IE 360 HW1

Baran Kırkgöz

2022-04-14

In this homework different time series data are selected from https://evds2.tcmb.gov.tr/ and related keyword data from https://trends.google.com/trends to analyze. From TCMB, US Dollar exchange rate, employment rate and clothing price index datas are chosen and from google trends search volume of "trendyol" keyword is selected.The datasets provide information from 01.2017 to 12.2021. The aim of this study is to investigate relationship between those datasets and the 2 questions are "Does changing in the dollar rate and employment rate effect clothing price index?" and "Is there any relation between clothing price index and 'trendyol' search volume on google?".

Firstly, the required packages are downloaded.

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Then, data file (hw) is uploaded and the first column's data type is changed to date and others to numeric. By this way, it will be possible to group the data by year and it will be easier to plot them.

```
library(readr)
hw <- read_csv("hw.csv", col_types = cols(Date = col_date(format = "%m/%Y"),
    `TP YISGUCU2 G7` = col_number(), `TP DK USD A YTL` = col_number(),
    `TP FG J031` = col_number()))
```

```
## Warning: One or more parsing issues, see `problems()` for details
```

After organizing data types of columns, data is checked to see if there is any unrelated rows. As there can be seen, the rows after 61 are not relavent.

```
str(hw)
```

```
## spec_tbl_df [82 x 4] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Date           : Date[1:82], format: "2017-01-01" "2017-02-01" ...
##  $ TP YISGUCU2 G7 : num [1:82] 44.1 45.2 45.8 46.8 48 47.4 47.7 48.1 47.7 47.6 ...
##  $ TP DK USD A YTL: num [1:82] 3.73 3.67 3.67 3.65 3.56 3.52 3.56 3.51 3.47 3.66 ...
##  $ TP FG J031     : num [1:82] 177 172 175 191 204 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Date = col_date(format = "%m/%Y"),
##   ..   `TP YISGUCU2 G7` = col_number(),
##   ..   `TP DK USD A YTL` = col_number(),
##   ..   `TP FG J031` = col_number()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Irrelavent rows are cleaned and the column names are changed to make them more understandable.

```
data<-hw[-c(61:82),]
colnames(data)<-c("date","employment.rate","usd","clothing")
```

Then a year column is added to data so that it can be visualized by years (A function from lubridate package is used).

```
data$Year<- year(data$date)
data
```

```
## # A tibble: 60 x 5
##    date       employment.rate   usd clothing  Year
##    <date>               <dbl> <dbl>    <dbl> <dbl>
##  1 2017-01-01            44.1  3.73     177.  2017
##  2 2017-02-01            45.2  3.67     172.  2017
##  3 2017-03-01            45.8  3.67     175.  2017
##  4 2017-04-01            46.8  3.65     191.  2017
##  5 2017-05-01            48    3.56     204.  2017
##  6 2017-06-01            47.4  3.52     201.  2017
##  7 2017-07-01            47.7  3.56     194.  2017
##  8 2017-08-01            48.1  3.51     187.  2017
##  9 2017-09-01            47.7  3.47     189.  2017
## 10 2017-10-01            47.6  3.66     211.  2017
## # ... with 50 more rows
```

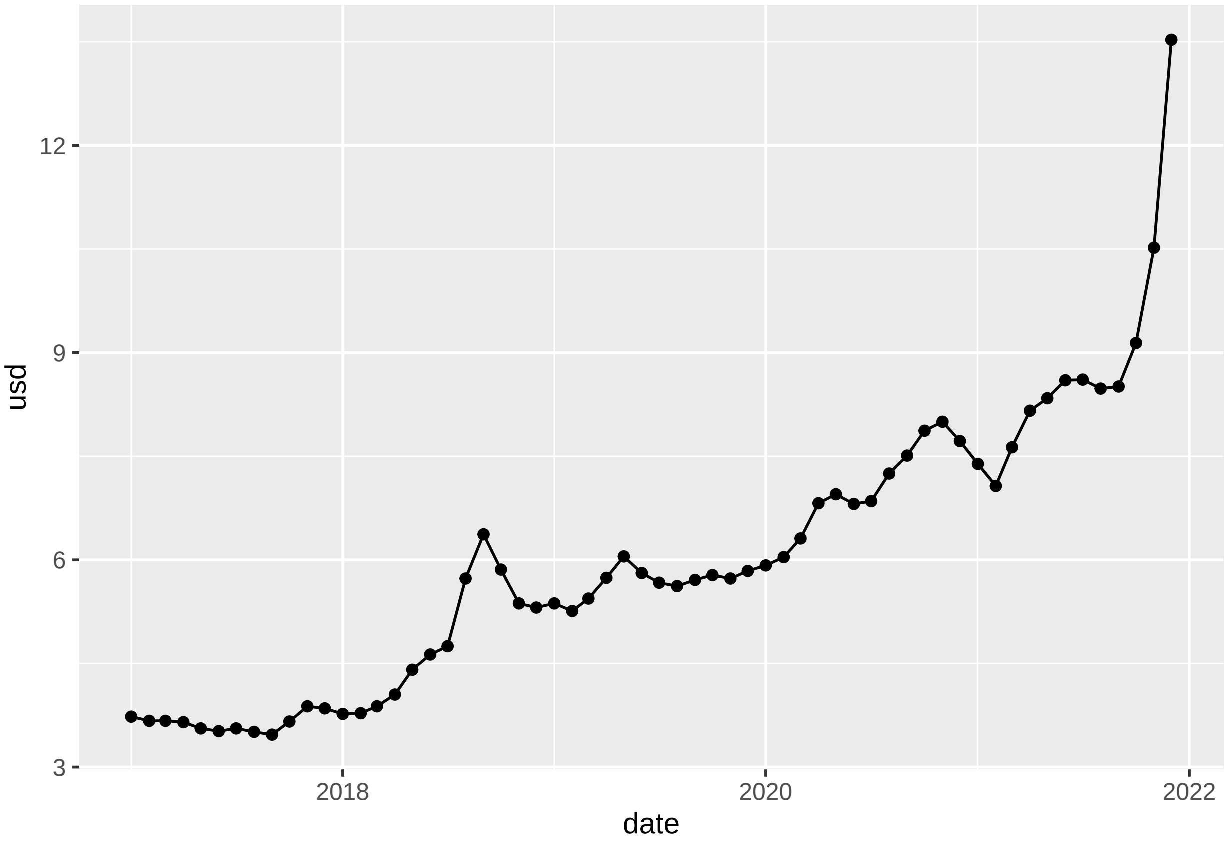
Before going into analyzing relationships, 4 different datas will be visualized indivisually.

USD

Line Graph

Firstly, the line graph is created to observe the pattern. As there can be seen easily, there is a strong trend in data especially after 2020. This is related to economic crisis in Turkey.

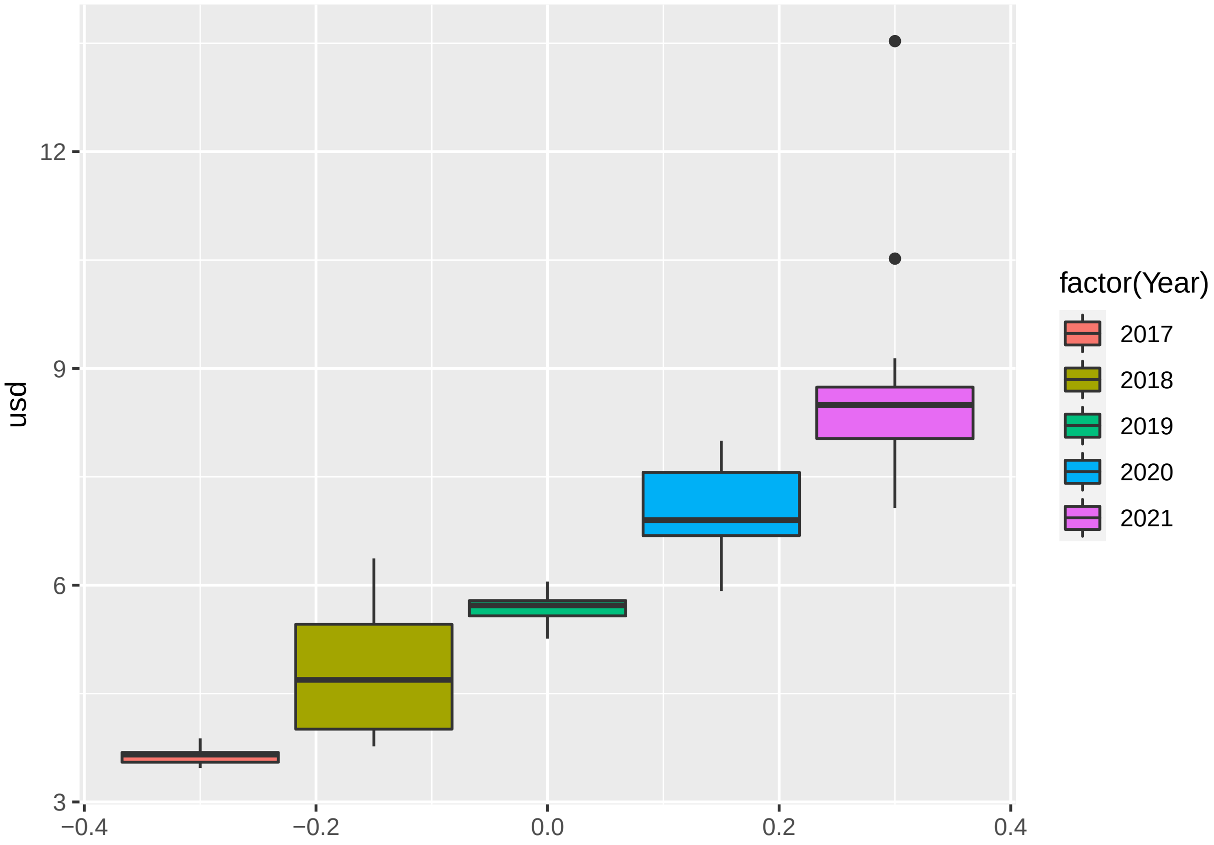
```
ggplot(data=data, aes(x=date, y=usd, group=1)) +
  geom_line()+
  geom_point()
```



Box plot

As there can be seen, the mean of the data always increases. To investigate deeper, mean values will be analyzed below. And also the standard deviation of every year is increasing slightly. IT will be analyzed too (functions from dplyr are used).

