Data624 - Project 1

Esteban Aramayo, Coffy Andrews-Guo, LeTicia Cancel, Joseph Connolly, Ian Costello

2022-06-21

Table of Contents

# Project Summary

De-identified data was provided to conduct a series of six forecasts of different variables of a provided data set. There are two major requirements of this project:

1. This written report
2. The forecasts and error rates

The data set does not appear to have any distinguishing labels that would indicate anything about the source or purpose of the data set. Under normal circumstances, context about data and use case is important for forecasting. Context may help identify the kinds of methods and models that would be best suited to produce an accurate result - following the “no free lunch” principle. Therefore, this dataset is completely for practice and exercising forecasting.

## Note on Libraries Used

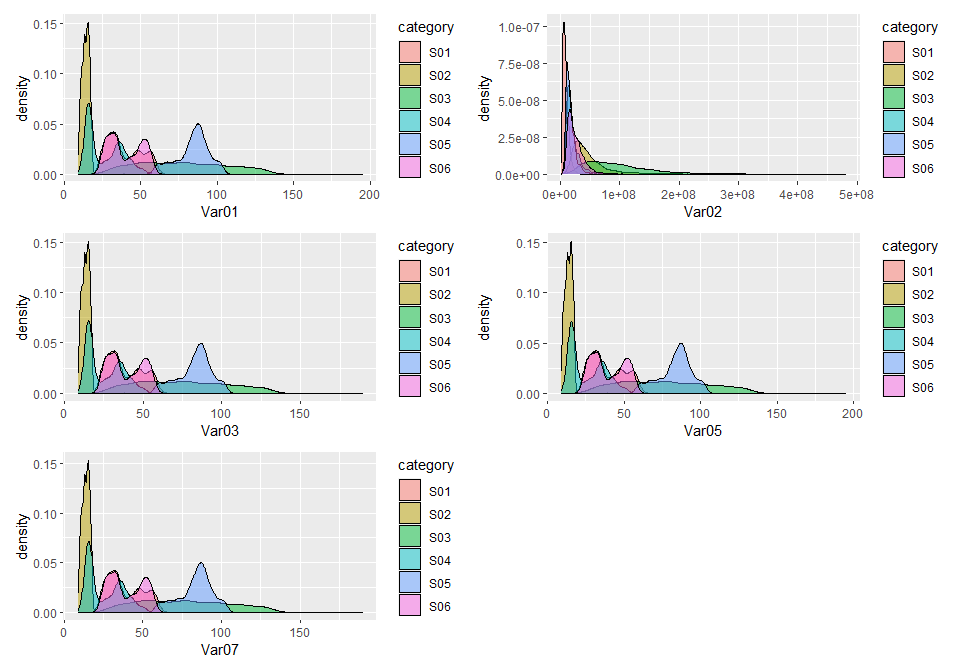
* Standard libraries
* Particularly useful libraries

# Data Exploration

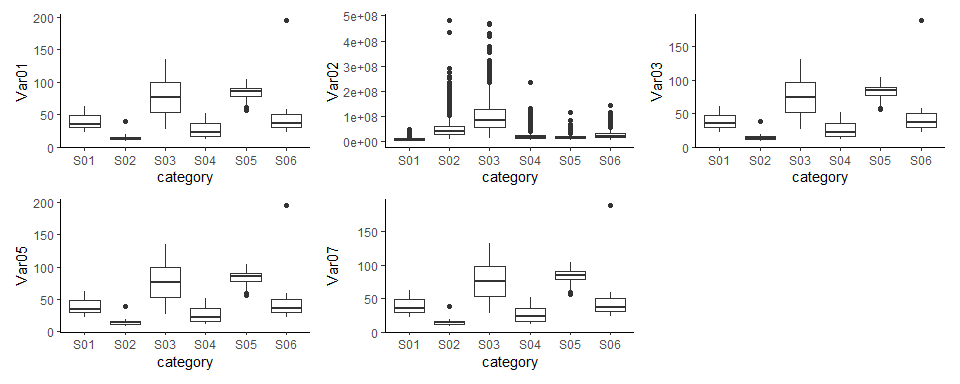
* The provided data set contains 7 columns and 10,572 rows. The first column, “SeriesInd”, is a datetime variable which can be converted to reflect a date. “category” is a classification of each row, of which there are 6 different ones, and the remaining 5 columns are variables to forecast and are the key ingredient for constructing timeseries.
* Variables missing 04 and 06… that’s a question mark

A time series that contain a list of numbers (the measurements), along with some information about what times those numbers were recorded (the index).

