**Received 6/22/24 – 1 Day Early**

**Overall Grade = 100, Congratulations! Best Grade in Class !**

**Rubric**

**REPORT – 70% Early, Report Grade 100**

1. You will turn in a written report. You need to write this report as if it the report will be routed in an office to personnel of vary different backgrounds. The ToC is great, Executive Summary is very good.
2. You need to be able to reach readers that have no data science background to full fledge data scientists. So, you need to explain what you have done, why and how you did it with both layman and technical terminology. The introduction is helpful here. Your report reads very well.
3. **Do not simply write this with solely me in mind**. Great job.
4. Visuals and output are expected, but **do not to include every bit of analysis you perform**. Your panel plots are nice, good examples without over doing it. Code offsets are nice.
5. A terse report with simple terminology get a higher score versus throwing in everything into a long, ad nauseam report. This is an art and a central theme of data science – the balance to automation and math traded with simplicity and transparency. I think you did a solid job.
6. Story telling is really taking on for data science, so please flex your muscles here. Okay
7. Your report will be graded on value of information presented, brevity, readability (easy to read), accuracy and demonstration of content knowledge. Very solid on the core components.

**Forecast Accuracy - 30 % Grade 100**

1. Forecasts, sending your Excel file in the **right** format is very important as I will calculate measures of accuracy (ask me if needed). You need to include training and forecast data.
2. Best Accuracy overall. Congratulations !

# **Data 624 Project 1**

By Zach Rose, Yina Qiao, Gavriel Steinmetz-Silber, and Noori Selina

6/23/2024

# Table of contents

* [Executive Summary](#_vvtgdm7pjbka)
* [Introduction](#_qv44jviroo5a)
* [Data Preprocessing](#_osd7rlgpqnir)
* [Exploratory Data Analysis (EDA)](#_80idp38z1vo2)
* [Decomposition](#_fl8pfg5uohnh)
* [Difference and Stationarity Checks](#_waf5d6kexqun)
* [Model Selection and Implementation](#_7o1lptezttm)
* [Forecasting](#_yk7mtl88n96o)
* [Results and Discussion](#_x78v5nkrfzd)
* [Conclusion](#_agbr8g1xza59)
* [Appendix](#_254y0zfmq3ij)

# Executive Summary

In this project, we analyzed 1622 periods of de-identified time series data to forecast 140 future periods using various models. We evaluated ARIMA vs ETS models using RMSE and MAPE metrics on unseen data, and chose ARIMA due to better residual diagnostics, capturing trends and seasonality effectively. These insights are valuable for decision-makers and stakeholders, optimizing future planning and strategies

# Introduction

Time series forecasting is a powerful tool used across a wide range of fields, from global economic markets to local environmental studies. In this project, we aim to thoroughly analyze a given set of data and accurately forecast the next 140 periods.

Our approach involves several steps:

* Decomposition: Breaking down the time series data into its basic components, such as trend, seasonality, and residuals, to understand underlying patterns.
* Differencing and Stationary Checks: Applying techniques to remove trends and seasonality, ensuring the time series data remains stable over time.
* Model Building: Constructing ETS (Error, Trend, Seasonality) and ARIMA (AutoRegressive Integrated Moving Average) models to capture the data's dynamics.
* Best Fit Model Selection: Evaluating different models and selecting the one that best fits the data.
* Forecasting: Using the selected model to predict future values for the next 140 periods.

Finally, we will review the results to understand the implications of our 140-period forecast. Through this comprehensive process, we aim to gain valuable insights into the data and provide reliable forecasts for future decision-making.

# Data Preprocessing

Each series utilized data cleaning methods to eliminate missing values and treating outliers. We group the series together based on the corresponding variables. We began by converting the ‘SeriesInd’ column to dates, which would allow us to produce better visualizations and analysis. Then, removing missing values was essential to create a continuous time series, and this was done so with ‘na\_interpolation’ functions. An example of this data transformation on Series1 can be seen below:

S01 <- data %>%

mutate(date = as.Date(SeriesInd, origin = '1899-12-30')) %>%

filter(SeriesInd < 43022 & category == 'S01') %>%

select(SeriesInd, date, category, Var01, Var02)

S01$Var01 <- na\_interpolation(S01$Var01)

S01$Var02 <- na\_interpolation(S01$Var02)

# Exploratory Data Analysis (EDA)

For the initial analysis we relied on the use of histograms and box plots to give us a better visualization of the data. The insights from these plots and charts allowed for an increased understanding into the structure of the skews and distributions of the data. It also assisted in discovering which variables would need to have outliers detected and replaced. With a large number of these plots and charts created, I have included some below that highlighted the different forms of results we collected:

**Figure 1: Figure 2:**

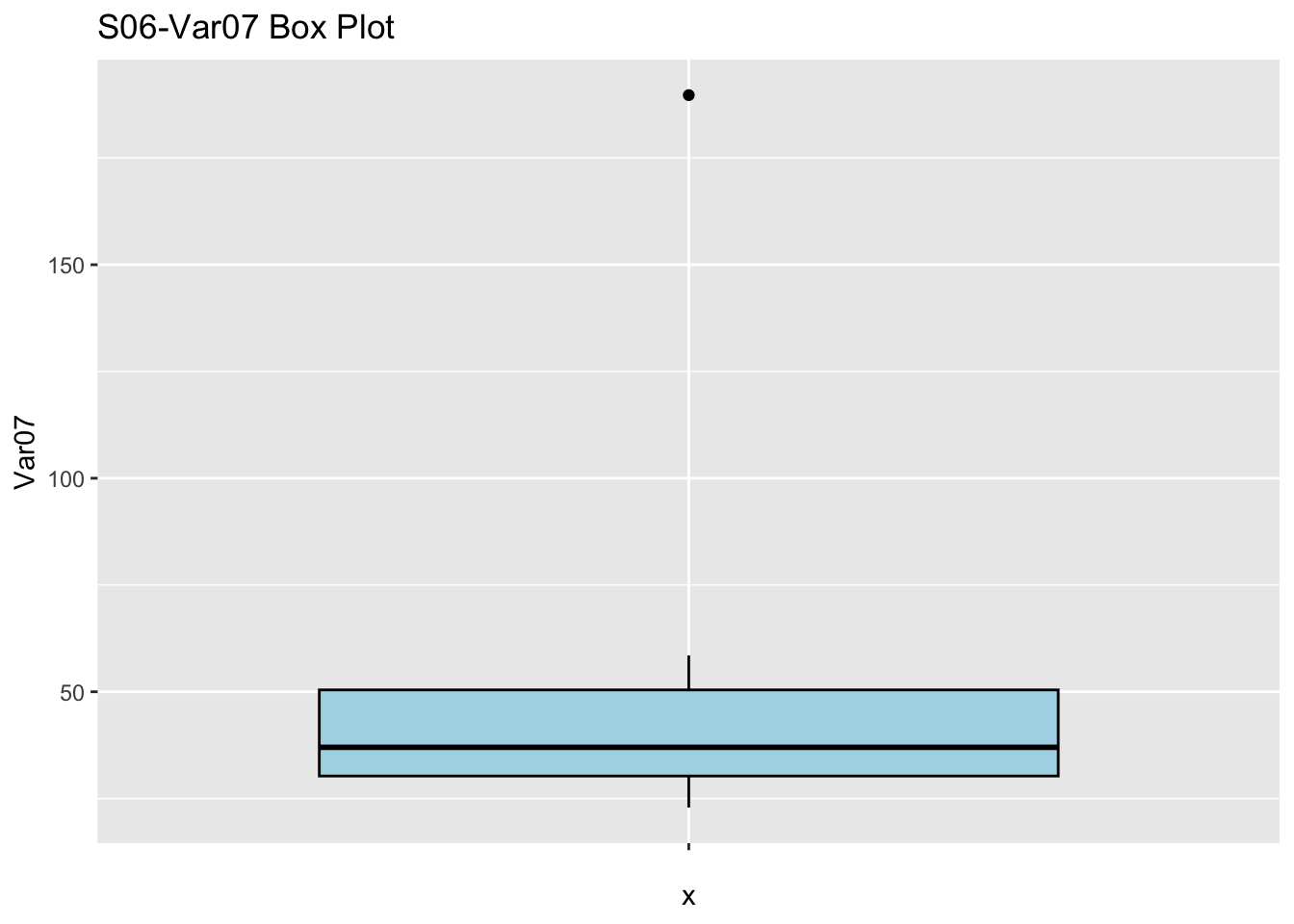
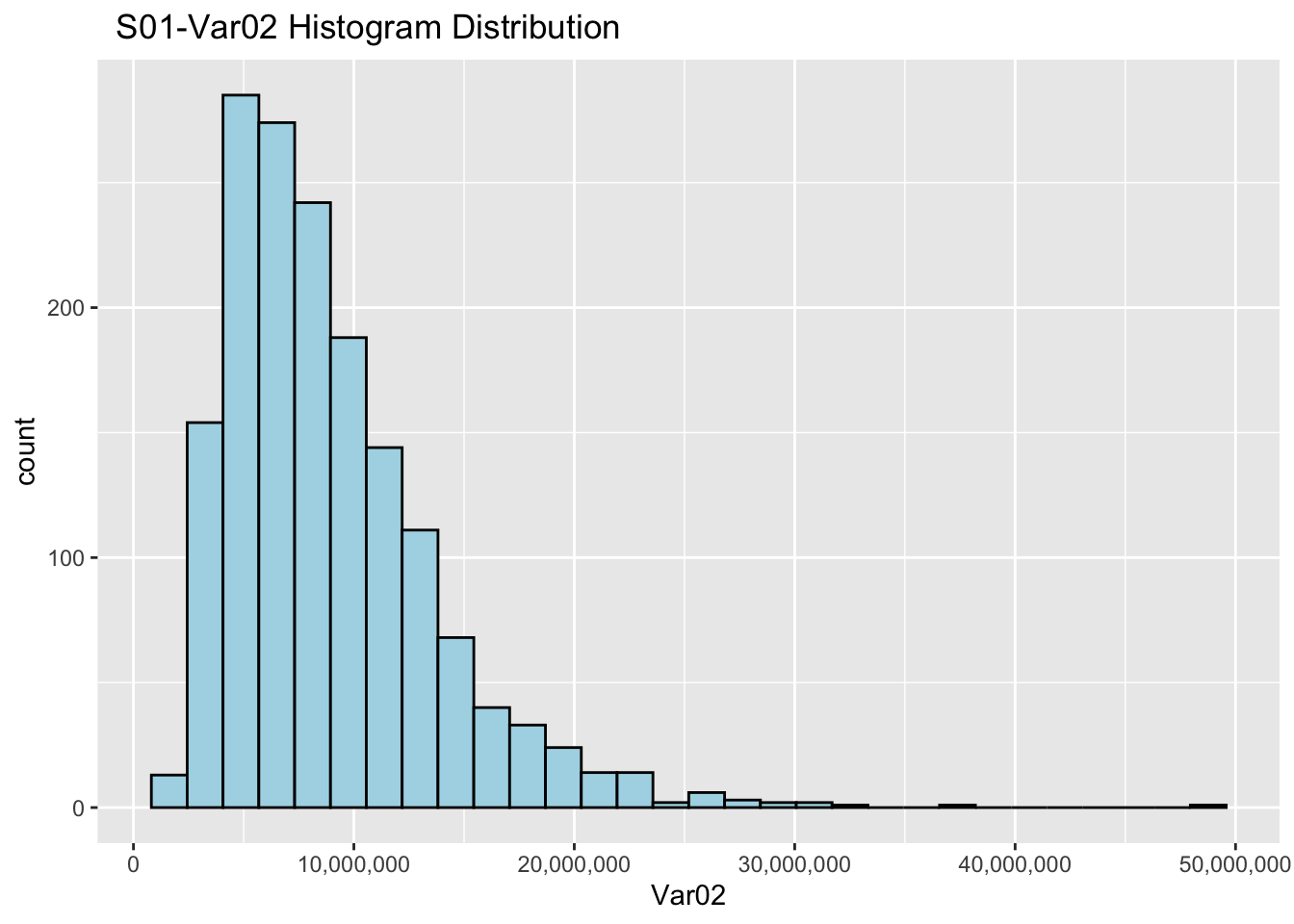


