T.C.

Muğla Sıtkı Koçman University

Computer Engineering

Students' Academic Performance

Statistical Computing Course Project

Ayşe Dilek

120709026

Project Supervisor

Asst. Prof. Eralp DOĞU

May 28, 2017

Students’ Academic Performance

Ayşe Dilek
Computer Engineering
Muğla Sıtkı Koçman University
April 15, 2017

Summary

In this project, it is aimed to examine the students’ academic performance according to educational data set which is collected from learning management system.

Table of Contents

1. Abstract

The performances of students in the school can be examined and compared with different features. Participation of students in the course, attendance rate of the course, the closeness of the parents and the interest are the factors that cause the change in success. The academic performance of 480 students was examined according to the data set that available in this project. In this data set, graphs were drawn and according to sex, lecture, nationality, participation in lessons, active in lessons, interest to the parents. These graphs were interpreted as results and the results were obtained.

2. Description of the Problem

The main problem is determined which features effect the students' achievement.

3. Description of the Data

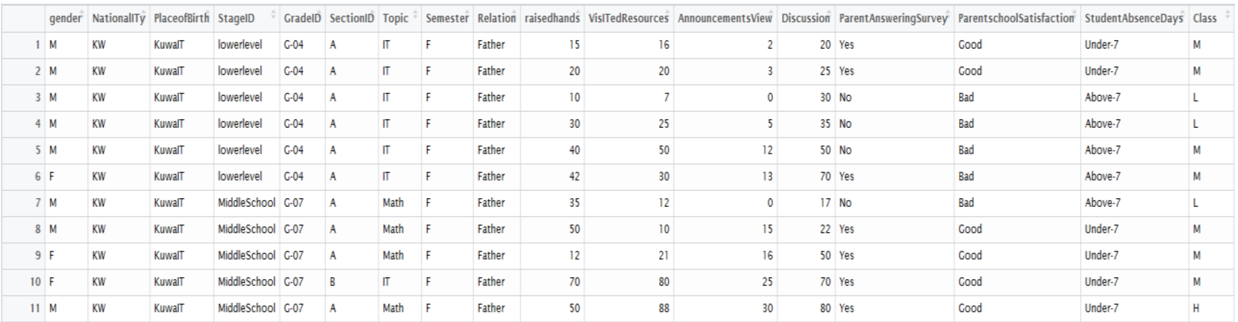
This is an educational data set which is collected from learning management system (LMS) called Kalboard 360. Kalboard 360 is a multi-agent LMS, which has been designed to facilitate learning through the use of leading-edge technology. Such system provides users with a synchronous access to educational resources from any device with Internet connection.

The dataset consists of 305 males and 175 females. The students come from different origins such as 179 students are from Kuwait, 172 students are from Jordan, 28 students from Palestine, 22 students are from Iraq, 17 students from Lebanon, 12 students from Tunis, 11 students from Saudi Arabia, 9 students from Egypt, 7 students from Syria, 6 students from USA, Iran and Libya, 4 students from Morocco and one student from Venezuela.

The dataset is collected through two educational semesters: 245 student records are collected during the first semester and 235 student records are collected during the second semester.

The data set includes also the school attendance feature such as the students are classified into two categories based on their absence days: 191 students exceed 7 absence days and 289 students their absence days under 7.

This dataset includes also a new category of features; this feature is parent parturition in the educational process. Parent participation feature have two sub features: Parent Answering Survey and Parent School Satisfaction. There are 270 of the parents answered survey and 210 are not, 292 of the parents are satisfied from the school and 188 are not.



	gender	NationalITy	PlaceofBirth	StageID	GradeID	SectionID	Topic	Semester	Relation	raisedhands	VisITedResources	AnnouncementsView	Discussion	ParentAnsweringSurvey	ParentschoolSatisfaction	StudentAbsenceDays	Class
1	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Father	15	16	2	20	Yes	Good	Under-7	M
2	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Father	20	20	3	25	Yes	Good	Under-7	M
3	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Father	10	7	0	30	No	Bad	Above-7	L
4	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Father	30	25	5	35	No	Bad	Above-7	L
5	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Father	40	50	12	50	No	Bad	Above-7	M
6	F	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Father	42	30	13	70	Yes	Bad	Above-7	M
7	M	KW	KuwaIT	MiddleSchool	G-07	A	Math	F	Father	35	12	0	17	No	Bad	Above-7	L
8	M	KW	KuwaIT	MiddleSchool	G-07	A	Math	F	Father	50	10	15	22	Yes	Good	Under-7	M
9	F	KW	KuwaIT	MiddleSchool	G-07	A	Math	F	Father	12	21	16	50	Yes	Good	Under-7	M
10	F	KW	KuwaIT	MiddleSchool	G-07	B	IT	F	Father	70	80	25	70	Yes	Good	Under-7	M
11	M	KW	KuwaIT	MiddleSchool	G-07	A	Math	F	Father	50	88	30	80	Yes	Good	Under-7	H

Table 1: Dataset

3.1. Attributes of dataset

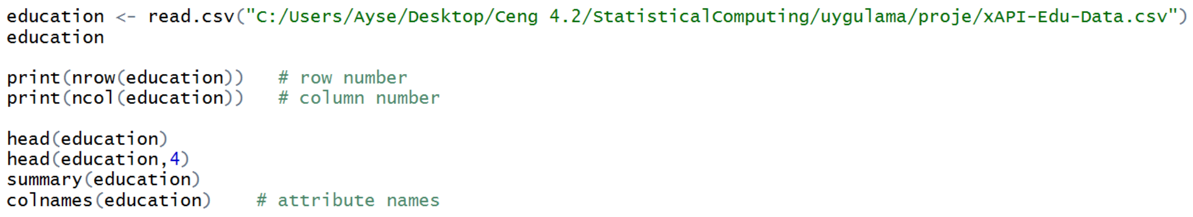
- Gender - student's gender (nominal: 'Male' or 'Female’)
- Nationality - student's nationality (nominal:’ Kuwait’,’ Lebanon’,’ Egypt’,’ SaudiArabia’,’ USA’,’ Jordan’,’ Venezuela’,’ Iran’,’ Tunis’,’ Morocco’,’ Syria’,’ Palestine’,’ Iraq’,’ Lybia’)
- Place of birth - student's Place of birth (nominal:’ Kuwait’,’ Lebanon’,’ Egypt’,’ SaudiArabia’,’ USA’,’ Jordan’,’ Venezuela’,’ Iran’,’ Tunis’,’ Morocco’,’ Syria’,’ Palestine’,’ Iraq’,’ Lybia’)
- Educational Stages - educational level student belongs (nominal: ‘lowerlevel’, ’MiddleSchool’, ’HighSchool’)
- Grade Levels - grade student belongs (nominal: ‘G-01’, ‘G-02’, ‘G-03’, ‘G-04’, ‘G-05’, ‘G-06’, ‘G-07’, ‘G-08’, ‘G-09’, ‘G-10’, ‘G-11’, ‘G-12 ‘)
- Section ID - classroom student belongs (nominal:’A’,’B’,’C’)
- Topic- course topic (nominal:’ English’,’ Spanish’, ‘French’,’ Arabic’,’ IT’,’ Math’,’ Chemistry’, ‘Biology’, ‘Science’,’ History’,’ Quran’,’ Geology’)
- Semester - school year semester (nominal:’ First’,’ Second’)
- Parent responsible for student (nominal:’mom’,’father’)
- Raised hand - how many times the student raises his/her hand on classroom (numeric:0-100)
- Visited resources - how many times the student visits a course content(numeric:0-100)
- Viewing announcements-how many times the student checks the new announcements(numeric:0-100)
- Discussion groups - how many times the student participate on discussion groups (numeric:0-100)
- Parent Answering Survey - parent answered the surveys which are provided from school or not (nominal:’Yes’,’No’)
- Parent School Satisfaction - the Degree of parent satisfaction from school(nominal:’Yes’,’No’)
- Student Absence Days - the number of absence days for each student (nominal: above-7, under-7)
- Class - Low-Level: interval includes values from 0 to 69, Middle-Level: interval includes values from 70 to 89, High-Level: interval includes values from 90-100.

4. Progress to Date

At the current point, data set and attributes have been examined. It was decided which graphics could be drawn and how the results could be drawn. And also some of them were made.

