Природно-математички факултет Крагујевац

УВОД У НАУКУ О ПОДАЦИМА

Семинарски рад

Професор:

др Бранко Арсић

Чланови тима:

Милош Милетић 66/2020

Емилија Јевремовић 47/2020

Садржај

1 Увод

У овом раду анализираћемо скуп података под називом „*Student Performance Dataset*“, помоћу кога ћемо имати увид у факторе који утичу на успех студената на испитима. Овај скуп података обухвата информације о навикама у учењу као што су: присуство предавањима, време које студенти проводе учећи, као и демографски подаци студената и њихових породица. Како бисмо боље разумели образовне процесе, као и који фактори највише утичу на постигнута академска достигнућа студената, извршићемо детаљну анализу ових података.

Циљ овог рада је да кроз анализу података откријемо који фактори највише утичу на успех студената. Посебну пажњу усмерићемо на њихове навике у учењу и ниво мотивације, јер се на прву лопту чини да ови аспекти имају директан утицај на резултате на испитима. Надамо се да ће резултати ове анализе помоћи студентима да постигну боље резултате и напредак.

Поред тога, бавићемо се и изградњом предиктивних модела који ће моћи да предвиде академски успех на основу унетих параметара. На тај начин, могу се идентификовати потенцијални проблеми код студената и омогућити да их на време увиде како би се побољшала њихова будућа постигнућа.

2 Представљање проблема

У овом раду користи се ***Student Performance Dataset***, који садржи податке о студентима и различитим факторима који утичу на њихов академски успех. *Dataset*, као што смо већ поменули, обухвата информације о навикама учења, присуству на часовима, ангажовању родитеља и демографским карактеристикама, а циљ је анализирати како ови фактори доприносе постигнућима на испитима (*Exam_Score*). Проблем којим ћемо се прво бавити је идентификација кључних фактора који највише утичу на успех студената и како се ти подаци могу искористити за предвиђање будућих резултата.

За почетак ћемо учитати потребне библиотеке за рад на пројекту.

```{r}
library(tidyverse)
library(Amelia)
library(moments)
library(reshape2)
library(rpart)
library(rpart.plot)
library(randomForest)
```

Затим ћемо импортовати *dataset*

```{r}
dataset = read.csv("StudentPerformanceFactors.csv")
View(dataset)
```

Представићемо *dataset* и објаснити сваку колону. Функцијом *str* проверавамо каква је структура датих колона. У примеру можемо видети да имамо 7 колона које су нумеричког типа (*int*) и 13 колона које су знаковног типа (*chr*).

```r
{r}
str(dataset)
```

```
'data.frame':   6607 obs. of  20 variables:
 $ Hours_Studied             : int  23 19 24 29 19 19 29 25 17 23 ...
 $ Attendance                : int  84 64 98 89 92 88 84 78 94 98 ...
 $ Parental_Involvement      : chr  "Low" "Low" "Medium" "Low" ...
 $ Access_to_Resources       : chr  "High" "Medium" "Medium" "Medium" ...
 $ Extracurricular_Activities: chr  "No" "No" "Yes" "Yes" ...
 $ Sleep_Hours               : int  7 8 7 8 6 8 7 6 6 8 ...
 $ Previous_Scores           : int  73 59 91 98 65 89 68 50 80 71 ...
 $ Motivation_Level          : chr  "Low" "Low" "Medium" "Medium" ...
 $ Internet_Access           : chr  "Yes" "Yes" "Yes" "Yes" ...
 $ Tutoring_Sessions         : int  0 2 2 1 3 3 1 1 0 0 ...
 $ Family_Income             : chr  "Low" "Medium" "Medium" "Medium" ...
 $ Teacher_Quality           : chr  "Medium" "Medium" "Medium" "Medium" ...
 $ School_Type               : chr  "Public" "Public" "Public" "Public" ...
 $ Peer_Influence            : chr  "Positive" "Negative" "Neutral" "Negative" ...
 $ Physical_Activity         : int  3 4 4 4 4 3 2 2 1 5 ...
 $ Learning_Disabilities     : chr  "No" "No" "No" "No" ...
 $ Parental_Education_Level  : chr  "High School" "College" "Postgraduate" "High
School" ...
 $ Distance_from_Home        : chr  "Near" "Moderate" "Near" "Moderate" ...
 $ Gender                    : chr  "Male" "Female" "Male" "Male" ...
 $ Exam_Score                : int  67 61 74 71 70 71 67 66 69 72 ...
```

Као и на слици изнад, функцијом *dim* можемо видети димензије скупа података. Види се да наш скуп података садржи 6607 редова и 20 колона.

```r
{r}
dim(dataset)
```

```
[1] 6607   20
```

Опис колона које садржи скуп података *Student Performance Factors*

- ***Hours_Studied*** – Број сати које је студент провео учећи недељно
- ***Attendance*** – Проценат долазака на часове
- ***Parental_Involvement*** – Ниво укључености родитеља у образовање студента (Low, Medium, High)
- ***Access_To_Resources*** – Доступност образовних ресурса (Low, Medium, High)
- ***Extracurricular_Activities*** - Учешће у ваннаставним активностима (Yes, No)
- ***Sleep_Hours*** – Просечан број сати сна по ноћи
- ***Previous_Scores*** – Оцене са претходних испита
- ***Motivation_Level*** – Ниво мотивације студента (Low, Medium, High)
- ***Internet_Access*** – Доступност приступа интернету (Yes, No)
- ***Tutoring_Sessions*** – Број посећених допунских часова месечно
- ***Family_Income*** – Ниво породичног прихода (Low, Medium, High)

- ***Teacher_Quality*** – Квалитет наставника (Low, Medium, High)
- ***School_Type*** – Врста школе коју похађа (Public, Private)
- ***Peer_Influence*** – Утицај вршњака на академски успех (Positive, Neutral, Negative)
- ***Physical_Activity*** – Просечан број сати физичке активности недељно
- ***Learning_Disabilities*** – Присуство сметњи у учењу (Yes, No)
- ***Parental_Education_Level*** – Највиши ниво образовања родитеља (High School, College, Postgraduate)
- ***Distance_from_Home*** – Удаљеност од куће до школе (Near, Moderate, Far)
- ***Gender*** – Пол студента (Male, Female)
- ***Exam_Score*** – Поени са завршног испита

3 Припрема података

Сада ћемо да проверимо колико наша база података има *NA* вредности. Након провере морамо утврдити да ли ћемо избацити неку од колона или покушати да попунимо неким приближним вредностима. Функција *summary* нам омогућава да видимо додатне информације о свакој колони као што су минималне и максималне вредности, просечна вредност, медијана као и први и трећи квартал за нумеричке колоне, док за категоријске колоне приказује учесталост сваке категорије.

```r
summary(dataset)
```

```
 Hours_Studied     Attendance     Parental_Involvement Access_to_Resources
Extracurricular_Activities  Sleep_Hours
 Min.   : 1.00   Min.   : 60.00   Length:6607          Length:6607
Length:6607                Min.   : 4.000
 1st Qu.:16.00   1st Qu.: 70.00   Class :character     Class :character    Class
:character           1st Qu.: 6.000
 Median :20.00   Median : 80.00   Mode  :character     Mode  :character    Mode
:character           Median : 7.000
 Mean   :19.98   Mean   : 79.98
Mean   : 7.029
 3rd Qu.:24.00   3rd Qu.: 90.00
3rd Qu.: 8.000
 Max.   :44.00   Max.   :100.00
Max.   :10.000
 Previous_Scores  Motivation_Level   Internet_Access    Tutoring_Sessions
Family_Income      Teacher_Quality
 Min.   : 50.00   Length:6607        Length:6607        Min.   :0.000
Length:6607        Length:6607
 1st Qu.: 63.00   Class :character   Class :character   1st Qu.:1.000     Class
:character   Class :character
 Median : 75.00   Mode  :character   Mode  :character   Median :1.000     Mode
:character   Mode  :character
 Mean   : 75.07                                        Mean   :1.494
 3rd Qu.: 88.00                                        3rd Qu.:2.000
 Max.   :100.00                                        Max.   :8.000
 School_Type        Peer_Influence     Physical_Activity Learning_Disabilities
Parental_Education_Level Distance_from_Home
 Length:6607        Length:6607        Min.   :0.000     Length:6607
Length:6607              Length:6607
 Class :character   Class :character   1st Qu.:2.000     Class :character
Class :character         Class :character
 Mode  :character   Mode  :character   Median :3.000     Mode  :character
Mode  :character         Mode  :character
                                      Mean   :2.968
                                      3rd Qu.:4.000
                                      Max.   :6.000
    Gender           Exam_Score
 Length:6607        Min.   : 55.00
 Class :character   1st Qu.: 65.00
 Mode  :character   Median : 67.00
                    Mean   : 67.24
                    3rd Qu.: 69.00
                    Max.   :101.00
```

Такође, функција *summary* може да нам покаже да ли у одређеној колони имамо присутне *NA* вредности. У нашем примеру можемо да видимо да се ни за једну колону не приказују информације о *NA* вредностима.

Након дубље анализе базе података дошли смо до закључка да *NA* вредности нису забележене у свом подразумеваном облику, већ је на тим местима остављено празно поље. Сада ћемо да прикажемо колоне које садрже празна поља као и њихов број.

```{r}
empty_counts <- sapply(dataset, function(x) sum(x == ""))
result <- data.frame(Count = empty_counts[empty_counts > 0])
print(result)
```

Description: df [3 × 1]

	Count <int>
Teacher_Quality	78
Parental_Educati...	90
Distance_from_...	67

3 rows

Прво смо уз помоћ функције *sapply* пролазили по колонама и за сваку колону рачунали колико има поља са празним вредностима. Онда смо приказали само оне колоне које имају више од 0 празних поља и дошли до закључка да имамо 3 колоне са *NA* вредностима. То су колоне: *Teacher_Quality, Parental_Education_Level* и *Distance_from_Home.*

Уместо празних поља у бази података, сада ћемо убацити да пише *NA* у тим пољима како бисмо касније могли да користимо додатне фунцкције које раде са тим вредностима и тиме олакшали рад.

```r
dataset[dataset == ""] <- NA
colSums(is.na(dataset))
```

```
             Hours_Studied                 Attendance        Parental_Involvement
Access_to_Resources
                         0                          0                           0
0
Extracurricular_Activities                Sleep_Hours           Previous_Scores
Motivation_Level
                         0                          0                           0
0
           Internet_Access          Tutoring_Sessions            Family_Income
Teacher_Quality
                         0                          0                           0
78
               School_Type             Peer_Influence         Physical_Activity
Learning_Disabilities
                         0                          0                           0
0
  Parental_Education_Level         Distance_from_Home                      Gender
Exam_Score
                        90                         67                           0
0
```

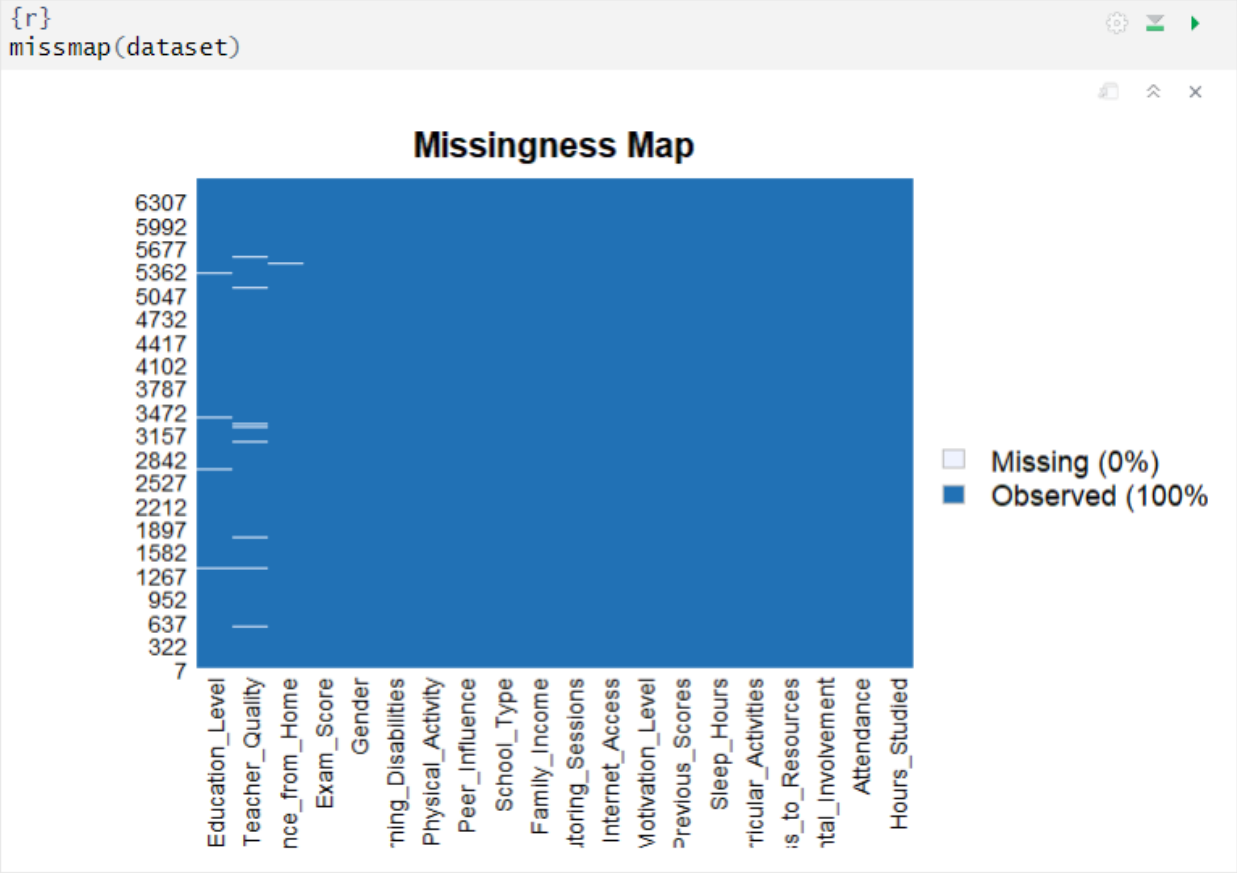
У овом примеру смо у свако празно поље додали *NA* вредност. Такође смо приказали све колоне и број *NA* вредности у свим колонама. Након тога ћемо да процентуално прикажемо колико има недостајућих вредности у колонама и на основу тога донети закључак шта да радимо са њима.

```r
colMeans(is.na(dataset)) * 100
```

```
             Hours_Studied                 Attendance        Parental_Involvement
Access_to_Resources
                  0.000000                   0.000000                    0.000000
0.000000
Extracurricular_Activities                Sleep_Hours           Previous_Scores
Motivation_Level
                  0.000000                   0.000000                    0.000000
0.000000
           Internet_Access          Tutoring_Sessions            Family_Income
Teacher_Quality
                  0.000000                   0.000000                    0.000000
1.180566
               School_Type             Peer_Influence         Physical_Activity
Learning_Disabilities
                  0.000000                   0.000000                    0.000000
0.000000
  Parental_Education_Level         Distance_from_Home                      Gender
Exam_Score
                  1.362192                   1.014076                    0.000000
0.000000
```

У примеру видимо да у колонама које имају недостајуће вредности постоји око 1% *NA* вредности за ту колону.

Функција *missmap* у пакету *Amelia* се користи за визуелни приказ *NA* вредности у подацима. Ту можемо преко графика видети како изгледа расподела *NA* вредности по колонама.



Пошто имамо укупно мање од 1% *NA* вредности, *Missing* је 0%.

Сада ћемо да проверимо да ли постоје неки редови у којима се налази више од једне *NA* вредности.

```r
{r}
rows_with_multiple_na <- rowSums(is.na(dataset)) > 1
dataset[rows_with_multiple_na, ]
```

Description: df [6 × 20]

	Hours_Studied <int>	Attendance <int>	Parental_Involvement <chr>
2225	16	76	High
3742	23	71	Medium
5248	11	78	Low
5257	18	89	Low
5636	24	74	Low
5822	22	98	Medium

6 rows | 1–4 of 20 columns

Пошто редови са више недостајућих вредности могу указивати на озбиљнији проблем у подацима, брисање тих редова би била добра опција. Такође број тих редова је веома мали (6/6607) тако да их хоћемо обрисати.

```r
{r}
dataset <- dataset[rowSums(is.na(dataset)) <= 1, ]
rows_with_multiple_na <- rowSums(is.na(dataset)) > 1
dataset[rows_with_multiple_na, ]
```

Description: df [0 × 20]

0 rows | 1–4 of 20 columns

Остале недостајуће вредности планирамо да попунимо са вредношћу која се најчешће појављује у тој колони. То је честа пракса када је број недостајућих података низак и пошто је наш *dataset* мали, не желимо да избацимо податке који могу садржати друге корисне информације.

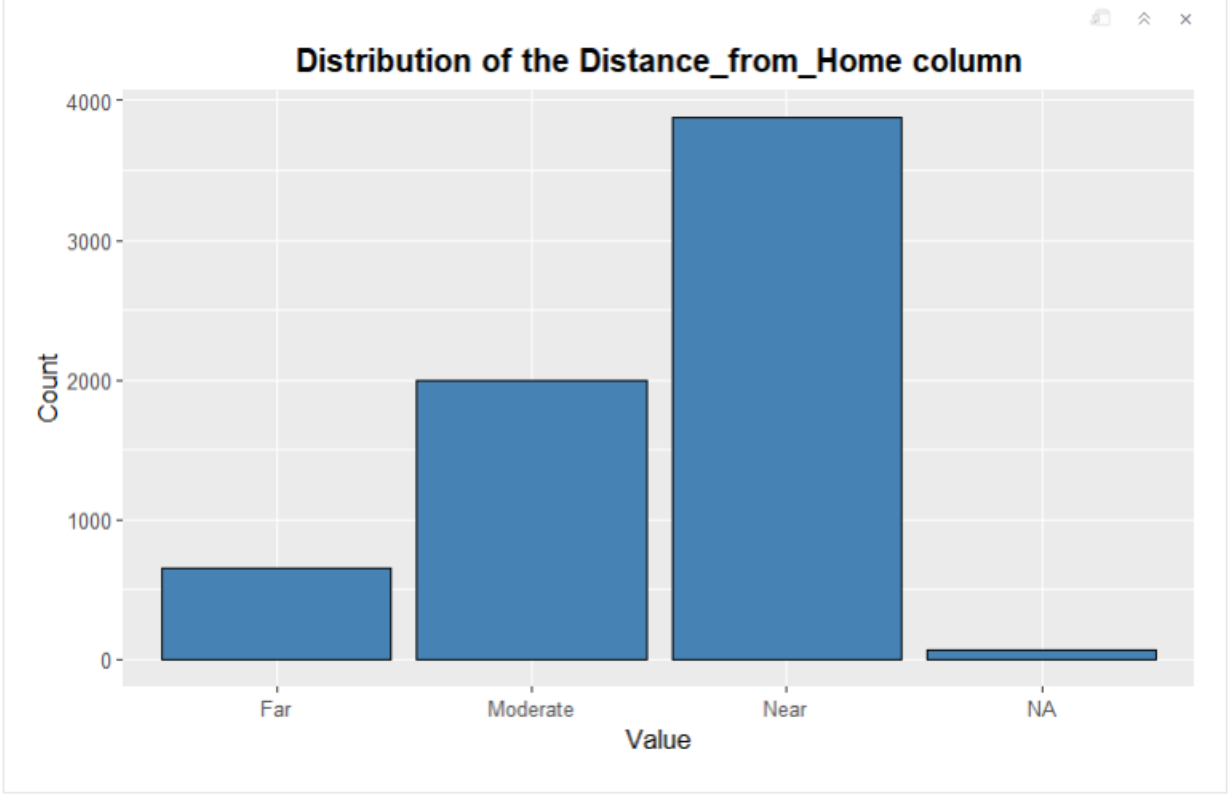
3.1 Колона *Distance_from_Home*

Сада ћемо вредности које су недостајуће да попунимо са вредношћу која се највише појављује у тој колони. Прво ћемо за колону *Distance_from_Home* да пронађемо ту вредност и након тога ћемо да је убацимо у *NA* поља. Користићемо библиотеку *tidyverse* и унутар ње библиотеку *ggplot* преко које ћемо графички да представимо расподелу колоне *Distance_from_Home.*

```r
table(dataset$Distance_from_Home)
```

```
     Far Moderate     Near
     658     1997     3881
```

```r
ggplot(dataset, aes(x = Distance_from_Home)) + geom_bar(fill = "steelblue", color = "black") + labs(title = "Distribution of the Distance_from_Home column") + xlab("Value") + ylab("Count") + theme(plot.title = element_text(hjust = 0.5, vjust = 1, face = "bold"))
```



На графику можемо da видimo да се највише појављује вредност *Near* у колони *Distance_from_Home*. *NA* вредности ћемо да попунимо том вредношћу јер је најзаступљенија. Та промена неће да утиче толико на податке јер имамо веома мало недостајућих података, а пуно *Near* вредности.

```r
{r}
dataset$Distance_from_Home[is.na(dataset$Distance_from_Home)] <- "Near"
table(dataset$Distance_from_Home)
```

```
     Far Moderate     Near 
     658     1997     3946 
```

Када смо попунили недостајуће вредности, расподела колоне *Distance_from_Home* изгледа као што видимо на слици изнад. Број вредности у *Near* категорији се повећао за број недостајућих вредности.

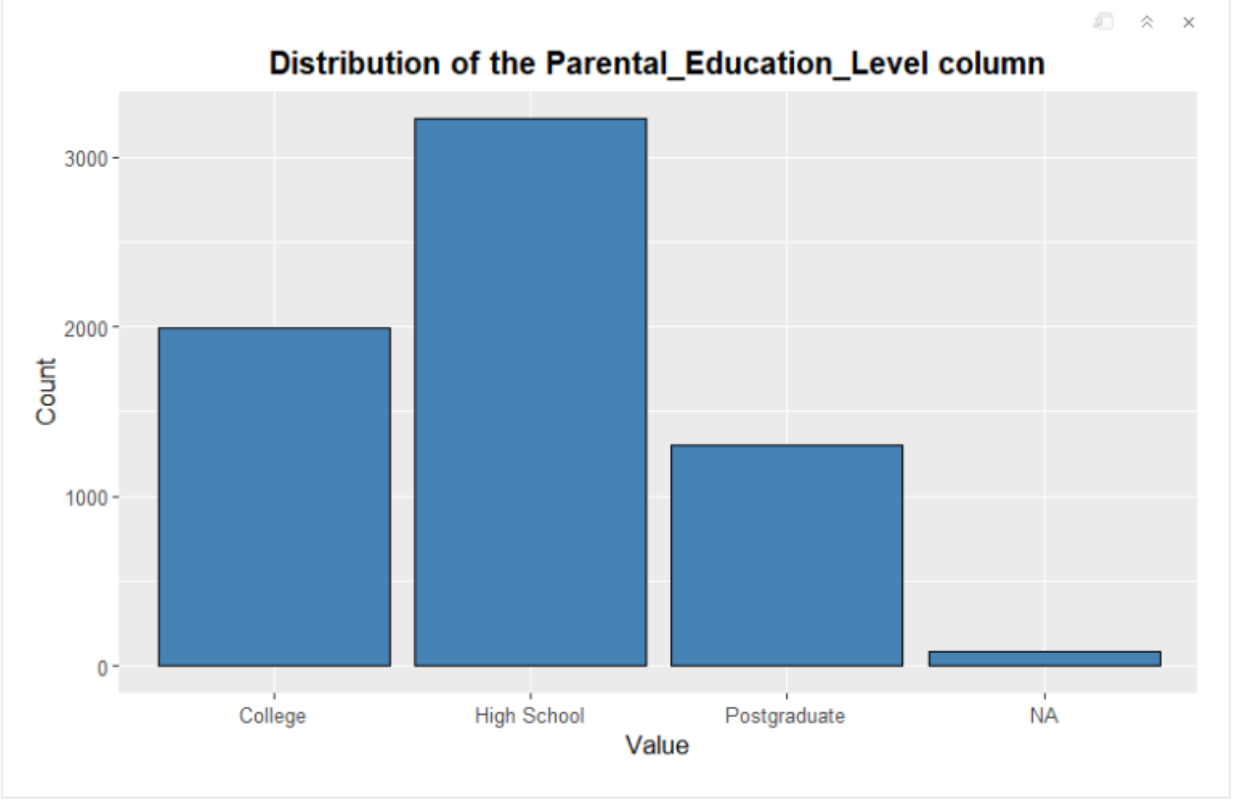
3.2 Колона *Parental_Education_Level*

Сада ћемо да поновимо поступак попуњавања недостајућих вредности и у колони *Parental_Education_Level*. Наћи ћемо најучесталију вредност и након тога приказати графички расподелу колоне и видећемо како ћемо да решимо проблем *NA* вредности у тој колони.

```r
{r}
table(dataset$Parental_Education_Level)
```

```
     College  High School Postgraduate 
        1989         3222         1304 
```

```{r}
ggplot(dataset, aes(x = Parental_Education_Level)) + geom_bar(fill = "steelblue",
color = "black") + labs(title = "Distribution of the Parental_Education_Level
column") + xlab("Value") + ylab("Count") + theme(plot.title = element_text(hjust =
0.5, vjust = 1, face = "bold"))
```



Видимо да се у овој колони највише понавља вредност *High School*, тако да ћемо наше недостајуће вредности у овој колони заменити том вредношћу. Због малог броја недостајућих вредности, ова промена неће значајно утицати на податке у нашем *dataset-y*.

```{r}
dataset$Parental_Education_Level[is.na(dataset$Parental_Education_Level)] <- "High
School"
table(dataset$Parental_Education_Level)
```

```
     College  High School Postgraduate
        1989         3308         1304
```

Када смо попунили недостајуће вредности у колони *Parental_Education_Level*, расподела колоне изгледа као на слици изнад. Број вредности у *High School* категорији се повећао за број недостајућих вредности.

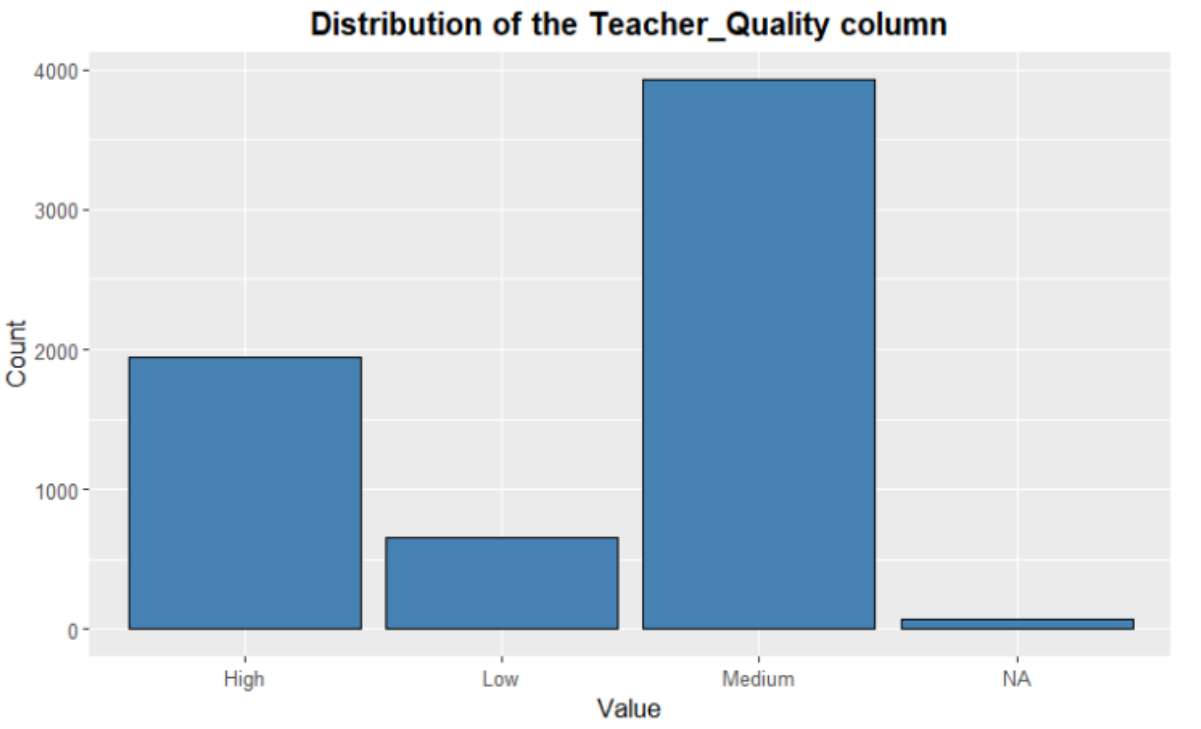
3.3 Колона *Teacher_Quality*

Остало нам је да решимо проблем недостајућих вредности у колони *Teacher_Quality*. Видећемо графички и бројчано каква је раподела ове колоне и на основу тога ћемо одлучити шта даље да радимо са недостајућим вредностима.

```r
table(dataset$Teacher_Quality)
```

```
High    Low Medium
1947    657   3925
```

```r
ggplot(dataset, aes(x = Teacher_Quality)) + geom_bar(fill = "steelblue", color = "black") + labs(title = "Distribution of the Teacher_Quality column") + xlab("Value") + ylab("Count") + theme(plot.title = element_text(hjust = 0.5, vjust = 1, face = "bold"))
```



Можемо видети да има доста више вредности *Medium* у колони од осталиих тако да ћемо поља са недостајућим вредностима да попунимо овом вредношћу. Тиме нећемо толико да утичемо на податке јер имамо мало недостајућих вредности у односу на остатак података.

```r
{r}
dataset$Teacher_Quality[is.na(dataset$Teacher_Quality)] <- "Medium"
table(dataset$Teacher_Quality)
```

```
High    Low Medium
1947    657   3997
```

Када смо попунили недостајуће податке одређеном вредношћу, расподела колоне *Teacher_Quality* изгледа као на слици изнад. Број вредности у *Medium* категорији се повећао за број недостајућих вредности.

4 Опис података

У овом поглављу ћемо се детаљније посветити анализи структуре података, представићемо атрибуте *dataset-a*, њихове типове и распон вредности, како бисмо боље разумели наш скуп података и касније извршили предикцију.

4.1 Описи нумеричких колона

Нумеричке колоне у *dataset-u* ћемо анализирати како би се стекло боље разумевање расподеле вредности и идентификовале евентуалне аномалије у подацима. За сваку колону ћемо испитати које су минималне и максималне вредности, квартили (25%, 50% и 75%), као и средња вредност и стандардна девијација.

4.1.1 Нумеричка колона *Hours_Studied*

Hours_Studied представља број сати које је студент провео учећи недељно.

