Time Series Analysis on Stock Market Data Using ARIMA

Anish Gillella[1], Devansh Pratap Singh[2], Dimple Neeluri Chandrasekhar[3], Karthik Kotam[4], Krishna Venkatesan[5], Kushimithaa Thimmaareddy[6], Medha Priyanga Saravanan[7], Sarthak Vajpayee[8], and Yash Agrawal[9]

[1]AXG220089, The University of Texas at Dallas, Richardson, TX - 75080
[2]DPS220001, The University of Texas at Dallas, Richardson, TX - 75080
[3]DXN210024, The University of Texas at Dallas, Richardson, TX - 75080
[4]KXK210073, The University of Texas at Dallas, Richardson, TX - 75080
[5]KXV220007, The University of Texas at Dallas, Richardson, TX - 75080
[6]KXT220026, The University of Texas at Dallas, Richardson, TX - 75080
[7]MXS220057, The University of Texas at Dallas, Richardson, TX - 75080
[8]SXV220020, The University of Texas at Dallas, Richardson, TX - 75080
[9]YXA220007, The University of Texas at Dallas, Richardson, TX – 75080

Abstract

This research focuses on developing predictive time series models for analyzing historical stock prices of 29 DJIA companies over the past 12 years, excluding 'V' (Visa). The dataset, a valuable resource for exploring stock market dynamics, is accompanied by diverse formats catering to various analytical needs. Despite the inherent complexity in handling vast financial data, the project presents an opportunity for substantial financial gains through successful model development. A dedicated GitHub repository houses scripts for data acquisition and continually updated modeling codes, ensuring accessibility and transparency in the research process. By leveraging this comprehensive dataset and repository, the study aims to address challenges in forecasting stock market trends, offering insights that contribute to effective decision-making in the dynamic realm of financial markets.

Keywords: Stock Market Forecasting, Time Series, ARIMA

1. Literature Review

This research endeavors to develop predictive time series models for the comprehensive analysis of historical stock prices encompassing 29 DJIA companies over the past 12 years, with Visa ('V') excluded from the dataset. The dataset proves to be a valuable resource for exploring the intricacies of stock market dynamics, providing diverse formats that cater to a range of analytical needs. Despite the inherent complexity associated with handling vast financial data, this project represents a promising opportunity for significant financial gains through the successful development of predictive models. The dedicated GitHub repository further enhances accessibility and transparency in the research process by housing scripts for data acquisition and continually updated modeling codes.

The dataset's rich structure encompasses key columns such as Date, Open, High, Low, Close, Volume, and Name. The 'Date' column captures the timeline in yy-mm-dd format, offering a detailed account of stock price movements over time. Meanwhile, 'Open,' 'High,' 'Low,' and 'Close' provide comprehensive information on the daily dynamics of stock prices, facilitating in-depth analysis.

Notably, the dataset is designed to accommodate diverse analytical needs, offering both extended (13 years) and condensed (past year) versions. Additionally, the data is uniformly formatted in USD for NYSE, promoting consistency and ease of analysis. The versatility of the dataset provides researchers and analysts with the flexibility to choose the temporal scope that aligns with their specific research objectives.

Acknowledging the collaborative nature of data science, this research adapts data from the 'S&P 500 Stock data' available on platforms like Kaggle and GitHub. The dataset is sourced from Google finance using the 'pandas_datareader' Python library, and the research extends gratitude to Kaggle, GitHub, and the broader market for their contributions to data accessibility.

The literature review contextualizes the research within the broader landscape of time series analysis, highlighting its significance in diverse fields such as economics, finance, engineering, and social sciences. The Comparative Study on Time Series Analysis by Salah et al. evaluates classical methods against machine learning algorithms for demand forecasting, providing nuanced insights into their respective strengths and weaknesses. Works such as "Forecasting: Principles and Practice" by Rob J. Hyndman and George Athanasopoulos, specifically Chapter 8 on ARIMA models, along with Kuvshinov, D., and Zimmermann, K.'s exploration of stock market capitalization impact, offer valuable perspectives on time series analysis. Furthermore, insights from "The Elements of Statistical Learning" by Hastie, T., Friedman, J., and Tisbshirani, specifically in Chapter 7, enrich our understanding of statistical learning, data mining, and prediction. Collectively, these studies contribute to a robust foundation for the research, emphasizing the need for a comprehensive understanding of both classical and machine learning-based algorithms in demand forecasting within the realm of time series analysis.

