Data\_624\_Project\_1

Group 4: Melvin Matanos, Claire Meyer, Chinedu Onyeka, Euclid Zhang, Jie Zou

6/11/2022

DATA 624 Summer 2022, Project #1

Project Description:

Your data is a de-identified Excel spreadsheet. Your assignment is to perform the appropriate analysis to forecast several series for 140 periods. You will have 1622 periods for your analysis.

# Data Exploration

Out data is the observations for 6 individual groups (S01, S02, S03, S04, S05, S06) during a period of length 1622. Each observation contains 5 variables (Var01, Var02, Var03, Var05, Var07).  
So there are 6\*5=30 time series (one per variable per group). In this analysis, we will focus on the following 12 time series and perform forecast the values for the next 140 periods:

* Group S01 – Var01, Var02
* Group S02 – Var02, Var03
* Group S03 – Var05, Var07
* Group S04 – Var01, Var02
* Group S05 – Var02, Var03
* Group S06 – Var05, Var07

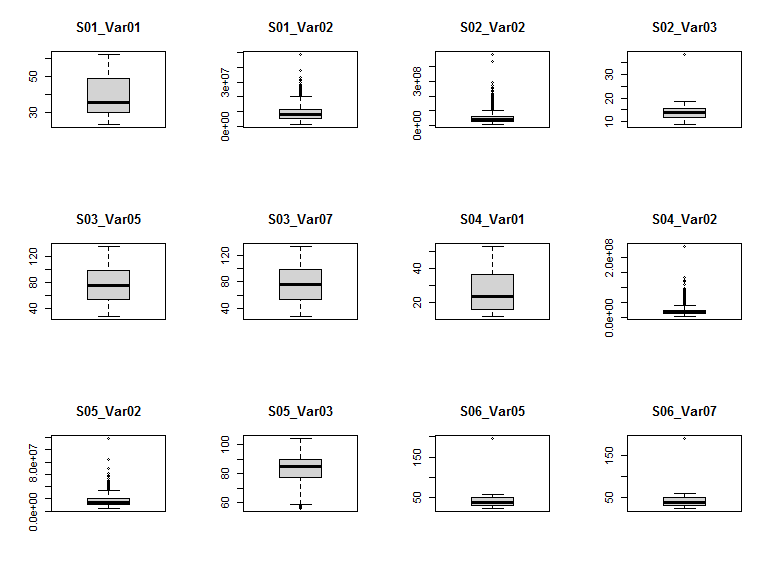
First, let look at the summary of the data:

## SeriesInd group Var01 Var02   
## Min. :40669 S01:1622 Min. : 9.03 Min. : 1339900   
## 1st Qu.:41253 S02:1622 1st Qu.: 23.10 1st Qu.: 12520675   
## Median :41846 S03:1622 Median : 38.44 Median : 21086550   
## Mean :41843 S04:1622 Mean : 46.98 Mean : 37035741   
## 3rd Qu.:42430 S05:1622 3rd Qu.: 66.78 3rd Qu.: 42486700   
## Max. :43021 S06:1622 Max. :195.18 Max. :480879500   
## NA's :14 NA's :2   
## Var03 Var05 Var07   
## Min. : 8.82 Min. : 8.99 Min. : 8.92   
## 1st Qu.: 22.59 1st Qu.: 22.91 1st Qu.: 22.88   
## Median : 37.66 Median : 38.05 Median : 38.05   
## Mean : 46.12 Mean : 46.55 Mean : 46.56   
## 3rd Qu.: 65.88 3rd Qu.: 66.38 3rd Qu.: 66.31   
## Max. :189.36 Max. :195.00 Max. :189.72   
## NA's :26 NA's :26 NA's :26

We have some missing values in each variables, these missing values will be imputed / filled with reasonable values before building our models.

The variables also have maximum values much larger than the 3rd quartile (the value that is larger than 75% of all values in the same variable), there may be extreme outliers in the data.

We can check the outliers and also the skewness of the variables form the boxplots.  
The following are the boxplots of the 12 time series.

 Var02 of all groups are highly right skewed that may need to be transformed to stabilize the variance.  
S02\_Var03, S06\_Var05, and S06\_Var07 have an extreme outlier.  
If the extreme outliers are excluded, all Variables except Var02 have stable variance and no transformation is needed.  
We will remove the extreme outliers and impute them with reasonable values along with other missing values.

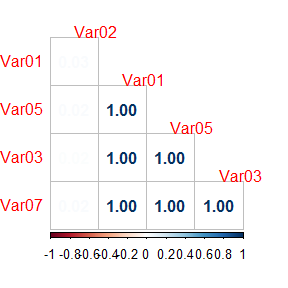
# Data Preparation

## Missing Values Imputation

The following are the records with missing value for our original data.

## SeriesInd group Var01 Var02 Var03 Var05 Var07  
## 118 40697 S06 NA NA NA NA NA  
## 4769 41821 S05 NA NA NA NA NA  
## 9217 42897 S03 NA 42343600 NA NA NA  
## 9218 42897 S02 NA 38160300 NA NA NA  
## 9219 42897 S01 NA 7329600 NA NA NA  
## 9220 42897 S06 NA 19885500 NA NA NA  
## 9221 42897 S05 NA 16610900 NA NA NA  
## 9222 42897 S04 NA 9098800 NA NA NA  
## 9223 42898 S03 NA 50074700 NA NA NA  
## 9224 42898 S02 NA 45801300 NA NA NA  
## 9225 42898 S01 NA 6121400 NA NA NA  
## 9226 42898 S06 NA 32570900 NA NA NA  
## 9227 42898 S05 NA 19331600 NA NA NA  
## 9228 42898 S04 NA 11188200 NA NA NA  
## 9637 42997 S03 95.43 32026000 NA NA NA  
## 9638 42997 S02 13.26 19465000 NA NA NA  
## 9639 42997 S01 58.83 6337000 NA NA NA  
## 9640 42997 S06 49.21 13222800 NA NA NA  
## 9641 42997 S05 90.40 13191900 NA NA NA  
## 9642 42997 S04 36.72 34330700 NA NA NA  
## 9643 43000 S03 97.19 38018600 NA NA NA  
## 9644 43000 S02 13.20 16234300 NA NA NA  
## 9645 43000 S01 59.28 3690900 NA NA NA  
## 9646 43000 S06 48.88 10644000 NA NA NA  
## 9647 43000 S05 89.90 11766100 NA NA NA  
## 9648 43000 S04 36.95 7785800 NA NA NA

We can checking the correlations between variables. A correlation close to 1 or -1 is considered as a strong correlation between two variables. That is, the value of one variable is highly dependent to the other variable. We can use linear models to impute the missing values of one variable using another variable.



From the correlation plot, Var03, Var05, and Var07 are highly correlated to Var01.  
Var02 seems to be independent to the other variables.  
We can impute the missing values of Var03, Var05, and Var07 where Var01 is available.  
For records with all Var01, Var03, Var05, and Var07 missing, We will fill in the missing values using the linear interpolation method.

