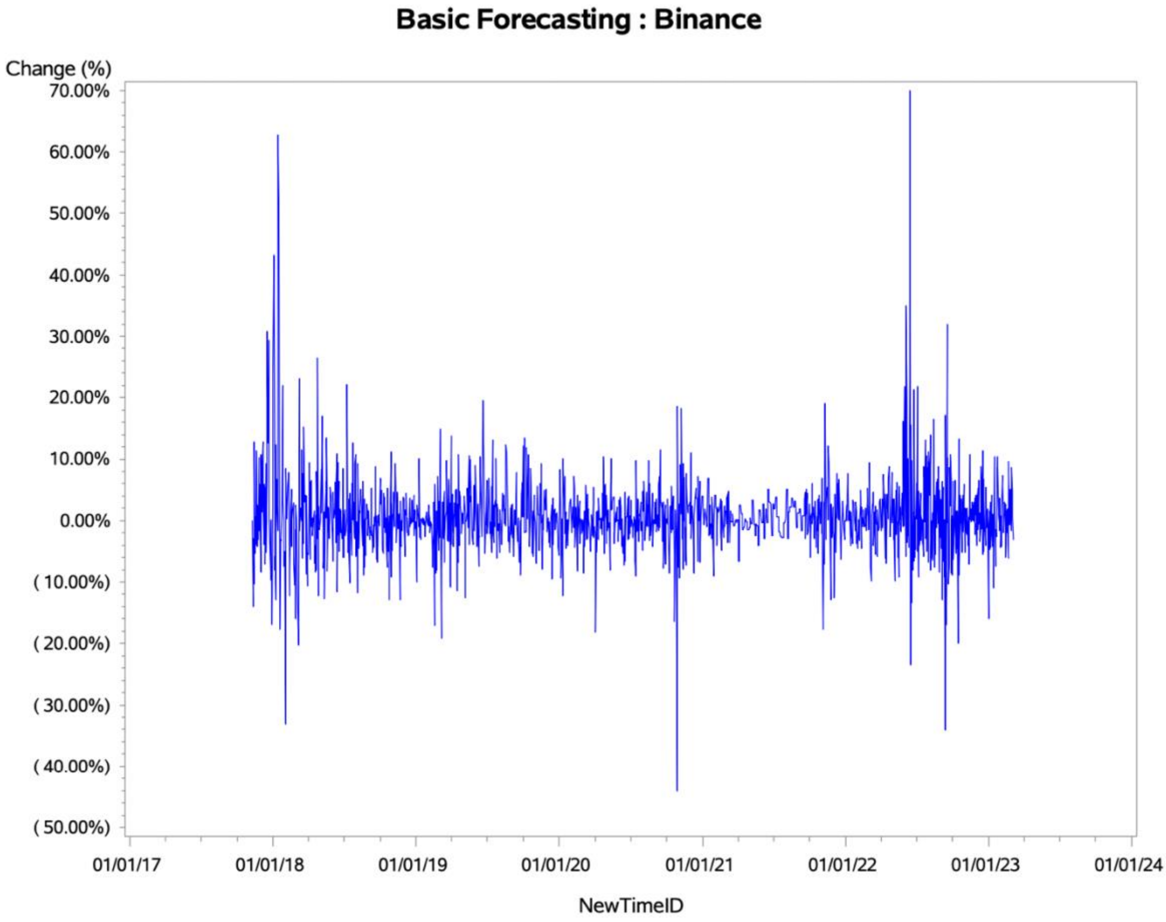

Fig 8: Basic Forecasting for the crypto Binance

◊ Bitcoin: The most profitable and expensive coin but also very volatile according to its analyzed trends.

