Statistical Computing
Data Analysis Report

May 21,2018

Department of Computer Engineering

College of Engineering,Mugla Sıtkı Kocman University

Mugla,TR 48000

Abstract

The purpose of the research is to investigate the reasons behind the influences of individuals between the ages of 18 and 30 in social media. Since the purpose of the study was to examine the participants' overviews of the subject, no individual evaluation was made and no information about identity was asked. With this research, the activity rates of the individuals whose ideas were taken at the specified age range in social media were estimated and the analysis results were documented.

Note:Question numbers starting from 1 to 15 and question2 as twitter,facebook and instagram groupped.Colnames are given as soru1,soru2, ….

Table of Contents

1. Abstract..... 2

2. Description of the problem..... 4

3. Description of the data..... 4

3.1 Dataset information 4

3.2 Attribute information 4

4. Progress to date 5

4.1 Dataset Manipulation..... 5

4.2 Social Media Spending Time vs Education Levels..... 5

4.3 Created Content vs Education Levels 6

4.4 Created Content vs Personal Mood..... 7

4.5 Social Media Account Numbers for Sleep Time as Personal Mood..... 7

4.6 Social Media Account Numbers for Sleep Time as Feeling Alone or not..... 8

4.7 Social Media Account Numbers for Sleep Time with Number of Sleep Time as Type of Usage Social Media Tool …..8

5. Hypothesis Testing 9

5.1 One Sample T-Test..... 9

5.2 Two Sample T-Test 10

6. Anova..... 11

7. Conclusion..... 12

8. References 12

2.Description of the Problem

People lives affected by many social medya contents and people spending so much time with social medya tools.I think this cause many problems such as sleep,focus,happiness,etc. I want to show to affect results with my collected and documented data.

3.Description of the Data

3.1 Data Set Information

This data collected from between 18-30 age people from Mugla. This data approach is the effects of social medya on between 18-30 age people at Mugla.This data aim to get information about social media affects on young people at Mugla.

3.2 Attribute Information

Person Indıvıdual Information

1-)Age-Person Age(numeric:from 18-30)

2-)Sex-Person Sex(String:”Man”,Woman”)

3-)Job-Person Job(String)

4-)Education Level-Person education(String:”Primary School”,”University”,”High School”)

5-)Place of Birth-Birth place of person(String)

Questions About the Search’s Subject

6-) What kind of content do you like from social media types?(String)

7-) Have you spent time in social media tools such as facebook, instagram and twitter?-Facebook(String)

8-) Have you spent time in social media tools such as facebook, instagram and twitter?-Twitter(String)

9-) Have you spent time in social media tools such as facebook, instagram and twitter?-Instagram(String)

10-) Is social media safe?(Strng:”Yes”,”No”)

11-) How many hours a day do you spend on social media?(Numeric)

12-) What is the number of content you share on a daily basis in social media?(Numeric)

13-) Before you use social media effectively, how much time did you spend on your daily basis for yourself(type in hours)?(Numeric)

14-) After you use social media effectively, how much time did you spend on your daily basis for yourself(type in hours)?(Numeric)

15-) How do you define your mood?- (String:“Aggressive”, “Happy”,” Sad”, “Unstable”)

16-) How many hours do you sleep per day?(Numeric)

17-) Do you think you feel alone?(String:”Yes”,”No”)

18-) Which tool do you use the most as a social media tool? (String:”TV”,” Phone”,” Tablet”)

19-) How many social media accounts do you have (with number)(Numeric)

20-) Are you a stressful person? (String:”Yes”,”No”)

21-) Before you start using social media tools, write down the time you have dedicated your work or your homework.(Numeric)

22-) After you start using social media tools, write down the time you have dedicated your work or your homework.(Numeric)

4.Progress to Date

4.1 Data Manipulation

Data was collected via face to face with people.There were some unecessary attributes inside of te data.I cleaned up and changed with the right attribute.Some values were missing so I filled up them with mean function.