The linear interpolation method connects the previous and next available values by a straight line and fill in the missing values by the points fall on the line. For example, we have the following 2 missing values for group S01 Variable 01:

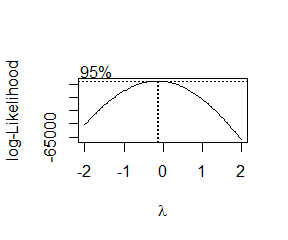
## S01\_Var01 S01\_Var02  
## 42891 51.62 9795800  
## 42892 51.41 7929800  
## 42895 51.23 15765600  
## 42896 51.63 7321700  
## 42897 NA 7329600  
## 42898 NA 6121400  
## 42899 51.40 8060600  
## 42903 52.77 6567400  
## 42904 53.72 6244700  
## 42905 53.82 6004200  
## 42906 52.94 6300700

The missing values are filled by 51.55 and 51.47 which are on the line connecting 51.63 and 51.4 with equal distance.

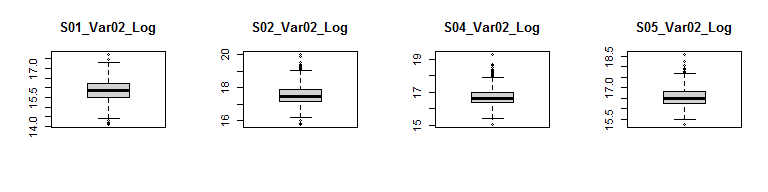
## S01\_Var01 S01\_Var02  
## 42891 51.62000 9795800  
## 42892 51.41000 7929800  
## 42895 51.23000 15765600  
## 42896 51.63000 7321700  
## 42897 51.55333 7329600  
## 42898 51.47667 6121400  
## 42899 51.40000 8060600  
## 42903 52.77000 6567400  
## 42904 53.72000 6244700  
## 42905 53.82000 6004200  
## 42906 52.94000 6300700

## Data Transformation

As we have seen from the boxplots above, Var02 is highly right skewed. We will transform the variable using the Box-Cox Transformation method. The Box-Cox Transformation finds the optimal power that can be applied to the data so that the transformed data is close to the normal distribution.

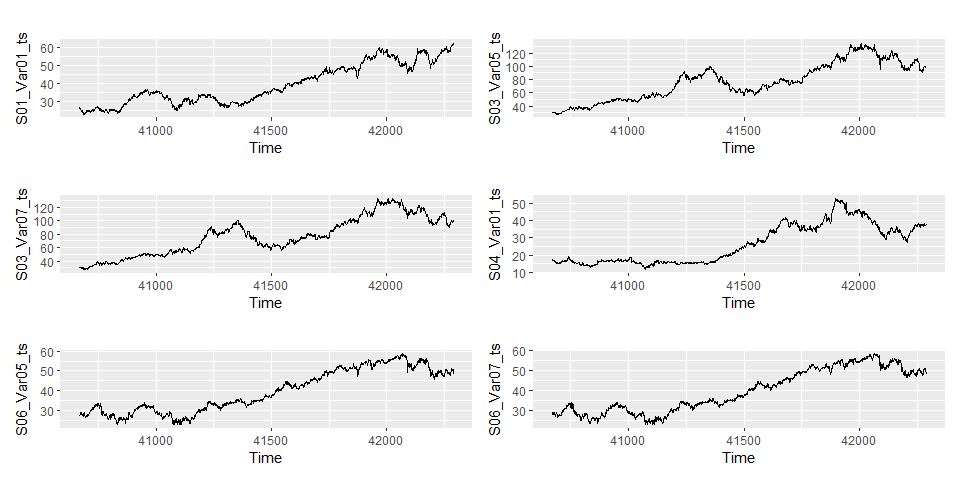


A parameter close to 0 suggests that a log-transformation is appropriate.

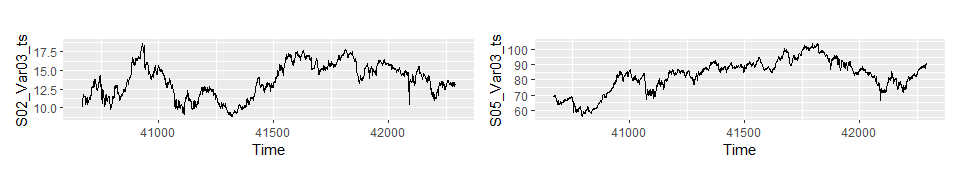
The followings are the boxplots of the log-transformated values for Var02 time series  The variance is much stabler than before.

# Time Series Exploration

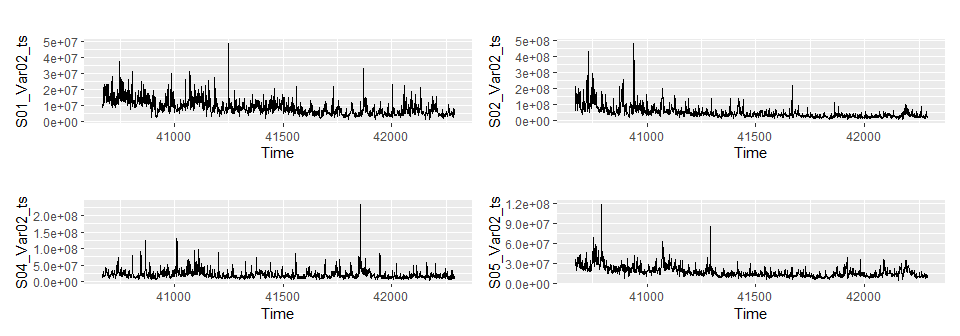
## Time Plots



* There is no apparent seasonal behaviours in the time series for Var01, Var05, and Var07.
* There are apparent trends in the time series for Var01, Var05, and Var07.