4.1. Dataset Manipulation

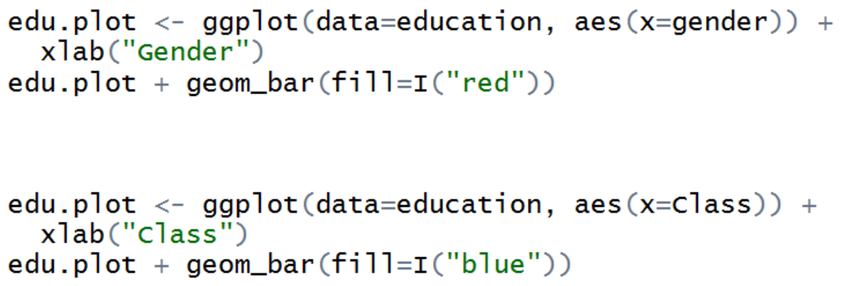
The data set used in this project has been downloaded in csv format via internet. The downloaded csv file is opened in the R program using the following code as the Figure 1.1



```
education <- read.csv("C:/Users/Ayse/Desktop/Ceng 4.2/StatisticalComputing/uygulama/proje/xAPI-Edu-Data.csv")
education

print(nrow(education))    # row number
print(ncol(education))    # column number

head(education)
head(education,4)
summary(education)
colnames(education)     # attribute names
```

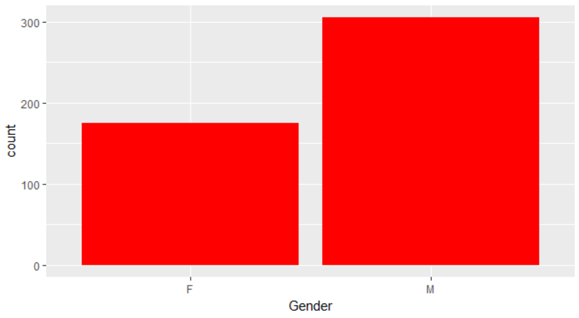
Figure 1.1: R Script

4.2. Gender, Class

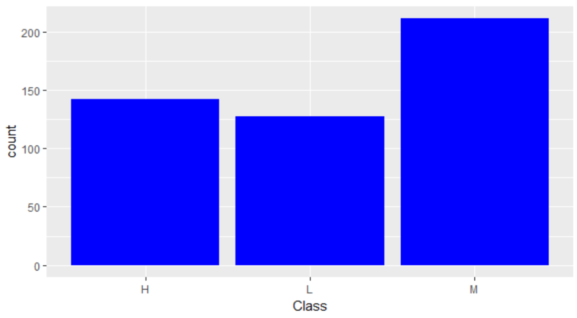


```
edu.plot <- ggplot(data=education, aes(x=gender)) +
  xlab("Gender")
edu.plot + geom_bar(fill=I("red"))


edu.plot <- ggplot(data=education, aes(x=Class)) +
  xlab("Class")
edu.plot + geom_bar(fill=I("blue"))
```

Figure 2.1: R Script



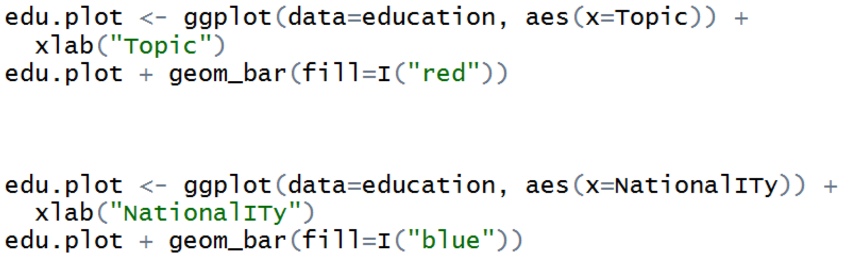
Graph 1.1: Gender Rate



Graph 1.2: Class Rate

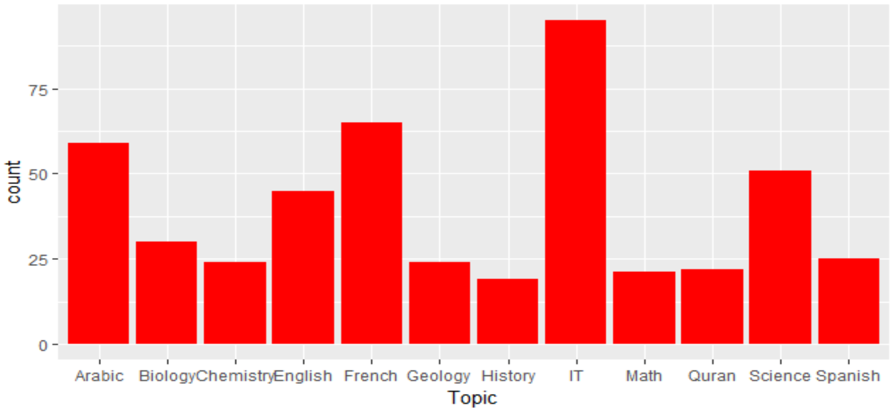
As seen in the graphs above, numbers of males are more than number of females and also majority of students have Middle-Level achievement that is interval includes values from 70 to 89.

4.3. Topic, Nationality

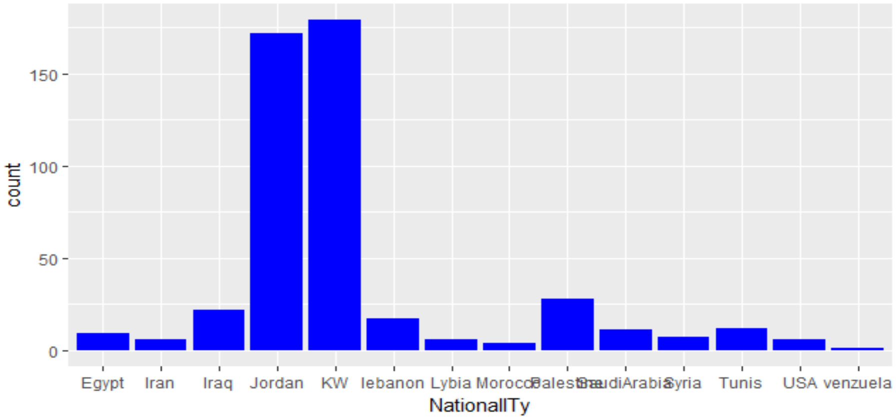


```
edu.plot <- ggplot(data=education, aes(x=Topic)) +
  xlab("Topic")
edu.plot + geom_bar(fill=I("red"))


edu.plot <- ggplot(data=education, aes(x=NationalITy)) +
  xlab("NationalITy")
edu.plot + geom_bar(fill=I("blue"))
```