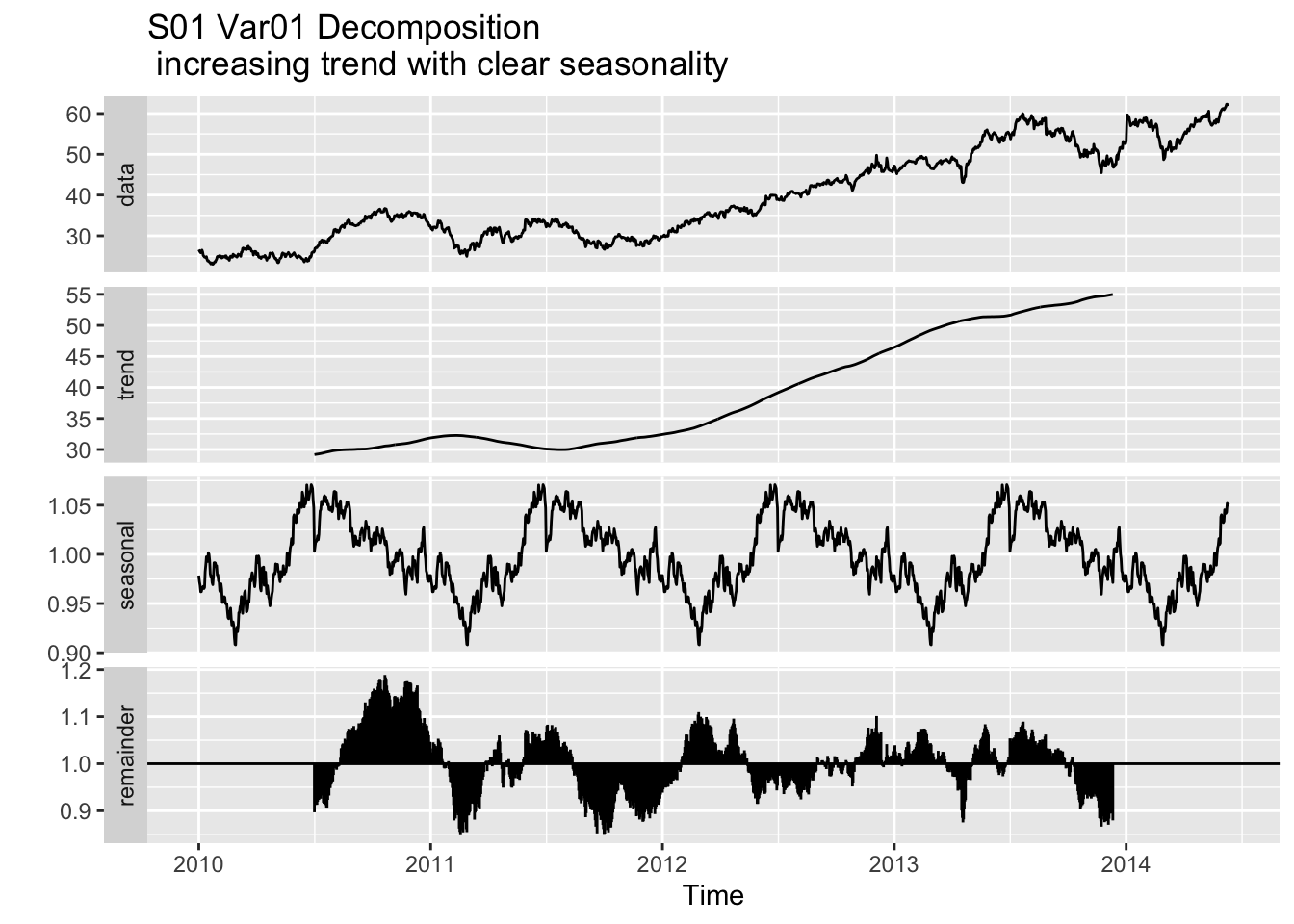
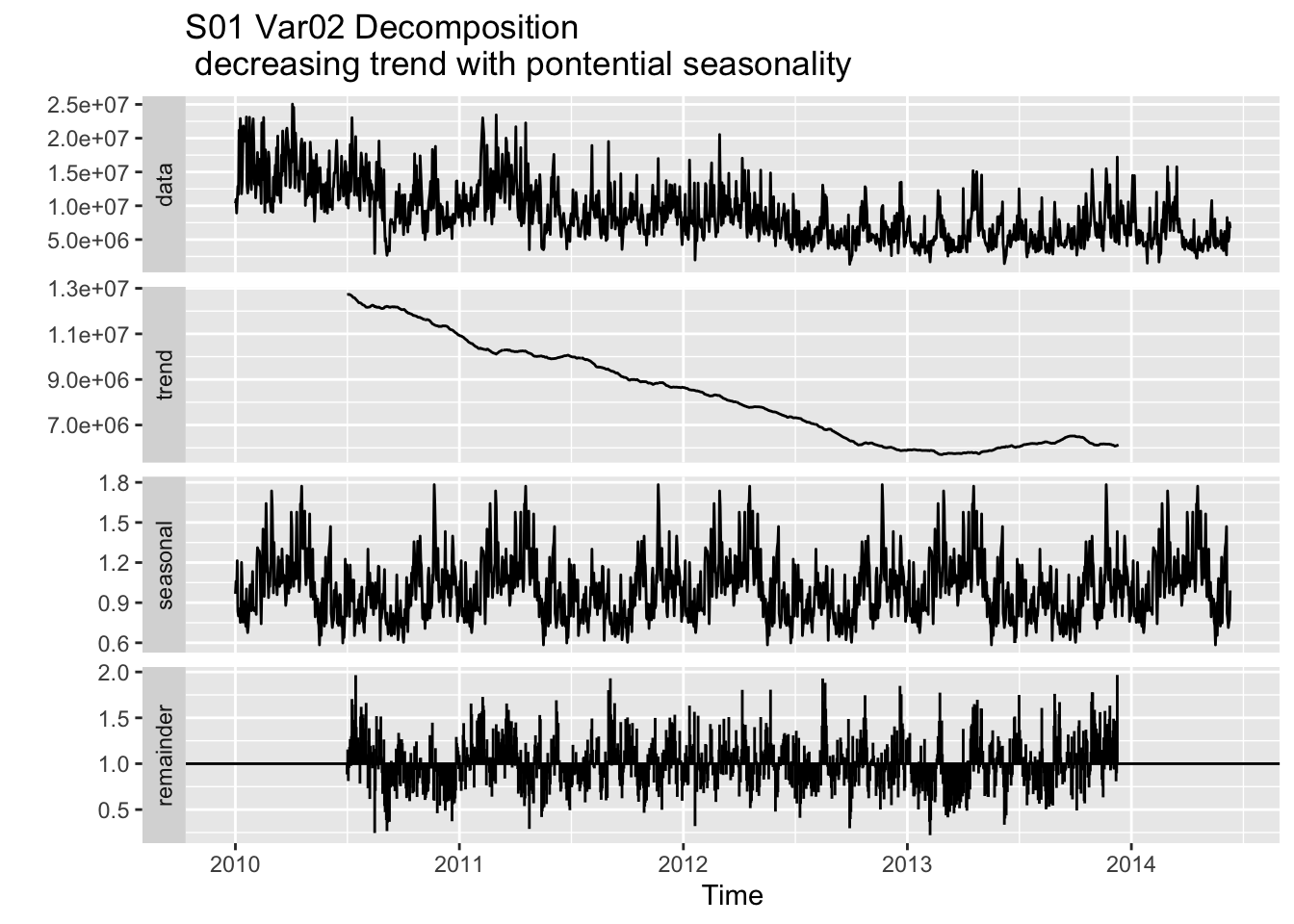
Figure 1’s histogram displays that the data from S01-Var02 was heavily right skewed with several outliers of high value.

Figure 2’s box plot displays a clustered distribution with a single high value outlier.

# Decomposition







We decompose the ts to see the trends and seasonality.

