# Problem Statement

Cryptocurrencies are one of the most emerging technological advancements in the financial sector of the past decade, and the leading cryptocurrency is Bitcoin. Cryptocurrencies, by nature, are extremely volatile, and Bitcoin is no different. The ability to predict the future price of bitcoin and forecast it is an extremely valuable asset, as it could guide investors on when to invest into said cryptocurrency. It also is beneficial to gather information about the world economy as a whole; since bitcoin is the biggest cryptocurrency, it shows and dictates patterns to other cryptocurrencies and other economical factors.

The aim of this project is to develop a data mining model utilizing the KDD process in RapidMiner Studio to preprocess, mine, and extract valuable information from its historical price data, and compare which model and data mining technique yields the most accurate results, ultimately leading to a model that is able to accurately predict and forecast the future price of Bitcoin.

This project addresses the challenges of predicting the future price of such a volatile concept through historical data in relation to it. Applying data preprocessing, model training and evaluation techniques in RapidMiner Studio to come up with the most accurate model to predict the Bitcoin price in the future.

By the end of the project, we expect to find the most accurate model throughout our testing, that can effectively predict future prices of Bitcoin, and gain more knowledge on what are the factors that affect its pricing overall.

# Selection of Data Set

The dataset that we’ve chosen is the Bitcoin dataset from Kaggle from the article “Cryptocurrency Historical Prices.” Which contains historical Bitcoin statistics collected over several years, namely from 2009 to 2017, which encompasses numerous values such as market performance and blockchain activity.

We have chosen this dataset for our project for a multitude of reasons, including but not limited to:

* It’s directly related to our topic in the financial technology sector
* It has more than sufficient data and is structured well enough to come up with predictive models and forecast bitcoin prices
* It has great depth for modeling as it covers many concepts such as blockchain and economic features

The dataset consists of 24 attributes, 23 of them being quantitative (float64) and 1 date attribute. There are over 1500 instances/records, more specifically, 1584 instances.

The dataset also has 478 missing values in the btc\_trade\_volume attribute, which is a highly important attribute for our analysis and modeling, so we would have to fill those in during preprocessing in Task 3.

The dataset also contains some redundant attributes, with some attributes having a correlation value of >0.99, alongside multiple transaction related attributes such as btc\_n\_transactions, btc\_n\_transactions\_excluding\_popular, and btc\_n\_transactions\_total, this will also be addressed in task 3 with either feature selection or applying PCA.

The following are some visualizations regarding the dataset to further understand the distribution and summary of the data.

