1.11.2023
Burak Çetiner

IE 582 Statistical Learning for Data Mining
Homework 1 Report

It is important to understand the dataset that you are aiming to analyze before starting the statistical calculations and operations. However, I wanted to state an important diffuculty that I have faced even before it. My preference was using RStudio and R programming language as a tool of statistical inference for two main reasons; first I had prior experience with this tool, and second it is more powerful and user-friendly in terms of visualize the finding comparing to Python. I decided to use RStudio and I had imported the dataset "all_ticks_wide". Everything works fine when I worked with a RScript. However, when I moved my codes to RMarkdown to create a html file, the software did not recognize (find) the dataset although it is saved in the environment. At this point, I decided to ask help from Chatgpt and after several trials I added the following code to download the dataset. (It was the only time I consulted to Chatgpt for this homework.)

```
github_url <- "https://github.com/BU-IE-582/fall-23-burakcetiner3/raw/main/all_ticks_wide.xlsx"
local_file_path <- "all_ticks_wide.xlsx"
download.file(url = github_url, destfile = local_file_path, mode = "wb")
Dataset <- read_excel(local_file_path)
```

I am aware that it is not the most suitable way to work with a dataset in RMarkdown because each time you update (knit) the file the dataset is redownloaded from the internet and it consume around one minute due to the low internet connection. So, I make a self-critism at this point, but I could not find any better solution.

4.1 Descriptive Statistics

After this long introduction, we can move to the data analysis. As we learned from the statistics courses, it is essential to be familiar with the dataset in order to make accurate inferences. The first step of this process is the descriptive statistics. I have started with "summary" statistics of the whole dataset. By this operation I realized that there are a lot of missing values (NAs) in the dataset. Firstly, I have deleted these rows. Then I checked the dataset again. Although it appeared that many observations are lost by this operation, there is still enough observations (around 9000) to make meaningful inferences.

After this operation, I chose 10 different stocks which I am already familiar with thanks to my brother's advices. Then I used "pairs" command to see if there is any obvious correlation within these variables. At the first glance, I realized that there may be positive correlation between AKBNK and GARAN, and also between AKSA and AKSEN. I decide to investigate these two pairs further to understand the relationships between them and interpret it.

