YILDIZ TECHNICAL UNIVERSITY
DATA ANALYSIS HOMEWORK

ANALYSIS OF DATASETS

YİĞİTHAN BAKIRCI - 20023602

PROF. DR. ALİ HAKAN BÜYÜKLÜ

CONTENTS

Question 1

For DATA 1

Summary Statistic Table

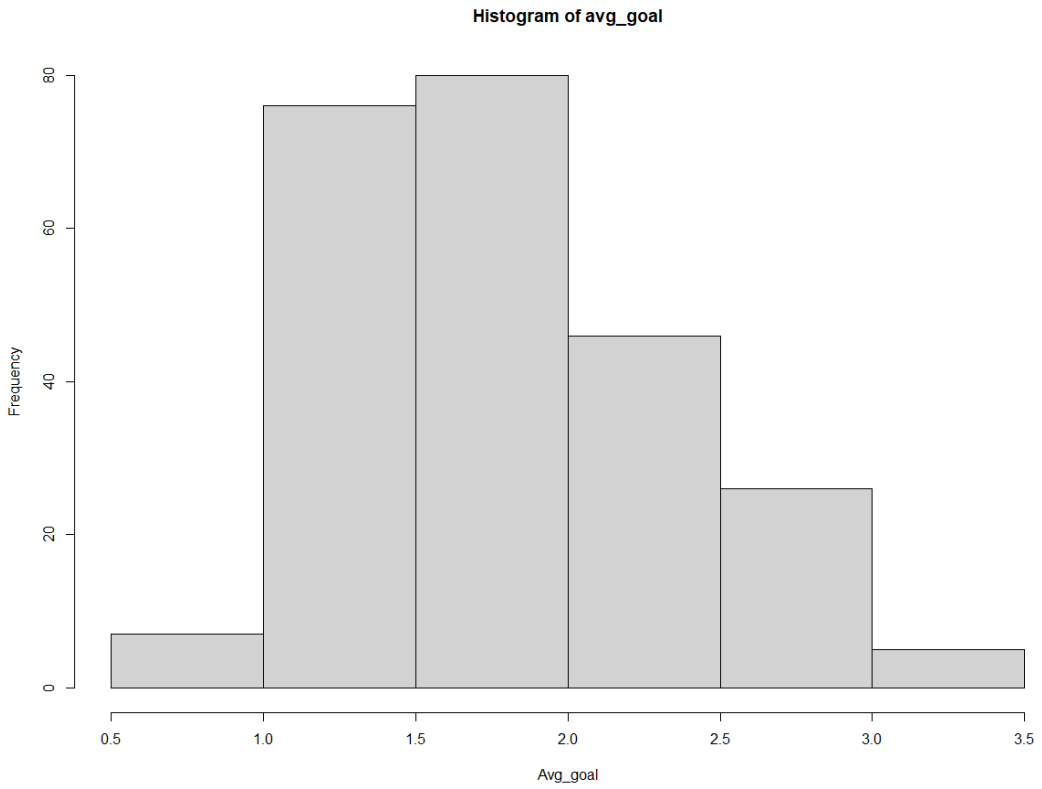
Mean	Mode	Median	Standart Deviation	Variance	Coefficient of variation	Skewness	Kurtosis
1.80666	2	1.67204	0.5342682	0.28544	29.57%	0.559638	-0.4502231

According to the summary statistic table, the kurtosis value of data1 is -0.45 and according to this value we can say that the curve is flatter than the normal distribution.

It is right-skewed because mean value is greater than median value.

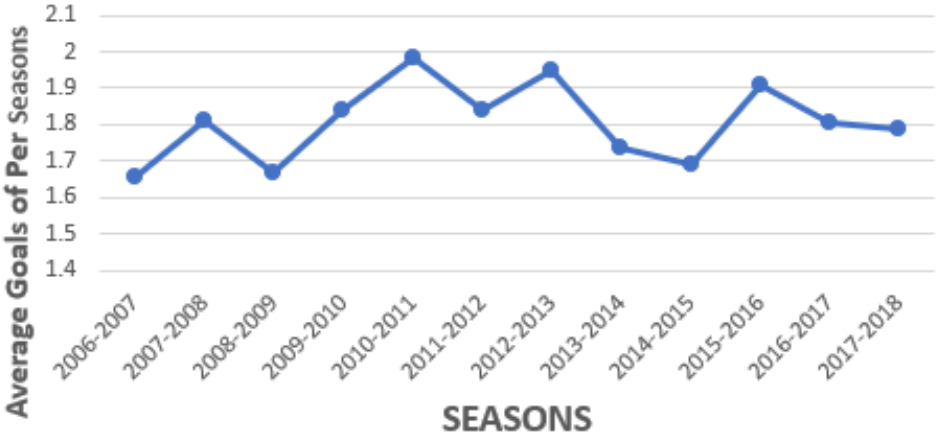
Histogram Graph

```
hist(data$avg_goal,main="Histogram of avg_goal",xlab="Avg_goal")
```



Values were collected mostly between 1.0-2.0. When we look at the histogram graph,we can say the distribution is similar to the normal distribution.

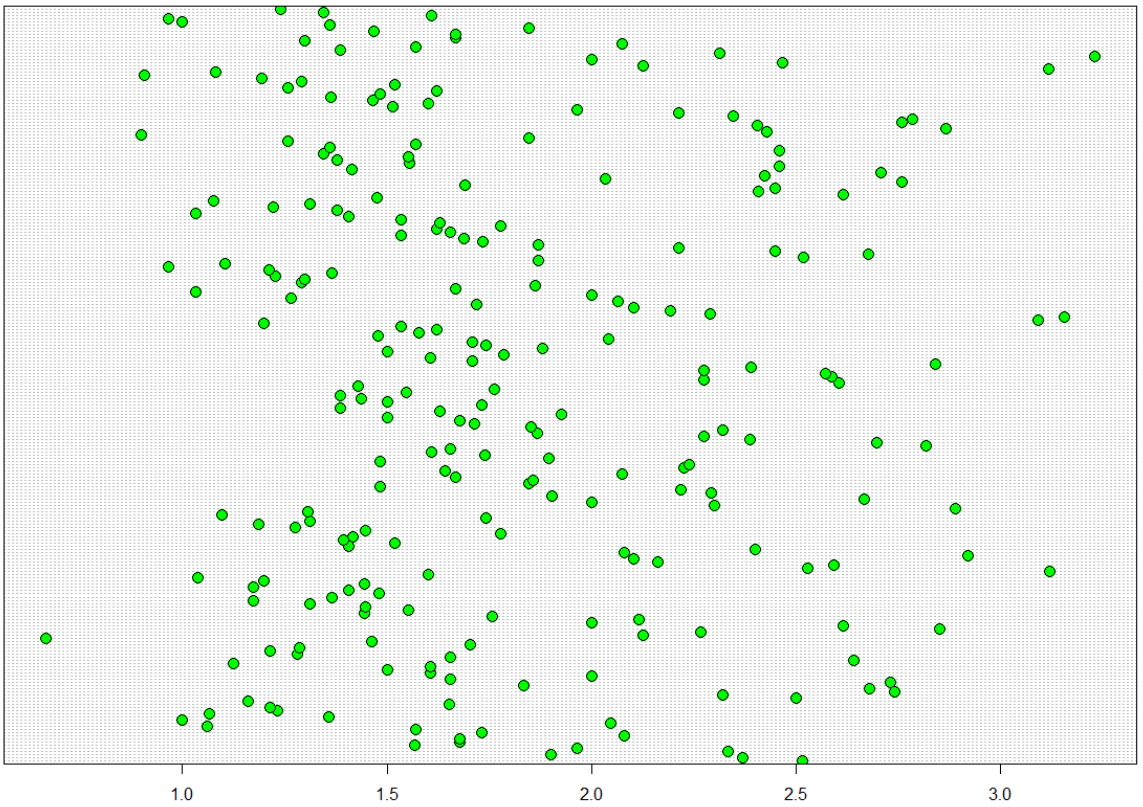
Polygon Graphic



According to polygon graph, the highest average number of goals was in the “2010-2011” season. The lowest average number of goals was in the “2006-2007” season.

Dot Plot Graph

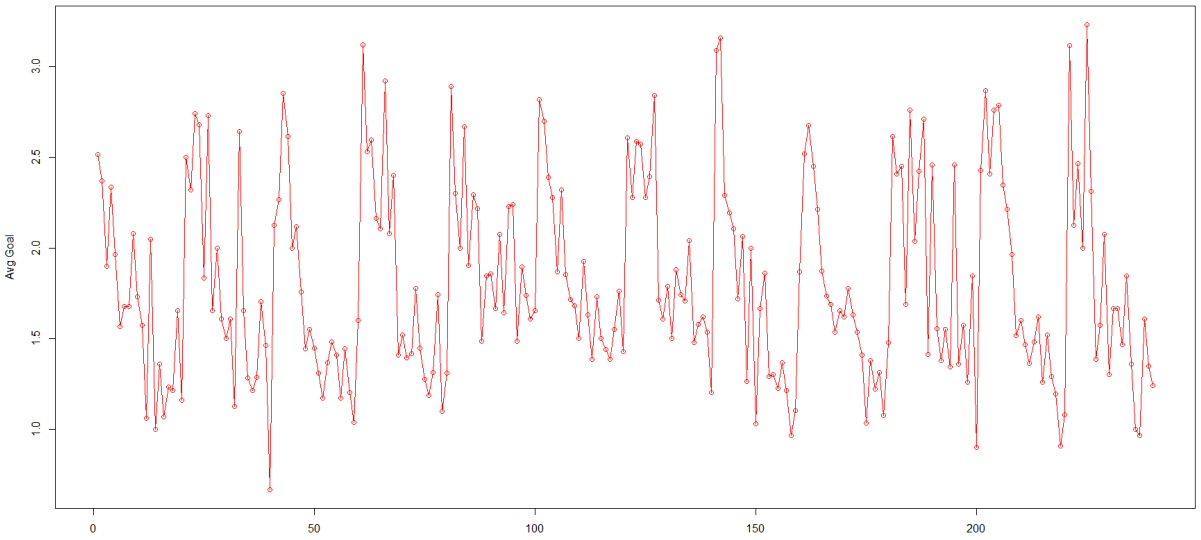
```
dotchart(data$avg_goal, pch = 21, bg = "green", pt.cex = 1.5)
```



As can be seen in dot plot graph, the values are concentrated between 1.0-2.0 but in general our dataset is very scattered.