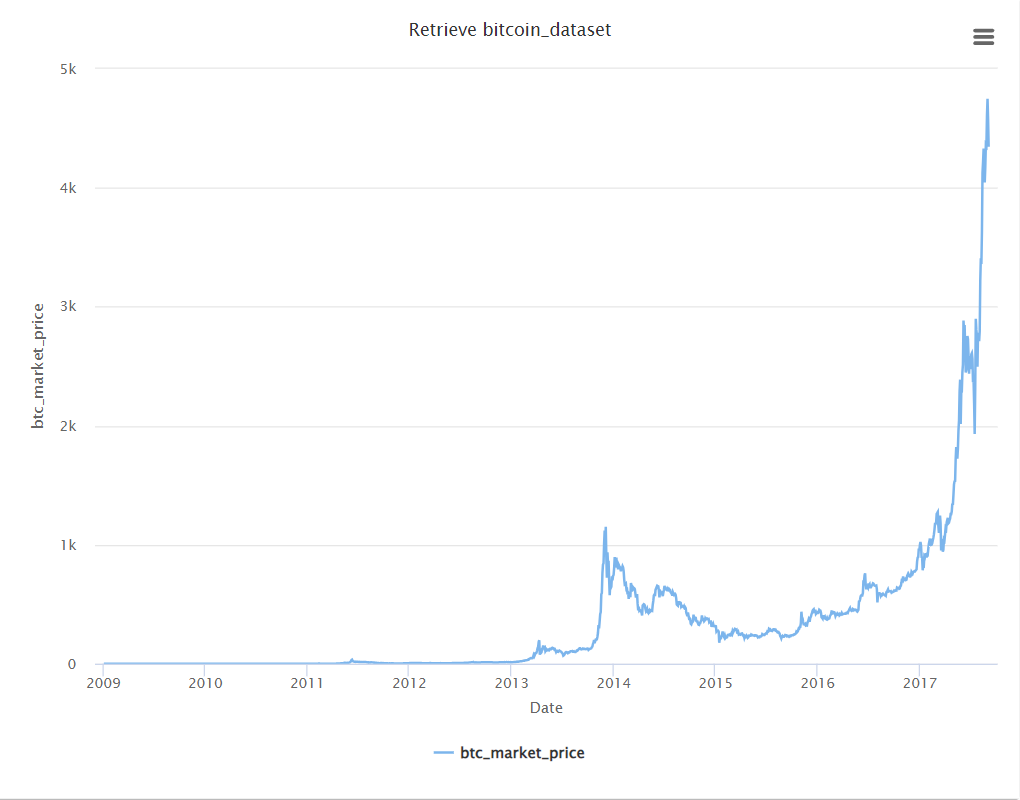


Figure 1 - The market price of bitcoin throughout the years

Figure 1 shows how volatile the price of bitcoin can be, and how it has had exponential growth in the later years.

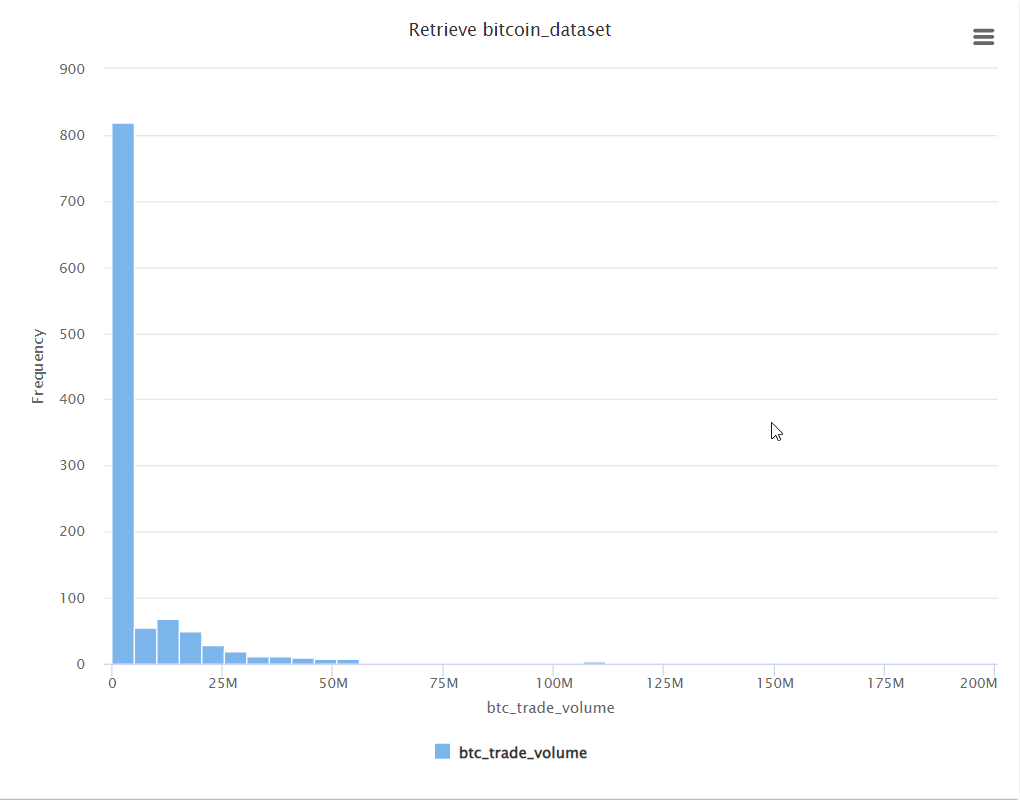


Figure 2 - A Histogram showcasing btc\_trade\_volume

As we can see from the histogram in figure 2, the btc\_trade\_volume attribute is extremely rightly skewed, and there are significant outliers in the data, showcasing the need for proper outlier treatment and normalizing the values.

