STA504 HOMEWORK1

Jessica Choe(Jaeseong Choe)

```
#0.preparing for analysis and Exploring the data

library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.6.2
```

```
## -- Attaching packages ----------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## Warning: package 'tidyr' was built under R version 3.6.2
```

```
## Warning: package 'readr' was built under R version 3.6.2
```

```
## Warning: package 'purrr' was built under R version 3.6.2
```

```
## Warning: package 'forcats' was built under R version 3.6.2
```

```
## -- Conflicts -------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
math <- read_csv("C:/Users/user/Desktop/2020spring/502 data visualization/hw/hw1/student-mat.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   age = col_double(),
##   Medu = col_double(),
##   Fedu = col_double(),
##   traveltime = col_double(),
##   studytime = col_double(),
##   failures = col_double(),
##   famrel = col_double(),
##   freetime = col_double(),
##   goout = col_double(),
##   Dalc = col_double(),
##   Walc = col_double(),
##   health = col_double(),
##   absences = col_double(),
##   G1 = col_double(),
##   G2 = col_double(),
##   G3 = col_double()
## )
```

```
## See spec(...) for full column specifications.
```

```r
is.data.frame(math)
```

```
## [1] TRUE
```

```r
ls(math)
```

```
##  [1] "absences"   "activities" "address"    "age"        "Dalc"
##  [6] "failures"   "famrel"     "famsize"    "famsup"     "Fedu"
## [11] "Fjob"       "freetime"   "G1"         "G2"         "G3"
## [16] "goout"      "guardian"   "health"     "higher"     "internet"
## [21] "Medu"       "Mjob"       "nursery"    "paid"       "Pstatus"
## [26] "reason"     "romantic"   "school"     "schoolsup"  "sex"
## [31] "studytime"  "traveltime" "Walc"
```

```r
head(math)
```

```
## # A tibble: 6 x 33
##   school sex     age address famsize Pstatus  Medu  Fedu Mjob  Fjob  reason
##   <chr>  <chr> <dbl> <chr>   <chr>   <chr>   <dbl> <dbl> <chr> <chr> <chr>
## 1 GP     F        18 U       GT3     A           4     4 at_h~ teac~ course
## 2 GP     F        17 U       GT3     T           1     1 at_h~ other course
## 3 GP     F        15 U       LE3     T           1     1 at_h~ other other
## 4 GP     F        15 U       GT3     T           4     2 heal~ serv~ home
## 5 GP     F        16 U       GT3     T           3     3 other other home
## 6 GP     M        16 U       LE3     T           4     3 serv~ other reput~
## # ... with 22 more variables: guardian <chr>, traveltime <dbl>,
## #   studytime <dbl>, failures <dbl>, schoolsup <chr>, famsup <chr>,
## #   paid <chr>, activities <chr>, nursery <chr>, higher <chr>,
## #   internet <chr>, romantic <chr>, famrel <dbl>, freetime <dbl>,
## #   goout <dbl>, Dalc <dbl>, Walc <dbl>, health <dbl>, absences <dbl>,
## #   G1 <dbl>, G2 <dbl>, G3 <dbl>
```

```r
#1.code:
math$school_labeled <- factor(math$school,
                    labels=c("Gabriel Pereira",
                             "Mousinho da Silveira"))
data.frame(table(math$school_labeled)) # One way to view the table- in a matrix format.
```

```
##                   Var1 Freq
## 1      Gabriel Pereira  349
## 2 Mousinho da Silveira   46
```

#1.answer: Gabriel Pereira(GP):349, Mousinho da Silveira(MS):46

```r
#2.code:
#(1)step1: calculate the count
reputation_count <- math %>% group_by(reason)%>%
                    summarise(total = n())
reputation_count
```

```
## # A tibble: 4 x 2
##   reason     total
##   <chr>      <int>
## 1 course       145
## 2 home         109
## 3 other         36
## 4 reputation   105
```

```
#(2)step2: calcute the percent for reputation.
reputation_percent <- reputation_count%>% mutate(percent = total / sum(total))
reputation_percent
```

```
## # A tibble: 4 x 3
##   reason     total percent
##   <chr>      <int>   <dbl>
## 1 course       145  0.367
## 2 home         109  0.276
## 3 other         36  0.0911
## 4 reputation   105  0.266
```

#2.answer: 26.58%(=0.26) of students chose their school based on reputation.

```
#3.code
math$sex_labeled <- factor(math$sex,
                   labels=c("Female","Male"))
Grade1_over10 <- math %>% dplyr::filter(G1>10)

F_Grade1_over10 <- Grade1_over10 %>%
                  group_by(sex_labeled) %>%
                  summarise(total = n()) %>%
                  mutate(percent = total / sum(total))
F_Grade1_over10
```

```
## # A tibble: 2 x 3
##   sex_labeled total percent
##   <fct>       <int>   <dbl>
## 1 Female         97   0.480
## 2 Male          105   0.520
```

#3.answer:The percent of female among students whose first period grade is greater than 10 is 48.0198%

```
#4. code:
#(1)change the class(type) from numeric to factor
math$studyT_labeled <- factor(math$studytime,
                    labels=c("Less than2", "2 to 5hrs",
                             "5 to 10hrs","More than10"))
#(2)set order for the categorical variables.
math$studyT_labeled <- ordered(math$studyT_labeled,
                    labels=c("Less than2", "2 to 5hrs",
                             "5 to 10hrs","More than10"))
head(math$studyT_labeled)
```

```
## [1] 2 to 5hrs  2 to 5hrs  2 to 5hrs  5 to 10hrs 2 to 5hrs  2 to 5hrs
## Levels: Less than2 < 2 to 5hrs < 5 to 10hrs < More than10
```

```
#(3) filter by studytime(=5 to 10 hours) and activities(=yes).
activityY_Study5to10 <-math%>%
          dplyr::filter(studyT_labeled=="5 to 10hrs" &
          activities=="yes")
#(4) summarize with mean and standard deviation for the 2nd grade
summary_2ndG <- activityY_Study5to10  %>%
              summarize(mean(G2),sd(G2))
summary_2ndG
```

```
## # A tibble: 1 x 2
##   `mean(G2)` `sd(G2)`
##        <dbl>    <dbl>
## 1       11.2     3.58
```

