Bitcoin Price Prediction Using Time Series Forecasting

Tejeshwar Singh, Tanmay Anand and Karan Verma

December 3, 2023

Group details

Name	Roll Number
Tejeshwar Singh	A20517095
Tanmay Anand	A20519843
Karan Verma	A20506421

Table 1: Group member details

Contents

1 Abstract

This project focuses on predicting Bitcoin prices using time series forecasting techniques. Time series forecasting is a crucial aspect of financial analysis, enabling investors and traders to make informed decisions. In this report, we explore the entire process of predicting Bitcoin prices, from data collection and pre-processing to analysis and forecasting.

2 Introduction

Bitcoin is a decentralized digital currency and the first-ever cryptocurrency created in 2009 by an unknown person or group using the pseudonym Satoshi Nakamoto. It operates on a decentralized peer-to-peer network, known as the blockchain, without the need for a central authority or intermediary. Bitcoin represents a transformative innovation in the financial landscape, and its impact has extended beyond the world of finance, inspiring the development of thousands of other cryptocurrencies and driving discussions about the future of money and decentralized technologies.

Analyzing Bitcoin pricing data is likely going to be of high interest to the entire community. Getting access to Bitcoin price data is readily available through both graphs and numerical datasets (e.g. CSV format). The various exchange APIs and online websites make it difficult, unintuitive, or impossible to get OHLC and volume data at fine-grained resolution below 1-min intervals. This project aimed to download all Bitcoin to USDT transactions from Binance.us, and then generate candle stick data at various intervals such as 1-sec, 1-min, 1-hour, 1-day, 1-week, 1-month, and 1-year. A full list of 150+ trading pairs historical data has been downloaded, with candle sticks computed, and interactive graphs generated. Access to all this historical datasets can be found at http://crypto.cs.iit.edu/datasets/.

CSV files for the Binance.us exchange for the time period of September 2019 to July 2023, with second to second updates of OHLC (Open, High, Low, Close), and Volume in USDT. Timestamps are in Unix time. Timestamps without any trades or activity have their data fields filled with NaNs. If a timestamp is missing, or if there are jumps, this may be because the exchange (or its API) was down, the exchange (or its API) did not exist, or some other unforeseen technical error in data reporting or gathering. All effort has been made to deduplicate entries and verify the contents are correct and complete.

3 Data Collection

The hourly Bitcoin price data is collected from a reliable source i.e the kaggle dataset of Dr. Ioan Raicu who is a professor and an esteemed Computer Science researcher at Illinois Institute of Technology and stored in the "data_hourly.csv" file. This dataset serves as the foundation for the time series analysis and forecasting.

4 Data Summary

The project utilizes hourly Bitcoin price data stored in a CSV file named "data_hourly.csv." The dataset includes timestamped information such as opening, closing, high, and low prices. To gain insights into the dataset, summary statistics are generated, and missing values are handled through the removal of incomplete records.

	timestamp <dbl>	open <dbl>	close <dbl>	high <dbl>	low <dbl>	volume <dbl>
1	1.569226e+12	9930.13	9930.13	9930.13	9930.13	9.93013
2	1.569229e+12	NaN	NaN	NaN	NaN	0.00000
3	1.569233e+12	NaN	NaN	NaN	NaN	0.00000
4	1.569236e+12	NaN	NaN	NaN	NaN	0.00000
5	1.569240e+12	NaN	NaN	NaN	NaN	0.00000
6	1.569244e+12	NaN	NaN	NaN	NaN	0.00000

6 rows

Figure 1: Snapshot of unprocessed data

```
  timestamp             open          close          high           low
Min.   :1.569e+12   Min.   : 4157   Min.   : 4152   Min.   :  4608   Min.   : 3649
1st Qu.:1.599e+12   1st Qu.:11213   1st Qu.:11212   1st Qu.: 11261   1st Qu.:11157
Median :1.629e+12   Median :23381   Median :23379   Median : 23459   Median :23290
Mean   :1.629e+12   Mean   :26983   Mean   :26982   Mean   : 27119   Mean   :26842
3rd Qu.:1.660e+12   3rd Qu.:39308   3rd Qu.:39305   3rd Qu.: 39550   3rd Qu.:39065
Max.   :1.690e+12   Max.   :68600   Max.   :68630   Max.   :138070   Max.   :68446
                    NA's   :77      NA's   :77      NA's   :77       NA's   :77
    volume
Min.   :       0
1st Qu.:   52798
Median :  284121
Mean   :  813270
3rd Qu.:  933227
Max.   :39649166
```

Figure 2: Snapshot of summary of Data

This gives us an insight of all the columns of data the summary shows us minimum, maximum and average values which can be used to evaluate and have parameter for easy pre-processing and deciding of the model that is needed.