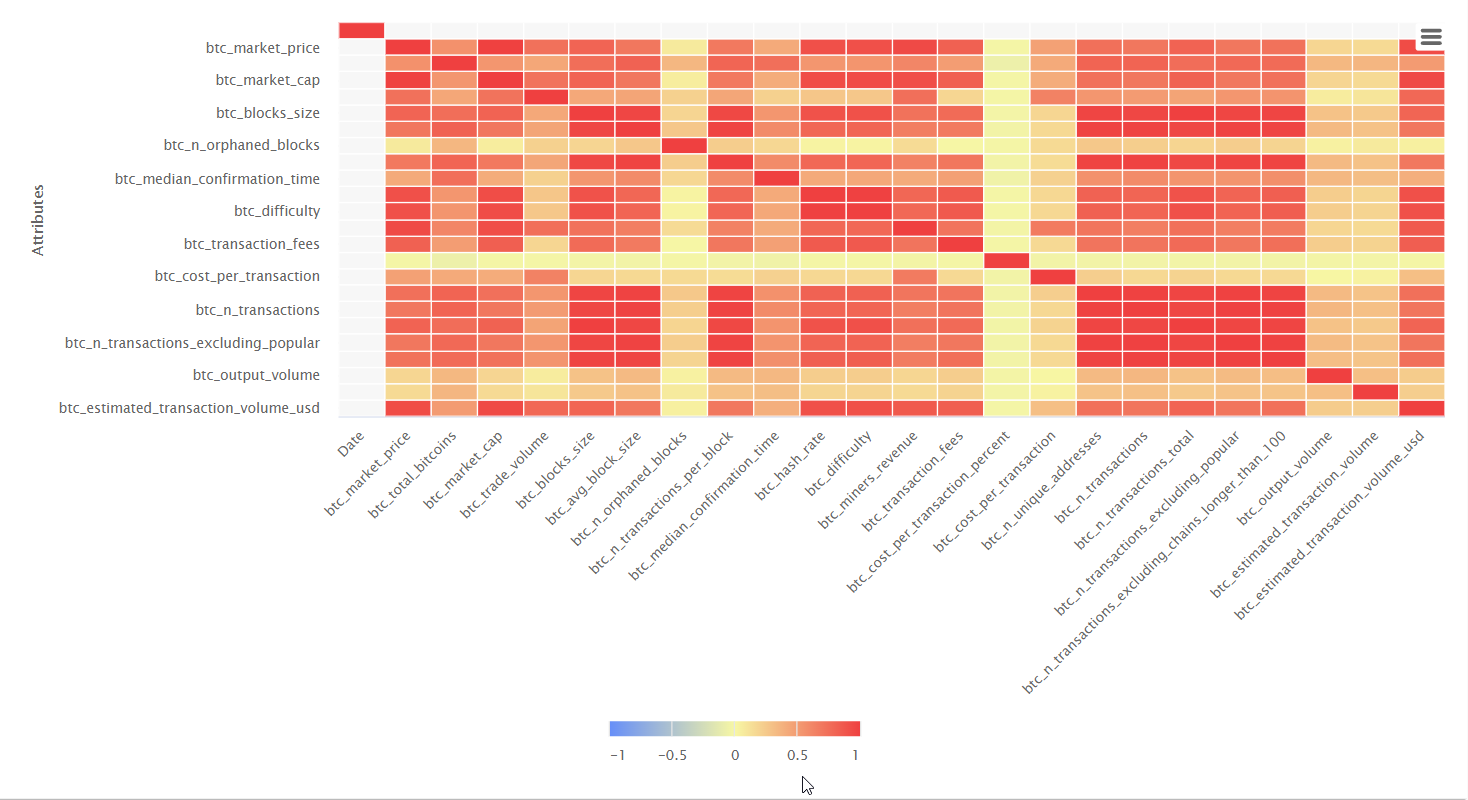


Figure 3 - Heatmap correlation Matrix

Figure 3 displays just how extremely highly correlated some attributes are, reinforcing the point we discussed earlier about redundant attributes, and the need for a dimenstionality reduction technique such as PCA.

A graph with blue lines

AI-generated content may be incorrect.

Figure 4 - Normalized Box plot of difficulty, hash rate, and market price

Figure 4 shows a normalized box plot of the 3 attributes to showcase the outliers and how the large scale attributes can severely skew the data.

This dataset is a time-series dataset, showcasing the evolution and volatility of the bitcoin cryptocurrency over time. Numerous attributes (such as market\_price and difficulty) display exponential growth in the later years, which goes hand in hand with the evolution in the financial technology sector. Within the attributes, there are large scale attributes, which are magnitudes bigger than others, portraying the need to normalize the data for better results.