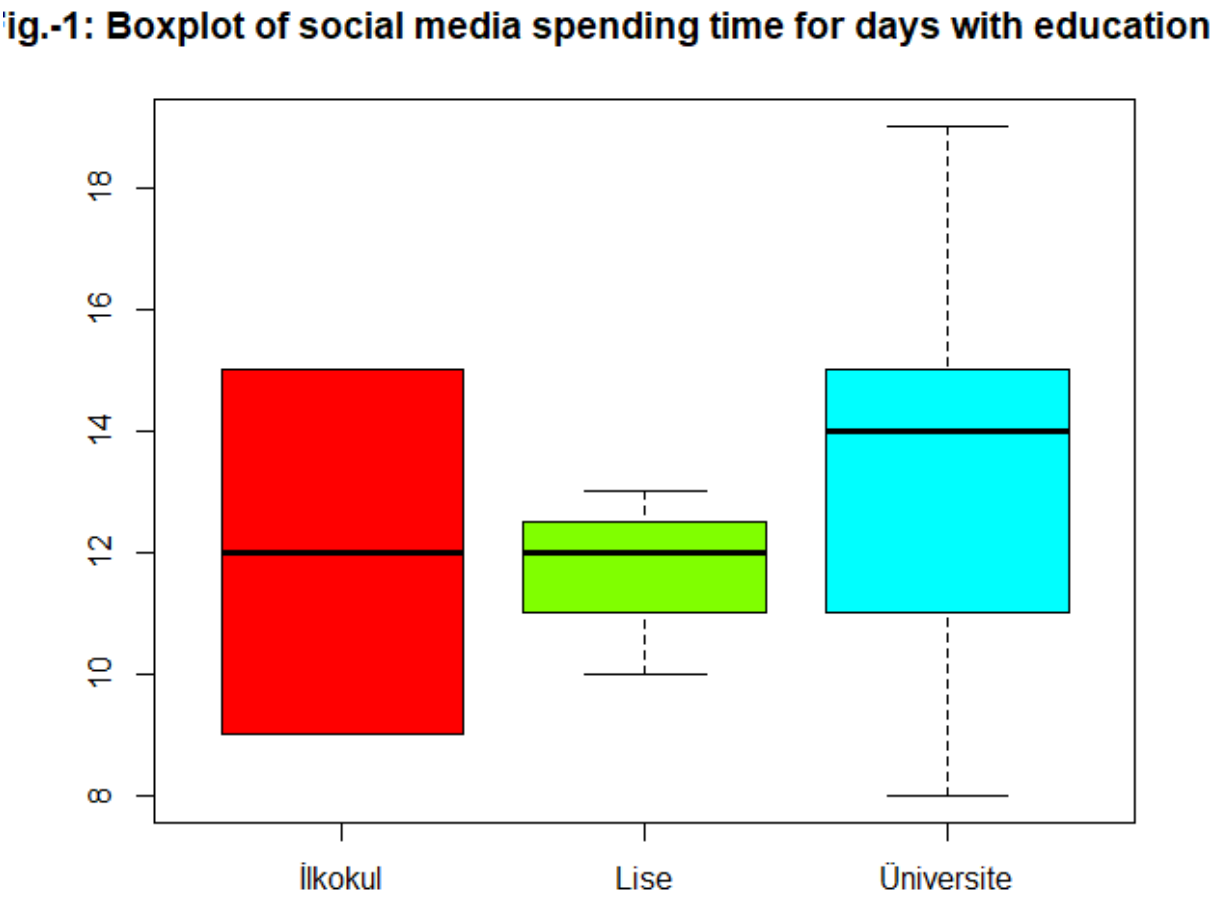
Example:

MyData <- read.csv(file="C:/Users/Moonster/Desktop/18-30 YAŞ ARASINDAKİ BİREYLERİN SOSYAL MEDYA KULLANIMI VE ETKİLENME SEVİYELERİ ANKET SONUÇLARI.csv", header=TRUE, sep=",")

(MyData$Soru6, na.r=TRUE) <- mean(MyData$Soru6, na.rm=TRUE)

