DATA 624 Spring 2019: Project-1

Ahmed Sajjad, Harpreet Shoker, Jagruti Solao, Chad Smith, Todd Weigel

April 16, 2019

Loading all the libraries we will be using in project

```
library(ggplot2)
library(tidyr)
library(dplyr)
library(fpp2)
library(lubridate)
library(imputeTS)
library(reshape2)
```

```
data <- read.csv("Project1data.csv",header=TRUE)
```

Here we are imported data from Excel and dates that are in a numeric format "SeriesInd". Using as.Date to import these, excel uses the origin date as December 30, 1899.

```
data$SeriesInd <- as.Date(data$SeriesInd,origin = "1899-12-30")
```

From the data we see starting row 9733 all the rows are blank. As part of data cleaning we are removing these rows.

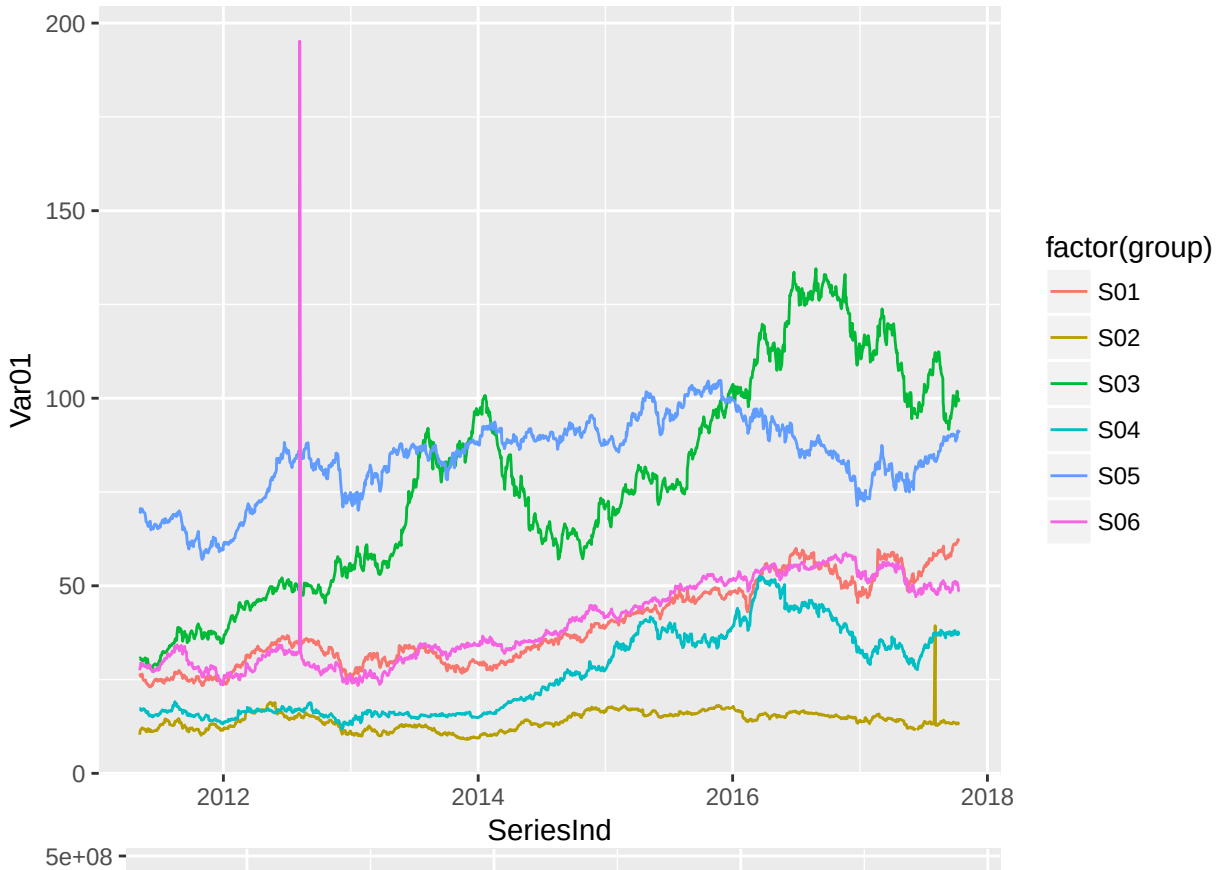
```
data <- data[1:9732,]
```

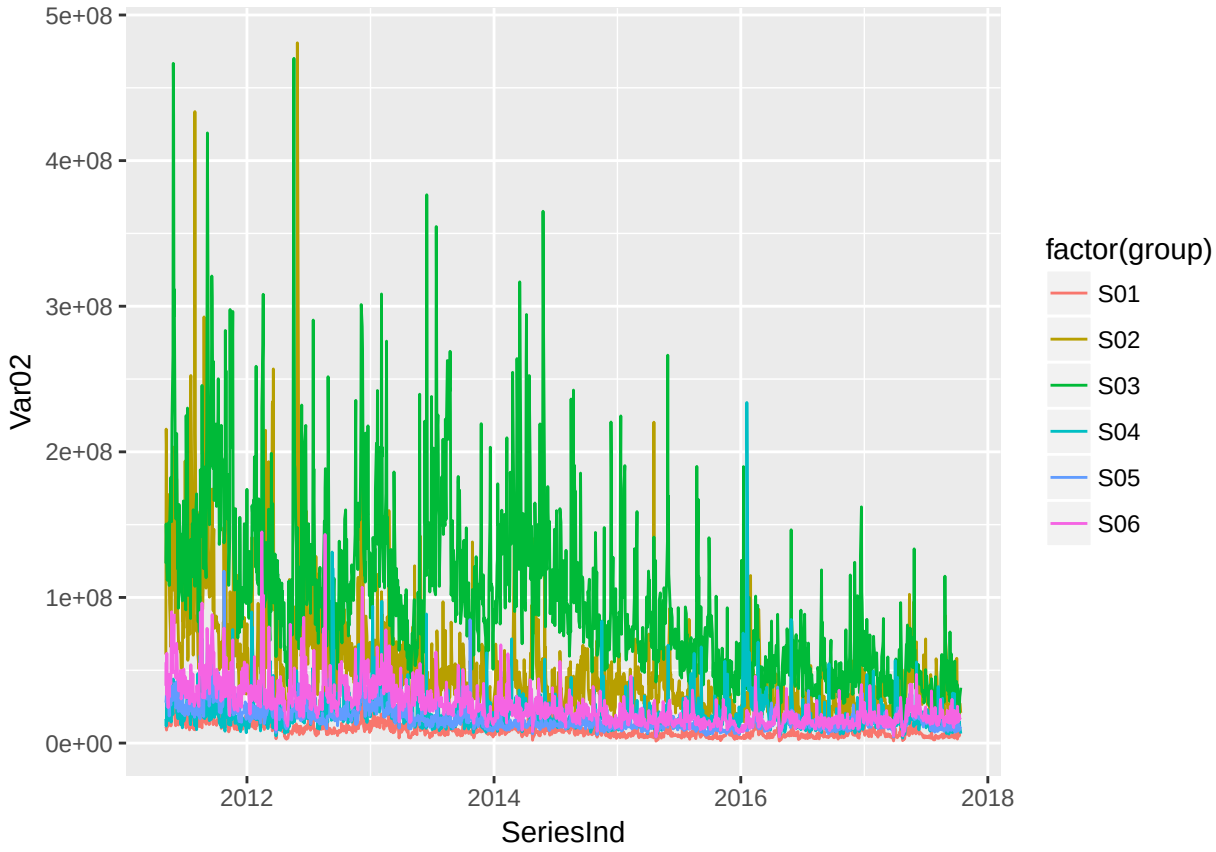
converting data to long format

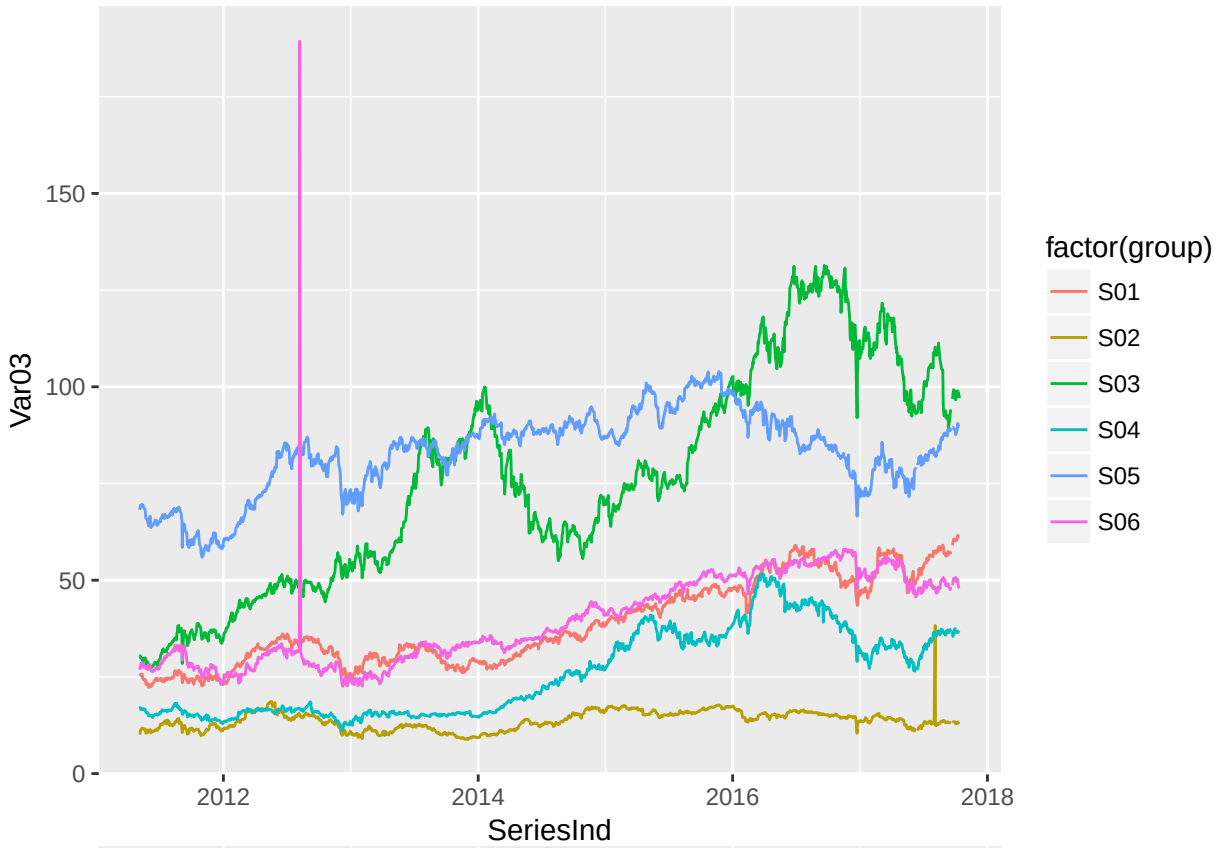
```
gathered_data <- gather(data,key="Variable",value="Value",-SeriesInd,-group)
```