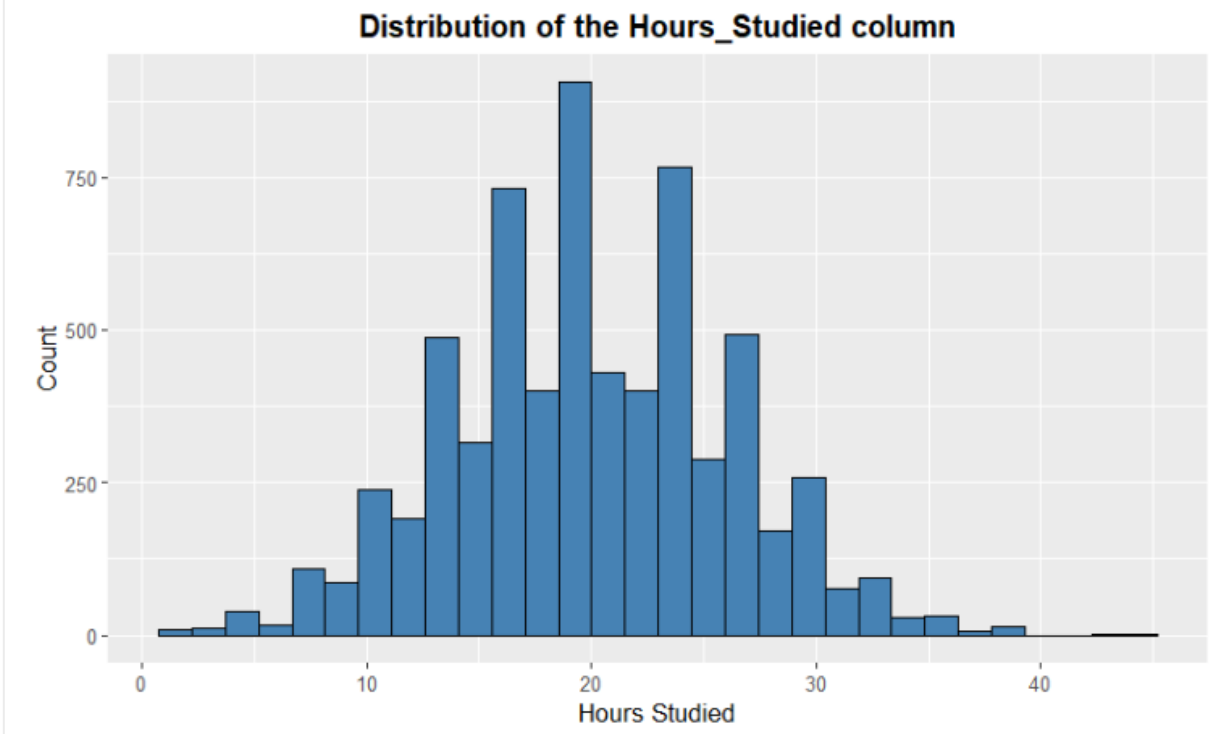
Функција *summary* нам приказује основне статистике о подацима из колоне *Hours Studied* као што су: минимална вредност, максимална вредност, просечна вредност, медијана (вредност која дели податке на два једнака дела, где је 50% података испод те вредности), први квантил (25% података испод те вредности) и трећи квантил (75% података испод те вредности).

```r
summary(dataset$Hours_Studied)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1.00   16.00   20.00   19.98   24.00   44.00 
```

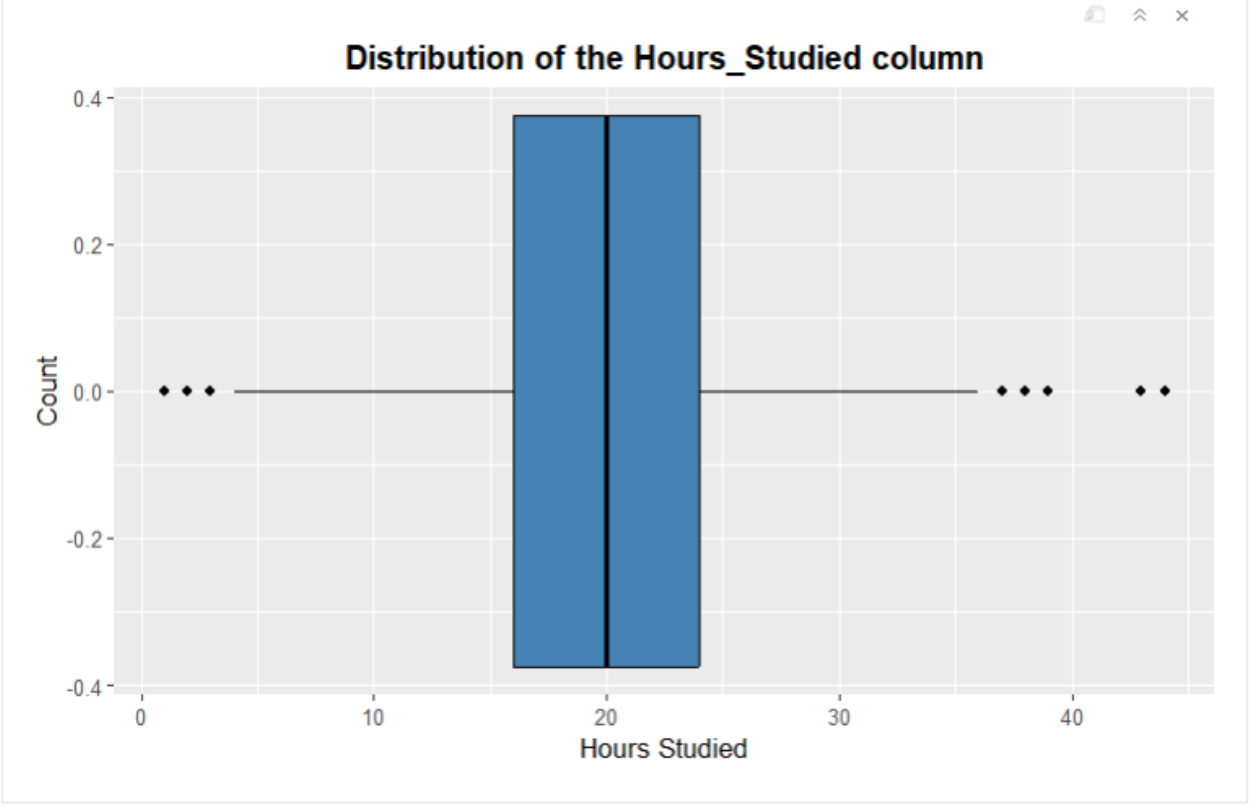
Ови подаци су релативно равномерно распоређени, али максимална вредност (44) изгледа као да одступа, што би могло бити вредно додатне анализе, као потенцијални *outlier*.

```r
ggplot(dataset, aes(x = Hours_Studied)) + geom_histogram(fill = "steelblue", color
= "black", bins = 30) + labs(title = "Distribution of the Hours_Studied column") +
xlab("Hours Studied") + ylab("Count") + theme(plot.title = element_text(hjust = 0.5
, vjust = 1, face = "bold"))
```



На основу хистограма можемо да видимо да је већина студената учила између 15 и 25 сати недељно, што значи да су подаци сконцентрисани у средњем делу. Мали број података одскаче од уобичајених вредности. Сада ћемо преко *boxplot-a* да прикажемо ове податке како бимсо утврдили постојање *oulier-a.*

```r
{r}
ggplot(dataset, aes(x = Hours_Studied)) +
  geom_boxplot(fill = "steelblue", color = "black") +
  labs(title = "Distribution of the Hours_Studied column") +
  xlab("Hours Studied") +
  ylab("Count") +
  theme(plot.title = element_text(hjust = 0.5, vjust = 1, face = "bold"))
```



Постоје *outlier-i* када је *Hours Studied* мањи од 4 и већи од 36. Међутим, уопште није немогуће да неки студенти проводе толико времена учећи и вероватно ће те вредности имати утицај на резултате тестова. Сматрамо да ће нам ове вредности бити од помоћи у креирању модела.

Користећи Z-score методу можемо идентификовати тачне вредности *outlier-a*.

```r
{r}
z_scores <- (dataset$Hours_Studied - mean(dataset$Hours_Studied)) / sd
(dataset$Hours_Studied)
outliers_z <- dataset$Hours_Studied[abs(z_scores) > 3]
print(outliers_z)
```

```
 [1] 43  1 38  2 39 39  2  2 39 38 38 44  2  2 39 39  2  1  1 38 39 38 38 39 38
```

Резултати су показали следеће вредности: 43, 1, 38, 2, 39, 44, као и неколико других поновљених вредности. Ове вредности су значајно одвојене од просека, што указује на то да представљају екстремне случајеве у овом скупу података.

4.1.2 Нумеричка колона *Attendance*

Attendance представља проценат долазака на часове.

Помовићемо поступак као и са претходном колоном и позвати функцију *summary* за колону *Attendance.*

```{r}
summary(dataset$Attendance)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  60.00   70.00   80.00   79.98   90.00  100.00 
```

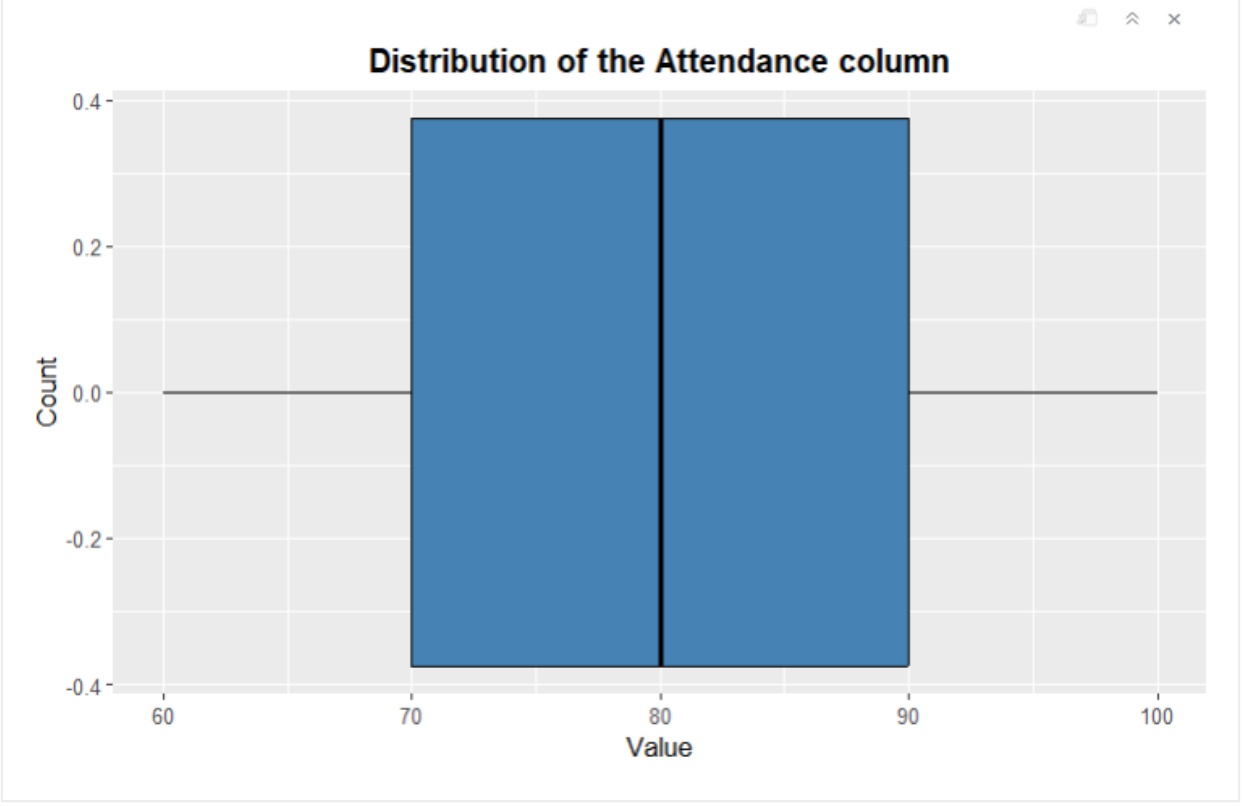
Према овим подацима можемо да закључимо да је присуство студената углавном високо, са минималном вредношћу од 60% и максималном од 100%. Медијана је 80%, што значи да половина студената има присуство веће од 80%. Просечна вредност је такође близу медијане, 79.98%, што указује на релативно симетричну дистрибуцију, што можемо да докажемо и помоћу функције *skewness* која израчунава асиметрију (секвенс) дистрибуције. Секвенс близу 0 (између -0.5 и 0.5) указује на симетричност.

```{r}
skewness_value <- skewness(dataset$Attendance)
print(skewness_value)
```

```
[1] 0.01336168
```

С обзиром на то да је *skewness* практично нула, можемо закључити да подаци о присуству студената немају екстремне вредности и да су равномерно распоређени око медијане. Ипак ћемо се уверити у то и преко графика, на пример преко *boxplot-a.*

```r
{r}
ggplot(dataset, aes(x = Attendance)) + geom_boxplot(fill = "steelblue", color =
"black") + labs(title = "Distribution of the Attendance column") + xlab("Value") +
ylab("Count") + theme(plot.title = element_text(hjust = 0.5, vjust = 1, face =
"bold"))
```



Овде заиста можемо да се уверимо да нема вредности које одскачу.

4.1.3 Нумеричка колона *Sleep_Hours*

Sleep_Hours представља просечан број сати сна по ноћи.

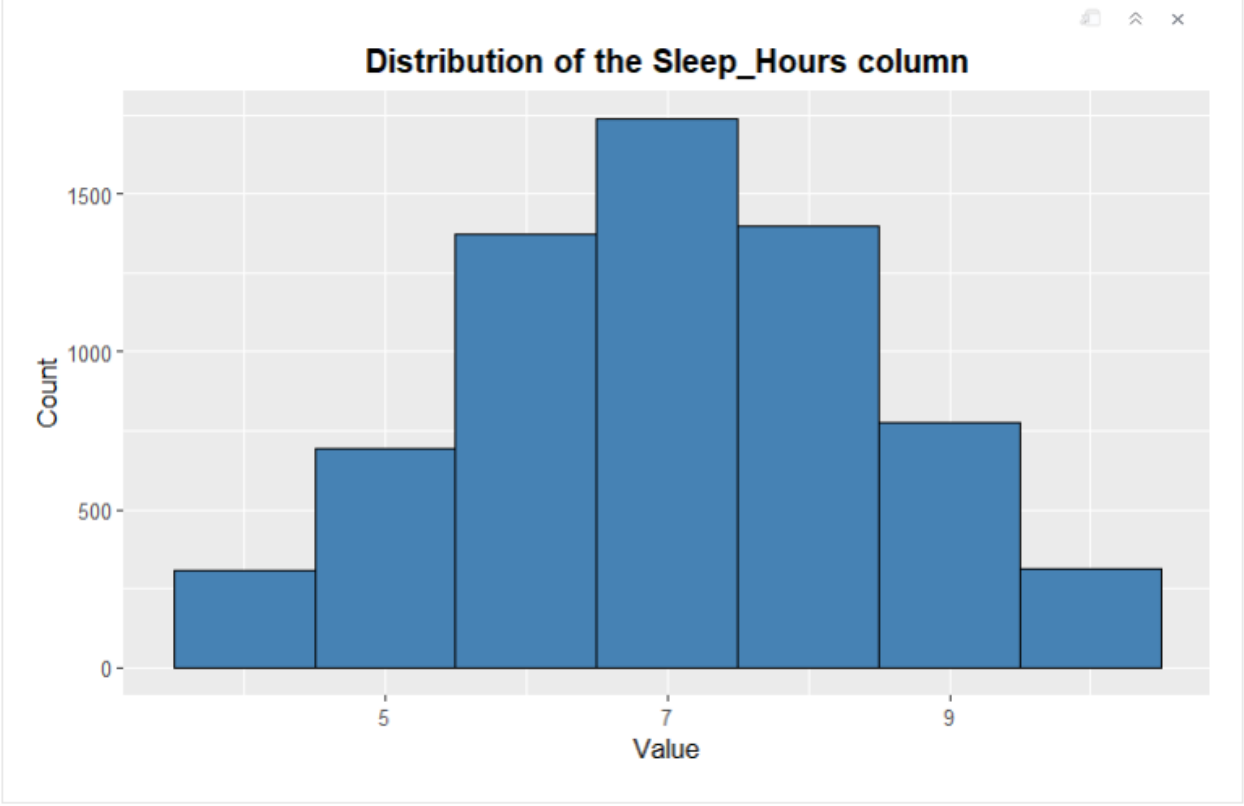
Помовићемо поступак као и са претходном колоном и позвати функцију *summary* за колону *Sleep_Hours.*

```r
{r}
summary(dataset$Sleep_Hours)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.000   6.000   7.000   7.029   8.000  10.000
```

Просек броја сати спавања је ~7 сати, што указује да већина има уобичајену количину сна, поготово за студенте. Слично као и за колону *Attendance*, делује да је симетрична расподела података. Проверићемо то и пркео *histogram-a.*

```{r}
ggplot(dataset, aes(x = Sleep_Hours)) +
  geom_histogram(fill = "steelblue", color = "black", binwidth = 1) +
  labs(title = "Distribution of the Sleep_Hours column") +
  xlab("Value") +
  ylab("Count") +
  theme(plot.title = element_text(hjust = 0.5, vjust = 1, face = "bold"))
```



Већина података се концентрише око 6-8 сати, што представља здрав опсег сна. Студенти који спавају 4 сата или 10 сати могу се сматрати изолованим случајевима. Ове вредности нису превише честе, али су у границама реалног понашања. Касније ћемо видети да ли и какав утицај сан може да има на перформансе на испитима.

Следећи корак јесте рачунање стандардне девијације која је важна статистичка мера и помаже нам да разумемо расподелу података у односу на средњу вредност.

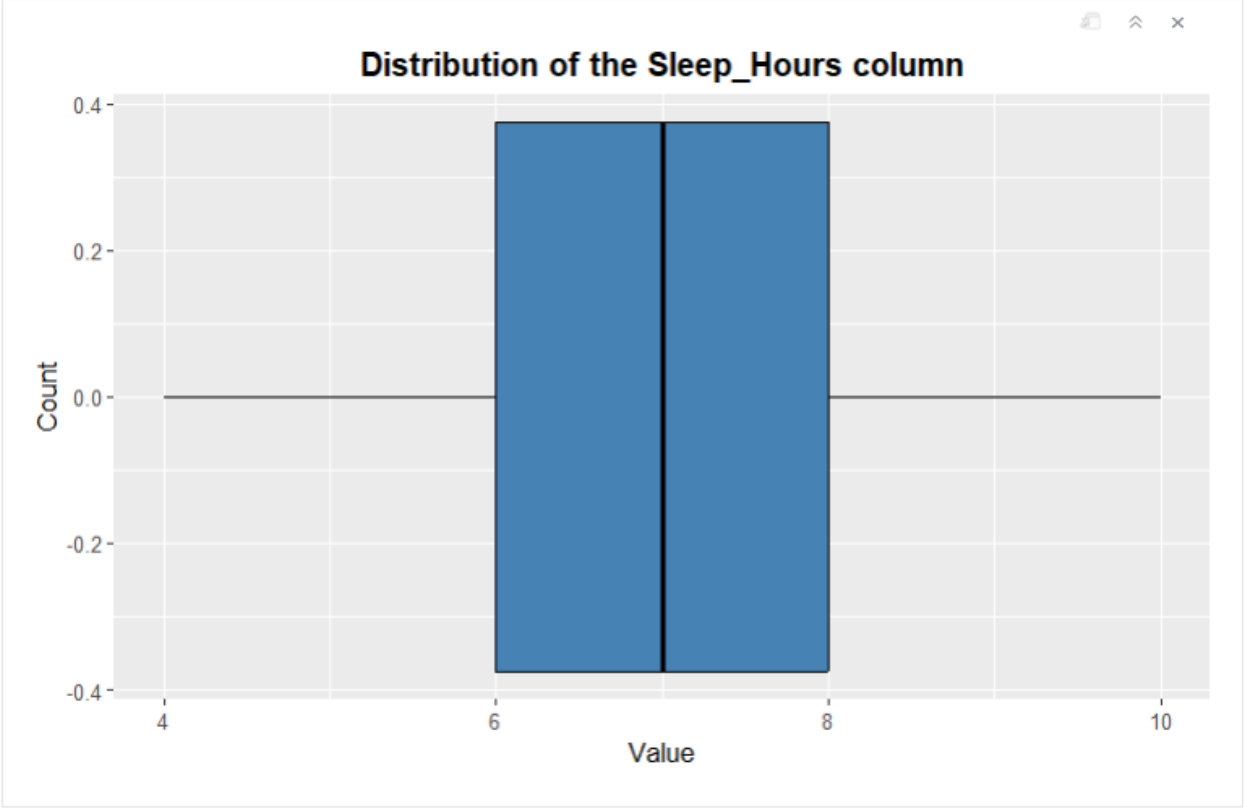
```{r}
std_dev <- sd(dataset$Sleep_Hours)
print(std_dev)
```

```
[1] 1.468435
```

Стандардна девијација износи 1.468 сати. Ова вредност указује на умерену варијабилност у расподели података. У овом случају, студенти спавају у широком распону сати, што

може указивати на различите навике спавања. Вредности које значајно одступају од овог просека могу бити предмет даљег истраживања, с тога ћемо користити *boxplot*.

```{r}
ggplot(dataset, aes(x = Sleep_Hours)) + geom_boxplot(fill = "steelblue", color = "black") + labs(title = "Distribution of the Sleep_Hours column") + xlab("Value") + ylab("Count") + theme(plot.title = element_text(hjust = 0.5, vjust = 1, face = "bold"))
```



Са овог графика можемо закључити да нема вредности које одскачу.

4.1.4 Нумеричка колона *Previous_Scores*

Previous_Scores представља оцене са претходних испита.

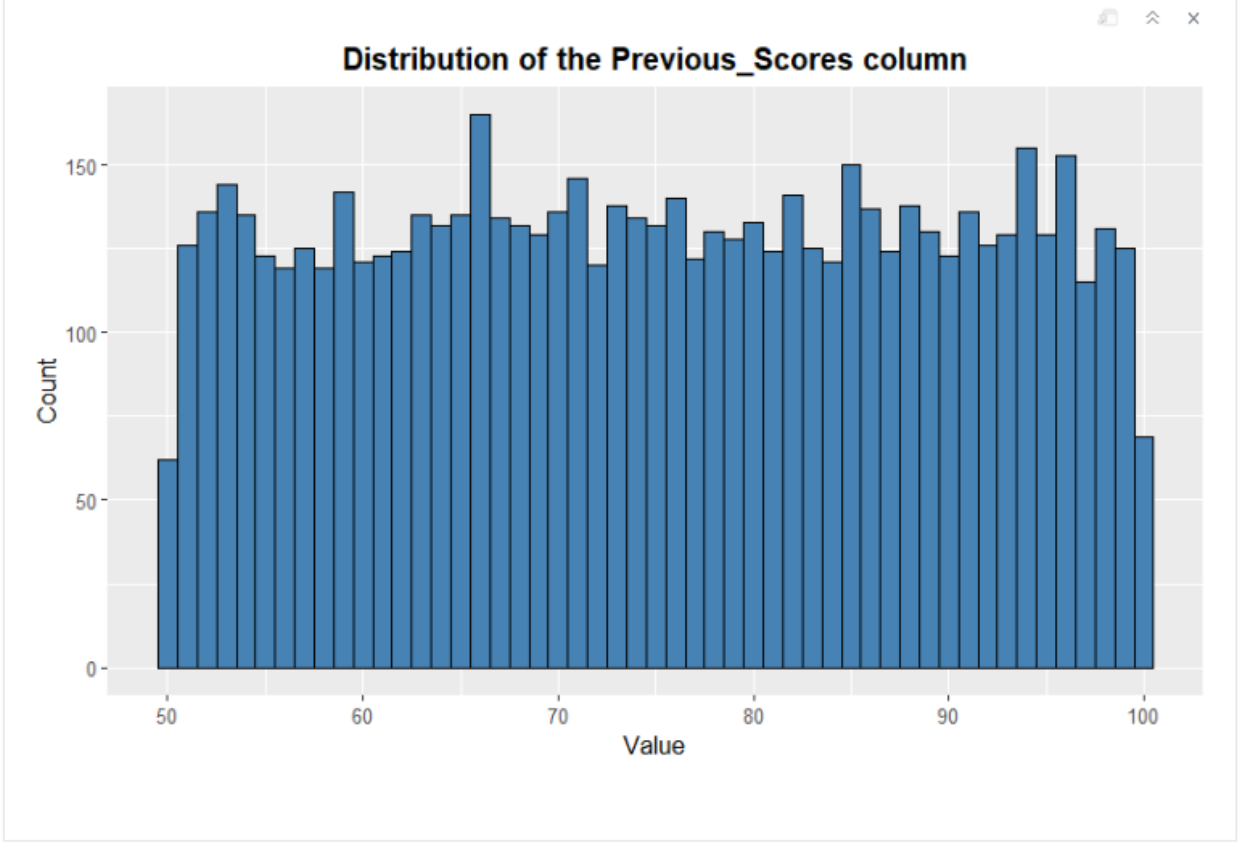
Поновићемо поступак као и са претходном колоном и позвати функцију *summary* за колону *Previous_Scores*.

```{r}
summary(dataset$Previous_Scores)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  50.00   63.00   75.00   75.08   88.00  100.00 
```

Укупно, расподела *Previous_Scores* изгледа као да је уравнотежена, са већином студената у распону од 63 до 88, а просек и медијана указују на то да је већина студената остварила добар резултат. Ипак, постоје и студенти са значајно нижим резултатима, најнижи резултат је 50.00, што може указивати на потребу за допунском наставом.

```{r}
ggplot(dataset, aes(x = Previous_Scores)) + geom_histogram(fill = "steelblue",
color = "black", binwidth = 1) + labs(title = "Distribution of the Previous_Scores
column") + xlab("Value") + ylab("Count") + theme(plot.title = element_text(hjust =
0.5, vjust = 1, face = "bold"))
```



Са графика изнад можемо да закључимо да су вредности приближно равномерно распоређене, свака вредност има сличну фреквенцију. Уочљива разлика је једино за вредности које су екстремно ниске или екстремно високе (за вредности 50 и 100), али нема доминантних вредности или великих одступања у појединим сегментима. Пошто је распон вредности велики вероватно је стандардна девијација висока.

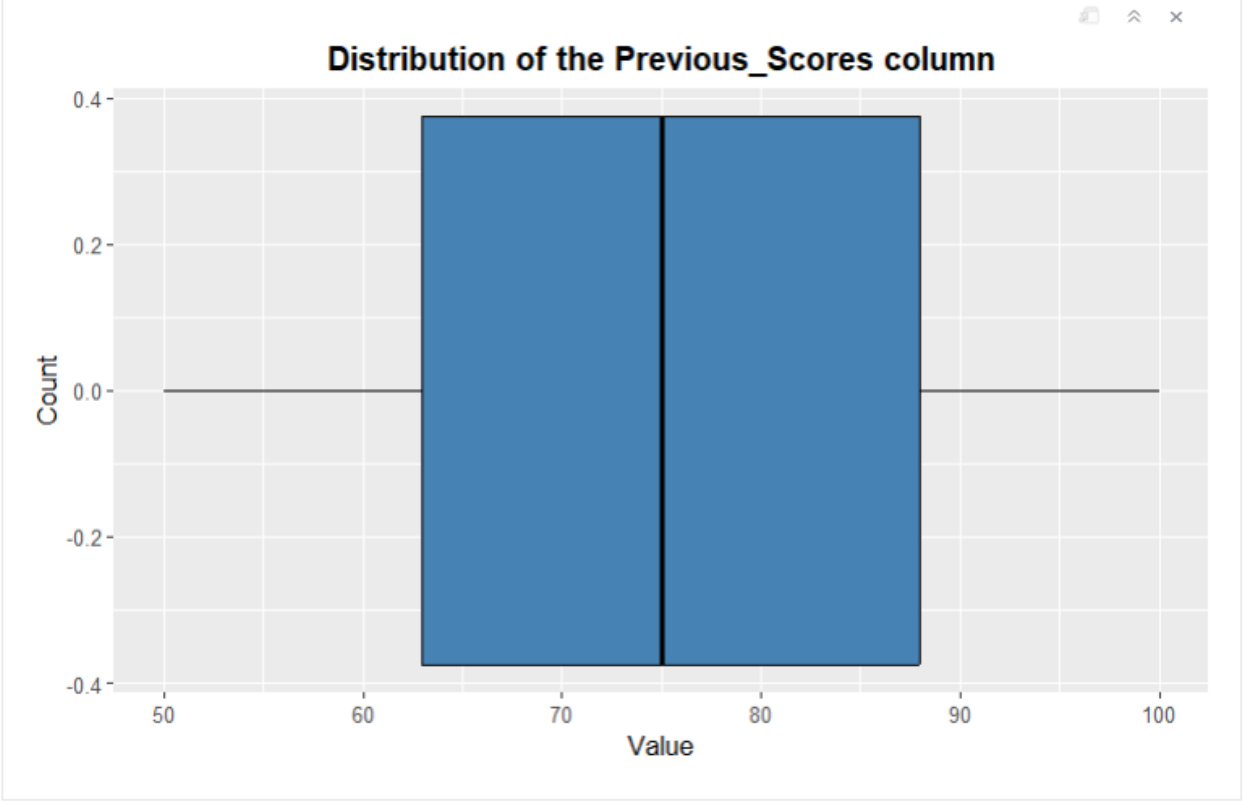
```{r}
std_dev <- sd(dataset$Previous_Scores)
print(std_dev)
```

```
[1] 14.39701
```

Као што смо и претпоставили вредност је висока, чак и у односу на вредност стандардне девијације за колону *Sleep_Hours* која је износила ~1.4.

Пошто су све вредности једнако заступљене, мислимо да нема екстремних вредности. Али ћемо свакако приказати то и на *boxplot-u*.

```{r}
ggplot(dataset, aes(x = Previous_Scores)) + geom_boxplot(fill = "steelblue", color
= "black") + labs(title = "Distribution of the Previous_Scores column") + xlab
("Value") + ylab("Count") + theme(plot.title = element_text(hjust = 0.5, vjust = 1,
face = "bold"))
```



Како смо и очекивали, нема *oulier-a.*

4.1.5 Нумеричка колона *Tutoring_Sessions*

Tutoring_Sessions представља број посећених допунских часова месечно.

Поновићемо поступак као и са претходном колоном и позвати функцију *summary* за колону *Tutoring_Sessions.*

```r
summary(dataset$Tutoring_Sessions)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   1.000   1.000   1.495   2.000   8.000 
```

Дистрибуција података показује да већина студената има мали број долазака на допунске часове, са 25% оних који немају ниједан долазак. Само неколико студената има већи број долазака, што указује на неуједначеност у коришћењу туторства.

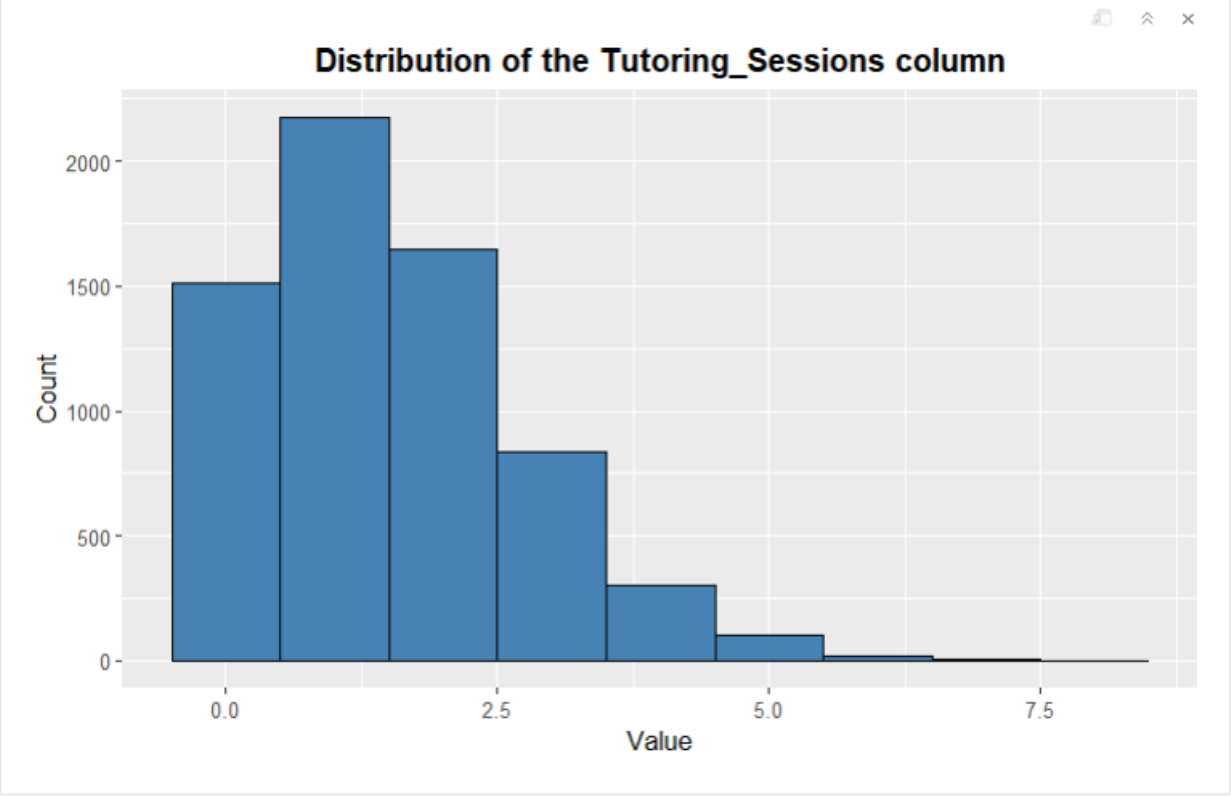
Израчунаћемо *skewness* да бисмо видели да ли су подаци у колони *Tutoring_Sessions* симетрични.

```r
skewness_value = skewness(dataset$Tutoring_Sessions)
print(skewness_value)
```

```
[1] 0.8145231
```

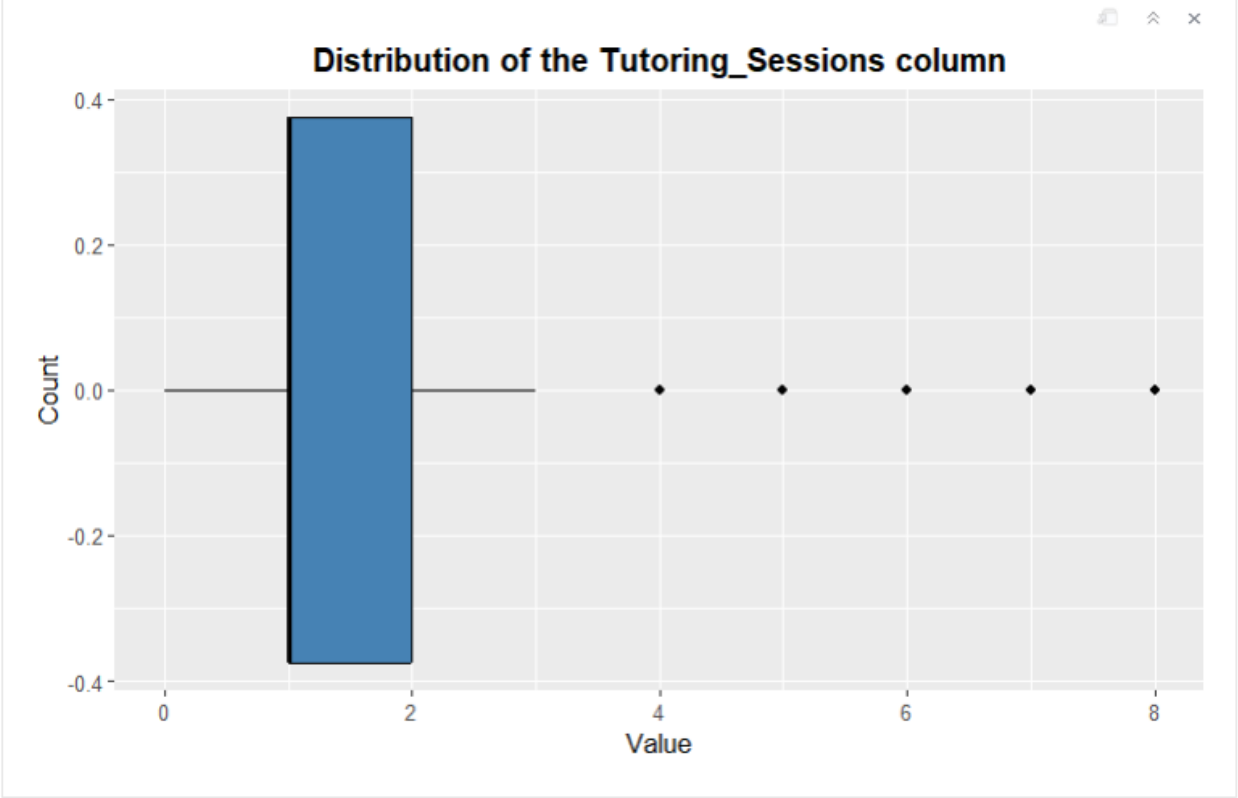
Пошто резултат ове функције мора да буде између -0.5 и 0.5 да би се подаци сматрали симетричним, у овом примеру можемо видети да подаци нису симетрични јер је резултат функције 0.81. Сада ћемо преко *histogram-a* визуелно приказати расподелу података у колони *Tutoring_Sessions.*

```r
{r}
ggplot(dataset, aes(x = Tutoring_Sessions)) + geom_histogram(fill = "steelblue",
color = "black", binwidth = 1) + labs(title = "Distribution of the
Tutoring_Sessions column") + xlab("Value") + ylab("Count") + theme(plot.title =
element_text(hjust = 0.5, vjust = 1, face = "bold"))
```



Са слике се јасно види да постоји солидан број студената (чак 25%) који нису ишли на допунске часове код тутора. Такође се види да је мали број студената ишло на више од 3 часа месечно. Остали број студената је углавном имао 1 до 2 часа месечно. Ова колона би била погодна за *Feature Engineering*, који ћемо одрадити касније.

```r
ggplot(dataset, aes(x = Tutoring_Sessions)) + geom_boxplot(fill = "steelblue",
color = "black") + labs(title = "Distribution of the Tutoring_Sessions column") +
xlab("Value") + ylab("Count") + theme(plot.title = element_text(hjust = 0.5, vjust
= 1, face = "bold"))
```



Преко *boxplot-a* јасно видимо које вредности одскачу од просека (4, 5, 6, 7, 8).