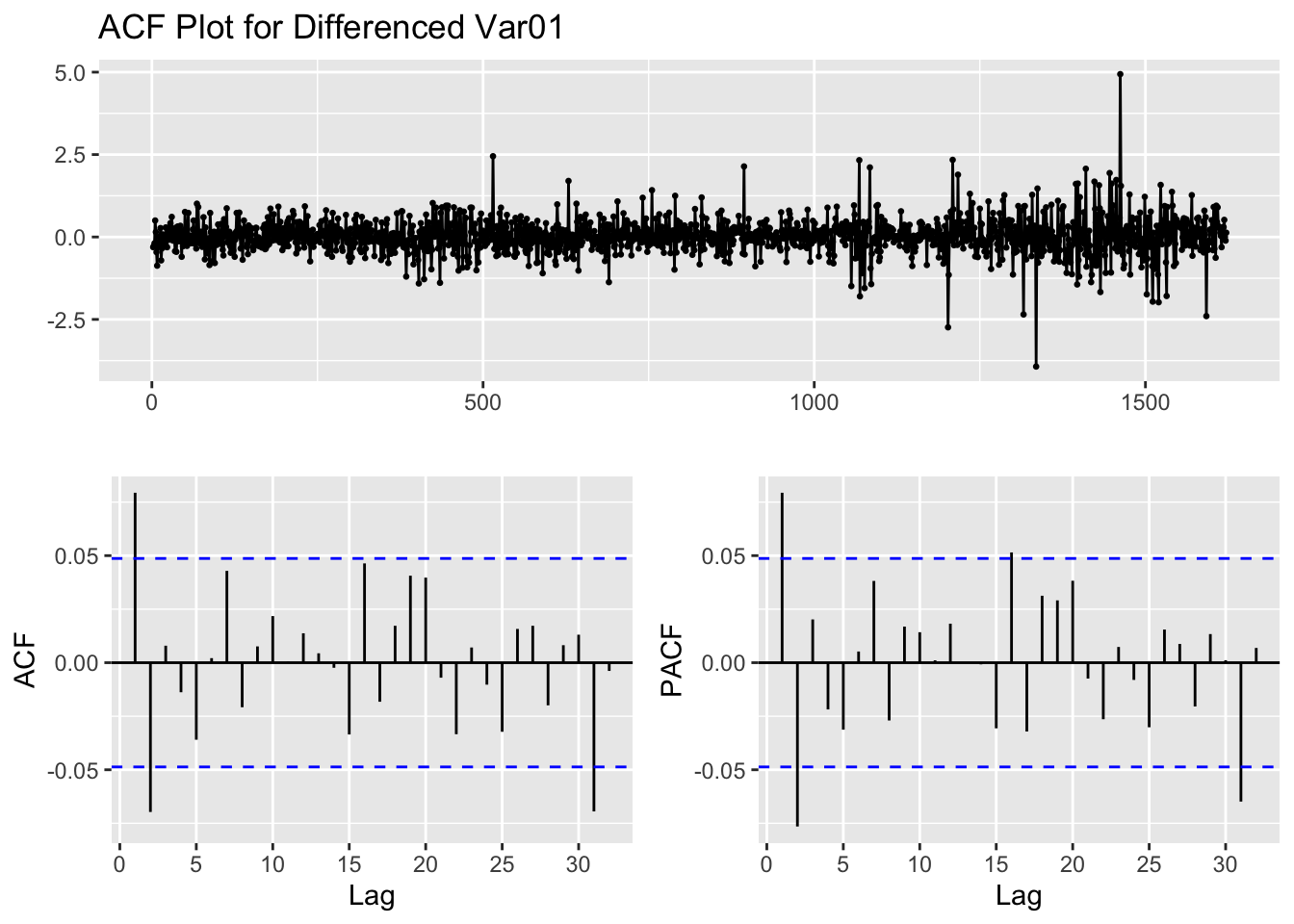
# Differencing and Stationarity Checks

The tests below check if the data we are working with is smooth (stationary) or bumpy (non-stationary). Most forecasting models work best with stationary data, so we check to ensure accurate predictions.



For Var01, initial test statistics (16.0589) showed the data was non-stationary. After applying a differencing technique, the test statistics (0.1001) showed the data became stationary.

The ACF and PACF plots reconfirm this. The quick drop-off to zero in the ACF plot confirms that there is no significant autocorrelation beyond the first few lags, which is a characteristic of stationary data. This implies that the data is suitable for applying forecasting models such as ARIMA and ETS. After making it smooth (differencing), it trains the model on this smooth data to make accurate predictions.



# Model Selection and Implementation

In our analysis of the time series data, we built several ARIMA and ETS models in hopes of finding the best fit to forecast the 140 future periods. These models were selected for their strong ability to forecast data. Below is a brief explanation of what these models do and why they were chosen.

**ARIMA Models (AutoRegressive Integrated Moving Average):**

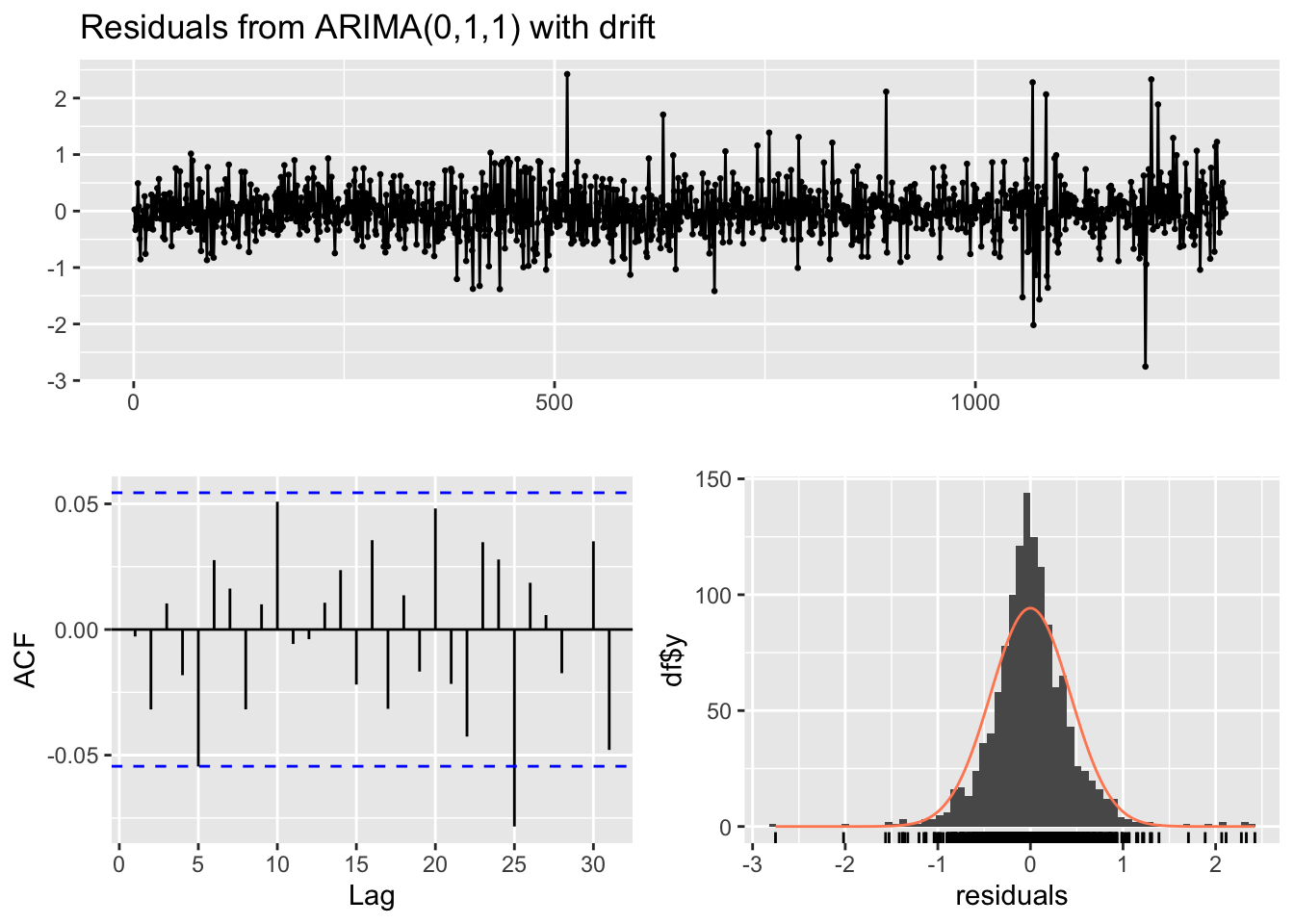
* ARIMA models forecast data by using the time series’s previous values, the difference between these previous values, and a moving average
* ARIMA models are unique in their ability to work with time series data featuring trends and seasonality
* Through the use of both moving average and auto regressive components, ARIMA models can capture both short term and long term discrepancies.
* Optimization tools such as AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion), both used in our models, allow the user to help identify the best fitting model by balancing model fit and model complexity.

**ETS Models (Error, Trend, Seasonality):**

* ETS models use a method called decomposition to break a time series into three parts, error, trend, and seasonality. This allows users to understand the components contributing to changes in the data.
* Three Parts:
  + Error: Displays fluctuations or spikes in data that the model can not account for
  + Trend: Represents long term direction of data
  + Seasonality: Displays what normal or regular patterns may be present in the data, and what sort of fluctuations they cause.
* ETS models can handle complex time series data as well as create smoothe forecasts for easy readability.

When deciding which model to use to forecast the 140 future periods, we relied partially on the results of each model’s RMSE (Root Mean Squared Error), MAPE (Mean Absolute Percentage Error), AIC (Akaike Information Criterion), and BIC (Bayesian Information Criterion) to assist in the decision. We also used residual plots to help in the model identification as they allowed us to visualize autocorrelation detection, and to see the size of errors. Below is an example of a residual plot for SO1 Var 01. For all the models in this project, we moved forward with selecting ARIMA models due to their consistently better RMSE and MAPE in both training and validation sets, as well as their frequently lower AIC and BIC values.

**Check residuals**



The residuals for the plot above are small and random with no significant autocorrelation detected, showing us that the model fits well.