**Figure 1 Density Plots**



**Figure 2 Box Plots**

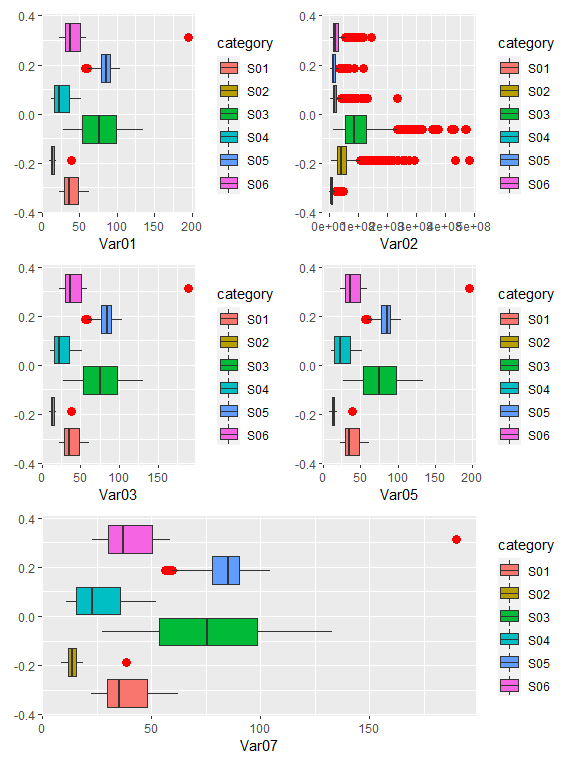


* Var02 has the most outliers. It’s also of a much larger a magnitude from the other predictors

##### Datetime conversion

* Datetime conversion is to be performed after imputation in order to ensure a successful imputation

**Figure 3 Outliers**



# Data Preparation

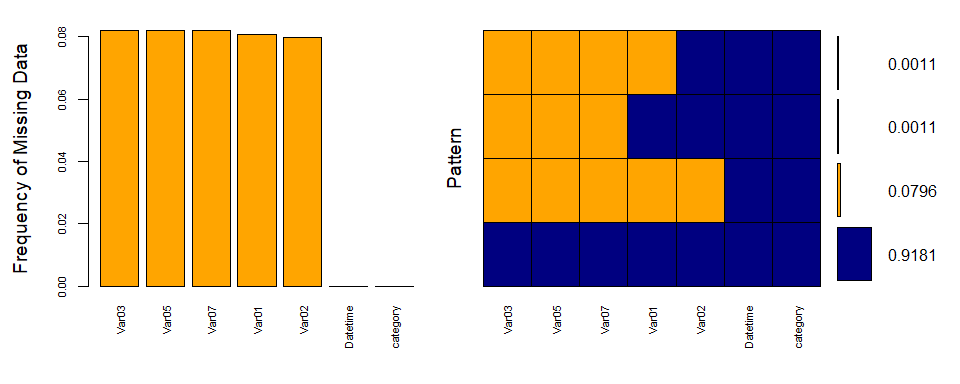
##### Handling Missing Data

#> [1] "4294 values missing from original set"

Looking at the summary generated above, columns 3-7 each have a range of 842-866 missing values. The dilemma is to decide whether or not it is appropriate to perform an analysis via imputing missing values, or to simply delete them. According to the plot below, generated via “VIM::aggr()”, 91.81% of the data is fulfilled. Var01, Var02, Var03, Var05, and Var07 are missing about 8% of data. At first impression, this seems like an insignificant amount of data that can be omitted from the set. Further investigation is needed to confirm this impression, and deletion will be deemed appropriate if and only the data is found to be missing completely at random (MCAR).