Line Graphic

```
plot(x=data$avg_goal,type = "o",col="red", ylab="Avg Goal")
```



According to line graph, we can say that our data is quite scattered.

For DATA 2

Summary Statistic Table

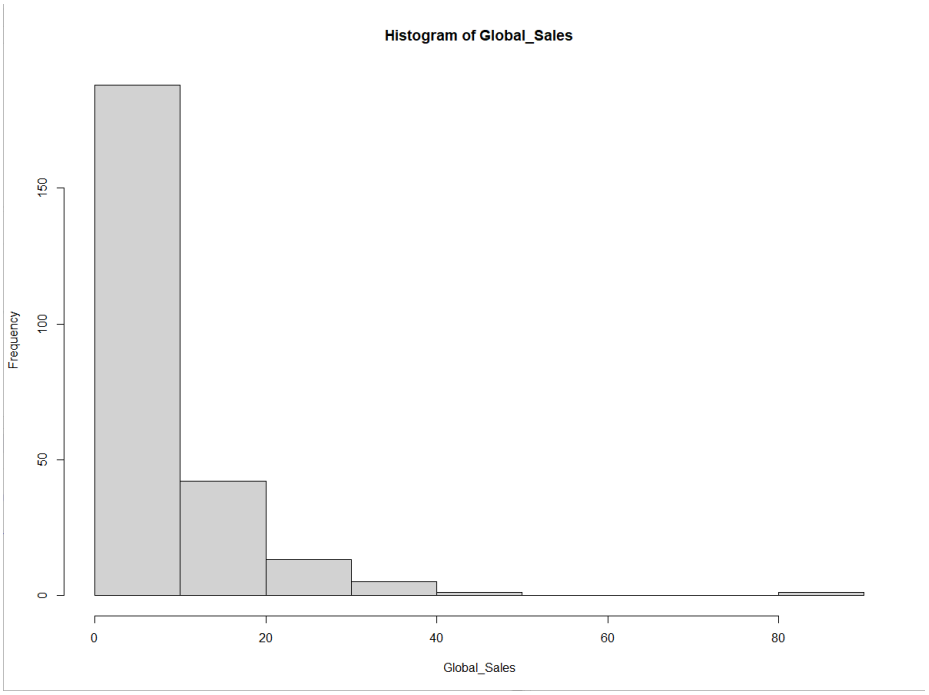
Mean	Mode	Median	Standart Deviation	Variance	Coefficient of Variation	Skewness	Kurtosis
9.17616	5.13	6.59	7.73	59.5	84.06%	4.68535714	34.69431556

According to the summary statistic table, the kurtosis value of data2 is 34.69 and according to this value we can say that the curve is steeper than the normal distribution.

It is right-skewed because mean value is greater than median value.

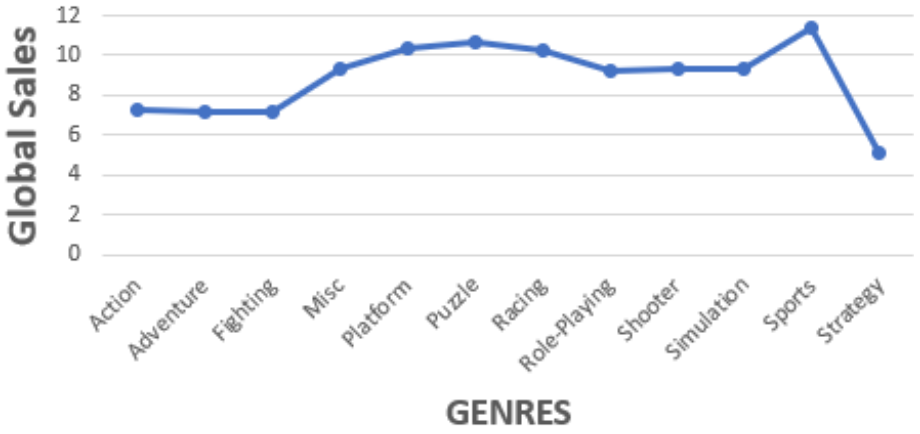
Histogram Graph

```
hist(data2$Global_Sales,main="Histogram of Global_Sales",xlab="Global_Sales")
```



Values were collected mostly between 0-20. When we look at the histogram graph,we can say the distribution is similar to the normal distribution.

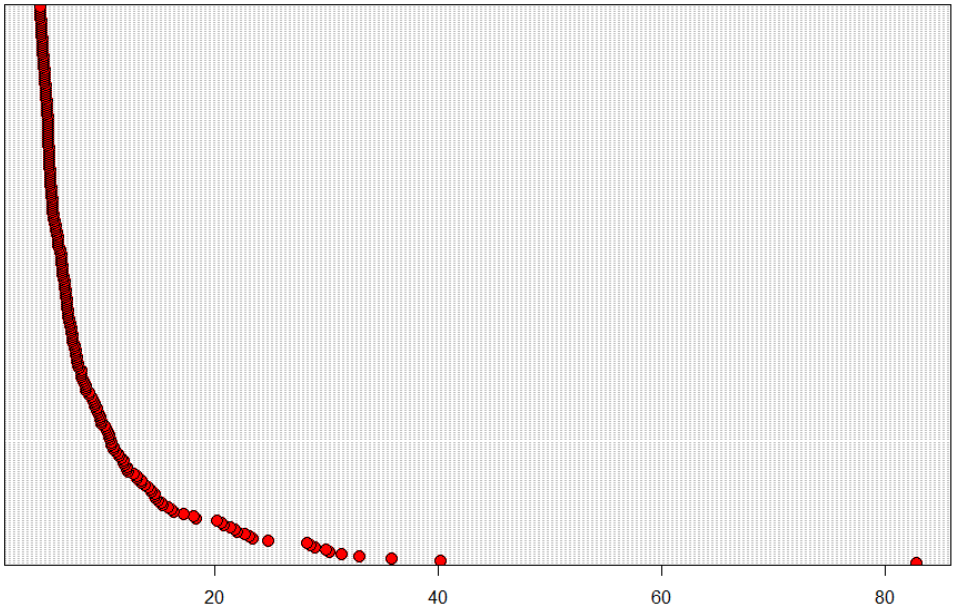
Polygon Graphic



According to polygon graph, the highest global sales was in the “Sports” genres. The lowest average number of goals was in the “Strategy” genres.

Dot Plot Graph

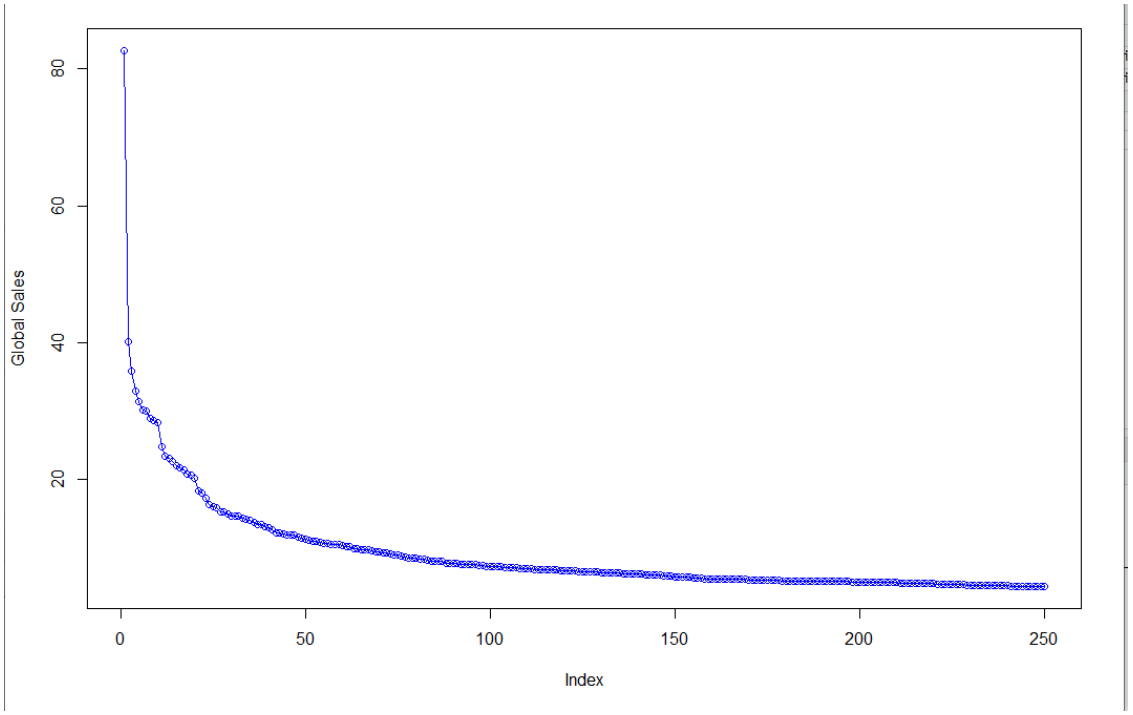
```
dotchart(data2$Global_Sales,pch = 21, bg = "red", pt.cex = 1.5)
```



As can be seen in dot plot graph, the values are concentrated between 0-20.