#4. answer:mean(Second grade):11.25 sd(Second grade):3.58

```
#5. code:
#(1) create new traveltime variable(=travelT_labeled) to change its type and labels.
#(study time is already labeled in problem 4)
class(math$traveltime)
```

```
## [1] "numeric"
```

```
math$travelT_labeled <- factor(math$traveltime,
                    labels=c("<15min","15 to 30m",
                    "30m to 1 hr",">1hr"))

math$travelT_labeled <- ordered(math$travelT_labeled,
                       labels=c("<15min","15 to 30m","30m to 1 hr",">1hr"))
class(math$travelT_labeled)
```

```
## [1] "ordered" "factor"
```

```
head(math$travelT_labeled)
```

```
## [1] 15 to 30m <15min    <15min    <15min    <15min    <15min
## Levels: <15min < 15 to 30m < 30m to 1 hr < >1hr
```

```
#the type of traveltime_labeled is factor (ordered)

#(2)find the highest average final grade:
#1.group the data by traveltime and study time.
#2.find mean for each group.
#3.sort it from highest final grade(=G3)
table(math$travelT_labeled,math$studyT_labeled)
```

```
##
##               Less than2 2 to 5hrs 5 to 10hrs More than10
##   <15min              60       133         47          17
##   15 to 30m           31        51         16           9
##   30m to 1 hr         11        11          1           0
##   >1hr                 3         3          1           1
```

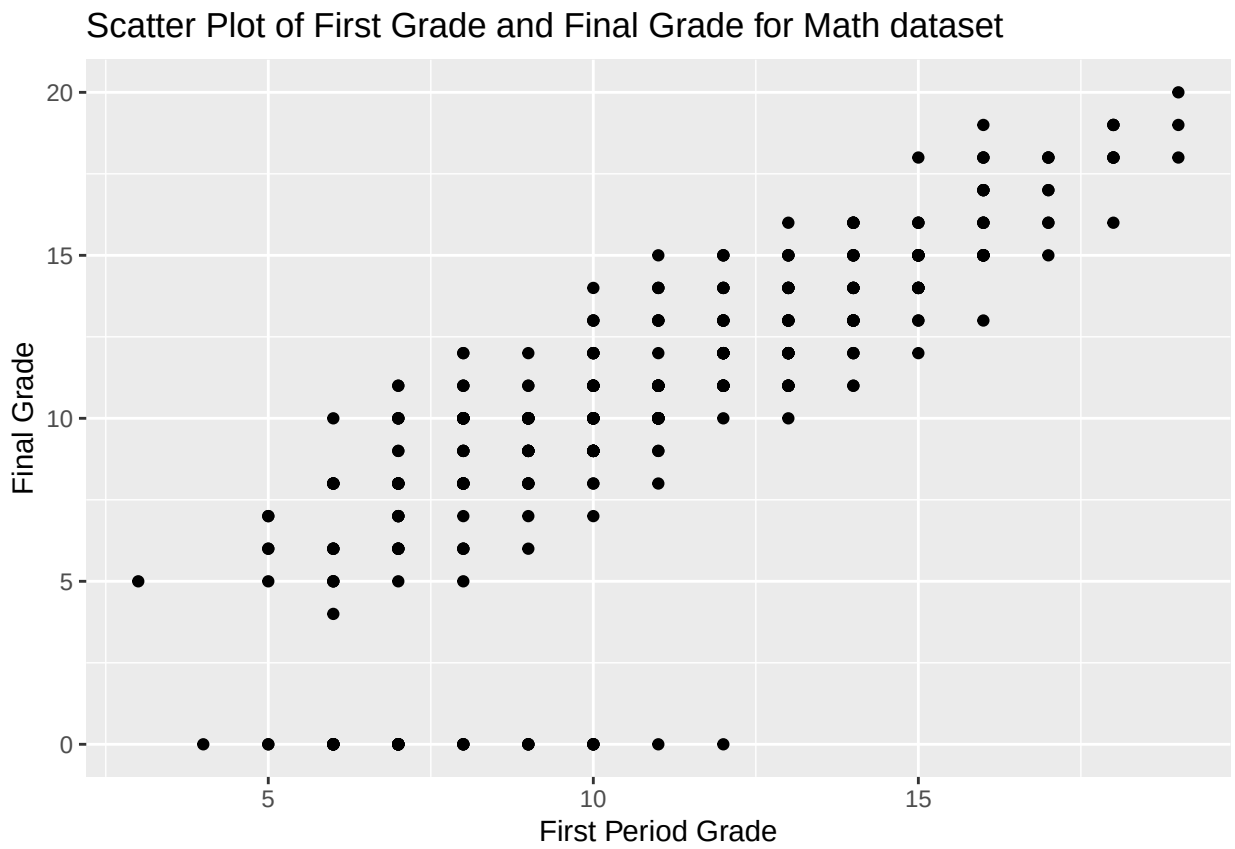
```
math2 <- math %>% group_by(travelT_labeled, studyT_labeled)
summary_desc <- math2 %>%
          summarise(mean.final=mean(G3),n=n())%>%
           arrange(desc(mean.final))
summary_desc
```

```
## # A tibble: 15 x 4
## # Groups:   travelT_labeled [4]
##    travelT_labeled studyT_labeled mean.final     n
##    <ord>           <ord>               <dbl> <int>
##  1 >1hr            More than10          13        1
##  2 30m to 1 hr     5 to 10hrs           12        1
##  3 <15min          More than10          11.7     17
##  4 15 to 30m       5 to 10hrs           11.5     16
##  5 <15min          5 to 10hrs           11.4     47
##  6 <15min          Less than2           10.9     60
##  7 <15min          2 to 5hrs            10.4    133
##  8 >1hr            Less than2           10.3      3
##  9 15 to 30m       More than10          10.2      9
## 10 30m to 1 hr     2 to 5hrs            10.2     11
## 11 >1hr            5 to 10hrs           10        1
## 12 15 to 30m       2 to 5hrs             9.82    51
## 13 15 to 30m       Less than2            9.13    31
## 14 30m to 1 hr     Less than2            8.09    11
## 15 >1hr            2 to 5hrs             5.33     3
```

#5. answer: study time more than 10 hours (4) and travel time more than 1 hour(4) has the highest average of final period grade of 13.