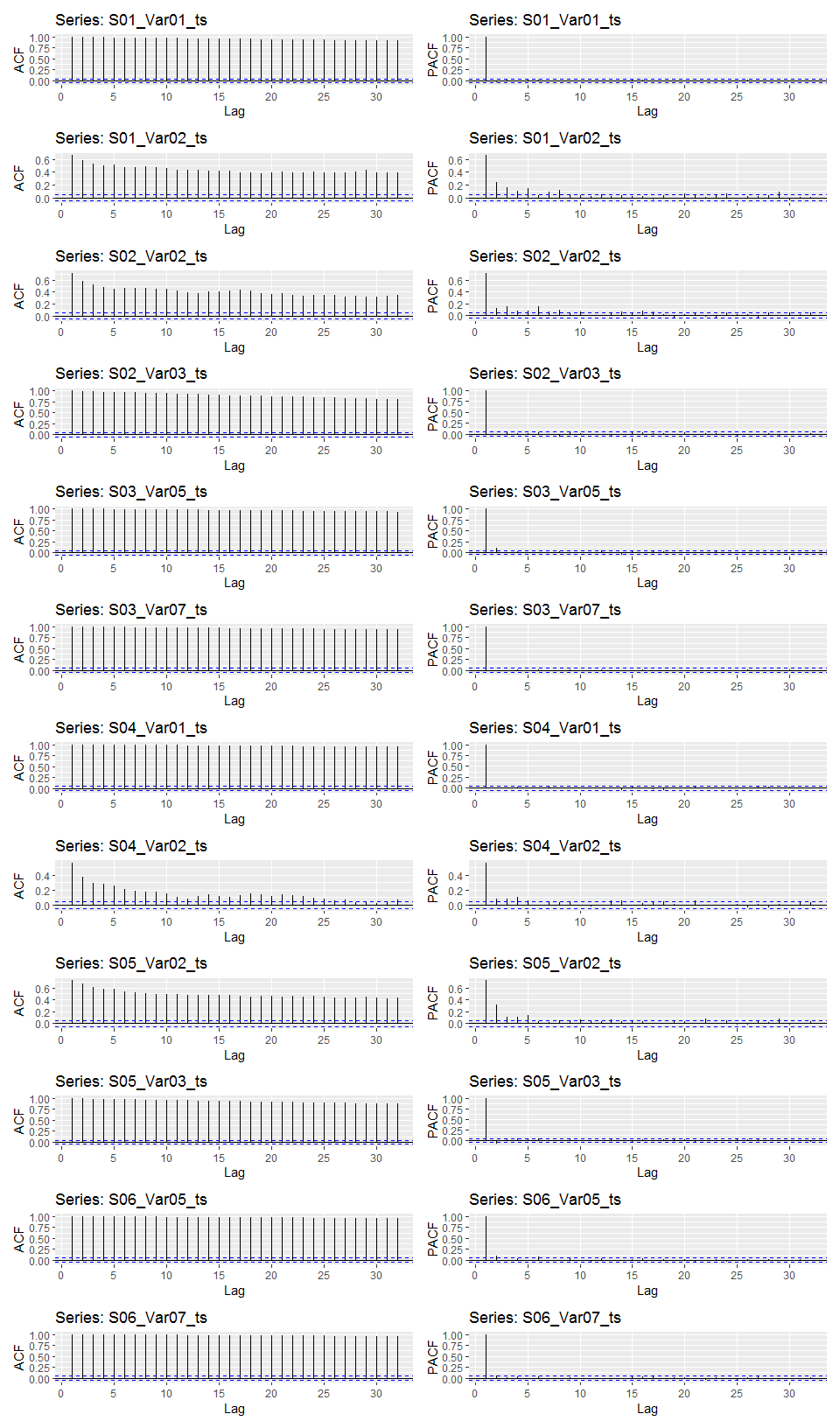
* There is no apparent seasonal behaviours in the time series for Var03.
* Var03 seems to have cyclic behaviours instead of trends.



The time series for Var02 do not have apparent patterns. We may need to check the autocorrelations to verify if they are stationary.

## ACF and PACF Plots

The ACF (Auto Correlation Function) plot and PACF (Partial Auto Correlation Function) show how strong that a value in a time series depends on its past values. A time series with high autocorrelation is non-stationary and hence there are predictable patterns.



All time series have significant autocorrelations (the ones with correlation higher than the threshold indicated by the blue dashed lines) for multiple lags in the ACF plots. The autocorrelations in the PACF plots are not so strong. We confirm that the time series are non-stationary.

# Building models

## Modeling Approach

The are generally two popular types of time series models: *Exponential Smoothing* (ETS) model and *ARIMA* (AutoRegressive Integrated Moving Average) model. The fitted values of an *Exponential Smoothing* model are affected by all past values, while the fitted values of an *ARIMA* model are affected by a number of most recent past values.

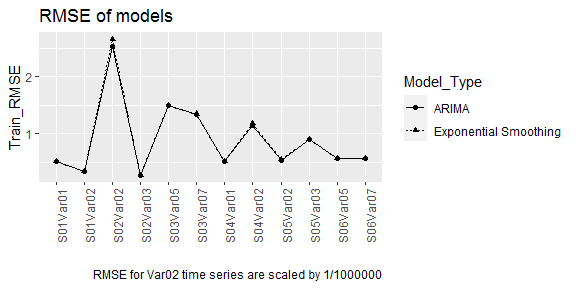
For each of the 12 time series, we perform the followings:

* Build the optimal *Exponential Smoothing* (ETS) model (log-transformation applied to the Var02 to stabilize the variance).
* Build the optimal *ARIMA* model (log-transformation applied to the Var02 to stabilize the variance).
* Perform cross-validation on the modeling method of *Exponential Smoothing*. Cross-validation is used to verify how well the model performs with unseen data.
* Perform cross-validation on the modeling method of *ARIMA*. Cross-validation is used to verify how well the model performs with unseen data.
* Compare the RMSE (Root Mean Squared Error) of models from the training data and cross-validation.
* Verify if there is any lack of fit by checking the residuals from the models.
* Select the most appropriate model for forecasting.

## Model Performance

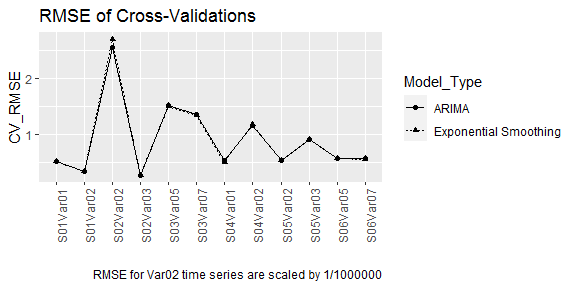
The RMSE (Root Mean Squared Error) measures the distance between a model’s fitted values and the actual values. Hence, smaller RMSE usually implies better performance.

The following are the RMSE values of the models



The RMSE from the models are used to check the performance of the models based on the *training data*. From the plot, the *ARIMA* models and \*Exponential Smoothing\*\* models have very close performance. The *ARIMA* models are slightly better.

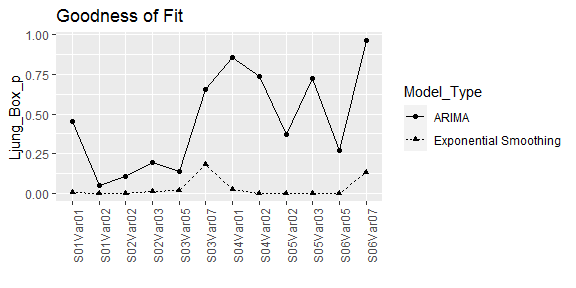
The following are the RMSE values from the outputs of Cross-Validation



The RMSE from the models are used to check the performance of the models with *unseen data*. From the plot, the *ARIMA* models and \*Exponential Smoothing\*\* models have very close performance. The *ARIMA* models are slightly better.