Line Graphic

```
plot(x=data2$Global_Sales,type = "o",col="blue",ylab="Global Sales")
```



According to line graph, we can say that our data is quite accumulated.

Question 2

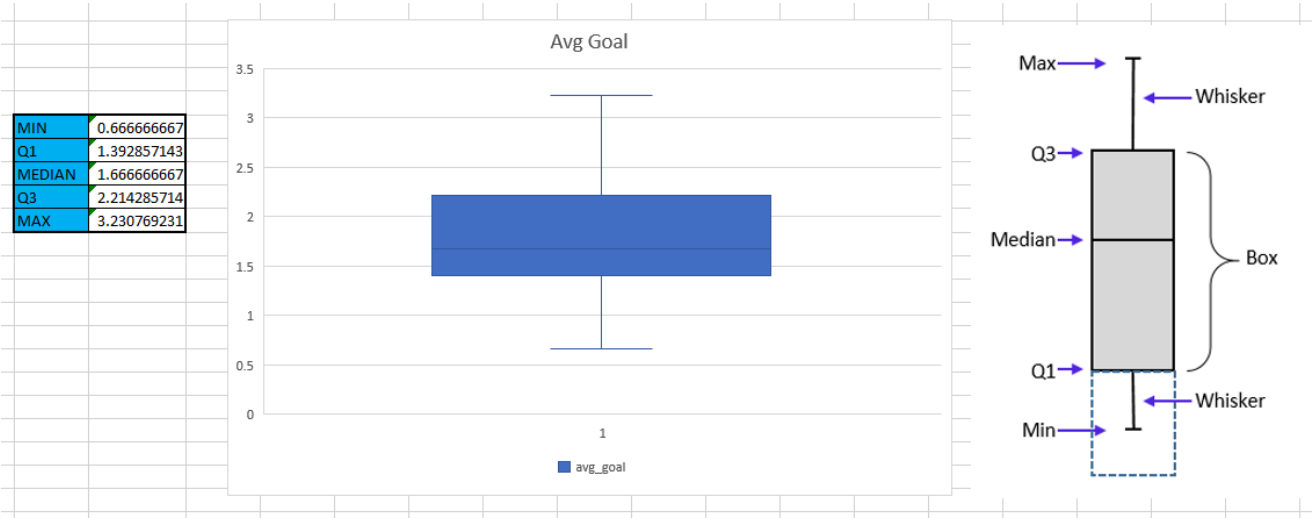
For DATA 1

Stem-and-Leaf Graph

```
stem(data$avg_goal)

The decimal point is 1 digit(s) to the left of the |

 6 | 7
 8 | 0177
10 | 00334678800367799
12 | 00111233466688999001111456666778888899
14 | 11112344455667888888000022233455567777
16 | 00111111222334555667777888990111233344 4
18 | 35555667778000367
20 | 000003456778800233 69
22 | 112347888990122357 99
24 | 011235566702237 99
26 | 12247880134669
28 | 245792
30 | 9226
32 | 3
```

According to stem-and-leave graph, the values in our dataset were at most 1.6, at least 0.6 and 3.2.

Box Plot (and box-and-whisker) Graph



According to box plot chart, its shows us “avg_goal” column approaching to the normal distribution. Datas are more concentrated in the upper part. It shows positive skewness and right-skewed. There is no outliers in the “avg_goal” column as seen in the box plot graph.

For DATA 2

Stem-and-Leaf Graph

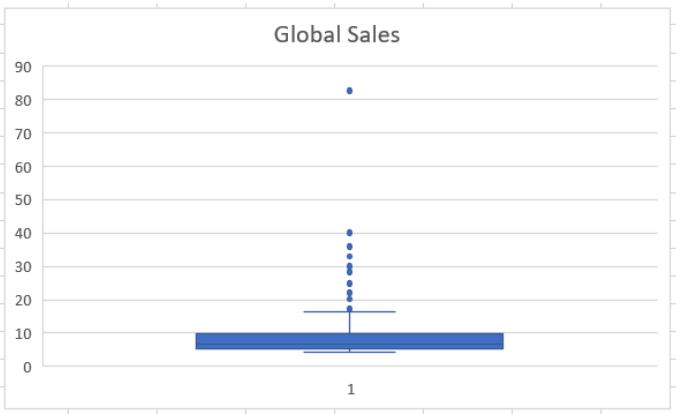
```
stem(data2$Global_Sales)
```

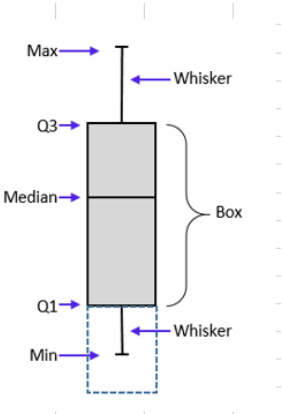
```
The decimal point is 1 digit(s) to the right of the |

0 | 4444444444
0 | 55555555555555555555555555555555555555555555555555555555555555555555555555555+91
1 | 000000000001111111112222222333344444
1 | 555555666788
2 | 011122333
2 | 5899
3 | 0013
3 | 6
4 | 0
4 |
5 |
5 |
6 |
6 |
7 |
7 |
8 | 3
```

According to stem-and-leave graph, the values in our dataset were at most 5.5.

Box Plot (and box-and-whisker) Graph

MIN	4.39
Q1	5.155
MEDIAN	6.59
Q3	9.8775
MAX	82.74



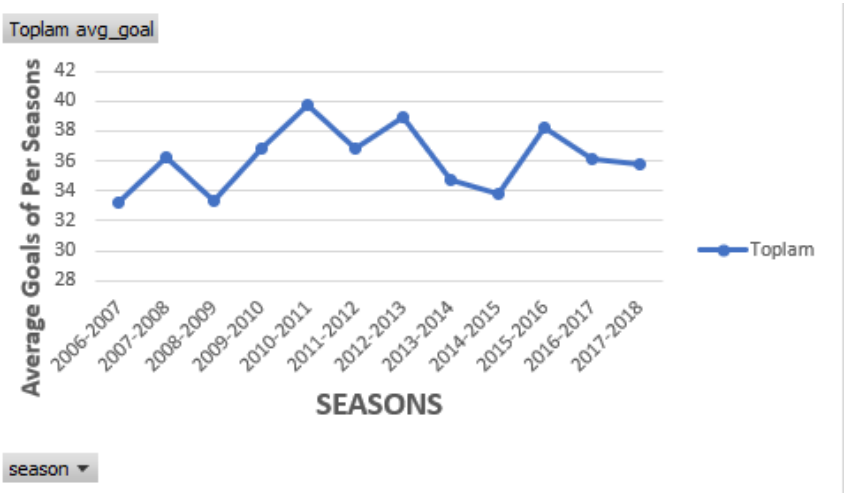


According to box plot chart, its shows us “Global_Sales” column deviating to the normal distribution. Datas are more concentrated in the lower part. It shows positive skewness and right-skewed. There are so many outliers in the “Global_Sales” column as seen in the box plot graph.

Question 3

Pivot Table For DATA 1

Satır Etiketleri	Toplam avg_goal
2006-2007	33.17990323
2007-2008	36.20773499
2008-2009	33.3685174
2009-2010	36.78937036
2010-2011	39.68797255
2011-2012	36.84531483
2012-2013	38.93158775
2013-2014	34.77588363
2014-2015	33.80684639
2015-2016	38.1841404
2016-2017	36.06544646
2017-2018	35.74922157
Genel Toplam	**433.5919395**



Total “avg_goal” numbers for every season are given in the table above. On the other hand, a line graph of these values is drawn on the graph.

Pivot Table For DATA 2

Satır Etiketleri	Say Global_Sales
Action	51
Adventure	3
Fighting	8
Misc	20
Platform	34
Puzzle	7
Racing	20
Role-Playing	34
Shooter	38
Simulation	8
Sports	24
Strategy	3
Genel Toplam	**250**

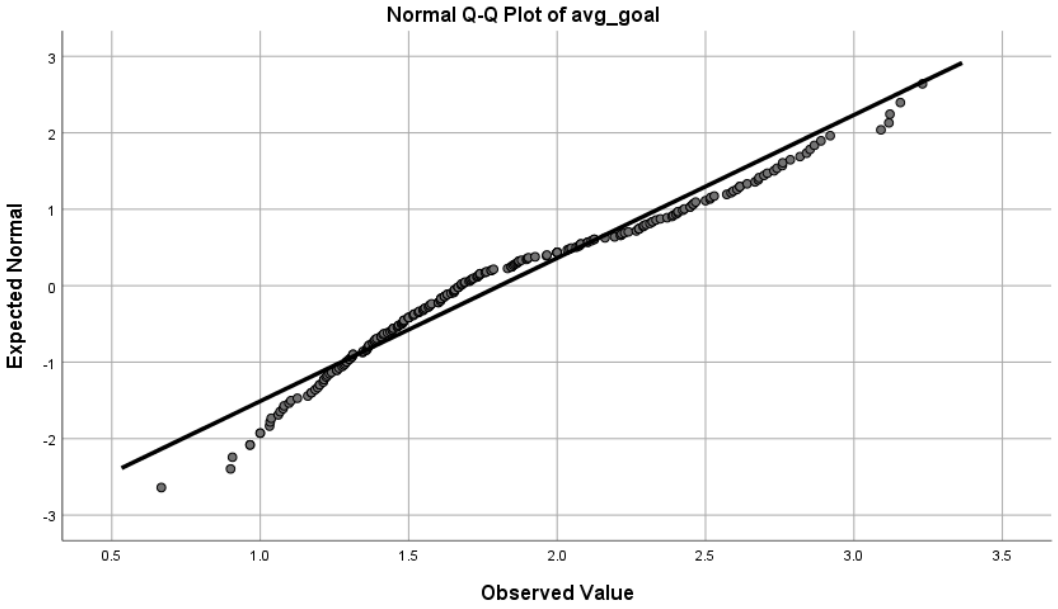
“Global_Sales” numbers for every categories are given in the table above. On the other hand, a line graph of these values is drawn on the graph.