```
ggplot(data, aes(y=usd, fill=factor(Year))) +
  geom_boxplot()
```



The means of the last 2 years are quite high.

```
data %>%
  mutate(year = year(date)) %>%
  group_by(year) %>%
  summarise(mean_var = mean(usd))
```

```
## # A tibble: 5 x 2
##    year mean_var
##   <dbl>    <dbl>
## 1  2017     3.64
## 2  2018     4.83
## 3  2019     5.67
## 4  2020     7.00
## 5  2021     8.83
```

The sd of last year is very high due to the jump in dollar rate in last months.

```
data %>%
  mutate(year = year(date)) %>%
  group_by(year) %>%
  summarise(sd_var = sd(usd))
```

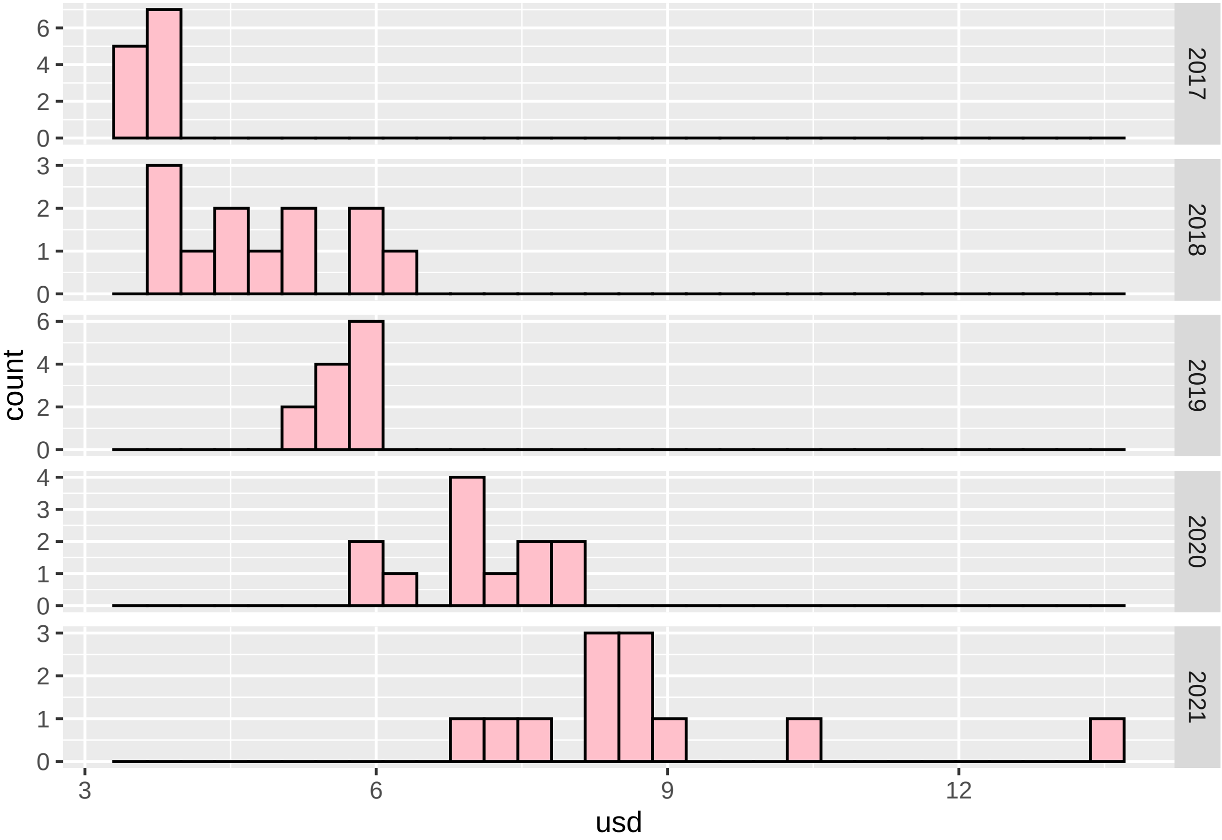
```
## # A tibble: 5 x 2
##    year sd_var
##   <dbl>  <dbl>
## 1  2017  0.130
## 2  2018  0.891
## 3  2019  0.219
## 4  2020  0.691
## 5  2021  1.72
```

Histogram

The shape of histograms change every year. There is a high fluctuation.