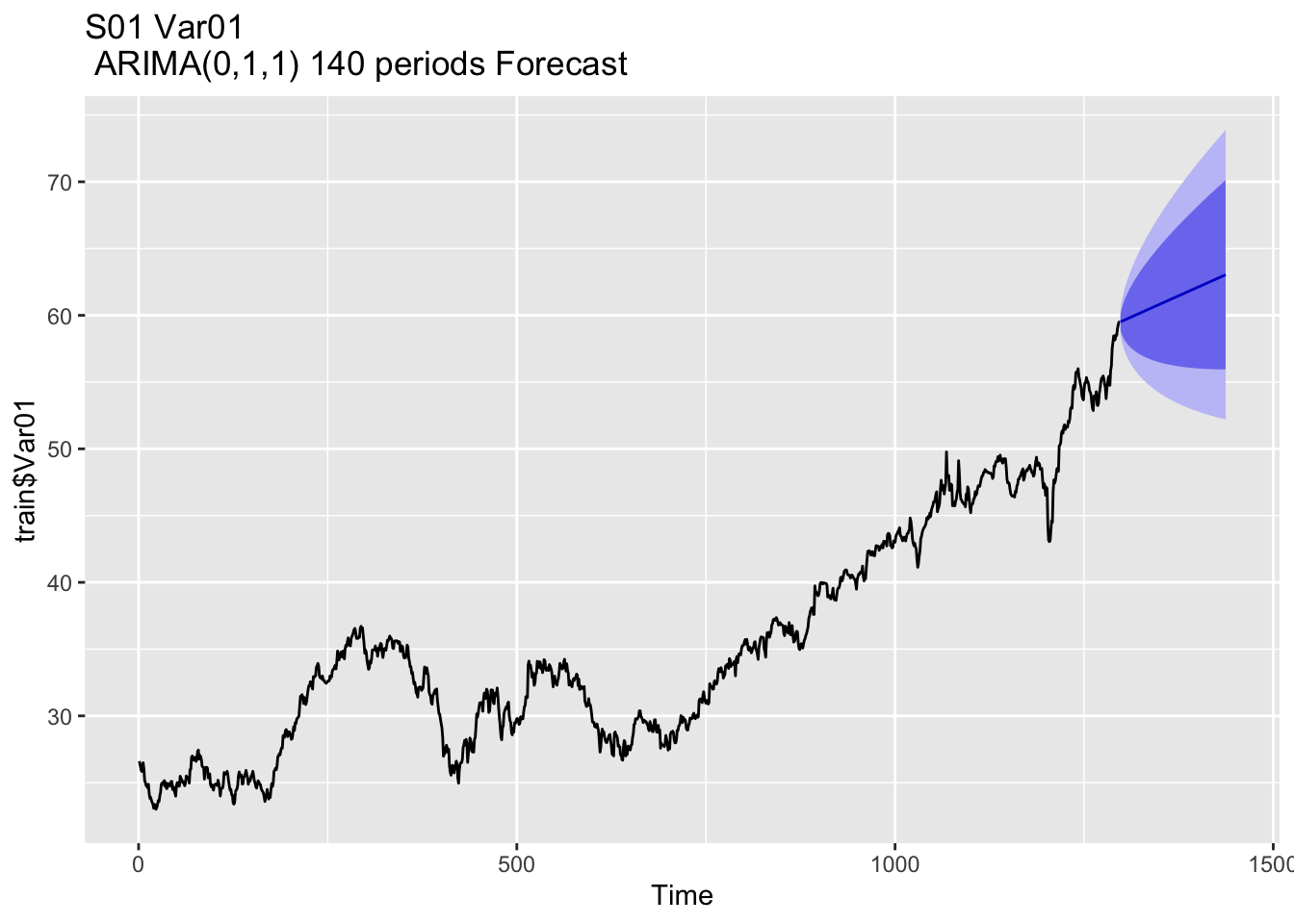
# Forecasting

In this section we will be reviewing and giving a basic overview of each time series forecast conducted.

**SO1:**

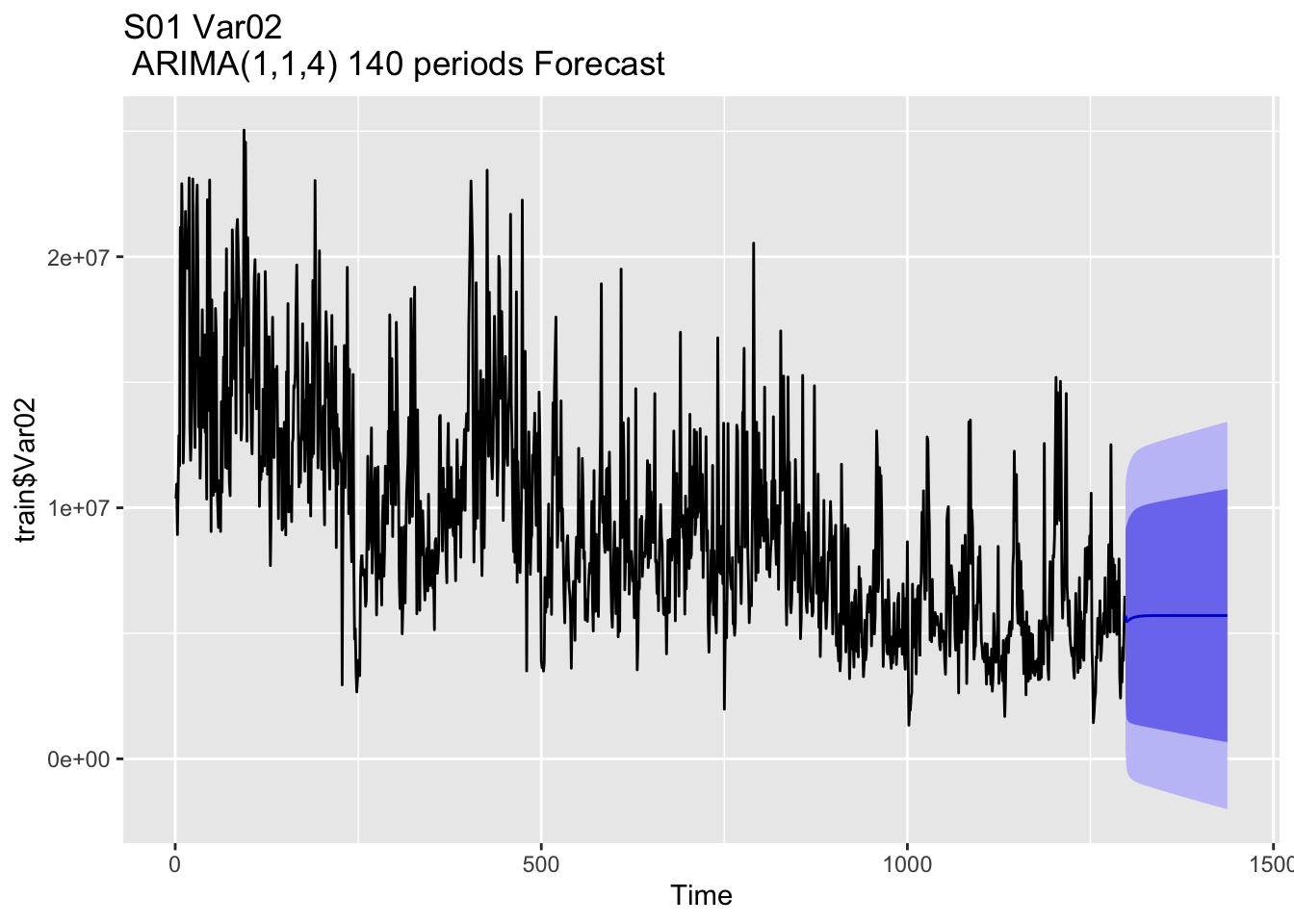
**Var 01:**

The forecast for Var 01 indicates a consistent upward trend over the 140 future periods. The confidence intervals are mostly narrow, suggesting that there is a strong likelihood that the predictions are accurate. Decision making using this model would be plausible.



**Var 02:**

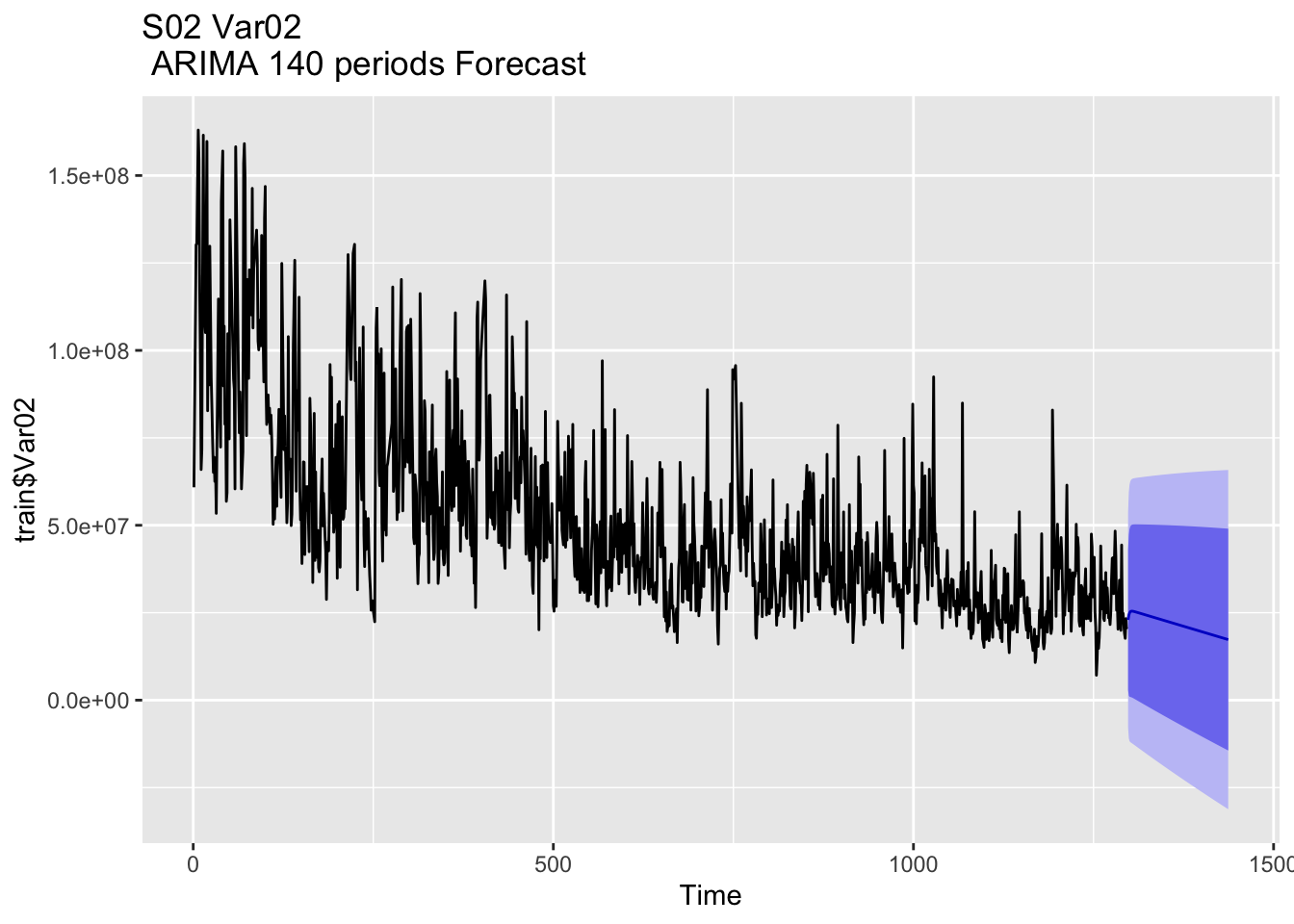
Var 02’s forecast displays greater variability and larger confidence intervals due to the increased complexity of the data provided. However, the model still provides solid insights into the future possibilities.