4.2 Social Media Spending Time vs Education Levels

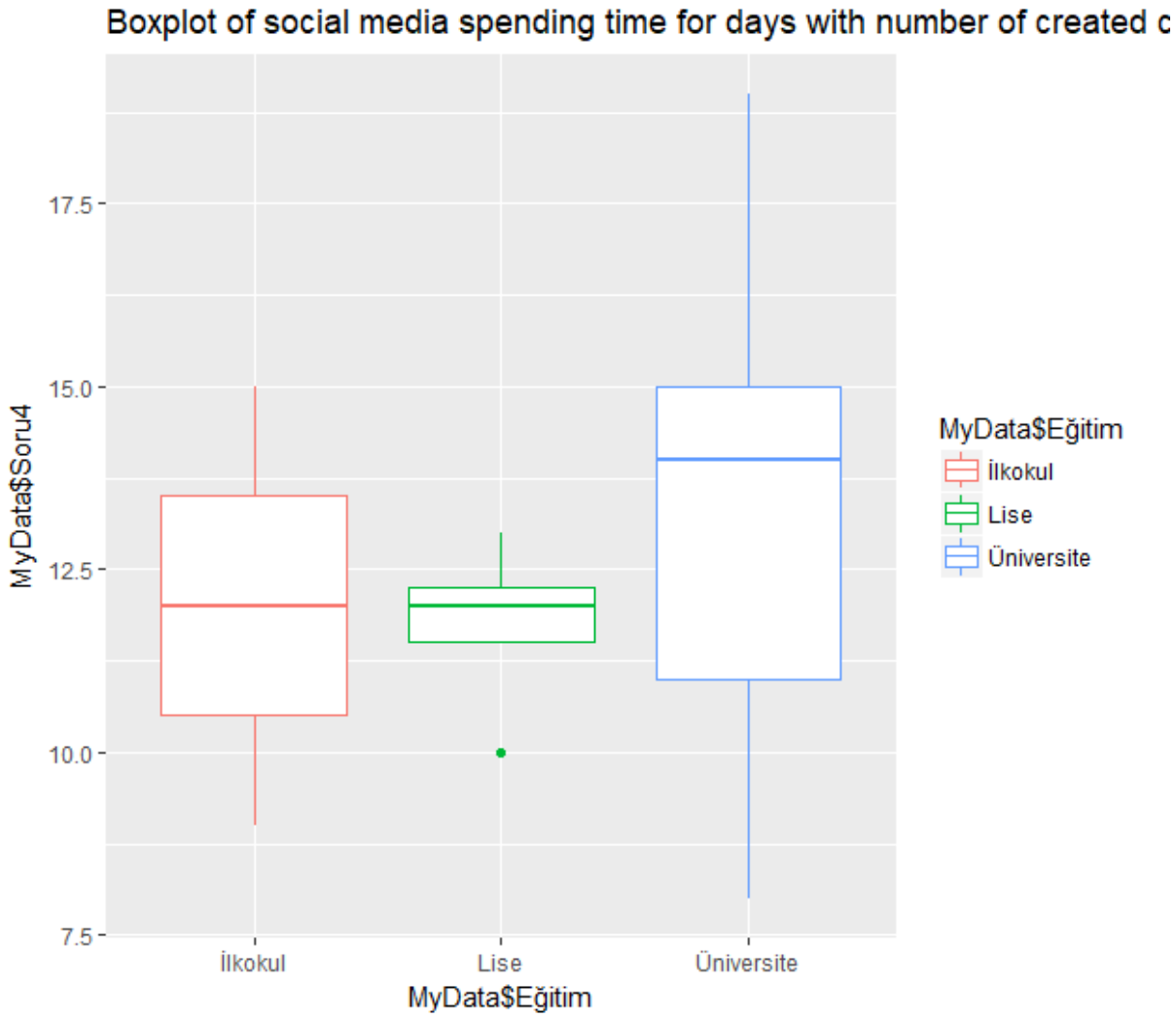
boxplot(MyData$Soru4~MyData$Eğitim, main="Fig.-1: Boxplot of social media spending time for days with education level", col= rainbow(4))



As you can see people Who have education level University and Primary School are spending more time on social media.