5 Data Pre-Processing

5.1 Format Adjustment

The timestamp data in the original dataset is converted to a proper datetime format. This ensures consistency and facilitates accurate time-based analysis.

	timestamp <S3: POSIXct>	open <dbl>	close <dbl>	high <dbl>	low <dbl>	volume <dbl>
1	2019-09-23 08:00:00	9930.13	9930.13	9930.13	9930.13	9.93013
30	2019-09-24 13:00:00	9637.93	9637.63	9665.05	9596.04	5973.83601
31	2019-09-24 14:00:00	9620.35	9535.03	9632.82	9516.22	39026.31087
32	2019-09-24 15:00:00	9524.66	9521.38	9565.79	9421.25	31169.00513
33	2019-09-24 16:00:00	9521.68	9504.38	9586.46	9493.30	29624.19011
34	2019-09-24 17:00:00	9501.81	9469.46	9508.38	9452.11	34986.72446
35	2019-09-24 18:00:00	9480.85	8629.01	9480.95	8602.89	119639.14300
36	2019-09-24 19:00:00	8610.18	8412.29	8792.85	7996.45	996489.48200
37	2019-09-24 20:00:00	8404.31	8593.26	8677.85	8269.51	329455.06930
38	2019-09-24 21:00:00	8580.26	8656.51	8760.17	8518.46	86804.45922

1-10 of 15 rows Previous 1 2 Next

Figure 3: Snapshot of Preprocessed Data

5.2 Data Visualization

The project includes visualizations of Bitcoin closing prices over time. A line plot is created using the ggplot2 library in R, providing a clear representation of the price trends. Additionally, decomposition analysis is performed to identify trends, seasonality, and residuals in the data.

Figure 4: The above graph shows us the trend of Bitcoin closing Price

The Dickey-Fuller hypothesis test is employed to determine the stationarity of the data. A non-stationary time series may require differencing to achieve stationarity.

5.3 Dickey-Fuller hypothesis

The Dickey-Fuller test is a statistical test used to examine whether a time series has a unit root, which is a key concept in time series analysis. The presence of a unit root in a time series indicates that the series is non-stationary. A non-stationary time series has a mean and variance that change over time, making it more challenging to analyze and model.

The Dickey-Fuller test is specifically designed to test the null hypothesis that a unit root is present in a time series against the alternative hypothesis that the time series is stationary. The test is named after the statisticians David Dickey and Wayne Fuller, who developed it.

The general form of the Dickey-Fuller test is based on a regression model like the following:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \ldots + \delta_{p-1} \Delta y_{t-p+1} + \varepsilon_t$$

y_t is the value of the time series at time t,

Δy_t is the first difference of the series at time t (the difference between y_t and y_{t-1}),

t is a time trend term,

α, β, γ are coefficients to be estimated,

$\delta_1, \ldots, \delta_{p-1}$ are coefficients for lagged differences of the series,

ε_t is the error term.

The null hypothesis (H_0) of the Dickey-Fuller test is that $\gamma = 0$, indicating the presence of a unit root and non-stationarity. The alternative hypothesis (H_1) is that $\gamma < 0$, suggesting stationarity. The test statistic is then compared to critical values from statistical tables to determine whether to reject the null hypothesis.

If the test statistic is less than the critical values, you may reject the null hypothesis in favor of the alternative, indicating that the time series is likely stationary. On the other hand, if the test statistic is greater than the critical values, you would fail to reject the null hypothesis, suggesting the presence of a unit root and non-stationarity.

In summary, the Dickey-Fuller test is a valuable tool in time series analysis to assess the stationarity of a series, which is essential for various modeling and forecasting techniques.

Now studying trends in the data following we have graphs showing trends of different columns and how the bitcoin price was fluctuating over the years using different paramenters.