```
ggplot(data, aes(x = usd)) +
  geom_histogram(fill = "pink", colour = "black") +
  facet_grid(Year ~ ., scales = "free")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

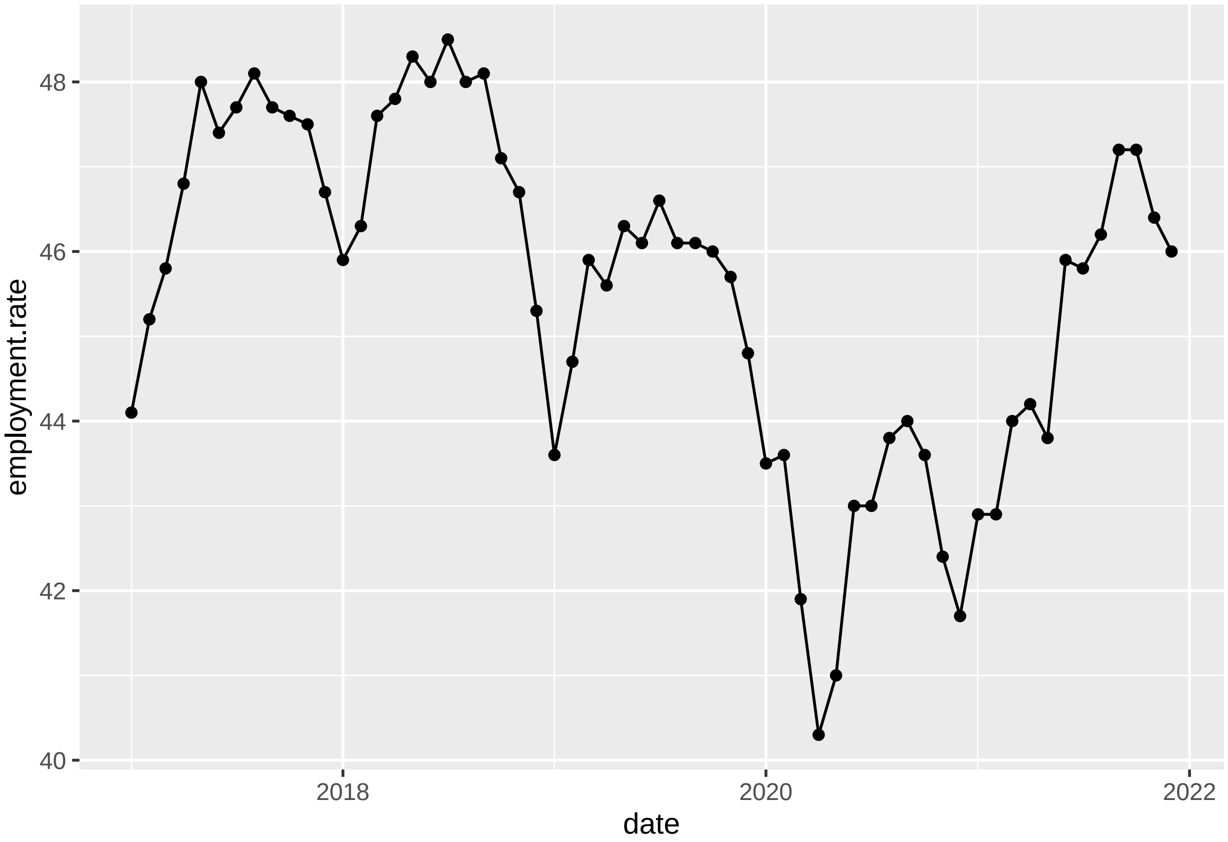


Employment Rate

Line Graph

There is no obvious trend or seasonality in the data but at the beginning of the 2020, employment rate falls deeply, this is related to beginning of the pandemic.

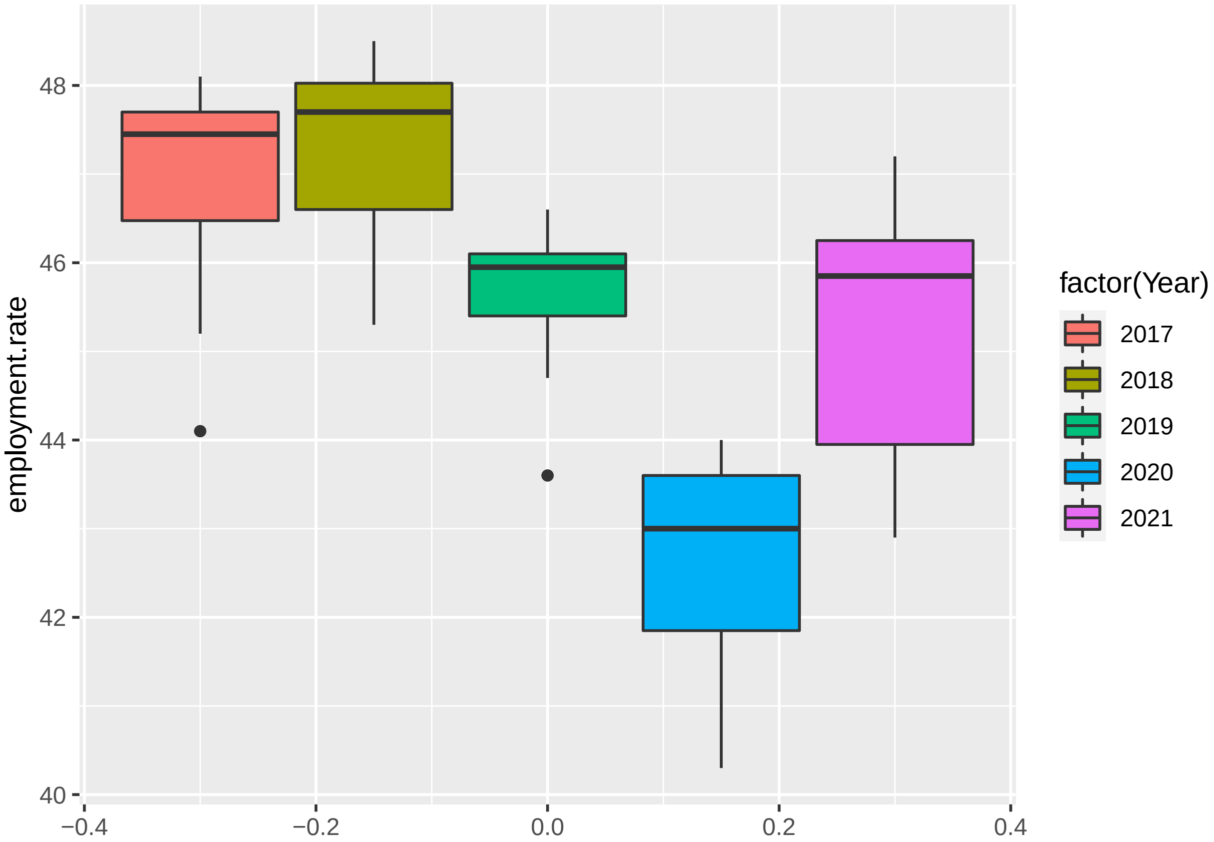
```
ggplot(data=data, aes(x=date, y=employment.rate, group=1)) +
  geom_line()+
  geom_point()
```