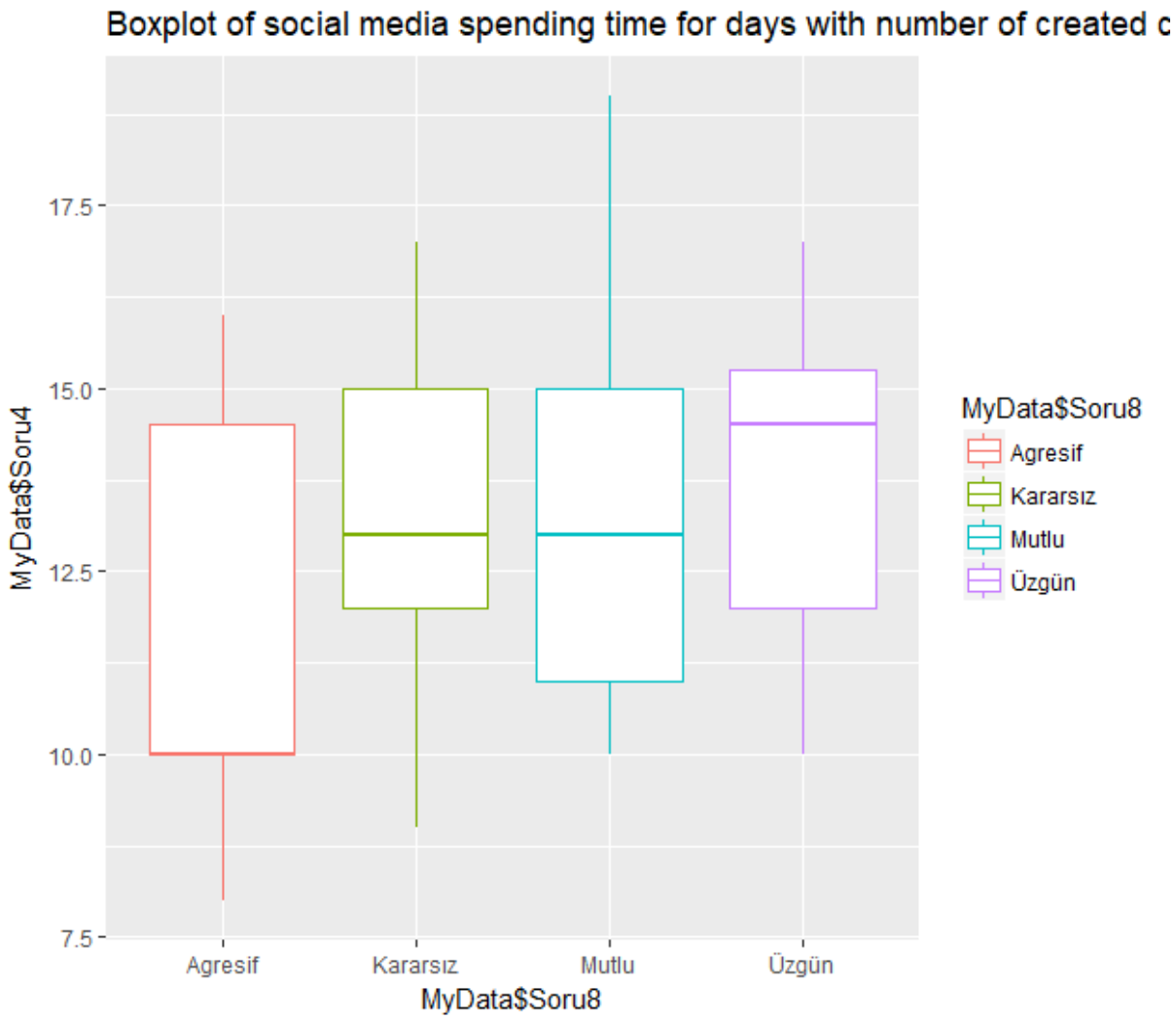
4.3 Created Content vs Education Levels

ggplot(MyData$soru5, aes(MyData$Eğitim,MyData$Soru4))+geom_boxplot(aes(col=MyData$Eğitim))+labs(title="Boxplot of social media spending time for days with number of created content as education level")