2. Data

The dataset under consideration boasts a robust structure, featuring key columns such as Date, Open, High, Low, Close, Volume, and Name. The 'Date' column meticulously captures the timeline in yy-mm-dd format, serving as a chronological record of stock price movements. Meanwhile, the 'Open,' 'High,' 'Low,' and 'Close' columns offer in-depth insights into the daily dynamics of stock prices, providing a comprehensive view of market trends and fluctuations. The 'Volume' column adds another layer of information, indicating the trading activity associated with each recorded data point, and the 'Name' column allows for clear identification of the specific stock under analysis.

This dataset caters to diverse analytical needs by offering both extended (13 years) and condensed (past year) versions. The temporal flexibility allows researchers and analysts to align the data with their specific research objectives. Moreover, the dataset is uniformly formatted in USD, specifically tailored for NYSE data, ensuring consistency and facilitating ease of analysis.

Acknowledging the collaborative nature of data science, this dataset is adapted from the 'S&P 500 Stock data' available on platforms like Kaggle and GitHub. The data is sourced from Google finance using the 'pandas_datareader' Python library, and sincere gratitude is extended to Kaggle, GitHub, and the broader market for their contributions to data accessibility. This acknowledgment underscores the collaborative and open nature of data sharing, fostering a culture of knowledge exchange within the research community.

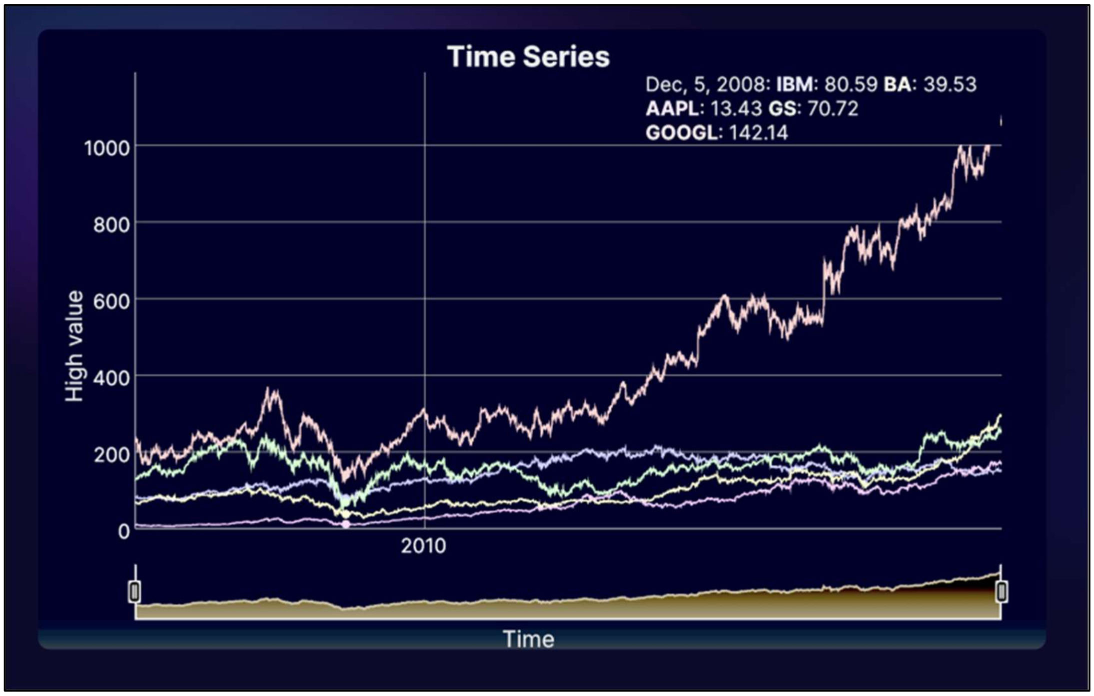
3. Empirical Method

First, let's break down these concepts in simpler terms. In the world of time series analysis, we use two main types of models: AutoRegressive (AR) and Moving Average (MA). Imagine you're tracking the temperature every day. An AutoRegressive (AR) model looks at how today's temperature is related to the temperatures on previous days. If it's an AR(p) model, it means we're considering the influence of the past p days.