Box Plot

Except 2020, the means are very close.

```
ggplot(data, aes(y=employment.rate, fill=factor(Year))) +
  geom_boxplot()
```



As there can be seen below, means are very close except 2020. In 2020, it is very low.

```
data %>%
  mutate(year = year(date)) %>%
  group_by(year) %>%
  summarise(mean_var = mean(employment.rate))
```

```
## # A tibble: 5 x 2
##    year mean_var
##   <dbl>    <dbl>
## 1  2017     46.9
## 2  2018     47.3
## 3  2019     45.6
## 4  2020     42.6
## 5  2021     45.2
```

The highest deviation belongs to year 2021, it is because pandemic started loose its affect and people started to get back their normal lifes.Because of that employment rate increases from beginning of 2021 till end of the year. This is why standard deviation is relatively high.

```
data %>%
  mutate(year = year(date)) %>%
  group_by(year) %>%
  summarise(sd_var = sd(employment.rate))
```

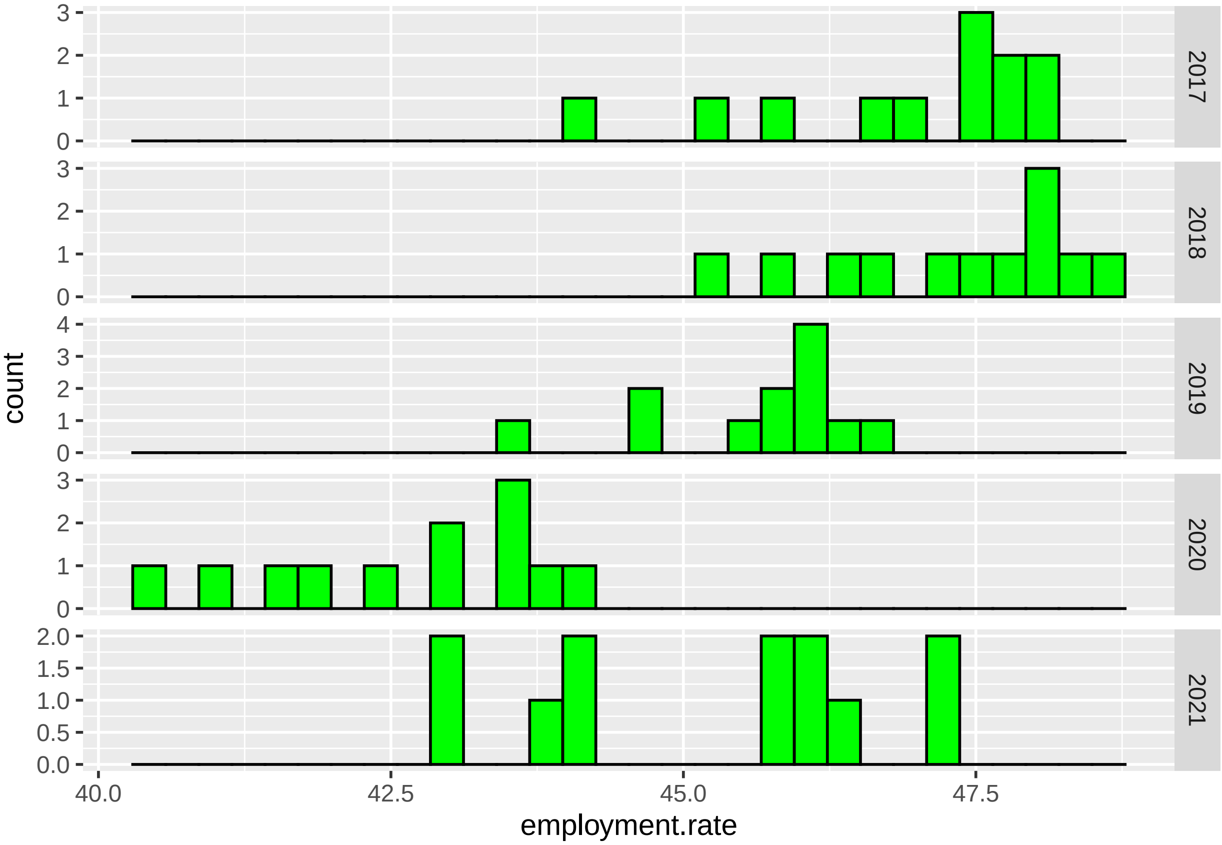
```
## # A tibble: 5 x 2
##    year sd_var
##   <dbl>  <dbl>
## 1  2017   1.24
## 2  2018  1.03
## 3  2019  0.850
## 4  2020  1.19
## 5  2021  1.56
```