Question 4

For DATA 1

Normal Probability Graph

```
qqnorm(data$avg_goal)
qqline(data$avg_goal)
```



Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
avg_goal	.116	240	.000	.960	240	.000

a. Lilliefors Significance Correction

When we look at the graph, the straight line shows us the normal distribution. The values in data set we used converge to normal as they approach the line. Apart from the extreme values, we can say that it has generally approached the normal distribution.

H0: Data is not normally distributed.

H1: Data is normally distributed.

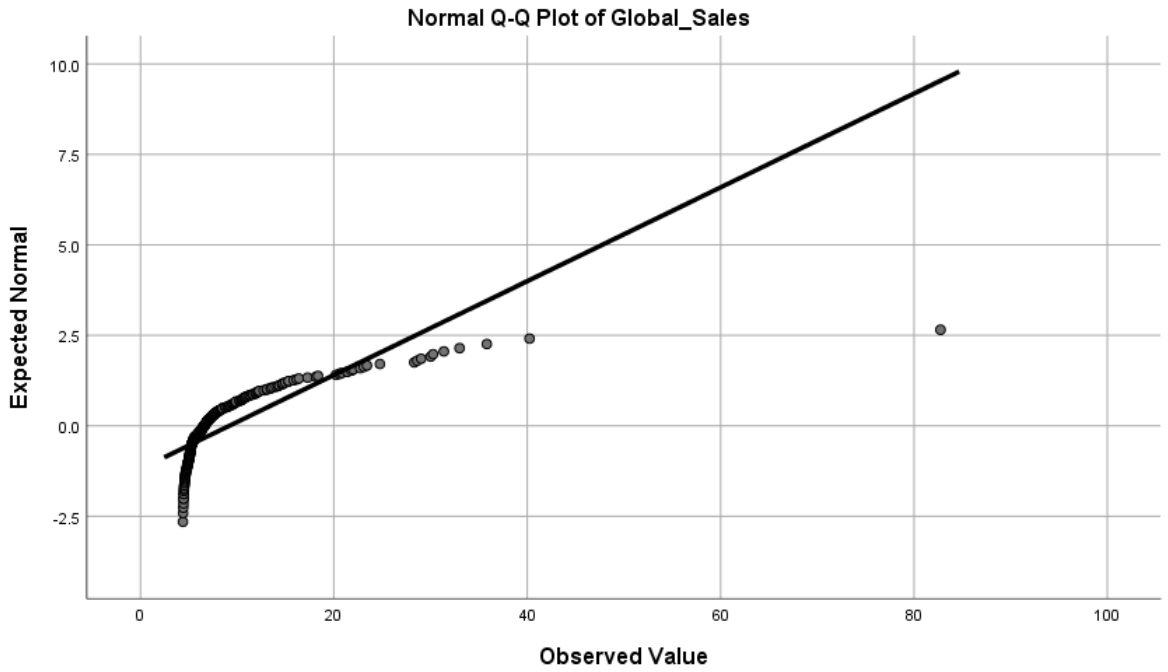
The significance value is greater than the alpha value (we'll use .05 as our alpha value), then there is no reason to think that our data differs significantly from a normal distribution. - i.e., we can reject the null hypothesis that it is non-normal.

As you can see above, both tests give a significance value that's greater than .05, therefore, we can be confident that our data is normally distributed.

For DATA 2

Normal Probability Graph

```
qqnorm(data2$Global_Sales)
qqline(data2$Global_Sales)
```



Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Global_Sales	.267	250	.000	.568	250	.000

a. Lilliefors Significance Correction

When we look at the graph, the straight line shows us the normal distribution. The values in data set we used converge to normal as they approach the line. Apart from the extreme values, we can say that it has generally approached the normal distribution.

H0: Data is not normally distributed.

H1: Data is normally distributed.

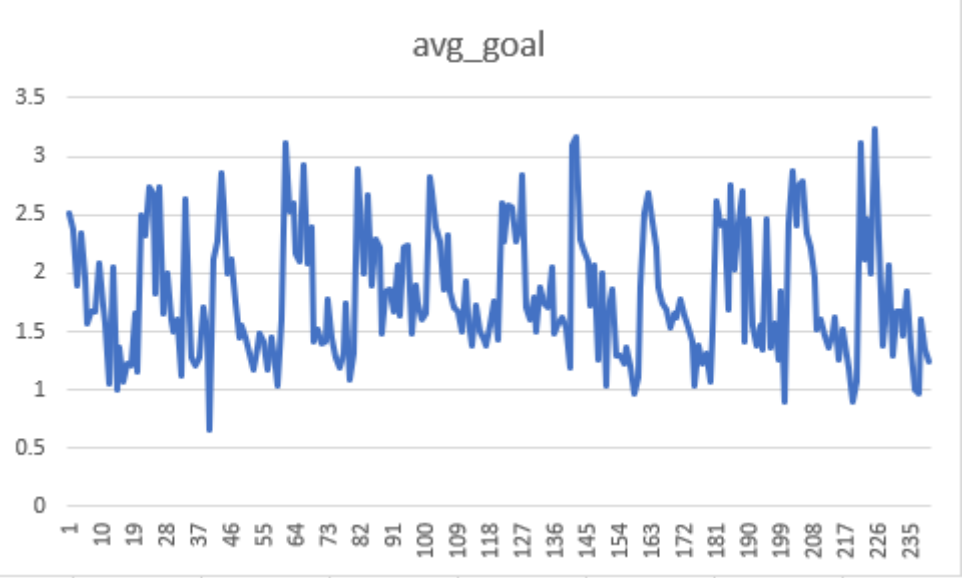
The significance value is greater than the alpha value (we'll use .05 as our alpha value), then there is no reason to think that our data differs significantly from a normal distribution. - i.e., we can reject the null hypothesis that it is non-normal.

As you can see above, both tests give a significance value that's greater than .05, therefore, we can be confident that our data is normally distributed.

Question 5

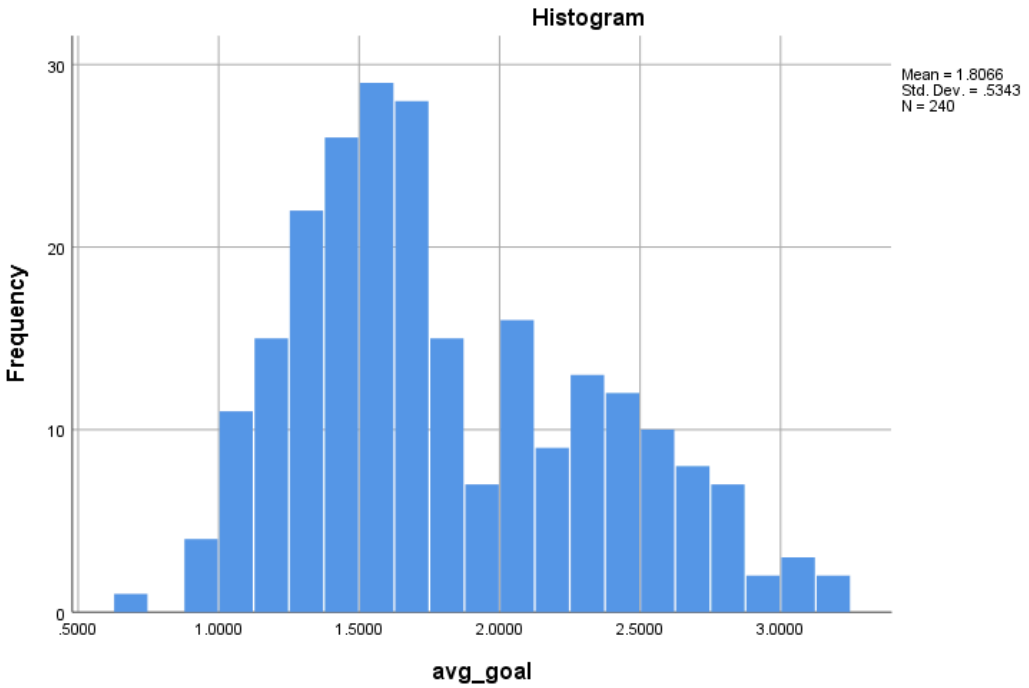
For DATA 1

1) Line Graph



In the line chart above, the average goal amounts of the premier league teams are given. The number of goals on the Y axis and the teams on the X axis are given. In this chart, we can say that the average goal per game varies between 1-3, without season limitation.

2) Histogram Graph



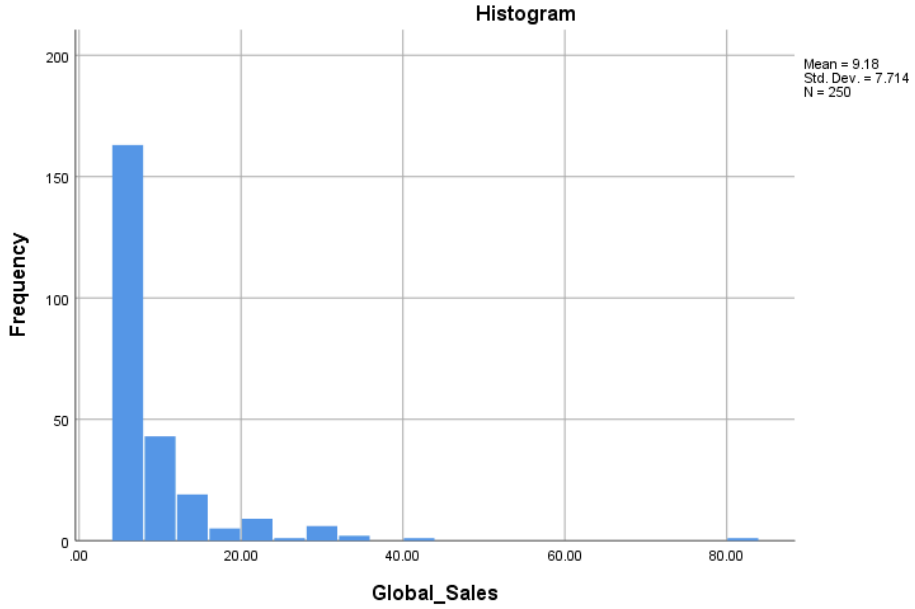
There are two graphics which is Line and Histogram graph. As seen in the two graph, there is not much deviation in our data. As we found in the previous questions, we can say that our data shows approaching the normal distribution.