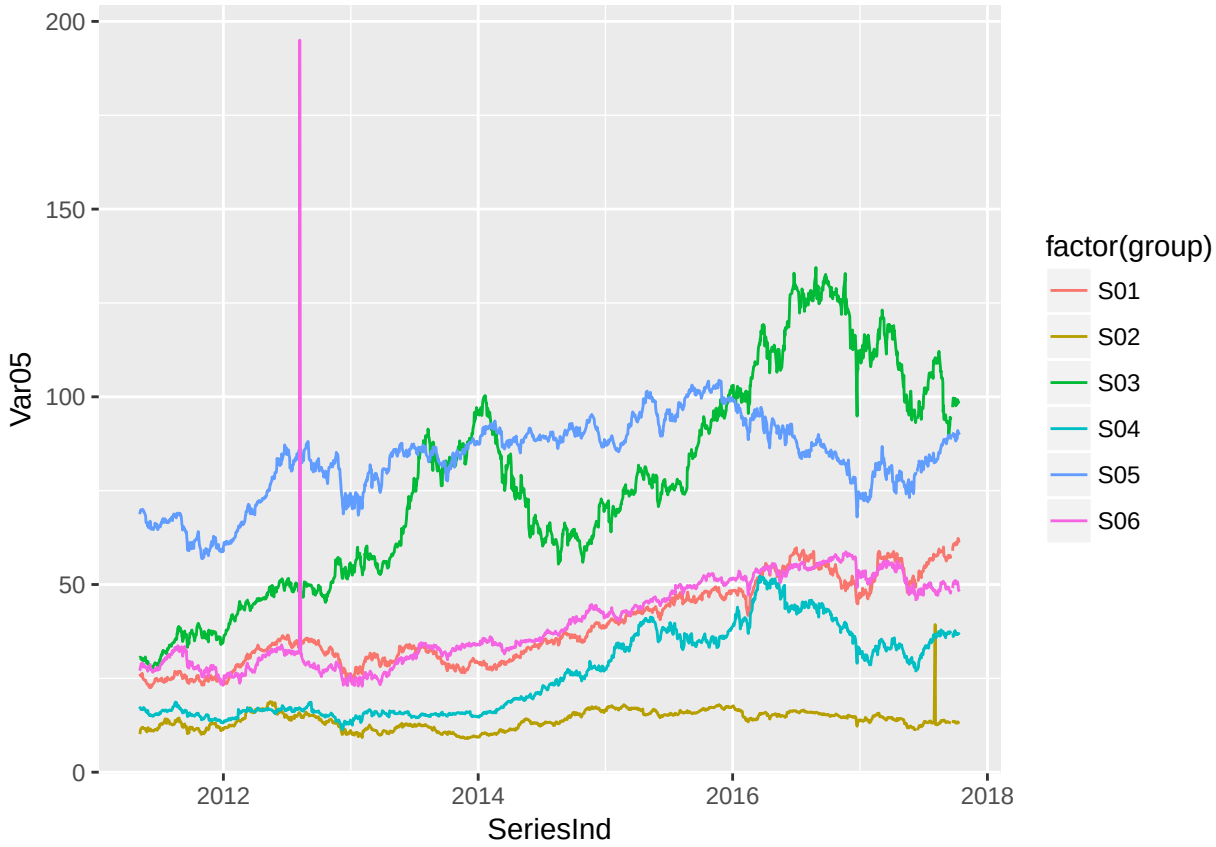
Lets plot data for all variables and group to do some analysis about data Here we are plotting line plots