4.1.6 Нумеричка колона *Physical_Activity*

Physical_Activity представља просечан број сати физичке активности недељно.

Поновићемо поступак као и са претходном колоном и позвати функцију *summary* за колону *Physical_Activity*.

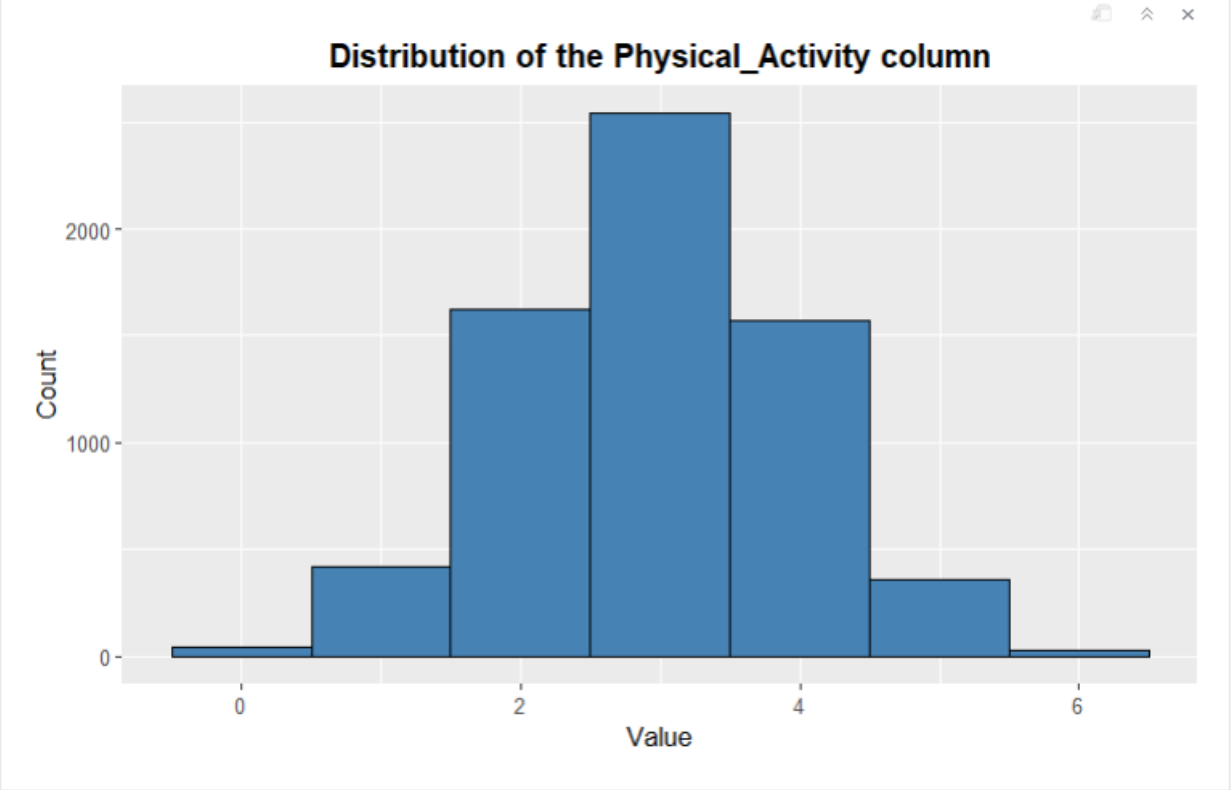
```r
summary(dataset$Physical_Activity)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   2.000   3.000   2.968   4.000   6.000
```

Можемо да видимо да је минимална вредност је 0, што значи да неки људи уопште немају физичку активност, али медијана (3 сата) указује да је пола студената активно најмање 3 сата недељно. Средња вредност (2.968) је слична медијани, што значи да су подаци

релативно симетрични и да не постоје екстремни *oulier-i* који би значајно утицали на просечне вредности.

```{r}
ggplot(dataset, aes(x = Physical_Activity)) + geom_histogram(fill = "steelblue",
color = "black", binwidth = 1) + labs(title = "Distribution of the
Physical_Activity column") + xlab("Value") + ylab("Count") + theme(plot.title =
element_text(hjust = 0.5, vjust = 1, face = "bold"))
```



Већина података је сконцентрисана на око 3 сата физичке активности недељно. Има студената који или немају никакву или имају високу физичку активност, али је број таквих студената знатно мањи у односу на остатак. Како су подаци врло симетрично распоређени, нема потребе да приказујемо *boxplot* у потрази за *outlier-има*.

4.1.7 Нумеричка колона *Exam_Score*

Exam_Score представља оцену завршног испита.

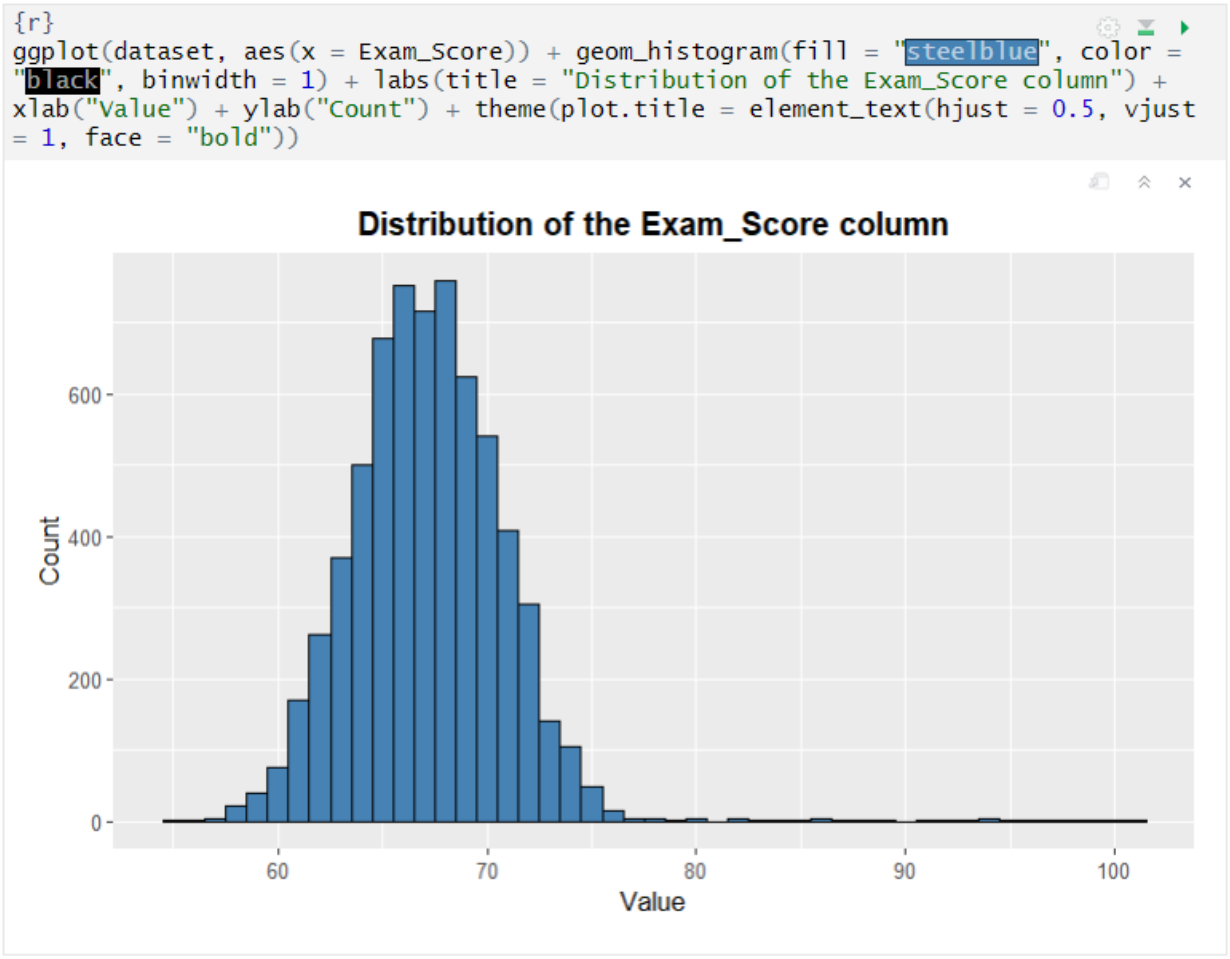
Поновићемо поступак као и са претходном колоном и позвати функцију *summary* за колону *Exam_Score*.

```r
summary(dataset$Exam_Score)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  55.00   65.00   67.00   67.24   69.00  101.00 
```

Медијана износи 67 и близу је средње вредности од 67.24, што указује на то да су резултати распоређени релативно симетрично. Први квартил и трећи квартил показују да је половина студената постигла резултате између 65 и 69. Дакле, већина студената је постигла резултате близу просека, а мали број студената је постигао резултате који су бољи од просека.

```r
ggplot(dataset, aes(x = Exam_Score)) + geom_histogram(fill = "steelblue", color = "black", binwidth = 1) + labs(title = "Distribution of the Exam_Score column") + xlab("Value") + ylab("Count") + theme(plot.title = element_text(hjust = 0.5, vjust = 1, face = "bold"))
```



Као што смо и претпоставили, на основу функције *summary*, мали број студената је постигао резултате боље од 75.

4.2 Описи категоријских колона

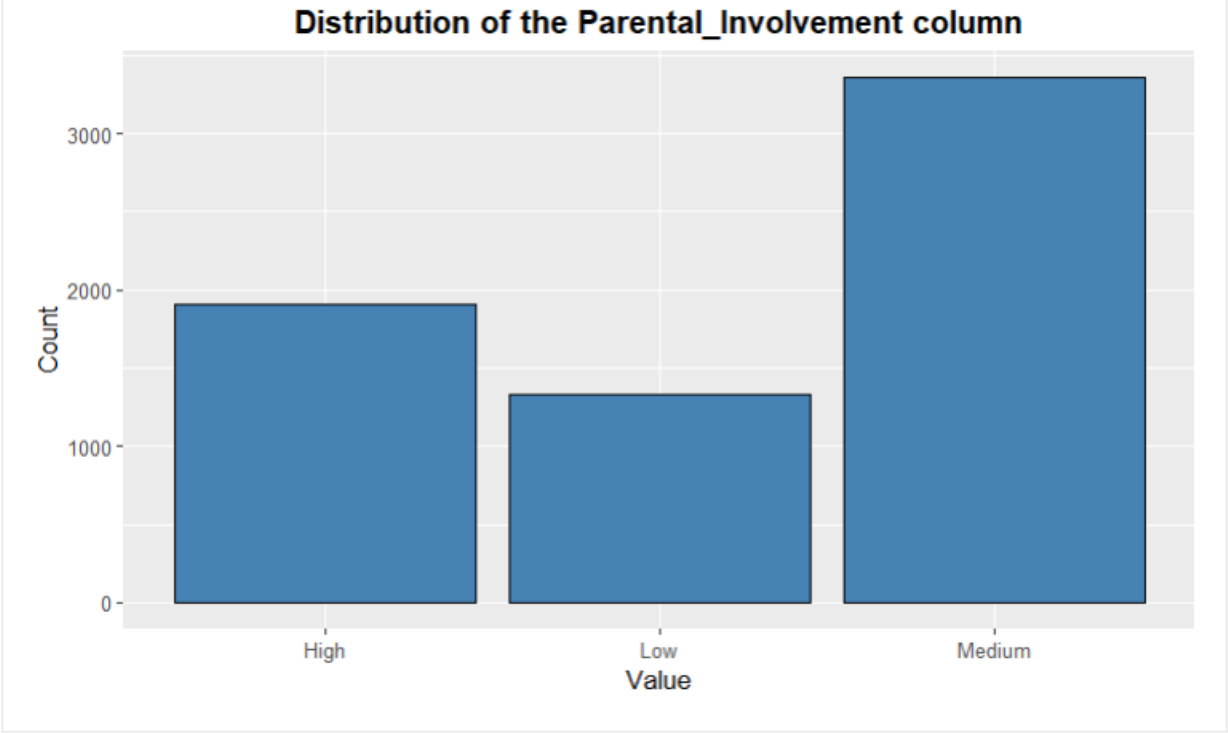
4.2.1 Категоријска колона *Parental_Involvement*

Parental_Involvement представља ниво укључености родитеља у образовање студента *(Low, Medium, High)*. Приказаћемо табелу учестаности да бисмо видели колико имамо категорија за ову колону.

```{r}
xtabs(~Parental_Involvement, dataset)
```

```
Parental_Involvement
  High    Low Medium
  1907   1334   3360
```

```{r}
ggplot(dataset, aes(x = Parental_Involvement)) + geom_bar(fill = "steelblue", color = "black") + labs(title = "Distribution of the Parental_Involvement column") + xlab("Value") + ylab("Count") + theme(plot.title = element_text(hjust = 0.5, vjust = 1, face = "bold"))
```



Са графика можемо видети да је укљученост родитеља у образовање студената најчешће средње (*Medium*). Касније ћемо видети да ли то има неки утицај на колону која се предвиђа.

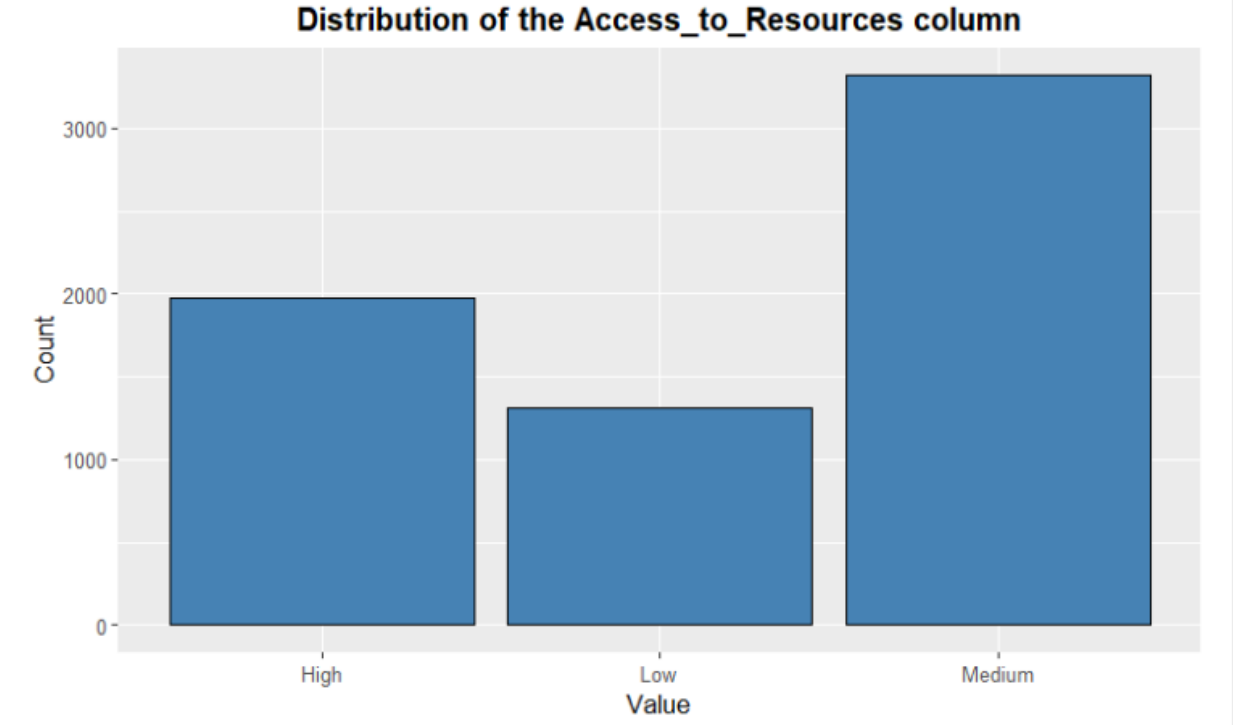
4.2.2 Категоријска колона *Access_to_Resources*

Access_To_Resources представља доступност образовних ресурса *(Low, Medium, High)*. Приказаћемо табелу учестаности да бисмо видели колико имамо категорија за ову колону.

```
{r}
xtabs(~Access_to_Resources, dataset)
```

```
Access_to_Resources
  High    Low Medium
  1972   1312   3317
```

```
{r}
ggplot(dataset, aes(x = Access_to_Resources)) + geom_bar(fill = "steelblue", color
= "black") + labs(title = "Distribution of the Access_to_Resources column") + xlab
("Value") + ylab("Count") + theme(plot.title = element_text(hjust = 0.5, vjust = 1,
face = "bold"))
```



Са графика можемо видети да је доступност образовних ресурса за студенте углавном средње (*Medium*).

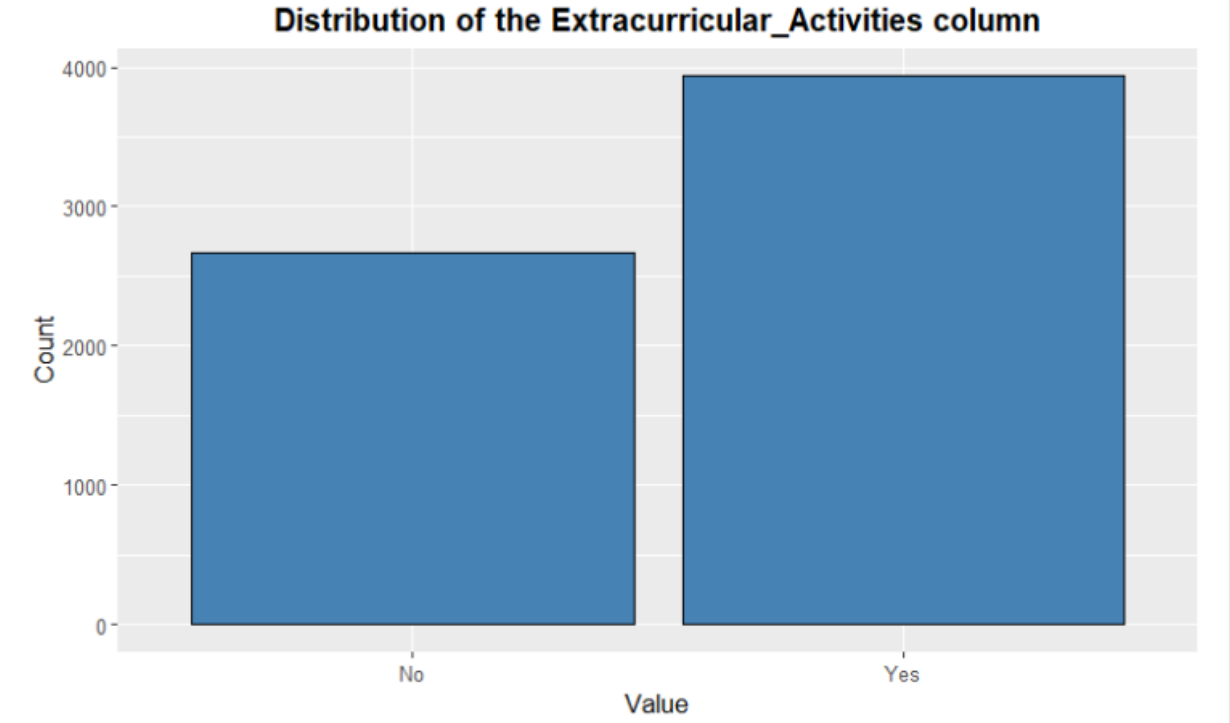
4.2.3 Категоријска колона *Extracurricular_Activities*

Extracurricular_Activities представља учешће у ваннаставним активностима *(Yes, No)*. Приказаћемо табелу учестаности да бисмо видели колико имамо категорија за ову колону.

```{r}
xtabs(~Extracurricular_Activities, dataset)
```

```
Extracurricular_Activities
  No  Yes
2667 3934
```

```{r}
ggplot(dataset, aes(x = Extracurricular_Activities)) + geom_bar(fill = "steelblue", color = "black") + labs(title = "Distribution of the Extracurricular_Activities column") + xlab("Value") + ylab("Count") + theme(plot.title = element_text(hjust = 0.5, vjust = 1, face = "bold"))
```



Са графика можемо видети да је више студената имало ваннаставне активности.

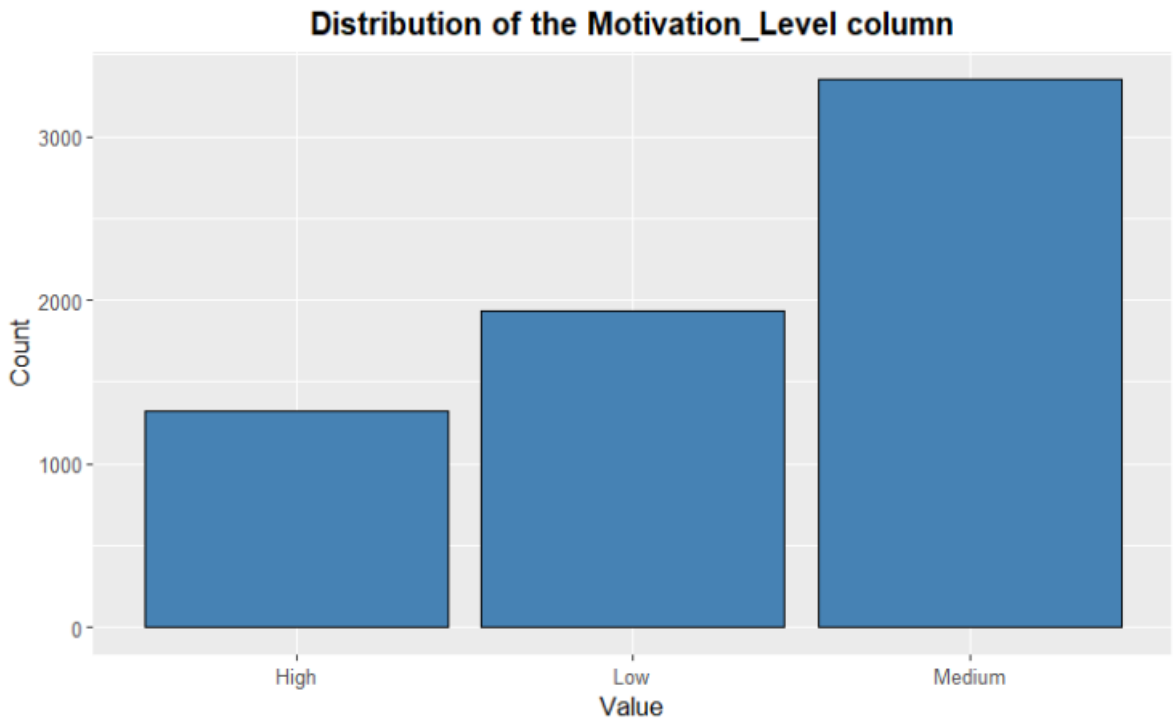
4.2.4 Категоријска колона *Motivation_Level*

Motivation_Level представља ниво мотивације студента *(Low, Medium, High)*. Приказаћемо табелу учестаности да бисмо видели колико имамо категорија за ову колону.

```r
{r}
xtabs(~Motivation_Level, dataset)
```

```
Motivation_Level
  High    Low Medium
  1319   1935   3347
```

```r
{r}
ggplot(dataset, aes(x = Motivation_Level)) + geom_bar(fill = "steelblue", color =
"black") + labs(title = "Distribution of the Motivation_Level column") + xlab
("Value") + ylab("Count") + theme(plot.title = element_text(hjust = 0.5, vjust = 1,
face = "bold"))
```



Са графика можемо видети да је мотивација студената углавном средње вредности (*Medium*).

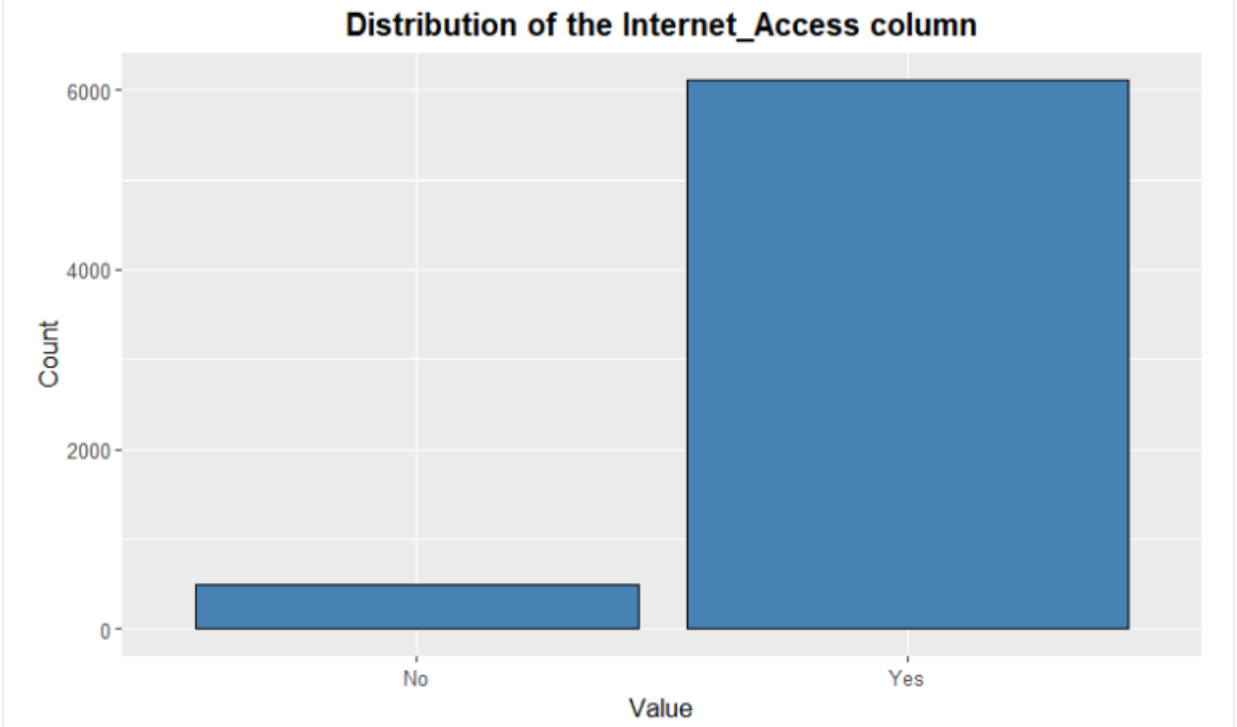
4.2.5 Категоријска колона *Internet_Access*

Internet_Access представља доступност приступа интернету *(Yes, No)*. Приказаћемо табелу учестаности да бисмо видели колико имамо категорија за ову колону.

```r
{r}
xtabs(~Internet_Access, dataset)
```

```
Internet_Access
  No  Yes
 498 6103
```

```r
{r}
ggplot(dataset, aes(x = Internet_Access)) + geom_bar(fill = "steelblue", color =
"black") + labs(title = "Distribution of the Internet_Access column") + xlab
("Value") + ylab("Count") + theme(plot.title = element_text(hjust = 0.5, vjust = 1,
face = "bold"))
```



Као што је и очекивано, велики број студената има приступ интернету наспрам веома малог броја оних који немају. Питамо се колико ће значајан фактор ово бити касније.

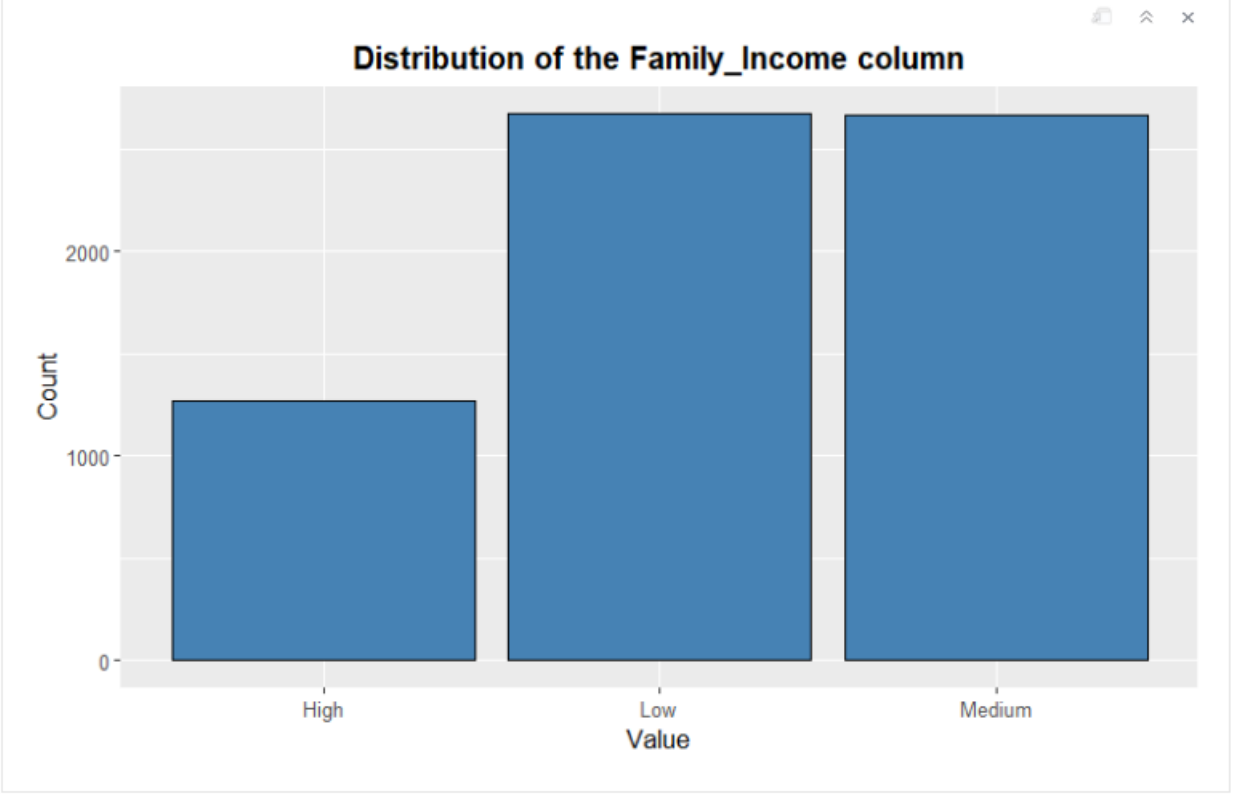
4.2.6 Категоријска колона *Family_Income*

Family_Income представља ниво породичног прихода *(Low, Medium, High)*. Приказаћемо табелу учестаности да бисмо видели колико имамо категорија за ову колону.

```r
{r}
xtabs(~Family_Income, dataset)
```

```
Family_Income
  High    Low Medium
  1268   2670   2663
```

```r
{r}
ggplot(dataset, aes(x = Family_Income)) + geom_bar(fill = "steelblue", color =
"black") + labs(title = "Distribution of the Family_Income column") + xlab("Value")
+ ylab("Count") + theme(plot.title = element_text(hjust = 0.5, vjust = 1, face =
"bold"))
```



Са графика можемо закључити да је у односу на *Low* и *Medium, High* вредност у колони *Family_Income* је доста мања од остале две категорије, што значи да постоји мањи број студената који имају висок породични приход.