For DATA 2

1) Line Graph

The line chart above shows the worldwide sales rank of video games. The number of sales is given on the Y axis, and the variety of games is given on the X axis. In this chart, game sales are ordered from most to least.

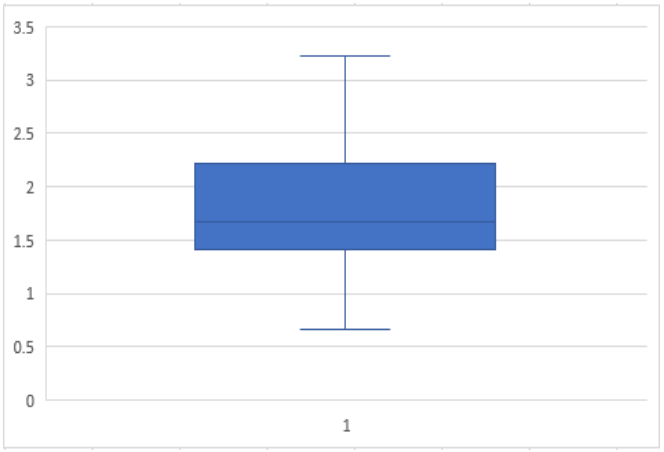
2) Histogram Graph

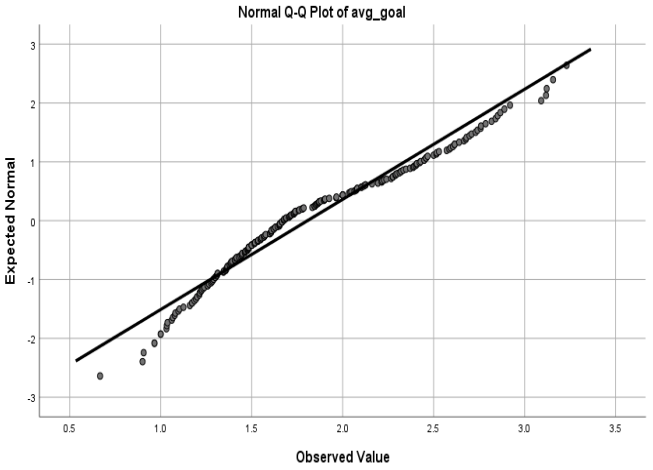


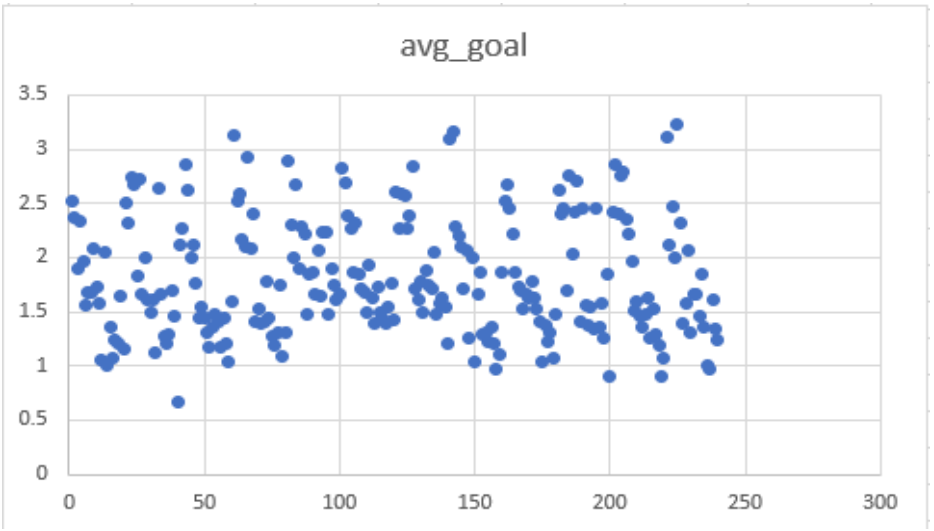
There are two graphics which is Line and Histogram graph. As seen in the two graph, there is not much deviation in our data. As we found in the previous questions, we can say that our data shows approaching the normal distribution.

Question 6

For DATA 1

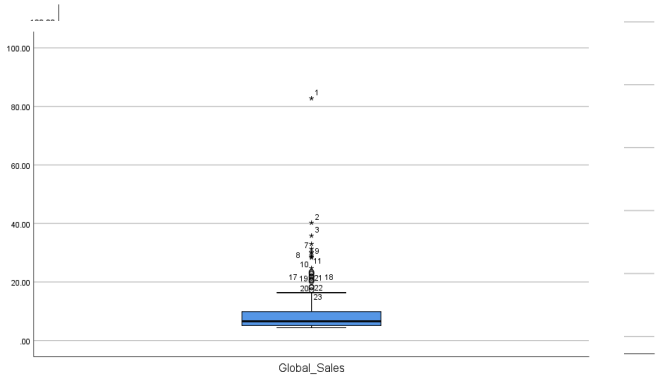


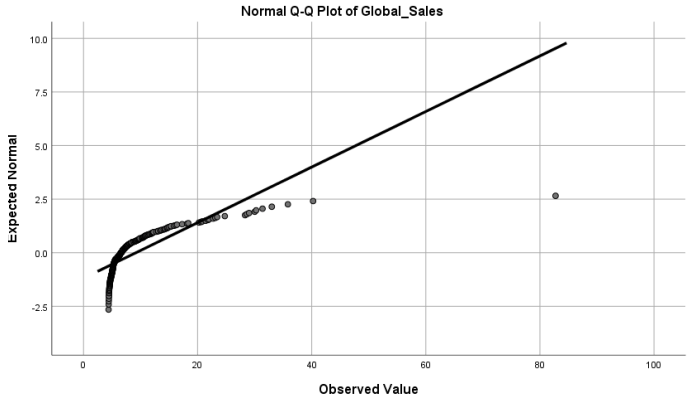


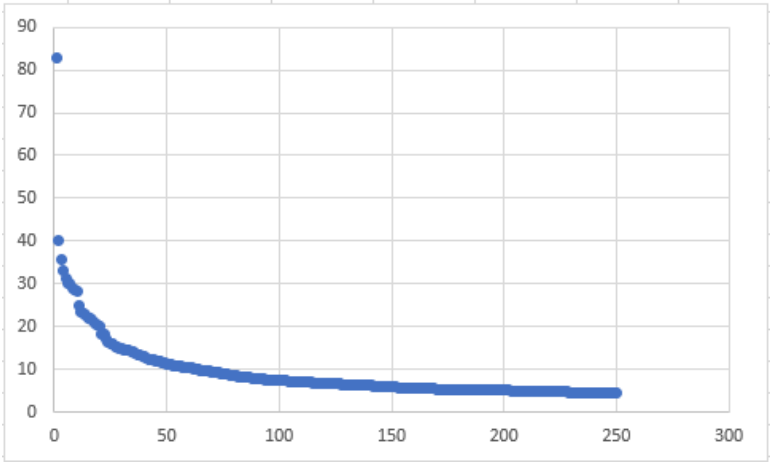


We draw 3 different graphs which are scatter plot, box plot, normal probability graph for looking are there any outliers in our data. According to the graphs there are no outliers in our data.

For DATA 2







We draw 3 different graphs which are scatter plot, box plot, normal probability graph for looking are there any outliers in our data. According to the graphs there are many outliers in our data.