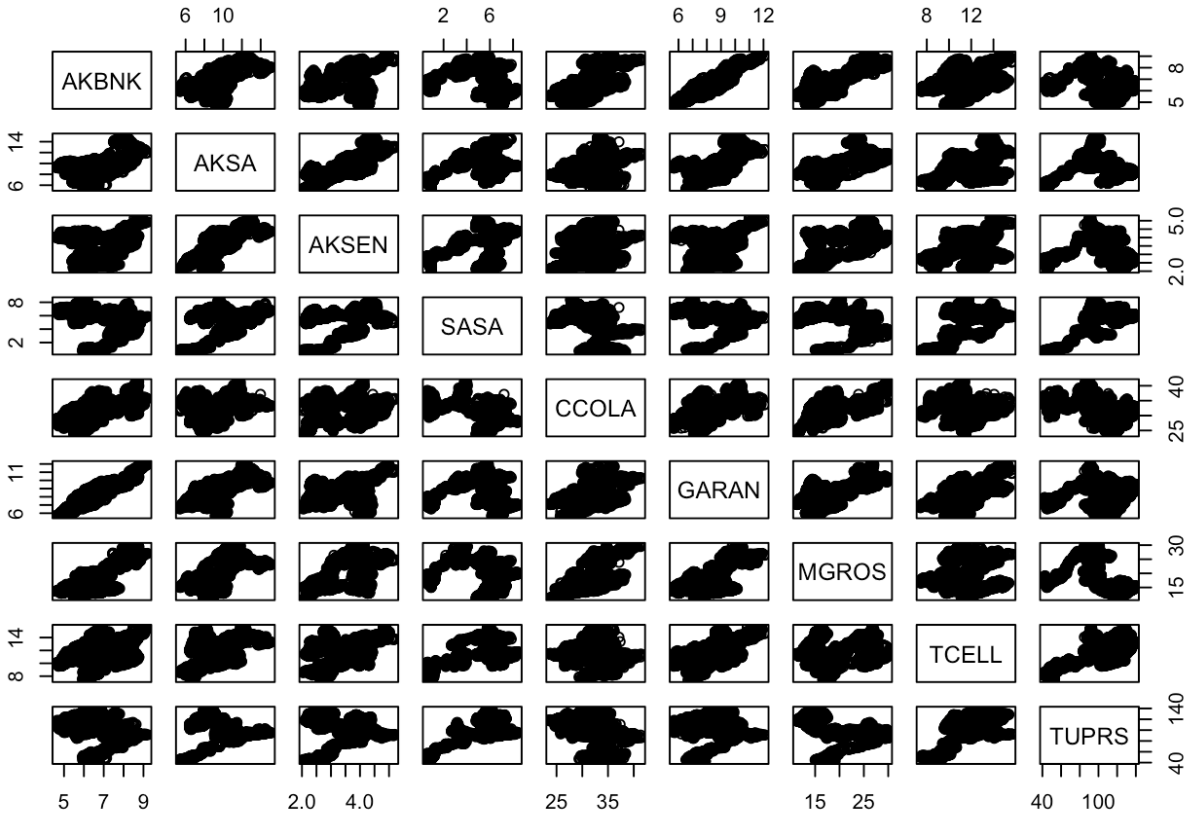


Figure 1

I have plotted the scatterplot AKBNK vs. GARAN and calculated correlation value. The correlation value 0.907 indicates a strong positive correlation between these stocks as well as the plot. It makes sense because these two stocks are owned by two banks, namely Akbank and Garanti and it is meaningful that these stocks have a strong correlation due to the belonging to the same sector which is finance.

On the other hand, there is also a positive relationship between AKSA and AKSEN. The correlation value is 0.636. In this case, the relationship is not as strong as the previous one and I had difficulty to understand the reason behind this correlation. Further investigation might be necessary for the interpretation. Or can it be just a coincidence?

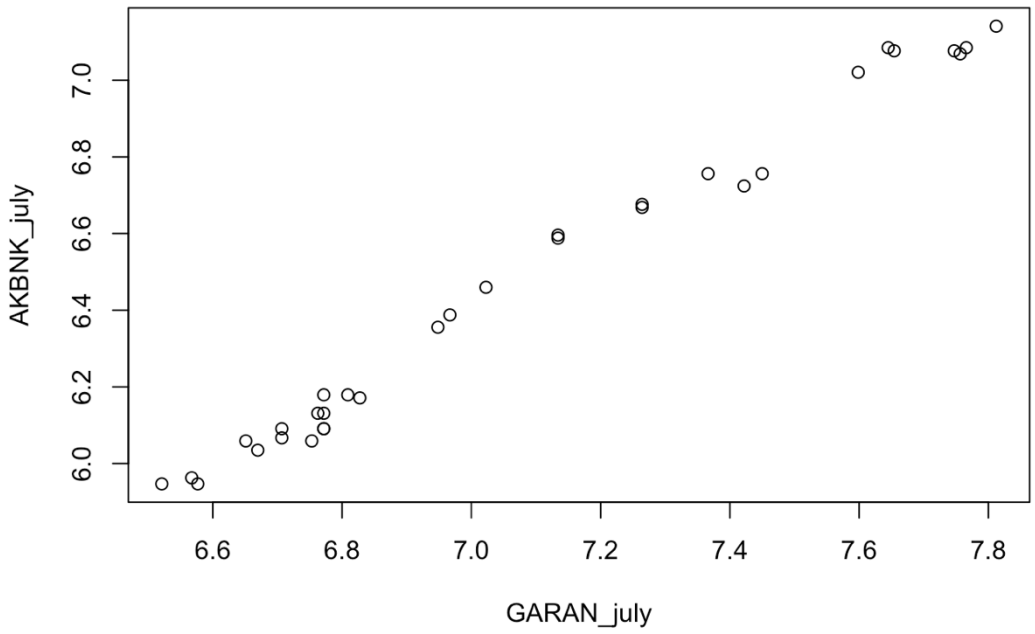
I have concluded the descriptive statistics by stating the standard deviation, variance, mode, interquartile range values of these four stocks. At the end, I plotted each stock's histogram to see the distribution of the stock values. At this point, the mode of each stock might be a meaningful measure for the estimations in the long run.