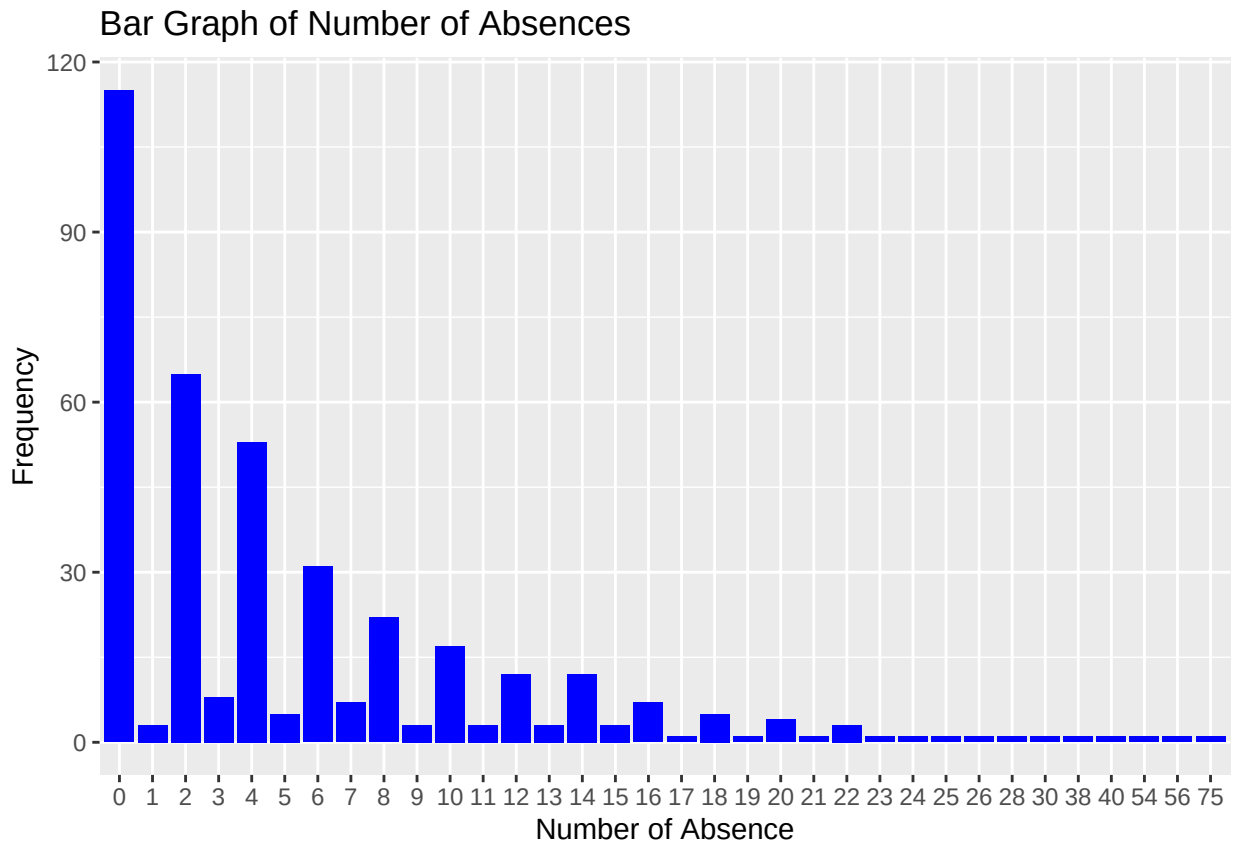
However, n=1(only 1 observation) in this case so we cannot say it is the best combination for the best grade as the observation is too small as only 1 person.

```
#6. code:
ggplot()+
  geom_point(aes(x=G1, y=G3),data=math)+
  labs(x="First Period Grade",y="Final Grade",
       title="Scatter Plot of First Grade and Final Grade for Math dataset")
```