**SO2:**

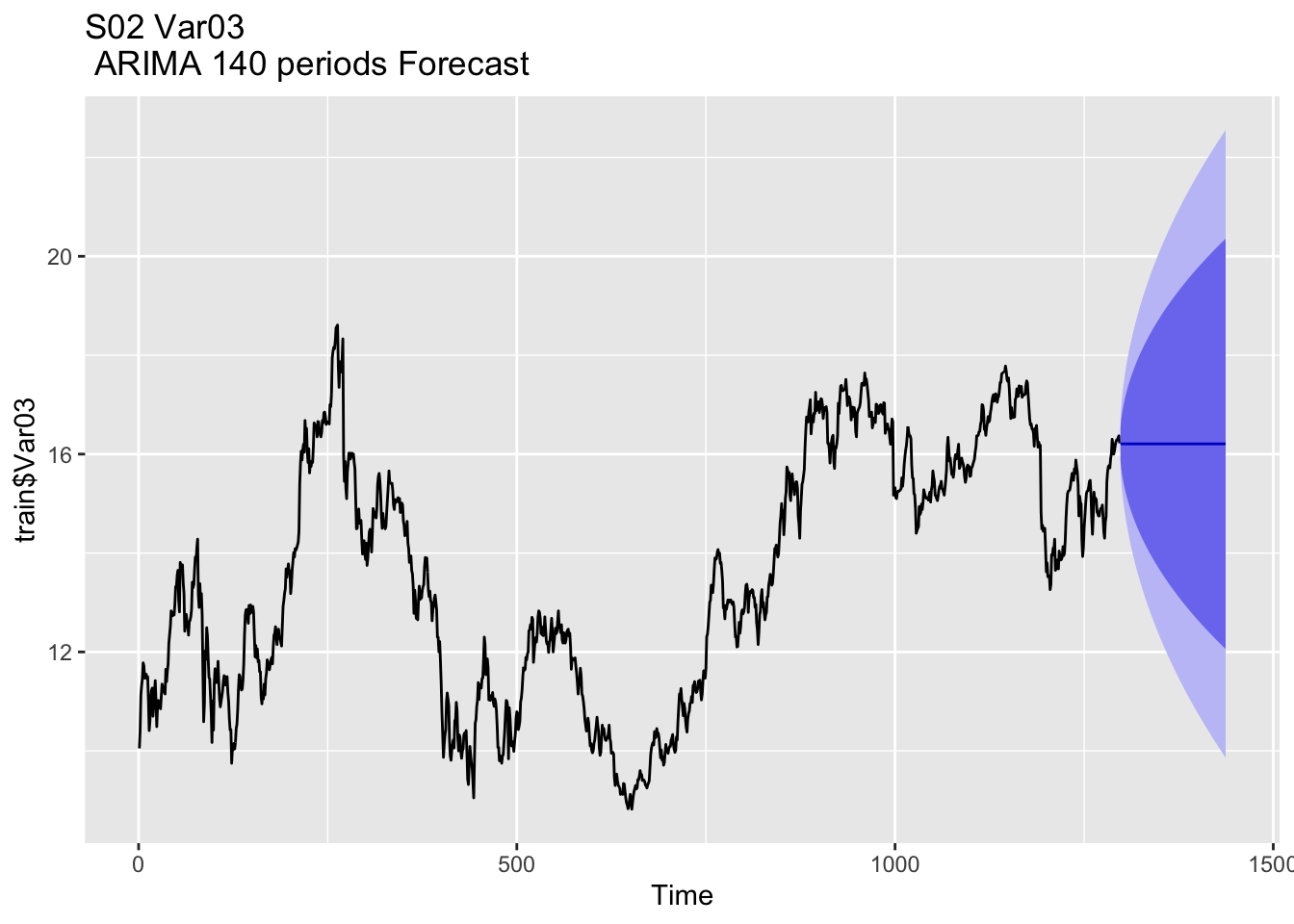
**Var 02:**

Var 02’s forecast features relatively small confidence intervals, indicating that the results of the model are reliable for predicting future outcomes.



**Var 03:**

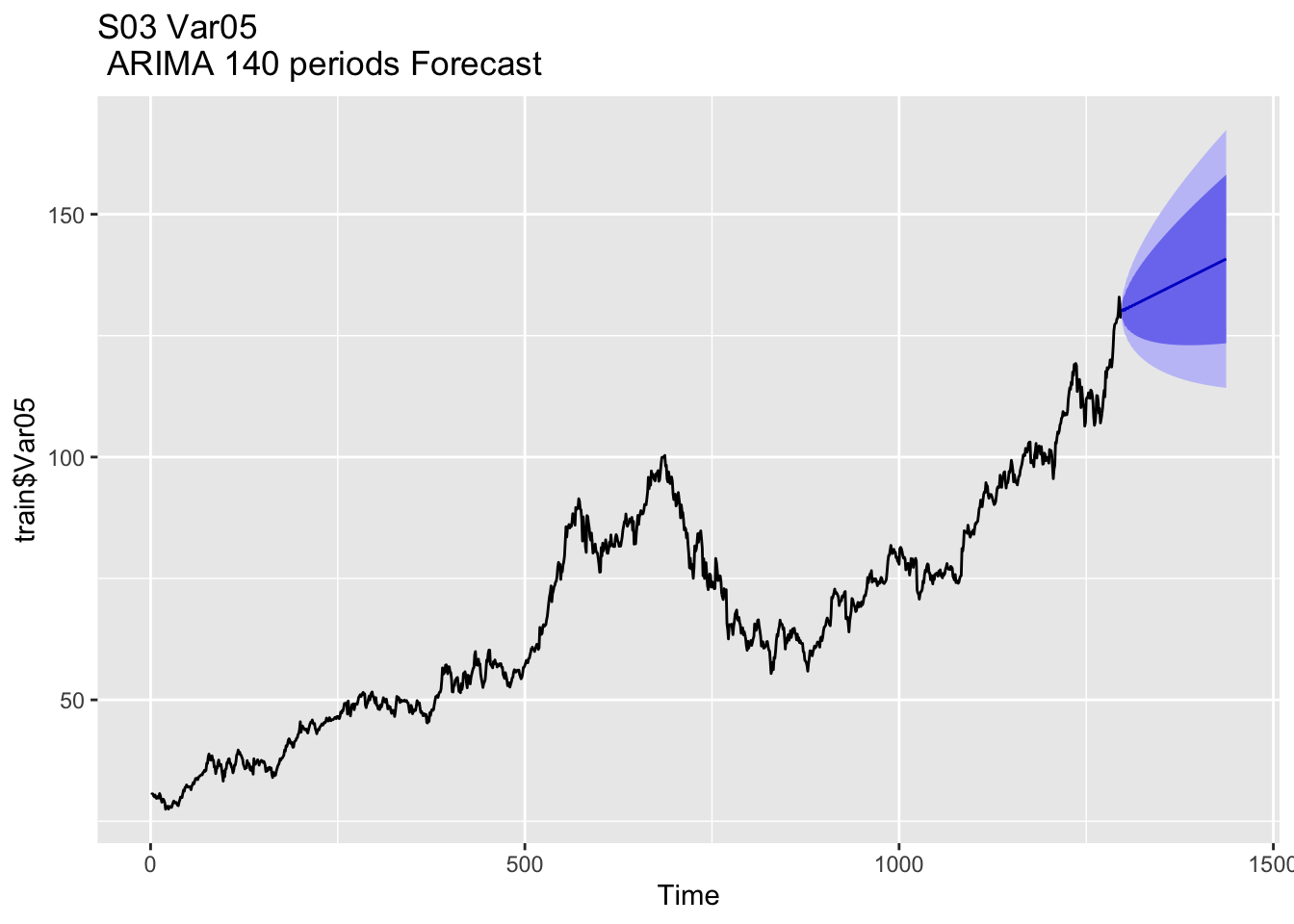
The confidence intervals for Var 03 are much larger than the previous models, indicating that the results and accuracy of the forecast should be taken with some uncertainty.