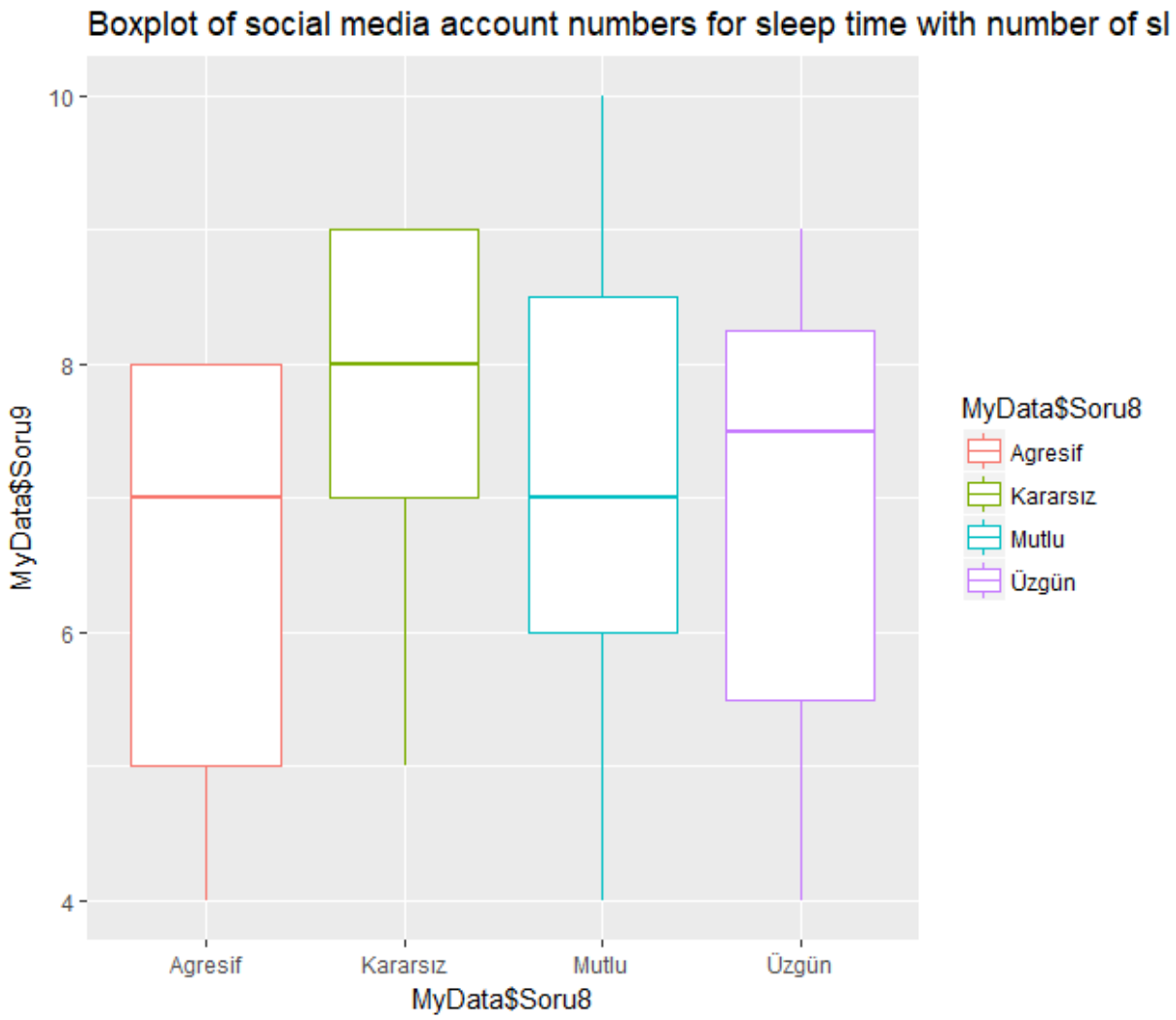
4.4 Created Content vs Personal Mood

ggplot(MyData$soru5, aes(MyData$Soru8,MyData$Soru4))+geom_boxplot(aes(col=MyData$Eğitim))+labs(title="Boxplot of social media spending time for days with number of created content as personal mood")