##### Impute or Delete?

An excerpt from the following paper, , by Hyun Kang, argues when deletion is appropriate or not from the following quote: *“…if the assumption of MCAR (missing completely at random) is satisfied, a listwise deletion is known to produce unbiased estimates and conservative results. When the data do not fulfill the assumption of MCAR, listwise deletion may cause bias in the estimates of the parameters. If there is a large enough sample, where power is not an issue, and the assumption of MCAR is satisfied, the listwise deletion may be a reasonable strategy. However, when there is not a large sample, or the assumption of MCAR is not satisfied, the listwise deletion is not the optimal strategy”*



#>   
#> Variables sorted by number of missings:   
#> Variable Count  
#> Var03 0.08191449  
#> Var05 0.08191449  
#> Var07 0.08191449  
#> Var01 0.08077942  
#> Var02 0.07964434  
#> Datetime 0.00000000  
#> category 0.00000000

By creating a shadow matrix to see a percentage (on a 0-1 scale) of missing values when all correlated among each other, this will help indicate whether or not the data is MCAR.

#> Var01 Var02 Var03 Var05 Var07  
#> Var01 1.0000000 0.9923369 0.9924341 0.9924341 0.9924341  
#> Var02 0.9923369 1.0000000 0.9848290 0.9848290 0.9848290  
#> Var03 0.9924341 0.9848290 1.0000000 1.0000000 1.0000000  
#> Var05 0.9924341 0.9848290 1.0000000 1.0000000 1.0000000  
#> Var07 0.9924341 0.9848290 1.0000000 1.0000000 1.0000000

Aside from considering values correlated with themselves, the following have no missing values when correlated with others:

* Var03 has no missing values when correlated with Var05 and Var07
* Var05 has no missing values when correlated with Var03 and Var07
* Var07 has no missing values when correlated with Var03 and Var05

Taking these observations into consideration, it seems there appears to be bias in the data in the context of missing values and therefore not missing completely at random. Imputing the values is an appropriate step to take.

##### Data Imputation

As stated, since the missing values are not MCAR, imputation will be utilized to represent missing values in our data set. There are numerous methods to impute data, especially in R. Within the caret package, the “preProcess()” function enables imputation, and allows for users to select a method of imputation. The method “medianImpute” was chosen because of the ease and efficiency.

#> [1] "0 values missing after imputation"

* No data are missing after imputation
* No discrepancies in the data after imputation.

## Data transformations

By using the skewness function within the moments package, we can see which predictor variable is skewed the most in quantifiable terms.

#> [1] "Var01 skewness: 0.833450080946069"

#> [1] "Var02 skewness: 3.20209362366419"

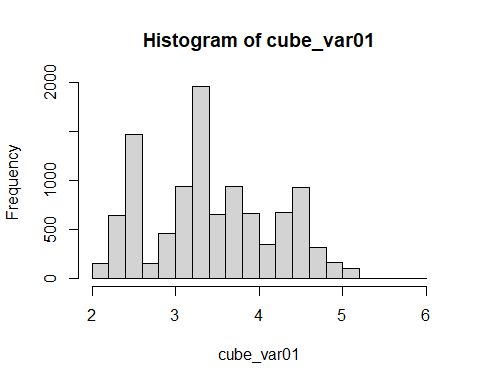
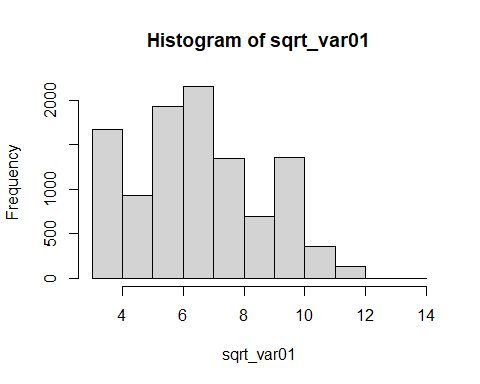
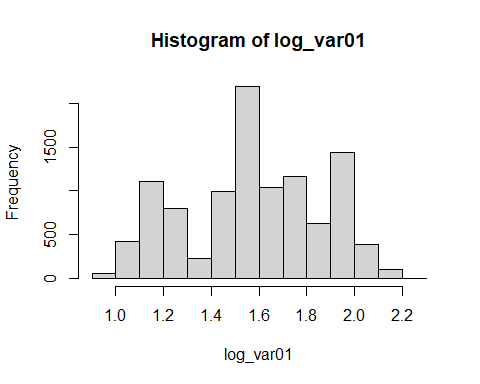
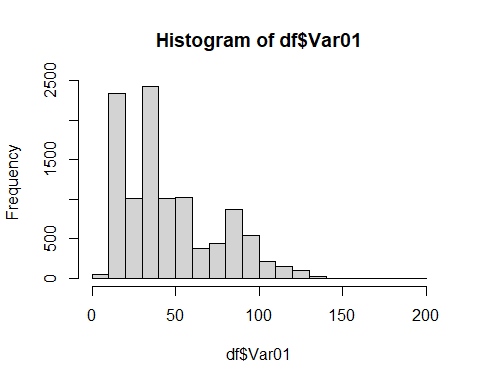
#> [1] "Var03 skewness: 0.834240845373782"

