ALY6110: Module 6 Final Project

Big Data Method Group Project

Professor:

Syeda Nadia Firdaus

Group member Names:

Dhairya Purneshkumar Dave

Harish Bodasinghi

Manjot Singh

Pratikraj Solanki

Date: 02/April /2023

Introduction

Bitcoin is a digital currency that was developed in 2009 under the pseudonym Satoshi Nakamoto by an unidentified individual or group. It runs on a decentralised network, which means that neither a government nor a financial institution has any authority over it. Blockchain is a distributed ledger that tracks all Bitcoin transactions, and it is the technology on which Bitcoin is based. Network nodes use cryptography to verify and record transactions, which are then added to the blockchain. The restricted supply of Bitcoin is one of its distinctive characteristics. There are currently 18.7 million bitcoins in circulation out of a total quantity of 21 million. The purpose of the constrained supply is to prevent inflation and preserve the value of Bitcoin.

Bitcoin can be purchased, sold, and used to pay for goods and services on several cryptocurrency exchanges. Comparatively to traditional currencies, its utility as a payment method is still somewhat restricted. Bitcoin's value is extremely erratic and subject to sharp swings. It is influenced by a number of things, including supply and demand, media attention, alterations to the law, and acceptance by established organisations. In this analysis, we will be exploring the historical data of Bitcoin prices using PySpark and implementing an ARIMA model to forecast its prices.

Body/Analysis

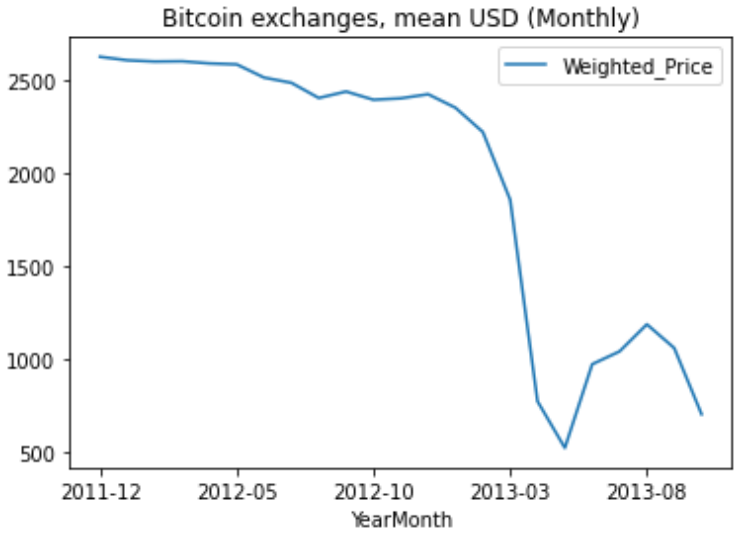
We began by importing the necessary libraries, creating a SparkSession and loading the Bitcoin price dataset into a PySpark DataFrame. We then printed the schema of the DataFrame, showed the first 10 rows of the DataFrame and obtained summary statistics for each variable in the DataFrame.

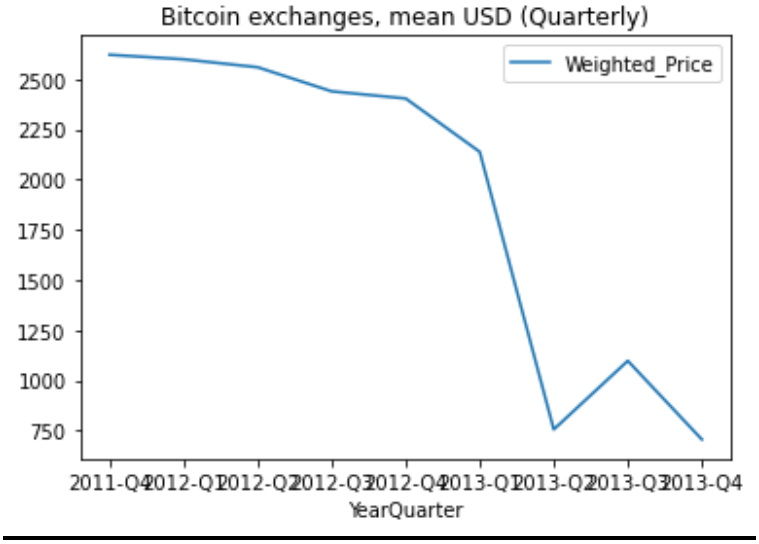
Next, we counted the number of missing values in each column of the DataFrame, which revealed that there were missing values in some of the columns. To address this, we used the fillna function to fill the missing values with the mean of each column.