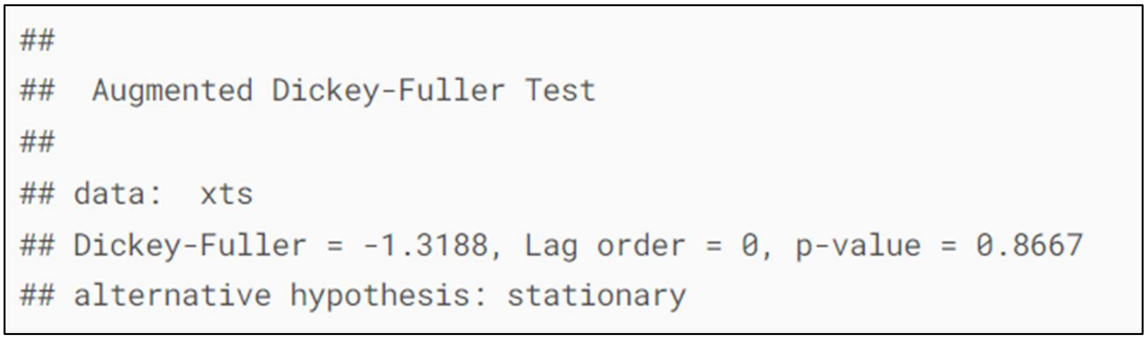
Now, let's talk about Moving Average (MA). This model thinks about the changes from day to day. If today is warmer or cooler than what we predicted based on the previous day, MA(q) takes that difference into account, considering the residuals from the last q days. Now, here's where it gets interesting. The ARMA model is like a combo of AR and MA. It's like having a buddy that looks at both the temperature trend over several days (AR part) and how much it deviates from the prediction based on recent days (MA part). So, ARMA(p,q) is just combining the insights from the last p days for the trend and the last q days for the deviations.

In a nutshell, these models help us make sense of patterns in data over time. AR focuses on past values affecting today, MA looks at unexpected changes, and ARMA combines both for a more complete picture. It's like having different lenses to see and understand the story that time series data is trying to tell us.

There are 5 time series in the data provided - (High, Low, Open, Close, Volume). We will look at the High values first.



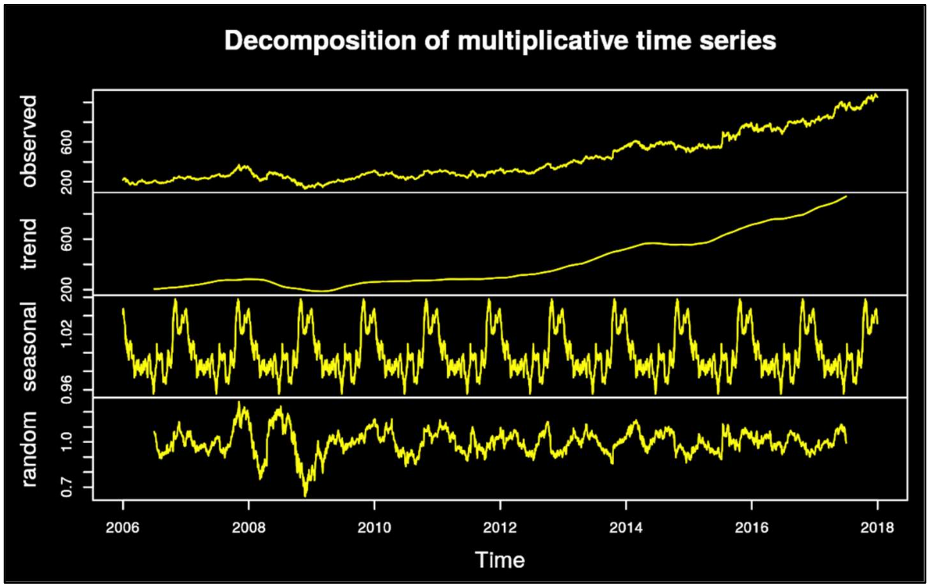
Next, we looked at stationarity in the data. A stationary time series is characterized by a consistent mean and variance over time, devoid of any discernible trend. Judging by the appearance, the mentioned time series does not exhibit stationarity. To formally confirm this, the stationarity of the time series was assessed using the Dickey-Fuller test.