## Goodness of Fit

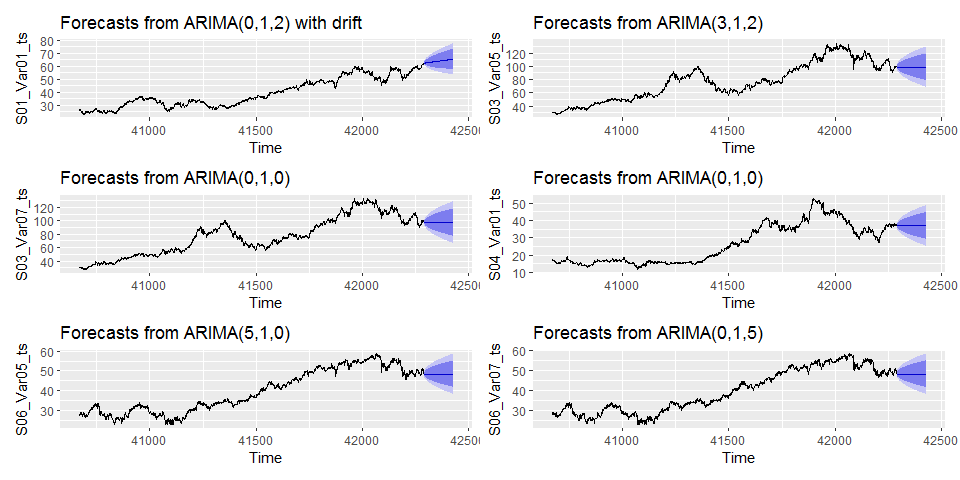
We can also verify how well the models fit in the data. We use the Ljung Box test on the residuals of the models. The null hypothesis of the test is that there is no significant autocorrelations in the model’s residuals, that is, the model explains all the variance / patterns of the data. A higher p-value implies that the model is fitting the data better. The p-values of the tests for the models are showed in the plot below.

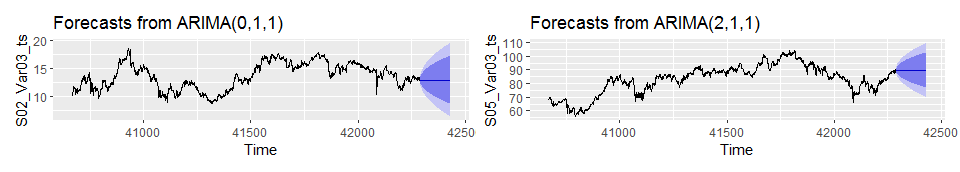


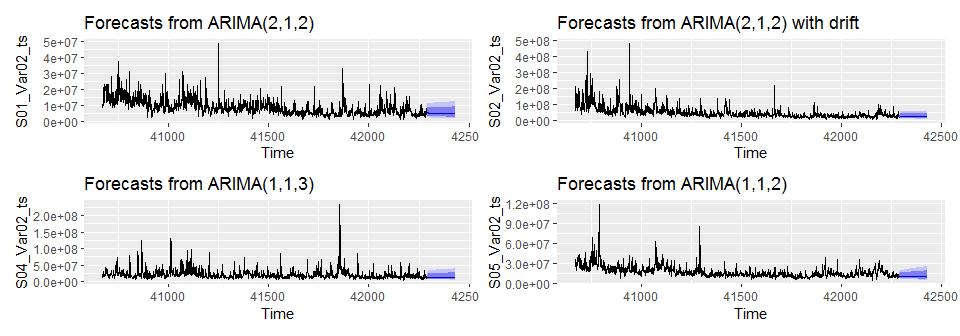
Based on the ljung-box test p-values, the ARIMA models are fitting to the data better so we will choose the ARIMA models for forecasting.

# Model Forecast

We forecast the values for the 12 times in the next 140 periods.  
The solid line represents the expected values. The dark blue area represents the 80% confidence interval, which indicates that there is 80% chance that the forecasted values will fall within that area. The light blue area represents the 95% confidence interval, there is 95% chance that the forecasted values will be in the area.







# Codes

# Loading packages  
library(fpp2)  
library(dplyr)  
library(tidyverse)  
library(corrplot)  
library(MASS)  
library(imputeTS)  
library(patchwork)  
library(ggplot2)  
  
  
  
# Loading the data set  
raw\_df <- readxl::read\_excel("Data Set for Class.xls")  
raw\_df <- as.data.frame(raw\_df)  
raw\_df$group <- as.factor(raw\_df$group)  
  
# Summary of the first 1622 periods. The remaining 140 periods are blank and need to be forecasted  
raw\_summary <- summary(raw\_df[c(1:(1622\*6)),])  
raw\_summary  
  
  
  
# Missing values in the observations  
raw\_na <- raw\_df[c(1:(1622\*6)),][apply(is.na(raw\_df[c(1:(1622\*6)),]),1,any),]  
raw\_na  
  
  
  
# Checking the correlations between variables. If the correlations are high, we can use linear models to impute the missing values of one variable using another variable.  
corrplot(cor(raw\_df[,c(3:7)], use = "na.or.complete"),   
 method = 'number', order = "hclust",  
 type = 'lower', diag = FALSE, tl.srt = 0.1)  
  
  
  
# Impute the missing values of Var03, Var05, Var07, where Var01 is available, using linear models  
  
var03\_lm <- lm(Var03~Var01,raw\_df)  
var05\_lm <- lm(Var05~Var01,raw\_df)  
var07\_lm <- lm(Var07~Var01,raw\_df)  
  
raw\_df$Var03[!is.na(raw\_df$Var01) & is.na(raw\_df$Var03)] <-  
 predict(var03\_lm,raw\_df[!is.na(raw\_df$Var01) & is.na(raw\_df$Var03),])  
raw\_df$Var05[!is.na(raw\_df$Var01) & is.na(raw\_df$Var05)] <-  
 predict(var03\_lm,raw\_df[!is.na(raw\_df$Var01) & is.na(raw\_df$Var05),])  
raw\_df$Var07[!is.na(raw\_df$Var01) & is.na(raw\_df$Var07)] <-  
 predict(var03\_lm,raw\_df[!is.na(raw\_df$Var01) & is.na(raw\_df$Var07),])  
  
  
  
# Gather data into one data frame, with one column per group per selected variable  
S01\_Var01 <- raw\_df %>% filter(group=="S01") %>% dplyr::select("SeriesInd","Var01")  
S01\_Var02 <- raw\_df %>% filter(group=="S01") %>% dplyr::select("SeriesInd","Var02")   
S02\_Var02 <- raw\_df %>% filter(group=="S02") %>% dplyr::select("SeriesInd","Var02")  
S02\_Var03 <- raw\_df %>% filter(group=="S02") %>% dplyr::select("SeriesInd","Var03")  
S03\_Var05 <- raw\_df %>% filter(group=="S03") %>% dplyr::select("SeriesInd","Var05")  
S03\_Var07 <- raw\_df %>% filter(group=="S03") %>% dplyr::select("SeriesInd","Var07")  
S04\_Var01 <- raw\_df %>% filter(group=="S04") %>% dplyr::select("SeriesInd","Var01")  
S04\_Var02 <- raw\_df %>% filter(group=="S04") %>% dplyr::select("SeriesInd","Var02")  
S05\_Var02 <- raw\_df %>% filter(group=="S05") %>% dplyr::select("SeriesInd","Var02")  
S05\_Var03 <- raw\_df %>% filter(group=="S05") %>% dplyr::select("SeriesInd","Var03")  
S06\_Var05 <- raw\_df %>% filter(group=="S06") %>% dplyr::select("SeriesInd","Var05")  
S06\_Var07 <- raw\_df %>% filter(group=="S06") %>% dplyr::select("SeriesInd","Var07")  
  
main\_df <- data.frame(S01\_Var01=S01\_Var01[,2],  
 S01\_Var02=S01\_Var02[,2],  
 S02\_Var02=S02\_Var02[,2],  
 S02\_Var03=S02\_Var03[,2],  
 S03\_Var05=S03\_Var05[,2],  
 S03\_Var07=S03\_Var07[,2],  
 S04\_Var01=S04\_Var01[,2],  
 S04\_Var02=S04\_Var02[,2],  
 S05\_Var02=S05\_Var02[,2],  
 S05\_Var03=S05\_Var03[,2],  
 S06\_Var05=S06\_Var05[,2],  
 S06\_Var07=S06\_Var07[,2])  
row.names(main\_df) <- S01\_Var01$SeriesInd  
  
main\_df  
  
  
  