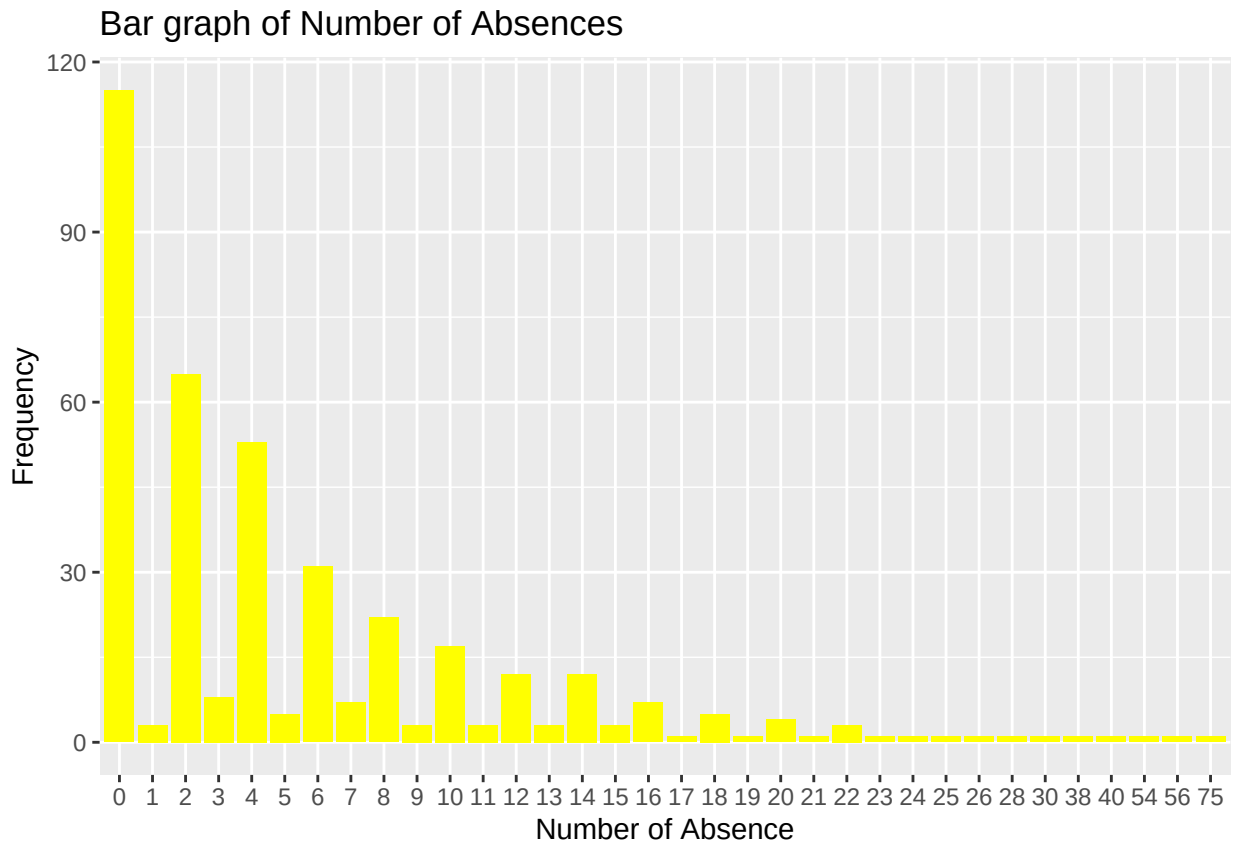


#6. answer: The relationship between first grade and final grade is pretty consistently linear and the scatter plot shows clear linear relationship. There are some outliers at final grade=0 and when I checked the data, I found that those students with final grade=0 had their scores below 10 for both the first and second grade, so they might have given up their studying after getting low grades at the earlier exams.

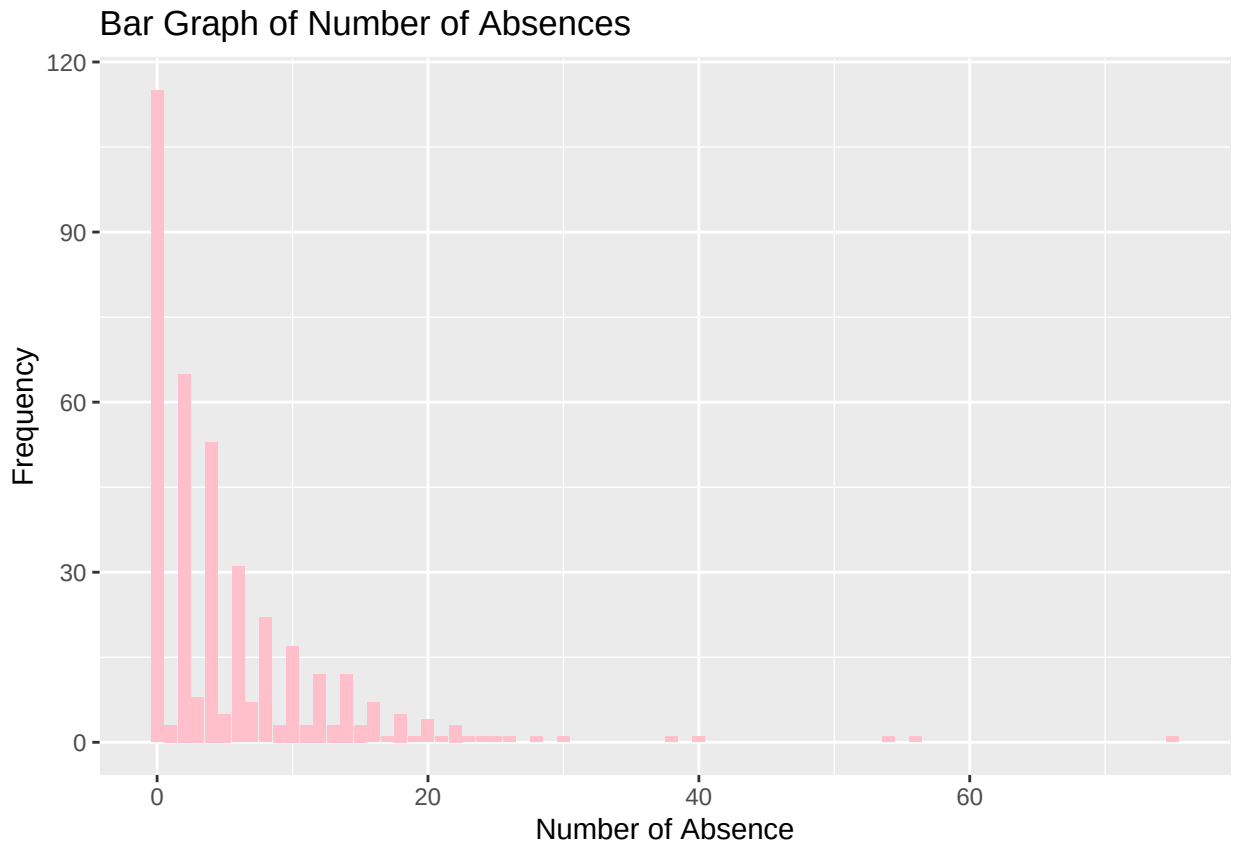
```
#7. code:
#5 ways to create a plot to show the distribution of school absences.
#Except for (5)geom_density showing distribution in y-axis, all the rest show the frequencies.
#(1)using geom_col
absences <- data.frame(table(math$absences))

ggplot() +
  geom_col(aes(x=Var1,y=Freq),fill="blue",data=absences)+
  labs(x="Number of Absence",y="Frequency",
       title="Bar Graph of Number of Absences")
```