**SO3:**

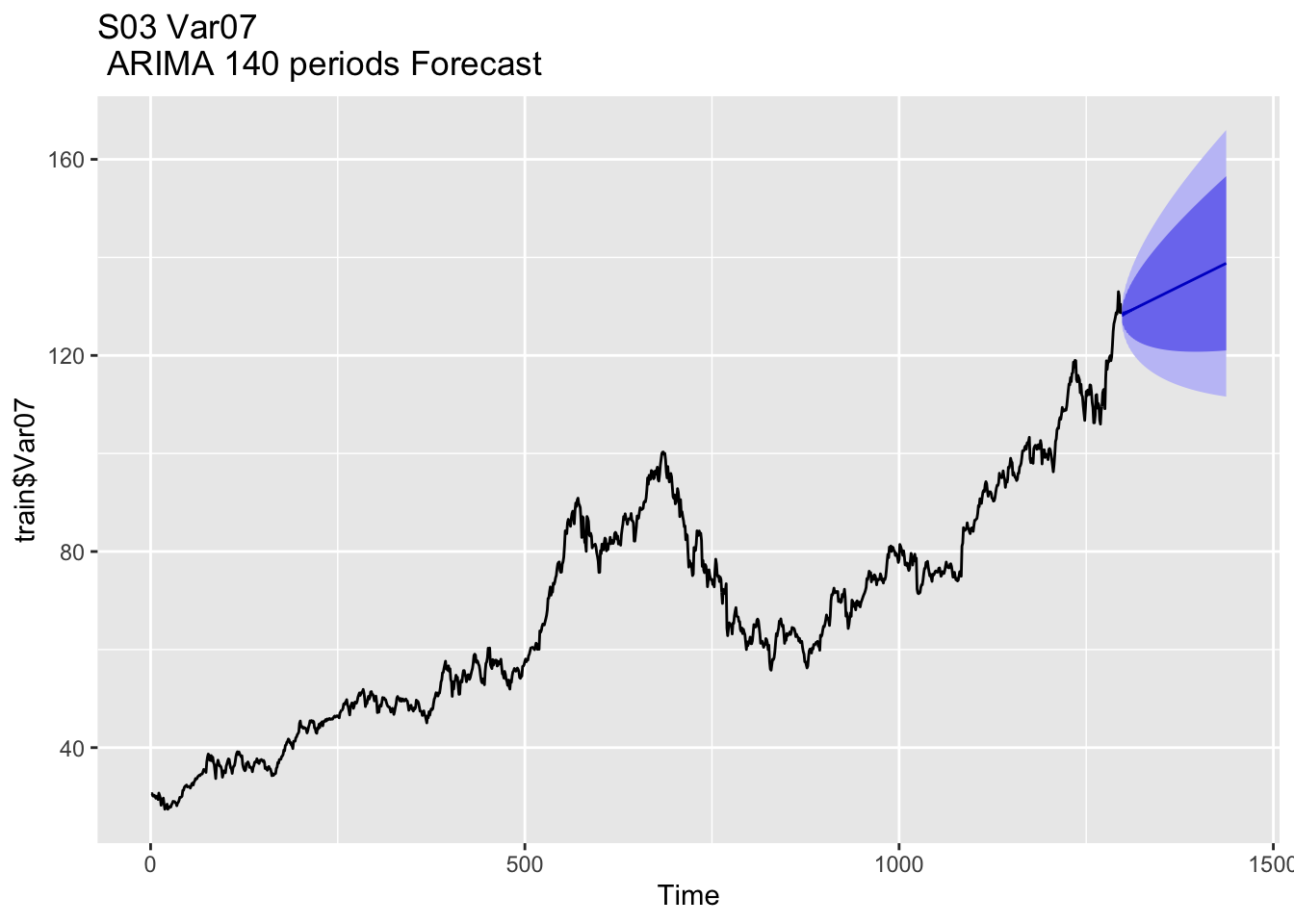
**Var 05:**

The forecast for Var 05 displays a strong upward trend and small confidence intervals, implying that the predicted forecast is likely to be accurate.



**Var 07:**

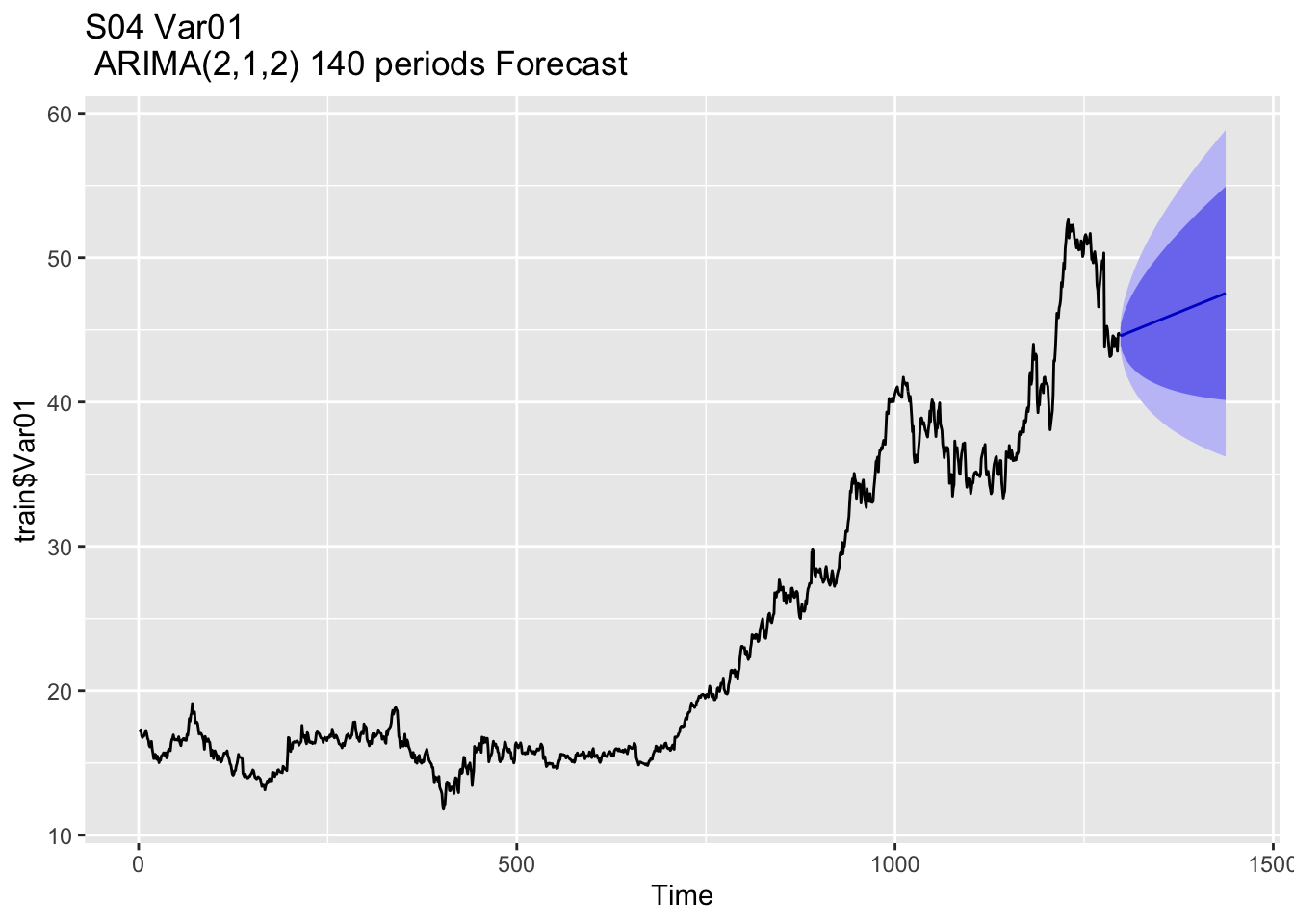
Similar to Var 05, Var 07’s forecast also displays an upward trend as well as small confidence intervals, suggesting similar accuracy.