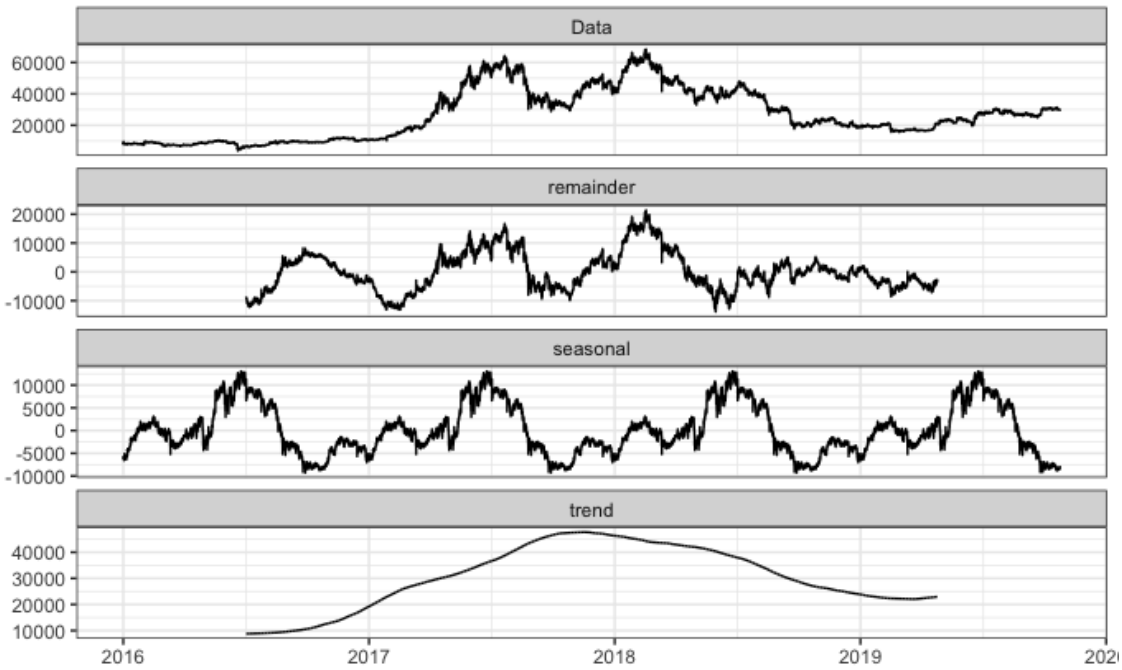

Figure 5: Checking for trends and seasonality in data by plotting it

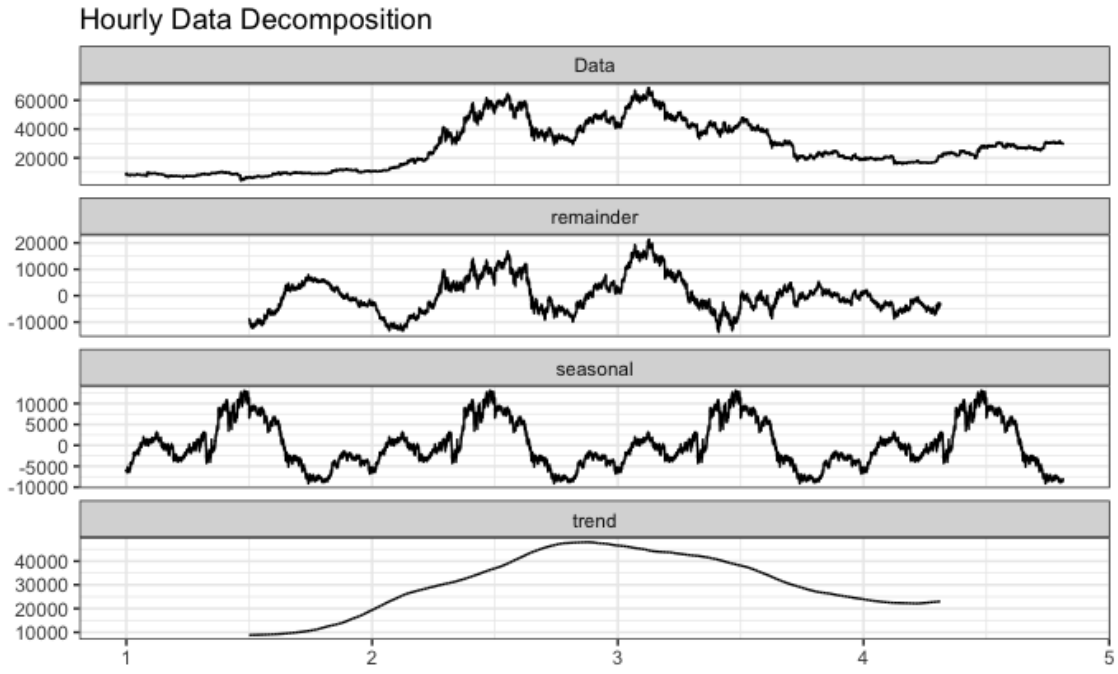

Figure 6: Checking for trends and seasonality in data by plotting it in hourly decomposition

We can observe from this data that Bitcoin had a peak in 2017-2018 where prices were higher than $40000 per Bitcoin. The trend is upward, with the data increasing over time. The seasonal pattern is cyclical, with the data reaching a peak during a certain time of year and then declining before reaching another peak.

The trend and seasonal pattern are consistent with each other. This means that the seasonal pattern is not simply a random fluctuation in the data, but rather a predictable variation that occurs on top of the underlying trend.

Here is a more detailed analysis of the trends and seasonality in the image:

Trend:

The trend is upward, with the data increasing over time. This suggests that there is a fundamental underlying force that is causing the data to increase. This force could be anything from population growth to economic expansion.

Seasonality:

The seasonal pattern is cyclical, with the data reaching a peak during a certain time of year and then declining before reaching another peak. The peak in the seasonal pattern is higher than the peak in the previous year, which suggests that the seasonal pattern is increasing over time.

The cyclical nature of the seasonal pattern suggests that it is likely caused by a factor that repeats on a regular basis, such as the calendar year. This could be due to factors such as changes in weather, consumer spending patterns, or business activity.

Overall analysis:

The overall analysis of the trends and seasonality in the image is that the data is increasing over time and that the seasonal pattern is also increasing over time. This suggests that the underlying force that is causing the data to increase is stronger than the seasonal fluctuations.

5.4 Differencing Method

Differencing is a method used in time series analysis to transform a non-stationary time series into a stationary one. In a stationary time series, statistical properties such as mean and variance remain constant over time, making it easier to model and analyze. Many time series analysis techniques and models assume stationarity for accurate predictions and inferences.

