StudentAnalysis

2023-04-28

First, we will load all necessary libraries.

```
library("readxl")
library("ggplot2")
library("dplyr")
```

```
## Warning: package 'dplyr' was built under R version 4.2.3
```

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

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

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

Inputting The Data Set

To load in the file.

```
dfStudentDat <- read.csv(file = "C:/Users/blueb/OneDrive/Desktop/StudentGradeAnalysis/student_data.csv")
head(dfStudentDat, n =  5)
```

```
##   school sex age address famsize Pstatus Medu Fedu    Mjob
Fjob reason
## 1     GP   F  18       U     GT3       A    4    4 at_home
teacher course
## 2     GP   F  17       U     GT3       T    1    1 at_home
other course
## 3     GP   F  15       U     LE3       T    1    1 at_home
other  other
## 4     GP   F  15       U     GT3       T    4    2  health
services   home
## 5     GP   F  16       U     GT3       T    3    3   other
other   home
##   guardian traveltime studytime failures schoolsup famsup
paid activities
## 1   mother          2         2        0       yes     no
no         no
## 2   father          1         2        0        no    yes
no         no
## 3   mother          1         2        3       yes     no
yes          no
## 4   mother          1         3        0        no    yes
yes         yes
## 5   father          1         2        0        no    yes
yes          no
##   nursery higher internet romantic famrel freetime goout D
alc Walc health
## 1     yes    yes       no       no      4        3     4
1    1      3
## 2      no    yes      yes       no      5        3     3
1    1      3
## 3     yes    yes      yes       no      4        3     2
2    3      3
## 4     yes    yes      yes      yes      3        2     2
1    1      5
## 5     yes    yes       no       no      4        3     2
1    2      5
##   absences G1 G2 G3
## 1        6  5  6  6
## 2        4  5  5  6
## 3       10  7  8 10
## 4        2 15 14 15
## 5        4  6 10 10
```

Confirming that a datframe is achieved.

```
class(dfStudentDat)
```

```
## [1] "data.frame"
```

Column information.

```
# Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets:
1 school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2 sex - student's sex (binary: 'F' - female or 'M' - male)
3 age - student's age (numeric: from 15 to 22)
4 address - student's home address type (binary: 'U' - urban or 'R' - rural)
5 famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6 Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7 Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)
8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)
9 Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10 Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
11 reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12 guardian - student's guardian (nominal: 'mother', 'father' or 'other')
13 traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15 failures - number of past class failures (numeric: n if 1<=n<3, else 4)
16 schoolsup - extra educational support (binary: yes or no)
17 famsup - family educational support (binary: yes or no)
18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19 activities - extra-curricular activities (binary: yes or no)
```

```
20 nursery - attended nursery school (binary: yes or no)
21 higher - wants to take higher education (binary: yes or no)
22 internet - Internet access at home (binary: yes or no)
23 romantic - with a romantic relationship (binary: yes or no)
24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29 health - current health status (numeric: from 1 - very bad to 5 - very good)
30 absences - number of school absences (numeric: from 0 to 93)

# these grades are related with the course subject, Math or Portuguese:
31 G1 - first period grade (numeric: from 0 to 20)
31 G2 - second period grade (numeric: from 0 to 20)
32 G3 - final grade (numeric: from 0 to 20, output target)
```