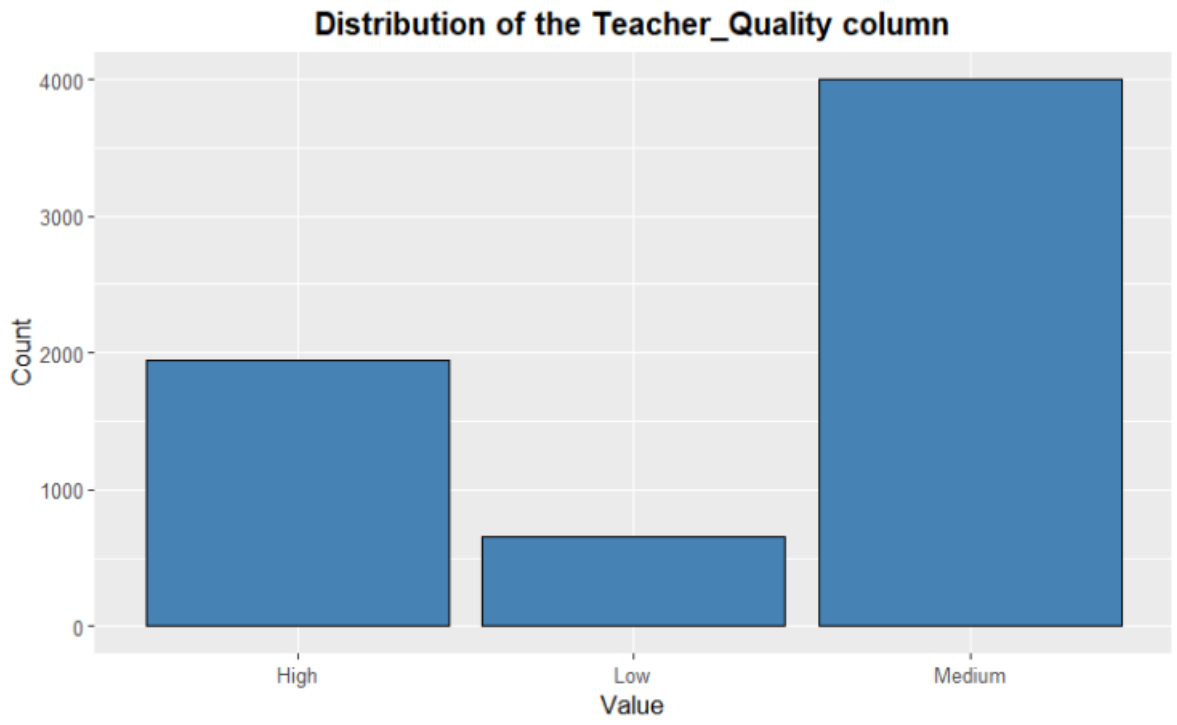
4.2.7 Категоријска колона *Teacher_Quality*

Teacher_Quality представља квалитет наставника *(Low, Medium, High).* Приказаћемо табелу учестаности да бисмо видели колико имамо категорија за ову колону.

```r
{r}
xtabs(~Teacher_Quality, dataset)
```

```
Teacher_Quality
  High    Low Medium
  1947    657   3997
```

```r
{r}
ggplot(dataset, aes(x = Teacher_Quality)) + geom_bar(fill = "steelblue", color =
"black") + labs(title = "Distribution of the Teacher_Quality column") + xlab
("Value") + ylab("Count") + theme(plot.title = element_text(hjust = 0.5, vjust = 1,
face = "bold"))
```



Видимо да највише има професора који имају средњу оцену (*Medium*), док је број оних који су лоше оцењени (*Low*) најмањи, што је добро. Видећемо колики утицај ово има на допунску наставу и крајњи резултат на испиту.

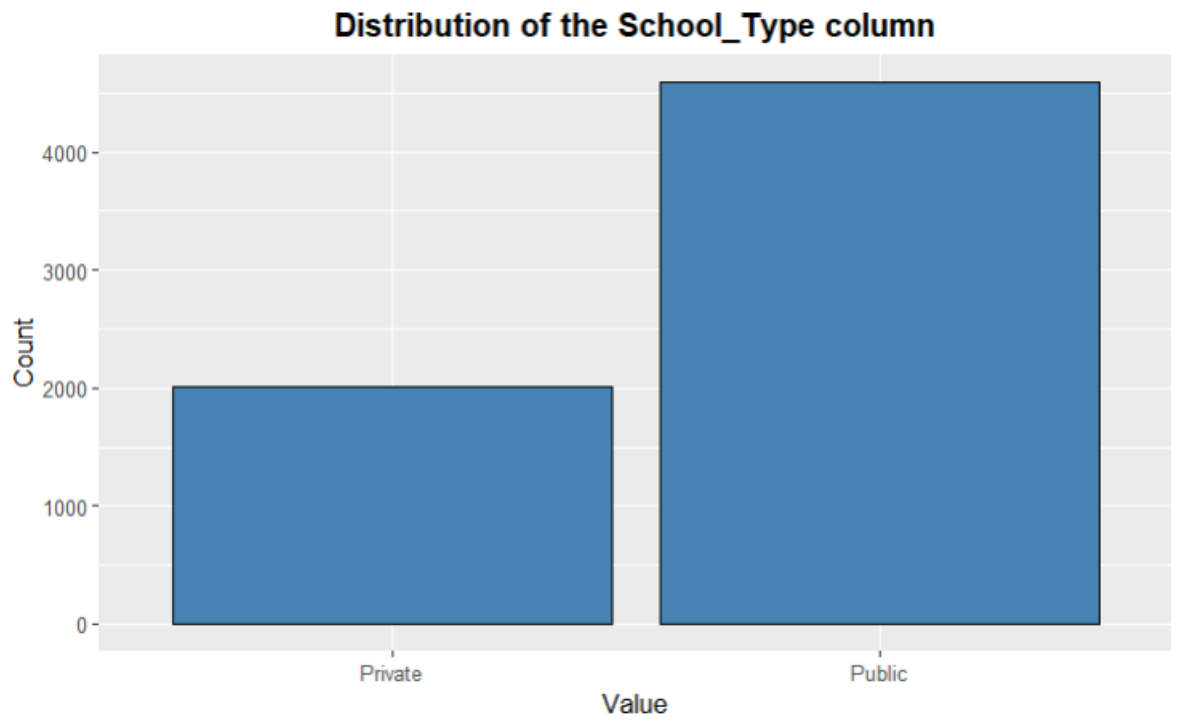
4.2.8 Категоријска колона *School_Type*

School_Type представља врсту школе коју студент похађа *(Public, Private).* Приказаћемо табелу учестаности да бисмо видели колико имамо категорија за ову колону.

```r
{r}
xtabs(~School_Type, dataset)
```

```
School_Type
Private  Public
   2006    4595
```

```r
{r}
ggplot(dataset, aes(x = School_Type)) + geom_bar(fill = "steelblue", color =
"black") + labs(title = "Distribution of the School_Type column") + xlab("Value") +
ylab("Count") + theme(plot.title = element_text(hjust = 0.5, vjust = 1, face =
"bold"))
```



Можемо видети да је већи број оних студената који похађају државну школу.

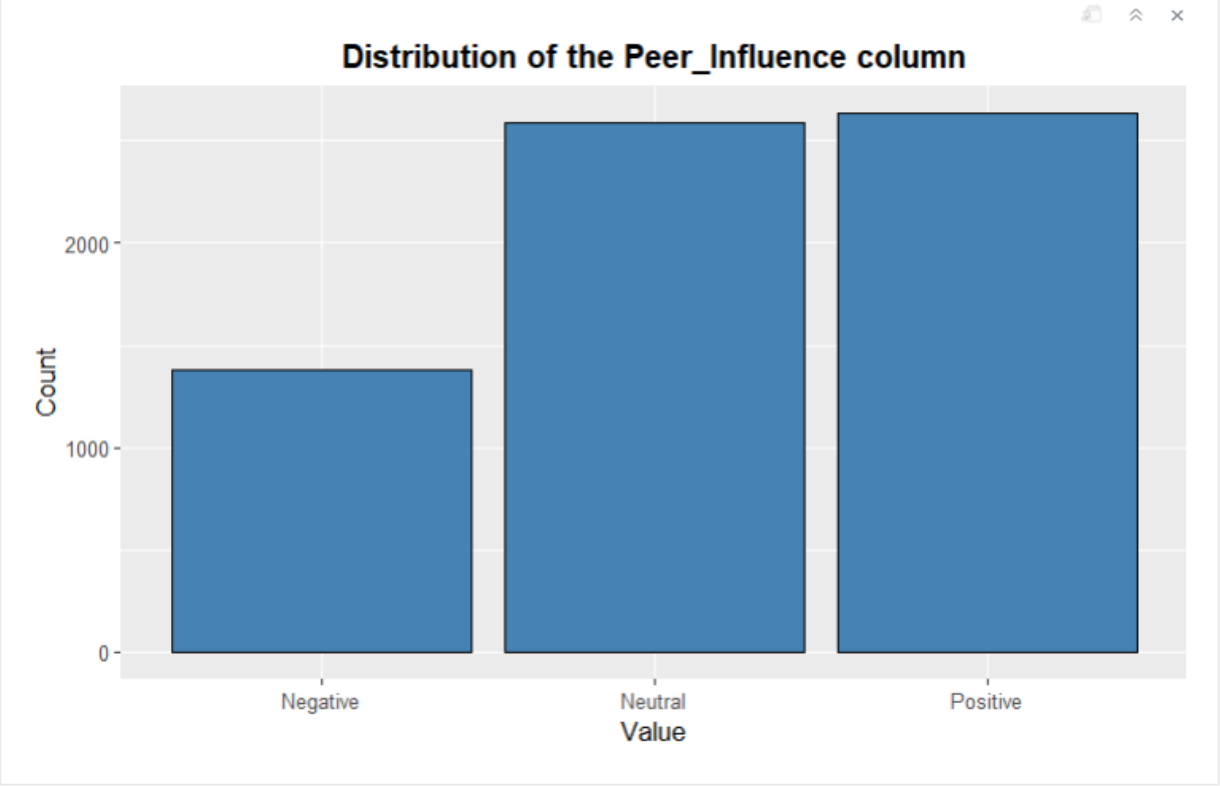
4.2.9 Категоријска колона *Peer_Influence*

Peer_Influence представља утицај вршњака на академски успех студента *(Positive, Neutral, Negative)*. Приказаћемо табелу учестаности да бисмо видели колико имамо категорија за ову колону.

```r
xtabs(~Peer_Influence, dataset)
```

```
Peer_Influence
Negative  Neutral Positive
    1376     2589     2636
```

```r
ggplot(dataset, aes(x = Peer_Influence)) + geom_bar(fill = "steelblue", color = "black") + labs(title = "Distribution of the Peer_Influence column") + xlab("Value") + ylab("Count") + theme(plot.title = element_text(hjust = 0.5, vjust = 1, face = "bold"))
```



Можемо приметити да је најмањи број утицаја негативан (*Negative*), док су неутрални (*Neutral*) и позитивни (*Positive*) утицаји скоро једнаки. Вероватно је да студенти на које колеге лоше утичу имају лошије резултате.

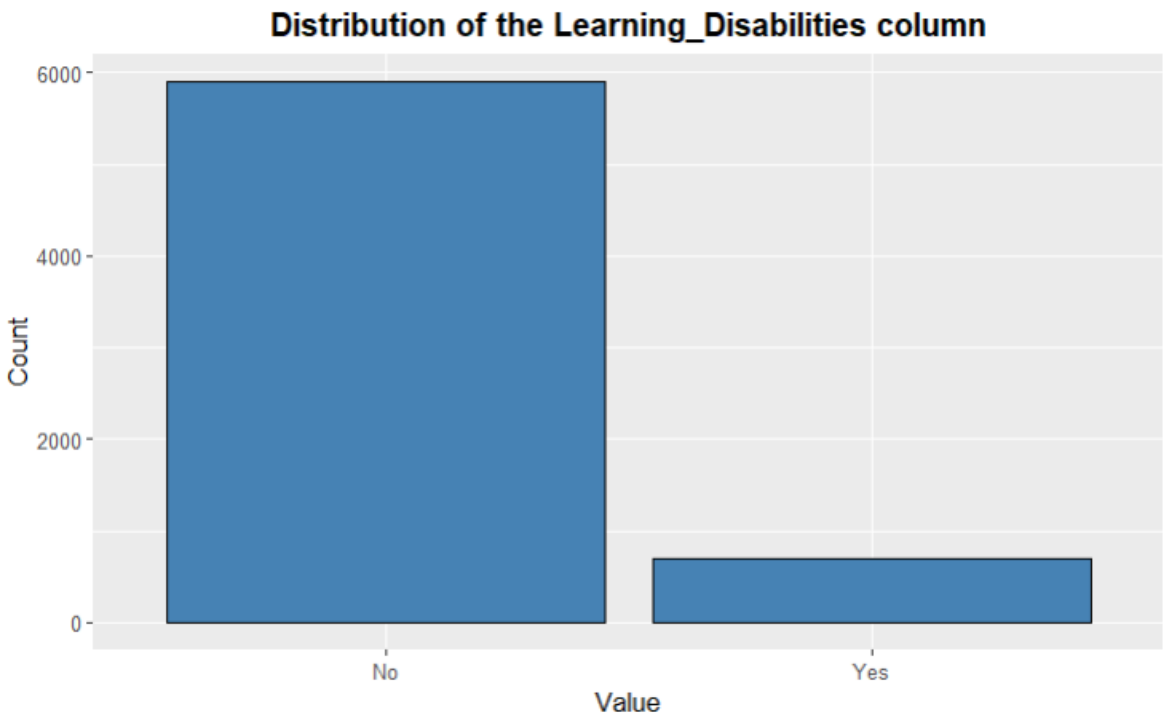
4.2.10 Категоријска колона *Learning_Disabilities*

Learning_Disabilities представља присуство сметњи у учењу *(Yes, No)*. Приказаћемо табелу учестаности да бисмо видели колико имамо категорија за ову колону.

```r
xtabs(~Learning_Disabilities, dataset)
```

```
Learning_Disabilities
  No  Yes
5906  695
```

```r
ggplot(dataset, aes(x = Learning_Disabilities)) + geom_bar(fill = "steelblue",
color = "black") + labs(title = "Distribution of the Learning_Disabilities column")
+ xlab("Value") + ylab("Count") + theme(plot.title = element_text(hjust = 0.5,
vjust = 1, face = "bold"))
```



Можемо приметити да постоје студенти који имају сметње у учењу, касније ћемо видети колики утицај оне имају на њихов крајни резултат на тестовима.

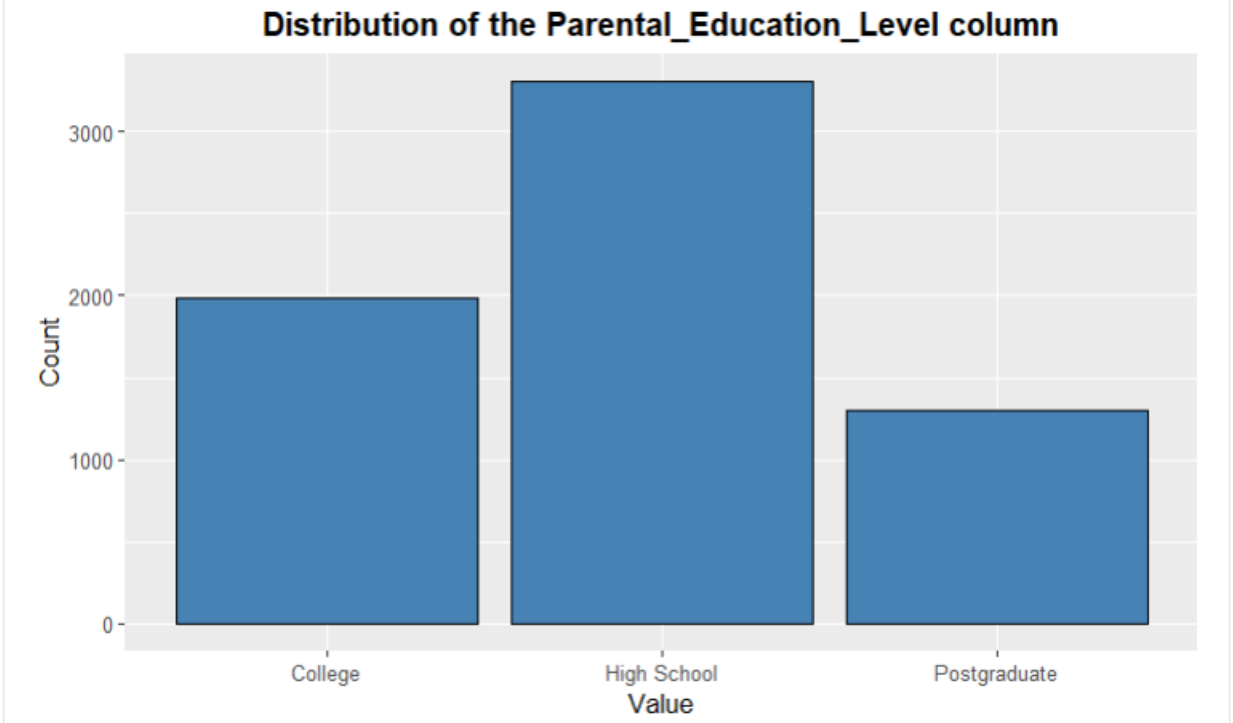
4.2.11 Категоријска колона *Parental_Education_Level*

Parental_Education_Level представља највиши ниво образовања родитеља *(High School, College, Postgraduate).* Приказаћемо табелу учестаности да бисмо видели колико имамо категорија за ову колону.

```r
{r}
xtabs(~Parental_Education_Level, dataset)
```

```
Parental_Education_Level
     College  High School Postgraduate
        1989         3308         1304
```

```r
{r}
ggplot(dataset, aes(x = Parental_Education_Level)) + geom_bar(fill = "steelblue",
color = "black") + labs(title = "Distribution of the Parental_Education_Level
column") + xlab("Value") + ylab("Count") + theme(plot.title = element_text(hjust =
0.5, vjust = 1, face = "bold"))
```



Примећујемо да највећи број студената има родитеље чији је највиши ниво едукације средња школа (*High School*), док је најмање оних који су наставили школовање након факултета на мастер или докторским студијама (*Postgraduate*).

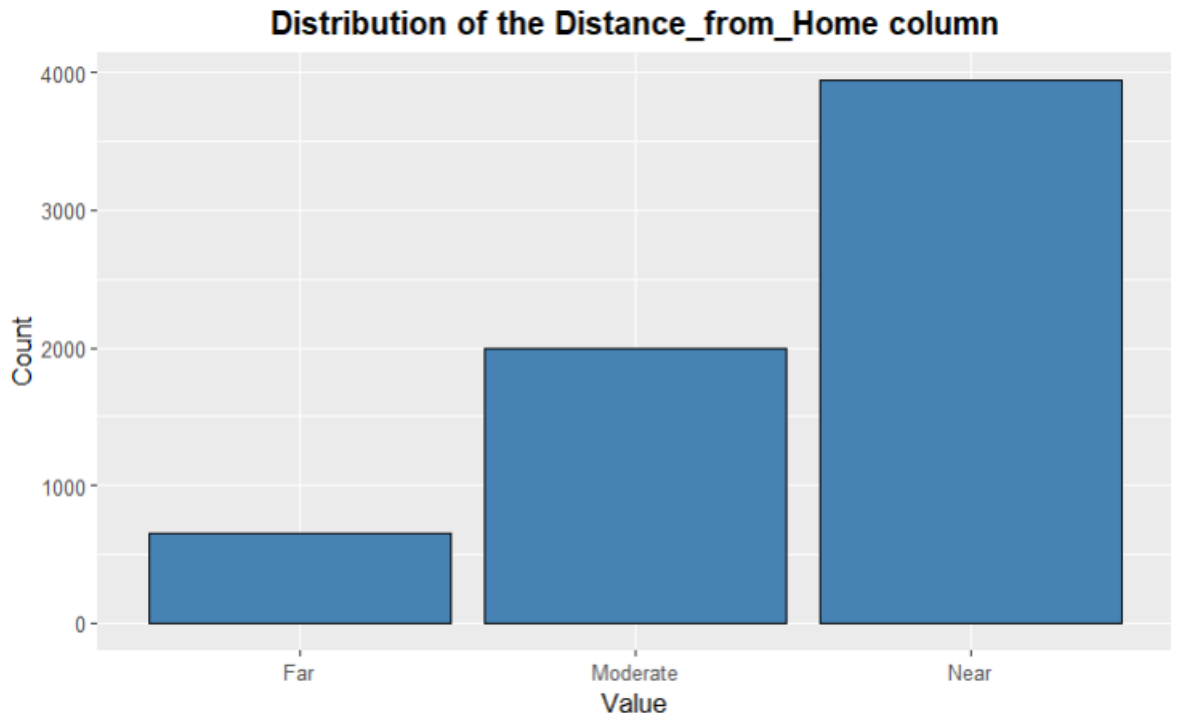
4.2.12 Категоријска колона *Distance_from_Home*

Distance_from_Home представља удаљеност од куће до школе *(Near, Moderate, Far).* Приказаћемо табелу учестаности да бисмо видели колико имамо категорија за ову колону.

```r
{r}
xtabs(~Distance_from_Home, dataset)
```

```
Distance_from_Home
     Far Moderate     Near
     658     1997     3946
```

```r
{r}
ggplot(dataset, aes(x = Distance_from_Home)) + geom_bar(fill = "steelblue", color = "black") + labs(title = "Distribution of the Distance_from_Home column") + xlab("Value") + ylab("Count") + theme(plot.title = element_text(hjust = 0.5, vjust = 1, face = "bold"))
```



Примећујемо да највећи број студената живи близу школе (*Near*), а да је број оних који живе далеко јако мали (*Far*).

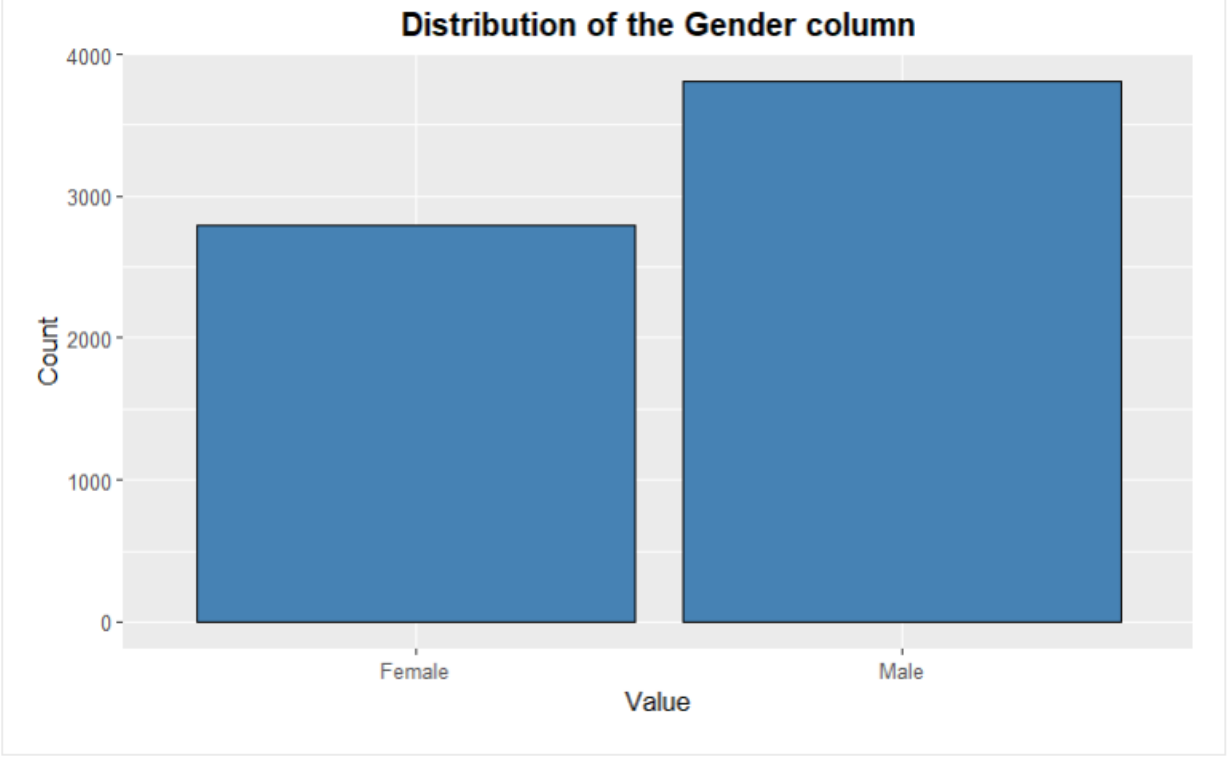
4.2.13 Категоријска колона *Gender*

Gender представља пол студента *(Male, Female).* Приказаћемо табелу учестаности да бисмо видели колико имамо категорија за ову колону.

```r
{r}
xtabs(~Gender, dataset)
```

```
Gender
Female   Male
  2790   3811
```

```r
{r}
ggplot(dataset, aes(x = Gender)) + geom_bar(fill = "steelblue", color = "black") +
labs(title = "Distribution of the Gender column") + xlab("Value") + ylab("Count") +
theme(plot.title = element_text(hjust = 0.5, vjust = 1, face = "bold"))
```



Евидентно је да је већи број студената мушког пола.

5 Анализа

У овом поглављу истражићемо различите факторе који могу имати утицаја на постигнуте резултате студената на испитима, како бисмо идентификовали кључне варијабле које највише доприносе академском успеху. Планирана анализа ће обухватати и нумеричке и категоријске факторе, како бисмо стекли дубље разумевање корелација и могућности предикције.

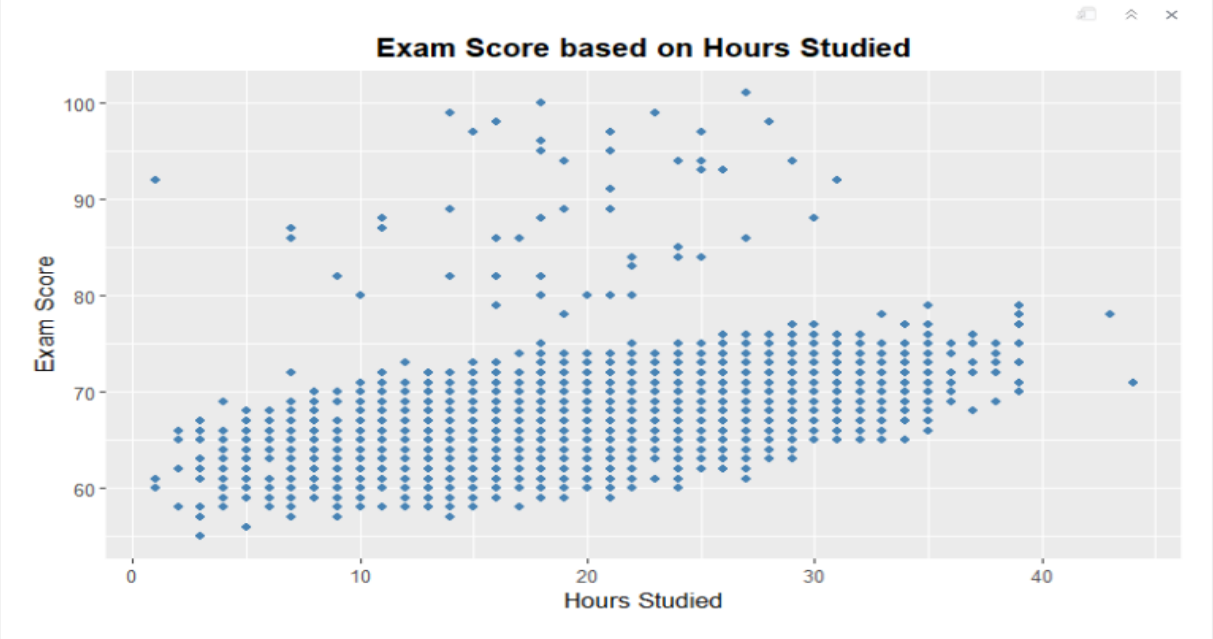
5.1 Анализа између предиктора и одговора

У овом делу ћемо да анализирамо оне колоне за које сматрамо да имају утицај на предикцију колоне *Exam_Score*. Колоне које ћемо обрађивати у даљем раду смо изабрали на основу доменског знања, док оне за оне које нису изабране сматрамо да немају утицај на излазну колону.

5.1.1 Утицај колоне *Hours_Studied*

Први фактор који анализирамо је број сати проведених у учењу недељно (*Hours_Studied*) у односу на оцену на завршном испиту (*Exam_Score*). Очекивано је да студенти који више времена проводе учећи остварују боље резултате.

```r
{r}
ggplot(dataset, aes(x = Hours_Studied, y = Exam_Score)) + geom_point(color =
"steelblue") + labs(title = "Exam Score based on Hours Studied") + xlab("Hours
Studied") + ylab("Exam Score") + theme(plot.title = element_text(hjust = 0.5, vjust
= 1, face = "bold"))
```



Са графика се може приметити позитиван однос између ове две променљиве, у већини случајева, како се повећава број сати у учењу, тако се повећава и резултат на испиту. Иако тренд није строго линеаран, делује да већи број сати у учењу води ка бољим резултатима, али подаци су прилично раштркани, што значи да број сати не гарантује директно већи успех. Такође, можемо да приметимо да има студената који проводе мање сати у учењу, а ипак постижу високе оцене, као и оних који више сати уче, а ипак немају сјајне резултате. Ово су потенцијални *outlier-i*.

```r
{r}
ggplot(dataset, aes(x = Hours_Studied, y = Exam_Score)) +
  geom_smooth(method = "loess", color = "steelblue") +
  labs(title = "Relationship between Hours Studied and Exam Score",
       x = "Hours Studied",
       y = "Exam Score") +
  theme_minimal()
```

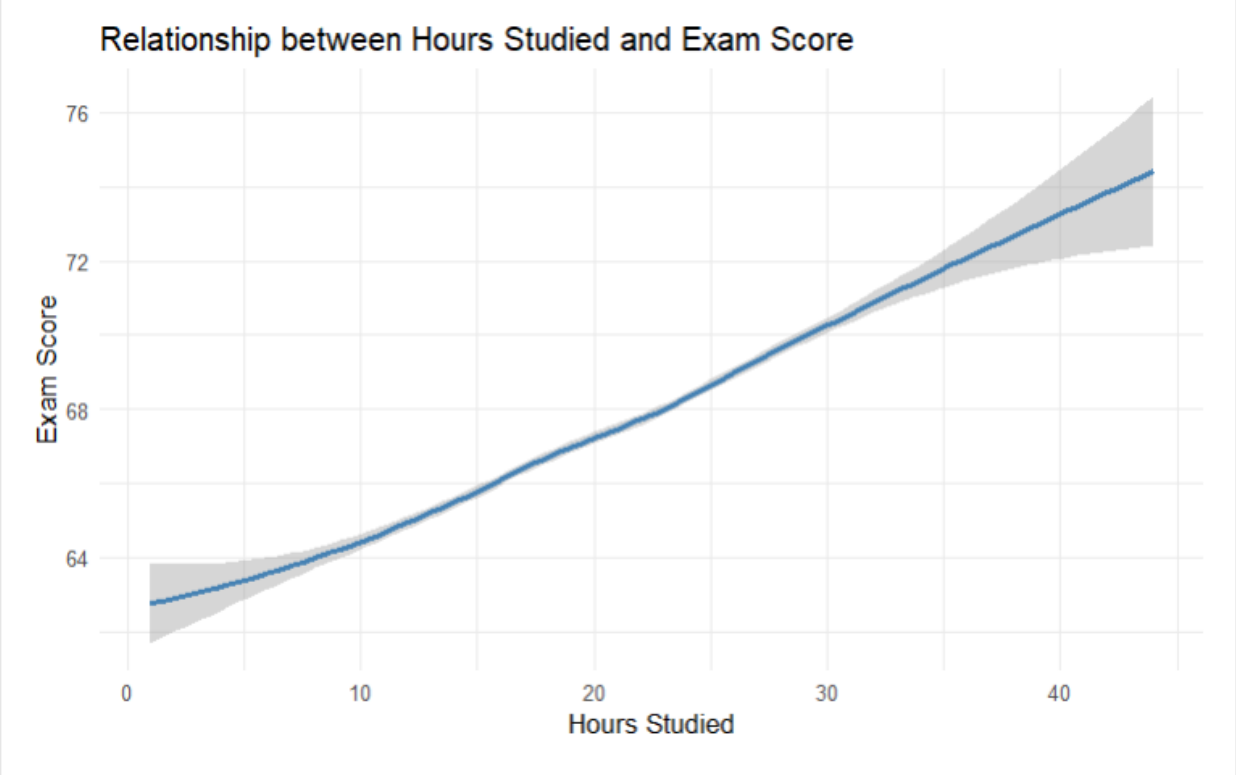


График приказује позитивну повезаност између броја сати проведених у учењу и резултата на испиту, више сати учења доводи до бољих оцена, али са нешто већим интервалом поверења, значи да постоји већа варијабилност у резултатима код студената који уче више од 35 сати недељно или код оних који уче мање од 10 сати недељно. Постоји могућност да су они који нису учили довољно, а остварили су веома добре резултате, или екстремно надарени или су варали на испитима.