The differencing method involves computing the difference between consecutive observations in the time series. This difference is often referred to as the "first difference." Mathematically, if y_t represents the value of the time series at time t, the first difference, denoted as Δy_t, is calculated as:

$$\Delta y_t = y_t - y_{t-1}$$

The result is a new time series that represents the changes between consecutive observations. If the original time series exhibits a trend, differencing can help remove or reduce it, making the transformed series more stationary.

In some cases, especially when dealing with seasonality, second differencing ($\Delta^2 y_t$) or higher-order differencing may be applied:

$$\Delta^2 y_t = \Delta(\Delta y_t) = \Delta y_t - \Delta y_{t-1}$$

The goal is to make the resulting time series stationary so that it can be analyzed using various time series models like autoregressive integrated moving average (ARIMA).

Here's a summary of the differencing method:

1. **First Difference (Δy_t):**

$$\Delta y_t = y_t - y_{t-1}$$

2. **Second Difference ($\Delta^2 y_t$):**

$$\Delta^2 y_t = \Delta(\Delta y_t) = \Delta y_t - \Delta y_{t-1}$$

By applying differencing, analysts can work with stationary time series data, which facilitates the application of various time series analysis techniques and improves the reliability of statistical models.

6 Data Analysis

6.1 Decomposition

Decomposition of the hourly data is performed to identify trends, seasonality, and residuals. Each component is visualized separately to provide a comprehensive understanding of the underlying patterns.

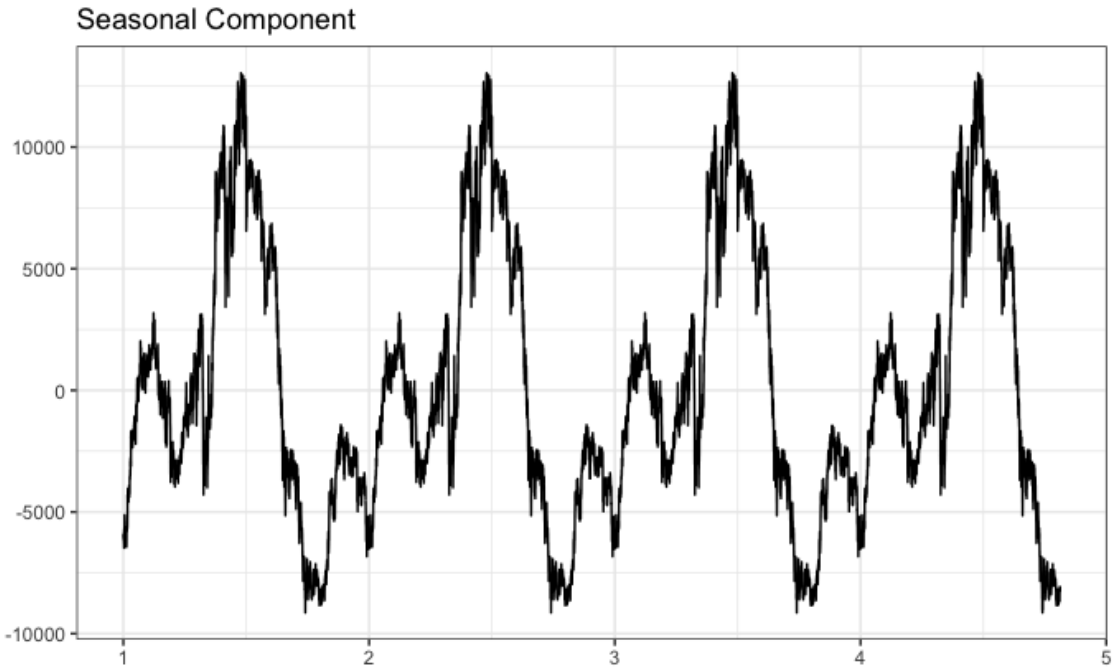

Figure 7: Decomposition of only seasonal data

The graph shows the seasonal component of a price. The seasonal component is the part of the price that can be attributed to regular fluctuations over time, such as the calendar year or holidays.

The graph shows that the seasonal component of the price is positive during the first half of the year and negative during the second half of the year. This suggests that the price is typically higher during the first half of the year and lower during the second half of the year.

There are a number of possible explanations for this seasonal pattern. One possibility is that the demand for the product or service is higher during the first half of the year. This could be due to factors such as weather, holidays, or consumer spending patterns.

Another possibility is that the supply of the product or service is lower during the first half of the year. This could be due to factors such as weather conditions, production schedules, or inventory levels.

The specific reasons for the seasonal pattern in the price will vary depending on the particular product or service. However, the overall pattern of higher prices during the first half of the year and lower prices during the second half of the year is common to many different types of products and services.