# Boxplots of the variables for checking outliers and skewness  
  
main\_df\_pre\_process <- main\_df  
  
par(mfrow=c(3,4))  
for (i in c(1:length(main\_df\_pre\_process))) {  
 boxplot(main\_df\_pre\_process[,i], main=colnames(main\_df\_pre\_process)[i])  
}  
  
  
  
# remove the extreme outliers to be imputed later  
main\_df$S02\_Var03[which.max(main\_df$S02\_Var03)] <- NA  
main\_df$S06\_Var05[which.max(main\_df$S06\_Var05)] <- NA  
main\_df$S06\_Var07[which.max(main\_df$S06\_Var07)] <- NA  
  
  
  
# Finding lambda for Box-Cox Transformation for Var02  
boxcox(lm(raw\_df$Var02 ~ 1))  
# A number near 0 suggested a log transformation should be used  
# Boxplot of Var02 after log transformation  
par(mfrow=c(1,4))  
boxplot(log(main\_df\_pre\_process$S01\_Var02), main="S01\_Var02\_Log")  
boxplot(log(main\_df\_pre\_process$S02\_Var02), main="S02\_Var02\_Log")  
boxplot(log(main\_df\_pre\_process$S04\_Var02), main="S04\_Var02\_Log")  
boxplot(log(main\_df\_pre\_process$S05\_Var02), main="S05\_Var02\_Log")  
  
  
  
# For remaining missing value, we will perform linear interpolation.  
# The following are examples of missing values before linear interpolation.  
main\_df\_pre\_interpolation <- main\_df[c(1533:1543),]  
main\_df\_pre\_interpolation  
  
# perform linear interpolation  
for (i in c(1:ncol(main\_df))) {  
 main\_df[c(1:1622),i] <- na\_interpolation(main\_df[c(1:1622),i])  
}  
  
# The following are the values after imputation by linear interpolation.  
main\_df\_post\_interpolation <- main\_df[c(1533:1543),]  
main\_df\_post\_interpolation  
  
  
  
# Data is ready for modeling  
# Create time series objects  
S01\_Var01\_ts <- ts(main\_df$S01\_Var01[1:1622],start=as.integer(raw\_df$SeriesInd[1]), frequency = 1)  
S01\_Var02\_ts <- ts(main\_df$S01\_Var02[1:1622],start=as.integer(raw\_df$SeriesInd[1]), frequency = 1)  
S02\_Var02\_ts <- ts(main\_df$S02\_Var02[1:1622],start=as.integer(raw\_df$SeriesInd[1]), frequency = 1)  
S02\_Var03\_ts <- ts(main\_df$S02\_Var03[1:1622],start=as.integer(raw\_df$SeriesInd[1]), frequency = 1)  
S03\_Var05\_ts <- ts(main\_df$S03\_Var05[1:1622],start=as.integer(raw\_df$SeriesInd[1]), frequency = 1)  
S03\_Var07\_ts <- ts(main\_df$S03\_Var07[1:1622],start=as.integer(raw\_df$SeriesInd[1]), frequency = 1)  
S04\_Var01\_ts <- ts(main\_df$S04\_Var01[1:1622],start=as.integer(raw\_df$SeriesInd[1]), frequency = 1)  
S04\_Var02\_ts <- ts(main\_df$S04\_Var02[1:1622],start=as.integer(raw\_df$SeriesInd[1]), frequency = 1)  
S05\_Var02\_ts <- ts(main\_df$S05\_Var02[1:1622],start=as.integer(raw\_df$SeriesInd[1]), frequency = 1)  
S05\_Var03\_ts <- ts(main\_df$S05\_Var03[1:1622],start=as.integer(raw\_df$SeriesInd[1]), frequency = 1)  
S06\_Var05\_ts <- ts(main\_df$S06\_Var05[1:1622],start=as.integer(raw\_df$SeriesInd[1]), frequency = 1)  
S06\_Var07\_ts <- ts(main\_df$S06\_Var07[1:1622],start=as.integer(raw\_df$SeriesInd[1]), frequency = 1)  
  
  
  
# Time Plot  
autoplot(S01\_Var01\_ts) +  
autoplot(S01\_Var02\_ts) +  
autoplot(S02\_Var02\_ts) +  
autoplot(S02\_Var03\_ts) +  
autoplot(S03\_Var05\_ts) +  
autoplot(S03\_Var07\_ts) +  
autoplot(S04\_Var01\_ts) +  
autoplot(S04\_Var02\_ts) +  
autoplot(S05\_Var02\_ts) +  
autoplot(S05\_Var03\_ts) +  
autoplot(S06\_Var05\_ts) +  
autoplot(S06\_Var07\_ts) +  
 plot\_layout(ncol = 1, guides = "collect")  
  
  
  
# ACF and PACF  
ggAcf(S01\_Var01\_ts) + ggPacf(S01\_Var01\_ts) +  
ggAcf(S01\_Var02\_ts) + ggPacf(S01\_Var02\_ts) +  
ggAcf(S02\_Var02\_ts) + ggPacf(S02\_Var02\_ts) +  
ggAcf(S02\_Var03\_ts) + ggPacf(S02\_Var03\_ts) +  
ggAcf(S03\_Var05\_ts) + ggPacf(S03\_Var05\_ts) +  
ggAcf(S03\_Var07\_ts) + ggPacf(S03\_Var07\_ts) +  
ggAcf(S04\_Var01\_ts) + ggPacf(S04\_Var01\_ts) +  
ggAcf(S04\_Var02\_ts) + ggPacf(S04\_Var02\_ts) +  
ggAcf(S05\_Var02\_ts) + ggPacf(S05\_Var02\_ts) +  
ggAcf(S05\_Var03\_ts) + ggPacf(S05\_Var03\_ts) +  
ggAcf(S06\_Var05\_ts) + ggPacf(S06\_Var05\_ts) +  
ggAcf(S06\_Var07\_ts) + ggPacf(S06\_Var07\_ts) +  
 plot\_layout(ncol = 2, guides = "collect")  
  
  
  
# Buildling models  
# For each time series, we build an optimal ETS model an an optimal ARIMA model based on the AIC scores.  
  