Издвојићемо случајеве када студенти имају више од 80 поена.

```r
high_scorers <- filter(dataset, Exam_Score > 80)
print(high_scorers)
```

Description: df [43 × 20]

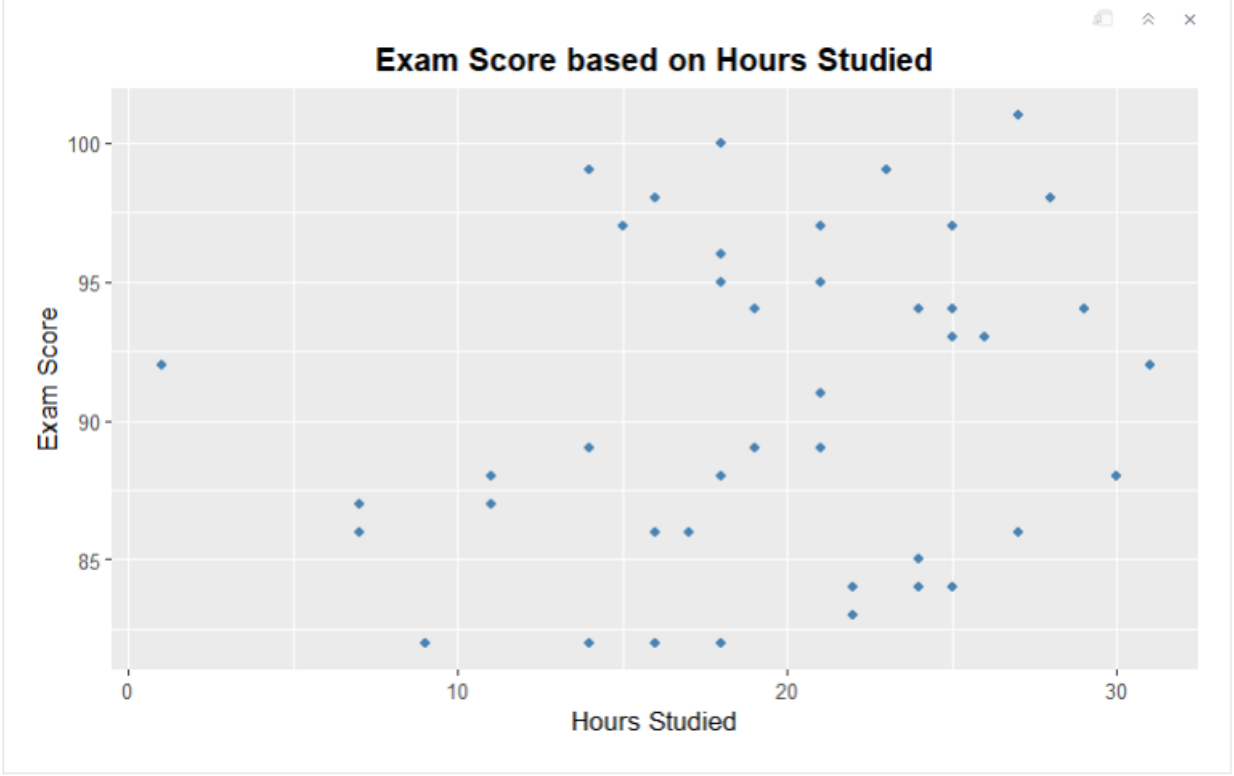
Hours_Studied <int>	Attendance <int>	Parental_Involvement <chr>	Access_to_Resources <chr>
18	89	High	Medium
19	70	Medium	Low
17	77	Low	High
15	83	Medium	Medium
22	70	Low	Medium
22	71	Low	High
24	96	Low	High
29	76	Medium	Medium
21	74	High	Medium
14	77	Low	High

1–10 of 43 rows | 1–4 of 20 columns Previous 1 2 3 4 5 Next

Анализирали смо овај подскуп у односу на целокупан *dataset* и не постоје неки значајни патерни или навике које студенти који постижу високе резултате имају у односу на остатак. Све вредности и медијане су јако сличне са минималним разликама.

Свакако ћемо приказати подскуп на графику.

```r
{r}
ggplot(high_scorers, aes(x = Hours_Studied, y = Exam_Score)) + geom_point(color =
"steelblue") + labs(title = "Exam Score based on Hours Studied") + xlab("Hours
Studied") + ylab("Exam Score") + theme(plot.title = element_text(hjust = 0.5, vjust
= 1, face = "bold"))
```



Примећујемо да ретко који студент који има добре резултате учи мање од 10 сати недељно, за студенте који ипак уче мање од 10 сати се поставља питање да ли су можда варали на тестовима, јер остали атрибути попут мотивације, присуства и претходних резултата немају одређен патерн, што се види на наредној слици.

```r
{r}
filtered_students <- dataset %>% filter(Hours_Studied < 10 & Exam_Score > 80)
filtered_students_extra <- filtered_students %>% select(Hours_Studied, Attendance,
Family_Income, Previous_Scores, Motivation_Level, Exam_Score)
print(filtered_students_extra)
```

Description: df [4 × 6]

Hours_Studied <int>	Attendance <int>	Family_Income <chr>	Previous_Scores <int>
9	61	High	77
7	69	Medium	54
7	63	Low	90
1	88	Medium	72

4 rows | 1–4 of 6 columns

Већина студената је сконцентрисана између 15 и 25 сати, па чак и ту има доста варијације у резултатима. Закључујемо да време проведено у учењу није фактор који самостално може да утиче на поене.

5.1.2 Утицај колоне *Attendance*

Овај график приказује однос између присуства на часовима (*Attendance*) и постигнуте оцене на завршном испиту (*Exam_Score*). Очекује се да студенти са већим присуством имају боље резултате на испиту.

```{r}
ggplot(dataset, aes(x = Attendance, y = Exam_Score)) + geom_point(color =
"steelblue") + labs(title = "Exam Score based on Attendance") +
scale_color_viridis_c() + xlab("Attendance") + ylab("Exam Score") + theme(plot
.title = element_text(hjust = 0.5, vjust = 1, face = "bold"))
```

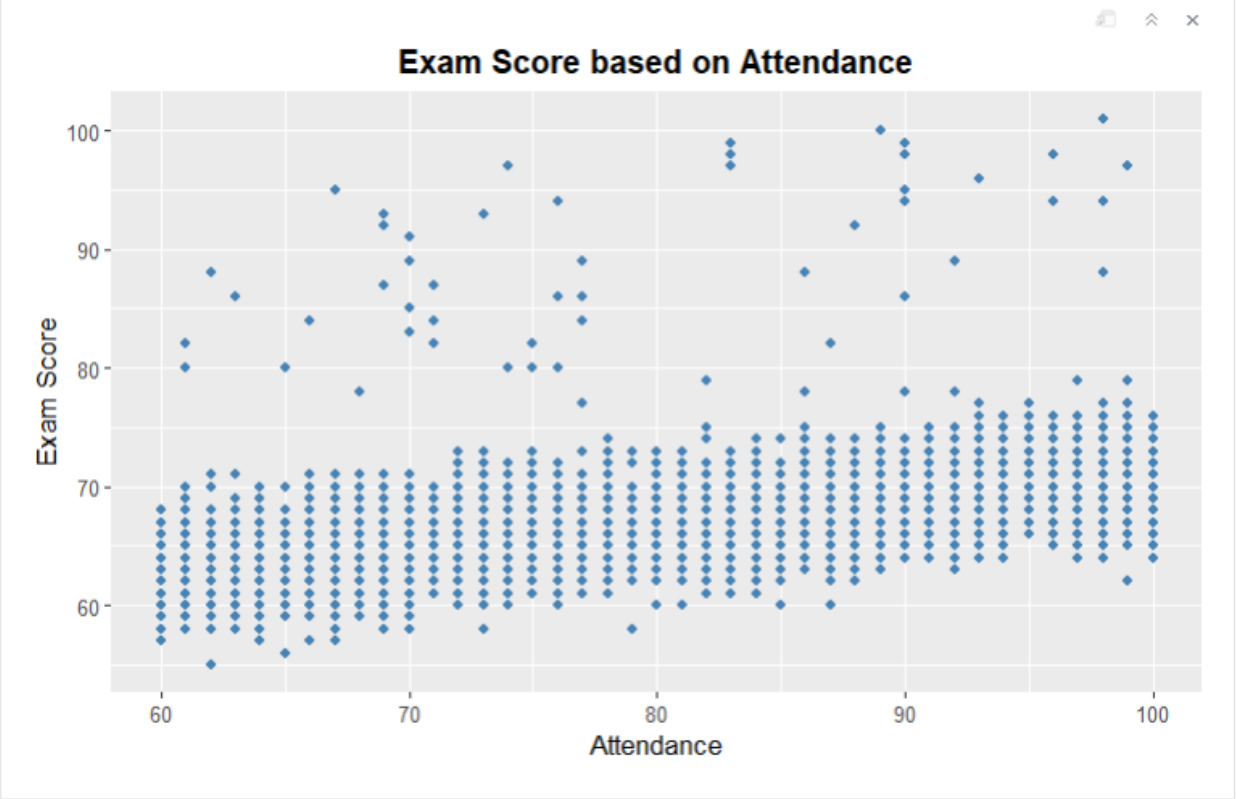


График нам показује да што више студенти присуствују часовима, углавном добијају боље оцене, али то није увек строго правило. Они који имају више од 90% присуства често постижу боље резултате у односу на оне са мањим присуством који остају у опсегу од 60 до 70 поена. Наравно постоје и изузеци, неки студенти са мањим присуством постижу високе оцене. Дакле, присуство је важно, али није једини фактор који утиче на успех.

```r
{r}
ggplot(dataset, aes(x = Attendance, y = Exam_Score)) +
  geom_smooth(method = "loess", color = "blue") +
  labs(title = "Relationship between Attendance and Exam Score",
       x = "Attendance",
       y = "Exam Score") +
  theme_minimal()
```

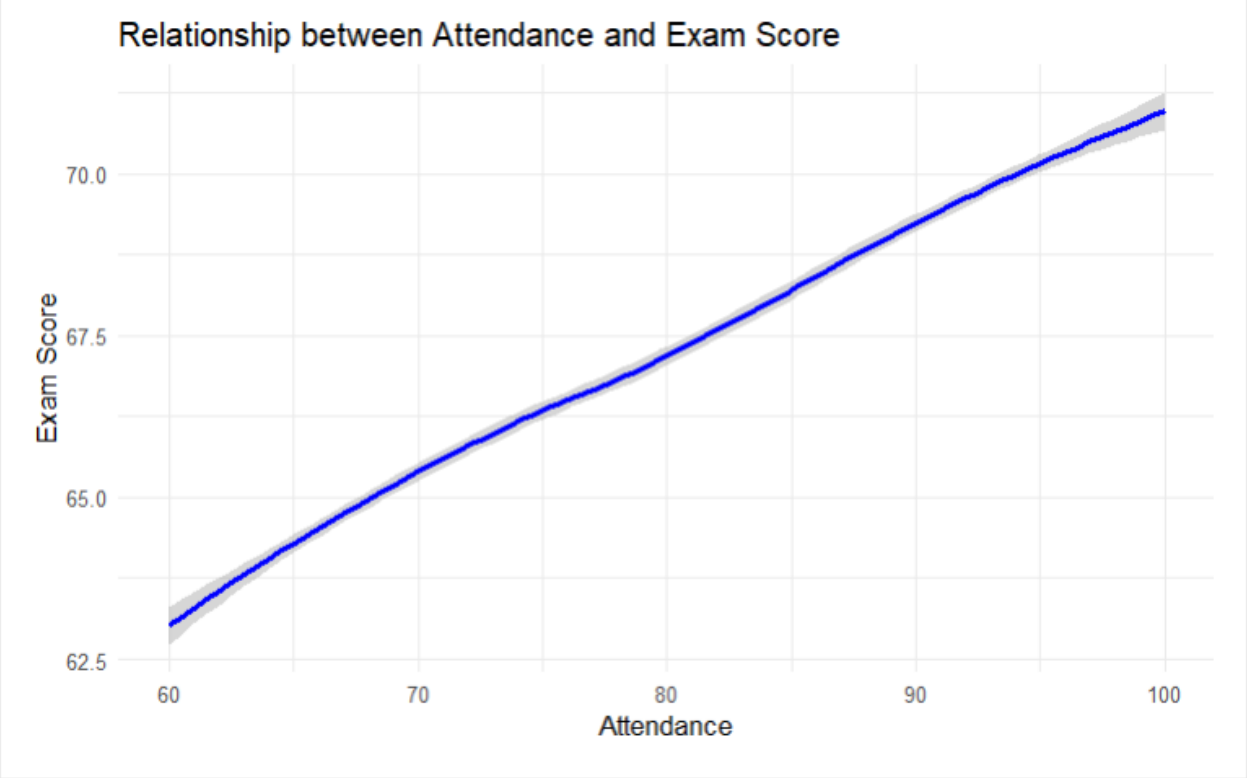
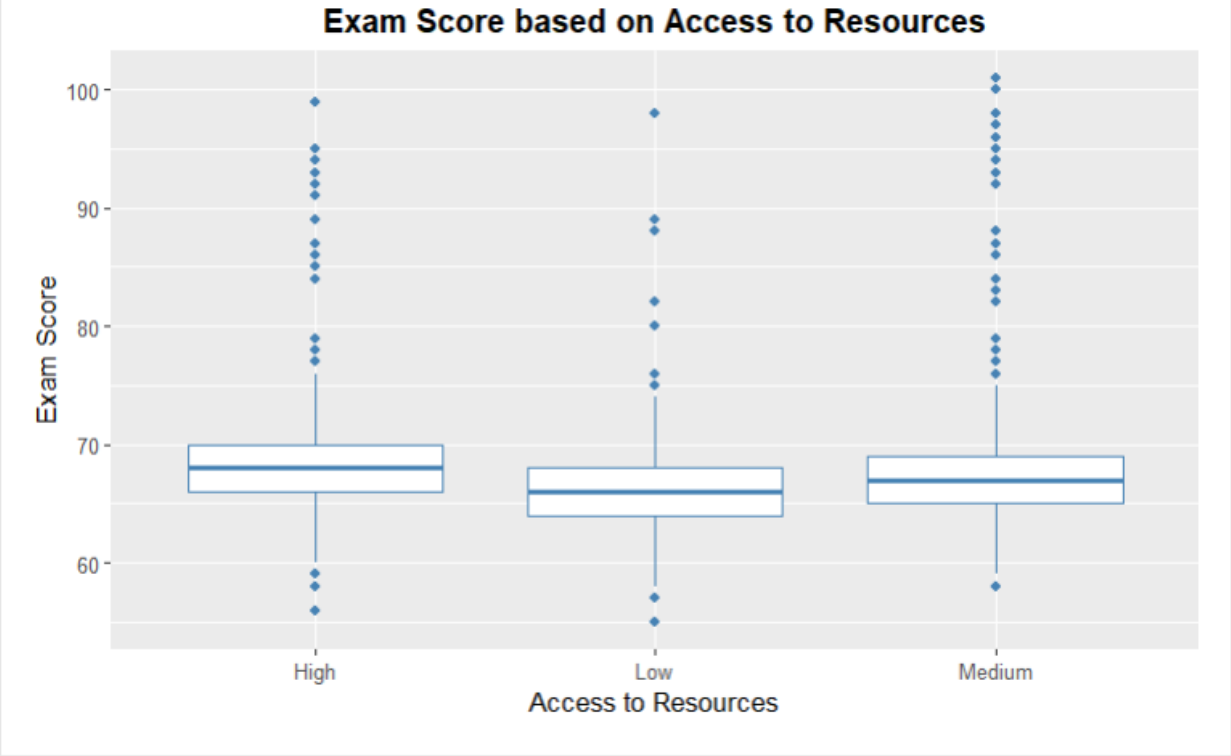


График показује позитивну повезаност између присуства студената и њихових резултата на испиту, у глобалу можемо рећи да веће присуство води ка бољим оценама.

5.1.3 Утицај колоне *Access_to_Resource*

Сада ћемо анализирати какав утицај има доступност образовних ресурса студента на његове резултате на испиту.

```r
{r}
ggplot(dataset, aes(x = Access_to_Resources, y = Exam_Score)) + geom_boxplot(color
= "steelblue") + labs(title = "Exam Score based on Access to Resources") +
scale_color_viridis_c() + xlab("Access to Resources") + ylab("Exam Score") + theme
(plot.title = element_text(hjust = 0.5, vjust = 1, face = "bold"))
```

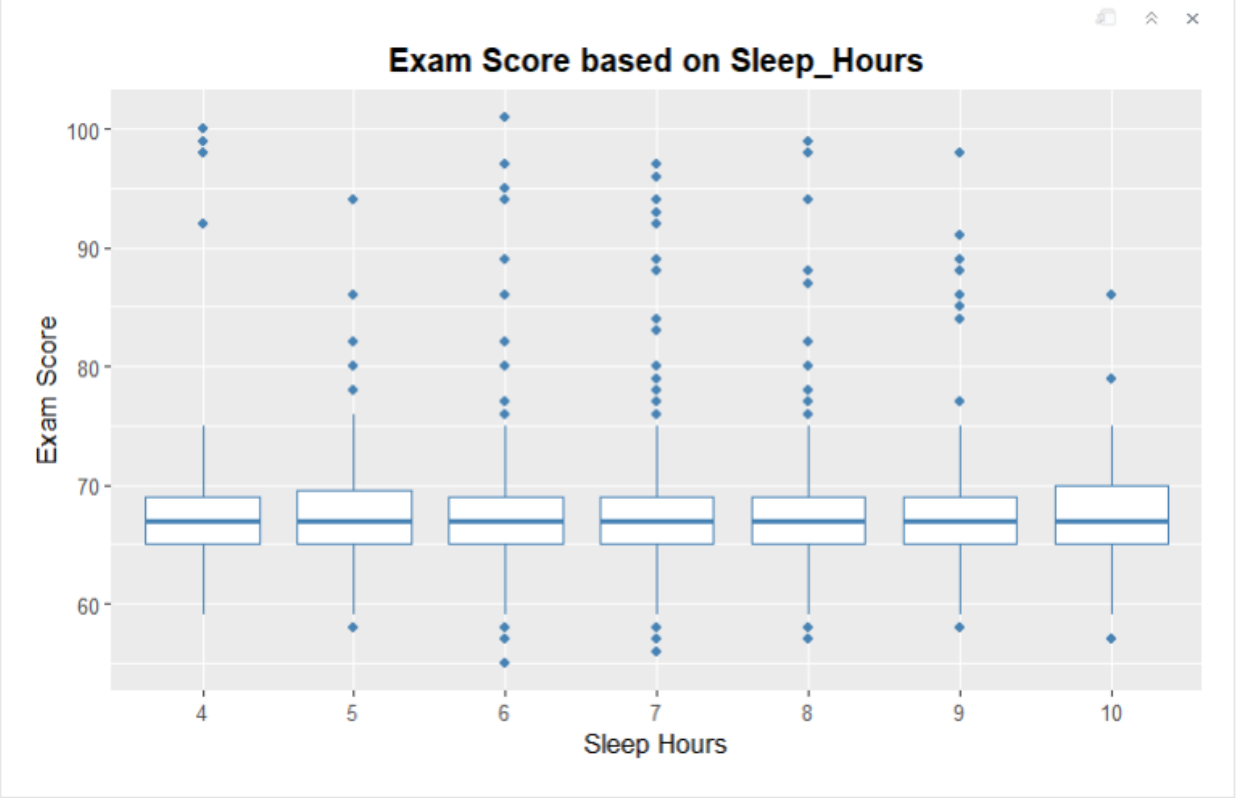


Закључујемо да нема значајних разлика у распону оцена између различитих нивоа доступности образовних ресурса. За све нивое *(High, Low, Medium)*, већина студената добија оцене у распону од 60 до 80 поена, без јасног тренда који би указивао на то да већа доступност ресурса доводи до бољих резултата. Постоји неколико *outlier-a* са високим оценама изнад 90, али они су присутни код свих категорија, што значи да овај фактор сам по себи можда није пресудан.

5.1.4 Утицај колоне *Sleep_Hours*

Сада ћемо анализирати колону *Sleep_Hours* и након анализе колоне видети да ли она има неки утицај на излазну колону *Exam_Score.*

```r
{r}
ggplot(dataset, aes(x = as.factor(Sleep_Hours), y = Exam_Score)) + geom_boxplot
(color = "steelblue") + labs(title = "Exam Score based on Sleep Hours") +
scale_color_viridis_c() + xlab("Sleep Hours") + ylab("Exam Score") + theme(plot
.title = element_text(hjust = 0.5, vjust = 1, face = "bold"))
```

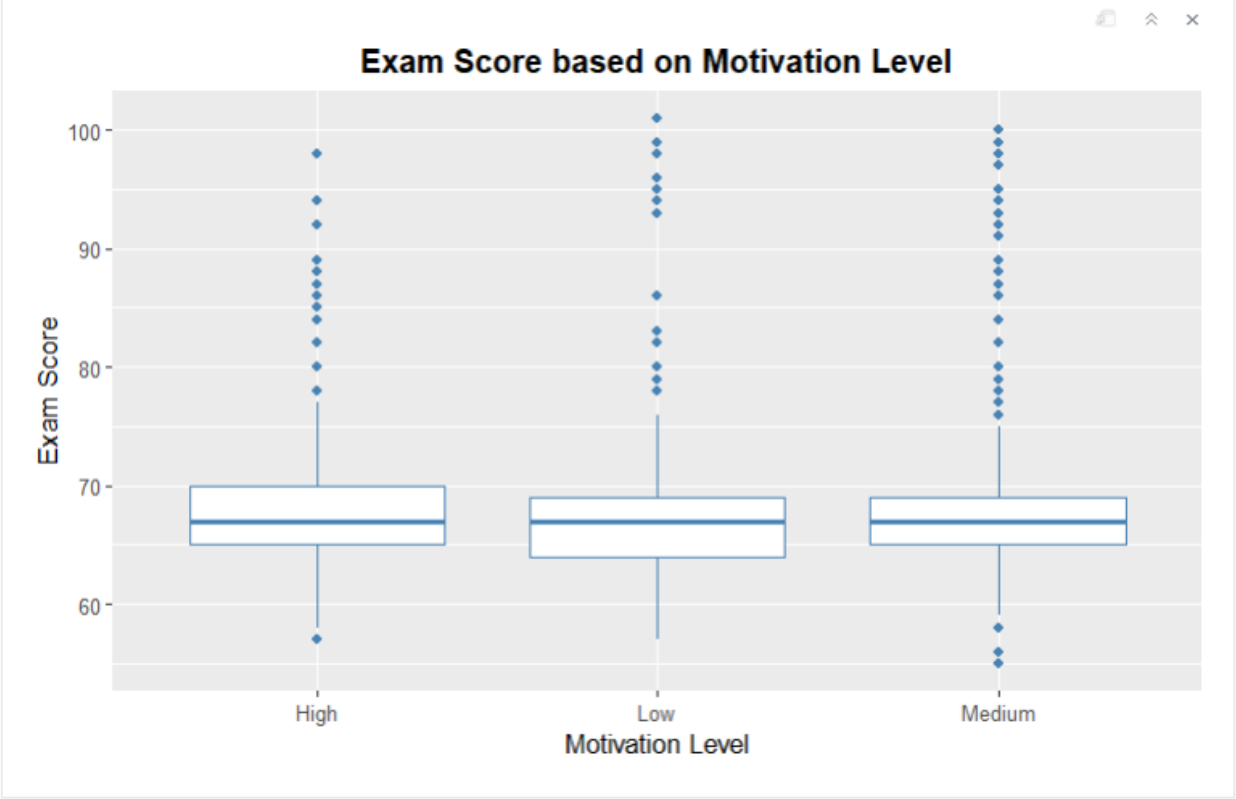


Видимо да већина студената, без обзира на број сати сна, има оцене концентрисане око 70 поена, што значи да нема значајних разлика у оценама у зависности од броја сати сна. Постоје неки *outlier-u* са изузетно високим резултатима, али они су равномерно распоређени кроз различите категорије, па бисмо рекли да количина сна можда нема велики утицај на резултате на испиту.

5.1.5 Утицај колоне *Motivaton_Level*

Анализираћемо колону *Motivation_Level* у односу на излазну колону и покушаћемо да донесемо закључак да ли ова колона има утицај на предикцију излазне колоне.

```r
{r}
ggplot(dataset, aes(x = Motivation_Level, y = Exam_Score)) + geom_boxplot(color =
"steelblue") + labs(title = "Exam Score based on Motivation Level") +
scale_color_viridis_c() + xlab("Motivation Level") + ylab("Exam Score") + theme
(plot.title = element_text(hjust = 0.5, vjust = 1, face = "bold"))
```

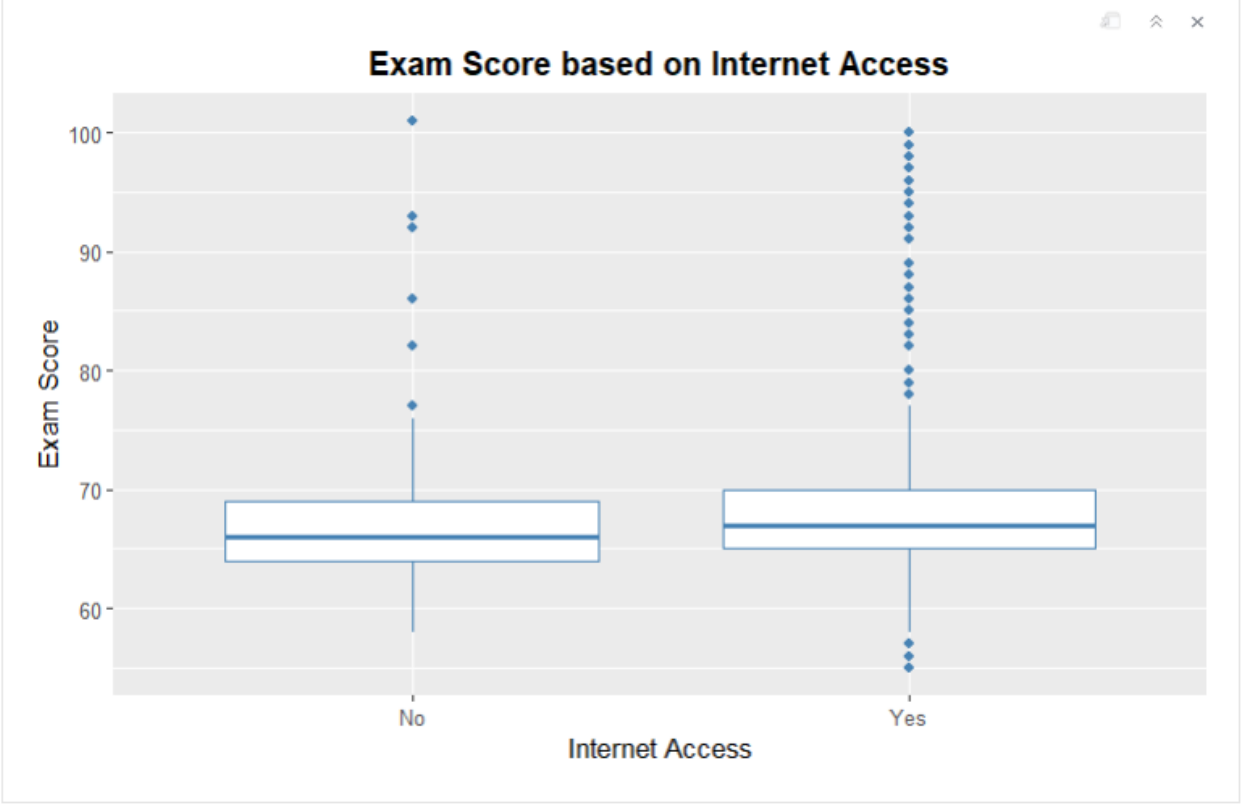


Можемо да видимо да постоји врло мала разлика у медијанама између различитих нивоа мотивације, па вероватно мотивација не игра значајну улогу у просечним резултатима. Међутим, студенти са високим нивоом мотивације имају нешто шири распон оцена, укључујући неколико изузетно високих резултата, док студенти са ниском мотивацијом такође имају појединачне *outlier-e* са високим оценама, али су у већини ближе просеку.

5.1.6 Утицај колоне *Internet_Access*

Сада ћемо анализирати какав утицај има приступ интернету на студентове резултате на испиту.

```r
{r}
ggplot(dataset, aes(x = Internet_Access, y = Exam_Score)) + geom_boxplot(color =
"steelblue") + labs(title = "Exam Score based on Internet Access") +
scale_color_viridis_c() + xlab("Internet Access") + ylab("Exam Score") + theme(plot
.title = element_text(hjust = 0.5, vjust = 1, face = "bold"))
```

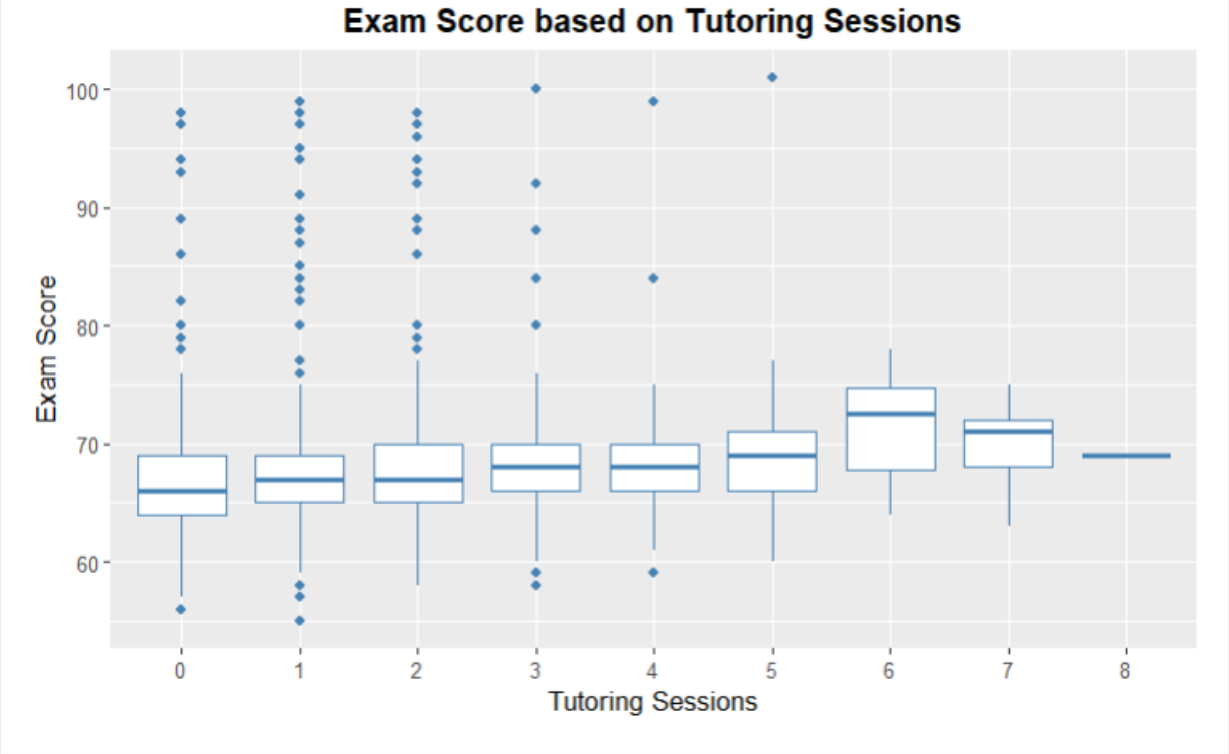


Видљиво је да студенти који имају приступ интернету имају нешто шири распон оцена, укључујући неколико изузетно високих резултата, док студенти без приступа интернету углавном постижу резултате ближе медијану око 70 поена. Иако постоје неки *outlier-u*, изгледа да приступ интернету нема утицај какав смо очекивали.

5.1.7 Утицај колоне *Tutoring_Sessions*

Сада ћемо видети какав утицај има колона *Tutoring_Sessions* на колону која се предвиђа.

```r
{r}
ggplot(dataset, aes(x = as.factor(Tutoring_Sessions), y = Exam_Score)) +
geom_boxplot(color = "steelblue") + labs(title = "Exam Score based on Tutoring
Sessions") + scale_color_viridis_c() + xlab("Tutoring Sessions") + ylab("Exam
Score") + theme(plot.title = element_text(hjust = 0.5, vjust = 1, face = "bold"))
```



Како се број сесија повећава, видимо благ пораст у медијани оцена, а посебно код студената који су имали 5 или више сесија. Постоје и неки outlier-и, али генерално, чини се да већи број сесија доводи до бољих резултата. Ова колона би могла бити корисна за *feature engineering*, јер видимо потенцијални позитиван утицај на резултате на испиту.