Figure 3.1: R Script



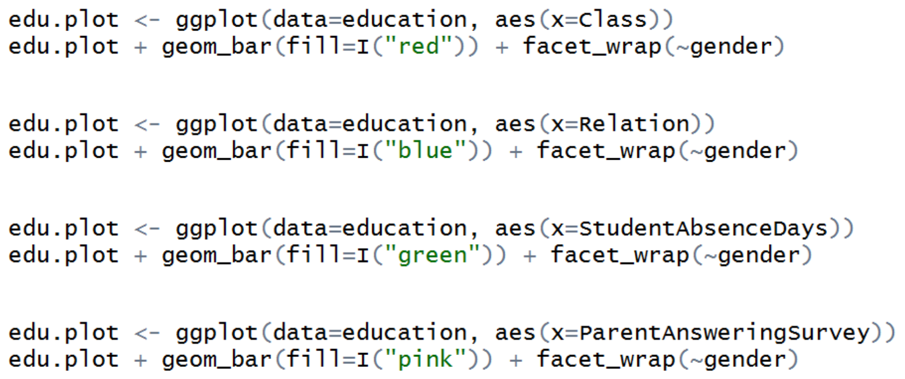
Graph 2.1: Topic Rate



Graph 2.2: Nationality Rate

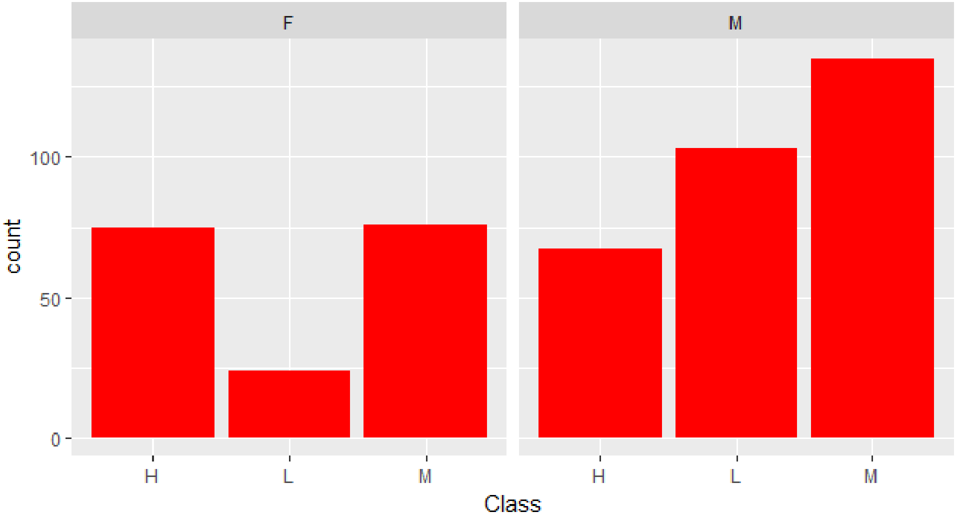
Most of these countries are in the Middle East (Islamic states), perhaps this explains the gender disparity.

4.4. Review by Gender



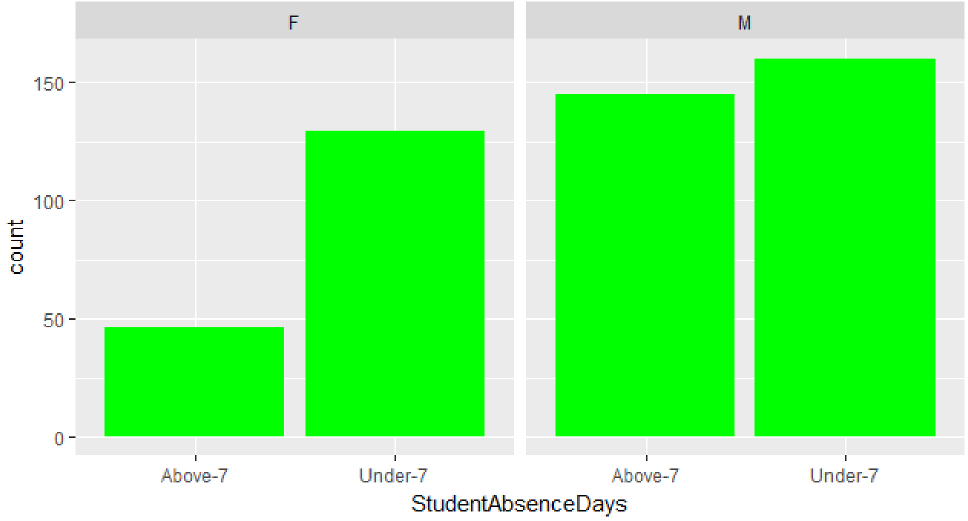
```
edu.plot <- ggplot(data=education, aes(x=Class))
edu.plot + geom_bar(fill=I("red")) + facet_wrap(~gender)

edu.plot <- ggplot(data=education, aes(x=Relation))
edu.plot + geom_bar(fill=I("blue")) + facet_wrap(~gender)

edu.plot <- ggplot(data=education, aes(x=StudentAbsenceDays))
edu.plot + geom_bar(fill=I("green")) + facet_wrap(~gender)

edu.plot <- ggplot(data=education, aes(x=ParentAnsweringSurvey))
edu.plot + geom_bar(fill=I("pink")) + facet_wrap(~gender)
```

Figure 4.1: R Script



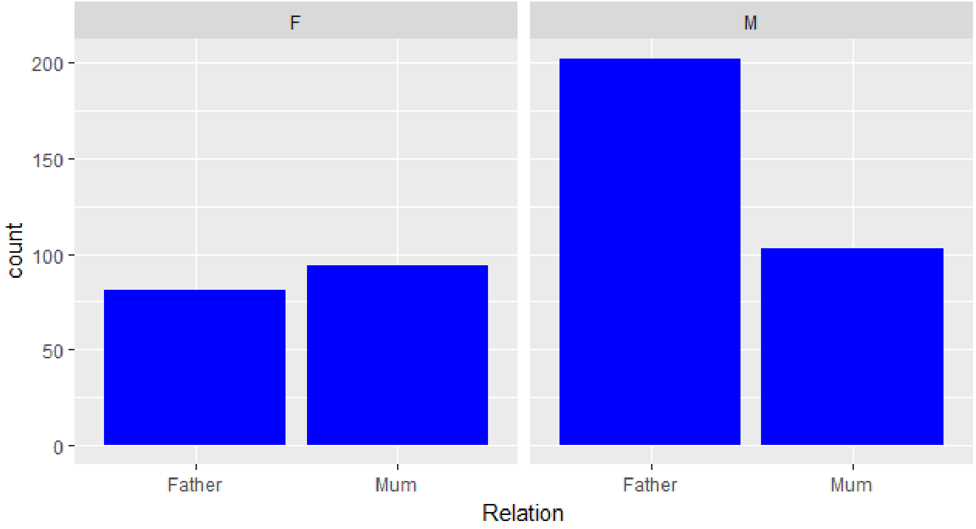
Graph 3.1: Class vs. Gender

As shown the ***Graph 3.1***, girls seem to have performed better than boys.



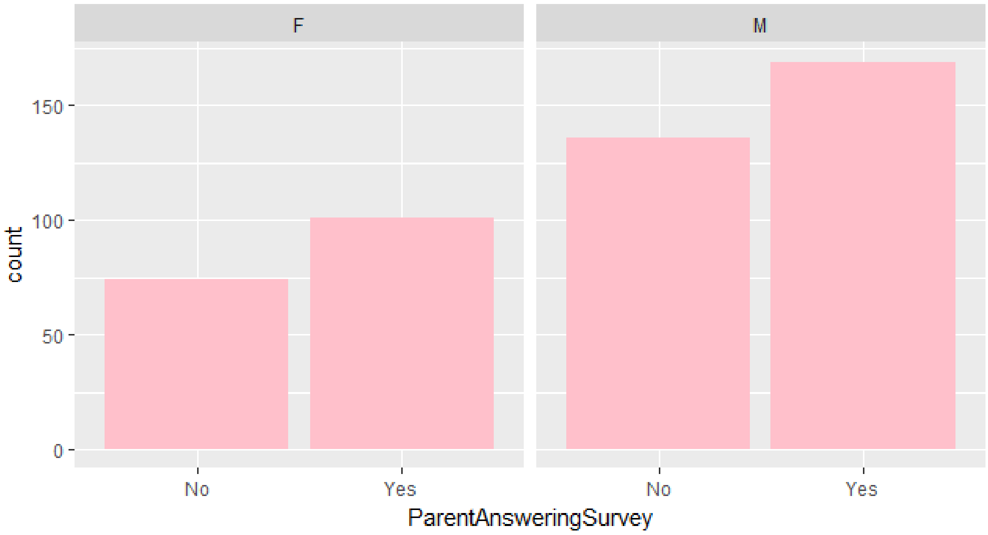
Graph 3.2: Student Absence Day vs. Gender

As shown the Graph 3.2, girls had much better attendance than boys.



Graph 3.3: Parent Relation vs. Gender

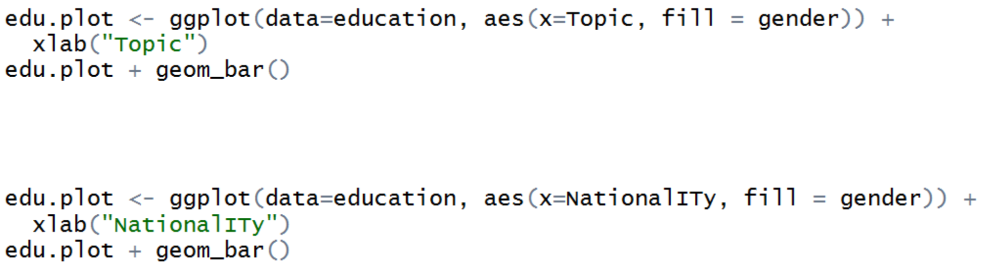
As shown the ***Graph 3.3***, in the case of girls, mothers seem to be more interested in their education than fathers. In boys this is the exact opposite.