S01\_Var01\_ets <- ets(S01\_Var01\_ts)  
S01\_Var01\_arima <- auto.arima(S01\_Var01\_ts, stepwise=FALSE, approximation=FALSE)  
  
S01\_Var02\_ets <- ets(S01\_Var02\_ts, lambda = 0)  
S01\_Var02\_arima <- auto.arima(S01\_Var02\_ts, lambda = 0, stepwise=FALSE, approximation=FALSE)  
  
S02\_Var02\_ets <- ets(S02\_Var02\_ts, lambda = 0)  
S02\_Var02\_arima <- auto.arima(S02\_Var02\_ts, lambda = 0, stepwise=FALSE, approximation=FALSE)  
  
S02\_Var03\_ets <- ets(S02\_Var03\_ts)  
S02\_Var03\_arima <- auto.arima(S02\_Var03\_ts, stepwise=FALSE, approximation=FALSE)  
  
S03\_Var05\_ets <- ets(S03\_Var05\_ts)  
S03\_Var05\_arima <- auto.arima(S03\_Var05\_ts, stepwise=FALSE, approximation=FALSE)  
  
S03\_Var07\_ets <- ets(S03\_Var07\_ts)  
S03\_Var07\_arima <- auto.arima(S03\_Var07\_ts, stepwise=FALSE, approximation=FALSE)  
  
S04\_Var01\_ets <- ets(S04\_Var01\_ts)  
S04\_Var01\_arima <- auto.arima(S04\_Var01\_ts, stepwise=FALSE, approximation=FALSE)  
  
S04\_Var02\_ets <- ets(S04\_Var02\_ts, lambda = 0)  
S04\_Var02\_arima <- auto.arima(S04\_Var02\_ts, lambda = 0, stepwise=FALSE, approximation=FALSE)  
  
S05\_Var02\_ets <- ets(S05\_Var02\_ts, lambda = 0)  
S05\_Var02\_arima <- auto.arima(S05\_Var02\_ts, lambda = 0, stepwise=FALSE, approximation=FALSE)  
  
S05\_Var03\_ets <- ets(S05\_Var03\_ts)  
S05\_Var03\_arima <- auto.arima(S05\_Var03\_ts, stepwise=FALSE, approximation=FALSE)  
  
S06\_Var05\_ets <- ets(S06\_Var05\_ts)  
S06\_Var05\_arima <- auto.arima(S06\_Var05\_ts, stepwise=FALSE, approximation=FALSE)  
  
S06\_Var07\_ets <- ets(S06\_Var07\_ts)  
S06\_Var07\_arima <- auto.arima(S06\_Var07\_ts, stepwise=FALSE, approximation=FALSE)  
  
  
  
# Perform Cross-Validation for both Exponential Smoothing (ETS) and ARIMA models  
# The process takes more than an hour so the pre-calculated results at the end of the block can be used to save time  
  
# fets <- function(x, h) {  
# forecast(ets(x), h = h)  
# }  
#   
# farima <- function(x, h) {  
# forecast(auto.arima(x), h=h)  
# }  
#   
# fets2 <- function(x, h) {  
# forecast(ets(x, lambda=0), h = h)  
# }  
#   
# farima2 <- function(x, h) {  
# forecast(auto.arima(x, lambda=0), h=h)  
# }  
#   
#   
# e1 <- tsCV(S01\_Var01\_ts, fets, h=1)  
# e2 <- tsCV(S01\_Var01\_ts, farima, h=1)  
# S01\_Var01\_ets\_cv <- sqrt(mean(e1^2, na.rm=TRUE))  
# S01\_Var01\_arima\_cv <- sqrt(mean(e2^2, na.rm=TRUE))  
#   
# e1 <- tsCV(S01\_Var02\_ts, fets2, h=1)  
# e2 <- tsCV(S01\_Var02\_ts, farima2, h=1)  
# S01\_Var02\_ets\_cv <- sqrt(mean(e1^2, na.rm=TRUE))  
# S01\_Var02\_arima\_cv <- sqrt(mean(e2^2, na.rm=TRUE))  
#   
# e1 <- tsCV(S02\_Var02\_ts, fets2, h=1)  
# e2 <- tsCV(S02\_Var02\_ts, farima2, h=1)  
# S02\_Var02\_ets\_cv <- sqrt(mean(e1^2, na.rm=TRUE))  
# S02\_Var02\_arima\_cv <- sqrt(mean(e2^2, na.rm=TRUE))  
#   
# e1 <- tsCV(S02\_Var03\_ts, fets, h=1)  
# e2 <- tsCV(S02\_Var03\_ts, farima, h=1)  
# S02\_Var03\_ets\_cv <- sqrt(mean(e1^2, na.rm=TRUE))  
# S02\_Var03\_arima\_cv <- sqrt(mean(e2^2, na.rm=TRUE))  
#   
# e1 <- tsCV(S03\_Var05\_ts, fets, h=1)  
# e2 <- tsCV(S03\_Var05\_ts, farima, h=1)  
# S03\_Var05\_ets\_cv <- sqrt(mean(e1^2, na.rm=TRUE))  
# S03\_Var05\_arima\_cv <- sqrt(mean(e2^2, na.rm=TRUE))  
#   
# e1 <- tsCV(S03\_Var07\_ts, fets, h=1)  
# e2 <- tsCV(S03\_Var07\_ts, farima, h=1)  
# S03\_Var07\_ets\_cv <- sqrt(mean(e1^2, na.rm=TRUE))  
# S03\_Var07\_arima\_cv <- sqrt(mean(e2^2, na.rm=TRUE))  
#   
# e1 <- tsCV(S04\_Var01\_ts, fets, h=1)  
# e2 <- tsCV(S04\_Var01\_ts, farima, h=1)  
# S04\_Var01\_ets\_cv <- sqrt(mean(e1^2, na.rm=TRUE))  
# S04\_Var01\_arima\_cv <- sqrt(mean(e2^2, na.rm=TRUE))  
#   
# e1 <- tsCV(S04\_Var02\_ts, fets2, h=1)  
# e2 <- tsCV(S04\_Var02\_ts, farima2, h=1)  
# S04\_Var02\_ets\_cv <- sqrt(mean(e1^2, na.rm=TRUE))  
# S04\_Var02\_arima\_cv <- sqrt(mean(e2^2, na.rm=TRUE))  
#   
# e1 <- tsCV(S05\_Var02\_ts, fets2, h=1)  
# e2 <- tsCV(S05\_Var02\_ts, farima2, h=1)  
# S05\_Var02\_ets\_cv <- sqrt(mean(e1^2, na.rm=TRUE))  
# S05\_Var02\_arima\_cv <- sqrt(mean(e2^2, na.rm=TRUE))  
#   
# e1 <- tsCV(S05\_Var03\_ts, fets, h=1)  
# e2 <- tsCV(S05\_Var03\_ts, farima, h=1)  
# S05\_Var03\_ets\_cv <- sqrt(mean(e1^2, na.rm=TRUE))  
# S05\_Var03\_arima\_cv <- sqrt(mean(e2^2, na.rm=TRUE))  
#   
# e1 <- tsCV(S06\_Var05\_ts, fets, h=1)  
# e2 <- tsCV(S06\_Var05\_ts, farima, h=1)  
# S06\_Var05\_ets\_cv <- sqrt(mean(e1^2, na.rm=TRUE))  
# S06\_Var05\_arima\_cv <- sqrt(mean(e2^2, na.rm=TRUE))  
#   
# e1 <- tsCV(S06\_Var07\_ts, fets, h=1)  
# e2 <- tsCV(S06\_Var07\_ts, farima, h=1)  
# S06\_Var07\_ets\_cv <- sqrt(mean(e1^2, na.rm=TRUE))  
# S06\_Var07\_arima\_cv <- sqrt(mean(e2^2, na.rm=TRUE))  
  