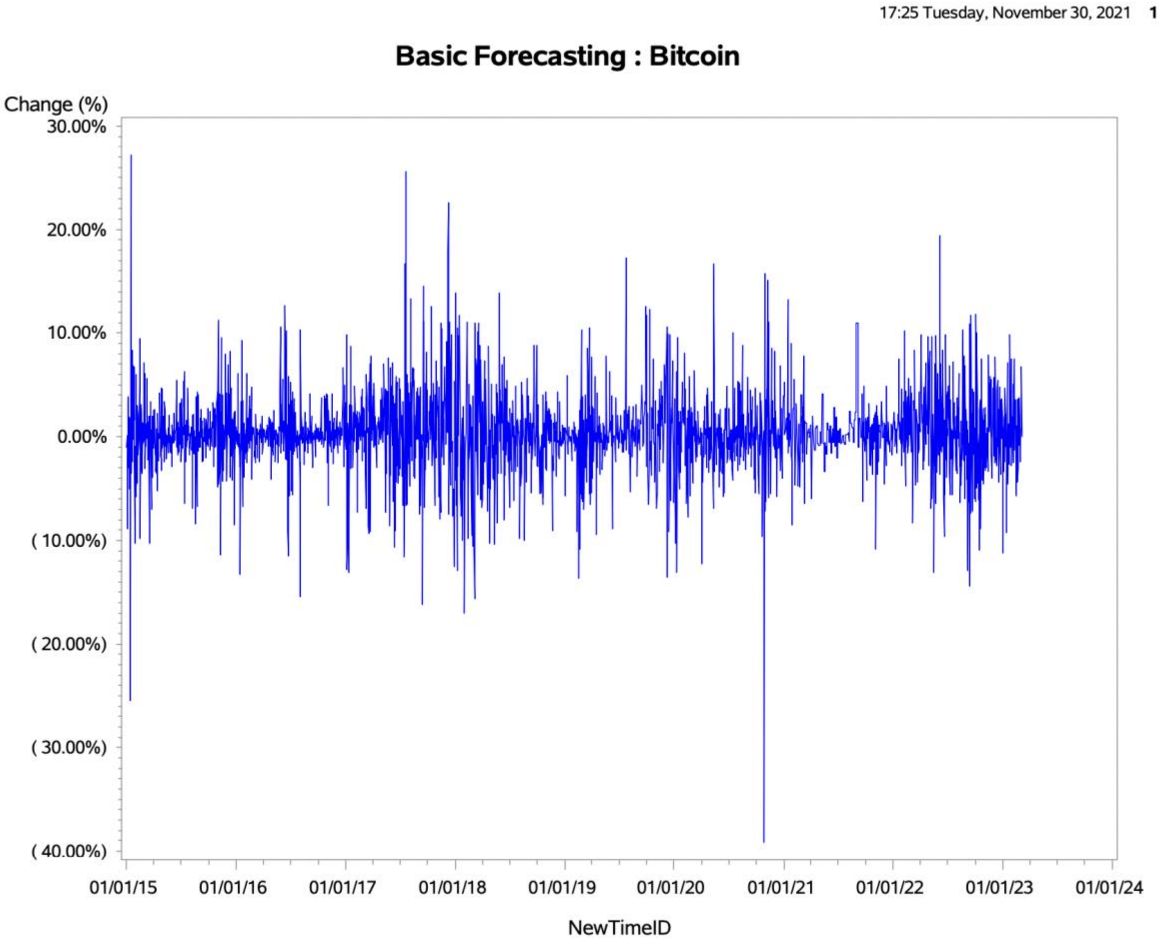

Fig 9: Basic Forecasting for the crypto Bitcoin

◊ Cardano: It's the second most stable coin amongst the group according to its long- and short-term analysis.