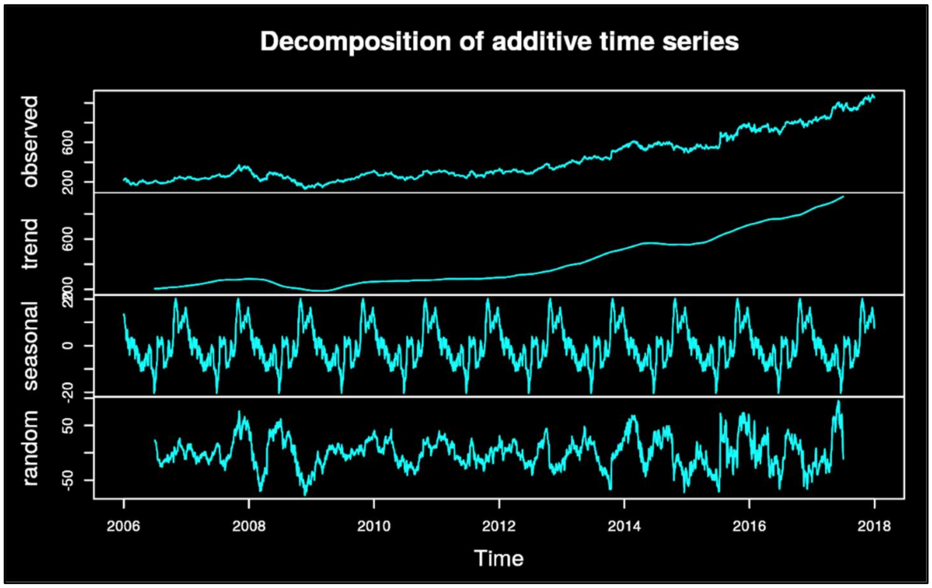


```
##
##  Augmented Dickey-Fuller Test
##
## data:  xts
## Dickey-Fuller = -1.3188, Lag order = 0, p-value = 0.8667
## alternative hypothesis: stationary
```

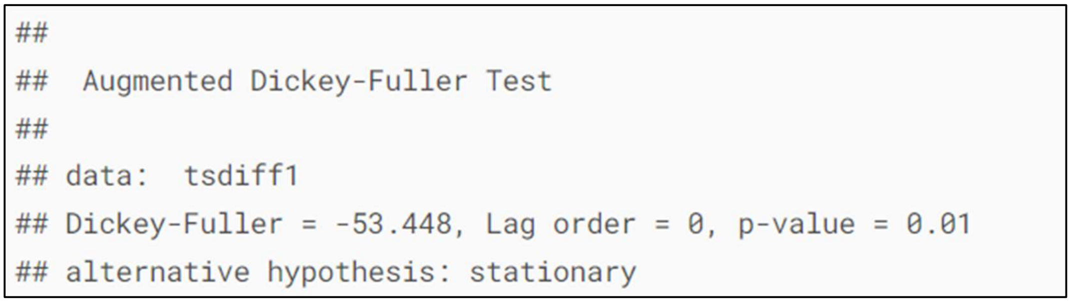
The test suggests that the time series is non-stationary, failing to reject the null hypothesis of non-stationarity. Therefore, we do not have sufficient evidence to support the alternative hypothesis of stationarity.

Next, we will use time series decomposition to dissect the data into trend and irregular components. This process involves testing both additive and multiplicative models to identify the most suitable framework for uncovering underlying patterns. This approach enhances our ability to understand and model complex temporal dynamics in diverse datasets.