# The followings are the pre-calculated results  
S01\_Var01\_ets\_cv <- 0.5152255  
S01\_Var01\_arima\_cv <- 0.5155231  
S01\_Var02\_ets\_cv <- 3412965  
S01\_Var02\_arima\_cv <- 3331920  
S02\_Var02\_ets\_cv <- 27002957  
S02\_Var02\_arima\_cv <- 25553578  
S02\_Var03\_ets\_cv <- 0.2693275  
S02\_Var03\_arima\_cv <- 0.2738209  
S03\_Var05\_ets\_cv <- 1.508795  
S03\_Var05\_arima\_cv <- 1.516405  
S03\_Var07\_ets\_cv <- 1.344844  
S03\_Var07\_arima\_cv <- 1.349914  
S04\_Var01\_ets\_cv <- 0.5057957  
S04\_Var01\_arima\_cv <- 0.5346757  
S04\_Var02\_ets\_cv <- 11819340  
S04\_Var02\_arima\_cv <- 11673370  
S05\_Var02\_ets\_cv <- 5419584  
S05\_Var02\_arima\_cv <- 5382840  
S05\_Var03\_ets\_cv <- 0.9069181  
S05\_Var03\_arima\_cv <- 0.9039646  
S06\_Var05\_ets\_cv <- 0.5676618  
S06\_Var05\_arima\_cv <- 0.5708166  
S06\_Var07\_ets\_cv <- 0.5618834  
S06\_Var07\_arima\_cv <- 0.5649096  
  
  
  
# Gather the performance results in one dataframe for comparison  
# The table includes the RMSE from the training data and the RMSE from the Cross-Validations  
model\_compare <-   
 data.frame(Group=c("S01","S01","S01","S01",  
 "S02","S02","S02","S02",  
 "S03","S03","S03","S03",  
 "S04","S04","S04","S04",  
 "S05","S05","S05","S05",  
 "S06","S06","S06","S06"),  
 Variable=c("Var01","Var01","Var02","Var02",  
 "Var02","Var02","Var03","Var03",  
 "Var05","Var05","Var07","Var07",  
 "Var01","Var01","Var02","Var02",  
 "Var02","Var02","Var03","Var03",  
 "Var05","Var05","Var07","Var07"),  
 Model\_Type=c("Exponential Smoothing","ARIMA","Exponential Smoothing","ARIMA",  
 "Exponential Smoothing","ARIMA","Exponential Smoothing","ARIMA",  
 "Exponential Smoothing","ARIMA","Exponential Smoothing","ARIMA",  
 "Exponential Smoothing","ARIMA","Exponential Smoothing","ARIMA",  
 "Exponential Smoothing","ARIMA","Exponential Smoothing","ARIMA",  
 "Exponential Smoothing","ARIMA","Exponential Smoothing","ARIMA"),  
 Model=c(as.character(S01\_Var01\_ets),as.character(S01\_Var01\_arima),  
 as.character(S01\_Var02\_ets),as.character(S01\_Var02\_arima),  
 as.character(S02\_Var02\_ets),as.character(S02\_Var02\_arima),  
 as.character(S02\_Var03\_ets),as.character(S02\_Var03\_arima),  
 as.character(S03\_Var05\_ets),as.character(S03\_Var05\_arima),  
 as.character(S03\_Var07\_ets),as.character(S03\_Var07\_arima),  
 as.character(S04\_Var01\_ets),as.character(S04\_Var01\_arima),  
 as.character(S04\_Var02\_ets),as.character(S04\_Var02\_arima),  
 as.character(S05\_Var02\_ets),as.character(S05\_Var02\_arima),  
 as.character(S05\_Var03\_ets),as.character(S05\_Var03\_arima),  
 as.character(S06\_Var05\_ets),as.character(S06\_Var05\_arima),  
 as.character(S06\_Var07\_ets),as.character(S06\_Var07\_arima)),  
 CV\_RMSE=c(S01\_Var01\_ets\_cv, S01\_Var01\_arima\_cv,  
 S01\_Var02\_ets\_cv, S01\_Var02\_arima\_cv,  
 S02\_Var02\_ets\_cv, S02\_Var02\_arima\_cv,  
 S02\_Var03\_ets\_cv, S02\_Var03\_arima\_cv,  
 S03\_Var05\_ets\_cv, S03\_Var05\_arima\_cv,  
 S03\_Var07\_ets\_cv, S03\_Var07\_arima\_cv,  
 S04\_Var01\_ets\_cv, S04\_Var01\_arima\_cv,  
 S04\_Var02\_ets\_cv, S04\_Var02\_arima\_cv,  
 S05\_Var02\_ets\_cv, S05\_Var02\_arima\_cv,  
 S05\_Var03\_ets\_cv, S05\_Var03\_arima\_cv,  
 S06\_Var05\_ets\_cv, S06\_Var05\_arima\_cv,  
 S06\_Var07\_ets\_cv, S06\_Var07\_arima\_cv),  
 Train\_RMSE=c(accuracy(S01\_Var01\_ets)[2],accuracy(S01\_Var01\_arima)[2],  
 accuracy(S01\_Var02\_ets)[2],accuracy(S01\_Var02\_arima)[2],  
 accuracy(S02\_Var02\_ets)[2],accuracy(S02\_Var02\_arima)[2],  
 accuracy(S02\_Var03\_ets)[2],accuracy(S02\_Var03\_arima)[2],  
 accuracy(S03\_Var05\_ets)[2],accuracy(S03\_Var05\_arima)[2],  
 accuracy(S03\_Var07\_ets)[2],accuracy(S03\_Var07\_arima)[2],  
 accuracy(S04\_Var01\_ets)[2],accuracy(S04\_Var01\_arima)[2],  
 accuracy(S04\_Var02\_ets)[2],accuracy(S04\_Var02\_arima)[2],  
 accuracy(S05\_Var02\_ets)[2],accuracy(S05\_Var02\_arima)[2],  
 accuracy(S05\_Var03\_ets)[2],accuracy(S05\_Var03\_arima)[2],  
 accuracy(S06\_Var05\_ets)[2],accuracy(S06\_Var05\_arima)[2],  
 accuracy(S06\_Var07\_ets)[2],accuracy(S06\_Var07\_arima)[2]))  
  