For this study, we will compare variables with the final scores, regardless of which school the student attended.

```
#Drop the first column, the school attended.
df_Data_1 <- dfStudentDat[-c(1)]
head(df_Data_1, n = 5)
```

```
##   sex age address famsize Pstatus Medu Fedu     Mjob     Fj
ob reason guardian
## 1   F  18       U     GT3       A    4    4 at_home  teach
er course   mother
## 2   F  17       U     GT3       T    1    1 at_home    oth
er course   father
## 3   F  15       U     LE3       T    1    1 at_home    oth
er  other   mother
## 4   F  15       U     GT3       T    4    2  health servic
es   home   mother
## 5   F  16       U     GT3       T    3    3   other    oth
er   home   father
##   traveltime studytime failures schoolsup famsup paid acti
vities nursery higher
## 1          2         2        0       yes     no   no
no      yes    yes
## 2          1         2        0        no    yes   no
no       no    yes
## 3          1         2        3       yes     no  yes
no      yes    yes
## 4          1         3        0        no    yes  yes
yes      yes    yes
## 5          1         2        0        no    yes  yes
no      yes    yes
##   internet romantic famrel freetime goout Dalc Walc health
absences G1 G2 G3
## 1       no       no      4        3     4    1    1      3
6  5  6  6
## 2      yes       no      5        3     3    1    1      3
4  5  5  6
## 3      yes       no      4        3     2    2    3      3
10  7  8 10
## 4      yes      yes      3        2     2    1    1      5
2 15 14 15
## 5       no       no      4        3     2    1    2      5
4  6 10 10
```

```
#number of rows -> number of students
nrow(df_Data_1)
```

```
## [1] 395
```

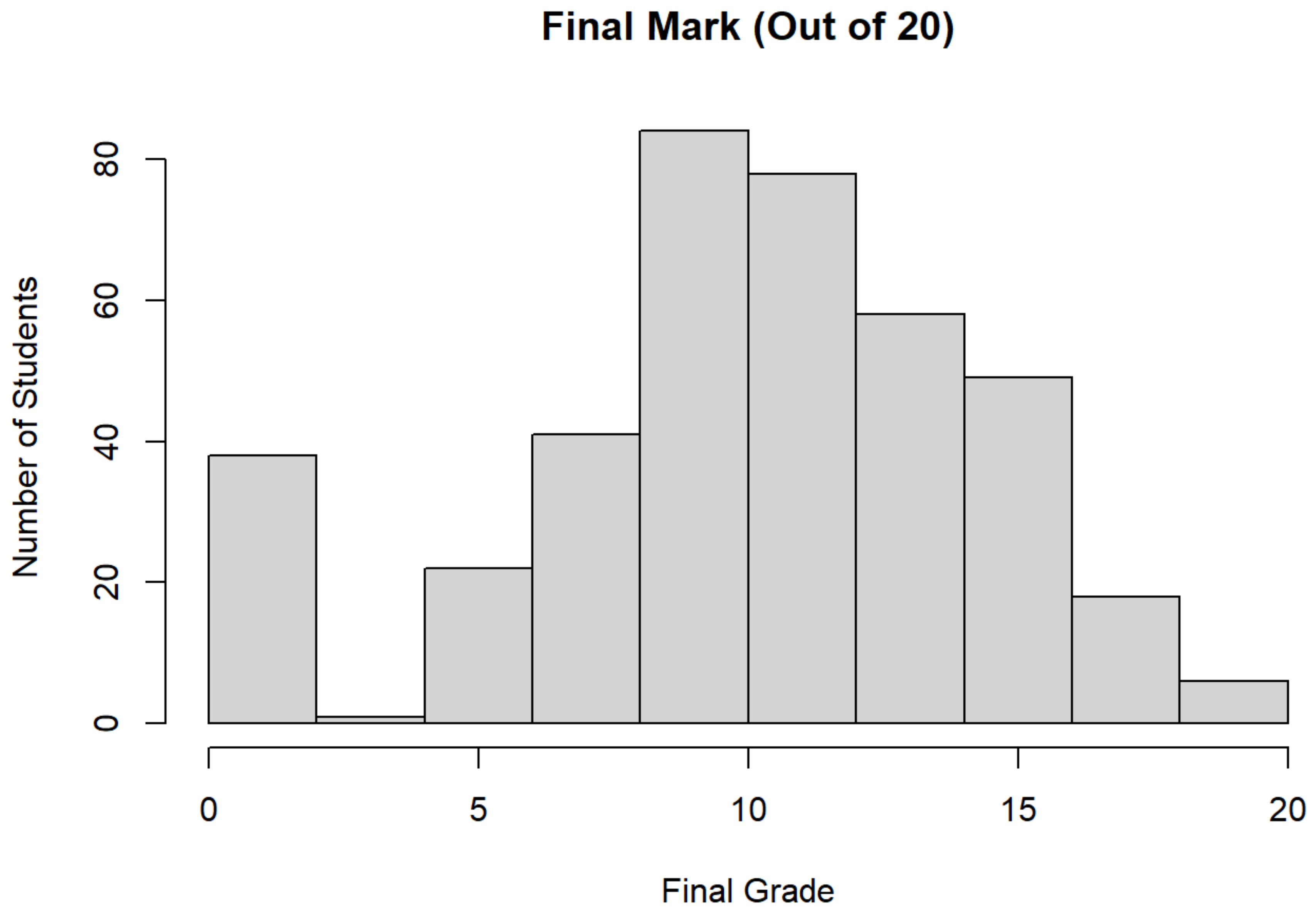
96% confidence interval estimation of the mean final grade (out of 20)