Histograms

The histograms' shape are quite similar except 2020.

```
ggplot(data, aes(x = employment.rate)) +
  geom_histogram(fill = "green", colour = "black") +
  facet_grid(Year ~ ., scales = "free")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

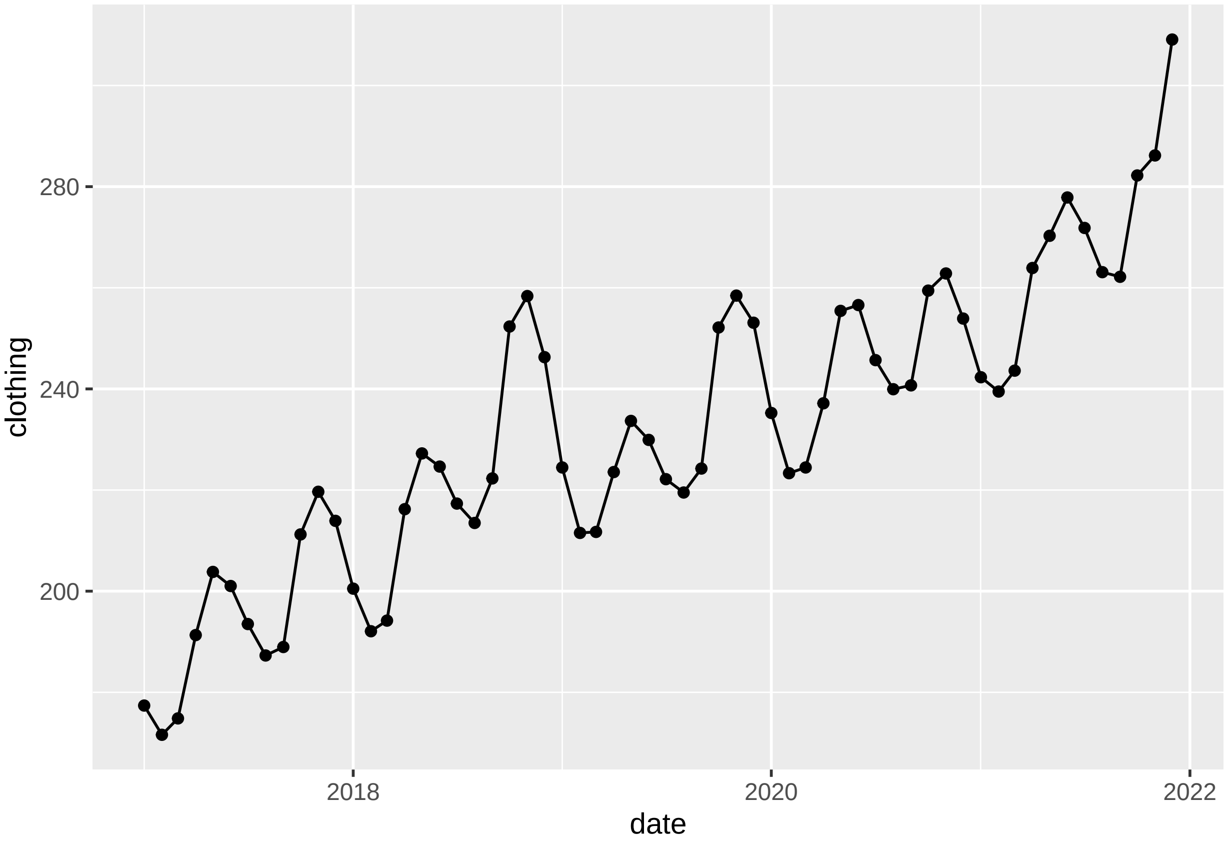


Clothing Price Index

Line Graph

There is a strong increasing trend and also strong seasonality. The graph looks very similar to usd's line graph. This can be a significant evidance to a relationship between these datasets.

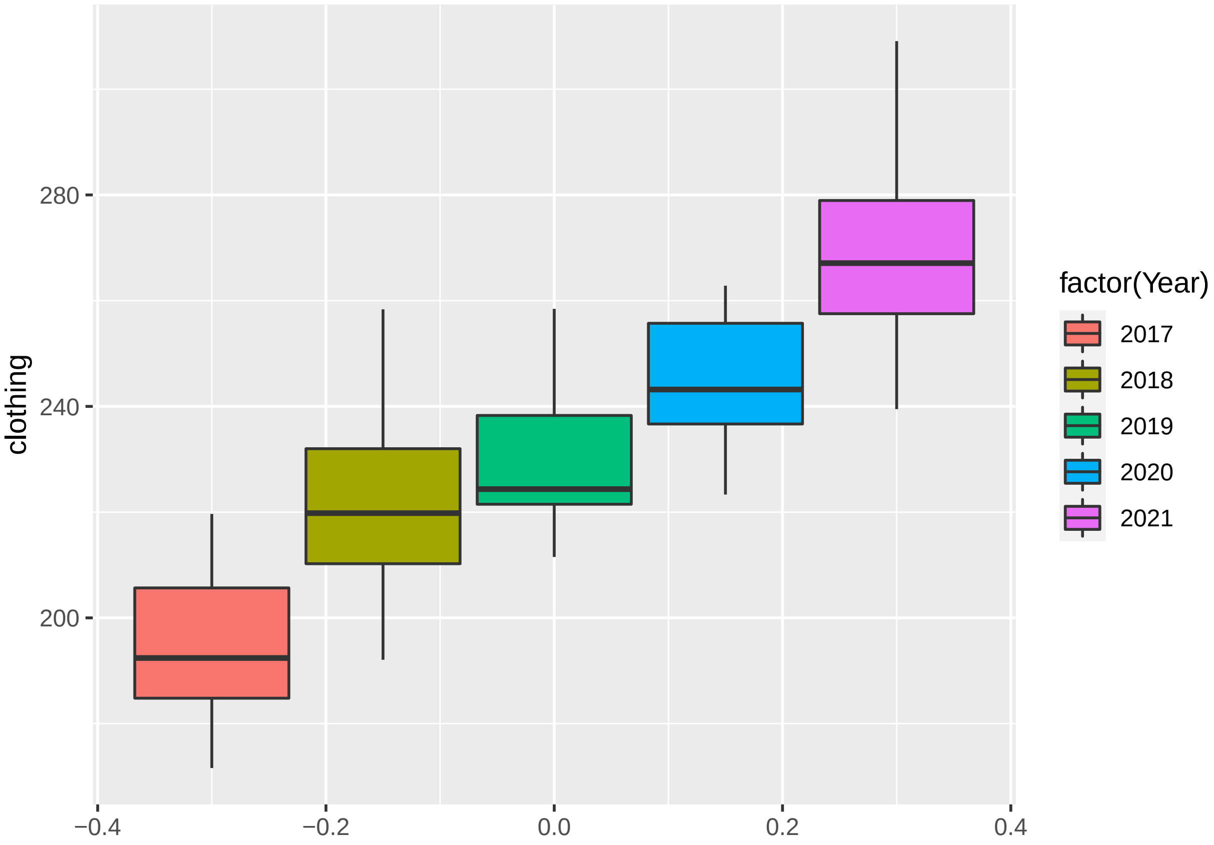
```
ggplot(data=data, aes(x=date, y=clothing, group=1)) +
  geom_line()+
  geom_point()
```



Box Plot

Means are increasing and deviations looks similar.

```
ggplot(data, aes(y=clothing, fill=factor(Year))) +
  geom_boxplot()
```



As there can be seen, clothing price index increases every year on average.

```
data %>%
  mutate(year = year(date)) %>%
  group_by(year) %>%
  summarise(mean_var = mean(clothing))
```

```
## # A tibble: 5 x 2
##    year mean_var
##   <dbl>    <dbl>
## 1  2017     195.
## 2  2018     222.
## 3  2019     230.
## 4  2020     245.
## 5  2021     268.
```

Deviations are close.