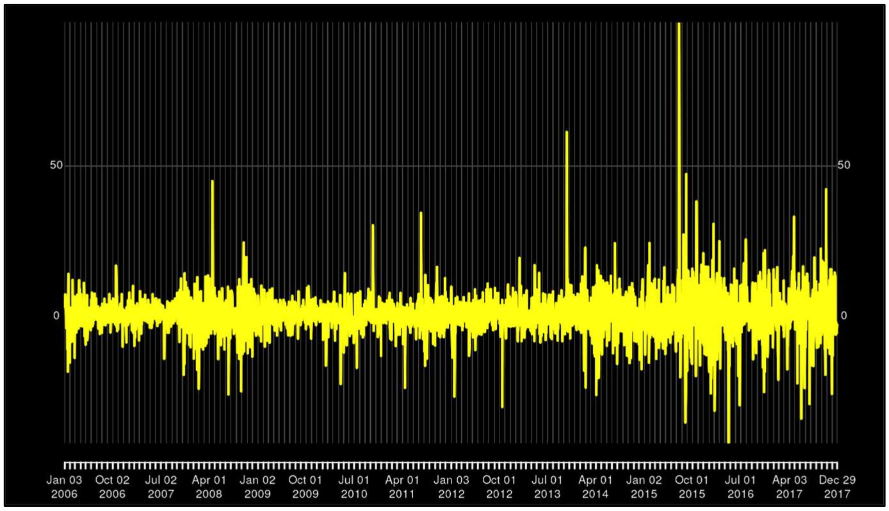




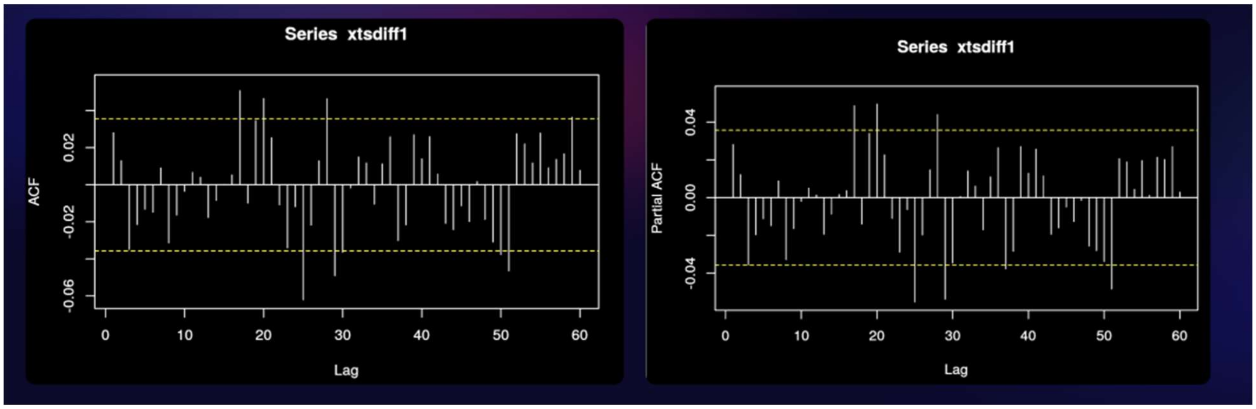
Now, we will utilize the R function 'diff,' differencing proves effective in achieving stationarity in variables. Beyond standard differencing, exploration of fractional differences yields a compelling result: a remarkably low p-value of 0.01. This robust evidence strongly supports the alternative hypothesis of stationarity, affirming the success of the transformation process. The application of these techniques enhances our ability to model and analyze time series data by ensuring the stability and consistency of underlying variables.



```
##
##  Augmented Dickey-Fuller Test
##
## data:  tsdiff1
## Dickey-Fuller = -53.448, Lag order = 0, p-value = 0.01
## alternative hypothesis: stationary
```

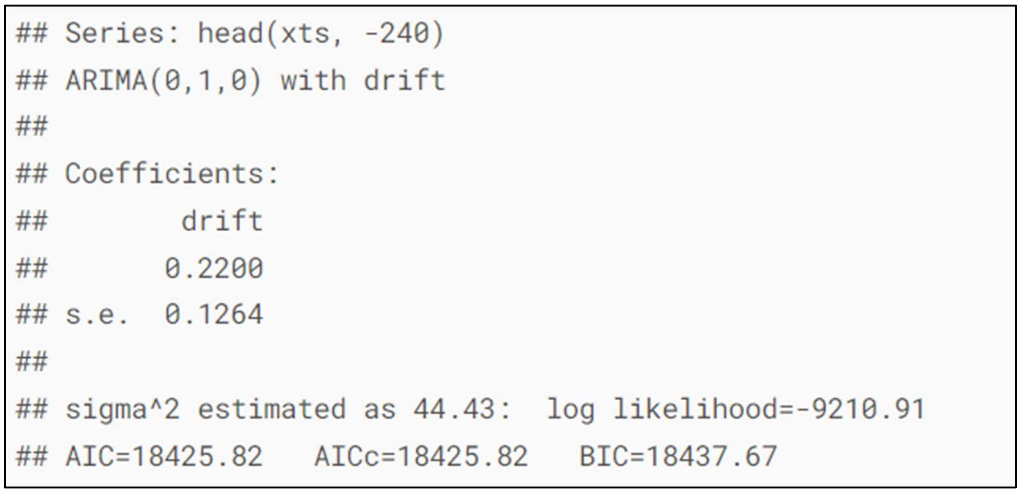


The next step is to select appropriate ARIMA model, which means finding the most appropriate values of p and q for an ARIMA(p,d,q) model. You usually need to examine the correlogram and partial correlogram of the stationary time series for this. To plot a correlogram and partial correlogram, we can use the acf() and pacf() functions in R, respectively.

