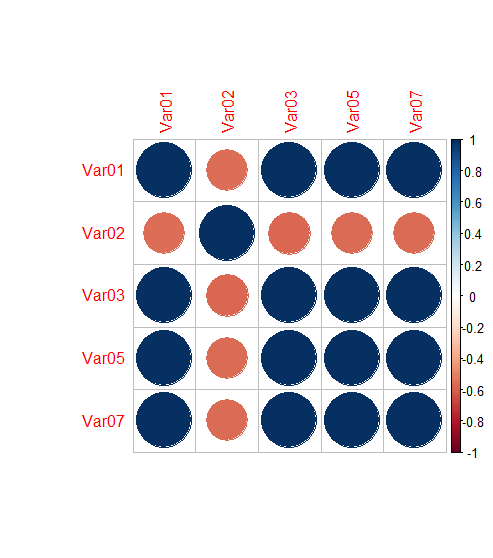
**Objectives**

The firm of *Beat the Market Brokerage* has commissioned *624 Group C Associates* to analyze, model and forecast the future direction of several data sets. To ensure strictest confidentiality in this work, our client has masked the data names and definitions. In this report, we will describe our findings related to these data sets. In particular, we will describe the data and its key characteristics, and then build models which we will use to forecast data for future time periods.

**Data Characteristics**

We first explored the data to get some basic details about the data types, frequencies, and quality. In particular, we wanted to identify any unusual characteristics or patterns that might influence how we prepared and cleaned the data and thereafter selected the model.

The full data set includes several series of data points (Var 1, 2, 3, 5, 7) organized into six groups (S01-06) each covering a period of 1622 dates. *Beat the Market* has asked us to forecast 12 data sets in all – two separate data sets within each groups – over 140 time periods. By observing trends according to the time series index, we observed that the data typically comes in sets of five followed by gaps of two-three. We have concluded that the data reflects a five day business week followed by weekends (regular or holiday.) Therefore, 1622 time periods covers approx. 6.5 years of data, and our forecasts will cover 140 business days, approx. 6.5 months of data. On average, there were 252 days of data per year.

We also noticed that within each group that nearly all the data sets – Var 1, 3, 5 and 7 (i.e. all except Var 2) were strongly correlated. Figure 1 shows that correlation – the darker blue circles signifying very strong correlation. While this will not affect our modeling process, since we model each data point separately, it raises the question of whether the data are related in some way – either variations of the same data set, or some strong relationship between the two. Var 2 was negatively related to the other data elements.

We have reviewed the data to assess any “missing” data. Each data set only has 2-3 missing values. We have imputed values for those data to ensure the cadence of the data remains consistent. Given the relatively low number of missing observations, it did not have a material impact on the substantive piece of our work – the modeling and forecasting.

Figure 1: Correlogram Displaying Correlation between Each Variable

**Data Trends and Manipulation**

To ensure a successful model, we analyzed the data to understand various key trends and separate out general trends from seasonal trends and random fluctuations.

We assessed the data for any major variance or fluctuations. Major fluctuations can significantly impact the success of a model. This allows us to determine whether we need to transform the data to make it more ready for analysis and not be unduly influenced by the outlier observations or variances. It was clear that Var 2 was very volatile while the others showed more stable trends over periods.

All the data sets showed some seasonal trends, but for some data sets, the trend was clearer than others. For data sets Var 1, 3, 5 and 7, we see clear seasonal trends, generally rising to a peak in the early months of the calendar year before declining to a low towards the end of the calendar year. For data set Var 2, we again see a lot more volatility in the seasonal trend similar to the general plot.

To illustrate these distinctions, below in Figures 2 and 3 we show two different data sets – Var 1 and Var 2 within group S01, “decomposed” into separate elements – the overall pattern of the data (row 1); broken into a general trend of the data (row 2), a seasonal trend (row 3) and other random fluctuations (row 4.) You will notice that Var 1 is relatively stable – both in general and per its seasonal trends; whereas Var 2 shows much more volatility generally and within the seasonal trends.