**SO4:**

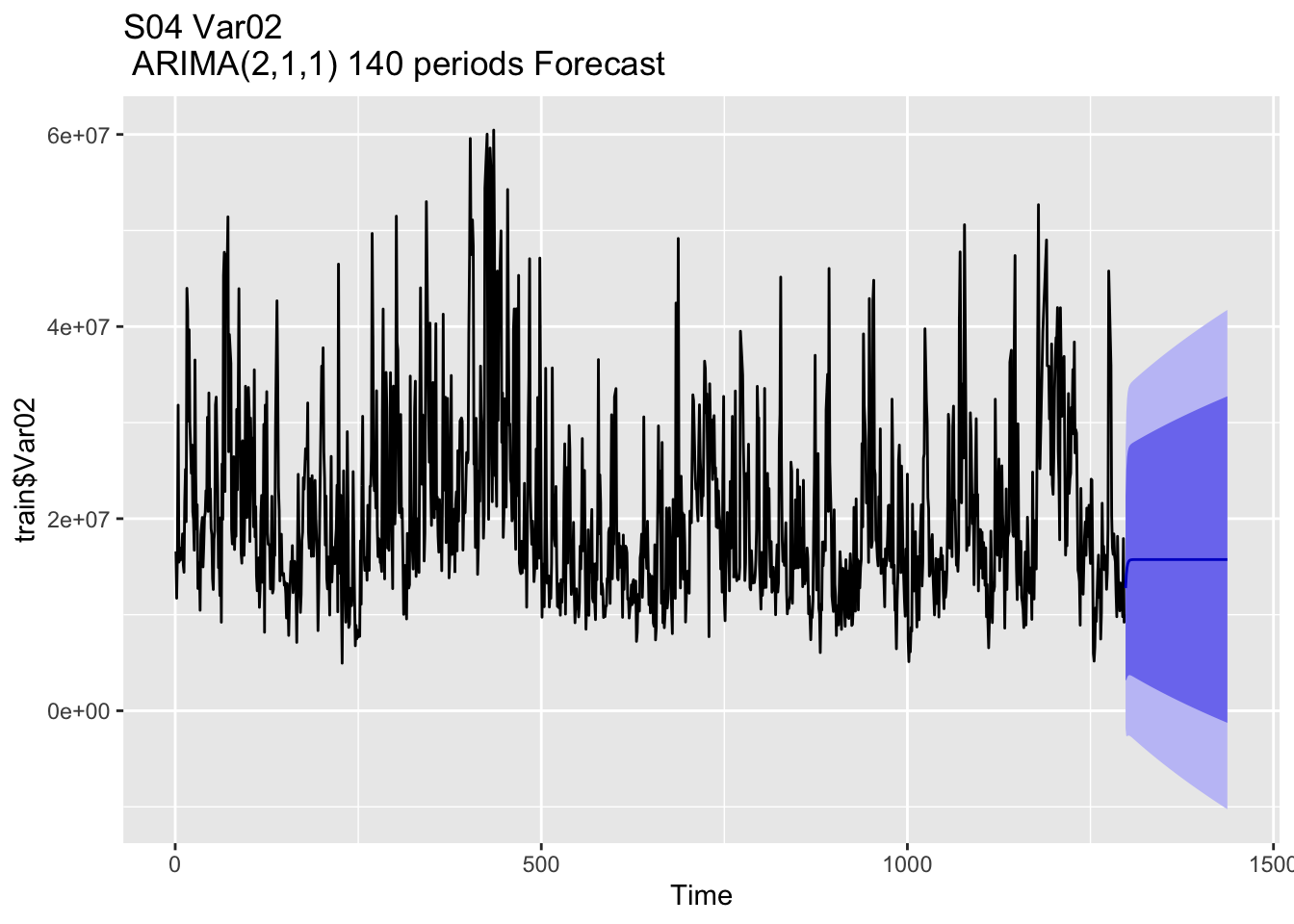
**Var 01:**

The forecast for Var 01 has relatively small confidence intervals and indicates a gradual upwards trend.



**Var 02:**

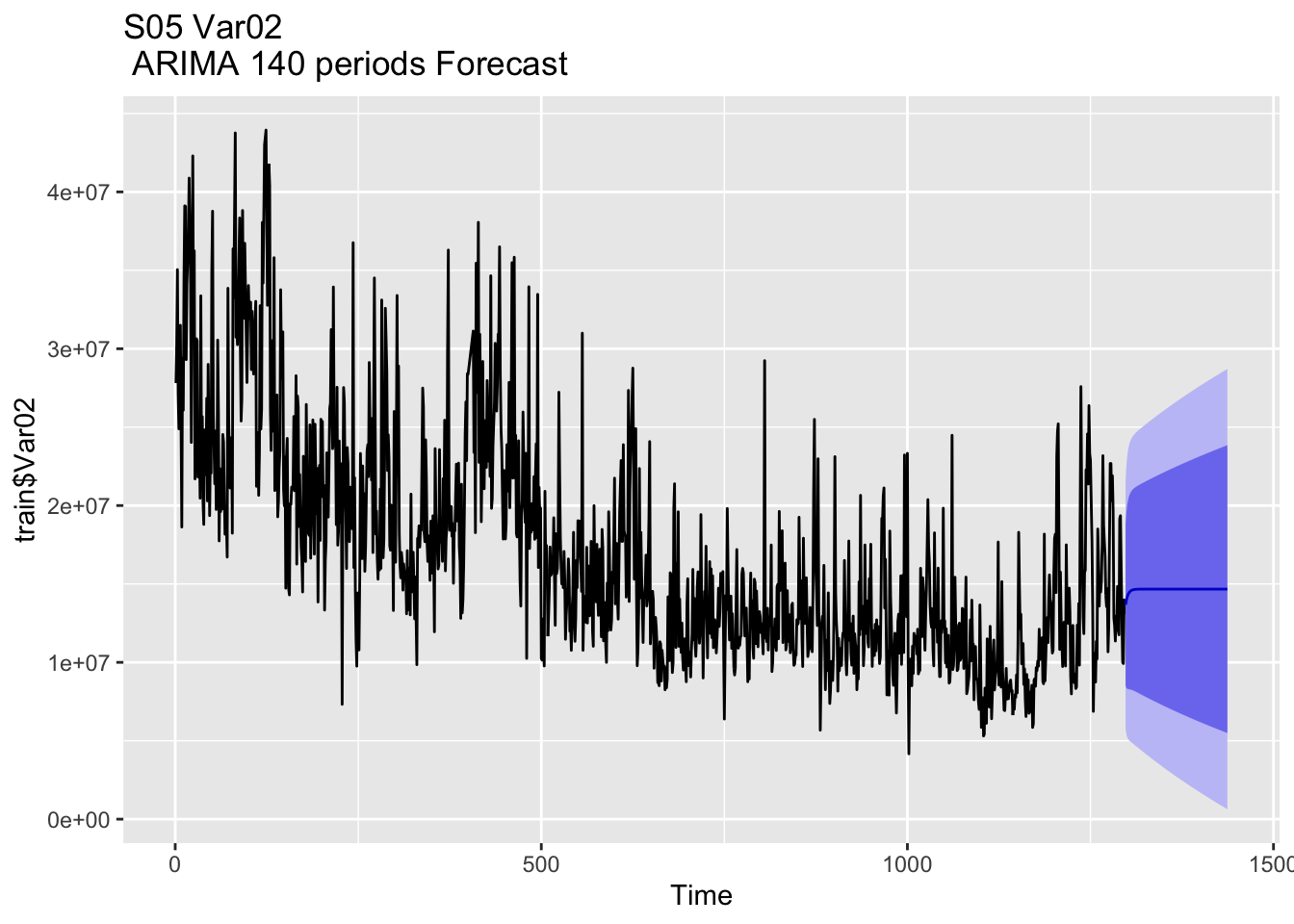
Var 02’s forecast shows the data staying somewhat consistent in the future periods, however the model also displays a relatively larger confidence interval.



**SO5:**

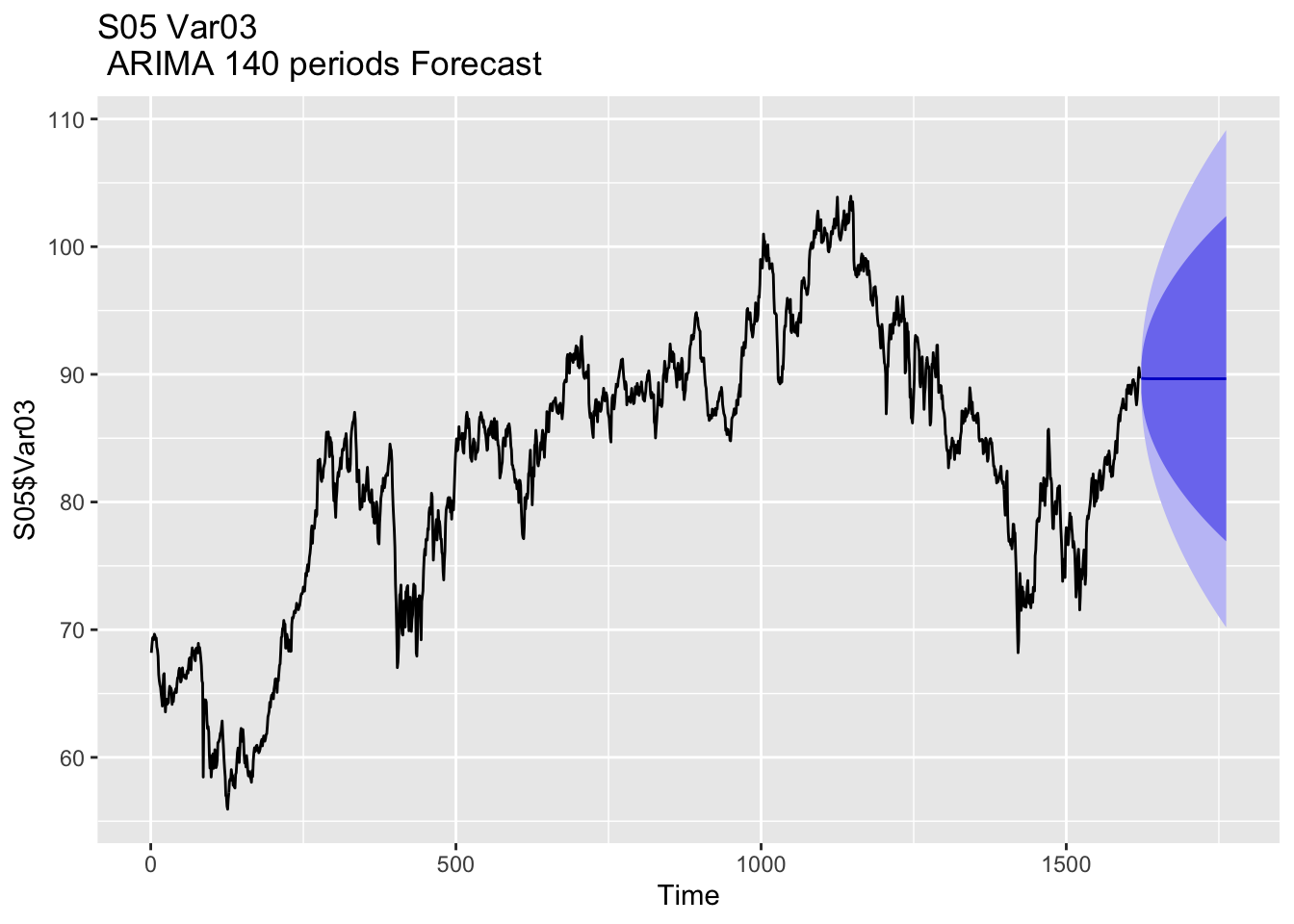
**Var 02:**

The forecast for this model expects the values to stabilize over the next 140 periods, however, the model also displays a relatively larger range of confidence intervals.