Question 7

Triangle (Ternary) Graph

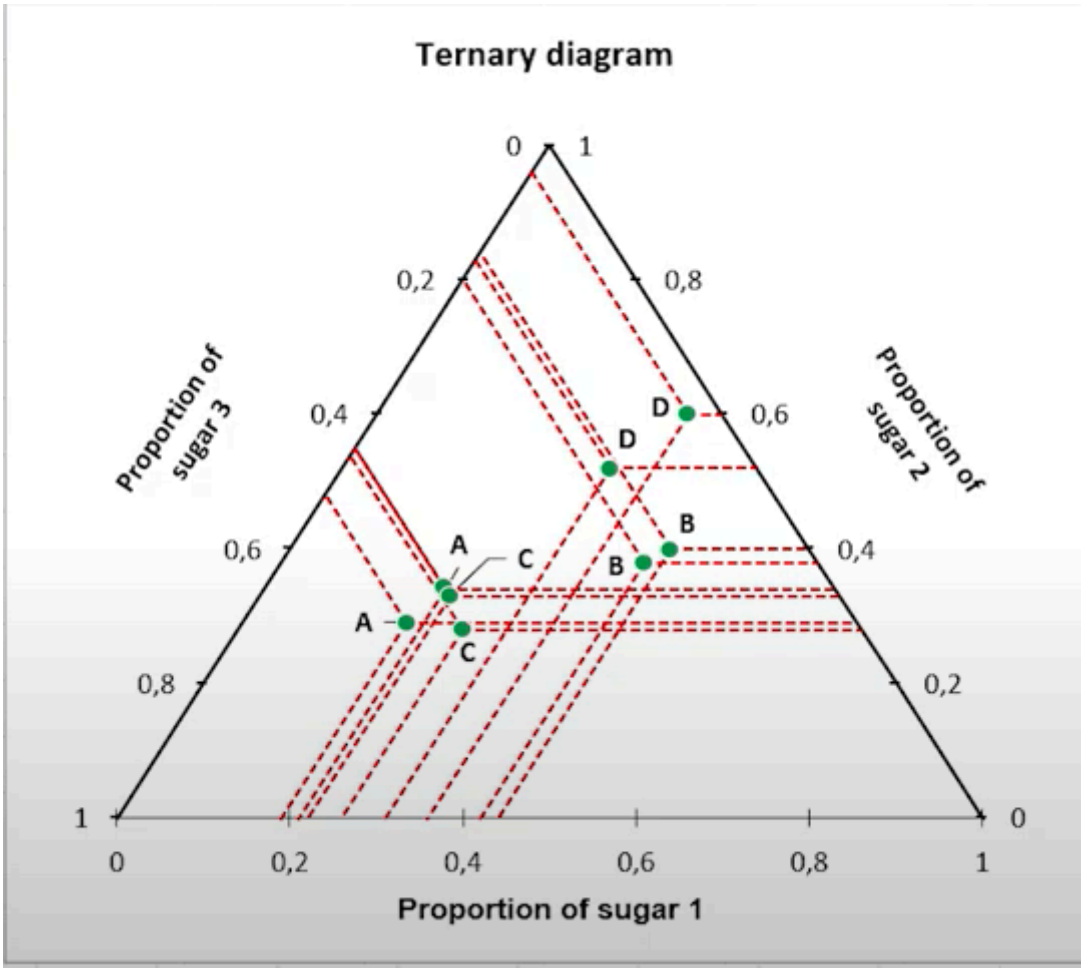
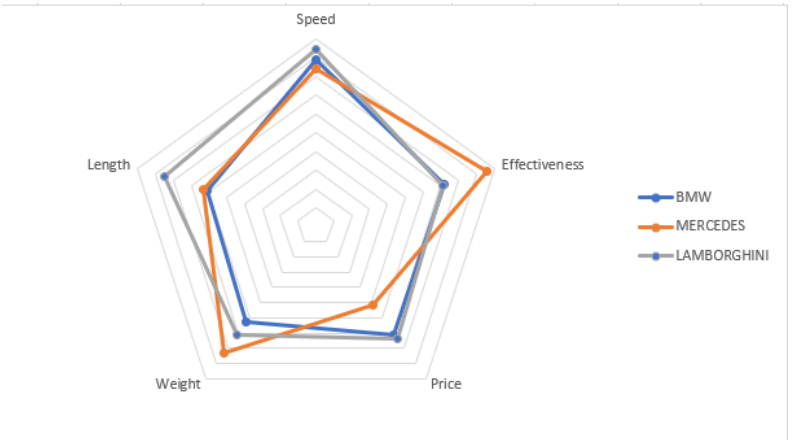


Diagram sites represent three variables. In this case our sugar concentrations. Points represent our fruits to know one's composition, simply follow the projection lines linking them to each variable. Take this one for example. Is proportion of sugar two is around? Oh point 6 proportion of sugar. Three is close to zero, and proportion of sugar one is under oh point 4. Globally speaking, varieties A and C seems to have a similar composition given their proximity on diagram. Varieties B and D, on the other hand, each constitute one relatively independent group.

Question 8

Star Graph



According to star graph, most effective,cheapest and heaviest car is Mercedes. The longest car is Lamborghini. The speed order is sorted by Lamborghini, BMW and Mercedes.

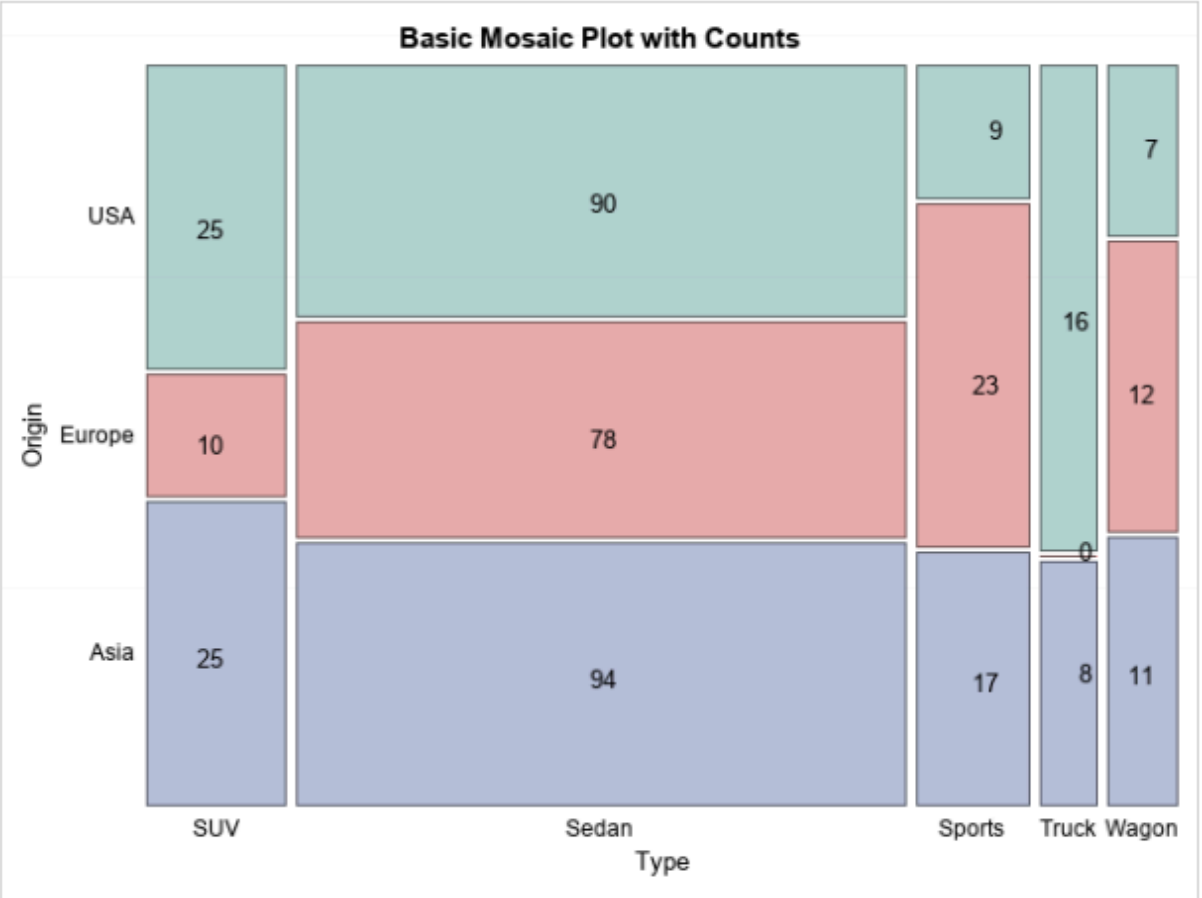
Question 9

Parallel Coordinates Plot

Parallel coordinate graphs were drawn about a night spent is each night a tourist of each countries. There is a negative correlation between July 2020 and September 2020. We can say that there is a positive relationship between April 2020 and June 2020.

Question 10

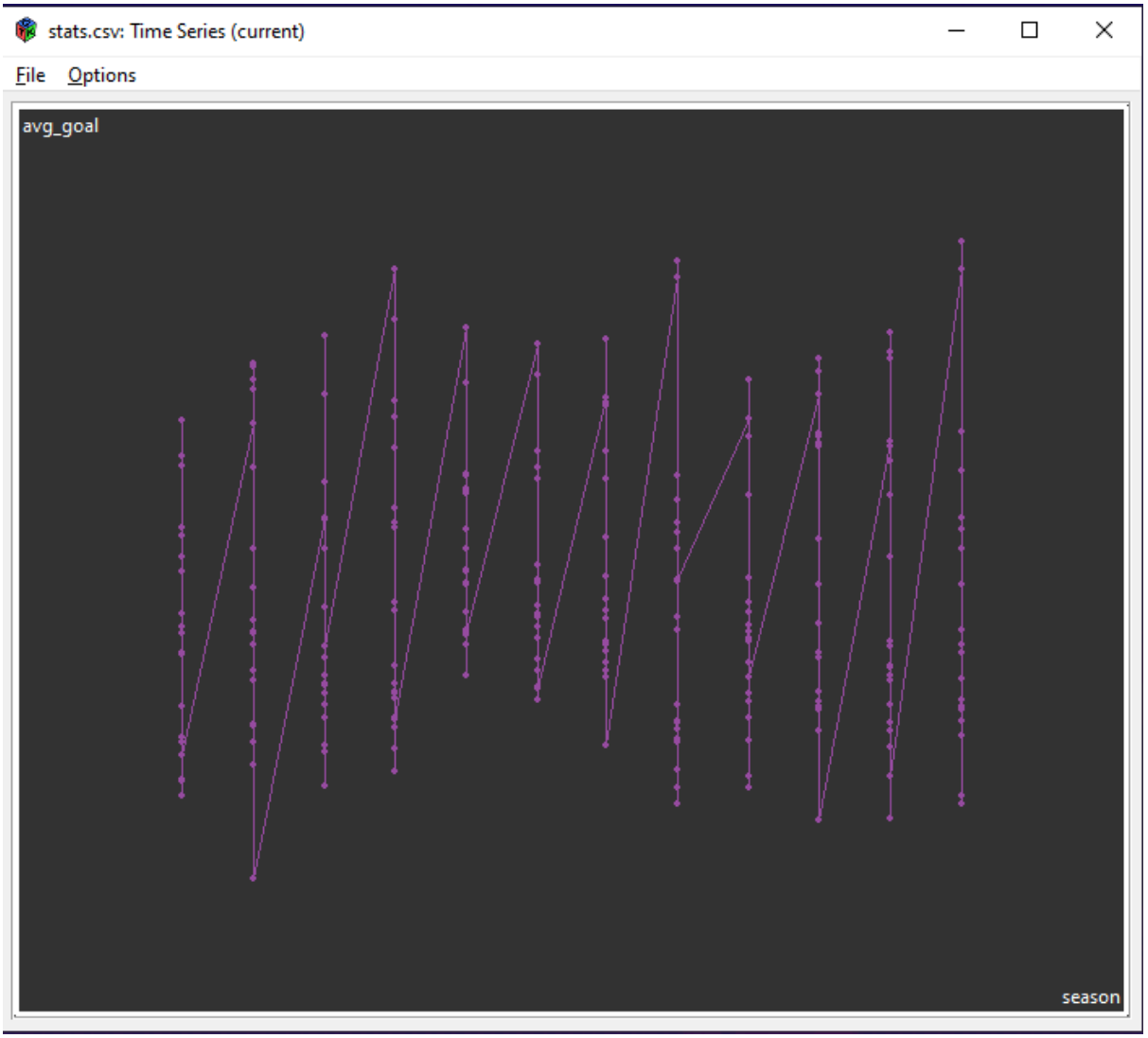
Mosaic Plot



We can see 3 types of cars which are SUV,Sedan,Sports,Truck,Wagon. Also we can see 3 continents. We can easily see Sedan cars have more drivers than the others. Also USA people more interested Truck cars than Europe and Asia.