Checking conditions for t-test: - Sample is random - Population standard deviation is unknown - sample size(n = 395) > 30

```
#FinalGrade will only have column of final grade
FinalGrade <- df_Data_1$G3
#t. test
t.test(FinalGrade, conf.level = 0.96)
```

```
##
##  One Sample t-test
##
## data:  FinalGrade
## t = 45.182, df = 394, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 96 percent confidence interval:
##   9.940193 10.890187
## sample estimates:
## mean of x
##  10.41519
```

```
hist(FinalGrade, ylab = "Number of Students", xlab = "Final G
rade", main ="Final Mark (Out of 20)")
```



We are 96% confident that the mean final grade in this data set is between 9.94 and 10.89.

Bivariate Analysis

Question: Are the final marks between male and female students the same?

Comparison between sex and final mark.

```
#obtaining the columns of sex and final mark(G3)
dfSexMark <- select(df_Data_1, sex, G3)
head(dfSexMark, n = 5)
```

```
##   sex G3
## 1   F  6
## 2   F  6
## 3   F 10
## 4   F 15
## 5   F 10
```

F test done in 95% confidence interval (a = 0.05).

$\boldsymbol{\sigma}1$ is the standard deviation of the final mark of male students. $\boldsymbol{\sigma}2$ is the standard deviation of the final mark of female students.

H0: $\boldsymbol{\sigma}1 = \boldsymbol{\sigma}2$ H1: $\boldsymbol{\sigma}1 \neq \boldsymbol{\sigma}2$

```
df_maleGrade <- dfSexMark[dfSexMark$sex == "M",]
# df_maleGrade

df_femaleGrade <- dfSexMark[dfSexMark$sex == "F",]
# df_femaleGrade

#F test to check type of t-test.
var.test(df_femaleGrade$G3, df_maleGrade$G3)
```

```
## 
##  F test to compare two variances
## 
## data:  df_femaleGrade$G3 and df_maleGrade$G3
## F = 1.0573, num df = 207, denom df = 186, p-value = 0.6989
## alternative hypothesis: true ratio of variances is not equ
al to 1
## 95 percent confidence interval:
##  0.7975698 1.3986777
## sample estimates:
## ratio of variances 
##           1.057321
```

p-value = 0.6989 a = 0.05

since p-value > a, we fail to reject the null hypothesis. Thus, a pooled t-test is to be used.