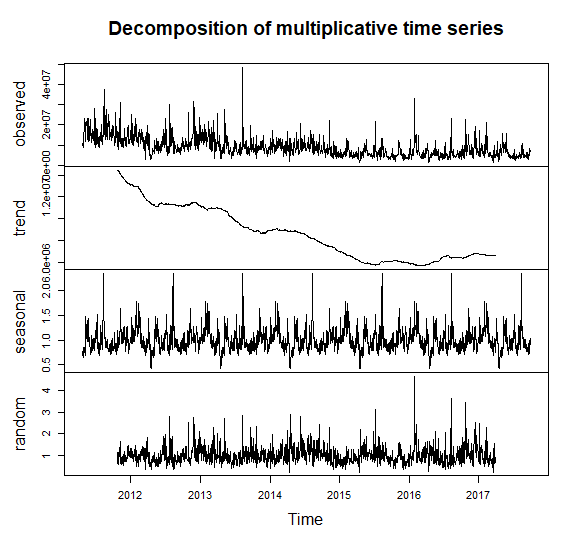
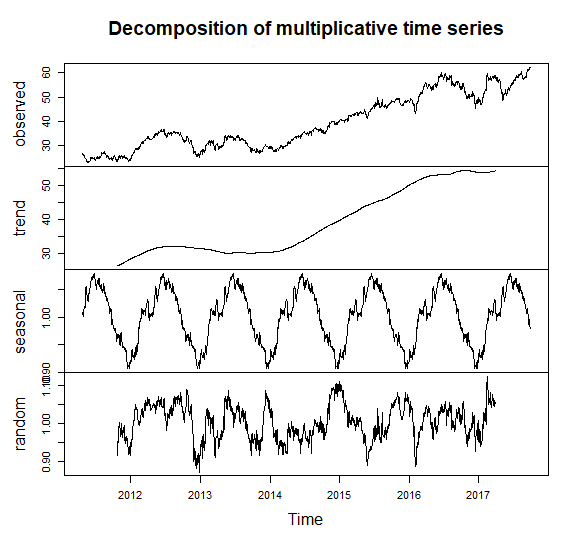


Figure 2: Decomposing Var 1 (S01) Figure 3: Decomposing Var 2 (S01)

**Models and Forecasts**

A successful model will use past data and the fluctuations and errors within forecast models to provide a strong forecast for future data sets, weighing more recent evidence more strongly than older observations. We developed several models using those elements as part of industry standard forecasting techniques (ARIMA, ETS, STL.)

We developed functions that allowed us to rapidly test all these different types of forecasting models using the past data and a variety of other settings. For example, we tested different time frequencies (e.g. daily (5) vs monthly (12) vs annual (252)), and multiple frequencies, the inclusion of a seasonal component, transformations to smooth the data, and other more technical inputs and assessments. We tested those models using industry-standard diagnostic models to ensure the models were appropriate.

In the end, we selected a model that could handle seasonality (STLF) together with ARIMA as the underlying method. We further transformed the data (Box Cox) to smooth it. Across all data sets, this model worked better than all the others we tested. Below is a summary of the forecasts based on the best performing models.