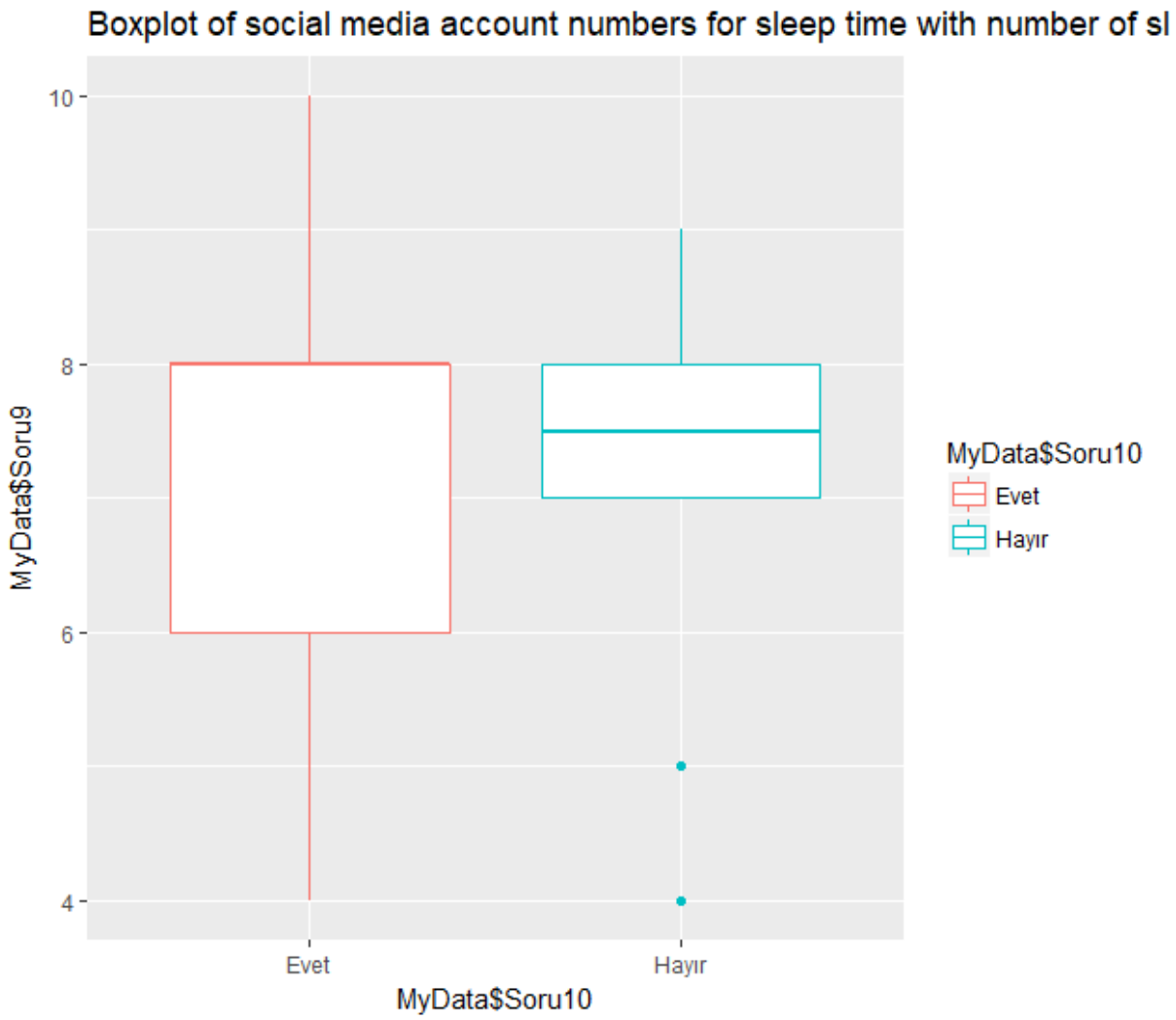
4.5 Social media account numbers for sleep time with number of sleep time as personal mood

ggplot(MyData$soru12, aes(MyData$Soru8,MyData$Soru9))+geom_boxplot(aes(col=MyData$Soru8))+labs(title="Boxplot of social media account numbers for sleep time with number of sleep time as personal mood")