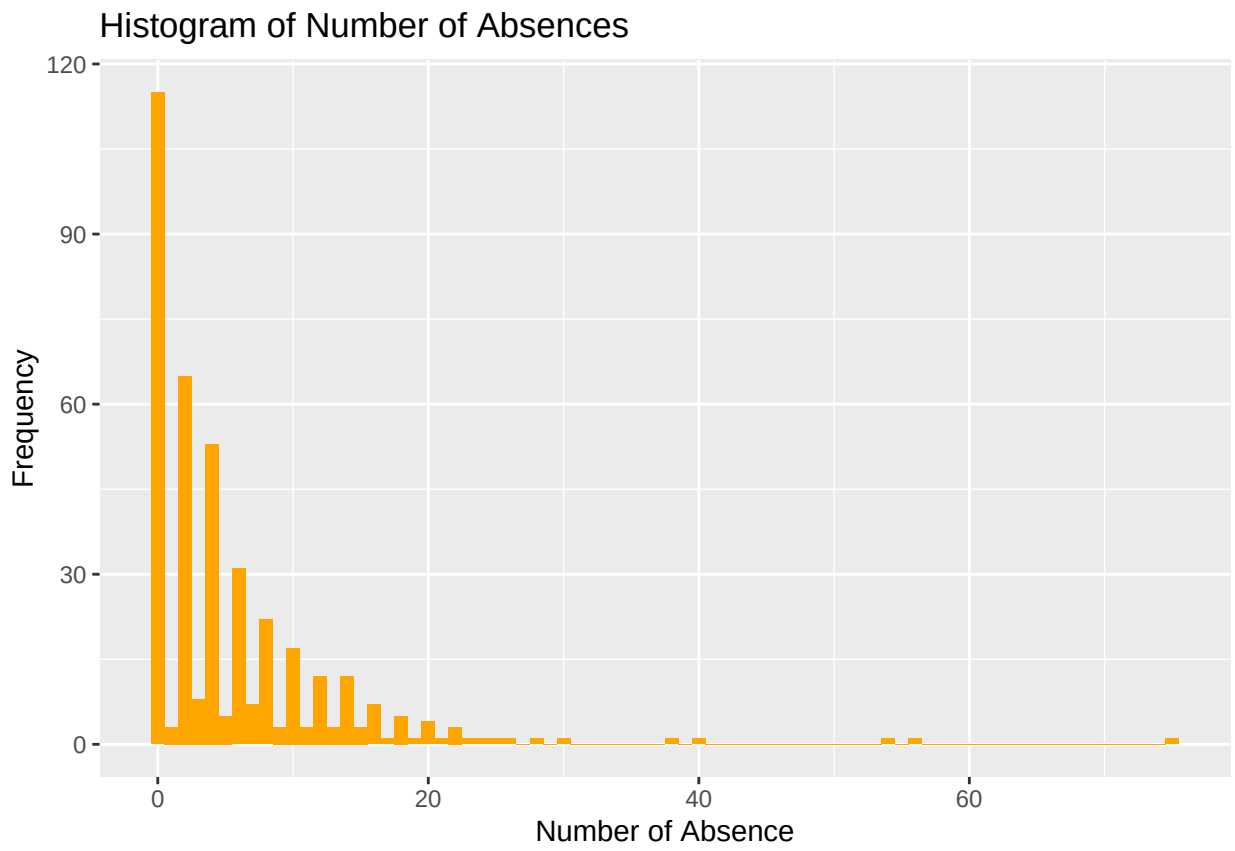

```
#(2)using geom_bar (data:absences)
ggplot() +
  geom_bar(aes(x=Var1,y=Freq),fill="yellow",data=absences,
           stat="identity") +
         labs(x="Number of Absence",y="Frequency",
       title="Bar graph of Number of Absences")
```



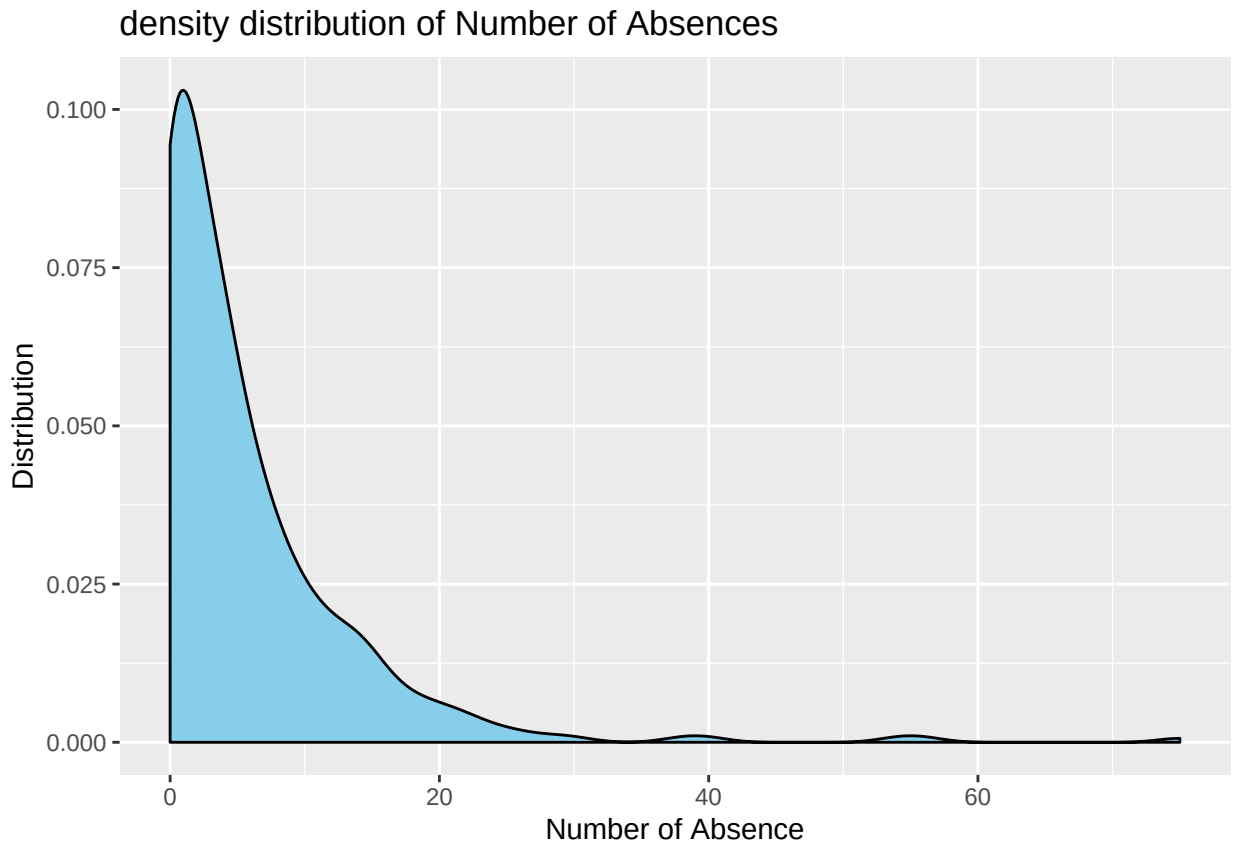

```
#(3)using geom_bar (data :math)
ggplot() +
  geom_bar(aes(x=absences),fill="pink",
           data=math)+
  labs(x="Number of Absence",y="Frequency",
       title="Bar Graph of Number of Absences")
```




```
#(4)using geom_histogram
ggplot() +
  geom_histogram(aes(x=absences),
                 fill="orange",
                 binwidth=1,data=math)+
  labs(x="Number of Absence",y="Frequency",
       title="Histogram of Number of Absences")
```