H0: The final mark of male students is the same as female students. H1: The final mark of female students is not the same as male students

Pooled T test in 95% confidence interval (a = 0.05).

```
#pooled t-test, var.equal set to 'TRUE'
t.test(df_femaleGrade$G3,df_maleGrade$G3, alternative = "two.
sided" ,var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  df_femaleGrade$G3 and df_maleGrade$G3
## t = -2.062, df = 393, p-value = 0.03987
## alternative hypothesis: true difference in means is not eq
ual to 0
## 95 percent confidence interval:
##  -1.85205632 -0.04412838
## sample estimates:
## mean of x mean of y
##  9.966346 10.914439
```

According to the pooled t-test, the p-value is 0.03987, which is smaller than the alpha value of 0.05. Since p-value < a, we reject the null hypothesis.

Therefore, we conclude that the final grade between male students and female students is not the same and a difference exists.

Question: Do students who live in urban areas score higher than students who live in rural areas?

```
df_rural_FinalGrade <- select(df_Data_1, address, G3)
head(df_rural_FinalGrade, n = 5)
```

```
##   address G3
## 1       U  6
## 2       U  6
## 3       U 10
## 4       U 15
## 5       U 10
```

F test done in 95% confidence interval (a = 0.05).

$\boldsymbol{\sigma}$1 is the standard deviation of the final mark of students with urban address. $\boldsymbol{\sigma}$2 is the standard deviation of the final mark of studetns with rural address.

H0: $\boldsymbol{\sigma}1 = \boldsymbol{\sigma}2$ H1: $\boldsymbol{\sigma}1 \neq \boldsymbol{\sigma}2$

```
df_urbanGrade <- df_rural_FinalGrade[df_rural_FinalGrade$addr
ess == "U",]


df_ruralGrade <- df_rural_FinalGrade[df_rural_FinalGrade$addr
ess == "R",]

# #F test to check type of t-test.
var.test(df_urbanGrade$G3, df_ruralGrade$G3)
```

```
## 
##  F test to compare two variances
## 
## data:  df_urbanGrade$G3 and df_ruralGrade$G3
## F = 1.003, num df = 306, denom df = 87, p-value = 0.9886
## alternative hypothesis: true ratio of variances is not equ
al to 1
## 95 percent confidence interval:
##  0.7035538 1.3846151
## sample estimates:
## ratio of variances 
##           1.003043
```

p-value = 0.886 a = 0.05

since p-value > a, we fail to reject the null hypothesis. Thus, a pooled t-test is to be used.

```
results = lm(G3 ~ absences, data = df_Data_1)
results
```

```
##
## Call:
## lm(formula = G3 ~ absences, data = df_Data_1)
##
## Coefficients:
## (Intercept)     absences
##    10.30327      0.01961
```

```
summary(results)
```