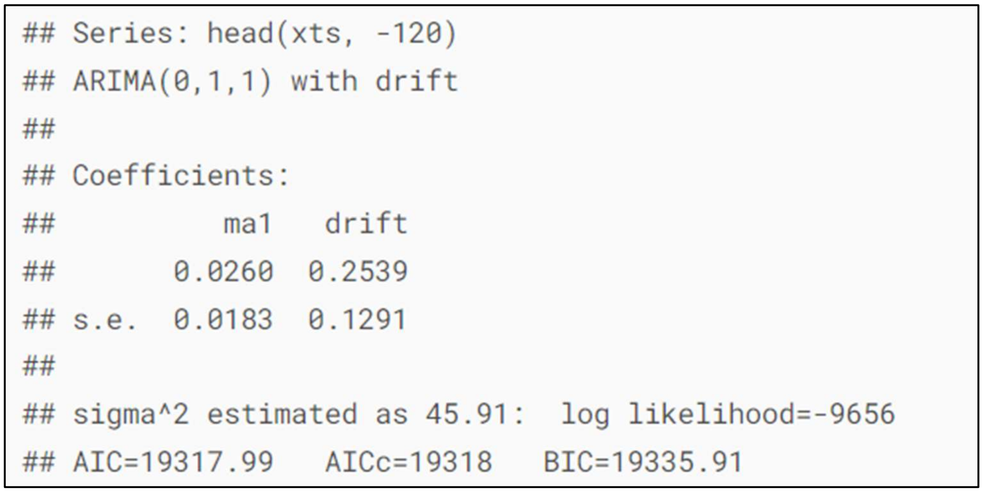


4. Results

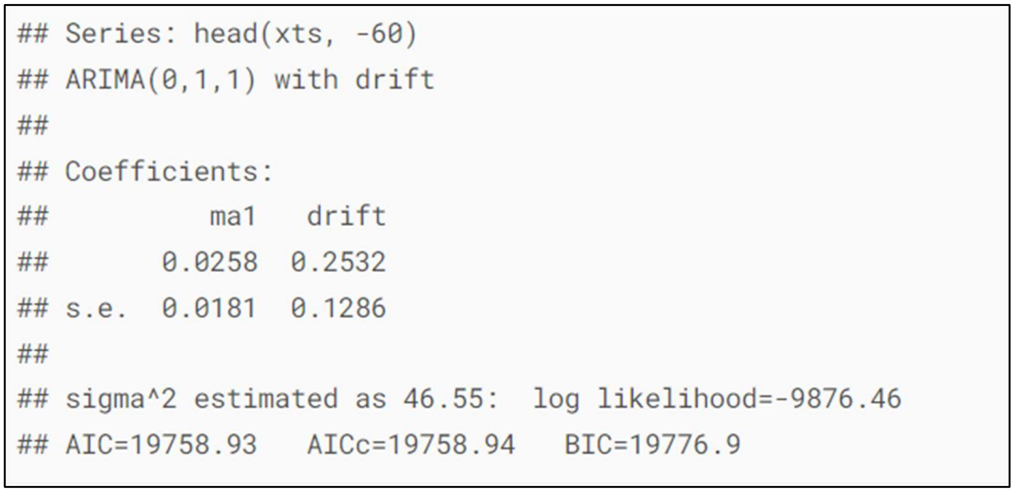
In R, the 'auto.arima' function is a powerful tool for time series modeling, selecting the optimal ARIMA model based on AIC, AICc or BIC values. This function systematically explores potential models within specified order constraints. To enhance model robustness, we train six distinct models, each utilizing different training data subsets. For instance, the 'tsarima240' model is trained with the entire time series, excluding the most recent 240 daily data points. This approach allows for comprehensive testing and validation, ensuring the selection of a well-suited ARIMA model tailored to the specific characteristics of the time series data.



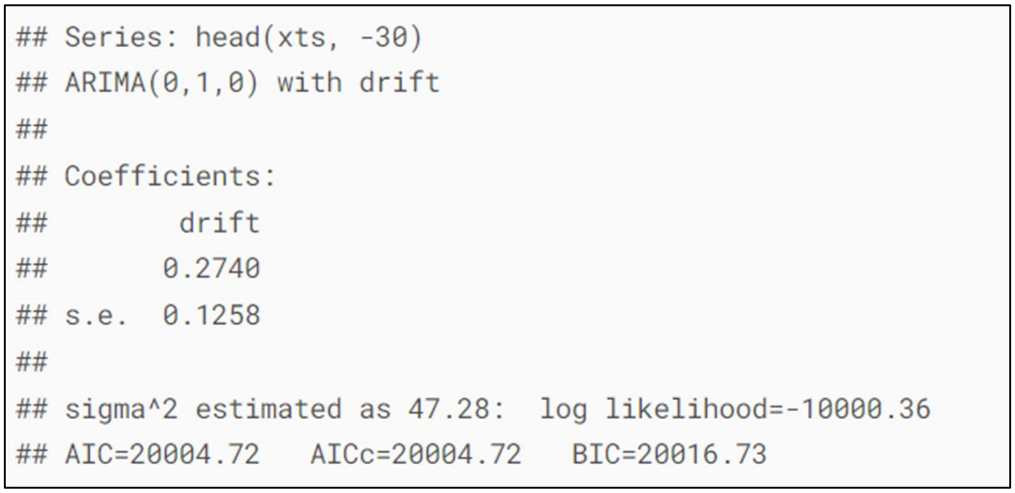
```
## Series: head(xts, -240)
## ARIMA(0,1,0) with drift
##
## Coefficients:
##        drift
##       0.2200
## s.e.  0.1264
##
## sigma^2 estimated as 44.43:  log likelihood=-9210.91
## AIC=18425.82   AICc=18425.82   BIC=18437.67
```