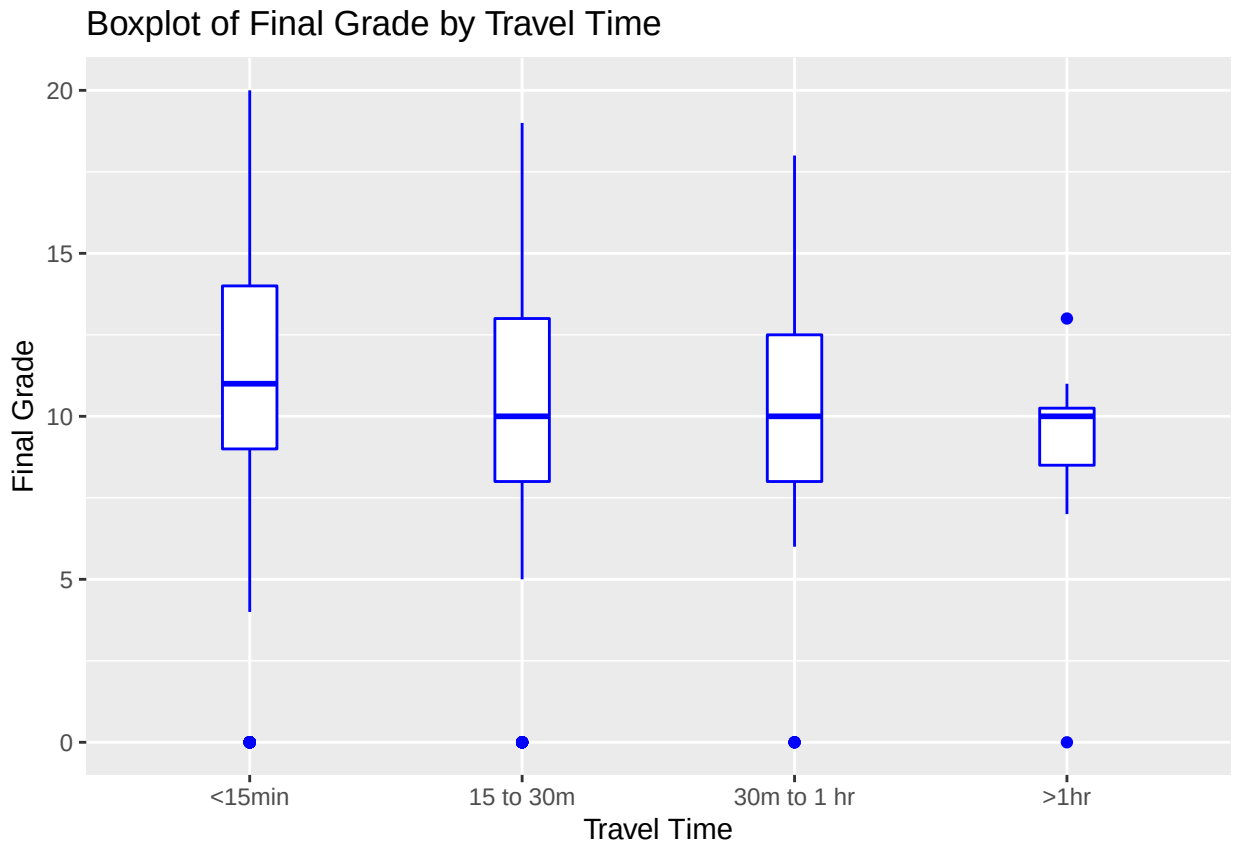



```
#(5)using geom_density
ggplot() +
  geom_density(aes(x=absences),fill="skyblue",data=math)+
  labs(x="Number of Absence",y="Distribution",
       title="density distribution of Number of Absences")
```



#7.answer: The distribution is very skewed to right (longer tail toward right.) and most of absences are gathered near zero and the shape is similiar to gamma distribution

```
#8. code:
#(1)traveltime_labeled was also created in problem5
# for better labeling and class change ( into "factor" variable).

math$traveltime_labeled <- factor(math$traveltime,
                                  labels=c("<15min","15 to 30m", "30m to 1 hr",">1hr"))
math$traveltime_labeled <- ordered(math$traveltime,
                    labels=c("<15min","15 to 30m",
                             "30m to 1 hr",">1hr"))

#(2)Create a boxplot to show the relationship between travel time and final grade.
ggplot()+
  geom_boxplot(aes(x=traveltime_labeled, y=G3),
  width=.2,fill="white", color="blue", data=math)+
  labs(x="Travel Time",
       y="Final Grade",
       title="Boxplot of Final Grade by Travel Time")
```



#8. answer: #There are outliers in final grade=0 for all groups. For those observations, maybe instead of travel time, other factors could have affect them and we can have a further investigation. (continued in problem 9)