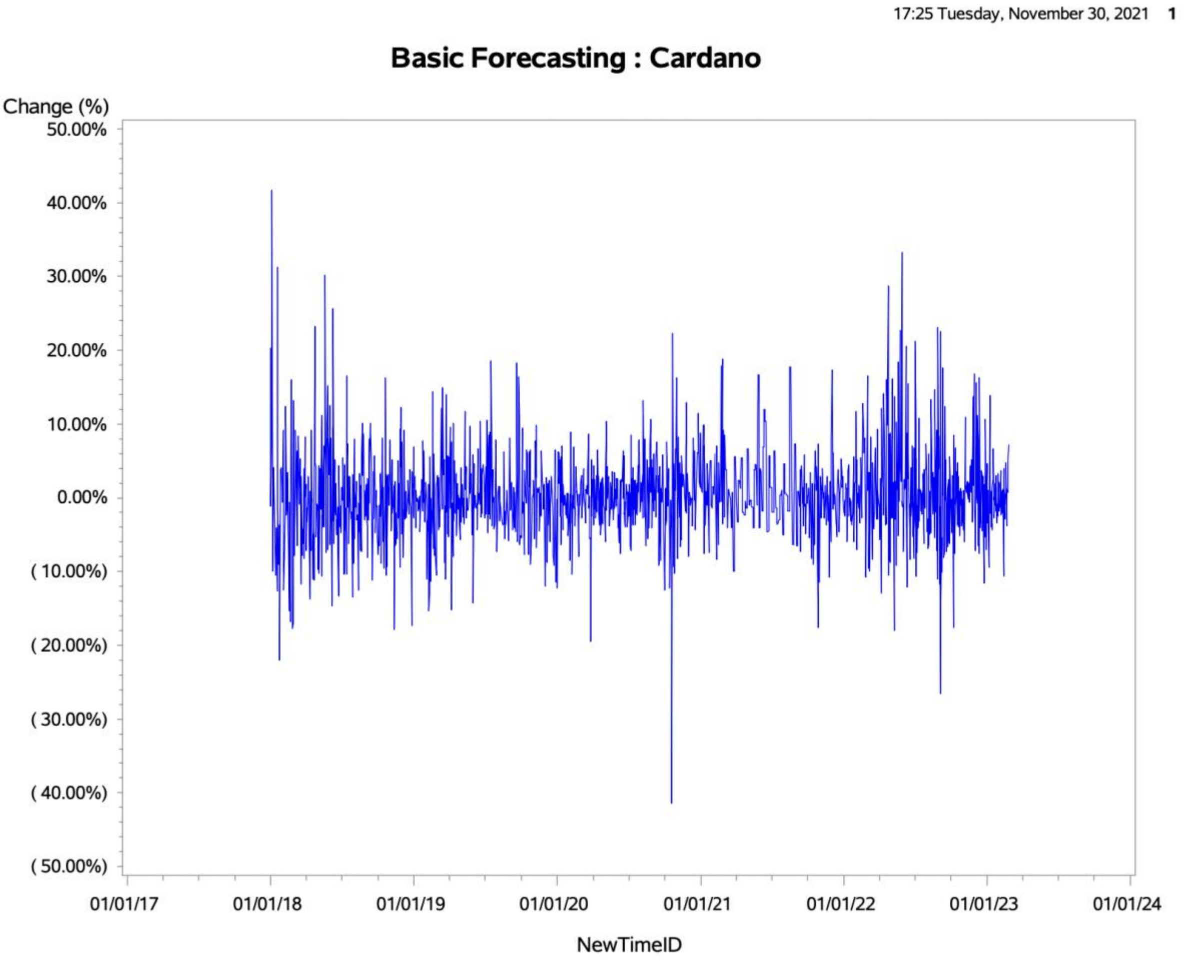

Fig 10: Basic Forecasting for the crypto Cardano

◊ Tether: It’s the most stable coin according to the trends and it has been forecasted to be stable in the coming year as well.