Graph 3.4: Parent Answering Survey vs. Gender

As shown the ***Graph 3.4***, in the case of girls and boys, parents answer the surveys mostly.

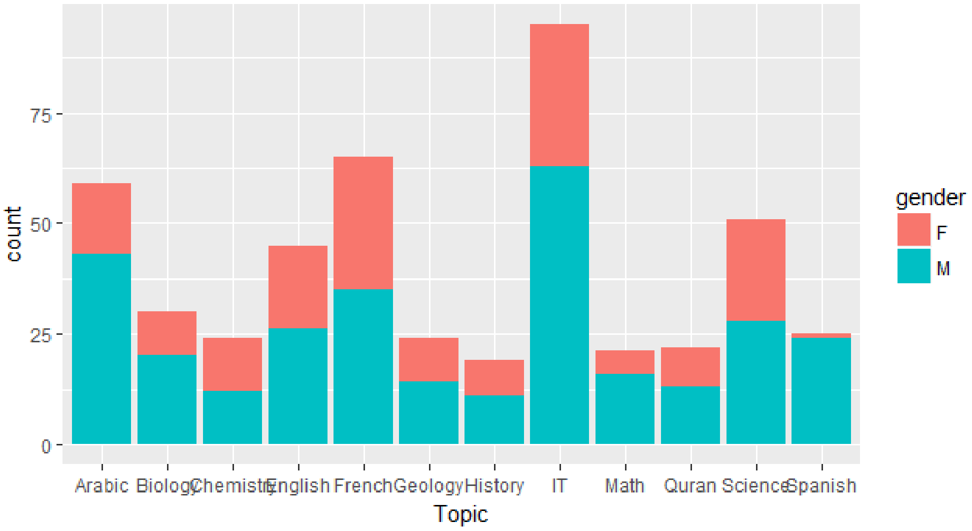
4.5. Topics and Nationalities vs. Gender



```
edu.plot <- ggplot(data=education, aes(x=Topic, fill = gender)) +
  xlab("Topic")
edu.plot + geom_bar()
```

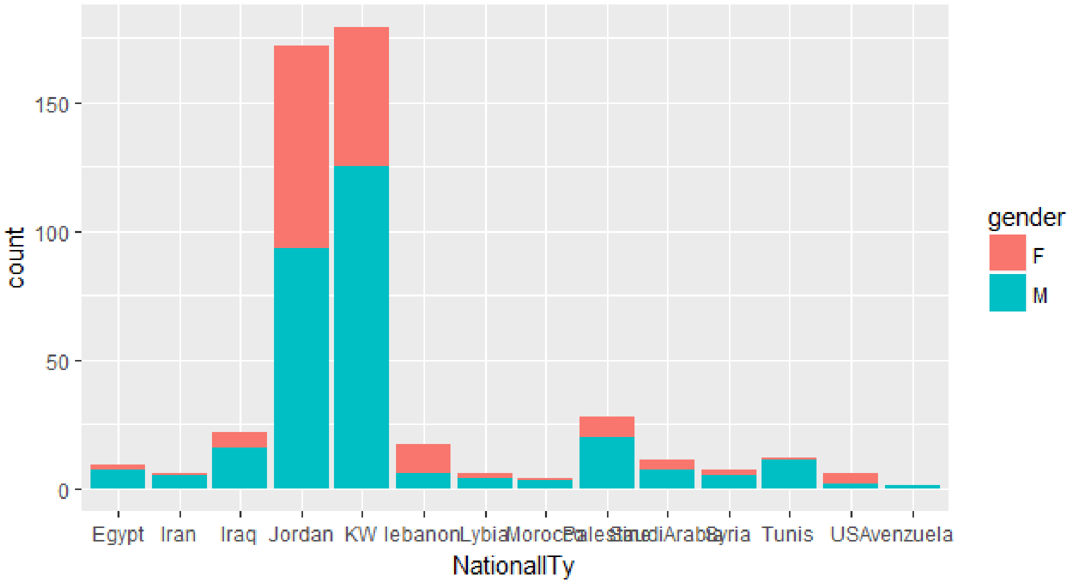
```
edu.plot <- ggplot(data=education, aes(x=NationalITy, fill = gender)) +
  xlab("NationalITy")
edu.plot + geom_bar()
```

Figure 5.1: R Script



Graph 4.1: Topic vs. Gender

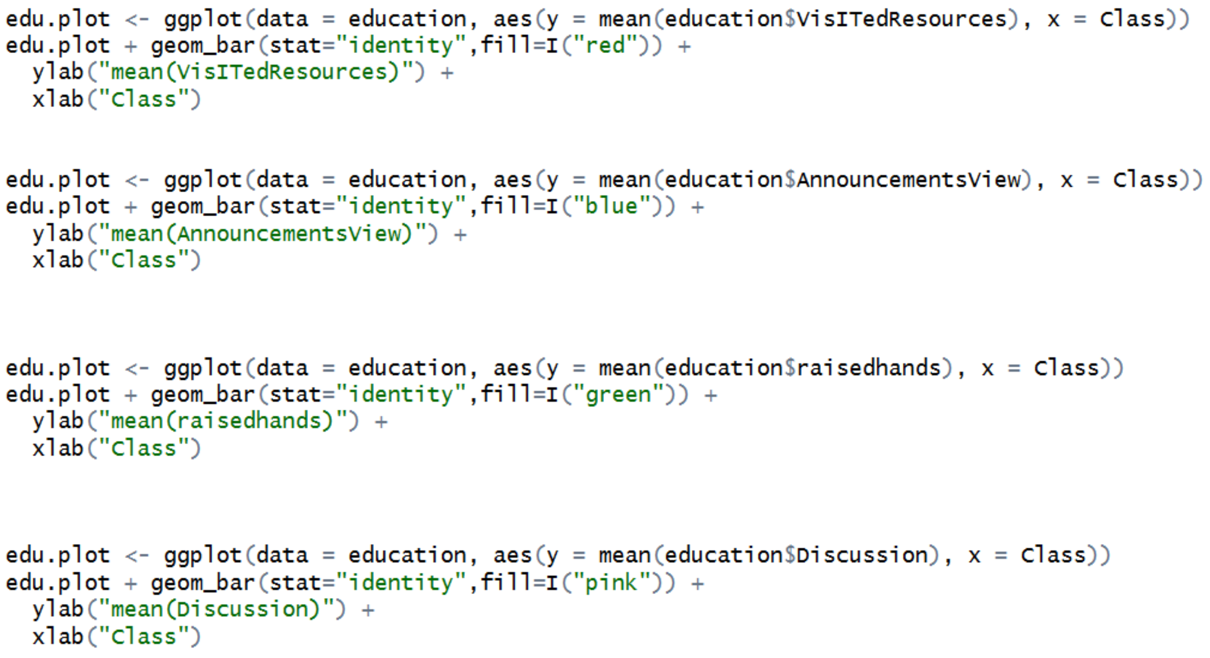
According to ***Graph 4.1***, no apparent gender bias when it comes to subject/topic choices, we cannot conclude that girls performed better because they perhaps took less technical subjects.



Graph 4.2: Nationality vs. Gender

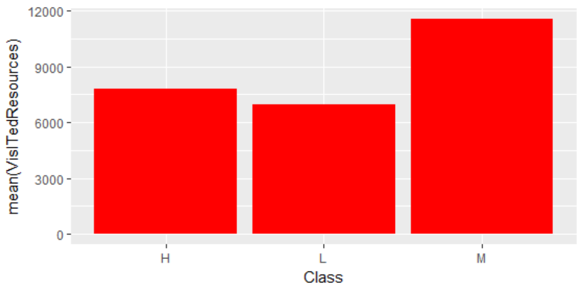
According to ***Graph 4.2***, gender disparity holds even at a country level.