4.2 Moving Window Correlation

It may be also useful to investigate the correlation over a moving window such as monthly time periods. Thus, I defined samples of each stock (GARAN; AKBNK) in a monthly time horizon. As I investigate the monthly correlation between AKBNK and GARAN, I realized that

it really differs month to month, but in general the strong positive correlation still holds when we analyze the plots and correlation values. One important finding might be the correlation value in month July (2016) is 0.9935 which indicates a very strong positive correlation (Figure 3). However, I was not satisfied with monthly analysis because of two reasons; first the sample sizes are relatively small, second the variance of correlation values is high. So, I decided to apply the same method in a yearly time period. After I conduct the yearly analysis, I see that the correlation values' variance is lower and they are closer to the general value (0.907) which is expected.

```
plot(GARAN_july, AKBNK_july)
```




```
cor(GARAN_july, AKBNK_july)
```

```
## [1] 0.9934856
```

Figure 2

4.3 Principal Component Analysis (PCA)

Principal component analysis is a useful to apply dimension reduction. In this case, I consult to the datacamp's PCA tutorial to conduct this analysis. By using this approach, we can see which variable has a stronger contribution. In order to apply we first normalize the dataset and then create the correlation matrix. However, the correlation matrix which is shown in Figure 3 is not really easy to interpret. Thus, I decide to use my preferred 10 stocks and create a correlation matrix as it is shown in Figure 4.