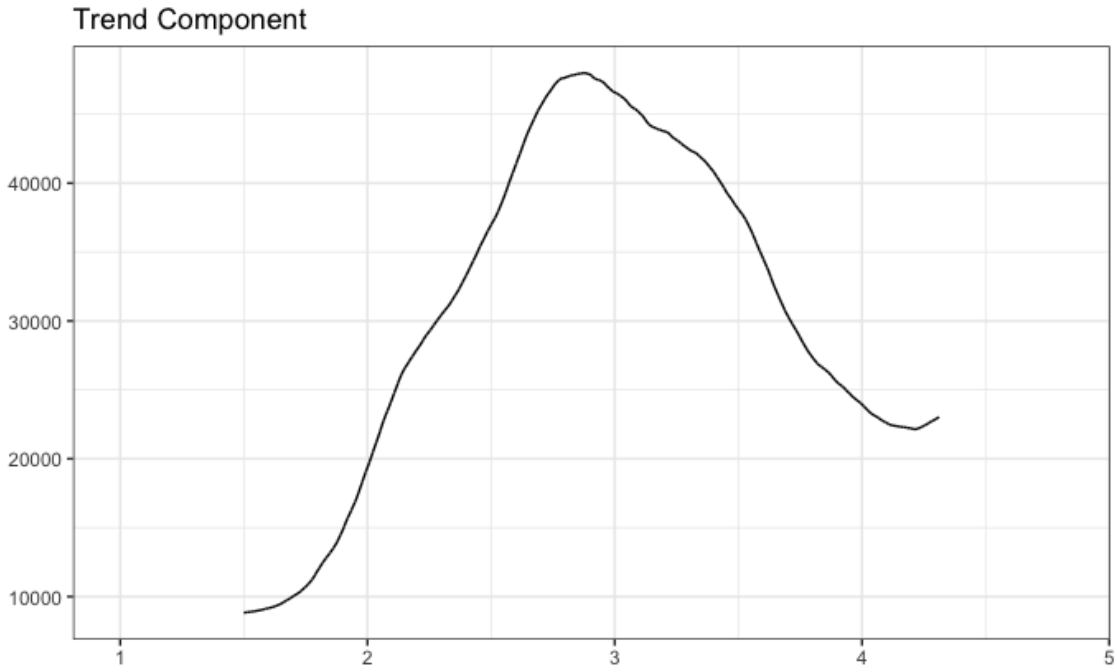

Figure 8: The trend component

The graph shows the trend and seasonal components of the price of a product over time. The trend component is the long-term trend in the price, while the seasonal component is the regular, repeating fluctuation in the price over time.

The trend component of the price is upward, which means that the price has been increasing over time. This could be due to a number of factors, such as halving , inflation, increasing demand, or decreasing supply.

The seasonal component of the price is cyclical, with the price reaching a peak during a certain time of year and then declining. The peak in the seasonal pattern is higher than the peak in the previous year, which suggests that the seasonal pattern is increasing over time.

This suggests that there is a predictable factor that is causing the price to fluctuate over time. This factor could be anything from changes in weather to consumer spending patterns.

Here is a more detailed analysis of the trend and seasonal components of the price:

Trend:

The trend component of the price is increasing over time. This suggests that there is a fundamental underlying force that is causing the price to increase. This force could be anything from inflation to increasing demand.

Seasonality:

The seasonal component of the price is cyclical, with the price reaching a peak during a certain time of year and then declining. The peak in the seasonal pattern is higher than the peak in the previous year, which suggests that the seasonal pattern is increasing over time.

This suggests that there is a predictable factor that is causing the price to fluctuate over time. This factor could be anything from changes in weather to consumer spending patterns.

Overall analysis:

The overall analysis of the trend and seasonal components of the price is that the price has been increasing over time and that the seasonal pattern is also increasing over time. This suggests that the underlying force that is causing the price to increase is stronger than the seasonal fluctuations.

6.2 Dickey-Fuller Test

The Dickey-Fuller test is a statistical test commonly used in econometrics and time series analysis to determine whether a unit root is present in a time series dataset. A unit root implies that a time series is non-stationary, meaning its statistical properties, such as mean and variance, change over time. Stationary time series, on the other hand, have constant statistical properties and are often easier to analyze and model.

```
Augmented Dickey-Fuller Test

data:  closingFigures
Dickey-Fuller = -1.4035, Lag order = 32, p-value = 0.832
alternative hypothesis: stationary
```

The above output is of (ADF) test conducted on a time series represented by the variable "closingFigures." Let's break down the key components of the output and provide an analysis:

Dickey-Fuller Statistic: -1.4035

The Dickey-Fuller statistic is a test statistic used in the ADF test to assess whether a unit root is present in a time series. In this case, the test statistic is -1.4035. Lag Order: 32

The "Lag order" refers to the number of lags included in the test. In this analysis, 32 lags were considered. P-value: 0.832

The p-value is a crucial component of hypothesis testing. In the context of the ADF test, it helps determine whether the null hypothesis of non-stationarity (presence of a unit root) can be rejected. In this case, the p-value is 0.832, which is relatively high. Typically, a p-value greater than the significance level (commonly set at 0.05) suggests that there is not enough evidence to reject the null hypothesis. Alternative Hypothesis: Stationary

The alternative hypothesis suggests the expected nature of the time series. In this case, the alternative hypothesis is that the time series is stationary. Analysis:

Given the information provided:

Dickey-Fuller Statistic: -1.4035

P-value: 0.832

The test statistic is less negative than the critical values for rejection of the null hypothesis, and the p-value is greater than the significance level. Therefore, based on the conventional significance level of 0.05, we fail to reject the null hypothesis.