Одлучили смо да колону *Tutoring_Sessions* поделимо у 3 категорије: *None* за оне студенте који уопште нису имали допунске часове, *Between 1 and 3* за оне који су имали између 1 и 3 часа месечно и *More than 3* за оне који су имали више од 3.

```r
{r}
dataset$Tutoring_Categories <- cut(dataset$Tutoring_Sessions,
breaks = c(-Inf, 0, 3, Inf),
labels = c("None", "Between 1 and 3", "More than 3"),
right = TRUE)

head(dataset[, c("Tutoring_Sessions", "Tutoring_Categories")])
```

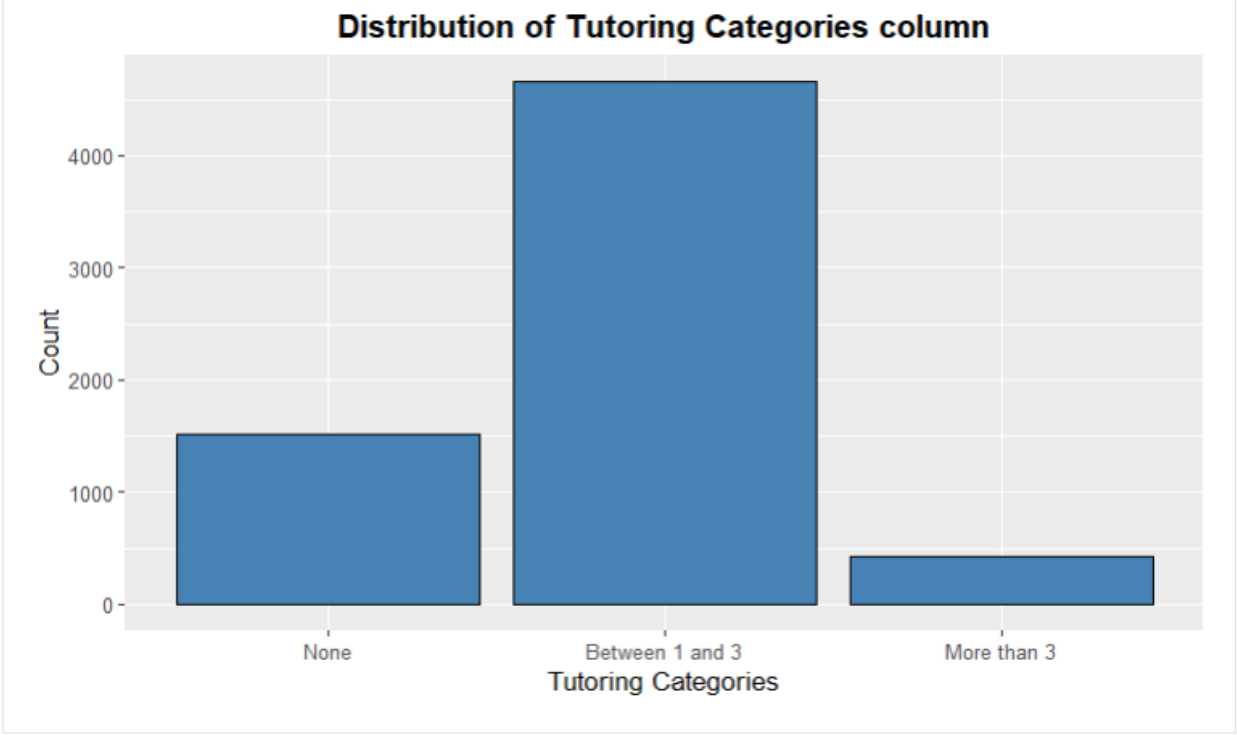
Description: df [6 × 2]

	Tutoring_Sessi... <int>	Tutoring_Cate... <fctr>
1	0	None
2	2	Between 1 and 3
3	2	Between 1 and 3
4	1	Between 1 and 3
5	3	Between 1 and 3
6	3	Between 1 and 3

6 rows

На графику испод ћемо приказати расподелу новокреиране категоријске променљиве *Tutoring_Categories.*

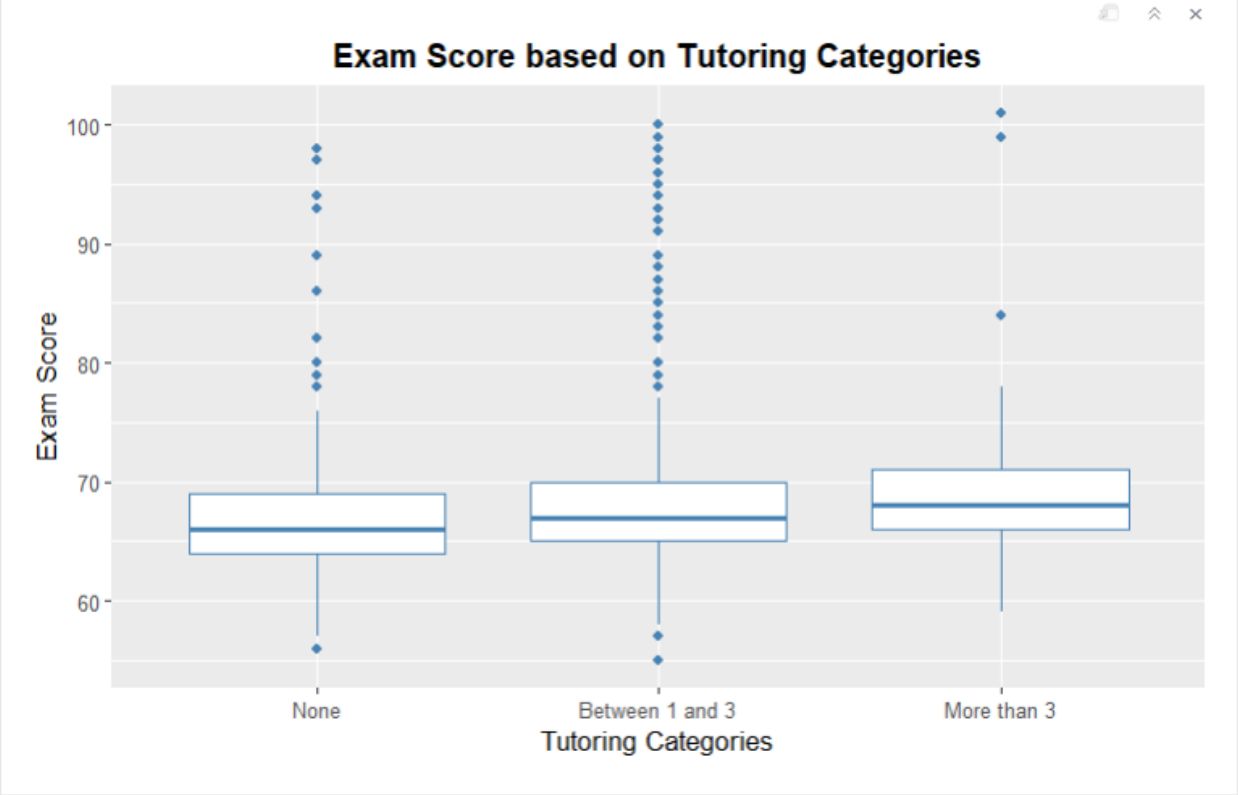
```r
{r}
ggplot(dataset, aes(x = Tutoring_Categories)) + geom_bar(fill = "steelblue", color
= "black") + labs(title = "Distribution of Tutoring Categories column") +
scale_color_viridis_c() + xlab("Tutoring Categories") + ylab("Count") + theme(plot
.title = element_text(hjust = 0.5, vjust = 1, face = "bold"))
```



Видимо да већина студената припада категорији *Between 1 and 3*, затим иду они који немају никакве допунске часове, а само мали број студената има већи број часова месечно.

Сада ћемо анализирати какав утицај допунски часови имају на крајњи резултат на испиту.

```r
{r}
ggplot(dataset, aes(x = as.factor(Tutoring_Categories), y = Exam_Score)) +
geom_boxplot(color = "steelblue") + labs(title = "Exam Score based on Tutoring
Categories") + scale_color_viridis_c() + xlab("Tutoring Categories") + ylab("Exam
Score") + theme(plot.title = element_text(hjust = 0.5, vjust = 1, face = "bold"))
```

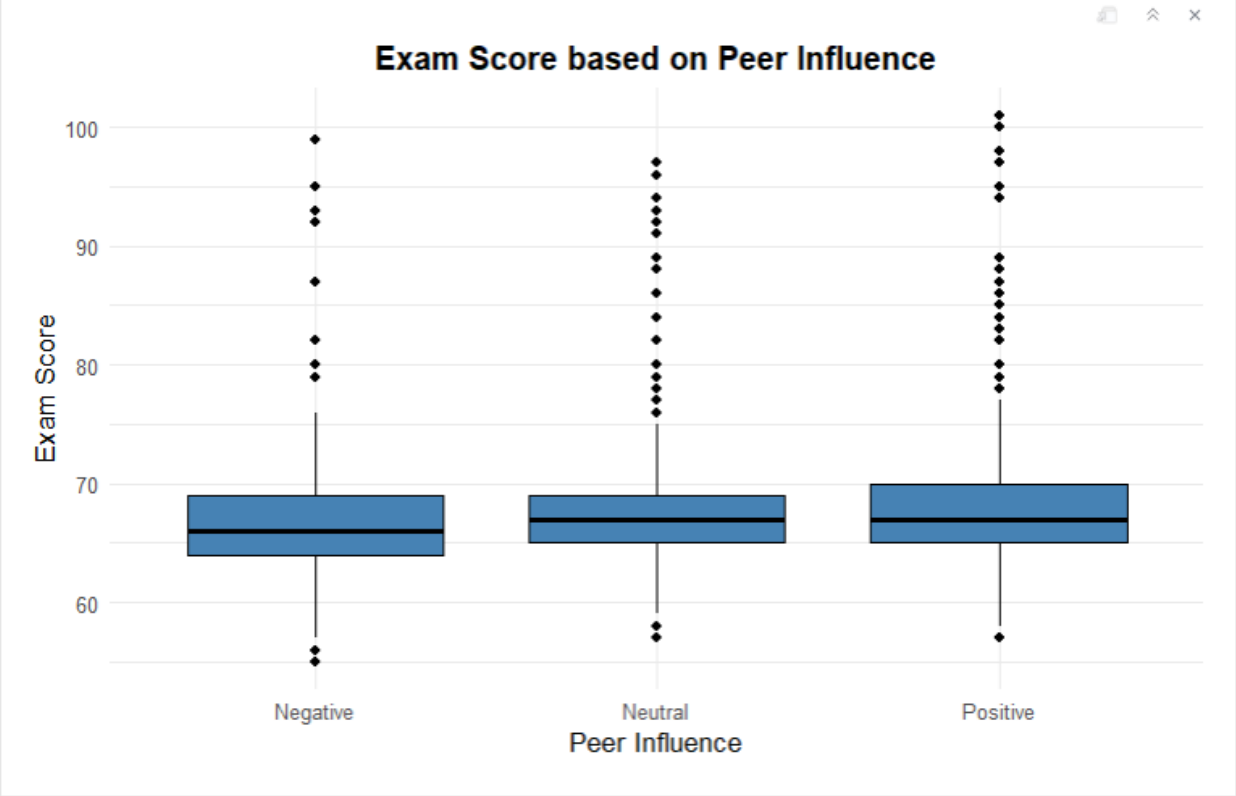


Овај график показује да студенти без допунских часова имају медијану око 70, али има неколико изузетака који постижу веома високе резултате. Слично томе, студенти који имају између 1 и 3 часа имају сличну медијану и распон резултата. Док студенти са више од 3 часа имају већу медијану и мањи распон поена, што указује на то да већи број сесија можда има позитиван утицај на успех, иако је узорак мањи и садржи неколико екстремних вредности.

Обрадили смо колоне за које смо сматрали да имају утицај на крајње резултате на испиту, остале колоне смо такође испитивали помоћу графика, али смо закључили да њихов утицај на резултате испита није значајан. Такође, на основу доменског знања, смо одлучили да неке од тих колона избацимо из даље анализе, јер њихово укључивање не би побољшало тачност предикције.

На пример за колону *Peer_Influence* смо претпостављали да ће имати већи утицај на крајни резултат, поготово за студенте на које колеге утичу лоше, међутим када смо направили график, видели смо да то баш и није тако.

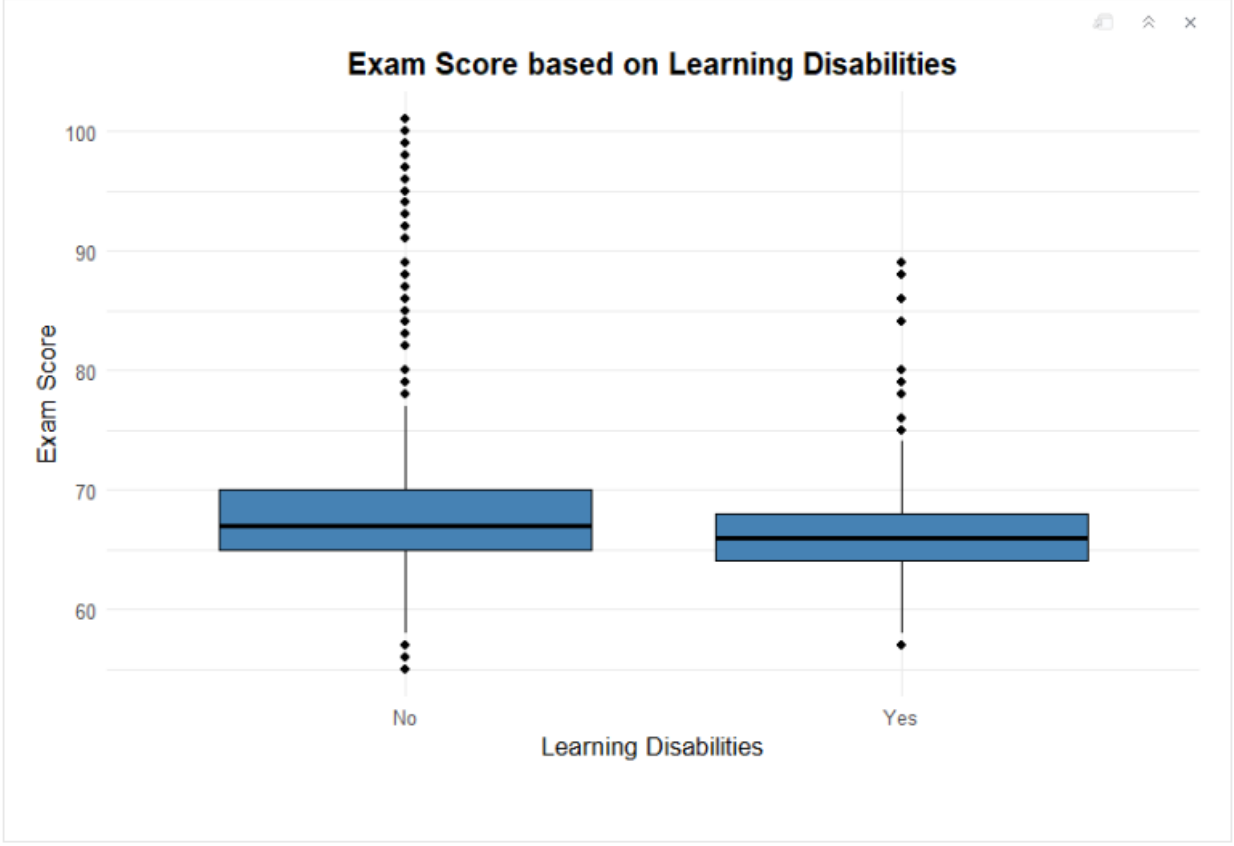
```{r}
ggplot(dataset, aes(x = Peer_Influence, y = Exam_Score)) + geom_boxplot(fill =
"steelblue", color = "black")+ labs(title = "Exam Score based on Peer Influence", x
= "Peer Influence", y = "Exam Score") + theme_minimal() + theme(plot.title =
element_text(hjust = 0.5, vjust = 1, face = "bold"))
```



Просечни резултати су веома слични, све три категорије имају сличан распон у резултатима, чак имају у сличан број изузетака. Нема очигледног обрасца или утицаја на поене на испиту.

Слично важи и за колону *Learning_Disabilities.* Очекивали смо да постоји видљива разлика у резултатима између студената који имају потешкоћа у учењу и оних који немају. Међутим већ на наредном графику се види да та претпоставка не важи. Медијана јесте нижа за студенте са потешкоћама, али и они могу да постижу врло добре резултате. Горња граница јесте нижа (89 поена) у односу на студенте без потешкоћа, али не постоји довољно јака веза која би указивала да потешкоће у учењу имају значајан утицај на крајњи исход.

```r
{r}
ggplot(dataset, aes(x = Learning_Disabilities, y = Exam_Score)) + geom_boxplot(fill
= "steelblue", color = "black")+ labs(title = "Exam Score based on Learning
Disabilities", x = "Learning Disabilities", y = "Exam Score") + theme_minimal() +
theme(plot.title = element_text(hjust = 0.5, vjust = 1, face = "bold"))
```



Колоне попут пола студента, типа факултета и укључења родитеља у образовање студента смо заобишли у обради на основу домеског знања. Сматрамо да пол нема утицаја на крајњи резултат, као ни ниво укључености родитеља у образовање. Из свакодневног живота можемо закључити да ови фактори немају утицај на успех студента.

5.2 Мултиваријантна анализа

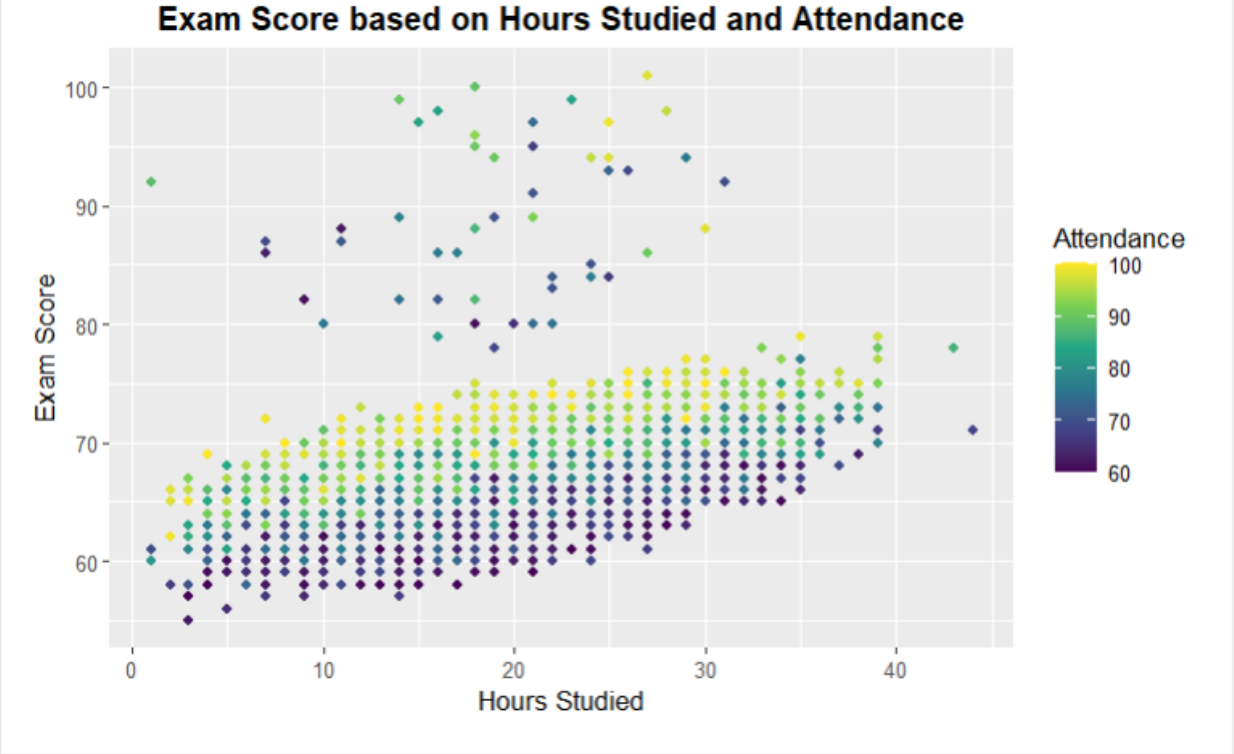
До сада смо се фокусирали на појединачне факторе који могу утицати на резултате студената на испиту, тако што смо их анализирали изоловано. У стварности академски успех зависи од комбинације више фактора. Зато је неопходно применити мултиваријантну анализу, како бисмо истражили међусобне односе између различитих променљивих и њихов заједнички утицај на резултат. Овај приступ ће нам омогућити да направимо бољи модел за предвиђање крајњих резултата на испиту.

5.2.1 Повезаност између *Hours_Studied* и *Attendance*

На основу анализа везаних за утицај различитих колона на излаз, закључили смо да би најбоље било да кренемо са комбинацијом колона *Hours_Studied* и *Attendance*. Зато што су ове две променљиве показале највећи индивидуални утицај на резултате на испиту.

Наредни график приказује однос између броја сати проведених у учењу и резултата на испиту, а боја представља присуство.

```{r}
ggplot(dataset, aes(x = Hours_Studied, y = Exam_Score, color = Attendance)) +
geom_point() + labs(title = "Exam Score based on Hours Studied and Attendance") +
scale_color_viridis_c() + xlab("Hours Studied") + ylab("Exam Score") + theme(plot
.title = element_text(hjust = 0.5, vjust = 1, face = "bold"))
```



Можемо да видимо да и време проведено у учењу и присуство јесу важни фактори утицаја на резултате испита. Веће присуство изгледа да надокнађује мањи број сати учења, док нижи ниво присуства захтева више сати учења да би се постигли бољи резултати.

Представићемо овај график на другачији начин и пошто је тешко протумачити присуство и број сати проведених у учењу, поделићемо ове вредности у интервале и покушати да одатле нешто закључимо.

```{r}
dataset <- dataset %>% mutate(Hours_Studied_Interval = cut(Hours_Studied, breaks =
seq(0, max(Hours_Studied) + 20, by = 20), right = FALSE), Attendance_Interval = cut
(Attendance, breaks = seq(0, 100, by = 20), right = TRUE))
avg_scores <- dataset %>% group_by(Hours_Studied_Interval, Attendance_Interval) %>%
summarize(Avg_Exam_Score = mean(Exam_Score), .groups = 'drop')
print(avg_scores)
```

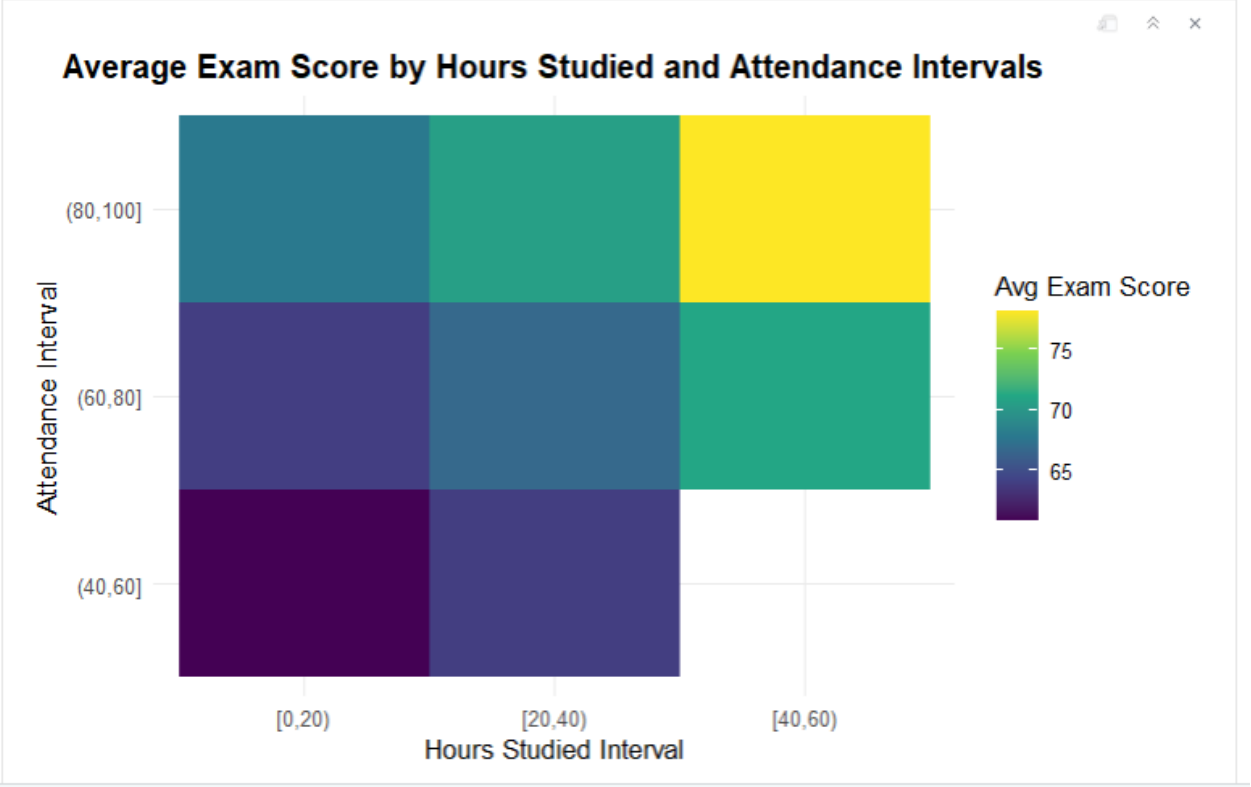
A tibble: 8 × 3

Hours_Studied_Interval <fctr>	Attendance_Interval <fctr>	Avg_Exam_Score <dbl>
[0,20)	(40,60]	60.88372
[0,20)	(60,80]	63.95416
[0,20)	(80,100]	67.80583
[20,40)	(40,60]	63.90909
[20,40)	(60,80]	66.69967
[20,40)	(80,100]	70.53584
[40,60)	(60,80]	71.00000
[40,60)	(80,100]	78.00000

8 rows

Видимо да веће присуство у комбинацији са више сати учења даје боље резултате, сада ћемо то представити на графику. Одлучили смо се за *heatmap* график.

```r
ggplot(dataset_summary, aes(x = Hours_Studied_Interval, y = Attendance_Interval,
fill = Avg_Exam_Score)) + geom_tile() + labs(title = "Average Exam Score by Hours
Studied and Attendance Intervals", x = "Hours Studied Interval", y = "Attendance
Interval", fill = "Avg Exam Score") +
scale_fill_viridis_c() + theme_minimal() + theme(plot.title = element_text(hjust =
0.5, face = "bold"))
```



Најбољи резултати се постижу када студенти уче између 40 и 60 сати недељно и имају присуство веће од 80% (просек је 78 поена). Такође, добре резултате (70-75) могу да постигну студенти који уче мање од 40 сати недељно, али имају веће присуство, или обрнуто, уколико је мање присуство (60-80) надокнади се са више сати учења. Такође примећујемо да не постоје студенти који уче више од 40 сати, а да имају долазност мању од 60%.

5.2.2 Повезаност између *Hours_Studied* и *Tutoring_Categories*

Сада ћемо истражити какав утицај заједно имају време проведено у учењу и број допунских часова на резултате испита.

Прво ћемо одредити просек освојених поена за комбинације различитих интервала за време проведено у учењу и број допунских часова.

```r
{r}
dataset_summary <- dataset %>%
  group_by(Hours_Studied_Interval, Tutoring_Categories) %>%
  summarise(Avg_Exam_Score = mean(Exam_Score, na.rm = TRUE), .groups = 'drop')
print(dataset_summary)
```

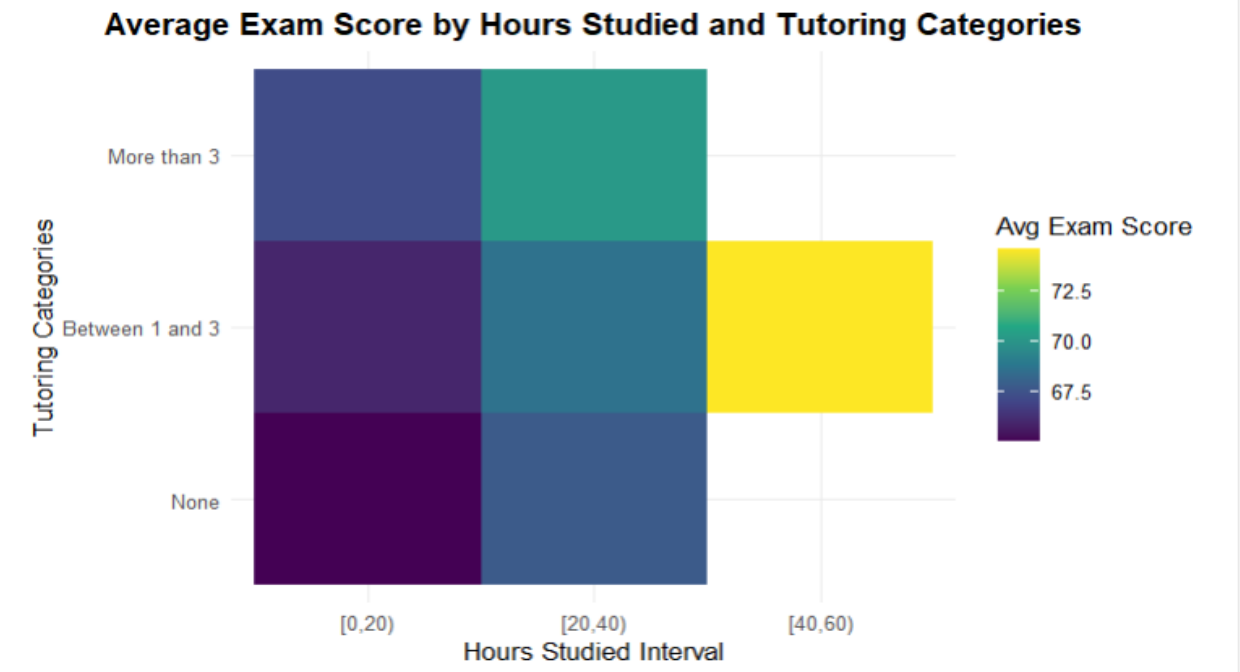
A tibble: 7 × 3

Hours_Studied_Interval <fctr>	Tutoring_Categories <fctr>	Avg_Exam_Score <dbl>
[0,20)	None	65.01149
[0,20)	Between 1 and 3	65.90976
[0,20)	More than 3	67.18919
[20,40)	None	67.74908
[20,40)	Between 1 and 3	68.59176
[20,40)	More than 3	70.11058
[40,60)	Between 1 and 3	74.50000

7 rows

Примећујемо да студенти који уопште не иду на допунску наставу имају боље резултате уз више учења. Генерално, са порастом сати учења у комбинацији са допунском наставом се виде бољи резултати.

```r
{r}
ggplot(dataset_summary, aes(x = Hours_Studied_Interval, y = Tutoring_Categories, fill = Avg_Exam_Score)) + geom_tile() + labs(title = "Average Exam Score by Hours Studied and Tutoring Categories", x = "Hours Studied Interval", y = "Tutoring Categories", fill = "Avg Exam Score") + scale_fill_viridis_c() + theme_minimal() + theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```

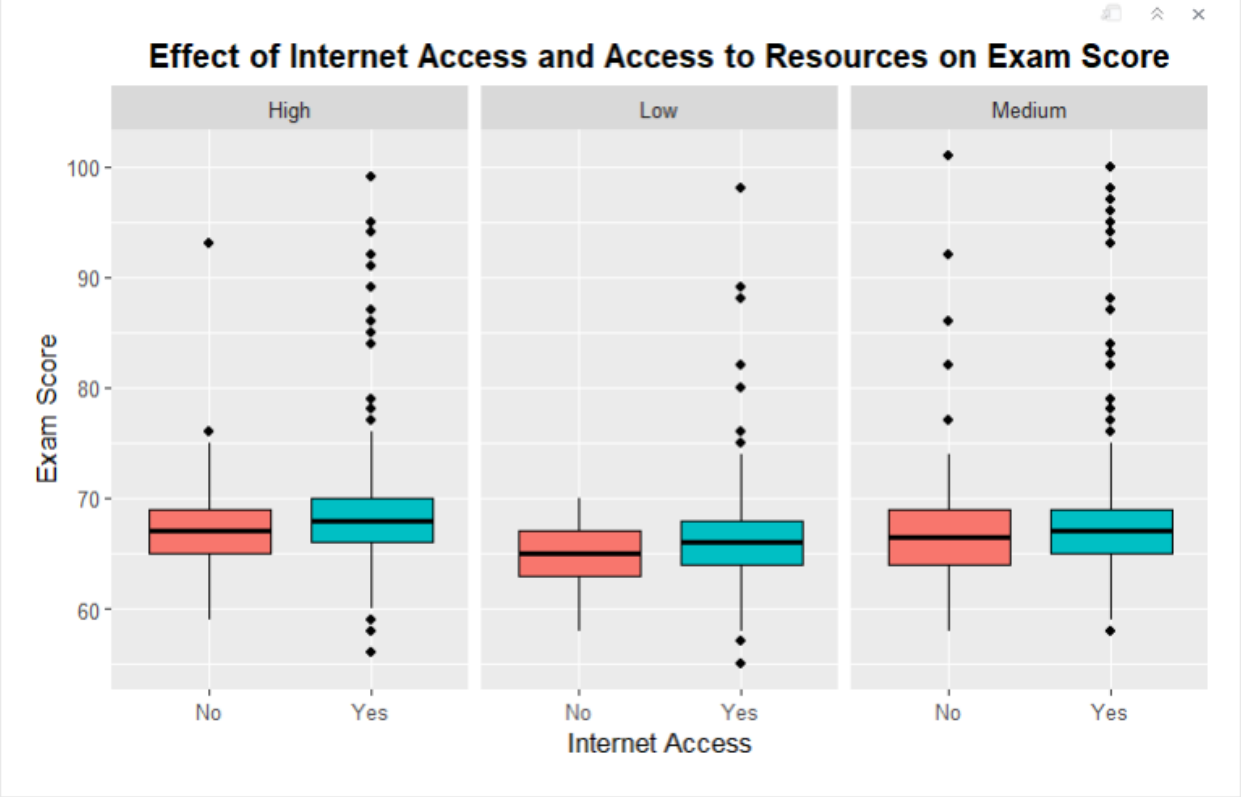


Најбоље резултате постижу студенти који су учили више од 40 сати недељно уз допунске часове (1-3), примећујемо да не постоје студенти који или нису ишли или су ишли на више од 3 допунска часа, а да су притом учили више од 40 сати недељно. Због тога не можемо за категорију од 40+ сати да видимо утицај допунске наставе. За остале интервале (0-20 и 20-40 сати недељно) можемо да приметимо да допунска настава има позитиван утицај на поене на испиту.

5.2.3 Повезаност између *Internet_Access* и *Access_to_Resources*

Занимало нас је да ли комбинација приступа интернету и осталим образовним ресурсима може да има значајнији утицај на резултате испита.