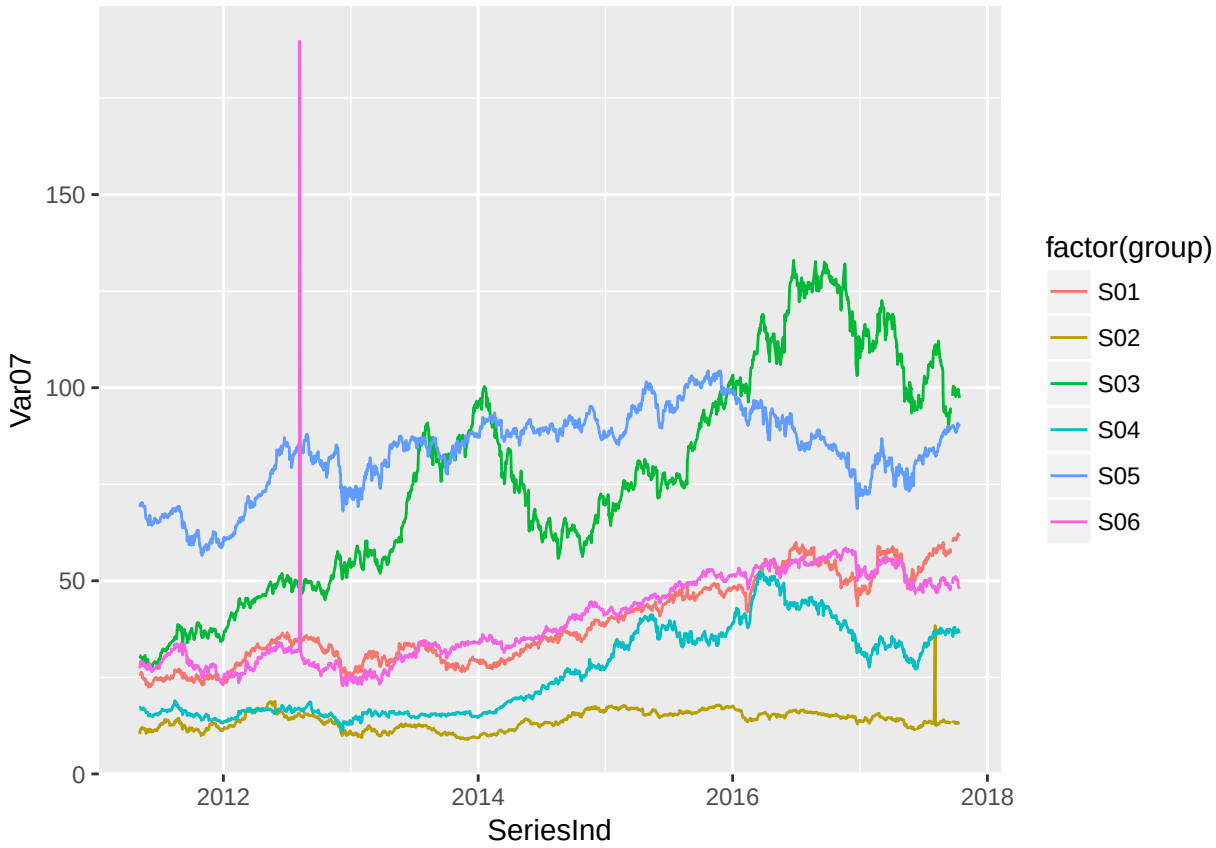
```
library(ggplot2)
vars <- unique(gathered_data$Variable)
for (i in 1:length(vars)) {
  plot <-  ggplot() +
         geom_line(data = subset(gathered_data, Variable == vars[[i]]),
                   aes(SeriesInd, Value, group = group, color = factor(group))) +
         ylab(as.character(vars[[i]]))
  print(plot)
}
```


200
150
100
50
0
Var01
2012
2014
2016
2018
SeriesInd
factor(group)
S01
S02
S03
S04
S05
S06


5e+08
4e+08
3e+08
2e+08
1e+08
0e+00
Var02
2012
2014
2016
2018
SeriesInd
factor(group)
S01
S02
S03
S04
S05
S06


Var03
150
100
50
0
2012
2014
2016
2018
SeriesInd
factor(group)
S01
S02
S03
S04
S05
S06


Var05
200
150
100
50
0
2012
2014
2016
2018
SeriesInd
factor(group)
S01
S02
S03
S04
S05
S06



We need to forecast for the follwing for our project S01 – Forecast Var01, Var02 S02 – Forecast Var02, Var03 S03 – Forecast Var05, Var07 S04 – Forecast Var01, Var02 S05 – Forecast Var02, Var03 S06 – Forecast Var05, Var07

And From the above plots we made the following observations

1. Series S01 - Variables(Var 01 , and Var02 looks similar)
2. Series S02 - Variables(var01, Var03,Var05,Var07) has outliers
3. Series S03 _ Variables(Var05 and Var 07 looks relatively similar)
4. Series S06 - Variables (var01,var03,var05,var07) has outliers

all the outliers needs to be fixed before forecasting

Lets check for missing null values in the data

```
missing_values <- gathered_data[which(is.na(gathered_data$Value) == TRUE),]
```

From the above there are missng values in data and these needs to be imputed before producing forecast model.We are using here na.interpolatin() to fill in the null values.

```
gathered_data <- na.interpolation(gathered_data)
```