**Var 03:**

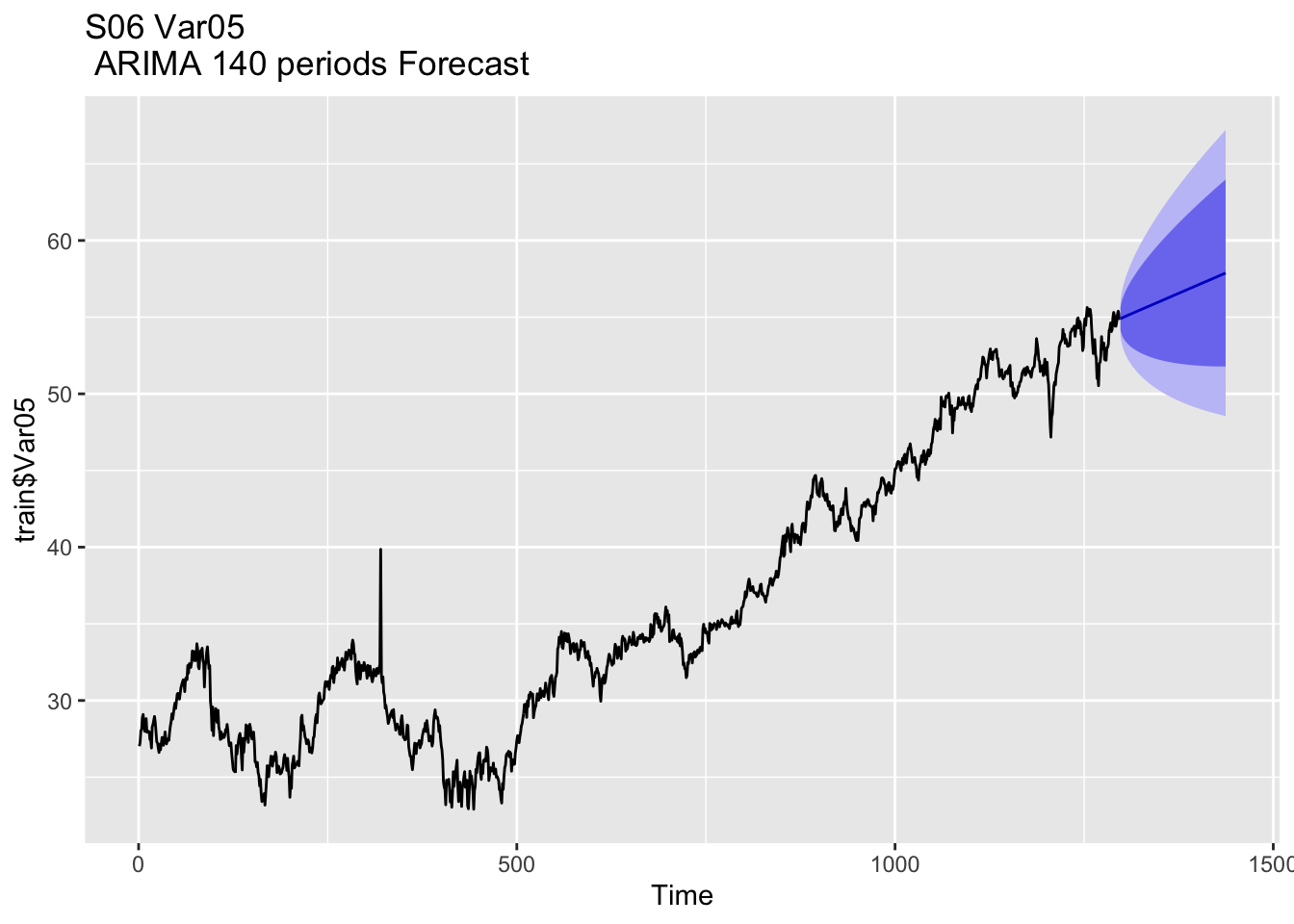
Var 03’s forecast also displays the data stabilizing over the predicted future periods, however the confidence intervals for this model are relatively smaller indicating greater likelihood of potential accuracy.



**SO6:**

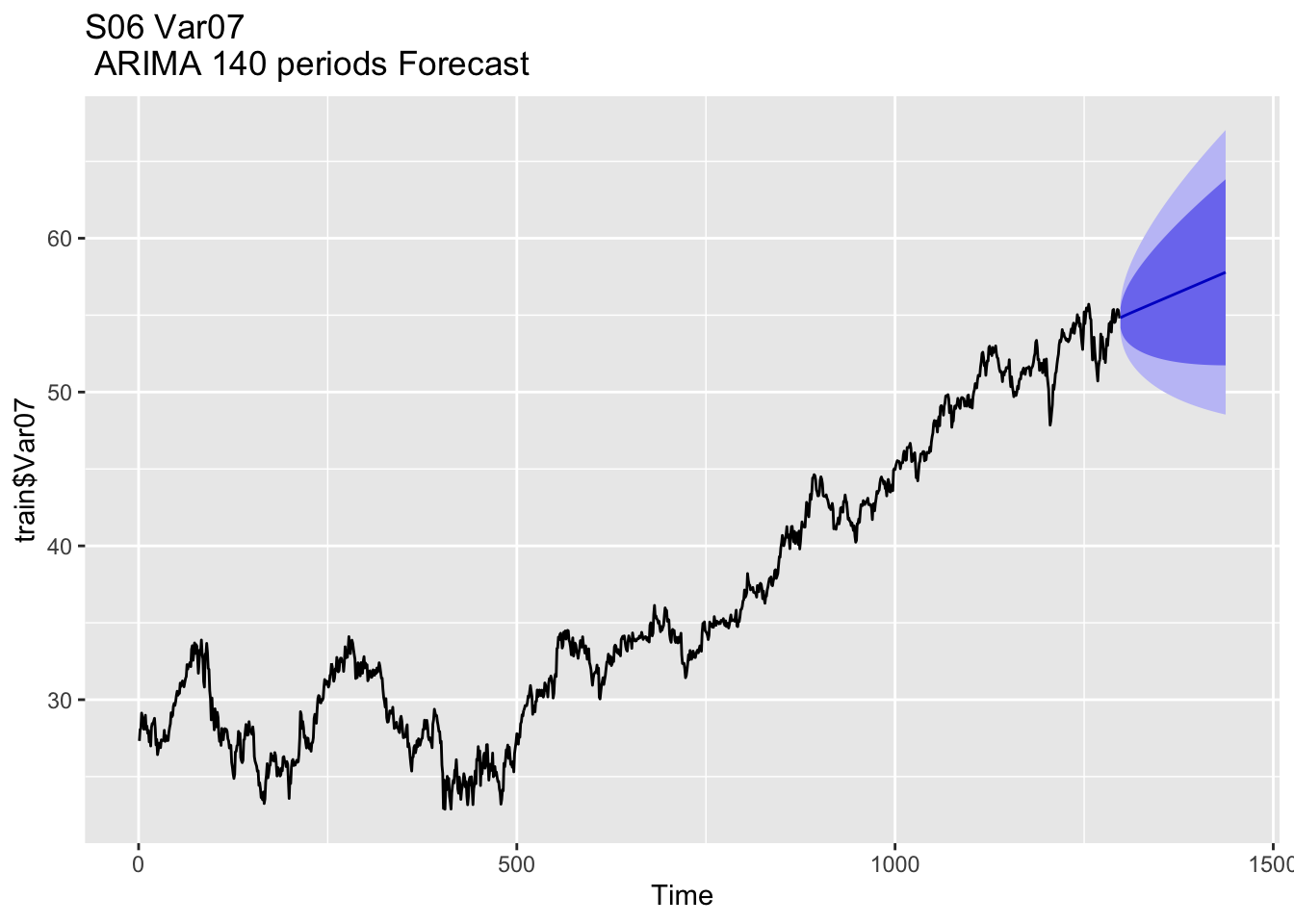
**Var 05:**

Var 05’s model predicts a gradual upward trend in the data with relatively high certainty due to small confidence intervals.



**Var 07:**

The results for Var 07 are very similar to Var 05, with a predicted gradual upward trend in the data as well as small confidence intervals suggesting strong certainty.



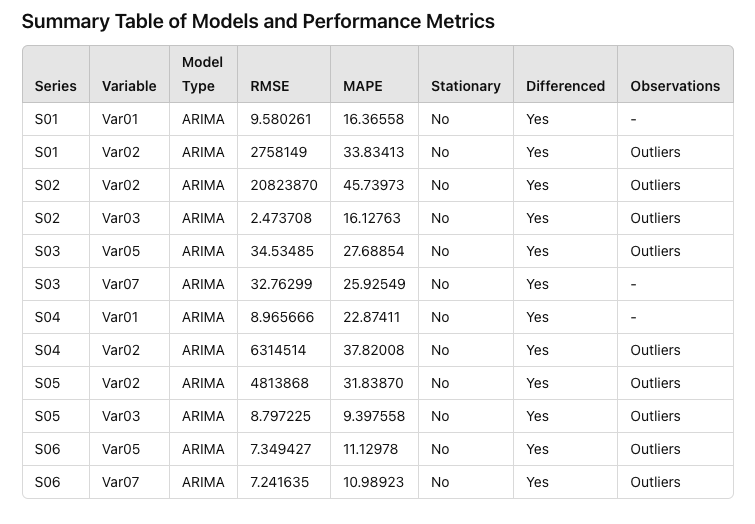
# Results & Conclusion

**RMSE:** Measures the average magnitude of errors. Lower values indicate better accuracy. Imagine throwing a ball at a target, RMSE tells you how far your throws are from the target on average. Smaller numbers mean you are closer to the target more often.

**MAPE:** Expresses forecast accuracy as a percentage. Lower percentages indicate more accurate forecasts. (Mean Absolute Percentage Error): Think about how frequently you miss the target compared to how far you threw the ball. Smaller percentages mean your throws are more accurate

With an extensive model selection process, and a thorough accuracy analysis, we believe that the ARIMA models chosen effectively encapsulate the trends and seasonality within the data. The forecasts conducted provide very valuable insight and if used correctly, could be very beneficial for stakeholders looking to gain insight into future possibilities. Overall, the models produced mixed levels of confidence, and this is important to consider when reviewing them.

Please see below the summary table for models used for each variable and performance metrics on validation sets.



# 

# Appendix

R code used for analysis

- [S01](https://rpubs.com/lindaqiao/1198027)

- [S02](https://rpubs.com/lindaqiao/1198038)

- [S03](https://rpubs.com/lindaqiao/1198129)

- [S04](https://rpubs.com/lindaqiao/1198078)

- [S05](https://rpubs.com/lindaqiao/1198114)

- [S06](https://rpubs.com/lindaqiao/1198135)