```
data %>%
  mutate(year = year(date)) %>%
  group_by(year) %>%
  summarise(sd_var = sd(clothing))
```

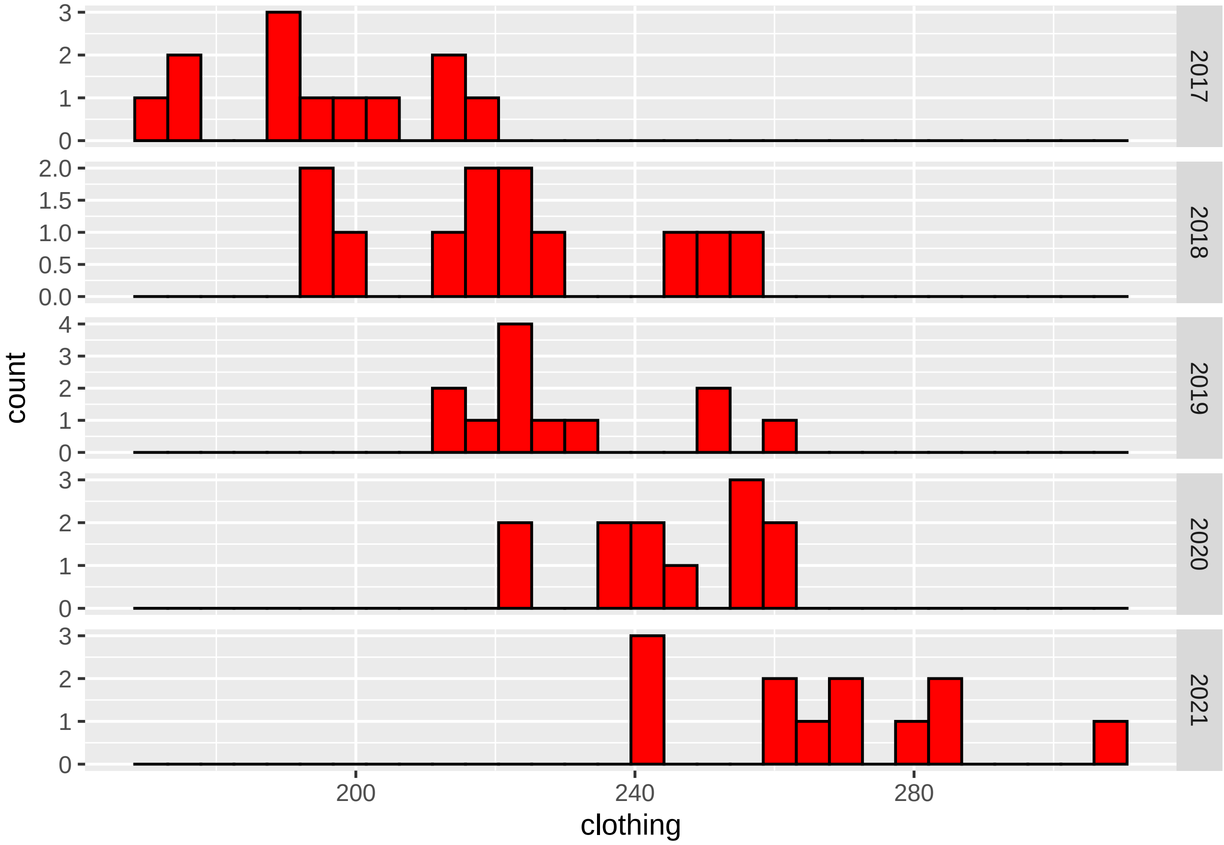
```
## # A tibble: 5 x 2
##    year sd_var
##   <dbl>  <dbl>
## 1  2017   15.7
## 2  2018   21.6
## 3  2019   15.9
## 4  2020   13.3
## 5  2021   20.2
```

Histograms

Also the histogram is very similar to usd's. There is a high fluctuation.

```
ggplot(data, aes(x = clothing)) +
  geom_histogram(fill = "red", colour = "black") +
  facet_grid(Year ~ ., scales = "free")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Google trends data

Now the search volume of keyword "trendyol" will be analyzed. Again, the data is prepared to deep analyze firstly. The data types are changed and year column is added. And also the name of the columns is Date and "trendol". The typo is to prevent the confusion between the name of the dataset and variable.

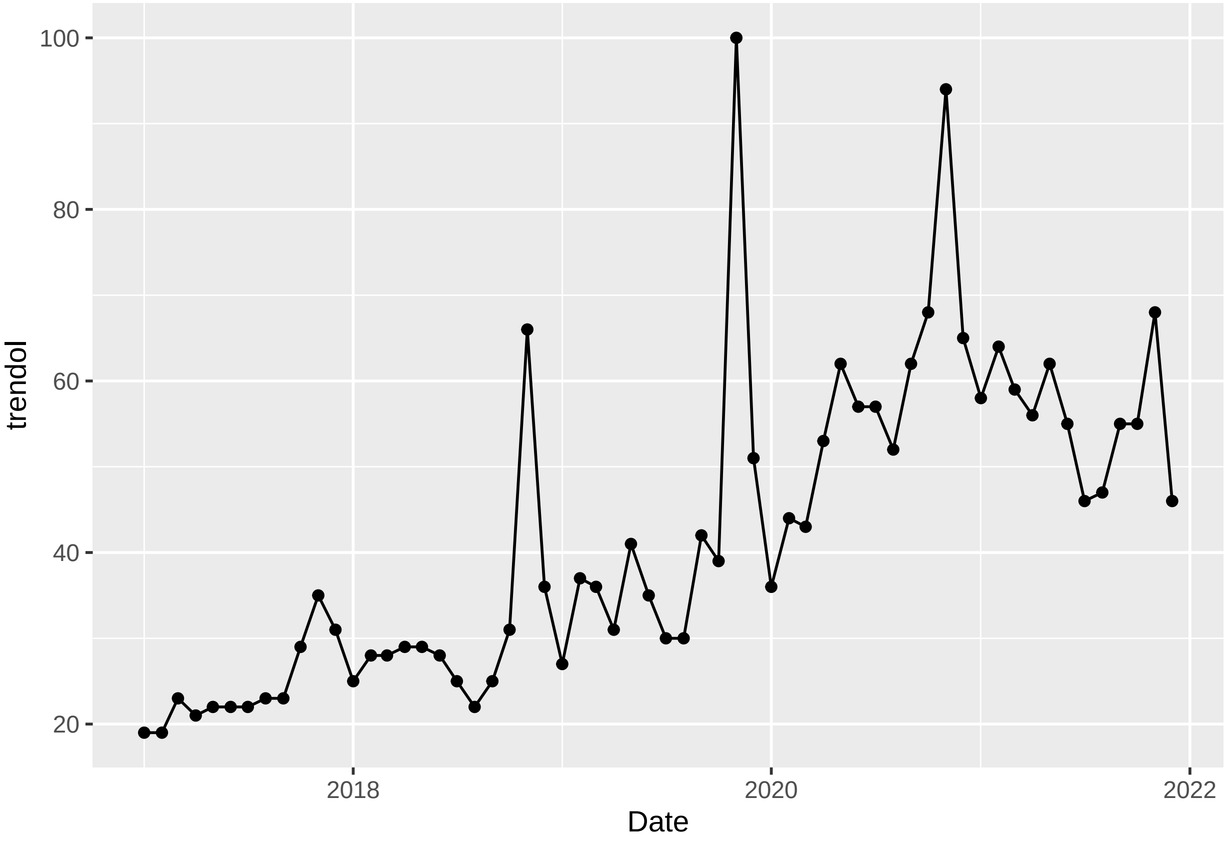
```
library(readr)
trendyol <- read_csv("multiTimeline (3).csv",
    col_types = cols(Date = col_date(format = "%m/%Y"),
        trendol = col_number()))
trendyol$Year<- year(trendyol$Date)

trendyol
```

```
## # A tibble: 60 x 3
##    Date       trendol  Year
##    <date>       <dbl> <dbl>
##  1 2017-01-01      19  2017
##  2 2017-02-01      19  2017
##  3 2017-03-01      23  2017
##  4 2017-04-01      21  2017
##  5 2017-05-01      22  2017
##  6 2017-06-01      22  2017
##  7 2017-07-01      22  2017
##  8 2017-08-01      23  2017
##  9 2017-09-01      23  2017
## 10 2017-10-01      29  2017
## # ... with 50 more rows
```

Line graph

There is an increasing trend and also seasonality in the data. As there can be seen, at the beginng of every year, volume is at maximum.