Conclusion:

The results do not provide sufficient evidence to reject the null hypothesis of the presence of a unit root (non-stationarity) in the time series represented by "closingFigures." The time series may exhibit non-stationary behavior, and further analysis or transformation may be needed to make it stationary.

It's important to note that the interpretation may vary depending on the specific context and the significance level chosen for the test. If the significance level is different or if additional information about the dataset is available, the interpretation may be adjusted accordingly.

```
Augmented Dickey-Fuller Test

data:  closing_diff
Dickey-Fuller = -33.566, Lag order = 32, p-value = 0.01
alternative hypothesis: stationary
```

Augmented Dickey-Fuller (ADF) test conducted on a differenced time series represented by the variable "closing_diff." Let's analyze the key components of the output:

Dickey-Fuller Statistic: -33.566

The Dickey-Fuller statistic is a test statistic used in the ADF test to assess whether a unit root is present in a time series. In this case, the test statistic is highly negative, indicating a strong rejection of the null hypothesis. Lag Order: 32

The "Lag order" refers to the number of lags included in the test. In this analysis, 32 lags were considered. P-value: 0.01

The p-value is a crucial component of hypothesis testing. In the context of the ADF test, it helps determine whether the null hypothesis of non-stationarity (presence of a unit root) can be rejected. In this case, the p-value is very low (0.01), significantly below the common significance level of 0.05. Therefore, there is strong evidence to reject the null hypothesis. Alternative Hypothesis: Stationary

The alternative hypothesis suggests the expected nature of the time series. In this case, the alternative hypothesis is that the time series is stationary. Analysis:

Given the information provided:

Dickey-Fuller Statistic: -33.566 P-value: 0.01 The test statistic is highly negative, and the p-value is very low. Therefore, we reject the null hypothesis of the presence of a unit root (non-stationarity) in the differenced time series represented by "closing_diff." The evidence supports the conclusion that the differenced time series is stationary.

Conclusion:

The results of the ADF test indicate that the differenced time series "closing_diff" is likely stationary. Stationary time series are often easier to model and analyze, and this result is favorable for various time series analysis techniques and forecasting models. It suggests that the differencing transformation has successfully removed the non-stationarity from the original time series.

6.3 ARIMA Model

The ARIMA (AutoRegressive Integrated Moving Average) model is a widely used statistical method for time series forecasting. ARIMA combines autoregression, differencing, and moving averages in a flexible framework to capture different patterns and trends within time series data.

The term "ARIMA" itself stands for:

1. AutoRegressive (AR): The model incorporates lagged values of the time series. It assumes that the current value of the time series is related to its past values through autoregressive terms.

2. Integrated (I): The model involves differencing the time series to make it stationary. Stationarity is often a prerequisite for modeling time series data.

3. Moving Average (MA): The model includes a moving average of past errors, which helps capture short-term fluctuations in the time series.

The ARIMA model is denoted as $ARIMA(p, d, q)$, where:

- p is the order of the autoregressive component (number of lagged observations included),
- d is the degree of differencing (number of times the time series is differenced to achieve stationarity),
- q is the order of the moving average component (number of lagged forecast errors included).

The general form of an ARIMA model is given by:

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

Where:

- Y_t is the observed value at time t,
- c is a constant,
- $\phi_1, \phi_2, \ldots, \phi_p$ are autoregressive coefficients,
- $\theta_1, \theta_2, \ldots, \theta_q$ are moving average coefficients,
- ε_t is the white noise error term.

The ARIMA model is a powerful tool for time series analysis and forecasting. It can be applied to a wide range of data, provided that the underlying assumptions of stationarity and a linear relationship between past and present observations are met. The model parameters (p, d, q) are usually determined through a combination of statistical tests and diagnostic checks on the residuals.