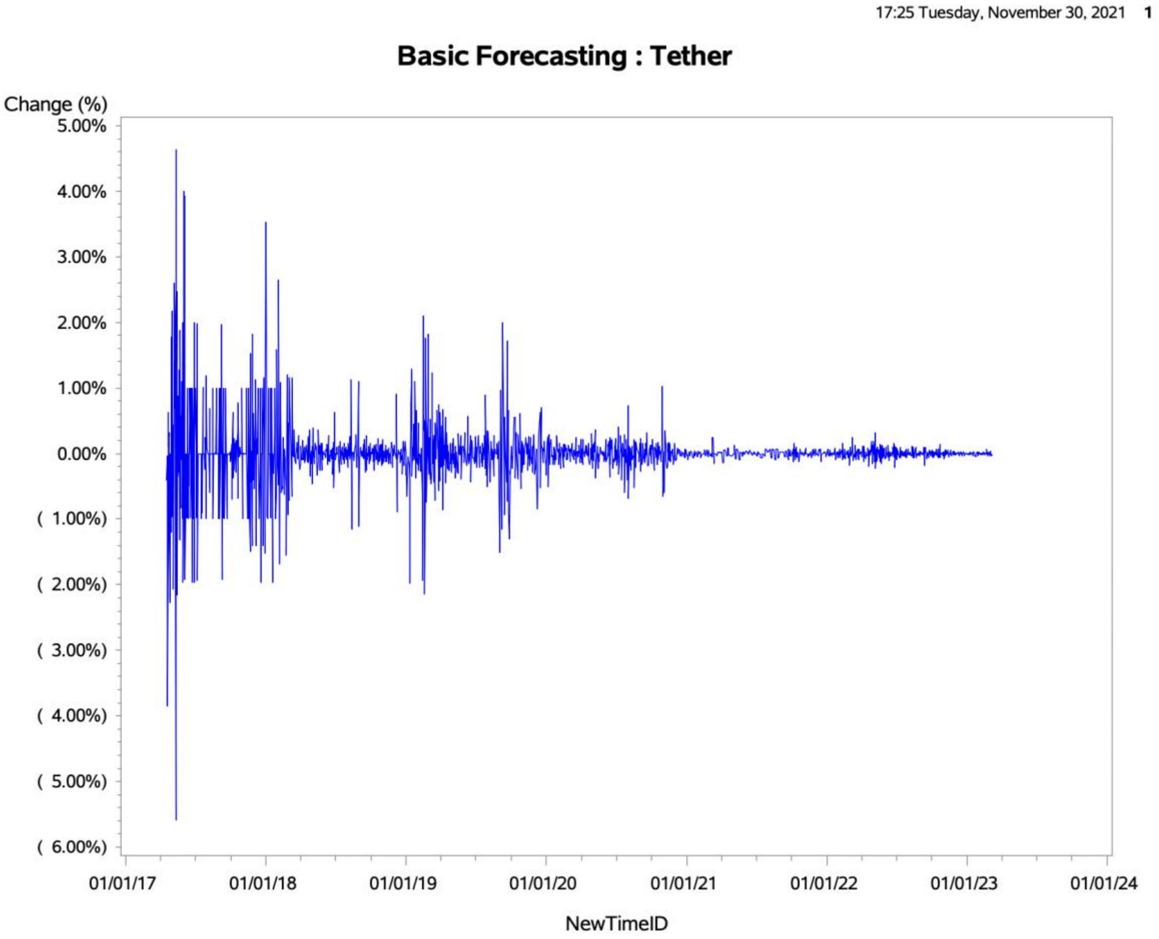

Fig 11: Basic Forecasting for the crypto Tether

◊ Ethereum: It is more unstable as a short-term investment, but it becomes more stable over time.