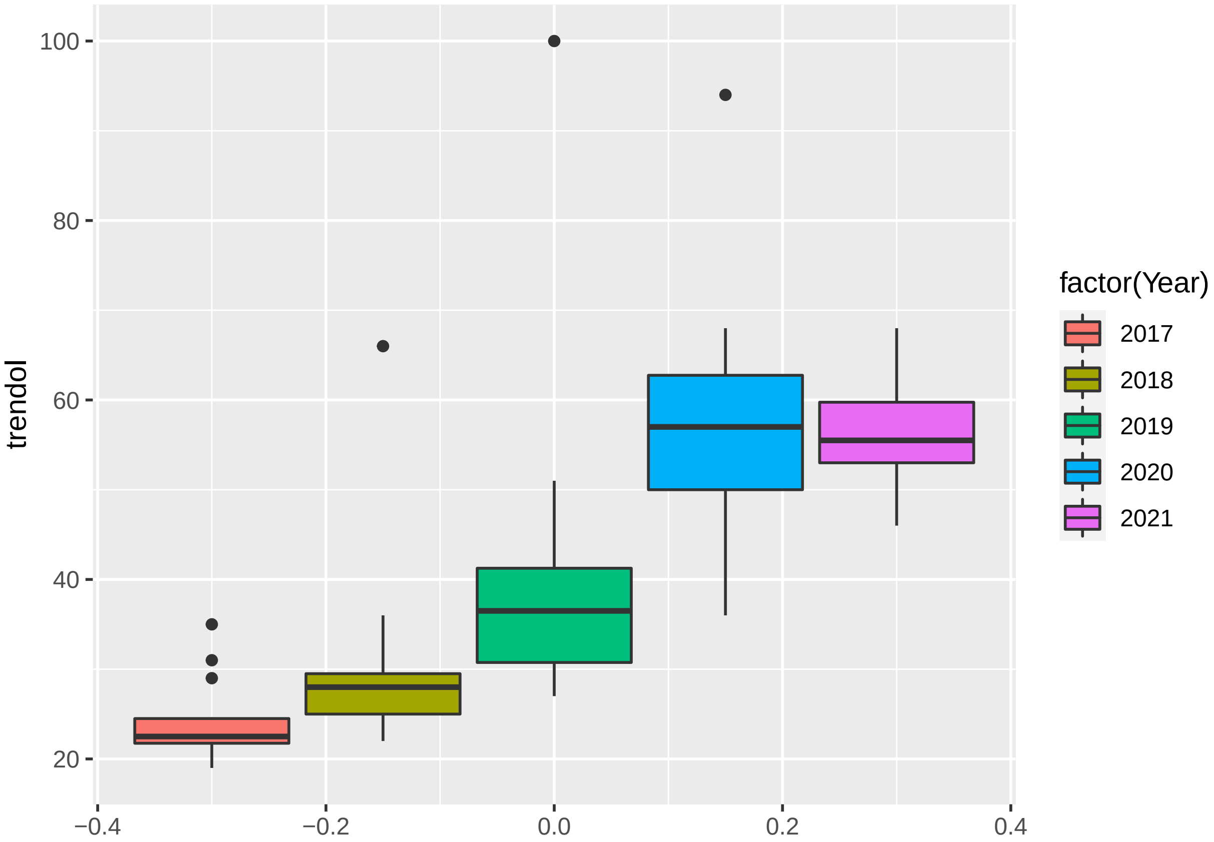
```
ggplot(data=trendyol, aes(x=Date, y=trendol, group=1)) +
  geom_line()+
  geom_point()
```



Box Plot

The means of the years increases every year except the last one. It can be caused by "back to normal life after covid". People started to shop offline again.

```
ggplot(trendyol, aes(y=trendol, fill=factor(Year))) +
  geom_boxplot()
```



Max mean belongs to 2020. It is because people were stucked at their home and there was no way but online shopping.

```
trendyol %>%
  mutate(year = year(Date)) %>%
  group_by(year) %>%
  summarise(mean_var = mean(trendol))
```

```
## # A tibble: 5 x 2
##    year mean_var
##   <dbl>    <dbl>
## 1  2017     24.1
## 2  2018     31
## 3  2019     41.6
## 4  2020     57.8
## 5  2021     55.9
```

The high sd in 2019 is related to big jump in the beginning of the year, and sd in 2020 is related to covid.

```
trendyol %>%
  mutate(year = year(Date)) %>%
  group_by(year) %>%
  summarise(sd_var = sd(trendol))
```

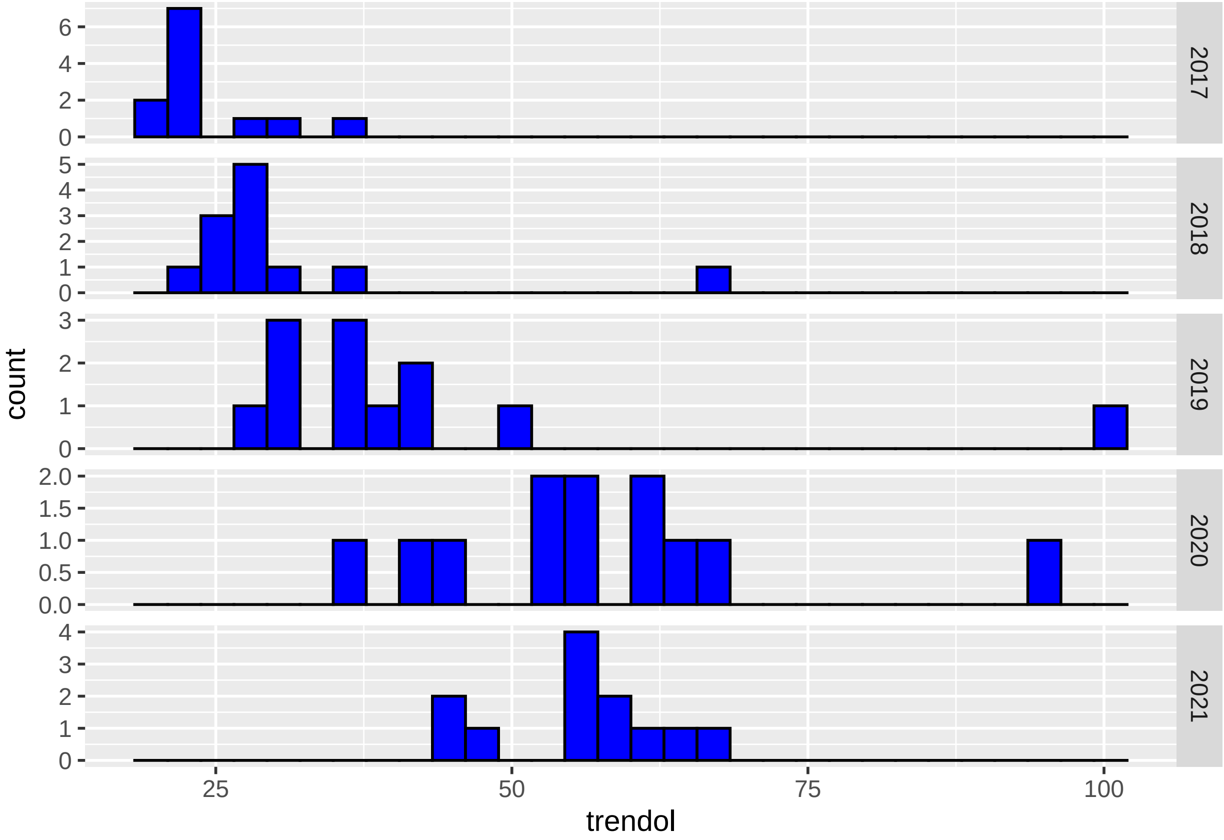
```
## # A tibble: 5 x 2
##    year sd_var
##    <dbl>  <dbl>
## 1  2017   4.94
## 2  2018  11.6
## 3  2019  19.5
## 4  2020  14.9
## 5  2021   7.01
```

Histogram

Histograms are similar to usds' and clothing price indexes' histograms. This can be an evidence for an relationship.

```
ggplot(trendyol, aes(x = trendol)) +
  geom_histogram(fill = "blue", colour = "black") +
  facet_grid(Year ~ ., scales = "free")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Second Part: Analyzing Relations

In this second part of homework, the relations between datasets will be analyzed. It will be done by using correlation. Correlation coefficients are indicators of the strength of the linear relationship between two different variables, x and y (investopedia.com). So if there is any relationship between two datasets, correlation coefficient will be close to -1 or 1. If there is no/weak relation between datasets, it will be close to 0.

Correlation between usd and clothing From above graphs it is observed that there can be a relation between usd exchange rate and clothing price index. This high correlation proves that there is a strong relationship between those.