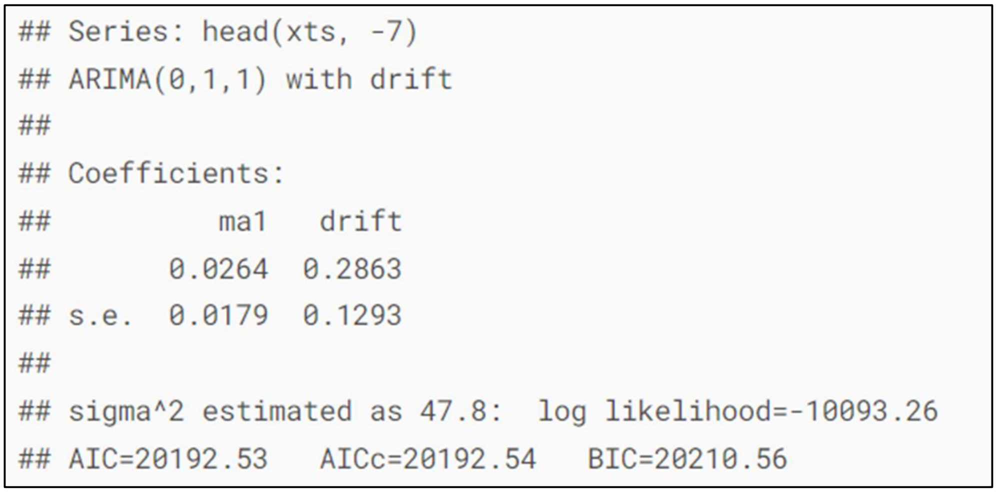
```
## Series: head(xts, -120)
## ARIMA(0,1,1) with drift
##
## Coefficients:
##          ma1   drift
##       0.0260  0.2539
## s.e.  0.0183  0.1291
##
## sigma^2 estimated as 45.91:  log likelihood=-9656
## AIC=19317.99   AICc=19318   BIC=19335.91
```



```
## Series: head(xts, -60)
## ARIMA(0,1,1) with drift
##
## Coefficients:
##          ma1   drift
##       0.0258  0.2532
## s.e.  0.0181  0.1286
##
## sigma^2 estimated as 46.55:  log likelihood=-9876.46
## AIC=19758.93   AICc=19758.94   BIC=19776.9
```