4.6. Review by Class

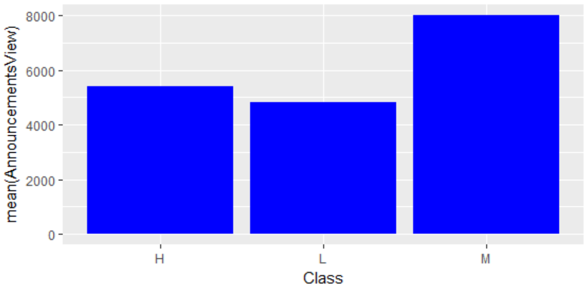


```
edu.plot <- ggplot(data = education, aes(y = mean(education$VisITedResources), x = Class))
edu.plot + geom_bar(stat="identity",fill=I("red")) +
  ylab("mean(VisITedResources)") +
  xlab("Class")


edu.plot <- ggplot(data = education, aes(y = mean(education$AnnouncementsView), x = Class))
edu.plot + geom_bar(stat="identity",fill=I("blue")) +
  ylab("mean(AnnouncementsView)") +
  xlab("Class")


edu.plot <- ggplot(data = education, aes(y = mean(education$raisedhands), x = Class))
edu.plot + geom_bar(stat="identity",fill=I("green")) +
  ylab("mean(raisedhands)") +
  xlab("Class")


edu.plot <- ggplot(data = education, aes(y = mean(education$Discussion), x = Class))
edu.plot + geom_bar(stat="identity",fill=I("pink")) +
  ylab("mean(Discussion)") +
  xlab("Class")
```

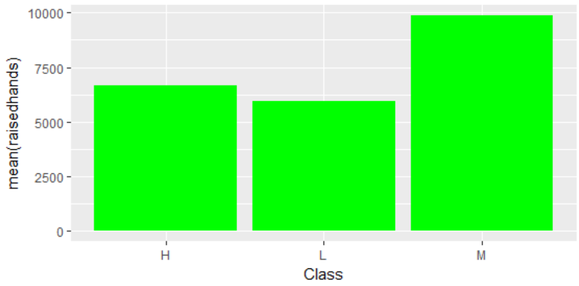
Figure 6.1: R Script



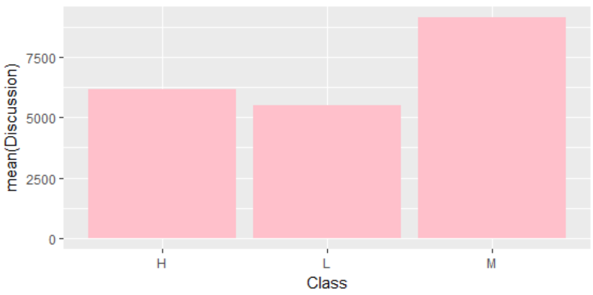
Graph 5.1: Visited Resource vs. Achievement



Graph 5.2: Announcements View vs. Achievement



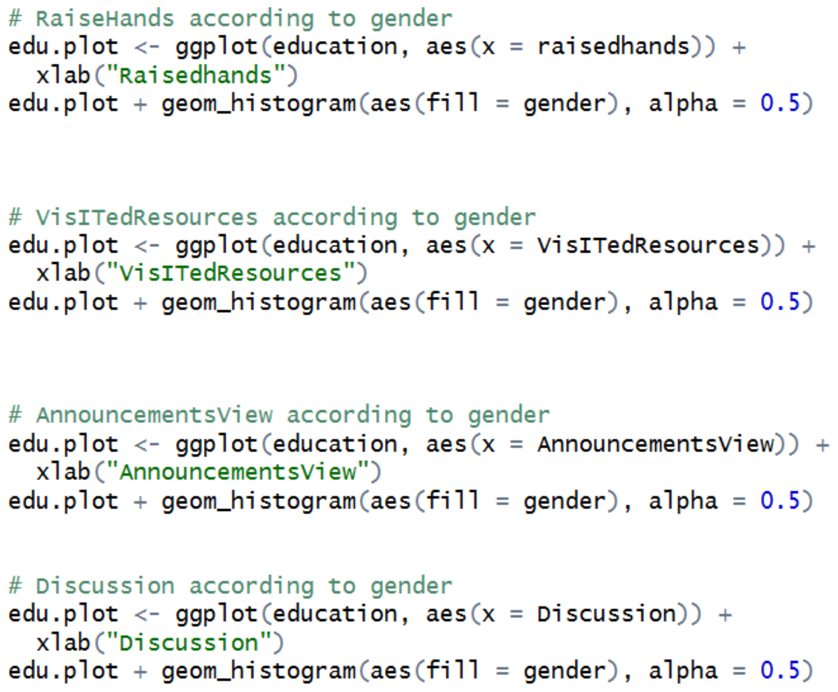
Graph 5.3: Raise Hands vs. Achievement



Graph 5.4: Discussion vs. Achievement

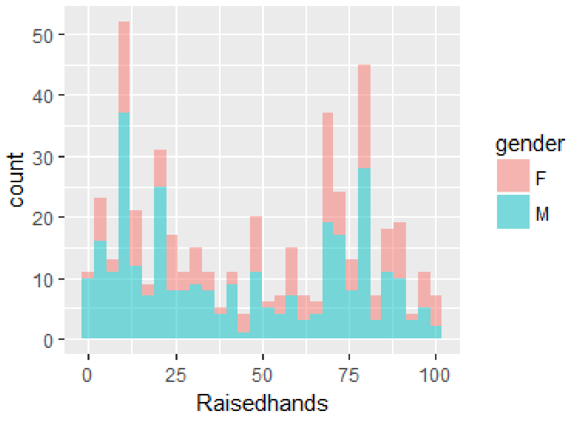
According to ***Graph* 5.1, 5.2, 5.3, 5.4**, as expected, those that participated more (higher counts in discussion, raised hands, announcement views and visited resource) performed better. It is may be about correlation and causation.

4.7. Review by Gender (2)

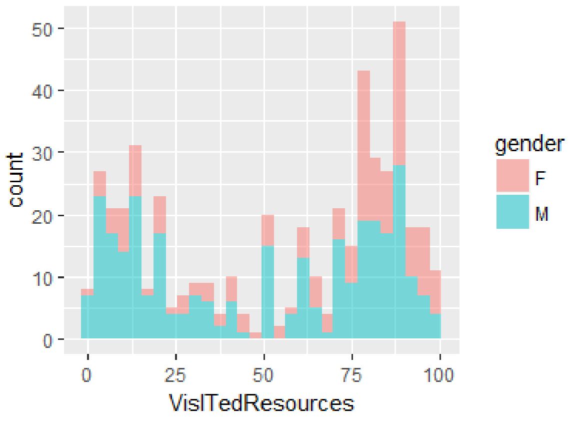


```
# RaiseHands according to gender
edu.plot <- ggplot(education, aes(x = raisedhands)) +
  xlab("Raisedhands")
edu.plot + geom_histogram(aes(fill = gender), alpha = 0.5)


# VisITedResources according to gender
edu.plot <- ggplot(education, aes(x = VisITedResources)) +
  xlab("VisITedResources")
edu.plot + geom_histogram(aes(fill = gender), alpha = 0.5)


# AnnouncementsView according to gender
edu.plot <- ggplot(education, aes(x = AnnouncementsView)) +
  xlab("AnnouncementsView")
edu.plot + geom_histogram(aes(fill = gender), alpha = 0.5)

# Discussion according to gender
edu.plot <- ggplot(education, aes(x = Discussion)) +
  xlab("Discussion")
edu.plot + geom_histogram(aes(fill = gender), alpha = 0.5)
```