```
cor.test(x = data$usd, y = data$clothing, method = "pearson", alternative = "greater")
```

```
##
##  Pearson's product-moment correlation
##
## data:  data$usd and data$clothing
## t = 15.063, df = 58, p-value < 2.2e-16
## alternative hypothesis: true correlation is greater than 0
## 95 percent confidence interval:
##  0.8384228 1.0000000
## sample estimates:
##       cor
## 0.8924236
```

Correlation between employment rate and clothing Between employment rate and clothing there is a negative correlation regarding to these calculation. It does makes sense because falling in the employment rate is a signal to problems in the economy and than can cause to increasing clothing price index.

```
cor.test(x = data$employment.rate, y = data$clothing, method = "pearson", alternative = "greater")
```

```
##
##  Pearson's product-moment correlation
##
## data:  data$employment.rate and data$clothing
## t = -2.4382, df = 58, p-value = 0.9911
## alternative hypothesis: true correlation is greater than 0
## 95 percent confidence interval:
##  -0.4875072  1.0000000
## sample estimates:
##        cor
## -0.3049054
```

Correlation between trendyol and clothing From the graphs it is observed that patterns of trendyol search volume and clothing price index are similar. Since correlation values is high, it could be said that there is a positive relationship between trendyol search volume and clothing price index.

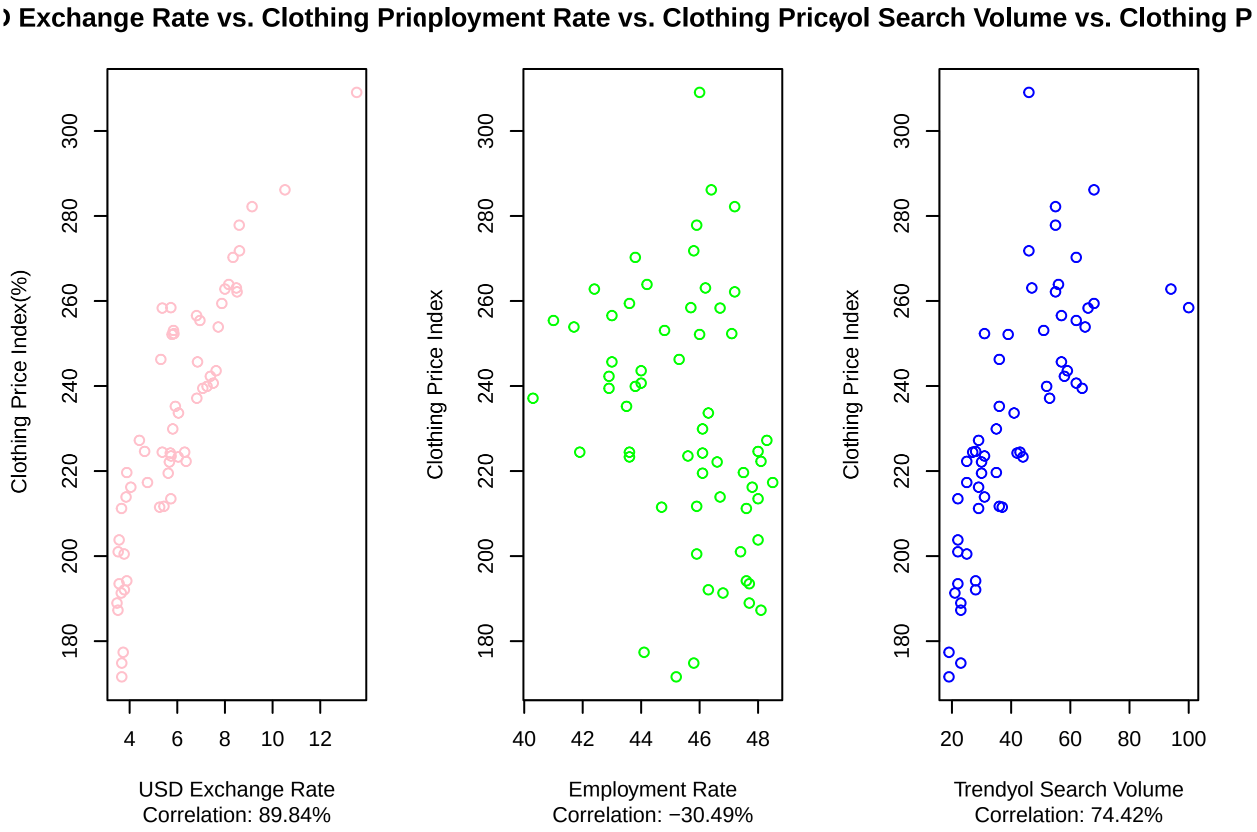
```
 cor.test(x = trendyol$trendol, y = data$clothing, method = "pearson", alternative = "greater")
```

```
##
##  Pearson's product-moment correlation
##
## data:  trendyol$trendol and data$clothing
## t = 8.4868, df = 58, p-value = 4.704e-12
## alternative hypothesis: true correlation is greater than 0
## 95 percent confidence interval:
##  0.6304205 1.0000000
## sample estimates:
##       cor
## 0.7442679
```

Pairwise Scatter Plot

In the last step, the pairwise scatter plots for each dataset duals are created to see the relations visually.

```
 par(mfrow=c(1,3))
plot(x = data$usd,y = data$clothing,
     main = "USD Exchange Rate vs. Clothing Price Index",
     xlab = "USD Exchange Rate", ylab = " Clothing Price Index(%)",
     sub = "Correlation: 89.84%", col="pink")
plot(x = data$employment.rate,y = data$clothing,
     main = "Employment Rate vs. Clothing Price Index",
     xlab = "Employment Rate", ylab = "Clothing Price Index",
     sub = "Correlation: -30.49%", col="green")
plot(x = trendyol$trendol,y = data$clothing,
     main = "Trendyol Search Volume vs. Clothing Price Index",
     xlab = "Trendyol Search Volume", ylab = "Clothing Price Index",
     sub = "Correlation: 74.42%", col = "blue")
```



In the first and third graphs, positive relationship is very strong. In the middle, the negative relation can be observed but it is weaker then first and third ones.

Conclusion

It is proved that there is a relation between clothing price index data and other datasets. That means a linear regression model can be created to predict clothing price index by using trendyol search volume, employment rate and usd dollar exchange rate.