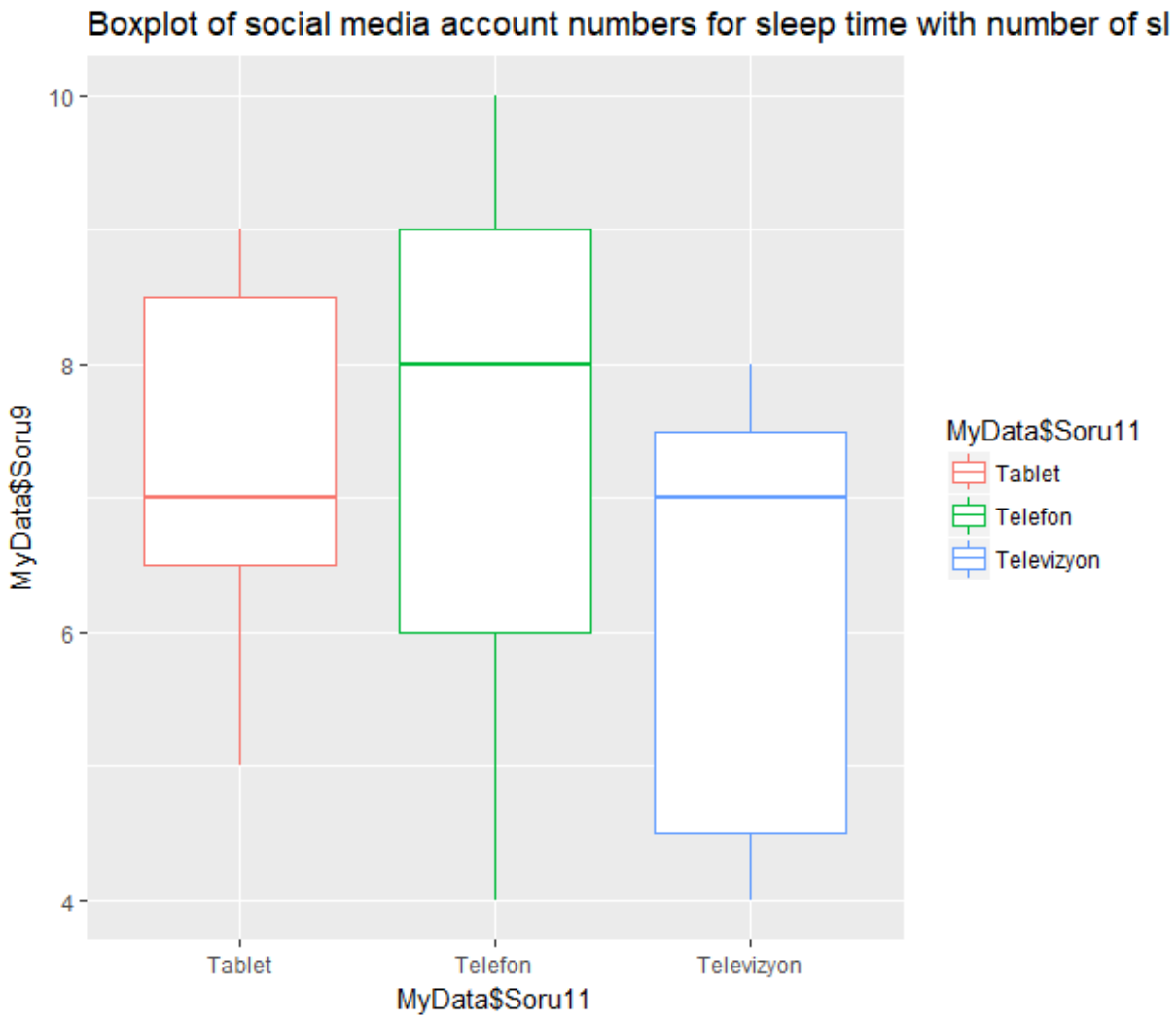
4.6 Social media account numbers for sleep time with number of sleep time as feeling alone or not

ggplot(MyData$soru12, aes(MyData$Soru10,MyData$Soru9))+geom_boxplot(aes(col=MyData$Soru10))+labs(title="Boxplot of social media account numbers for sleep time with number of sleep time as feeling alone or not")



4.7 Social media account numbers for sleep time with number of sleep time as type of usage social media tools

ggplot(MyData$soru12, aes(MyData$Soru11,MyData$Soru9))+geom_boxplot(aes(col=MyData$Soru11))+labs(title="Boxplot of social media account numbers for sleep time with number of sleep time as type of usage social medya tools")



5.Hypothesis Testing

a. One Sample T-Test

Let assume that the mean of shared content in social media greater than 10 when person mood is Happy.

1-)Write the hypothesis:

Ho: = 10

Ha: >10

2-)Test the hypothesis for 95% confidence level:

```
> person_have_happy_mood<-subset(MyData,MyData$Soru8=="Mutlu")
There were 32 warnings (use warnings() to see them)
> warning()
Warning message:

> result<-c(person_have_happy_mood$Soru5)
> t.test(result, alternative = "greater", mu=10)

        One Sample t-test

data:  result
t = 5.4268, df = 14, p-value = 4.46e-05
alternative hypothesis: true mean is greater than 10
95 percent confidence interval:
 31.02872      Inf
sample estimates:
mean of x
 41.13333

>
```

We see the p value =4.46 .We accept the the main hypothesis.So the people has happy mood sharing more than 10 content on social media.

5.2 Two Sample T-Test

1-)Write down a hypothesis if educating level affects the sharing content in social media.

Ho:M1<=M2

Ha:M1=M2

2-)Write down a hypothesis if education level effects sharing content in social media negatively.