# Adding the p=value from the ljung-box test to compare the goodness of fit for each model  
model\_compare$Ljung\_Box\_p[1] <- checkresiduals(S01\_Var01\_ets, plot=FALSE)$p.value  
model\_compare$Ljung\_Box\_p[2] <- checkresiduals(S01\_Var01\_arima, plot=FALSE)$p.value  
model\_compare$Ljung\_Box\_p[3] <- checkresiduals(S01\_Var02\_ets, plot=FALSE)$p.value  
model\_compare$Ljung\_Box\_p[4] <- checkresiduals(S01\_Var02\_arima, plot=FALSE)$p.value  
model\_compare$Ljung\_Box\_p[5] <- checkresiduals(S02\_Var02\_ets, plot=FALSE)$p.value  
model\_compare$Ljung\_Box\_p[6] <- checkresiduals(S02\_Var02\_arima, plot=FALSE)$p.value  
model\_compare$Ljung\_Box\_p[7] <- checkresiduals(S02\_Var03\_ets, plot=FALSE)$p.value  
model\_compare$Ljung\_Box\_p[8] <- checkresiduals(S02\_Var03\_arima, plot=FALSE)$p.value  
model\_compare$Ljung\_Box\_p[9] <- checkresiduals(S03\_Var05\_ets, plot=FALSE)$p.value  
model\_compare$Ljung\_Box\_p[10] <- checkresiduals(S03\_Var05\_arima, plot=FALSE)$p.value  
model\_compare$Ljung\_Box\_p[11] <- checkresiduals(S03\_Var07\_ets, plot=FALSE)$p.value  
model\_compare$Ljung\_Box\_p[12] <- checkresiduals(S03\_Var07\_arima, plot=FALSE)$p.value  
model\_compare$Ljung\_Box\_p[13] <- checkresiduals(S04\_Var01\_ets, plot=FALSE)$p.value  
model\_compare$Ljung\_Box\_p[14] <- checkresiduals(S04\_Var01\_arima, plot=FALSE)$p.value  
model\_compare$Ljung\_Box\_p[15] <- checkresiduals(S04\_Var02\_ets, plot=FALSE)$p.value  
model\_compare$Ljung\_Box\_p[16] <- checkresiduals(S04\_Var02\_arima, plot=FALSE)$p.value  
model\_compare$Ljung\_Box\_p[17] <- checkresiduals(S05\_Var02\_ets, plot=FALSE)$p.value  
model\_compare$Ljung\_Box\_p[18] <- checkresiduals(S05\_Var02\_arima, plot=FALSE)$p.value  
model\_compare$Ljung\_Box\_p[19] <- checkresiduals(S05\_Var03\_ets, plot=FALSE)$p.value  
model\_compare$Ljung\_Box\_p[20] <- checkresiduals(S05\_Var03\_arima, plot=FALSE)$p.value  
model\_compare$Ljung\_Box\_p[21] <- checkresiduals(S06\_Var05\_ets, plot=FALSE)$p.value  
model\_compare$Ljung\_Box\_p[22] <- checkresiduals(S06\_Var05\_arima, plot=FALSE)$p.value  
model\_compare$Ljung\_Box\_p[23] <- checkresiduals(S06\_Var07\_ets, plot=FALSE)$p.value  
model\_compare$Ljung\_Box\_p[24] <- checkresiduals(S06\_Var07\_arima, plot=FALSE)$p.value  
  
model\_compare  
  
  
  
# Prepare data to plot the RMSE.  
# Since Var02 has a number scale much larger than the other variables, we have to scale the RMSE for Var02 models by multiplying 1/10000000 so they can be plotted in the same graph.  
model\_compare2 <- model\_compare  
model\_compare2$CV\_RMSE <- ifelse(model\_compare2$Variable=="Var02",  
 model\_compare2$CV\_RMSE/10000000,  
 model\_compare2$CV\_RMSE)  
model\_compare2$Train\_RMSE <- ifelse(model\_compare2$Variable=="Var02",  
 model\_compare2$Train\_RMSE/10000000,  
 model\_compare2$Train\_RMSE)  
  
# Plot the Cross-Validation RMSE.   
ggplot(model\_compare2, aes(x=paste0(Group, Variable), y=CV\_RMSE, group=Model\_Type)) +  
 geom\_line(aes(linetype=Model\_Type))+  
 geom\_point(aes(shape=Model\_Type))+  
 theme(axis.text.x = element\_text(angle = 90))+  
 xlab("")  
  
# Plot the training data RMSE.   
ggplot(model\_compare2, aes(x=paste0(Group, Variable), y=Train\_RMSE, group=Model\_Type)) +  
 geom\_line(aes(linetype=Model\_Type))+  
 geom\_point(aes(shape=Model\_Type))+  
 theme(axis.text.x = element\_text(angle = 90))+  
 xlab("")  
  
# Plot the ljung-box test p-value  
ggplot(model\_compare2, aes(x=paste0(Group, Variable), y=Ljung\_Box\_p, group=Model\_Type)) +  
 geom\_line(aes(linetype=Model\_Type))+  
 geom\_point(aes(shape=Model\_Type))+  
 theme(axis.text.x = element\_text(angle = 90))+  
 xlab("")  
# The RMSE for ETS and ARIMA models are very close, with the ARIMA models perform slightly better.  
# The the ljung-box test p-values, the ARIMA models are fitting to the data better so we will choose the ARIMA models for forecasting  
  
  
  
# Forcasting  
S01\_Var01\_forecast <- S01\_Var01\_arima %>% forecast(h=140)  
S01\_Var02\_forecast <- S01\_Var02\_arima %>% forecast(h=140)  
S02\_Var02\_forecast <- S02\_Var02\_arima %>% forecast(h=140)  
S02\_Var03\_forecast <- S02\_Var03\_arima %>% forecast(h=140)  
S03\_Var05\_forecast <- S03\_Var05\_arima %>% forecast(h=140)  
S03\_Var07\_forecast <- S03\_Var07\_arima %>% forecast(h=140)  
S04\_Var01\_forecast <- S04\_Var01\_arima %>% forecast(h=140)  
S04\_Var02\_forecast <- S04\_Var02\_arima %>% forecast(h=140)  
S05\_Var02\_forecast <- S05\_Var02\_arima %>% forecast(h=140)  
S05\_Var03\_forecast <- S05\_Var03\_arima %>% forecast(h=140)  
S06\_Var05\_forecast <- S06\_Var05\_arima %>% forecast(h=140)  
S06\_Var07\_forecast <- S06\_Var07\_arima %>% forecast(h=140)  
  
  
  
# Forecast Plot  
S01\_Var01\_forecast %>% autoplot()  
S01\_Var02\_forecast %>% autoplot()  
S02\_Var02\_forecast %>% autoplot()  
S02\_Var03\_forecast %>% autoplot()  
S03\_Var05\_forecast %>% autoplot()  
S03\_Var07\_forecast %>% autoplot()  
S04\_Var01\_forecast %>% autoplot()  
S04\_Var02\_forecast %>% autoplot()  
S05\_Var02\_forecast %>% autoplot()  
S05\_Var03\_forecast %>% autoplot()  
S06\_Var05\_forecast %>% autoplot()  
S06\_Var07\_forecast %>% autoplot()