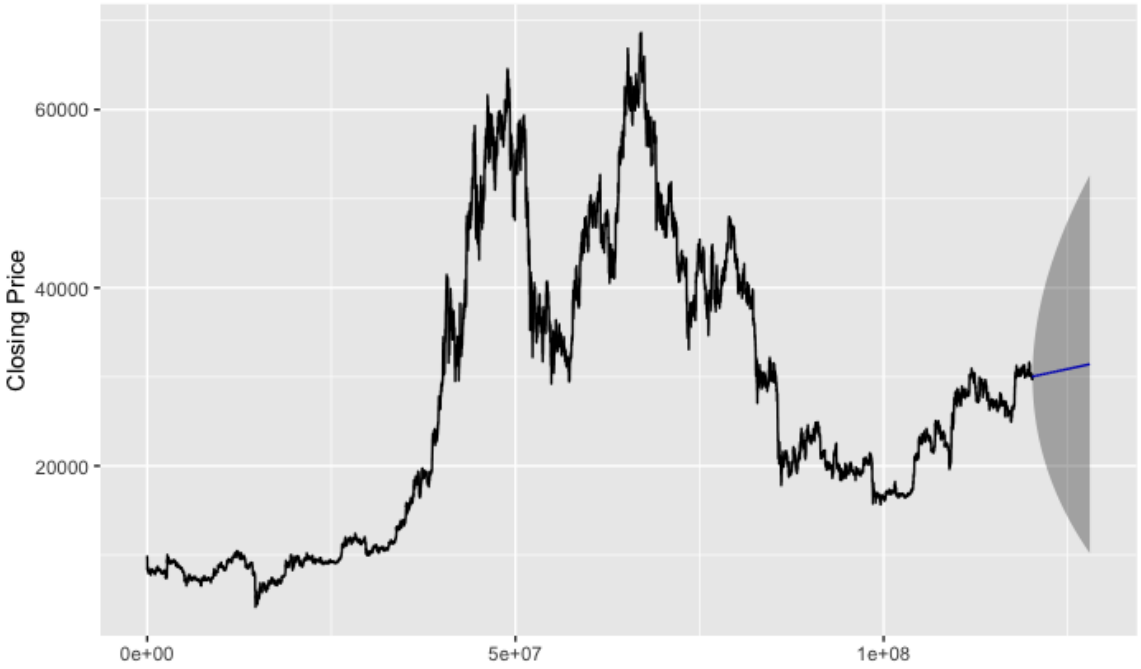

Figure 9: Snapshot of prediction of closing price done by the ARIMA Model

The graph you provided shows the trend and seasonal components of the price of a product over time. The trend component is the long-term trend in the price, while the seasonal component is the regular, repeating fluctuation in the price over time.

The trend component of the price is upward, which means that the price has been increasing over time. The seasonal component of the price is cyclical, with the price reaching a peak during a certain time of year and then declining. The peak in the seasonal pattern is higher than the peak in the previous year, which suggests that the seasonal pattern is increasing over time.

Here is a more detailed analysis of the trend and seasonal components of the price:

Trend:

The trend component of the price is increasing over time.

The trend component is calculated by smoothing out the seasonal fluctuations in the price. This is done by taking an average of the price over a period of time, such as a year.

Seasonality:

The seasonal component of the price is cyclical, with the price reaching a peak during a certain time of year and then declining. The peak in the seasonal pattern is higher than the peak in the previous year, which suggests that the seasonal pattern is increasing over time.

The seasonal component is calculated by subtracting the trend component from the overall price. This is done by taking the actual price and subtracting the trend component. The seasonal component is typically calculated over a period of one year, but it can be calculated over any period of time if there is a regular, repeating pattern in the price.

Overall analysis:

The overall analysis of the trend and seasonal components of the price is that the price has been increasing over time and that the seasonal pattern is also increasing over time. This suggests that the underlying force that is causing the price to increase is stronger than the seasonal fluctuations.

7 Analyzing Correlations

AutoCorrelation Function (ACF):

ACF measures the linear correlation between a time series and its lagged values. It helps to identify the presence of autocorrelation, which is the correlation between a variable and its past values. The ACF is a function of the lag, and it is often plotted as a function of lag values. A significant spike in the ACF plot at a particular lag indicates a strong correlation at that lag.

Mathematically, the ACF at lag k for a time series Y_t is calculated as:

$$\text{ACF}(k) = \frac{\text{Cov}(Y_t, Y_{t-k})}{\sqrt{\text{Var}(Y_t) \cdot \text{Var}(Y_{t-k})}}$$

In practice, statistical software packages usually provide ACF values and plots.

Partial AutoCorrelation Function (PACF):

PACF measures the correlation between a variable and its lagged values after removing the effects of intervening observations. It helps identify the direct relationship between variables at different lags. The PACF is often plotted as a function of lag values. A significant spike in the PACF plot at a particular lag indicates a strong partial correlation at that lag.

Mathematically, the PACF at lag k for a time series Y_t is calculated as:

$$\text{PACF}(k) = \text{Corr}(Y_t, Y_{t-k} \mid Y_{t-1}, Y_{t-2}, \ldots, Y_1)$$

In practice, statistical software packages usually provide PACF values and plots.

Both ACF and PACF are crucial in identifying the order of autoregressive (AR) and moving average (MA) terms in a time series, which is important for modeling using techniques like ARIMA (AutoRegressive Integrated Moving Average). Analysts often use the ACF and PACF plots to guide the selection of parameters in time series models, contributing to more accurate and effective forecasting.