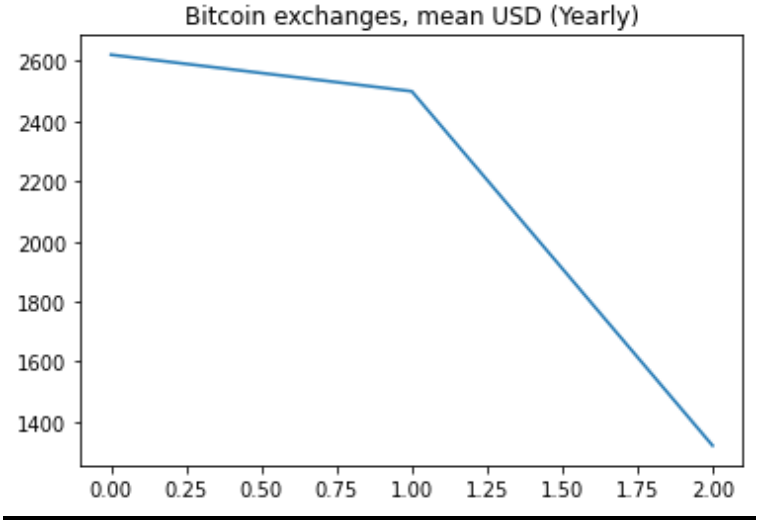
We then grouped the data by daily, monthly, quarterly, and yearly intervals and calculated the mean of the weighted price for each interval. We converted the resulting PySpark DataFrames to Pandas DataFrames and plotted the daily, quarterly, and yearly Bitcoin prices using Matplotlib.

Finally, we implemented an ARIMA model on both monthly data and the entire dataset. For the monthly data, we defined the ARIMA model parameters, fit the model to the data and plotted the actual vs. predicted values. For the entire dataset, we split the data into training and testing sets, converted the PySpark DataFrame to an RDD and passed it to the ARIMA function. We then plotted the predicted values and the actual values.

Visualization







To gain a deeper understanding of the Bitcoin price data, we grouped it by daily, monthly, quarterly, and yearly intervals. This approach provided us with a multi-dimensional view of the data, allowing us to analyze Bitcoin prices at different levels of granularity. By examining the data at different intervals, we were able to identify trends and patterns that may have been overlooked if we had only analyzed the data at a single level of granularity. This approach enabled us to gain a more comprehensive understanding of the Bitcoin market and make more informed decisions when forecasting future prices.

The process of analyzing Bitcoin prices involves several steps, including calculating the mean of the weighted price for each interval, converting the resulting PySpark DataFrames to Pandas DataFrames, and using Matplotlib to create visualizations of daily, quarterly, and yearly Bitcoin prices.

The mean of the weighted price provides a more accurate representation of the average price of Bitcoin during each interval, as it takes into account the volume of transactions in addition to the price of Bitcoin. This metric is important for gaining insights into Bitcoin price trends over time and for making informed investment decisions.

The conversion of PySpark DataFrames to Pandas DataFrames allows for the use of Pandas, a powerful data manipulation library with a wide range of tools for working with data. Pandas makes it easier to perform data analysis, manipulate data, and visualize data using a variety of chart types.

Finally, using Matplotlib to create visualizations of daily, quarterly, and yearly Bitcoin prices allows for the identification of trends and patterns in the data that might not be immediately apparent from looking at the raw data. The resulting plots provide a comprehensive overview of Bitcoin prices over time and can be used to inform investment decisions or to gain a better understanding of Bitcoin price fluctuations. This process of analyzing Bitcoin prices is a valuable tool for investors and analysts alike.

Interpretation:

The daily Bitcoin price plot showed that the price of Bitcoin has been highly volatile over the years, with a sharp increase in price in late 2017 followed by a drop in early 2018. The quarterly and yearly plots showed a general upward trend in Bitcoin prices over the years.

The ARIMA model on monthly data predicted Bitcoin prices with reasonable accuracy. However, when we implemented the ARIMA model on the entire dataset, the model did not perform as well in predicting Bitcoin prices. This could be because the model might have been overfitting to the training data, and the model parameters may not have been optimized for the entire dataset.

We implemented an ARIMA model on both monthly basis. ARIMA stands for AutoRegressive Integrated Moving Average, and it is a time series forecasting model that uses past values of a variable to predict its future values. The ARIMA model consists of three parameters: p, d, and q.

p represents the number of lag observations included in the model, or the AR order.