```r
{r}
ggplot(dataset, aes(x = Internet_Access, y = Exam_Score)) +
  geom_boxplot(aes(fill = Internet_Access), color = "black") +
  facet_wrap(~ Access_to_Resources) +
  labs(title = "Effect of Internet Access and Access to Resources on Exam Score",
       x = "Internet Access",
       y = "Exam Score") +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"),
        legend.position = "none")
```



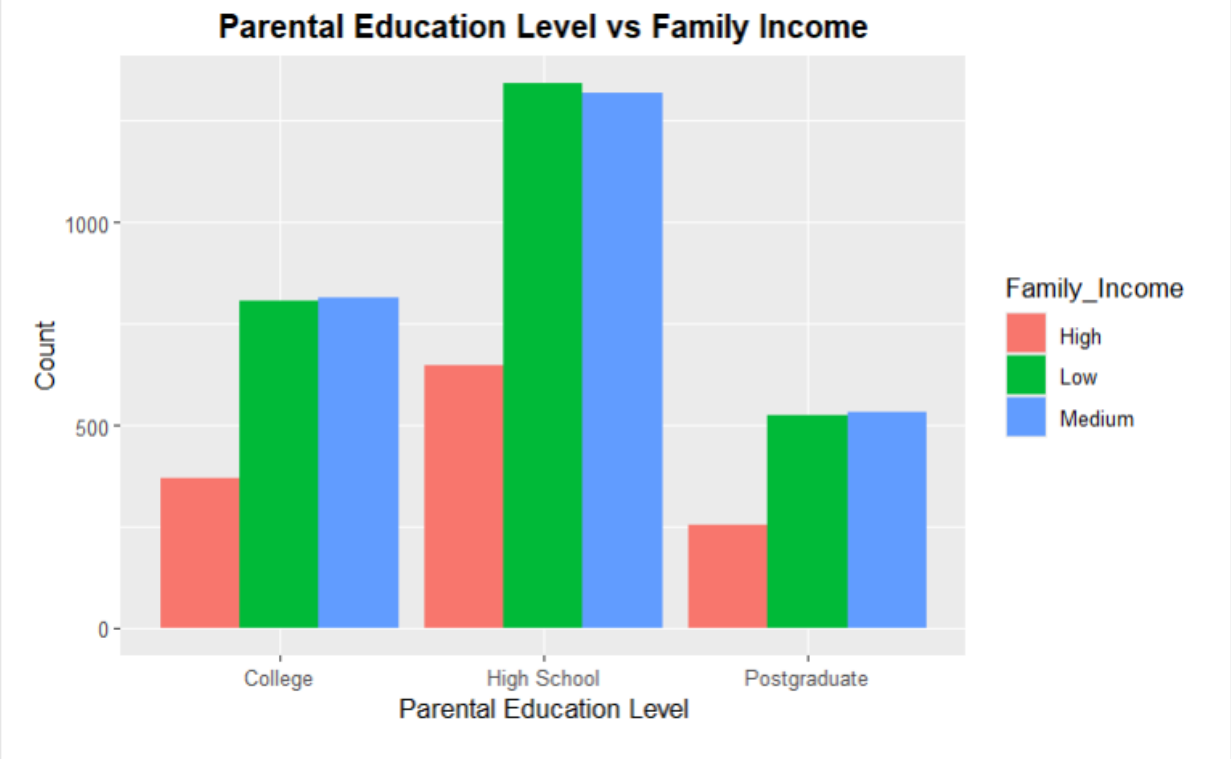
Једино што се истиче на овом графику је то да када студенти немају приступ интернету и имају низак ниво доступности образовним ресурсима, медијана је уочљиво нижа од

осталих комбинација и нема вредности које одступају. Највећи резултат који је постигнут је ~70 поена.

5.2.4 Повезаност између *Parental_Education_Level* и *Family_Income*

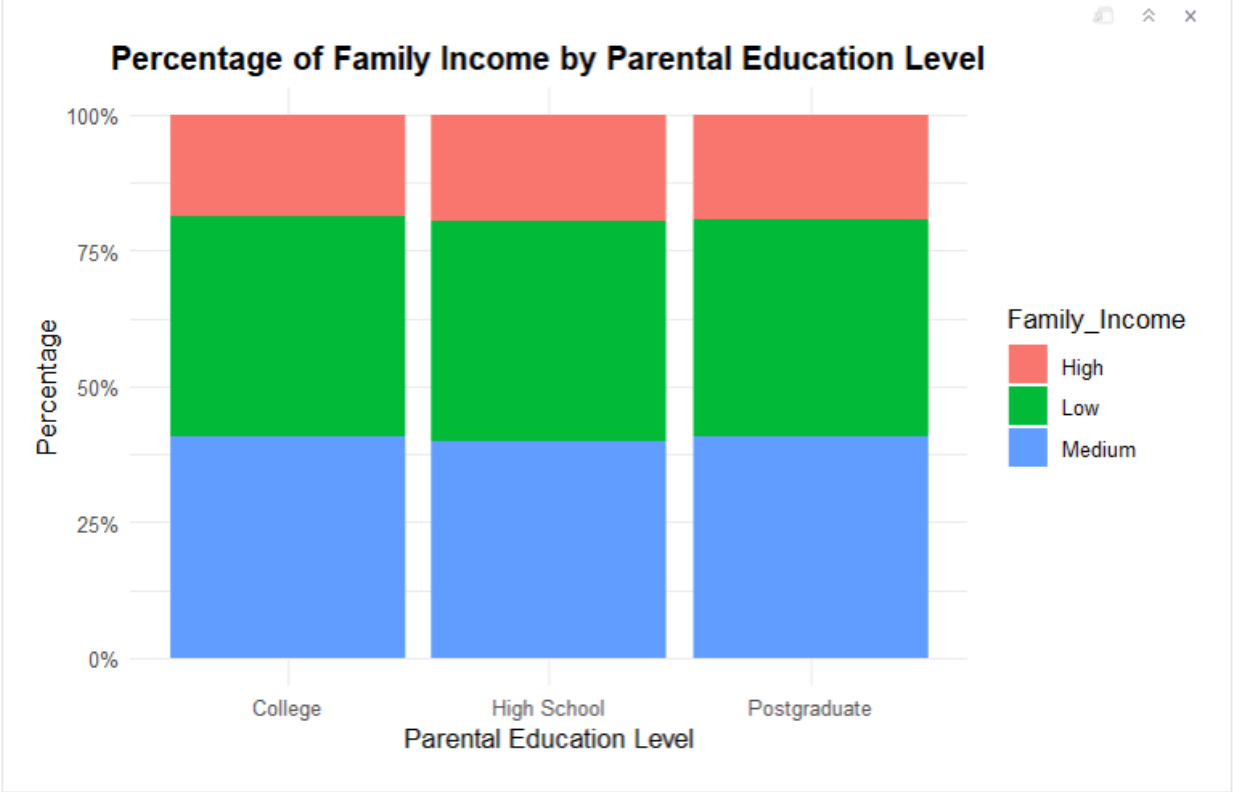
Претпоставили смо да образовање родитеља може утицати на економски статус породице. Виши ниво образовања често је повезан са бољим радним позицијама и већим приходима, па смо желели да испитамо да ли ове две варијабле имају неку корелацију.

```{r}
ggplot(dataset, aes(x = Parental_Education_Level, fill = Family_Income)) + geom_bar
(position = "dodge") + labs(title = "Parental Education Level vs Family Income", x
= "Parental Education Level", y = "Count") + theme(plot.title = element_text(hjust
= 0.5, vjust = 1, face = "bold"))
```



Родитељи са средњошколским образовањем углавном припадају нижим и средњим приходима, иако категорија *Postgraduate* имаја нешто већи број у категорији средњих и високих прихода у поређењу са осталим групама, образовање не гарантује висок приход. Да бисмо мало боље анализирали и упоредили различите категорије, сагледаћемо процентуалну расподелу.

```r
{r}
ggplot(dataset, aes(x = Parental_Education_Level, fill = Family_Income)) +
geom_bar(position = "fill") + labs(title = "Percentage of Family Income by Parental
Education Level", x = "Parental Education Level", y = "Percentage") +
scale_y_continuous(labels = scales::percent) + theme_minimal() + theme(plot.title =
element_text(hjust = 0.5, face = "bold"))
```

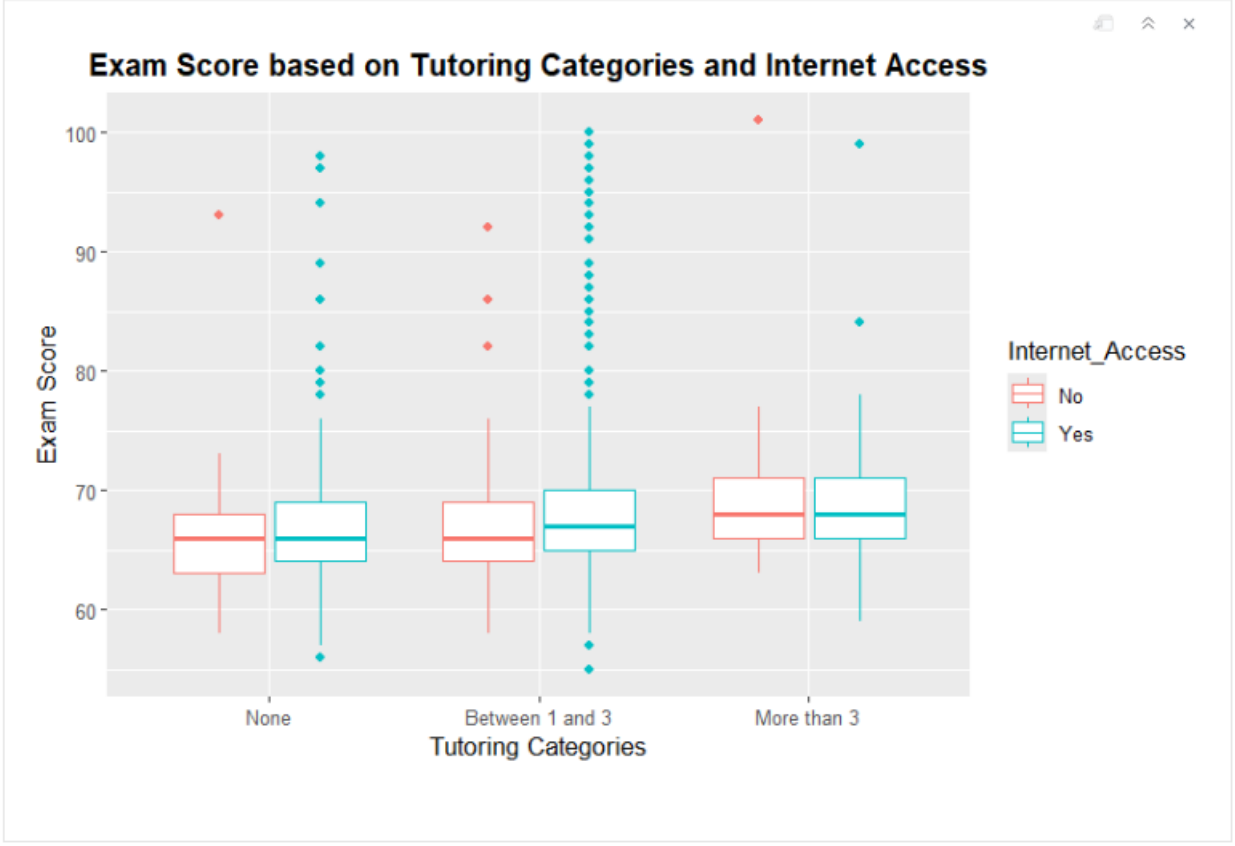


Са овог графика се заиста може уочити да ниво образовања нема устицај на приход. Разлике су минималне.

5.2.5 Повезаност између *Tutoring_Categories* и *Internet_Access*

Раније смо истражили заједнички утицај интернета и приступа образовним материјалима, сада ћемо видети да ли приступ интернету може имати утицај на то да ли студенти имају потребу за допунском наставом.

```r
{r}
ggplot(dataset, aes(x = Tutoring_Categories, y = Exam_Score, color =
Internet_Access)) + geom_boxplot() + labs(title = "Exam Score based on Tutoring
Categories and Internet Access") + xlab("Tutoring Categories") + ylab("Exam Score")
+ theme(plot.title = element_text(hjust = 0.5, vjust = 1, face = "bold"))
```

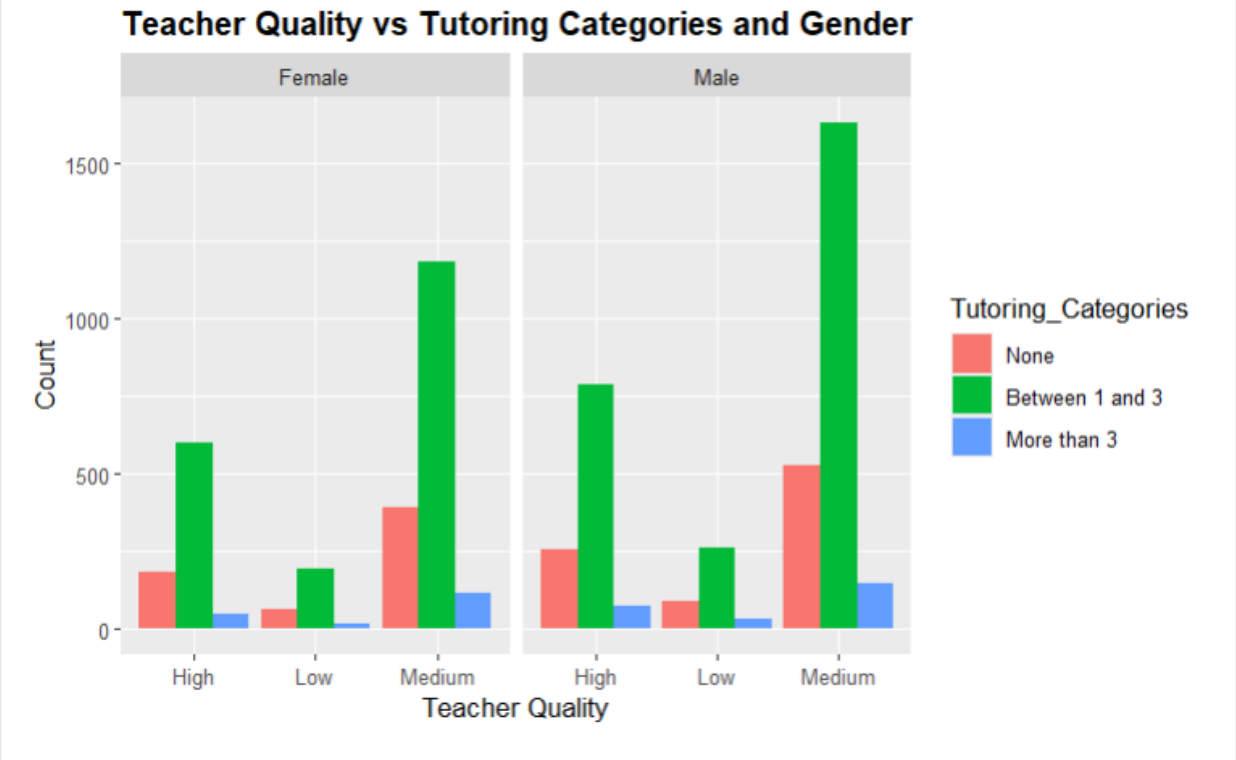


Приметили смо да приступ интернету има значај у случају када студенти немају ниједан допунски час, иако је медијана слична распон поена за оне који имају интернет је доста већи и има више изузетака који постижу јако добре резултате. Слично важи и за категорију студената који имају између 1 и 3 допунска часа, где се и на графику очигледно види да имају већу медијану они који имају приступ интернету, а имају и више изузетака.

5.2.6 Повезаност између *Teacher_Quality*, *Tutoring_Categories* и *Gender*

Замисао је да квалитет професора може да има значајан утицај на број допунских часова, односно, уколико студент сматра да је квалитет професора лош то би значило да му је потребно више допунских часова. И на крају ћемо видети да ли и пол има додатни утицај.

```r
{r}
ggplot(dataset, aes(x = Teacher_Quality, fill = Tutoring_Categories)) + geom_bar
(position = "dodge") + labs(title = "Teacher Quality vs Tutoring Categories and
Gender", x = "Teacher Quality", y = "Count") + facet_wrap(~Gender) + theme(plot
.title = element_text(hjust = 0.5, vjust = 1, face = "bold"))
```

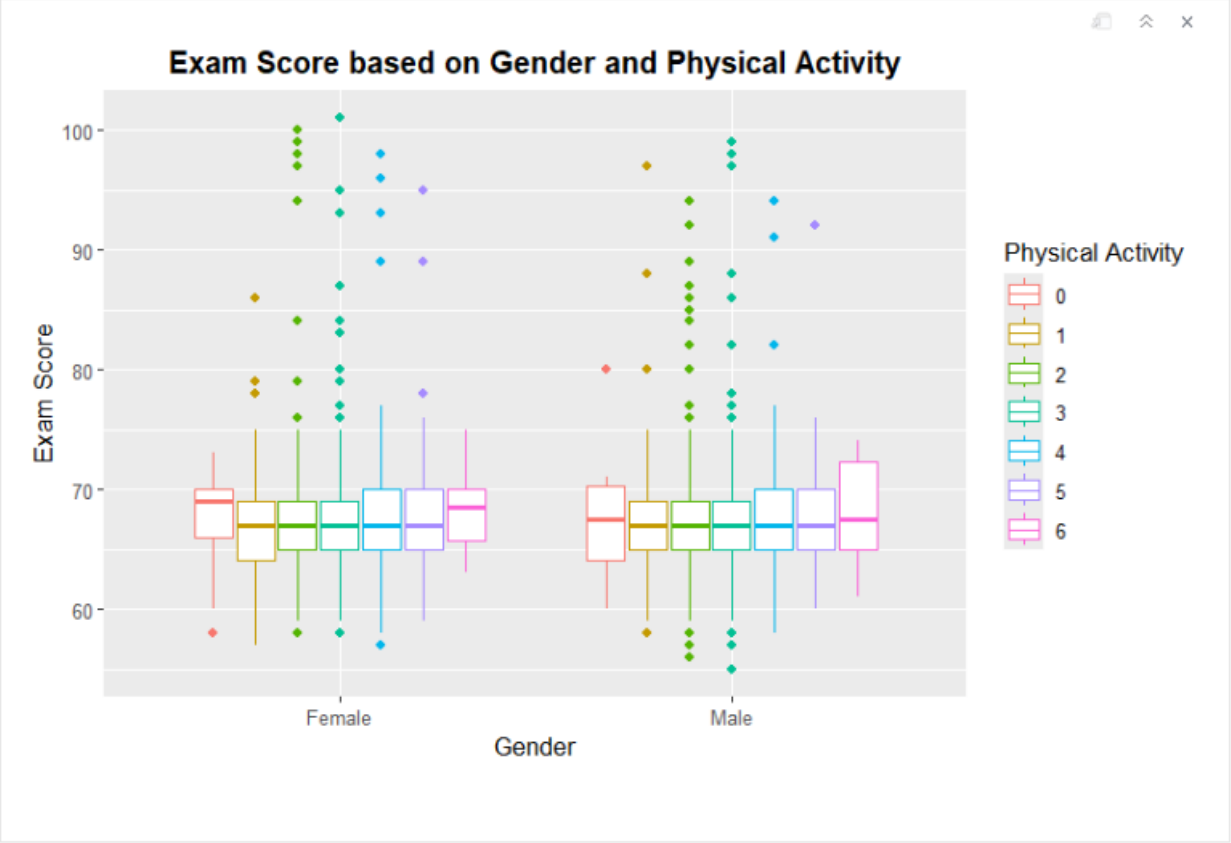


Прво примећујемо да пол нема никакав утицај на број допунских часова, јер је распоред врло сличан. Такође видимо да иако неки студенти оцењују професоре као лоше, не похађају више допунских часова у односу на оне који су их ставили у категорију средњег квалитета. Сматрамо да је повезаност минимална до непостојећа.

5.2.7 Повезаност између *Physical_Activity* и *Gender*

Физичка активност може да утиче на концентрацију и опште ментално здравље, што на крају може утицати и на резултате тестова. Видећемо да ли је то истина и да ли пол има везе са тим.

```r
{r}
ggplot(dataset, aes(x = Gender, y = Exam_Score, color = as.factor(Physical_Activity
))) + geom_boxplot() + labs(title = "Exam Score based on Gender and Physical
Activity", color = "Physical Activity") + xlab("Gender") + ylab("Exam Score") +
theme(plot.title = element_text(hjust = 0.5, vjust = 1, face = "bold"))
```

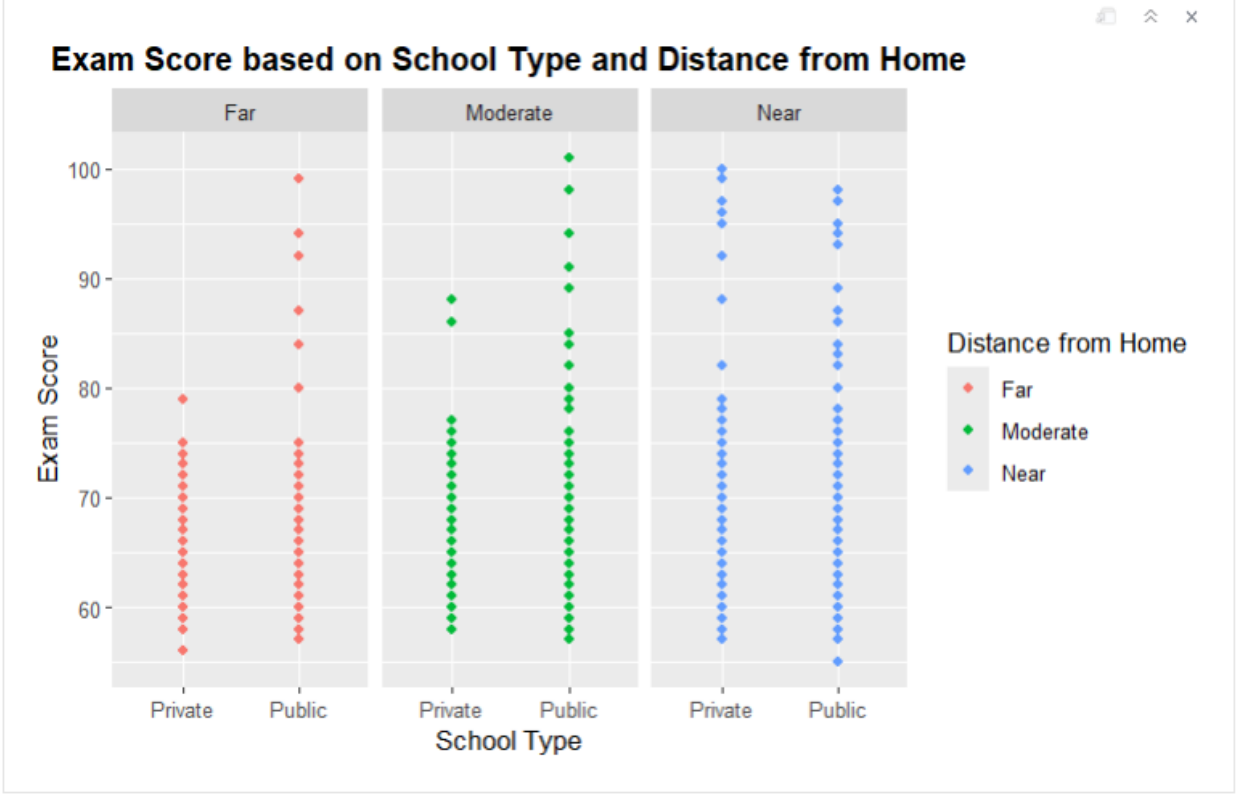


Једина приметна разлика се види код студенткиња које немају физичку активност, оне имају бољу медијану у односу на остале студенткиње. Следеће су студенткиње које имају највишу физичку активност, њихова медијана је већа од остатка групе. Код мушкараца постоји сличан патерн али га је теже уочити, што значи да физичка активност нема велики утицај на њихов успех.

5.2.8 Повезаност између *School_Type* и *Distance_From_Home*

Занима нас да ли удаљеност од куће има утицај на тип школе који студенти уписују, да ли већа удаљеност значи да се студенти пре одлуче за приватни или јавни факултет. Видећемо да ли заједно имају утицај на резултате испита.

```r
ggplot(dataset, aes(x = School_Type, y = Exam_Score)) + geom_point(aes(color =
Distance_from_Home)) + facet_wrap(~ Distance_from_Home) + labs(title = "Exam Score
based on School Type and Distance from Home", color = "Distance from Home") + xlab
("School Type") + ylab("Exam Score") + theme(plot.title = element_text(hjust = 0.5,
vjust = 1, face = "bold"))
```



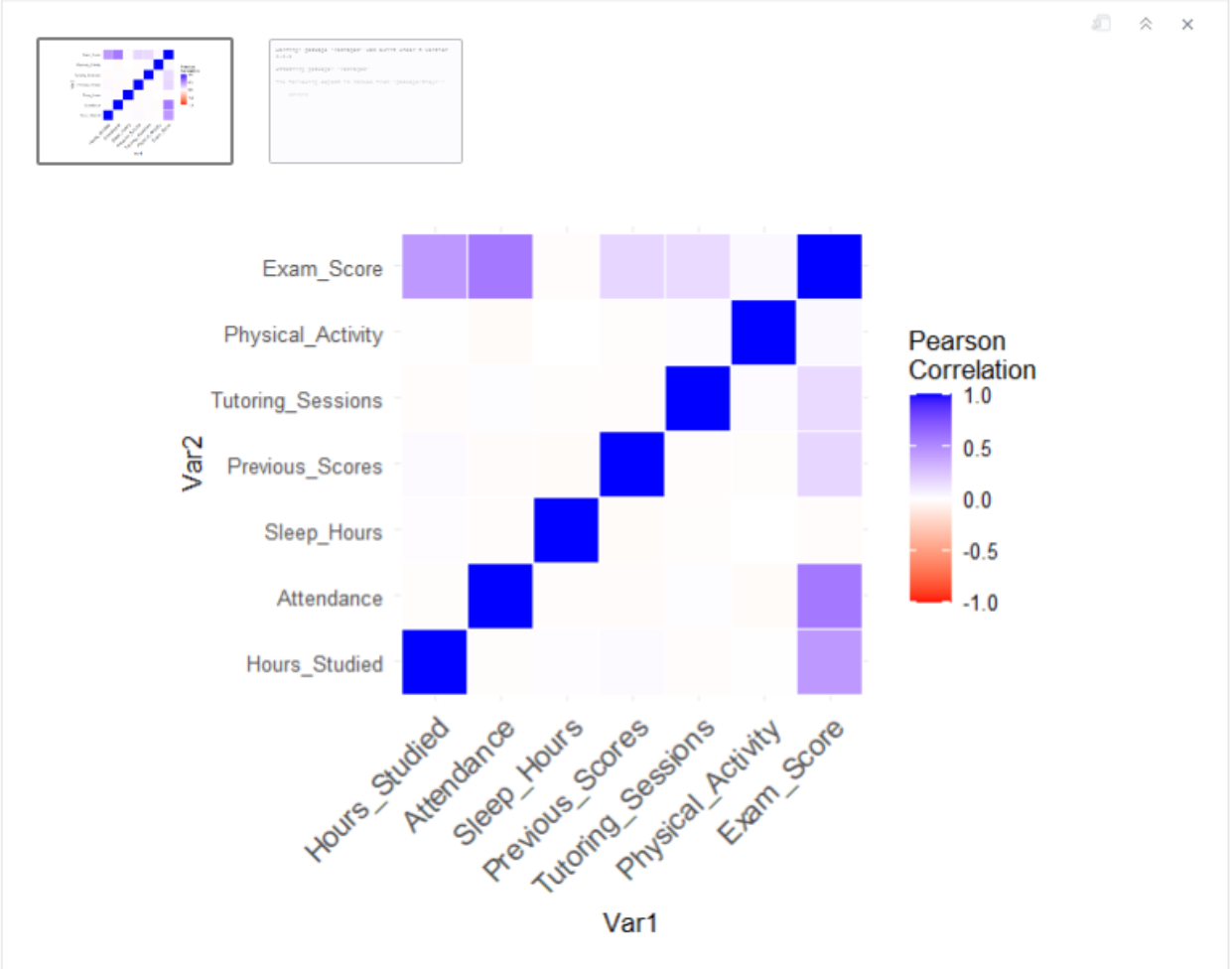
Највећа разлика се види када је удаљеност велика, студенти на приватним факултетима имају нешто лошији просек у односу на студенте са државних факултета. Док, када је у питању мала удаљеност од куће, мало бољи успех имају студенти који су на приватним факултетима. Ове разлике нису много значајне у односу на неке претходне примере, и не указују на јасан тренд који има утицај на успех на испиту.

6 Креирање модела

Након што смо истражили везе између самих предиктора, као и између предиктора и одговора започећемо креирање модела.

Прво ћемо приказати матрицу корелације како бисмо имали бољи увид у корелацију између нумеричких колона.

```{r}
numeric_df <- dataset[sapply(dataset, is.numeric)]
correlation_matrix <- cor(numeric_df)
melted_corr <- melt(correlation_matrix)
ggplot(data = melted_corr, aes(x = Var1, y = Var2, fill = value)) + geom_tile(color
= "white") + scale_fill_gradient2(low = "red", high = "blue", mid = "white",
midpoint = 0, limit = c(-1, 1), space = "Lab", name = "Pearson Correlation") +
theme_minimal() + theme(axis.text.x = element_text(angle = 45, vjust = 1, size = 12
, hjust = 1)) + coord_fixed()
```



Примећујемо да *Exam_Score* има најбољу корелацију са *Attendance*, а потом са *Hours_Studied*. Нешто мању корелацију има са *Previous_Scores* и *Tutoring_Sessions*.

Затим ћемо трансформисати све категоријске колоне у фактор варијабле.

```{r}
dataset$Parental_Involvement <- factor(dataset$Parental_Involvement)
dataset$Access_to_Resources <- factor(dataset$Access_to_Resources)
dataset$Extracurricular_Activities <- factor(dataset$Extracurricular_Activities)
dataset$Motivation_Level <- factor(dataset$Motivation_Level)
dataset$Internet_Access <- factor(dataset$Internet_Access)
dataset$Family_Income <- factor(dataset$Family_Income)
dataset$Teacher_Quality <- factor(dataset$Teacher_Quality)
dataset$School_Type <- factor(dataset$School_Type)
dataset$Peer_Influence <- factor(dataset$Peer_Influence)
dataset$Learning_Disabilities <- factor(dataset$Learning_Disabilities)
dataset$Parental_Education_Level <- factor(dataset$Parental_Education_Level)
dataset$Distance_from_Home <- factor(dataset$Distance_from_Home)
dataset$Gender <- factor(dataset$Gender)
str(dataset)
```

Потом се уверавамо да су колоне заправо конвертоване.