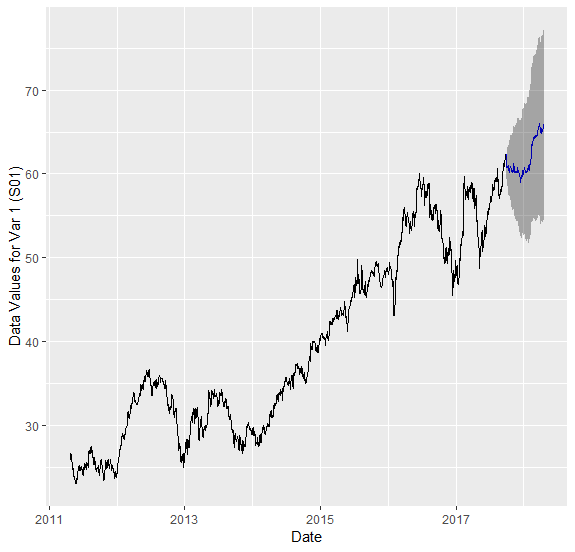
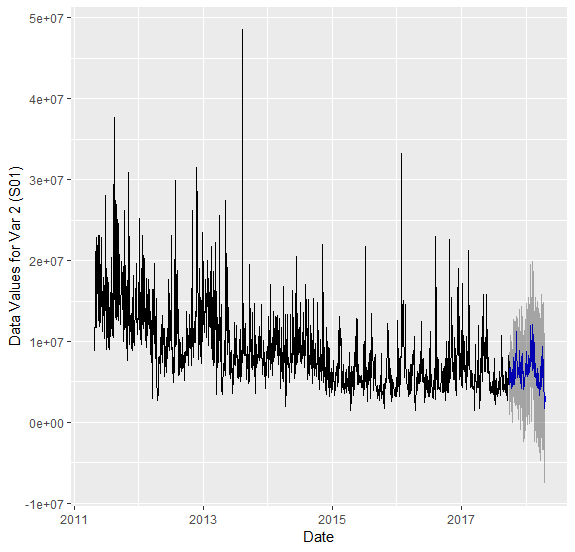
|  |  |  |  |
| --- | --- | --- | --- |
| Group | Var | Forecast | Projection |
| 1 | 1 |  | The data will continue its upward trend |
| 2 |  | The data will remain volatile and eventually decline |
| 2 | 2 |  | The data will remain volatile but hover around the same value |
| 3 |  | The data will remain relatively constant |
| 3 | 5 |  | The data will rise modestly but not yet to previous highs |
| 7 |  | The data will rise modestly but not yet to previous highs |
| 4 | 1 |  | The data will rise modestly but not yet to previous highs |
| 2 |  | The data will remain volatile but hover around the same value |
| 5 | 2 |  | The data will remain volatile but hover around the same value |
| 3 |  | The data will fall and then rise again to current values |
| 6 | 5 |  | The data will rise modestly but not yet to previous highs |
| 7 |  | The data will rise modestly but not yet to previous highs |

In the enclosed spreadsheet you will find the forecasted values for each data set and below you will find graphs forecasting the data trends for each of the data sets. Since these are forecasts we are also providing ranges of potential outcomes for those data so that decision makers can assess a broad swath of potential outcomes.

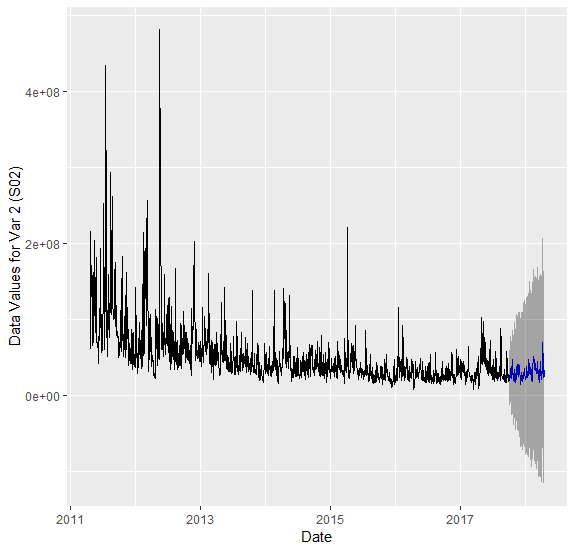
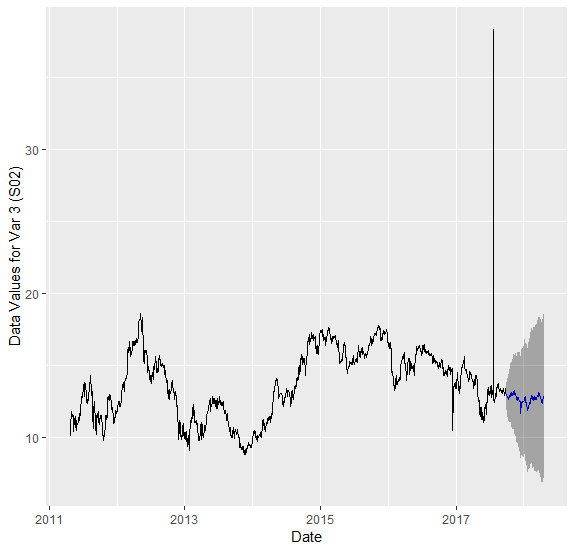
**Limitations**