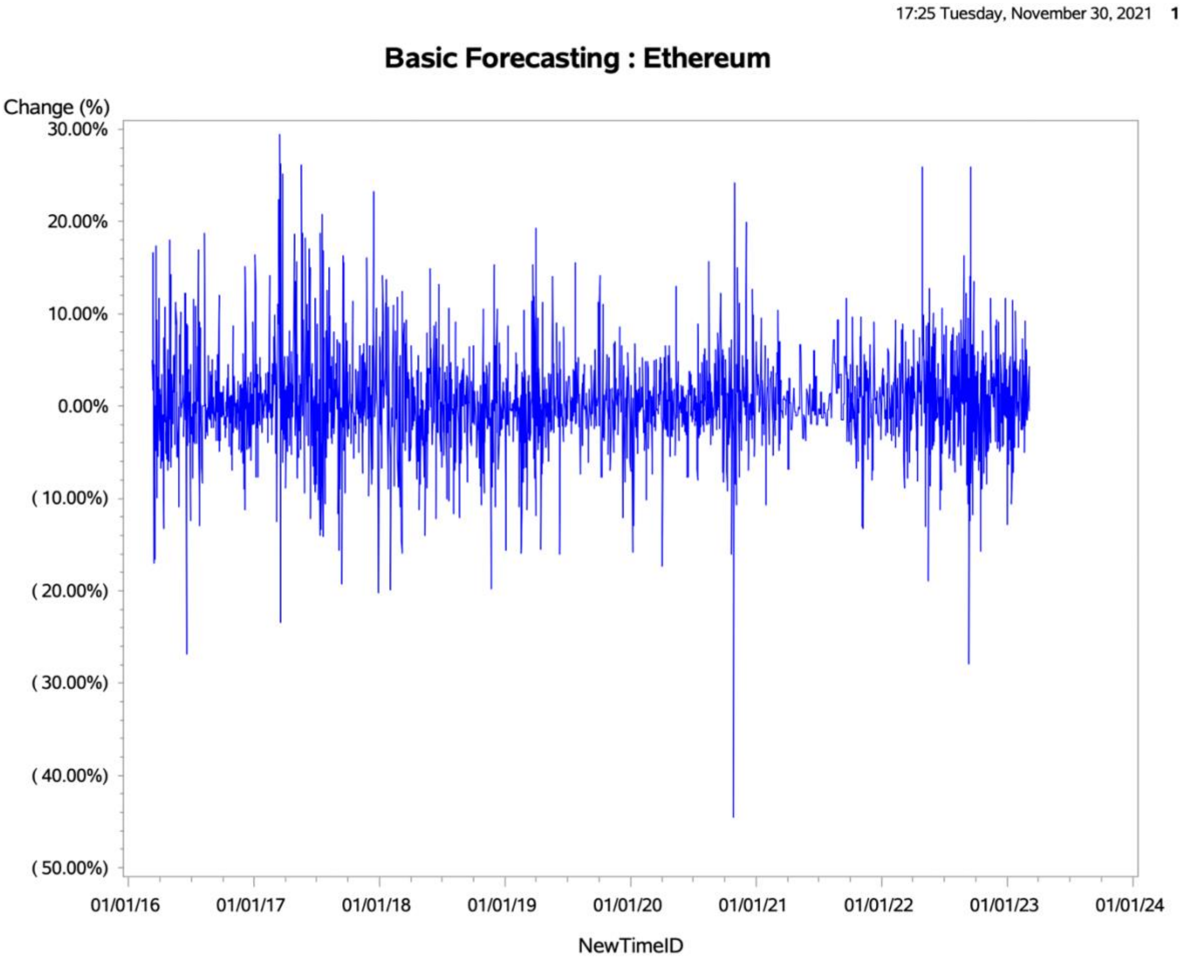

Fig 12: Basic Forecasting for the crypto Ethereum

The common trend we see from the line graphs is that as popularity increases the coins become more stable since more people invest in it.

Recommendations

There were a few limitations that we came across when working with the projects. The limitations for the project have been listed below:

- ◊ System could hold about 10000 rows of scraped data of tweets.
- ◊ The time range of the tweets fluctuates because of the limitation to scaping of the data.
- ◊ This was due to some of the coins (like Bitcoin) that was trending a lot which allowed for the data to be scraped within a limited timeframe.
- ◊ Not all the coins were launched at the same time. Example, Cardano = 2017, Bitcoin = 2009. This also affected the popularity since one coin had more time to cement itself in the market before another.

◊ SAS takes a long time to load the current dataset (since it’s 10,000 rows of tweets for each crypto). Use of more data could give better predictions.

◊ The data exhibits desirable properties such as stationarity dataset and some classical time series prediction methods exploits this behavior, such as ARIMA models, which produces poor predictions and lacks a probabilistic interpretation.

Some recommendations that can be made for the projects are:

◊ Taking datasets that are more probable regarding the time frame.

◊ Taking proper tweets as datasets regarding the chosen cryptocurrencies.

◊ Using proper functionalities to include the extremities of the factors that affects the trends of the cryptocurrencies.

Conclusion

Therefore, the project analyzed the trends in the selected five cryptocurrencies to analyze how the historical data for the cryptocurrencies have been till date. As such, this project helped us to organize our skills and capabilities to an extent to implement it in a real-life project. It helped us to clear our understanding on the concepts and to implement our learnings in a real environment. The project also helped us to understand concepts on time-series forecasting using ARIMA model as well as to perform some basic forecasting on the datasets.