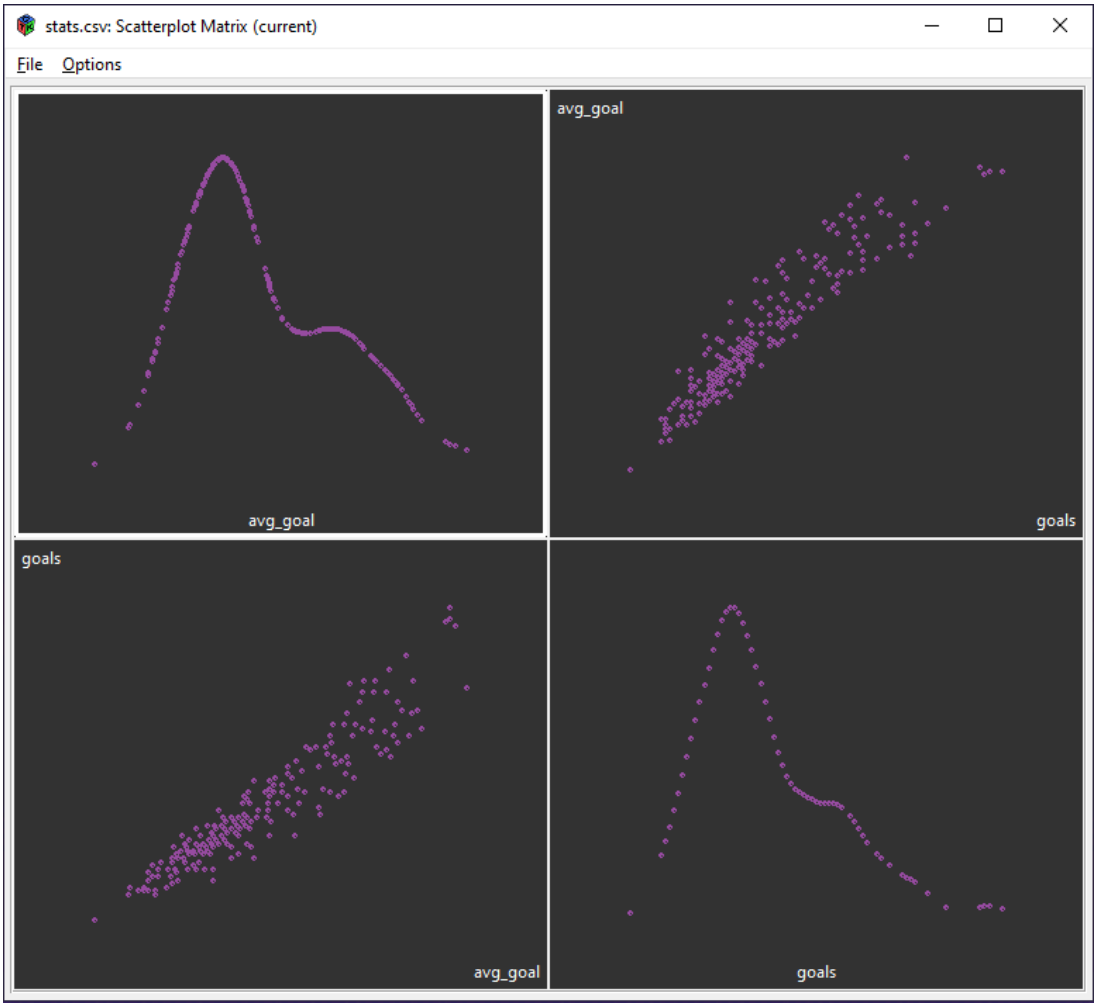
Question 11

Time Series



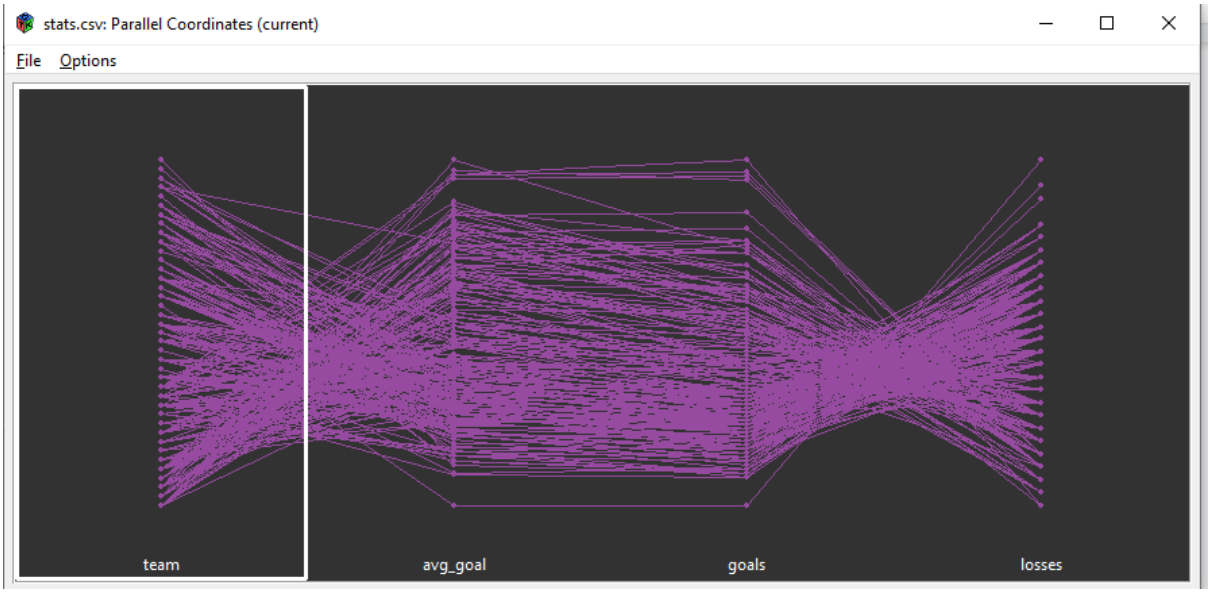
The average number of goals scored per game in the premier league is given in the chart above. When we look at the 4th vertical line here (2009-2010 season), we can say that the average goal score of the teams is high. The average number of goals scored by a single team is the Liverpool team on the last vertical line (2017-2018 season). The minimum average number of goals scored by a single team is Derby County team on the 2nd straight line (2007-2008 season).

Scatter Plot



When we look at top left and bottom right graphs, we can say that these graphs are approaching to the normal distribution. When we look at other two graphs, we can say data don't deviate too much and have regular structure.

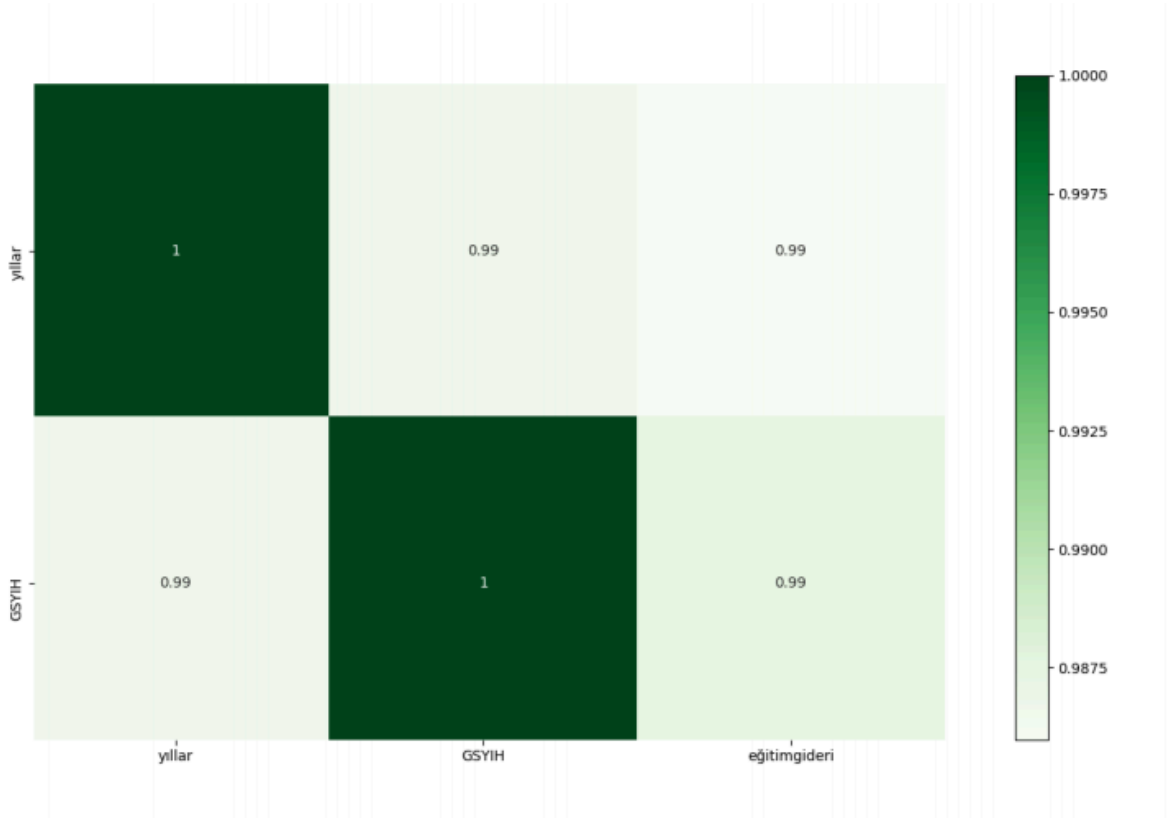
Parallel Coordinates Plot



When we look at the graph there is a negative correlation between team between avg_goal and goals between losses. We can say that there is a positive relationship between avg_goal and goals.