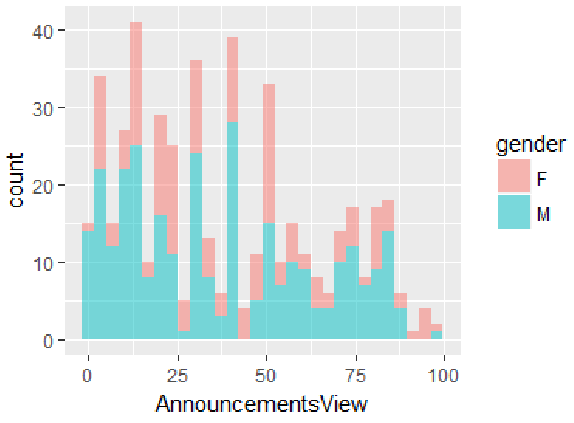
Figure 7.1: R Script

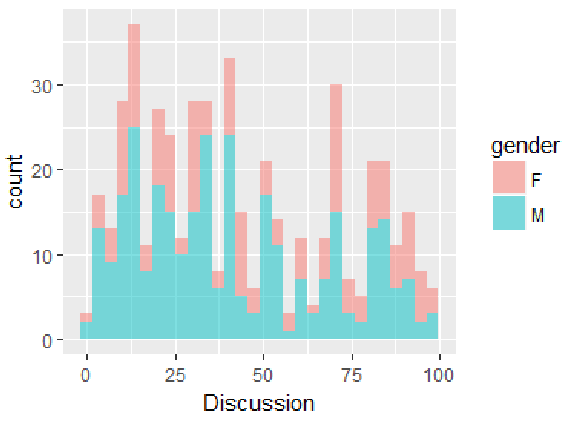


Graph 6.1: Raise Hands vs. Gender



Graph 6.2: Visited Resources vs. Gender



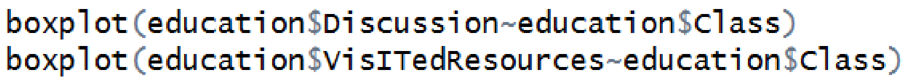


Graph 6.3: Announcements View vs. Gender

Graph 6.4: Discussion vs. Gender

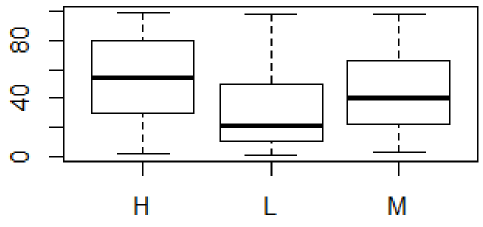
According to ***Graph 6.1, 6.2, 6.3, 6.4***, females are more active and more successful than males.

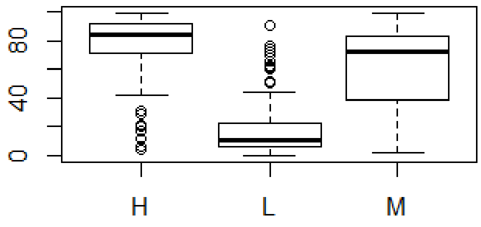
4.8. Review by Class Activities



```
boxplot(education$Discussion~education$Class)
boxplot(education$VisITedResources~education$Class)
```

Figure 8.1: R Script



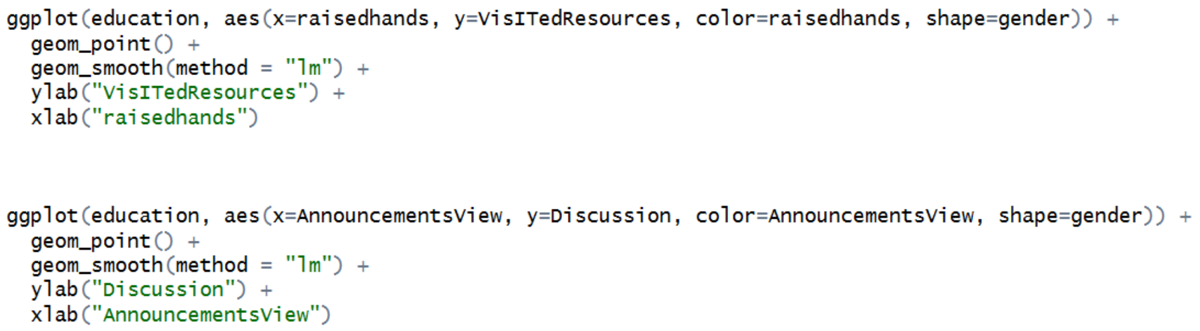


Graph 7.1: Discussion vs. Achievement

Graph 7.2: Visited Resources vs. Achievement

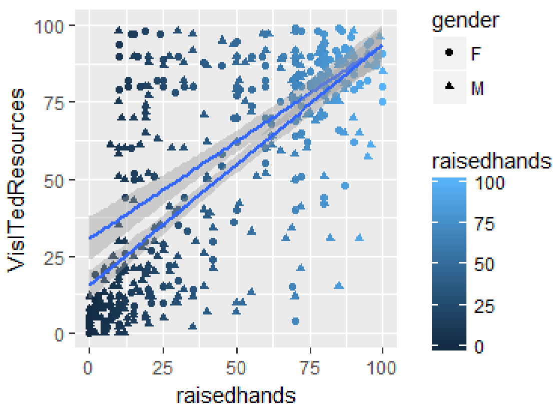
According to ***Graph 7.1 and 7.2***, visiting resources may not be as sure a path to performing well as discussions.

4.9. Review by Class Activities (2)

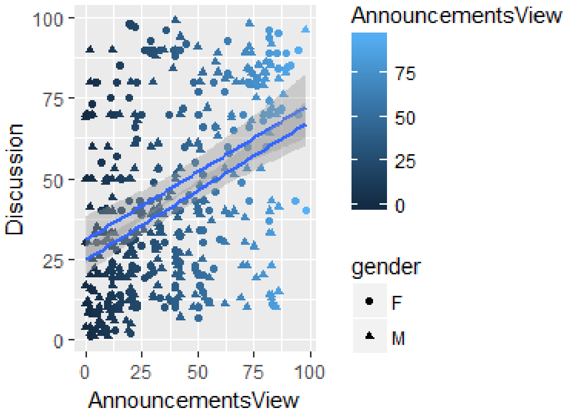


```
ggplot(education, aes(x=raisedhands, y=VisITedResources, color=raisedhands, shape=gender)) +
  geom_point() +
  geom_smooth(method = "lm") +
  ylab("VisITedResources") +
  xlab("raisedhands")


ggplot(education, aes(x=AnnouncementsView, y=Discussion, color=AnnouncementsView, shape=gender)) +
  geom_point() +
  geom_smooth(method = "lm") +
  ylab("Discussion") +
  xlab("AnnouncementsView")
```

Figure 9.1: R Script



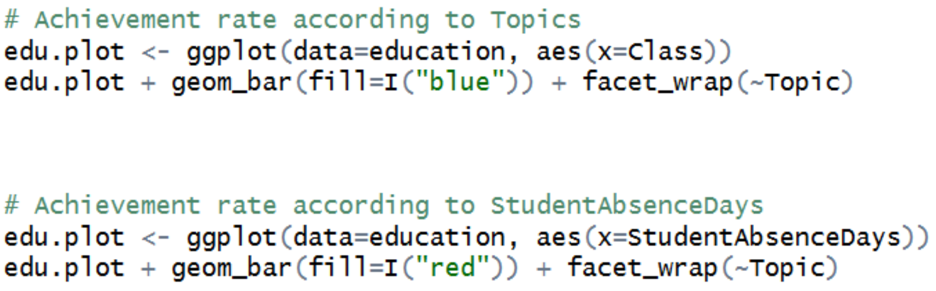
Graph 8.1: Visited Resources vs. Raise Hands



Graph 8.2: Discussion vs. Announcements View

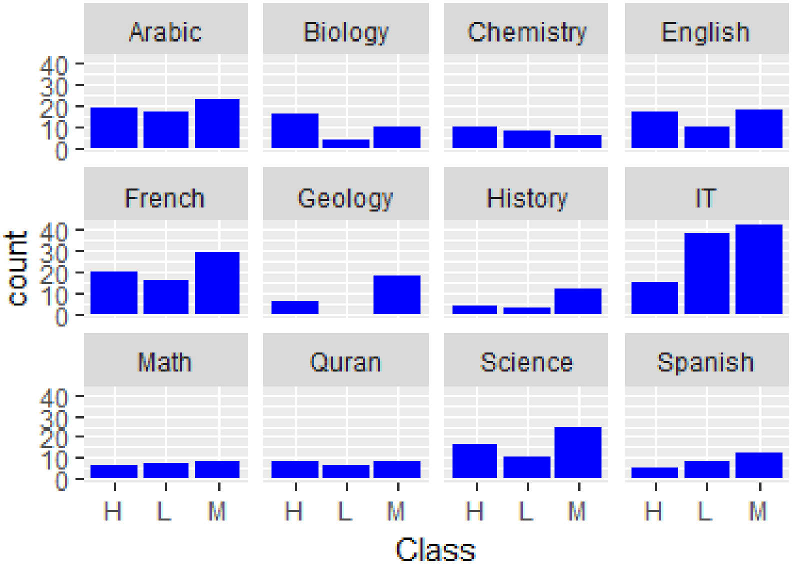
According to ***Graph 8.1 and 8.2***, there does not appear to be much of a linear relationship between the numerical features.