The project in highlight, concluded that among the five cryptocurrencies, bitcoin is the crypto that is to invest in for a profitable result. However, from the up-and-coming trends, we can see that the most stable coin out of all is Tether followed by Cardano. This sums up the main aim of the project to analyze the historical trends of the data to get results to determine some of the most table cryptocurrency throughout as well as which crypto have been unstable or will be profitable.

Appendix

Team Members

All the team members of this project with their respective UTSA IDs are:

- Diego Aldo Pettorossi [zho125]
- Saket Mishra [jjw317]
- Manoj Siva Gannamani
- Ibteaz Hasan [uni298]
- Shreya Budhathoki

Meeting Minutes

This section holds the meeting minutes and the progress of each meeting that was scheduled over zoom. The meetings done in person has not been recorded for this section.

Meeting Minute 1:

Team Number: 4

Meeting Duration: 45 Minutes

Subject: Discussion on the probable ideas for the project

Attendees: Manoj, Diego, Shreya, Ibteaz and Saket

Meeting Progress:

- Came up with two ideas for the project.
- Went through probable datasets for the projects.
- Discussion on the main problem statement of the project.

Meeting Outcome:

- Finalized on two main ideas.
- Emailed the professor to get her input on the idea for the projects to finalize the project topic.

Next Scheduled meeting: A week Later.

Meeting Minute 2:

Team Number: 4

Meeting Duration: 30 Minutes

Subject: Discussion on the project and task division

Attendees: Manoj, Diego, Shreya, Ibteaz and Saket

Meeting Progress:

- ◊ Finalizing the project.
- ◊ Discussion about division of tasks
- ◊ Exploring the datasets.

Meeting Outcome:

- ◊ Finalized on the main idea of the project.
- ◊ Developed a problem statement for the project.
- ◊ Explored datasets to finalize on a few of them.
- ◊ Division of tasks among the members.

Next Scheduled meeting: 3 days later

Meeting Minute 3:

Team Number: 4

Meeting Duration: 1 hour 45 minutes

Subject: Project Progress

Attendees: Manoj, Diego, Shreya, Ibteaz and Saket

Meeting Progress:

- ◊ Finalizing the datasets.
- ◊ Understanding the concepts of scraping, summarized columns
- ◊ Deciding on the concepts of the summarized columns.
- ◊ Discussion on the types of visualization and analysis to be performed.
- ◊ Final division of tasks.

Meeting Outcome:

- ◊ Finalized dataset.
- ◊ Research task divided for scraping, recoding columns.
- ◊ Research task divided for visualization and analysis to be done on the dataset.

Next Scheduled meeting: A week later

Meeting Minute 4:

Team Number: 4

Meeting Duration: 2 hour 30 Minutes

Subject: Project Progress I

Attendees: Manoj, Diego, Shreya, Ibteaz and Saket

Meeting Progress:

- ◊ Progress for the project was discussed.
- ◊ Limitations faced were discussed.
- ◊ Probable solutions were suggested.

Meeting Outcome:

- ◊ Tracking the progress of the project.
- ◊ Helping with the progress of the project.
- ◊ Brainstorming for solutions to the limitations faced.

Next Scheduled meeting: a week later

Meeting Minute 5:

Team Number: 4

Meeting Duration: 2 hours

Subject: Finalizing the project

Attendees: Manoj, Diego, Shreya, Ibteaz and Saket

Meeting Progress:

- ◊ Finalizing the entire dataset.
- ◊ Finalizing the analysis that was already done.
- ◊ Finalizing the analysis if there were more to be done.
- ◊ Finalizing the project presentation.
- ◊ Discussion on the recording of the final presentation.

Meeting Outcome:

- ◊ Finalized dataset.
- ◊ Finalized analysis for the project.
- ◊ Finalized forecasting and visualizations.
- ◊ Finalized the presentation details before recording.

Next Scheduled meeting: The next day [for recording the presentation]