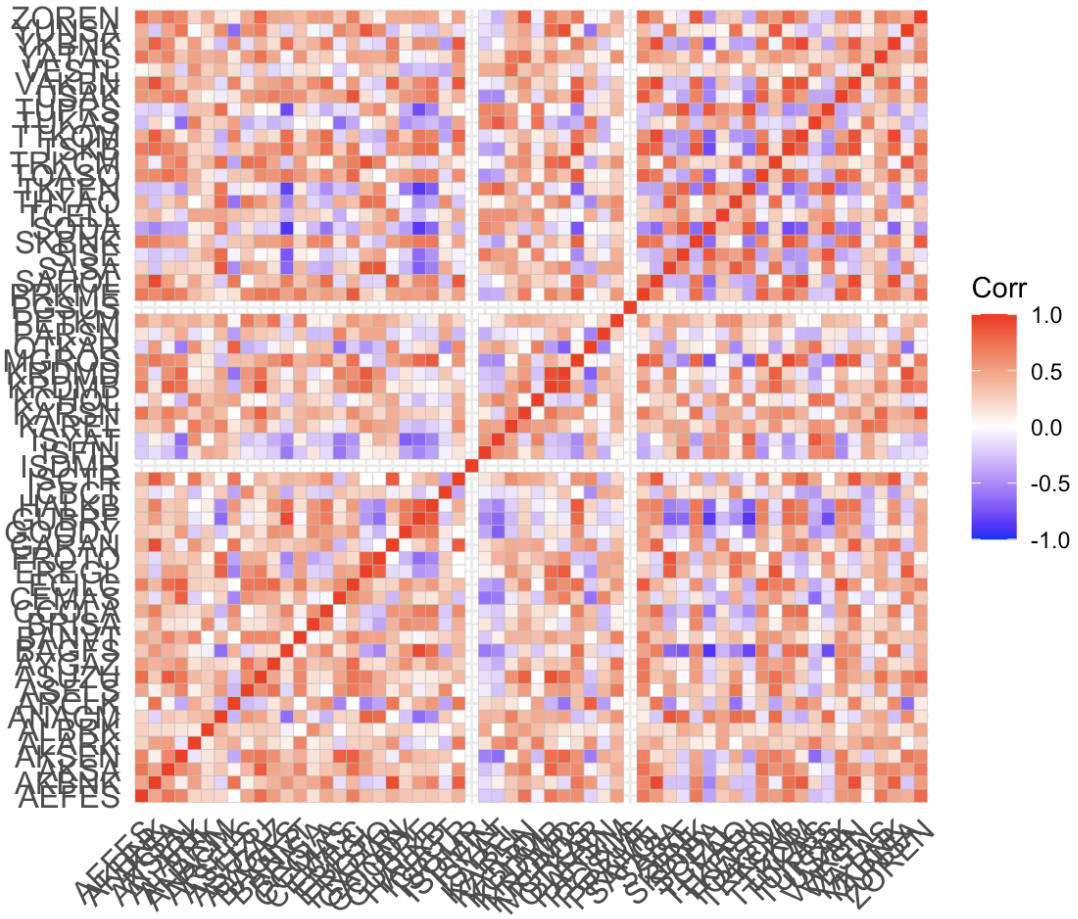


Figure 3

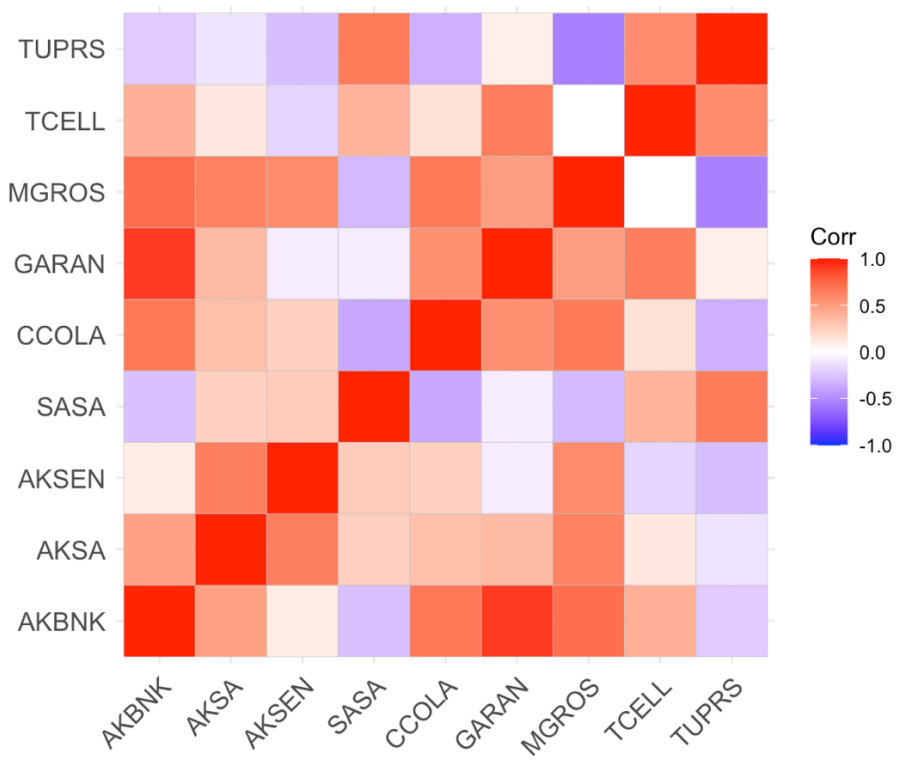


Figure 4

Each component explains a percentage of the total variance in the data set. As the cumulative proportion indicates he first principal component explains %69.5 of the total variance. The second one explains 25.73% of the total variance. The cumulative proportion of Comp.1 and Comp.2 explains nearly 95% of the total variance. This means that the first two principal components can accurately represent the data.