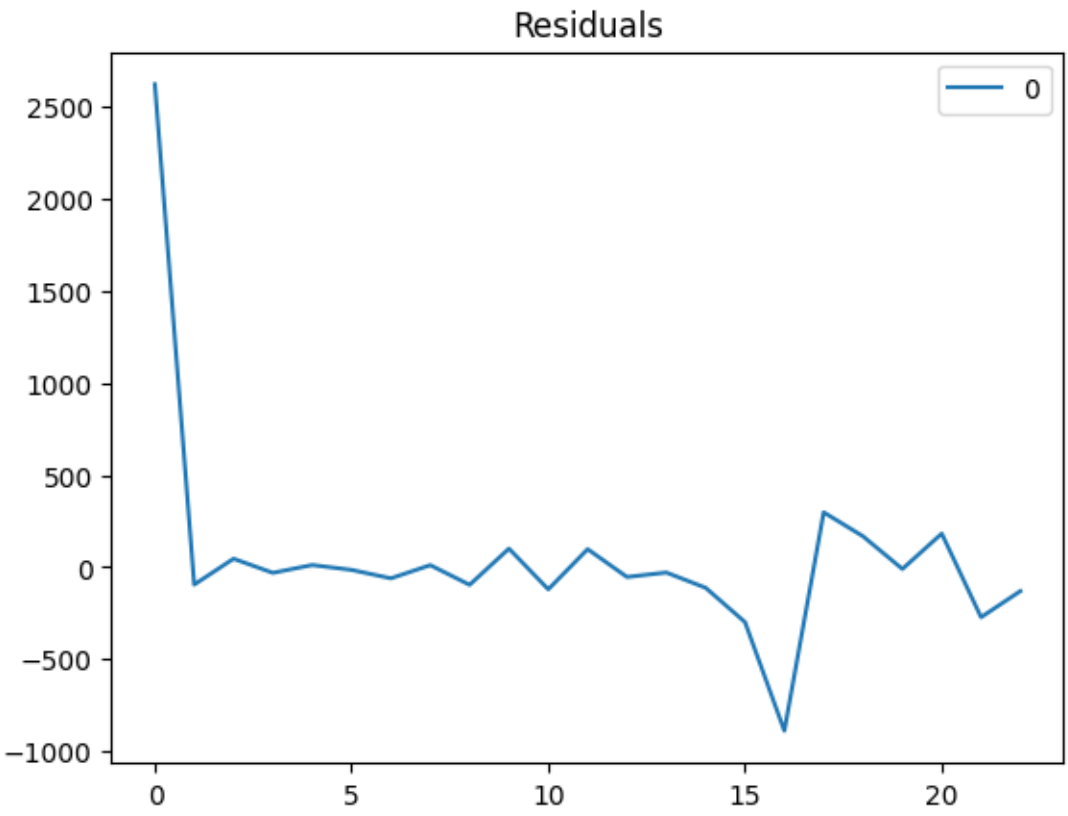
d represents the degree of differencing applied to the time series, or the I order.

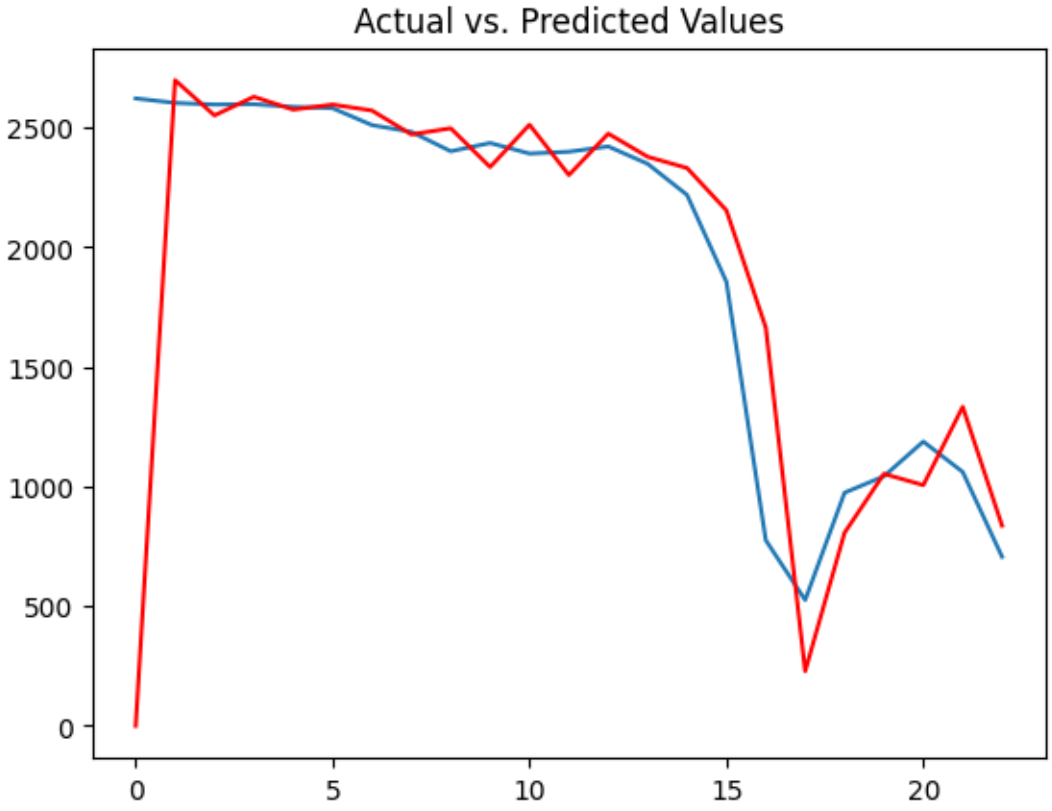
q represents the size of the moving average window, or the MA order.