Forecasts are based in large part on historical data. Naturally, past performance cannot guarantee future performance. Unforeseen events can have a big impact on the future values of these data sets. Since we are not fully familiar with the source of the data, we are not equipped to comment on the quality of the data collection or the reliability of the source data.

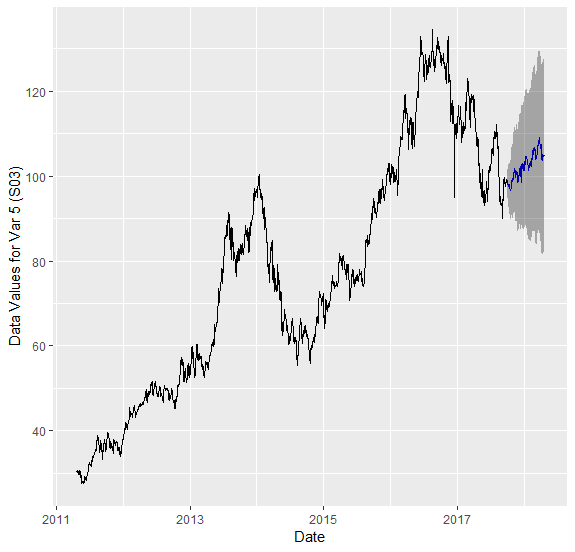
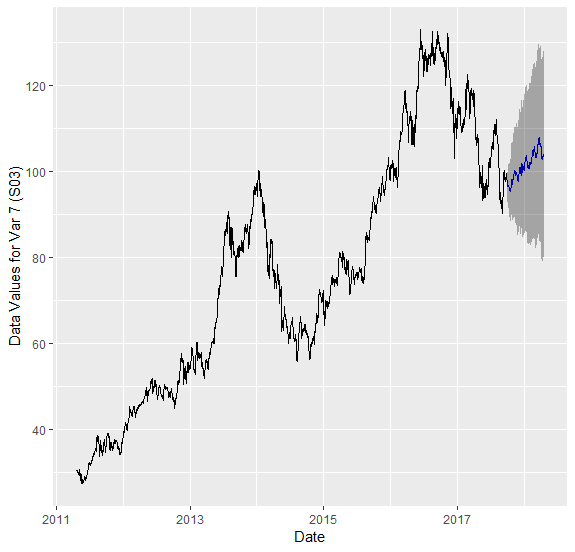
Group 1 Forecasts