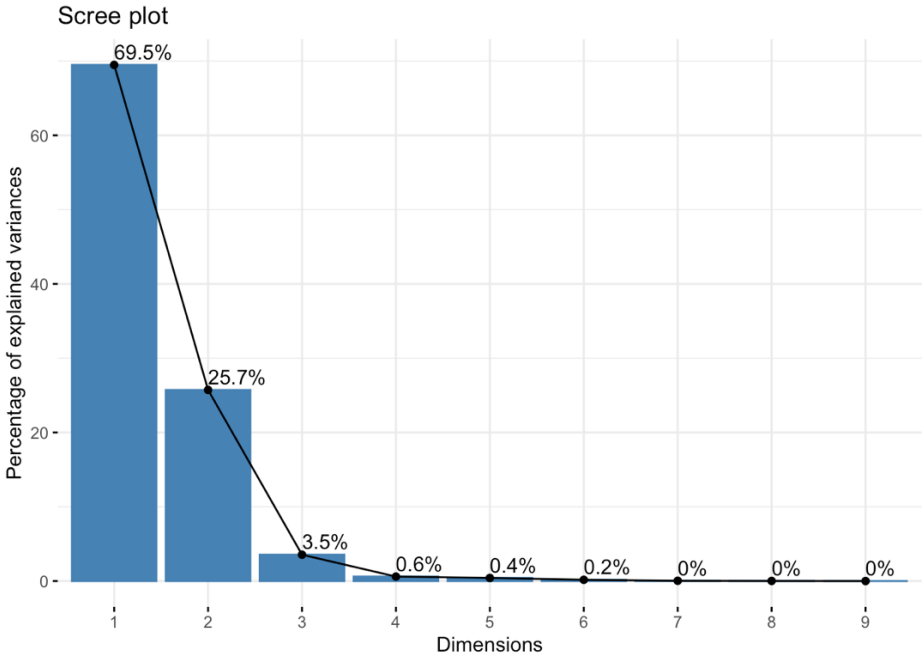


Figure 5

Moreover, we can visualize the PCA analysis with the help of scree plot and biplot of the variables as they are shown in Figure 5 and 6. It is possible to visualize the similarities and dissimilarities between the samples, and further shows the impact of each variable on each of the principal components.

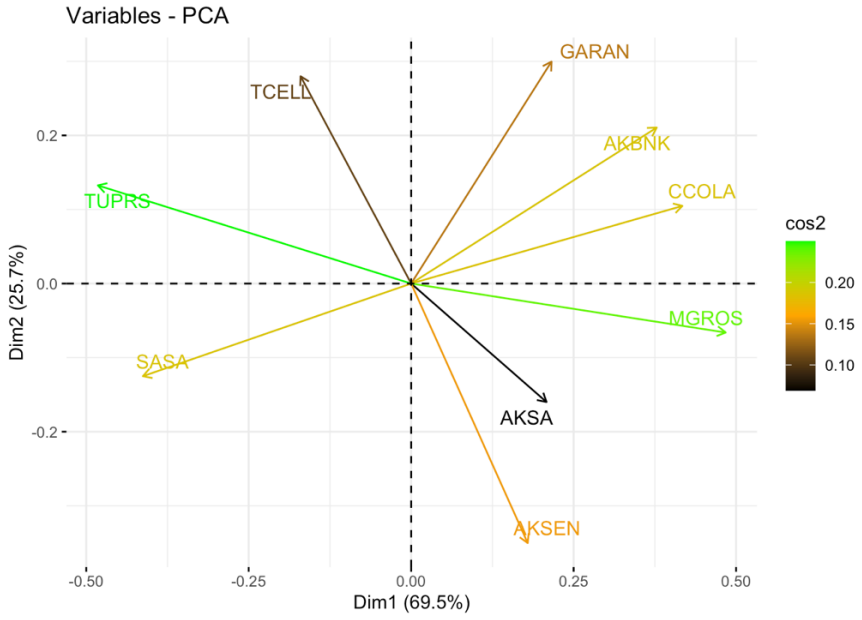


Figure 6

4.4 Inference with Google Trends

Lastly, we conduct an analysis using google trends. We are expecting to see similar relationships that we find out in the descriptive statistics part. In other words, we observed a positive strong correlation between AKBNK and GARAN, and we also observed a positive correlation between AKSA and AKSEN. However, we cannot really such a relationship when we analyze the google trends plot between AKBNK and GARAN (Figure 7). Although there are some common ups and downs, we cannot really detect a strong positive correlation.