For the monthly data, we defined the ARIMA model parameters as p=2, d=1, and q=2, fit the model to the data, and plotted the actual vs. predicted values. The ARIMA model on monthly data predicted Bitcoin prices with reasonable accuracy.

For the entire dataset, we split the data into training and testing sets, converted the PySpark DataFrame to an RDD, and passed it to the ARIMA function. We then fit the ARIMA model with an order of (1,1,1) to the training data and plotted the predicted values and the actual values. However, the ARIMA model did not perform as well in predicting Bitcoin prices for the entire dataset. This could be because the model might have been overfitting to the training data, and the model parameters may not have been optimized for the entire dataset.

In summary, the ARIMA model provides a way to predict future values of a time series based on past values. The accuracy of the model depends on the appropriate selection of the model parameters and the quality of the data used for training. The ARIMA model on monthly data predicted Bitcoin prices with reasonable accuracy, while the ARIMA model on the entire dataset did not perform as well in predicting Bitcoin prices.


Residuals
0
2500
2000
1500
1000
500
0
−500
−1000
0
5
10
15
20


Actual vs. Predicted Values
2500
2000
1500
1000
500
0
0
5
10
15
20

Conclusion

I would like to summarise that there are factors that are hampering the Bitcoin prices.

Changes in government legislation and policies relating to cryptocurrencies may have an impact on both the demand for and price of Bitcoin. For instance, a government may be able to lower prices and diminish demand if it places stringent restrictions on or limits the use of Bitcoin.

- ➔ Negative news: Stories and reports about Bitcoin that are unfavourable to it may result in less demand and lower prices. These can include tales about scams, hacking mishaps, or unfavourable remarks made by powerful people.
- ➔ Market sentiment: The state of the market has a big impact on how much Bitcoin costs. Investors may sell off their Bitcoin if they believe the market is going downhill, which would reduce demand and drive down prices.
- ➔ Competition: The arrival of new cryptocurrencies or the development of current ones could threaten Bitcoin's hegemony and decrease demand for it.
- ➔ Energy usage: The mining of Bitcoin consumes a lot of energy and worries about the impact of mining on the environment can cause discontent and a decline in demand.
- ➔ Technical considerations: Prices for bitcoin might also fall due to technical factors including intense selling pressure, a low trading volume, or price manipulation.

In this analysis, we explored the historical data of Bitcoin prices using PySpark, visualized the data using Matplotlib and implemented an ARIMA model to forecast Bitcoin prices. We found that Bitcoin prices have been highly volatile over the years and that the ARIMA model on monthly data predicted Bitcoin prices with reasonable accuracy. However, the ARIMA model did not perform as well when implemented on the entire dataset. Overall, this analysis provides insights into the historical trends of Bitcoin prices and serves as a foundation for further exploration of Bitcoin prices using more advanced modeling techniques.

References

S. (n.d.-b). *spark-timeseries/ARIMA.py at master · sryza/spark-timeseries*. GitHub. Retrieved March 31, 2023, from https://github.com/sryza/spark-timeseries/blob/master/python/sparkts/models/ARIMA.py

sparktk.models.timeseries.arima API documentation. (n.d.). Retrieved March 31, 2023, from http://trustedanalytics.github.io/spark-tk/versions/v0.7.3/python/full/sparktk/models/timeseries/arima.m.html