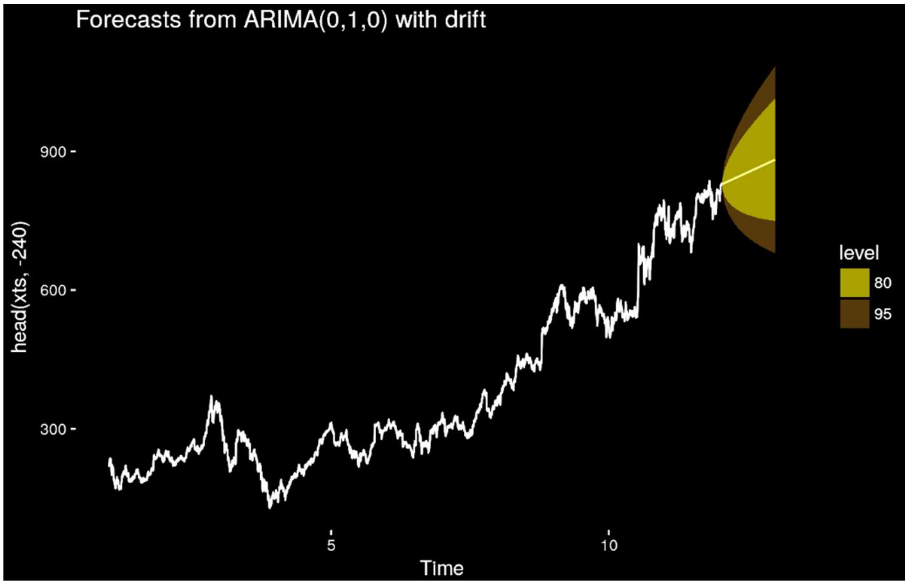


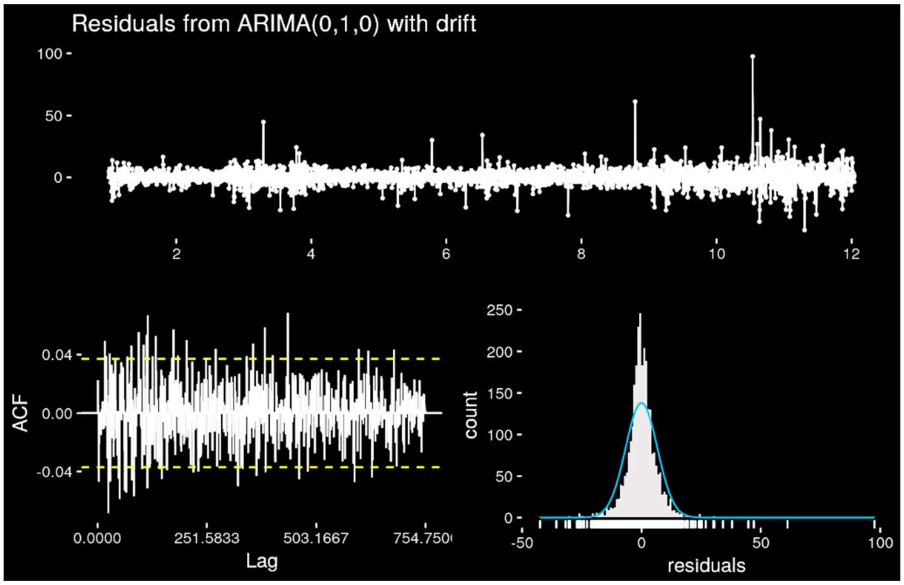
```
## Series: head(xts, -30)
## ARIMA(0,1,0) with drift
##
## Coefficients:
##         drift
##        0.2740
## s.e.   0.1258
##
## sigma^2 estimated as 47.28:  log likelihood=-10000.36
## AIC=20004.72   AICc=20004.72   BIC=20016.73
```



```
## Series: head(xts, -7)
## ARIMA(0,1,1) with drift
##
## Coefficients:
##          ma1   drift
##       0.0264  0.2863
## s.e.  0.0179  0.1293
##
## sigma^2 estimated as 47.8:  log likelihood=-10093.26
## AIC=20192.53   AICc=20192.54   BIC=20210.56
```

The optimal model, ARIMA(0,1,0) with drift, is characterized by a drift coefficient of 0.2200 with standard error 0.126. The estimated variance, sigma^2, stands at 44.43. The model's log likelihood is -9210.91, resulting in AIC and AICc values of 18425.82, while the BIC is 18437.67. These metrics provide insights into the model's goodness of fit and predictive performance.





5. Conclusion

Embarking on a comprehensive exploration of time series mysteries, our journey concludes with the triumphant unveiling of insights through the lens of ARIMA modeling. This transformative process has empowered us to convert raw data into a predictive powerhouse, offering a profound understanding of temporal patterns. Among the array of models considered, the selected ARIMA(0,1,0) configuration, enriched with a drift term, emerged as the beacon of accuracy in forecasting. This victory is not merely the result of chance but is rooted in meticulous training on diverse data subsets, ensuring our model's adaptability to various temporal nuances.

The Ljung-Box test stands as a testament to the robustness of our chosen ARIMA model, validating its predictive prowess and asserting its reliability in capturing and explaining the underlying time series dynamics. The forecast graph, a graphical representation of our optimized approach, paints a vivid picture of a future characterized by predictive precision and informed decision-making.

In essence, this project has gone beyond mere demystification of time series intricacies; it has armed us with a potent tool for making data-driven decisions. Each data point now holds the potential to tell a compelling story, guiding us forward in our pursuit of understanding and harnessing the valuable insights embedded in time series data. As we propel forward, our refined analytical capabilities promise to unlock new dimensions of knowledge, shaping a future where every data point contributes to a narrative of informed and strategic decision-making.