#> [1] "Var05 skewness: 0.83686517614705"

#> [1] "Var07 skewness: 0.833896603742934"

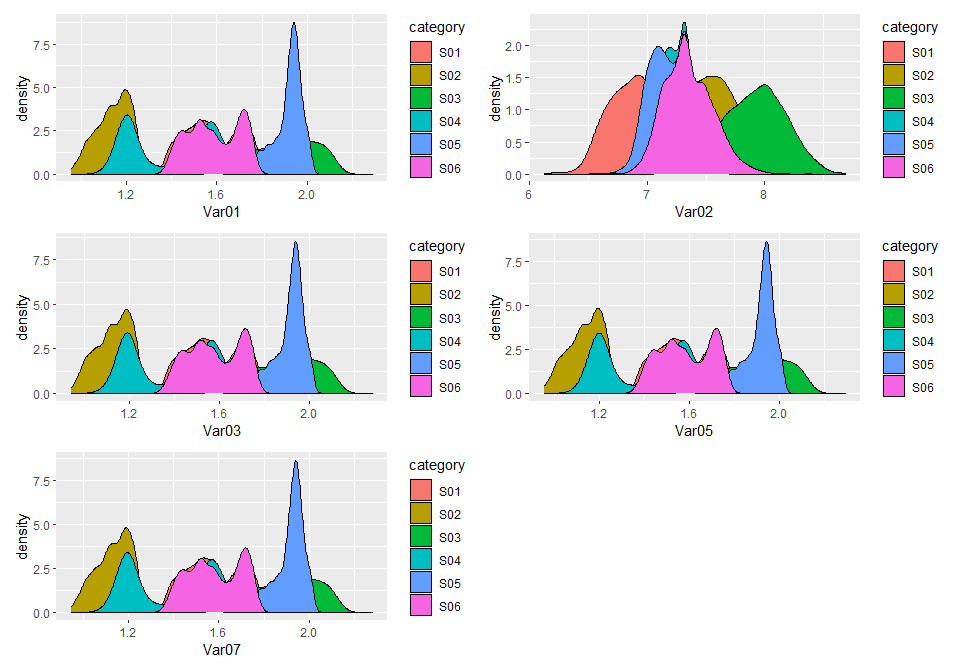
* Var02 is the most skewed

In order to improve the distribution for the predictors, we can investigate with applying 3 different transformations; log, square root, and cube root. Applying these transformations to one predictor will indicate which will provide the most normal distribution



It’s revealed the log transformation appears to follow the most normal distribution. Therefore, the log transformation will be applied to all predictor variables.

New plots using transformed dataframe. Var02 is the most normalize so we will try different transformations with the other columns



Final step before forecasting, create time series for each subset using the data frame with dates.