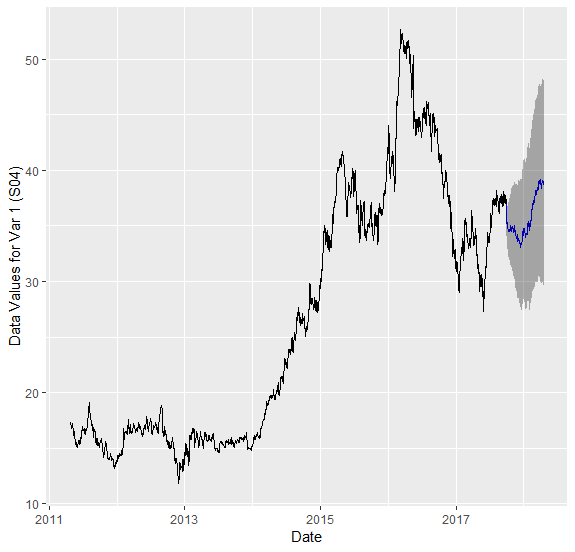
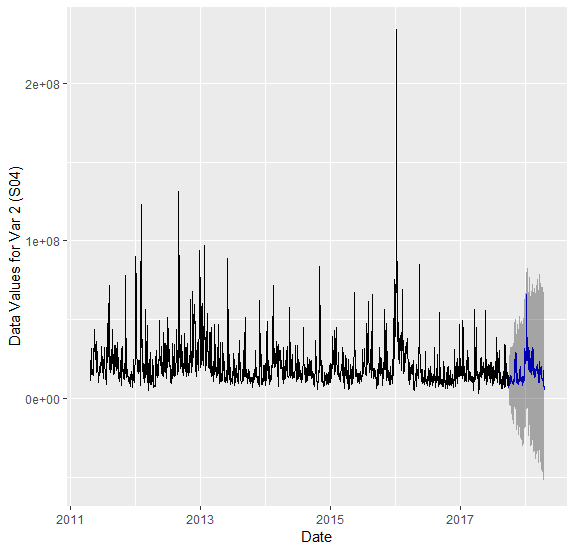
Group 2 Forecasts

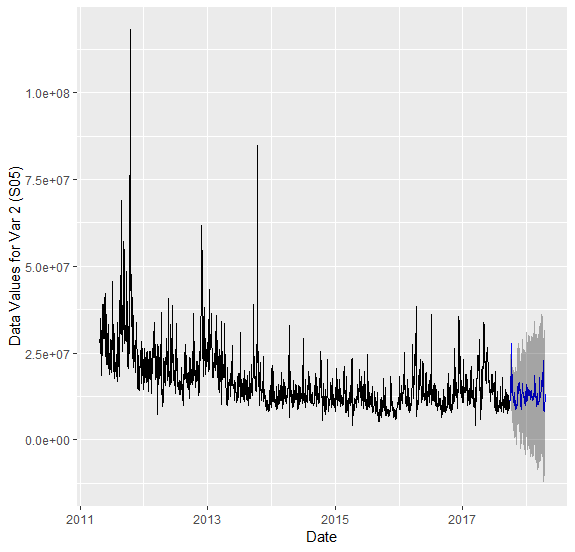
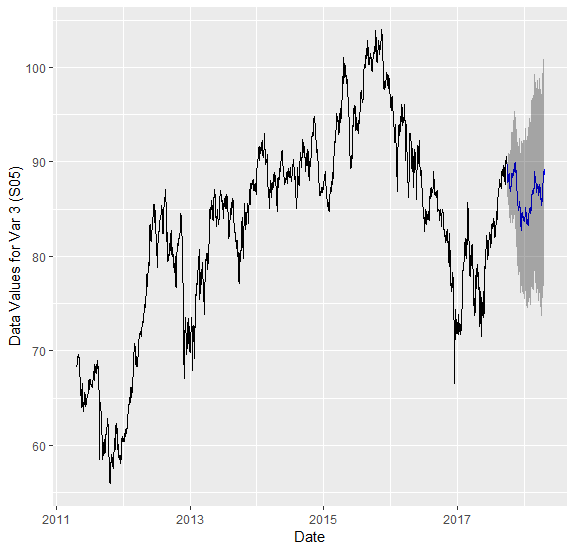
Group 3 Forecasts

Group 4 Forecasts

Group 5 Forecasts

Group 6 Forecasts