4.10. Review by Achievement Rate



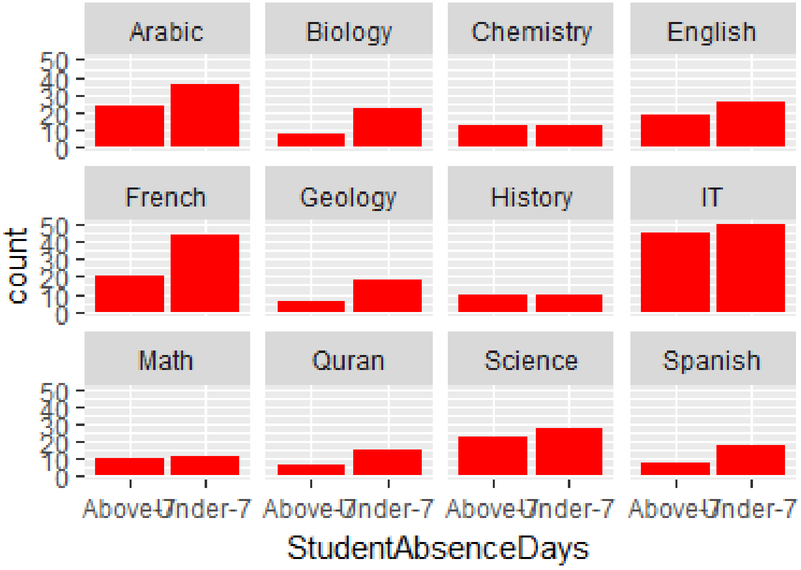
```
# Achievement rate according to Topics
edu.plot <- ggplot(data=education, aes(x=Class))
edu.plot + geom_bar(fill=I("blue")) + facet_wrap(~Topic)


# Achievement rate according to StudentAbsenceDays
edu.plot <- ggplot(data=education, aes(x=StudentAbsenceDays))
edu.plot + geom_bar(fill=I("red")) + facet_wrap(~Topic)
```

Figure 10.1: R Script



Graph 9.1: Achievement Rate by Topics

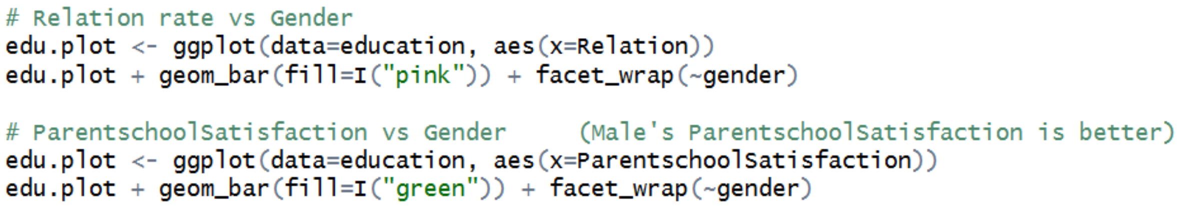
According to ***Graph 9.1***, Biology, English and Chemistry have higher achievement rates.



Graph 9.2: Achievement Rate by Student Absence Days

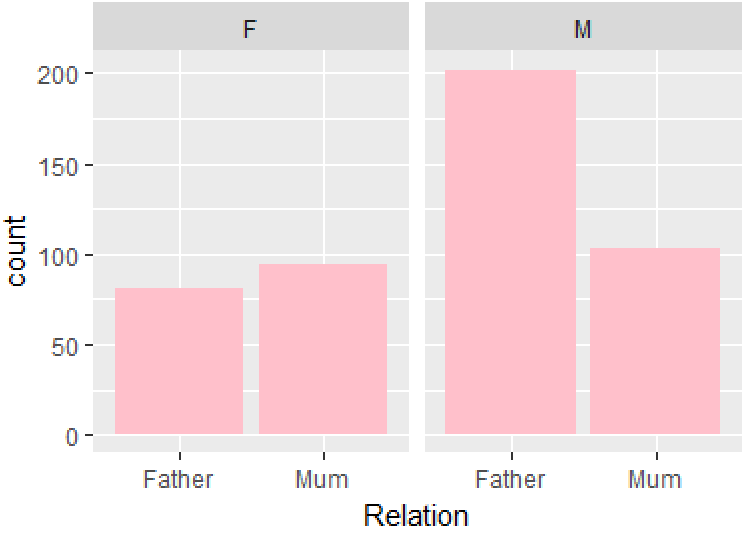
According to ***Graph 9.2***, Student Absence Days is low for Arabic, Math, Geology and Biology.

4.11. Review by Parents Effect



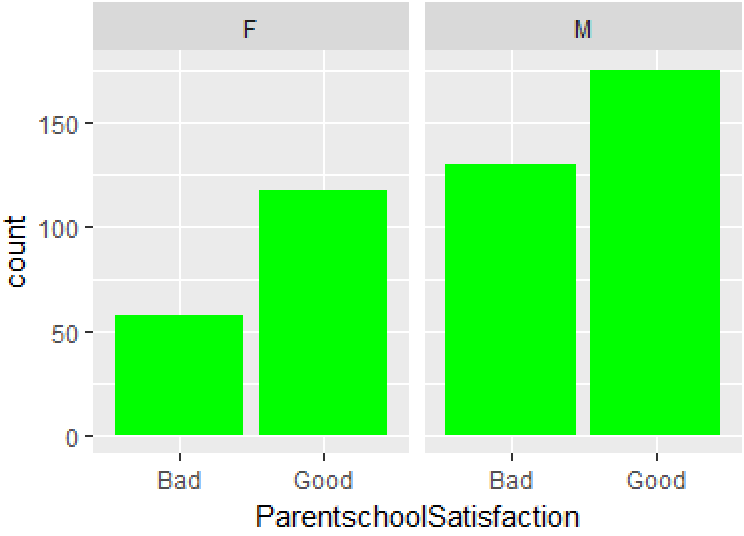
```
# Relation rate vs Gender
edu.plot <- ggplot(data=education, aes(x=Relation))
edu.plot + geom_bar(fill=I("pink")) + facet_wrap(~gender)

# ParentschoolSatisfaction vs Gender      (Male's ParentschoolSatisfaction is better)
edu.plot <- ggplot(data=education, aes(x=ParentschoolSatisfaction))
edu.plot + geom_bar(fill=I("green")) + facet_wrap(~gender)
```

Figure 11.1: R Script



Graph 10.1: Relation Rate by Gender

According to ***Graph 10.1***, male's relation is father mostly.

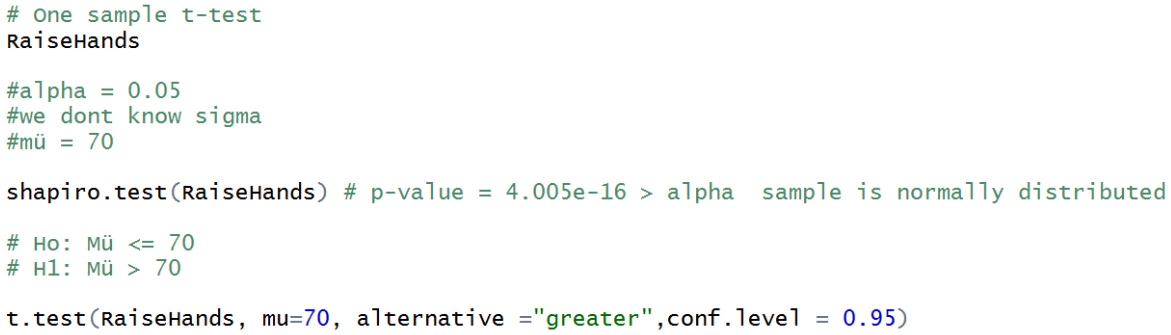


Graph 10.2: Parent School Satisfaction by Gender

According to ***Graph 10.2***, male's parent school satisfaction is better.