```
$ Hours_Studied             : int  23 19 24 29 19 19 29 25 17 23 ...
$ Attendance                : int  84 64 98 89 92 88 84 78 94 98 ...
$ Parental_Involvement      : Factor w/ 3 levels "High","Low","Medium": 2 2 3 2 3 3 3 2 3 3 ...
$ Access_to_Resources       : Factor w/ 3 levels "High","Low","Medium": 1 3 3 3 3 3 2 1 1 3 ...
$ Extracurricular_Activities: Factor w/ 2 levels "No","Yes": 1 1 2 2 2 2 2 2 1 2 ...
$ Sleep_Hours               : int  7 8 7 8 6 8 7 6 6 8 ...
$ Previous_Scores           : int  73 59 91 98 65 89 68 50 80 71 ...
$ Motivation_Level          : Factor w/ 3 levels "High","Low","Medium": 2 2 3 3 3 3 2 3 1 3 ...
$ Internet_Access           : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
$ Tutoring_Sessions         : int  0 2 2 1 3 3 1 1 0 0 ...
$ Family_Income             : Factor w/ 3 levels "High","Low","Medium": 2 3 3 3 3 3 2 1 3 1 ...
$ Teacher_Quality           : Factor w/ 3 levels "High","Low","Medium": 3 3 3 3 1 3 3 1 2 1 ...
$ School_Type               : Factor w/ 2 levels "Private","Public": 2 2 2 2 2 2 1 2 1 2 ...
$ Peer_Influence            : Factor w/ 3 levels "Negative","Neutral",..: 3 1 2 1 2 3 2 1 2 3 ...
$ Physical_Activity         : int  3 4 4 4 4 3 2 2 1 5 ...
$ Learning_Disabilities     : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
$ Parental_Education_Level  : Factor w/ 3 levels "College","High School",..: 2 1 3 2 1 3 2 2 1 2 ...
$ Distance_from_Home        : Factor w/ 3 levels "Far","Moderate",..: 3 2 3 2 3 3 2 1 3 2 ...
$ Gender                    : Factor w/ 2 levels "Female","Male": 2 1 2 2 1 2 2 2 2 2 ...
$ Exam_Score                : int  67 61 74 71 70 71 67 66 69 72 ...
$ Tutoring_Categories       : Factor w/ 3 levels "None","Between 1 and 3",..: 1 2 2 2 2 2 2 2 1 1 ...
```

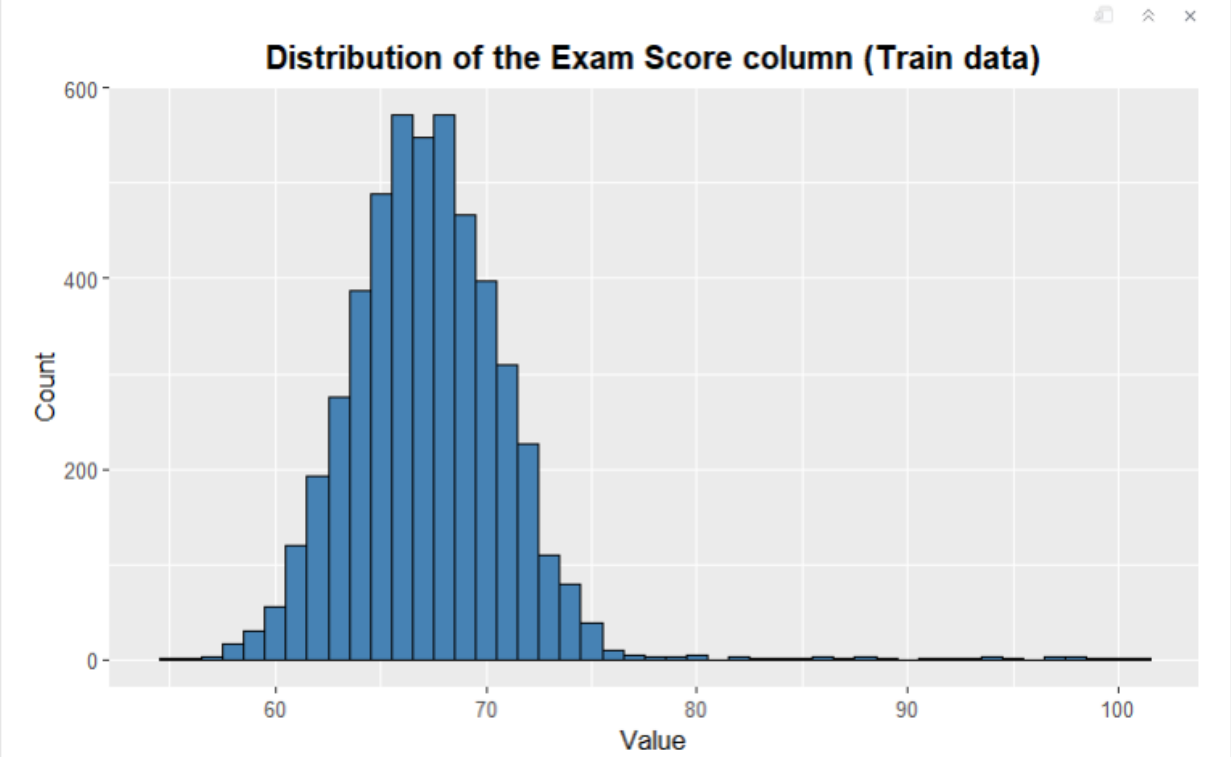
Наредни корак јесте да наш *dataset* поделимо да тренинг и тест скупове, у односу 75:25.

```{r}
set.seed(14)
train_indices <- sample(1:nrow(dataset), size = 0.75 * nrow(dataset))
train_data <- dataset[train_indices, ]
test_data <- dataset[-train_indices, ]
cat("Veličina trening skupa:", nrow(train_data), "\n")
cat("Veličina test skupa:", nrow(test_data), "\n")
```

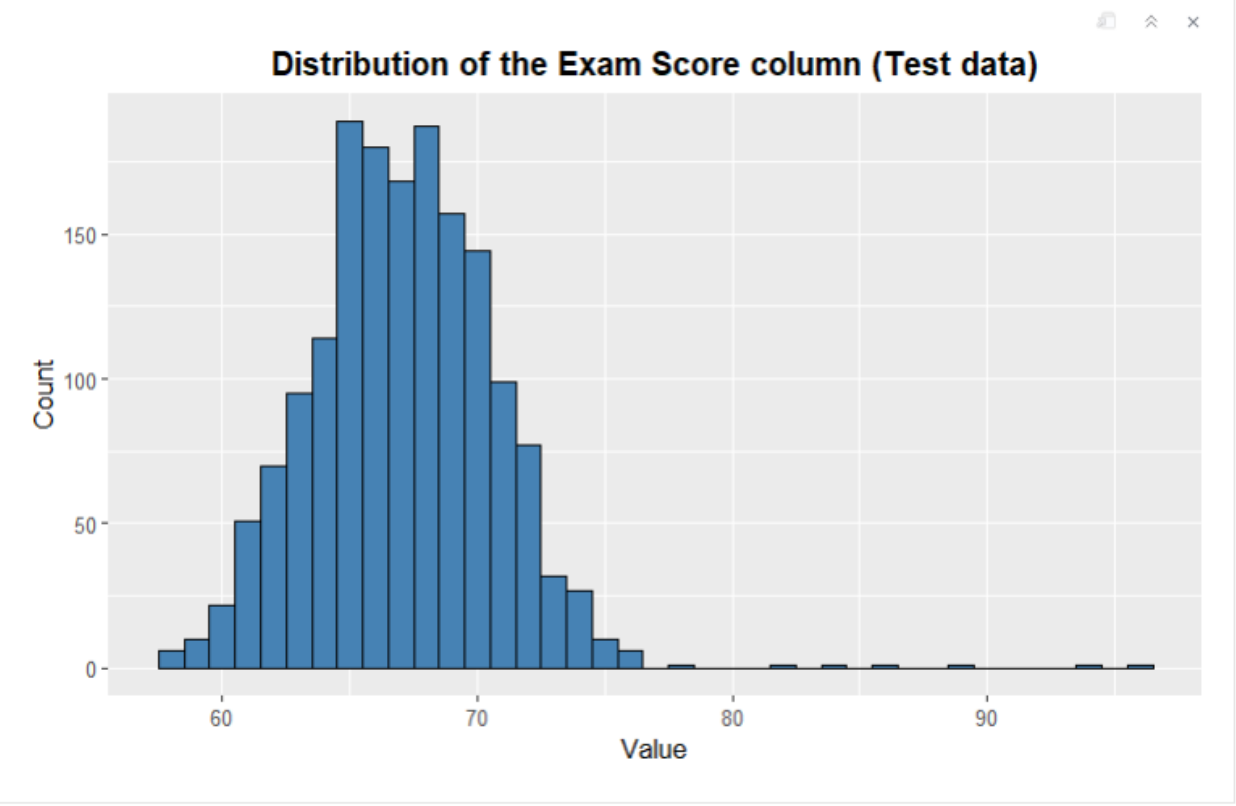
```
Veličina trening skupa: 4950
Veličina test skupa: 1651
```

Приказаћемо расподелу резултата испита у тренинг и тест скупу понаособ.

```{r}
ggplot(train_data, aes(x = Exam_Score)) + geom_histogram(fill = "steelblue", color = "black", binwidth = 1) + labs(title = "Distribution of the Exam Score column (Train data)") + xlab("Value") + ylab("Count") + theme(plot.title = element_text(hjust = 0.5, vjust = 1, face = "bold"))
```



```r
{r}
ggplot(test_data, aes(x = Exam_Score)) + geom_histogram(fill = "steelblue", color = "black", binwidth = 1) + labs(title = "Distribution of the Exam Score column (Test data)") + xlab("Value") + ylab("Count") + theme(plot.title = element_text(hjust = 0.5, vjust = 1, face = "bold"))
```



Помоћу графика боље видимо расподеле резултата испита унутар оба скупа и можемо да се уверимо да су слично подељени и да је све спремно за даљу анализу.

6.1 Линеарна регресија (*Linear Regression*)

Користићемо линеарну регресију како бисмо анализирали однос између предиктора и одговора. Прво ћемо у модел убацити предикторе за које смо се уверили да имају највећи утицај.

```
{r}
lm.fit1 <- lm(Exam_Score ~ Hours_Studied + Attendance, data = train_data)
summary(lm.fit1)
```

```
Call:
lm(formula = Exam_Score ~ Hours_Studied + Attendance, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-6.0861 -1.3556 -0.1956  1.0192 31.5333

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   45.783576   0.310051  147.66   <2e-16 ***
Hours_Studied  0.289995   0.006599   43.94   <2e-16 ***
Attendance     0.196215   0.003432   57.18   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.795 on 4947 degrees of freedom
Multiple R-squared:  0.5069,    Adjusted R-squared:  0.5067
F-statistic:  2542 on 2 and 4947 DF,  p-value: < 2.2e-16
```

Тачност модела је ~51% што и није толико добар резултат, али је F-статистика > 1 и p-вредност < 0.5 из чега видимо да постоји повезаност између предиктора и одговора. Покушаћемо да убацимо и колоне *Previous_Scores* и *Tutoring_Sessions* јер су оне следеће дале најбоље резултате и надамо се да ће се тачност модела повећати.

```r
{r}
lm.fit2 <- lm(Exam_Score ~ Hours_Studied + Attendance + Previous_Scores +
Tutoring_Sessions, data = train_data)
summary(lm.fit2)
```

```
Call:
lm(formula = Exam_Score ~ Hours_Studied + Attendance + Previous_Scores +
    Tutoring_Sessions, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6236 -1.1697 -0.1863  0.8279 30.8757

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       41.331071   0.357912  115.48   <2e-16 ***
Hours_Studied      0.288125   0.006236   46.21   <2e-16 ***
Attendance         0.197543   0.003242   60.92   <2e-16 ***
Previous_Scores    0.048330   0.002608   18.54   <2e-16 ***
Tutoring_Sessions  0.499220   0.030393   16.43   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.64 on 4945 degrees of freedom
Multiple R-squared:  0.5602,    Adjusted R-squared:  0.5599
F-statistic:  1575 on 4 and 4945 DF,  p-value: < 2.2e-16
```

Тачност се повећала на ~56% што је боље, такође видимо да *Adjusted* R^2 не одступа превише од *Multiple* R^2 што нам је битно да не бисмо *overfit*-овали податке, а и стандардна грешка се смањила.

У модел ћемо додати још две колоне *Parental_Involvement* и *Access_to_Resources*. Како бисмо покушали да повећамо тачност.

```r
{r}
lm.fit3 <- lm(Exam_Score ~ Hours_Studied + Attendance + Previous_Scores +
Tutoring_Sessions + Parental_Involvement + Access_to_Resources, data = train_data)
summary(lm.fit3)
```

```
Call:
lm(formula = Exam_Score ~ Hours_Studied + Attendance + Previous_Scores +
    Tutoring_Sessions + Parental_Involvement + Access_to_Resources,
    data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3779 -0.9000 -0.1729  0.5338 31.3496

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               43.000377   0.340046  126.45   <2e-16 ***
Hours_Studied              0.291729   0.005786   50.42   <2e-16 ***
Attendance                 0.198258   0.003008   65.92   <2e-16 ***
Previous_Scores            0.048628   0.002419   20.10   <2e-16 ***
Tutoring_Sessions          0.507184   0.028190   17.99   <2e-16 ***
Parental_InvolvementLow   -2.015250   0.100493  -20.05   <2e-16 ***
Parental_InvolvementMedium -1.059129   0.081198  -13.04   <2e-16 ***
Access_to_ResourcesLow     -2.060387   0.100680  -20.46   <2e-16 ***
Access_to_ResourcesMedium  -0.946026   0.080303  -11.78   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.448 on 4941 degrees of freedom
Multiple R-squared:  0.6221,    Adjusted R-squared:  0.6215
F-statistic:  1017 on 8 and 4941 DF,  p-value: < 2.2e-16
```

Сада тачност износи ~62% што је доста боље у односу на почетни модел. Свих осам предиктора су статистички значајни, са p-вредношћу мањом од 2e-16. Повећањем *Hours_Studied* и *Tutoring_Sessions* се повећавају и резултати испита, док негативни коефицијенти за *Parental_Involvement* и *Access_to_Resources* указују на смањење резултата.

Пробаћемо још мало да побољшамо модел.

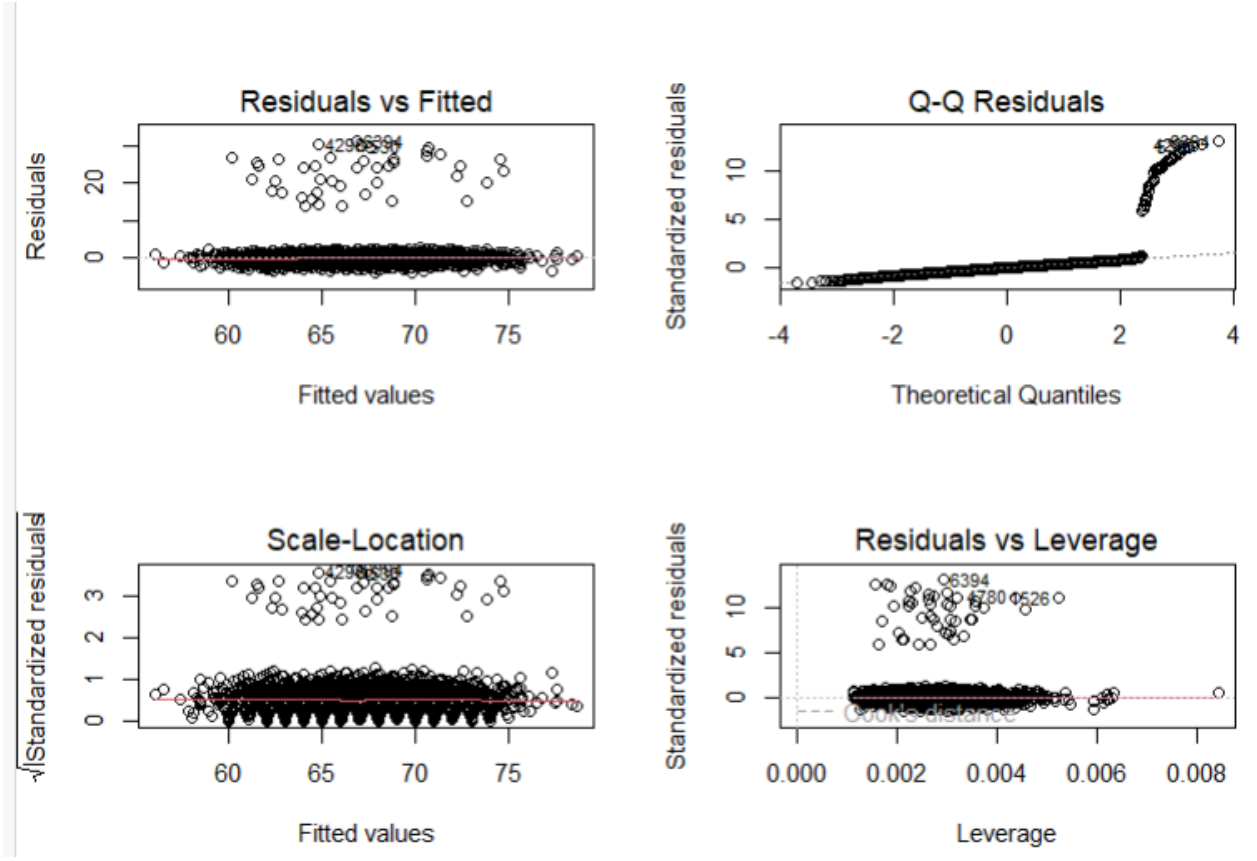
```r
{r}
lm.fit4 <- lm(Exam_Score ~ Hours_Studied + Attendance + Previous_Scores +
Tutoring_Sessions + Parental_Involvement + Access_to_Resources + Family_Income +
Motivation_Level, data = train_data)
summary(lm.fit4)
```

```
Call:
lm(formula = Exam_Score ~ Hours_Studied + Attendance + Previous_Scores +
    Tutoring_Sessions + Parental_Involvement + Access_to_Resources +
    Family_Income + Motivation_Level, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8346 -0.8135 -0.1581  0.4776 31.1768

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                44.229322   0.344178 128.507  < 2e-16 ***
Hours_Studied               0.291957   0.005640  51.765  < 2e-16 ***
Attendance                  0.198754   0.002931  67.817  < 2e-16 ***
Previous_Scores             0.048754   0.002358  20.675  < 2e-16 ***
Tutoring_Sessions           0.504041   0.027464  18.353  < 2e-16 ***
Parental_InvolvementLow    -2.031292   0.097918 -20.745  < 2e-16 ***
Parental_InvolvementMedium -1.078655   0.079138 -13.630  < 2e-16 ***
Access_to_ResourcesLow     -2.049700   0.098120 -20.890  < 2e-16 ***
Access_to_ResourcesMedium  -0.962180   0.078261 -12.294  < 2e-16 ***
Family_IncomeLow           -1.106481   0.094236 -11.742  < 2e-16 ***
Family_IncomeMedium        -0.572366   0.094035  -6.087 1.24e-09 ***
Motivation_LevelLow        -1.073897   0.098140 -10.942  < 2e-16 ***
Motivation_LevelMedium     -0.531422   0.089166  -5.960 2.70e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.385 on 4937 degrees of freedom
Multiple R-squared:  0.6416,    Adjusted R-squared:  0.6407
F-statistic: 736.5 on 12 and 4937 DF,  p-value: < 2.2e-16
```

Додавањем *Family_Income* и *Motivation_Level* добили смо нешто бољи модел са тачношћу од ~64%, што је одлично у односу на почетни модел.

Приказаћемо графички наш модел.

```r
{r}
par(mfrow=c(2,2))
plot(lm.fit4)
```



Помоћу *Residuals vs Fitted* графика проверавамо линеаран однос између предиктора и зависне промењиве. Идеално, резидуали треба да буду равномерно распоређени око хоризонталне линије без јасних образаца. Овде је већина тачака распоређена око линије, али постоје и неке тачке које одступају.

За *Normal Q-Q* график важи да уколико резидуали прате линију онда су нормално распоређени. Видимо да има одступања на крајевима, вероватно због екстремних вредности.

Scale-Location график служи за проверу хомоскедастичности. Потребно је да тачке буду једнако распршене око линије. Има доста одступања па можемо да кажемо да варијанса резидуала није константна.

Residuals vs Leverage график се користи за откривање *outlier-a*. Као што смо и претпоставили, на основу *Normal Q-Q* графика, имамо неколико екстремних вредности.

6.2 Стабло одлучивања (*Decision Tree*)

Стабло одлучивања омогућава моделирање сложенијих и нелинеарних односа између променљивих, па је корисно ако линеарна регресија не даје довољно прецизне резултате.

```r
{r}
tree_model <- rpart(Exam_Score ~ Hours_Studied + Attendance + Previous_Scores +
                    Sleep_Hours + Tutoring_Sessions + Parental_Involvement +
                    Access_to_Resources + Motivation_Level + Family_Income +
                    Teacher_Quality,
                    data = train_data, method = "anova")
tree_model
```

```
n= 4950

node), split, n, deviance, yval
      * denotes terminal node

 1) root 4950 78343.500 67.28727
   2) Attendance< 82.5 2794 33140.840 65.58805
     4) Hours_Studied< 20.5 1475 14952.920 64.31254
       8) Attendance< 69.5 638  6136.171 63.03605 *
       9) Attendance>=69.5 837  6984.755 65.28554 *
     5) Hours_Studied>=20.5 1319 13104.730 67.01440
      10) Attendance< 72.5 737  7113.373 66.05970 *
      11) Attendance>=72.5 582  4468.962 68.22337 *
   3) Attendance>=82.5 2156 26680.750 69.48933
     6) Hours_Studied< 21.5 1323 12620.670 68.30159
      12) Hours_Studied< 13.5 308  2021.880 66.60065 *
      13) Hours_Studied>=13.5 1015  9437.281 68.81773 *
     7) Hours_Studied>=21.5 833  9229.390 71.37575
      14) Attendance< 94.5 579  4146.418 70.68739 *
      15) Attendance>=94.5 254  4183.228 72.94488 *
```

Прва подела у стаблу је на основу *Attendance* < 82.5, где се подаци деле на оне са вредностима мањим од 82.5 и оне са вредностима већим или једнаким од 82.5, затим је подела према *Hours_Studied* итд. На пример, у чвору где је *Attendance* < 69.5 и *Hours_Studied* < 20.5, просечан резултат испита је ~63.

```r
summary(tree_model)
```

```
Call:
rpart(formula = Exam_Score ~ Hours_Studied + Attendance + Previous_Scores +
    Sleep_Hours + Tutoring_Sessions + Parental_Involvement +
    Access_to_Resources + Motivation_Level + Family_Income +
    Teacher_Quality, data = train_data, method = "anova")
  n= 4950

          CP nsplit rel error    xerror       xstd
1 0.23641914      0 1.0000000 1.0005161 0.05254582
2 0.06488345      1 0.7635809 0.7661485 0.04976860
3 0.06166048      2 0.6986974 0.7167918 0.05008614
4 0.02338410      3 0.6370369 0.6400654 0.04854360
5 0.01943226      4 0.6136528 0.6201609 0.04873305
6 0.01482581      5 0.5942206 0.6006503 0.04893268
7 0.01148460      6 0.5793948 0.5897399 0.04923622
8 0.01000000      7 0.5679102 0.5804390 0.04864995

Variable importance
   Attendance Hours_Studied
           67            33

Node number 1: 4950 observations,    complexity param=0.2364191
  mean=67.28727, MSE=15.82697
  left son=2 (2794 obs) right son=3 (2156 obs)
  Primary splits:
      Attendance           < 82.5 to the left,  improve=0.23641910, (0 missing)
      Hours_Studied        < 20.5 to the left,  improve=0.11648970, (0 missing)
      Previous_Scores      < 83.5 to the left,  improve=0.02480398, (0 missing)
      Access_to_Resources  splits as  RLL,      improve=0.02089025, (0 missing)
      Parental_Involvement splits as  RLL,      improve=0.01785503, (0 missing)
  Surrogate splits:
      Hours_Studied     < 2.5  to the right, agree=0.565, adj=0.001, (0 split)
      Tutoring_Sessions < 5.5  to the left,  agree=0.565, adj=0.001, (0 split)

Node number 2: 2794 observations,    complexity param=0.06488345
  mean=65.58805, MSE=11.86143
  left son=4 (1475 obs) right son=5 (1319 obs)
  Primary splits:
      Hours_Studied        < 20.5 to the left,  improve=0.15338160, (0 missing)
      Attendance           < 69.5 to the left,  improve=0.10318070, (0 missing)
      Access_to_Resources  splits as  RLL,      improve=0.02645182, (0 missing)
      Previous_Scores      < 74.5 to the left,  improve=0.02551846, (0 missing)
      Parental_Involvement splits as  RLR,      improve=0.01907318, (0 missing)
  Surrogate splits:
      Access_to_Resources  splits as  LRL,      agree=0.533, adj=0.011, (0 split)
      Parental_Involvement splits as  LRL,      agree=0.531, adj=0.007, (0 split)
      Sleep_Hours          < 9.5  to the left,  agree=0.530, adj=0.005, (0 split)
```

```
Node number 3: 2156 observations,    complexity param=0.06166048
  mean=69.48933, MSE=12.37512
  left son=6 (1323 obs) right son=7 (833 obs)
  Primary splits:
      Hours_Studied        < 21.5 to the left,  improve=0.18105550, (0 missing)
      Attendance           < 89.5 to the left,  improve=0.05843870, (0 missing)
      Previous_Scores      < 75.5 to the left,  improve=0.04632299, (0 missing)
      Access_to_Resources  splits as  RLR,      improve=0.03492933, (0 missing)
      Parental_Involvement splits as  RLL,      improve=0.02465049, (0 missing)
  Surrogate splits:
      Tutoring_Sessions < 5.5  to the left,  agree=0.616, adj=0.006, (0 split)

Node number 4: 1475 observations,    complexity param=0.0233841
  mean=64.31254, MSE=10.13757
  left son=8 (638 obs) right son=9 (837 obs)
  Primary splits:
      Attendance          < 69.5 to the left,  improve=0.12251740, (0 missing)
      Hours_Studied       < 15.5 to the left,  improve=0.07139868, (0 missing)
      Previous_Scores     < 74.5 to the left,  improve=0.03631959, (0 missing)
      Access_to_Resources splits as  RLL,      improve=0.02512834, (0 missing)
      Motivation_Level    splits as  RLR,      improve=0.02061749, (0 missing)
  Surrogate splits:
      Hours_Studied   < 7.5  to the left,  agree=0.576, adj=0.020, (0 split)
      Previous_Scores < 50.5 to the left,  agree=0.569, adj=0.005, (0 split)
```

```
Node number 5: 1319 observations,    complexity param=0.01943226
  mean=67.0144, MSE=9.93535
  left son=10 (737 obs) right son=11 (582 obs)
  Primary splits:
      Attendance           < 72.5 to the left,  improve=0.11617110, (0 missing)
      Hours_Studied        < 27.5 to the left,  improve=0.07521435, (0 missing)
      Parental_Involvement splits as  RLR,      improve=0.04010902, (0 missing)
      Access_to_Resources  splits as  RLL,      improve=0.03988631, (0 missing)
      Previous_Scores      < 83.5 to the left,  improve=0.03429267, (0 missing)
  Surrogate splits:
      Previous_Scores < 98.5 to the left,  agree=0.564, adj=0.012, (0 split)
      Hours_Studied   < 36.5 to the left,  agree=0.561, adj=0.005, (0 split)

Node number 6: 1323 observations,    complexity param=0.01482581
  mean=68.30159, MSE=9.539431
  left son=12 (308 obs) right son=13 (1015 obs)
  Primary splits:
      Hours_Studied        < 13.5 to the left,  improve=0.09203206, (0 missing)
      Attendance           < 87.5 to the left,  improve=0.06386966, (0 missing)
      Access_to_Resources  splits as  RLL,      improve=0.05054173, (0 missing)
      Previous_Scores      < 71.5 to the left,  improve=0.04016879, (0 missing)
      Parental_Involvement splits as  RLL,      improve=0.03184620, (0 missing)
```

```
Node number 7: 833 observations,    complexity param=0.0114846
  mean=71.37575, MSE=11.0797
  left son=14 (579 obs) right son=15 (254 obs)
  Primary splits:
      Attendance          < 94.5 to the left,  improve=0.09748681, (0 missing)
      Previous_Scores     < 84.5 to the left,  improve=0.08144272, (0 missing)
      Hours_Studied       < 26.5 to the left,  improve=0.07388407, (0 missing)
      Parental_Involvement splits as  RLL,     improve=0.03477103, (0 missing)
      Access_to_Resources  splits as  RLR,     improve=0.03444381, (0 missing)
  Surrogate splits:
      Tutoring_Sessions < 6.5  to the left,  agree=0.697, adj=0.008, (0 split)

Node number 8: 638 observations
  mean=63.03605, MSE=9.617823

Node number 9: 837 observations
  mean=65.28554, MSE=8.344988

Node number 10: 737 observations
  mean=66.0597, MSE=9.651795

Node number 11: 582 observations
  mean=68.22337, MSE=7.678629

Node number 12: 308 observations
  mean=66.60065, MSE=6.564545

Node number 13: 1015 observations
  mean=68.81773, MSE=9.297814

Node number 14: 579 observations
  mean=70.68739, MSE=7.161344

Node number 15: 254 observations
  mean=72.94488, MSE=16.4694
```

На основу *summary* функције, очиглендно је да су најважнији предиктори *Attendance* (67%) и *Hours_Studied* (33%). Модел се побољшава кроз неколико фаза цепања, а варијанса (MSE) у резидуалима се смањује, па можемо да кажемо да су ови предиктори релевантни за предвиђање резултата.

Следећи корак је израчунавање метрика тачности модела као што су: *RMSE* (Root Mean Squared Error), *MAE* (Mean Absolute Error), *MSE* (Mean Squared Error), и R^2 (*coefficient of determination*). На основу њих процењујемо колико добро модел предвиђа резултате испита.

```r
{r}
tree_predictions <- predict(tree_model, newdata = test_data)
# RMSE
tree_rmse <- sqrt(mean((tree_predictions - test_data$Exam_Score)^2))
# MAE (Mean Absolute Error)
tree_mae <- mean(abs(tree_predictions - test_data$Exam_Score))
# MSE (Mean Squared Error)
tree_mse <- mean((tree_predictions - test_data$Exam_Score)^2)
# R-squared (R²)
ss_total <- sum((test_data$Exam_Score - mean(test_data$Exam_Score))^2)
ss_residual <- sum((test_data$Exam_Score - tree_predictions)^2)
tree_r2 <- 1 - (ss_residual / ss_total)

cat("RMSE:", tree_rmse, "\n")
cat("MAE:", tree_mae, "\n")
cat("MSE:", tree_mse, "\n")
cat("R²:", tree_r2, "\n")
```

```
RMSE: 2.475139
MAE: 1.78487
MSE: 6.126313
R²: 0.5302175
```

RMSE износи ~2.47, што значи да предвиђене вредности у просеку одступају од стварних за око 2.48 поена. *MAE* је ~1.78, значи да је просечно апсолутно одступање предвиђених од стварних вредности релативно мало. *MSE* је 6.12, с обзиром да укључује квадрирање, корисна је да наглси веће грешке. R^2 је ~0.53, што значи да модел објашњава око 53% варијације у резултатима испита.

```r
{r}
print(tree_model$variable.importance)
```

```
          Attendance        Hours_Studied    Tutoring_Sessions  Access_to_Resources
Parental_Involvement
         22776.02968          11146.34924             61.85297             53.95356
34.68443
     Previous_Scores          Sleep_Hours
            26.92493             23.12296
```

Видимо да су *Hours_Studied* и *Attendance* најбитнији предиктори и да додавање нових заправо и не прави разлику јер се стабло одлучивања ослања само на ова два предиктора. Међутим исти резултати се добијају и када су само *Hours_Studied* и *Attendance* укључени у модел, па га нећемо приказивати.

6.3 *Random Forest*

Покушаћемо да направимо и *Random Forest* модел, он прави више стабала одлучивања и комбинује њихове предвиђене вредности. Требало би да повећа стабилност и тачност предвиђања.

```
{r}
rf_model <- randomForest(Exam_Score ~ Hours_Studied + Attendance + Previous_Scores
+
                      Sleep_Hours + Tutoring_Sessions + Parental_Involvement +
                      Access_to_Resources + Motivation_Level + Family_Income +
                      Teacher_Quality, data = train_data, ntree = 100)

rf_model
```

```
Call:
 randomForest(formula = Exam_Score ~ Hours_Studied + Attendance +
Previous_Scores + Sleep_Hours + Tutoring_Sessions + Parental_Involvement +
Access_to_Resources + Motivation_Level + Family_Income +      Teacher_Quality,
data = train_data, ntree = 100)
               Type of random forest: regression
                     Number of trees: 100
No. of variables tried at each split: 3

          Mean of squared residuals: 6.395021
                    % Var explained: 59.59
```

Овде смо креирали *Random Forest* модел са 100 стабала и предвиђамо резултат испита на основу броја сати учења и присуства. Модел објашњава 59.59% варијације у подацима, можемо да приметимо побољшање у односу на претходни модел стабла одлучивања.

Сада ћемо видети да ли су се и метрике тачности побољшале.

```r
{r}
rf_predictions <- predict(rf_model, newdata = test_data)

# MAE (Mean Absolute Error)
mae <- mean(abs(rf_predictions - test_data$Exam_Score))
# RMSE (Root Mean Squared Error)
rmse <- sqrt(mean((rf_predictions - test_data$Exam_Score)^2))
# MSE (Mean Squared Error)
mse <- mean((rf_predictions - test_data$Exam_Score)^2)
# R² (R-squared)
ss_total <- sum((test_data$Exam_Score - mean(test_data$Exam_Score))^2)
ss_residual <- sum((test_data$Exam_Score - rf_predictions)^2)
r2 <- 1 - (ss_residual / ss_total)

cat("MAE:", mae, "\n")
cat("RMSE:", rmse, "\n")
cat("MSE:", mse, "\n")
cat("R²:", r2, "\n")
```

```
MAE: 1.023761
RMSE: 1.73702
MSE: 3.017239
R²: 0.7686298
```

Видимо да су и метрике доста боље у односу на претходне моделе. *MAE*, *RMSE* и *MSE* су се смањиле, па су и просечне грешке у предвиђањима мање. R^2 је повећан, што значи да модел објашњава скоро 77% варијансе у резултатима испита.

7 Закључак

Циљ овог рада био је да се анализом података утврди који фактори највише утичу на резултате студената на испитима. Након свих извршених анализа, дошли смо до закључка да су сати учења и присуство на часовима кључни предиктори. Кроз примену различитих модела, најбољи резултати су добијени помоћу Random Forest модела, који је објаснио 76.86% варијације у резултатима испита. Кобминација труда и рада је добитна комбинација за добре резултате на факултету.

8 Литература

Увод у науку о подацима – вежбе и предавања

https://www.r-bloggers.com/2021/04/random-forest-in-r/

https://www.kaggle.com/code/mushei/eda-on-student-performance

https://www.r-bloggers.com/2021/04/decision-trees-in-r/

https://r-graph-gallery.com/