Ho:M1<=M2

Ha:M1>=M2

3-) Test the Hypothesis for 95% confidence interval.

```
> st<-subset(MyData,Eğitim=="Üniversite")
There were 24 warnings (use warnings() to see them)
> View(st)
> edu<- (st$Soru5)
> st2<-subset(MyData, Eğitim=="Lise")
> edu2<-(st2$Soru5)
> #Normality Test
> shapiro.test(st)
Error in shapiro.test(st) : is.numeric(x) is not TRUE
> shapiro.test(as.numeric(st))
Error in is.numeric(x) : (list) object cannot be coerced to type 'double'
> View(st)
> shapiro.test(edu)

        Shapiro-Wilk normality test

data:  edu
W = 0.94849, p-value = 0.1202

> shapiro.test(edu2)

        Shapiro-Wilk normality test

data:  edu2
W = 0.834, p-value = 0.1785

> #Variance Test
> var.test
function (x, ...)
UseMethod("var.test")
<bytecode: 0x000000000d932ab8>
<environment: namespace:stats>
> var.test(edu,edu2)

        F test to compare two variances

data:  edu and edu2
F = 2.2153, num df = 32, denom df = 3, p-value = 0.5639
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1574512 7.8804150
sample estimates:
ratio of variances
          2.215269

> #Testing the hypothesis for 95% confidence interval
> t.test(edu,edu2,var.equal = TRUE, alternative = "less")

        Two Sample t-test

data:  edu and edu2
t = 1.9829, df = 35, p-value = 0.9724
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 40.50664
sample estimates:
mean of x mean of y
 49.12121  27.25000

>
```

According the p value its normaly distrubuted

We can see that variances are not equal.

P value 0.9724 is greater than 0.05, don’t reject Ho.It means that educating level affect the sharing content in social media negatively.When the educating level is university student sharing content number is high. When the educating level is High School ,shared content number is less.

6. ANOVA

#Estimation of model

model1<- aov(MyData$Soru4~MyData$Eğitim)

summary(model1)

TukeyHSD(model1, conf.level = 0.99)

plot(TukeyHSD(model1, conf.level = 0.99),las=1, col = "red")

library(gplots)

plotmeans(MyData$Soru4~MyData$Eğitim, main="Fig.-3: Mean Plot with 95% Confidence Interval", ylab = "Social Media Spending Time(Hour)", xlab = "Created Content's Education Level")

```
library(ggplot2)

ggplot(MyData$soru5, aes(MyData$Eğitim,MyData$Soru4))+geom_boxplot(aes(col=MyData$Eğitim))+labs(title="Boxplot of social media spending time for days

boxplot.stats(MyData$Soru4[MyData$Eğitim=="Üniversite"])

#Estimation of model

model1<- aov(MyData$Soru4~MyData$Eğitim)

summary(model1)

TukeyHSD(model1, conf.level = 0.99)

plot(TukeyHSD(model1, conf.level = 0.99),las=1, col = "red")

library(gplots)
plotmeans(MyData$Soru4~MyData$Eğitim, main="Fig.-3: Mean Plot with 95% Confidence Interval", ylab = "Social Media Spending Time(Hour)", xlab = "Create

#Diagnostic Checking

par(mfrow=c(2,2))
plot(model1)

uhat <-resid(model1)

shapiro.test(uhat)

bartlett.test(MyData$Soru4~MyData$Eğitim)

library(car)
leveneTest(MyData$Soru4~MyData$Eğitim)
```

```
> #Estimation of model
>
> model1<- aov(MyData$Soru4~MyData$Eğitim)
There were 12 warnings (use warnings() to see them)
>
> summary(model1)
             Df Sum Sq Mean Sq F value Pr(>F)
MyData$Eğitim  2  12.45   6.224    0.86  0.432
Residuals     36 260.63   7.240
>
> TukeyHSD(model1, conf.level = 0.99)
  Tukey multiple comparisons of means
    99% family-wise confidence level

Fit: aov(formula = MyData$Soru4 ~ MyData$Eğitim)

$`MyData$Eğitim`
                     diff       lwr      upr     p adj
Lise-İlkokul       -0.250000 -7.493402 6.993402 0.9936752
Üniversite-İlkokul  1.393939 -4.696856 7.484734 0.7583667
Üniversite-Lise     1.643939 -2.784246 6.072125 0.4878847

>
```

Wee can see last 2 group are strongly same.but there is a difference between first group and other groups.

7. CONCLUSION

In conclusion, people social medya usage,spending time and affects on person are searched according to some factors.There are many things effect the social media activity.In the beginning, I mentioned about social medya effects on people.If we look at the statistics, we can see that the most effective one is personal mood on social medya content sharing and spending time.I applied hypothesis testing but some tests are not suitable form y data because of the data entity number and normality distrubution.

8. REFERENCES

[3] Lecture Notes, Dr. Eralp DOGU, 2018.

https://rcompanion.org/rcompanion/c_04.html

http://staff.pubhealth.ku.dk/~tag/Teaching/share/R-tutorials/ConfidenceIntervals.html

https://datascienceplus.com/one-way-anova-in-r/