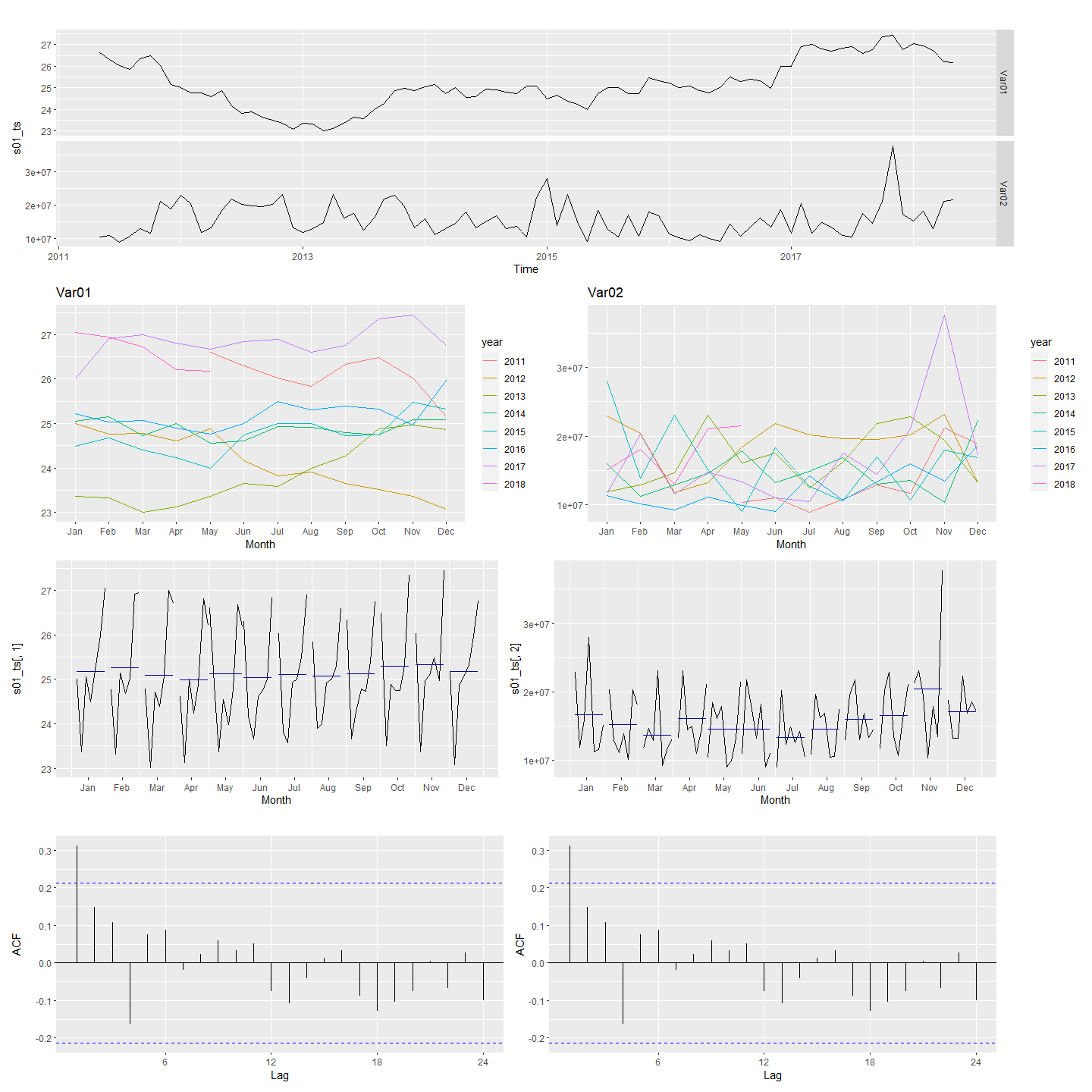
# Forecasting

Now that the data has been transformed, we can begin forecasting

##### Auto and Seasonal Plots for S01: Var01, 02

For Var01, there are a few patterns to be pointed out. Annually, the value follows a slight valley pattern as it steadily decreases from the beginning to gradually rise towards the end. Monthly, there appears to be a pattern; toward the beginning of each month, the value decreases to a significant low only to lead up to a high-point approximately 2.5 points above the bottom, with a slight correction along the way. This timeseries is also the closest to a white noise series.