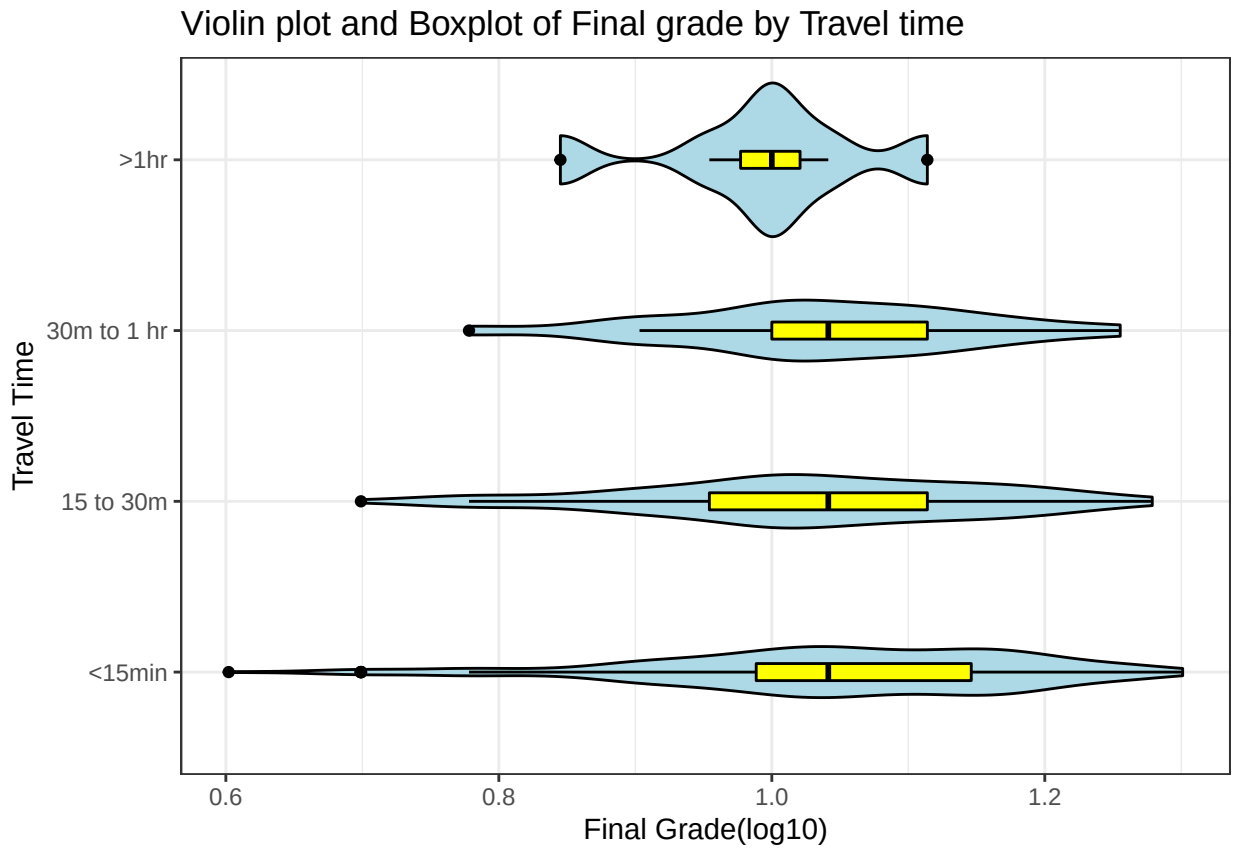
#The medians for final grade are all around 10 or a little above for all travel times and the first and third quartiles do not seem very different (except the longer than 1 hour.) Therefore, it does not look like there is much impact of travel time on final grade.

#Lastly, there is longer whiskers(The two lines extending from the boxes) for shorter travel time implying there is bigger range in shorter travel time while the longest travel time group has very short whiskers meaning shorter range. We can assume that there are not so many observations for longer travel time.

```
#9. code:
#(1)Violine plot(log10(G3))+Boxplot(log10(G3))+flipping
ggplot()+
  geom_violin(aes(x=traveltime_labeled, y=log10(G3)),
              color="black", fill="lightblue",
              data=math) +
  geom_boxplot(aes(x=traveltime_labeled, y=log10(G3)),
               width=.1, color="black", fill="yellow",
               data=math)+
                theme_bw()+
       labs(x="Travel Time",y="Final Grade(log10)",
      title="Violin plot and Boxplot of Final grade by Travel time")+
  coord_flip()
```

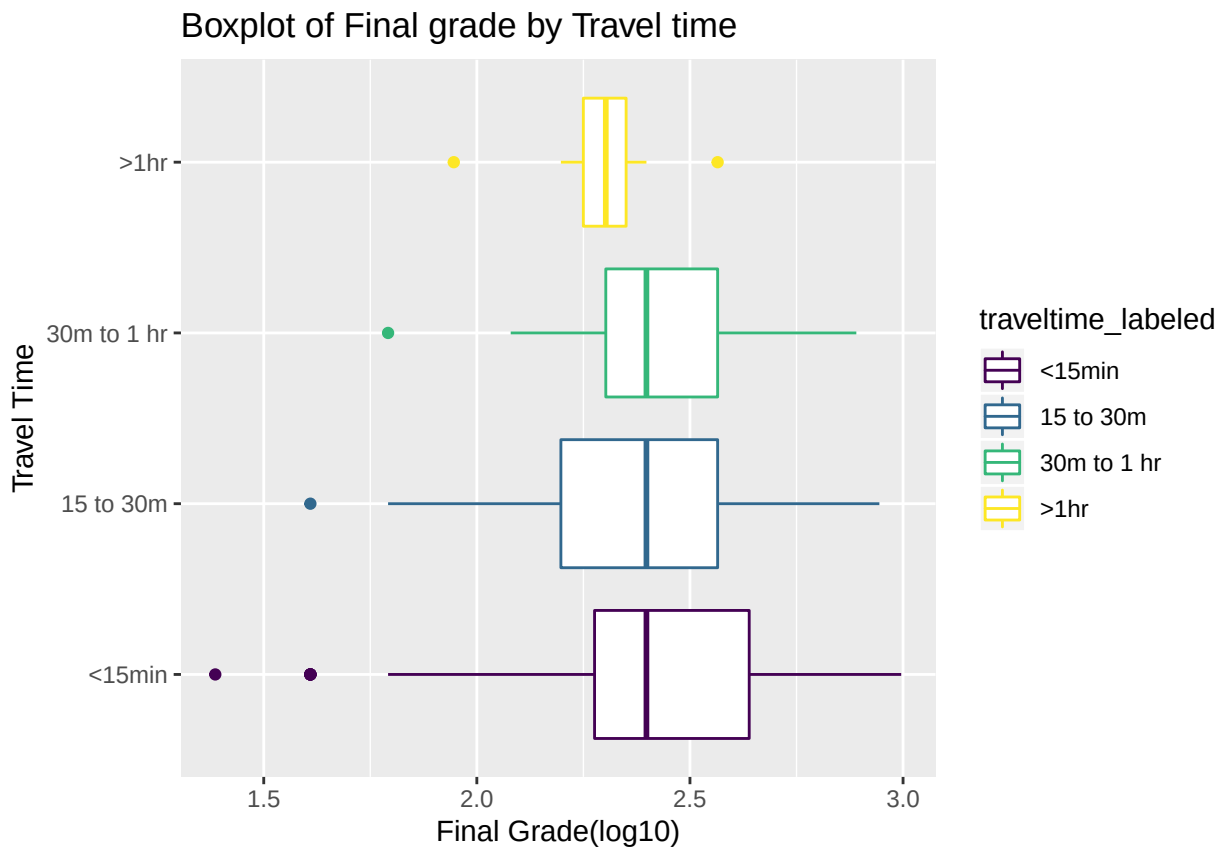
```
## Warning: Removed 38 rows containing non-finite values (stat_ydensity).
```

```
## Warning: Removed 38 rows containing non-finite values (stat_boxplot).
```




```
#(2) Box plot only (to see the boxplot more clearly after transformation with log(last grade))
ggplot()+
  geom_boxplot(aes(x=traveltime_labeled, y=log(G3),color=traveltime_labeled), data=math) +
  coord_flip() +
  labs(x="Travel Time",y="Final Grade(log10)",
       title="Boxplot of Final grade by Travel time")
```

```
## Warning: Removed 38 rows containing non-finite values (stat_boxplot).
```