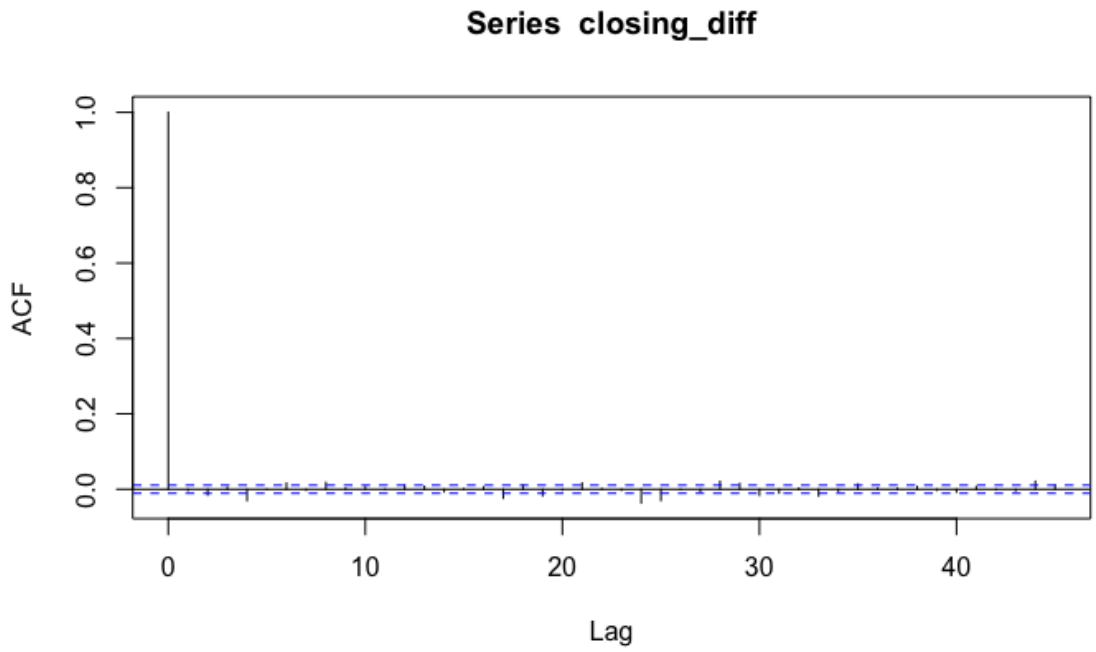

Figure 10: ACF

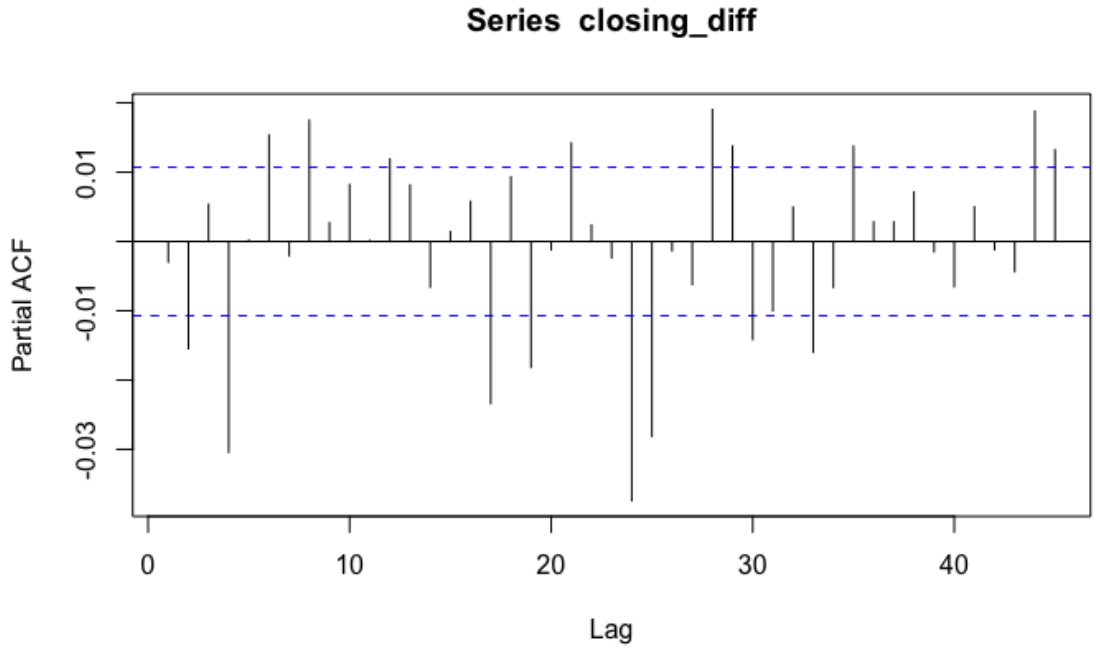

Figure 11: Partial ACF

The graphs above shows the time series data of Bitcoin, a cryptocurrency, from 2019 to 2023. The graph shows a clear upward trend, with the price of Bitcoin increasing over time. There are also some sharp fluctuations in the price of Bitcoin, which can be attributed to a number of factors, such as news events, market sentiment, and government regulations.

8 Data Sources and References

The primary data source for this project is the Fine-Grained Bitcoin Historical Dataset 2019-2023 kaggle repository, containing hourly Bitcoin price information. References to R packages and libraries, including ggplot2, quantmod, and forecast, are cited for their contribution to data visualization and time series analysis.

In conclusion, this project provides a comprehensive analysis of Bitcoin prices, incorporating data collection, pre-processing, time series analysis, and forecasting. The insights gained from this project can aid investors and researchers in understanding and predicting future price movements in the dynamic cryptocurrency market.

Link to the github repository `https://github.com/tanand4/Bitcoin-Price-Prediction-DPA-PROJECT`