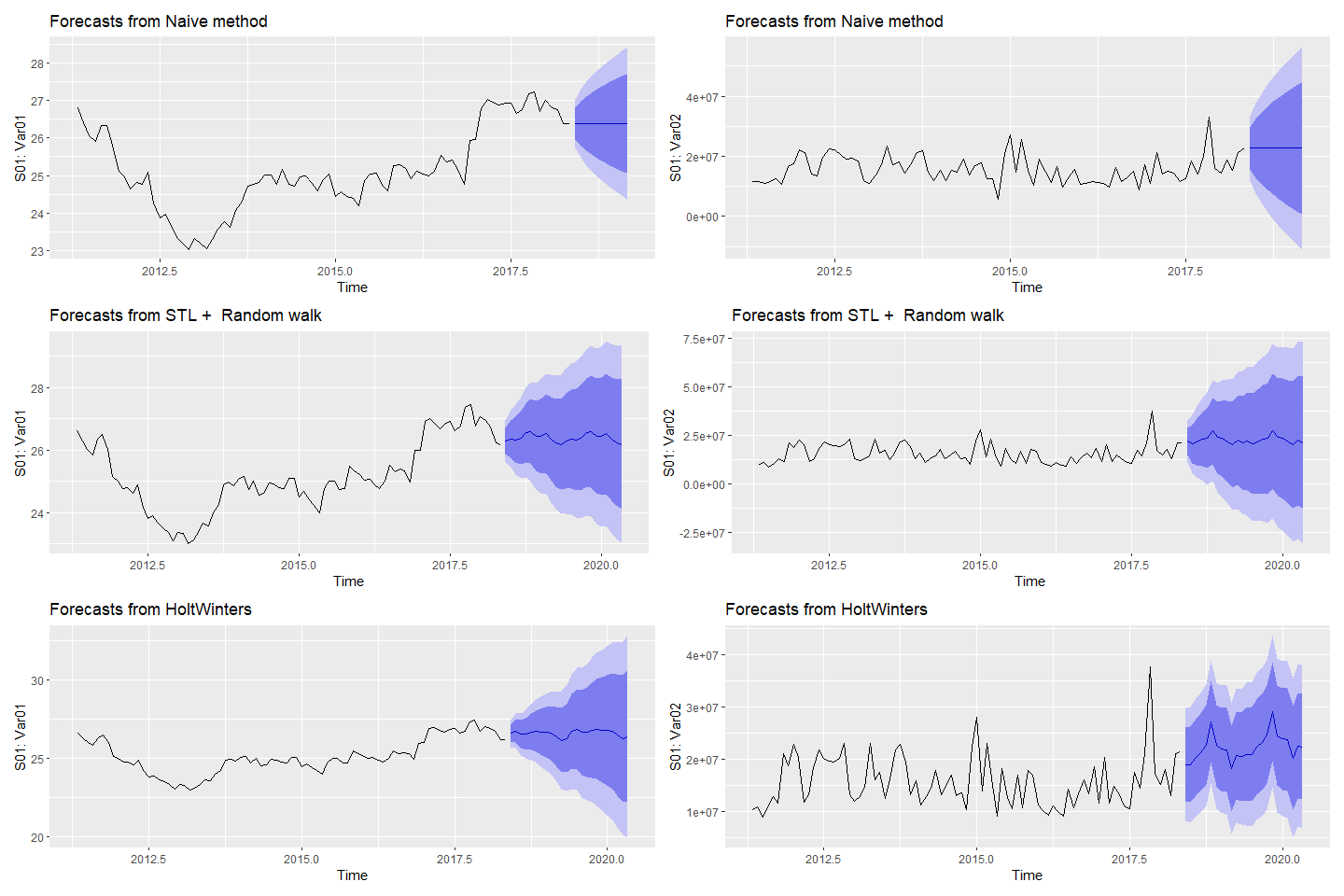
* Same approach was applied to all other variables
* Generally similar to variable 1 to illustrate approach
* Only showing variable 1 for brevity



## Data Modeling and Forecasting

* Explain Models used - Naive method, STL + Random Walk + Holt Winters
* Confidence intervals are narrow with HoltWinters, and variable on the line point forecast
* Show category 1 plots for brevity, but all other categories followed similar approach
* Final selection on Holt Winters because…

**Forecasting S01: Var01 & Var02 with decomposition**



## Other Forecast Models

* SES is… 7.1 for broad defintion
* Point forecast is straightline
* ETS
* ARIMA

| Category | Variable | ARIMA autoplot() Selection |
| --- | --- | --- |
| s01 | Var01 | ARIMA(0,1,0) |
| s01 | Var02 | ARIMA(1,0,0) w/ non-zero mean |
| s02 | Var02 | ARIMA(0,0,0) w/ non-zero mean |
| s02 | Var03 | ARIMA(0,1,2) |
| s03 | Var05 | ARIMA(0,1,0) |
| s03 | Var07 | ARIMA(0,1,0) |
| s04 | Var01 | ARIMA(0,1,0) |
| s04 | Var02 | ARIMA(1,0,0)(0,0,1)[12] w/ non-zero mean |
| s05 | Var02 | ARIMA(2,0,0) w/ non-zero mean |
| s05 | Var03 | ARIMA(0,1,2) |
| s06 | Var05 | ARIMA(0,1,1) |
| s06 | Var07 | ARIMA(0,1,1) |