Question 12

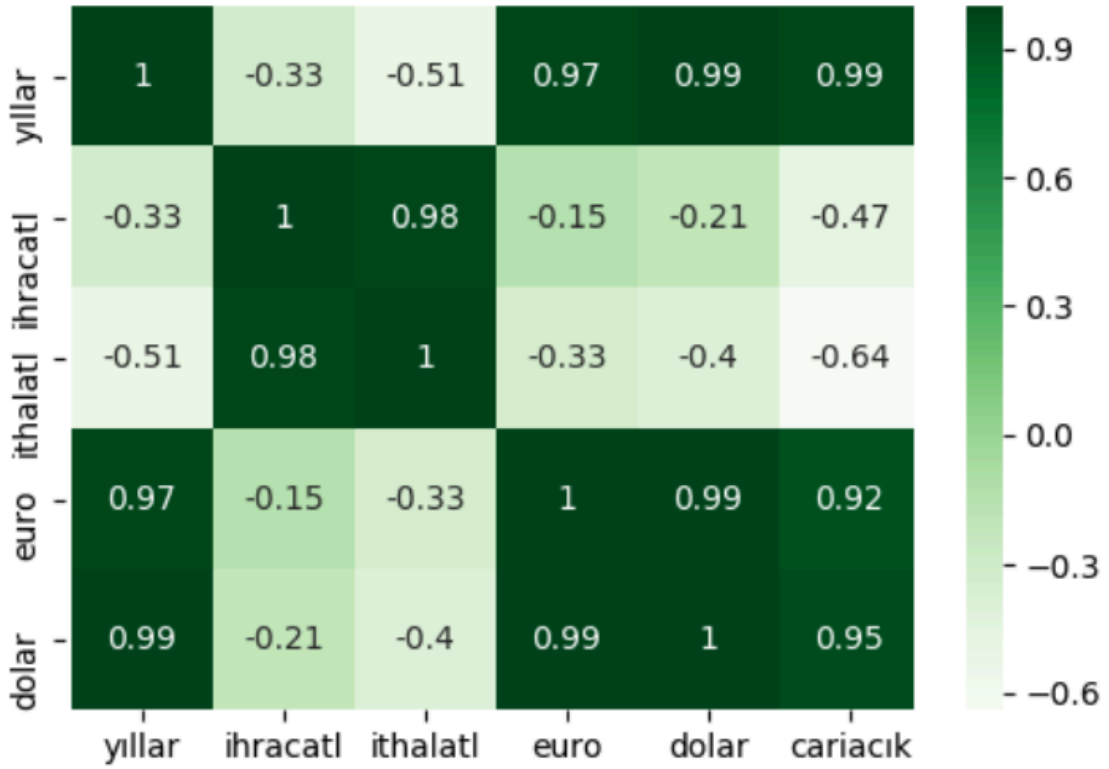
e) There is a small correlation between per capita GDP and education expenditure per capita.



We chose high correlation between per capita GDP and education expenditure per capita data for Turkey. So we used heatmap graph.

We can see for last 5 years there is a high correlation between GDP and education expenditure.

g) As the current account deficit of Turkey changes, the dollar exchange rate changes. There is negative correlation between them.

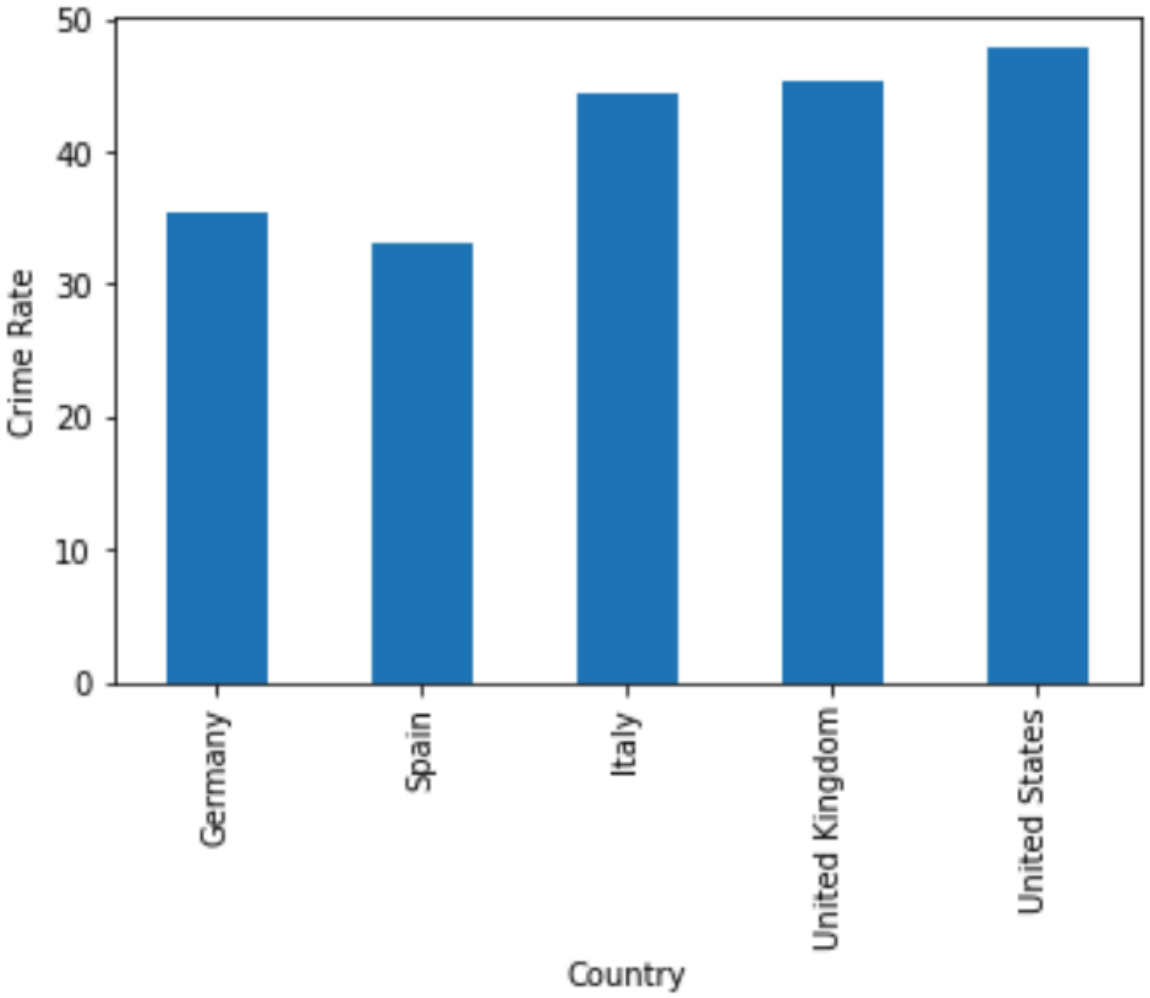


We chose the data which is current account deficit of Turkey's correlation between exchange rate changes. So we can used heatmap graph again.

Dark tones are positive correlation,light tones are negative correlation in this graph.

So, we can see if current account deficit of Turkey changes, exchange rate also varies. This is positive a change.

d) Crimes (or crime rate) in Germany, England, Italy and Spain are increasing more slowly than the US crime.



We can see Spain has lowest crime rate. United States has highest crime rate.

If we want to compare countries by crime rates. It is United States>United Kingdom>Italy>Germany>Spain.

As you can see crime rate in Germany, England, Italy and Spain are increasing more slowly than the US crime.

REFERENCES

- https://www.numbeo.com/crime/rankings_by_country.jsp?title=2021&displayColumn=0
- https://www.kaggle.com/zaeemnalla/premier-league?select=stats.csv
- https://www.kaggle.com/andrewsundberg/college-basketball-dataset?select=cbb21.csv
- https://www.kaggle.com/gregorut/videogamesales
- http://ggobi.org/docs/parallel-coordinates.html
- https://en.wikipedia.org/wiki/Mosaic_plot#:~:text=A%20mosaic%20plot%20(also%20known,information%20for%20only%20one%20variable.
- https://en.wikipedia.org/wiki/Normal_probability_plot
- https://ceaksan.com/tr/pivot-table-nedir
- https://www.youtube.com/watch?v=mRherx56vp8
- https://www.youtube.com/watch?v=d2V5P_KhNLc
- https://www.youtube.com/watch?v=CDIvHSuCpmg
- https://www.youtube.com/watch?v=cTI3rPLP1cE
- https://www.xlstat.com/en/
- https://www.youtube.com/watch?v=K74_FNnIIF8