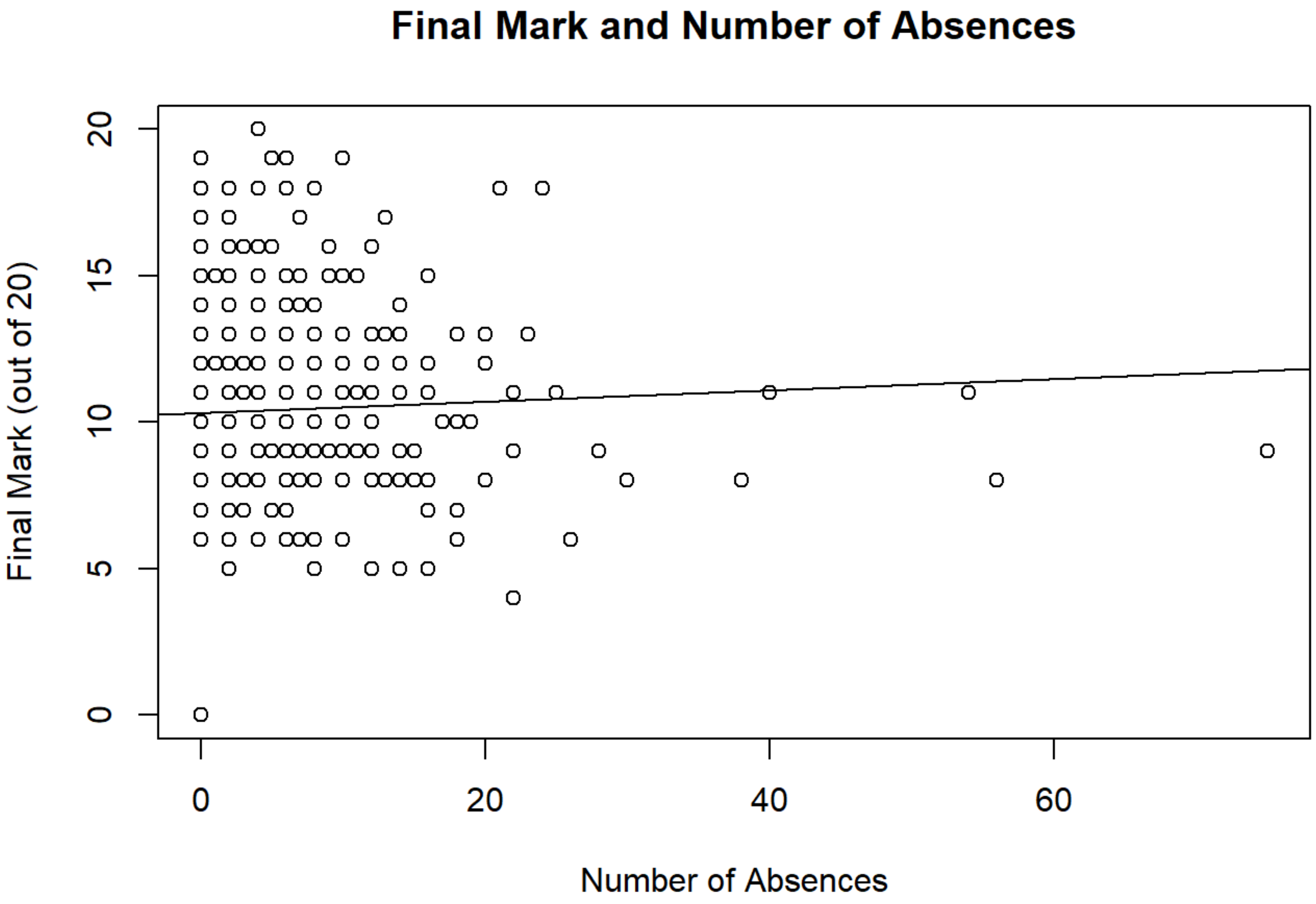
```
##
## Call:
## lm(formula = G3 ~ absences, data = df_Data_1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3033  -2.3033   0.5007   3.4811   9.6183
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.30327    0.28347  36.347   <2e-16 ***
## absences     0.01961    0.02886   0.679    0.497
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1
##
## Residual standard error: 4.585 on 393 degrees of freedom
## Multiple R-squared:  0.001173,   Adjusted R-squared:  -0.0
01369
## F-statistic: 0.4615 on 1 and 393 DF,  p-value: 0.4973
```

```
anova(results)
```

```
## Analysis of Variance Table
##
## Response: G3
##            Df Sum Sq Mean Sq F value Pr(>F)
## absences    1    9.7  9.6996  0.4615 0.4973
## Residuals 393 8260.2 21.0183
```