#9. answer: # For each travel time, there were outliers at zero score in final grade(G3) in problem 8, and their final grades were relatively not very related/affected by the travel time in problem 8. Total 38 outliers at zero score were therefore removed by the transformation of final grade.(log(G3))

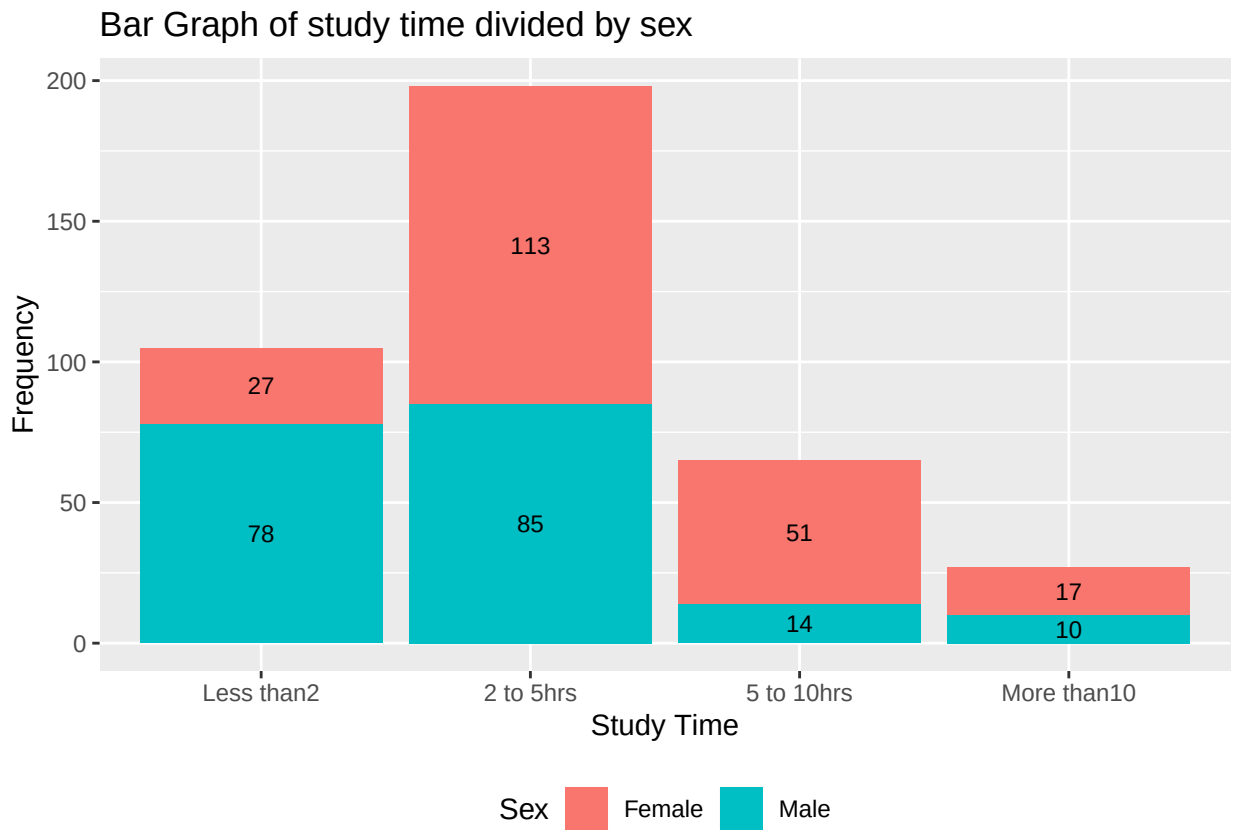
#In the violin plot (as well as boxplot), we can observe that the shorter the travel time is the longer the range of distribution(higher variance) we get.

10. my own topic:

(1) Create a geom bar graph to show the frequency of study hours whose group is break down into sex.
(2) Indicate frequencies(counts) for each group in the bar and put them on readable positions.
(3) Put title, x-axis, and y-axis. Also, place legend at the bottom.
(4) For any inputs used multiple times, only enter them one time put them by putting them on the top commend in order to make the code as simple as possible.

```
#10. code:
# aes(x=studyT_labled), data=math can appear multiple times, so put them in ggplot()
math$sex_labeled <- factor(math$sex,
                    labels=c("Female","Male"))
ggplot(aes(x=studyT_labeled),data=math)+
  geom_bar(aes(fill=sex_labeled))+
  labs(x="Study Time", y="Frequency",
      title="Bar Graph of study time divided by sex")+
  geom_text(aes(label=..count..,group=sex),
            stat="count",
            position=position_stack(vjust=0.5),size=3)+
```

```
guides(fill=guide_legend(title="Sex"))+
theme(legend.position= "bottom")
```



#Most frequent group is study time 2 to 5 hours group. Both male and female has highst frequency in 2 to 5 hours of study. Overll, the plot is skewed to the right and female students has more frequency in studying longer hours than male students.