5. Hypothesis Testing

5.1. One Sample t-test

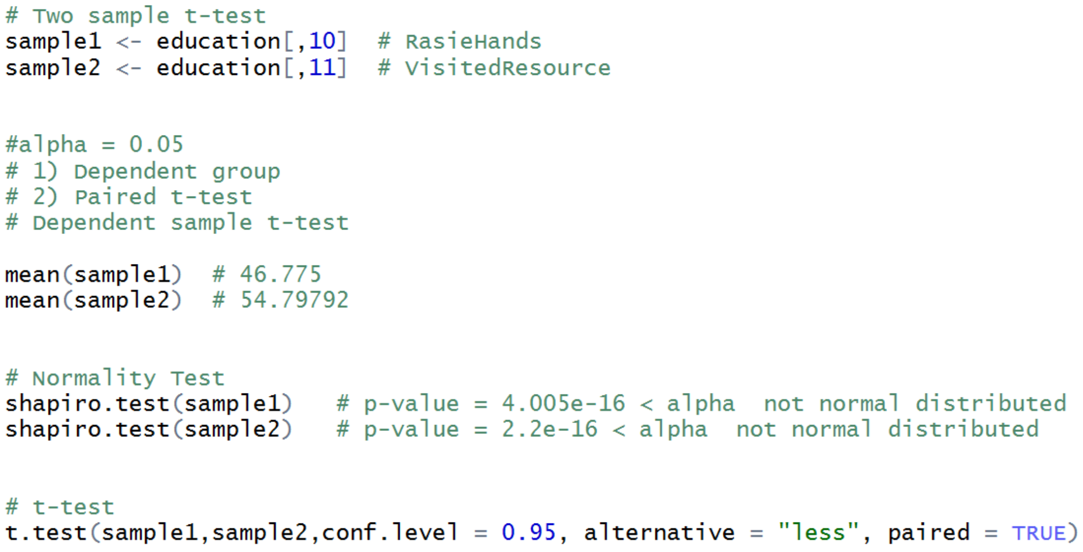


```
# One sample t-test
RaiseHands

#alpha = 0.05
#we dont know sigma
#mü = 70

shapiro.test(RaiseHands) # p-value = 4.005e-16 > alpha  sample is normally distributed

# Ho: Mü <= 70
# H1: Mü > 70

t.test(RaiseHands, mu=70, alternative ="greater",conf.level = 0.95)
```

Figure 12.1: R Script

According to ***R Script 12.1***, p-value = 1 > alpha=0.05 so you cannot reject the null hypothesis, it means rate of Raise Hands is not greater than 70.

5.2. Two Sample t-test

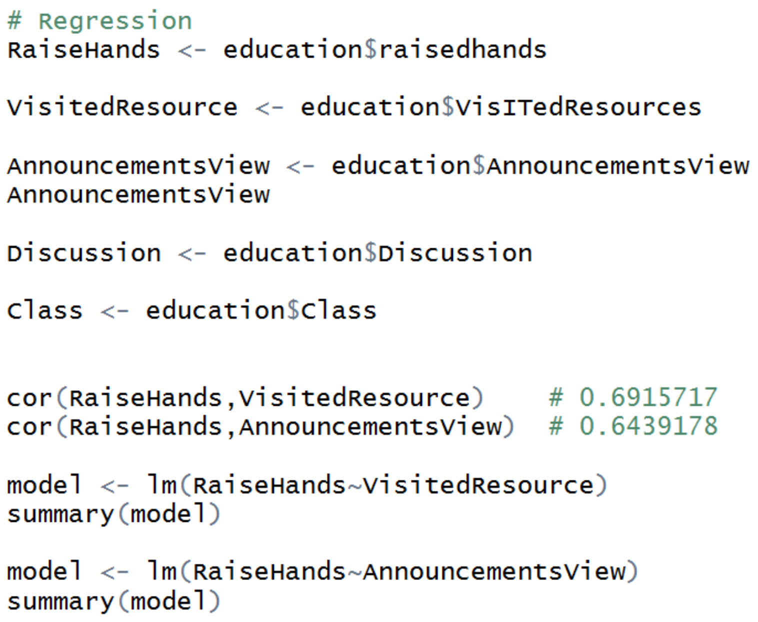


```
# Two sample t-test
sample1 <- education[,10]  # RasieHands
sample2 <- education[,11]  # VisitedResource


#alpha = 0.05
# 1) Dependent group
# 2) Paired t-test
# Dependent sample t-test

mean(sample1)  # 46.775
mean(sample2)  # 54.79792


# Normality Test
shapiro.test(sample1)   # p-value = 4.005e-16 < alpha  not normal distributed
shapiro.test(sample2)   # p-value = 2.2e-16 < alpha  not normal distributed


# t-test
t.test(sample1,sample2,conf.level = 0.95, alternative = "less", paired = TRUE)
```

Figure 13.1: R Script

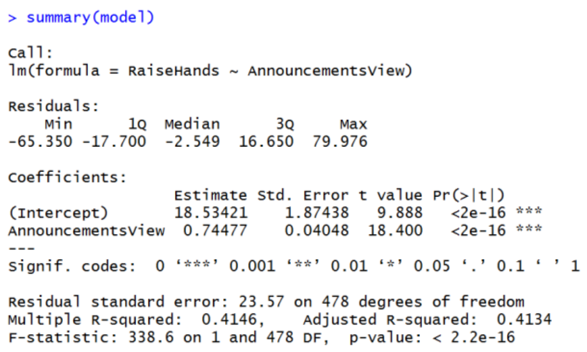
According to ***R Script 13.1***, p-value = 4.806e-12 < alpha = 0.05. So reject the null hypothesis (Ho: Mü1 = Mü2). Visited Resource is not less than Raise Hands.

6. Regression



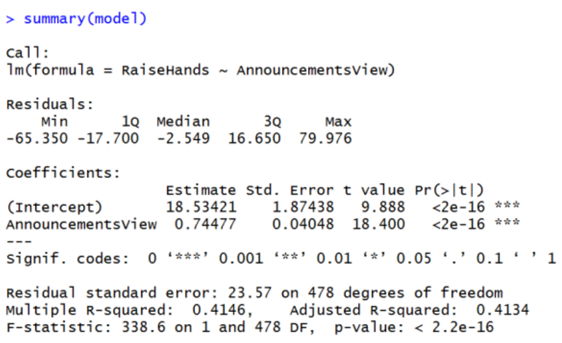
```
# Regression
RaiseHands <- education$raisedhands

VisitedResource <- education$VisITedResources

AnnouncementsView <- education$AnnouncementsView
AnnouncementsView

Discussion <- education$Discussion

Class <- education$Class


cor(RaiseHands,VisitedResource)     # 0.6915717
cor(RaiseHands,AnnouncementsView)   # 0.6439178

model <- lm(RaiseHands~VisitedResource)
summary(model)

model <- lm(RaiseHands~AnnouncementsView)
summary(model)
```

Figure 14.1: R Script



```
> summary(model)

Call:
lm(formula = RaiseHands ~ AnnouncementsView)

Residuals:
    Min      1Q  Median      3Q     Max 
-65.350 -17.700  -2.549  16.650  79.976 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       18.53421    1.87438   9.888   <2e-16 ***
AnnouncementsView  0.74477    0.04048  18.400   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23.57 on 478 degrees of freedom
Multiple R-squared:  0.4146,	Adjusted R-squared:  0.4134 
F-statistic: 338.6 on 1 and 478 DF,  p-value: < 2.2e-16
```

Figure 14.2: Result for RaiseHands~VisitedResources



```
> summary(model)

Call:
lm(formula = RaiseHands ~ AnnouncementsView)

Residuals:
    Min      1Q  Median      3Q     Max 
-65.350 -17.700  -2.549  16.650  79.976 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       18.53421    1.87438   9.888   <2e-16 ***
AnnouncementsView  0.74477    0.04048  18.400   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23.57 on 478 degrees of freedom
Multiple R-squared:  0.4146,	Adjusted R-squared:  0.4134 
F-statistic: 338.6 on 1 and 478 DF,  p-value: < 2.2e-16
```

Figure 14.3: for RaiseHands~AnnouncementsView

According to ***Figure 14.2***, Rate of Visited Resource affects the rate of Raise Hands positively. And, according to ***Figure 14.3***, Rate of Announcements View affects the rate of Raise Hands positively.

7. Conclusions

To summarize, in this project, it is aimed to examine to the students' achievements in the class. All R scripts are available in my github page, https://github.com/aysedilekk/Students_Academic_Performance

8. References

[1] Students' Academic Performance Dataset, https://www.kaggle.com/aljarah/xAPI-Edu-Data, November 11, 2016

[2] R Tutorial, http://www.cyclismo.org/tutorial/R

[3] Lecture Notes, Dr. Eralp DOGU, 2017.