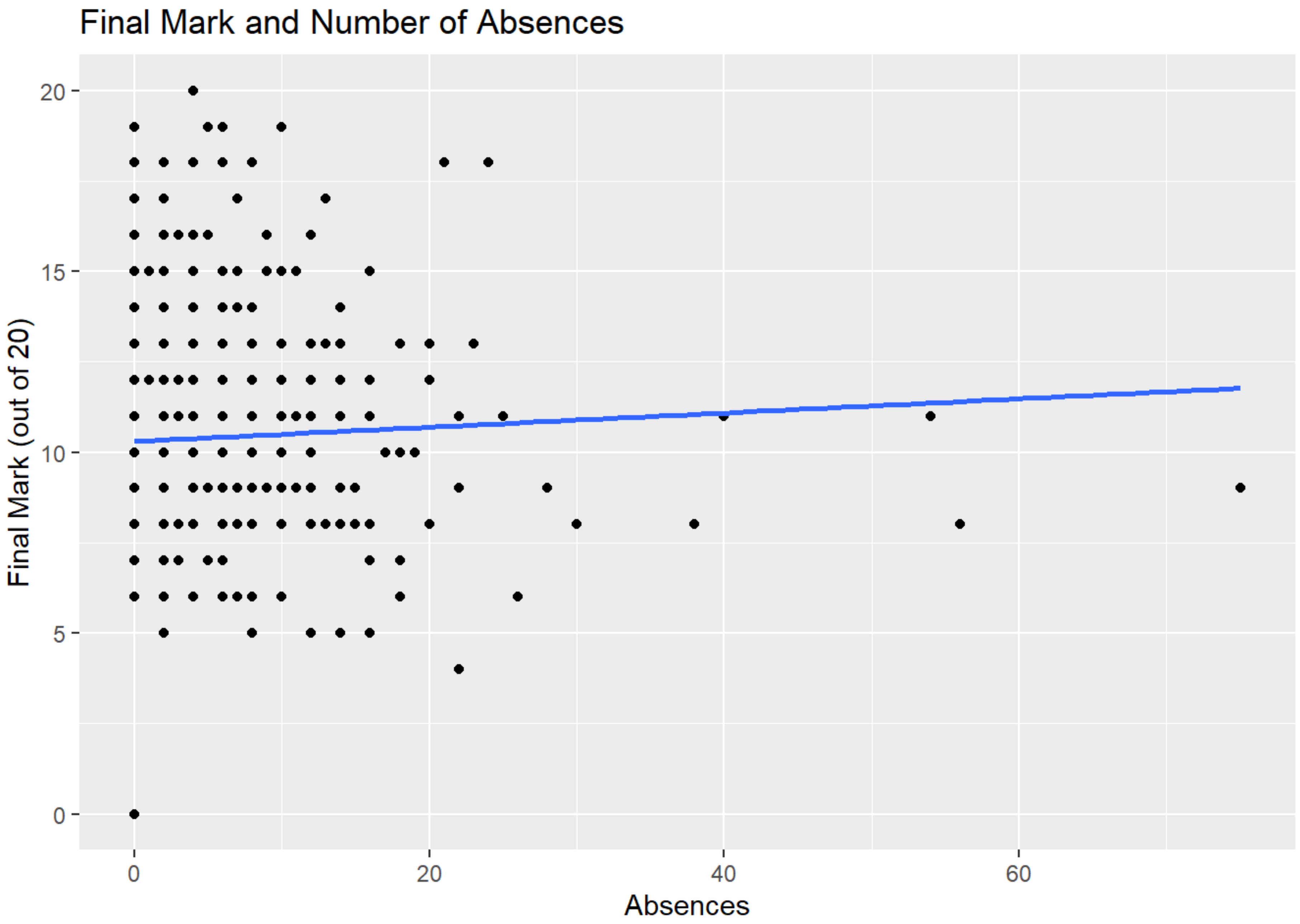
```r
plot(df_Data_1$absences, df_Data_1$G3,
     main = "Final Mark and Number of Absences",
     xlab = "Number of Absences",
     ylab = "Final Mark (out of 20)")

abline(lm(df_Data_1$G3 ~ df_Data_1$absences))
```



```r
ggplot(df_Data_1, aes(x=absences, y= G3)) +
    geom_point() +
    ggtitle("Final Mark and Number of Absences")+
    xlab("Absences") +
    ylab("Final Mark (out of 20)") +
    geom_smooth(method=lm, se=FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```


Final Mark and Number of Absences
Final Mark (out of 20)
20
15
10
5
0
0
20
40
60
Absences