Project 1

S03 Var05, S03 Var07, S05 Var03

Group 3: Shaya Engelman, Julia Ferris, Amanda Fox, Jean Jimenez

2024-06-23

# Introduction

This project is an attempt to forecast various variables in a dataset. The dataset is completely deidentified and contains 6 ‘categories’ with 5 time series for each category. The dataset consists of of 1622 observations and another 140 empty observations to be forecasted.

Here, the focus is on forecasting the variables S03 Var05, S03 Var07, and S05 Var03.

# S03

## Data

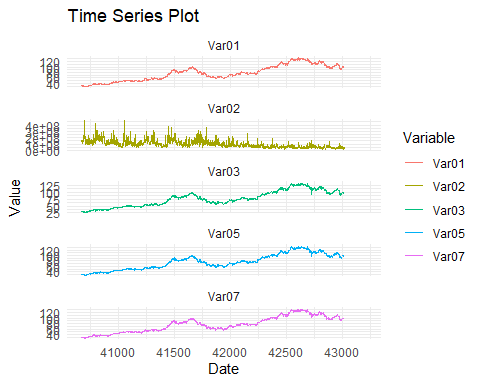
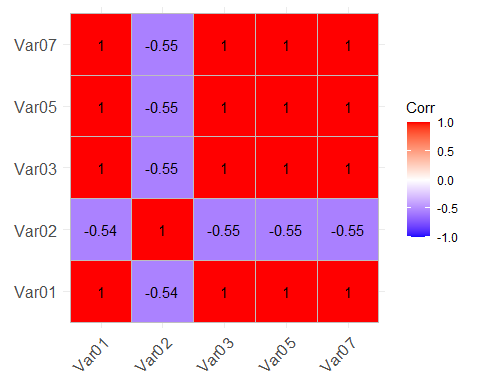
The data is loaded and the category of interest is extracted. The following table is the first 6 rows of the data. The SeriesInd column is the index of the time series, the other columns are the various time series. As mentioned, the data, including the index is completely deidentified, so the actual meaning of both the variables and what the time step being measured is unknown.

First 6 rows of the S03 data

| SeriesInd | Var01 | Var02 | Var03 | Var05 | Var07 |
| --- | --- | --- | --- | --- | --- |
| 40669 | 30.64286 | 123432400 | 30.34000 | 30.49000 | 30.57286 |
| 40670 | 30.79857 | 150476200 | 30.46428 | 30.65714 | 30.62571 |
| 40671 | 30.74714 | 138040000 | 30.10714 | 30.62571 | 30.13857 |
| 40672 | 30.28572 | 119282800 | 29.86429 | 30.25000 | 30.08286 |
| 40673 | 30.28572 | 111902700 | 29.86572 | 30.04286 | 30.28286 |
| 40676 | 30.42857 | 115557400 | 29.77857 | 30.40000 | 30.01571 |

While only two variables, Var05 and Var07, are of interest, is is importnant to take a glance at the other variables in the dataset. By visualizing the time series and the correlation between the variables, it becomes apparent that many of these variables are not just highly correlated, but also have almost identical patterns. In particular, the two variables of interest, Var05 and Var07, seem to be almost identical. Only Var02 seems to be onan entirely different order of magnitude with much less correlation to the other variables.

## Warning: Removed 700 rows containing missing values or values outside the scale range  
## (`geom\_line()`).

The summary statistics reveal some interesting information. The two variables of interest, Var05 and Var07, have almost identical values for the mean, median, and standard deviation. The minimum and maximum values are also very similar.

Since the data is completely deidentified, it is not known what the ‘SeriesInd’ column represents. The actual values appear to be Excel date serials (the number of days since December 30, 1899), but since this is not confirmed, it will be left as a blank index unless there is a reason to change it. An interesting insight from the summary statistics

There are 144 missing values for both ‘Var05’ and ‘Var07’. The majority of the missing values are in the last 140 rows of the dataset, which are the rows to be forecasted. There are 4 rows of missing data in the first 1482 rows. By checking those rows, it is observed that both ‘Var05’ and ‘Var07’ are missing for the same rows. This does not allow us to use the other variable to fill in the missing values. Since there are only 4 rows out of 1622 missing, the method of imputation should not have a significant impact on the forecast. The method chosen was to use an interpolation method to fill in the missing values with the average of the two closest non-missing values.

## [1] 1762

SeriesInd

Var05

Var07

Min. :40669

Min. : 27.48

Min. : 27.44

1st Qu.:41304

1st Qu.: 53.30

1st Qu.: 53.46

Median :41946

Median : 75.59

Median : 75.71

Mean :41945

Mean : 76.90

Mean : 76.87

3rd Qu.:42586

3rd Qu.: 98.55

3rd Qu.: 98.61

Max. :43221

Max. :134.46

Max. :133.00

NA

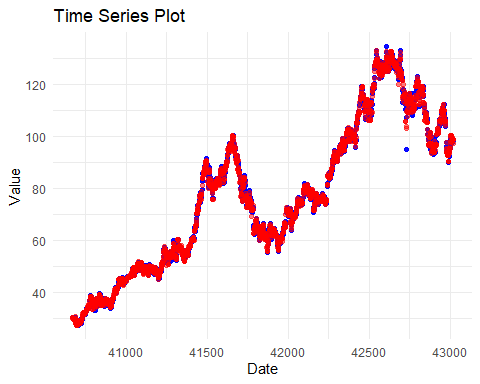
NA’s :144

NA’s :144

Additionally, to fill in the implicitly missing rows to make the data continuous, a different imputation method was used. For this, the rows were filled with the most recent non-missing value. This method was chosen because the data appears to have zero change for those observations based on the reasoning why they might be missing, such as only operaing on business days.

The individual time series for ‘Var05’ and ‘Var07’ are plotted below and it is noted that there does not seem to be any major outliers that need to be addressed. The data does display an overall upward trend with significant variability. Initially, the data seems to have very minor fluctuationsm, but later on, the variability increases significantly. Additionally, the data seems to have an either seasonal or cyclical component, with the data appearing to have a pattern of peaks and valleys.

## Don't know how to automatically pick scale for object of type <ts>. Defaulting  
## to continuous.



## Data Preparation