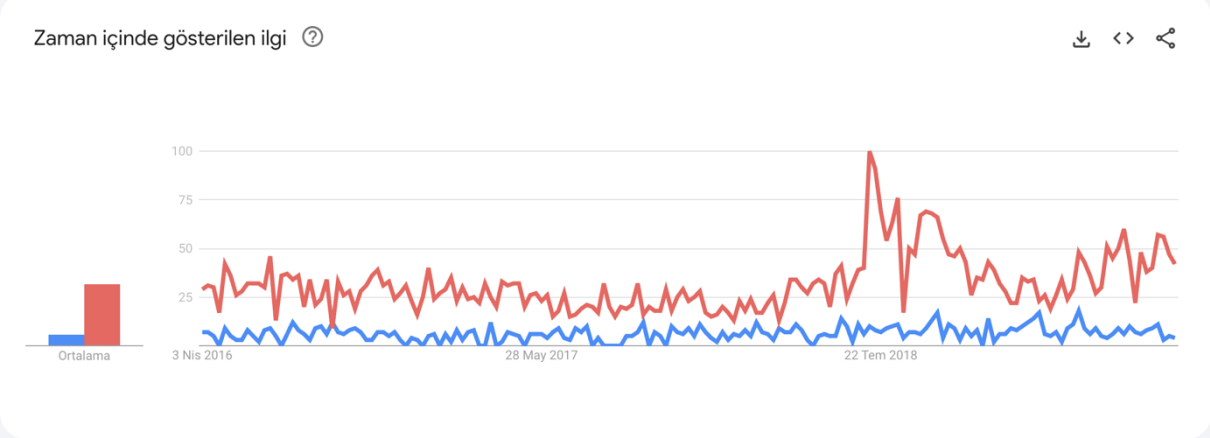


Figure 7

Similarly, there is not enough evidence to claim that AKSA and AKSEN stocks are positively correlated according to google searches. We may find some similar patterns which may imply a weak correlation between them as it can been seen in Figure 8. So, as a result of this analysis we may claim that the fluctuations of the stocks are not directly represented in the google trends. Another tool can be used to catch more significant patterns.

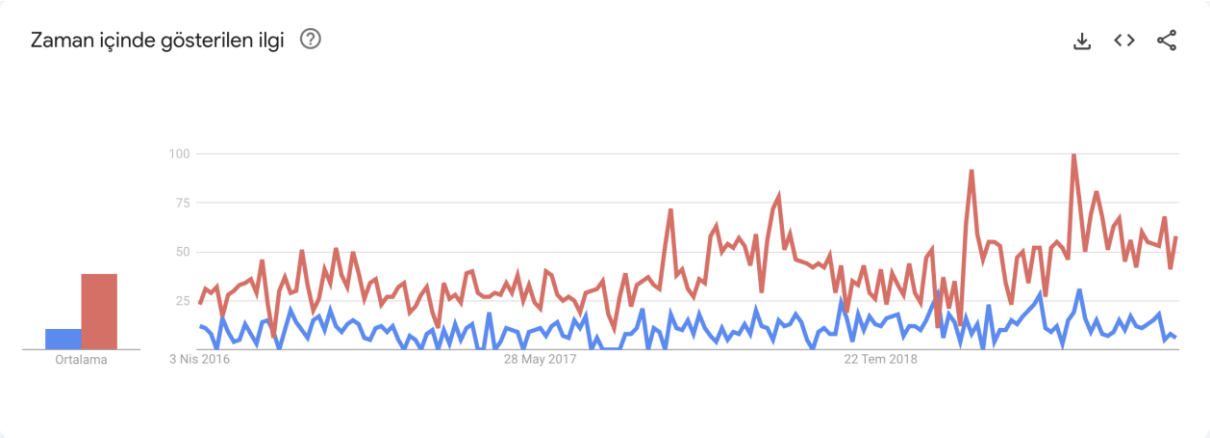


Figure 8