STA404/504 HW2

Ethan Gutknecht

09/08/2022

Learning Objectives:

- Data manipulation and computation
- Scatter plot and aesthetic settings

Submission Requirement (Apply to all the HW)

Please upload the following files separately to the canvas website. Please **do not** upload a zip file. Files do not follow the requirements will receive points deduction.

(1) An **Rmarkdown** file that can reproduce your knitted file. Please be sure to properly document and organize your code so that the data manipulation and plot creation processes are clear. Please also mark the questions clearly and have necessary answers and discussions.

(2) A knitted word/pdf/html file that containing the code, the outputs (the plots) and appropriate labels, discussion if applicable. For those who did not install latex, you may want to knit as html file.

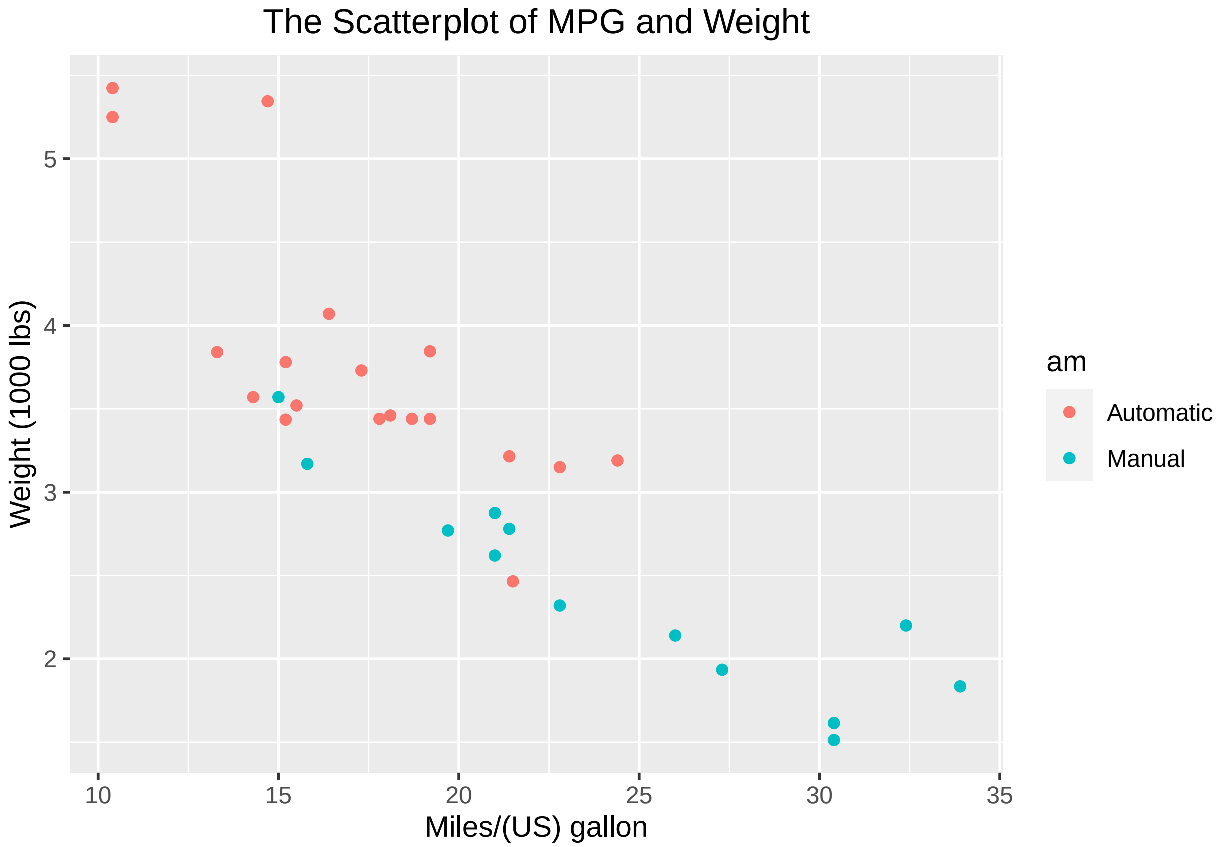
Question 1.

The first question is related to Question 1 in HW1. We will use the R built-in dataset mtcars to answer the following question:

(1) Now we have learned many tools to change the aesthetic setting of a scatterplot during class, revisit the plot you have created in Question 1 (5) of HW1. Edit and refine your plot by adding a third variables' information to it. Make sure the aesthetic settings, background color and theme, title (please center the title), labels, etc. are clear and appropriate.

```r
# Change the transmission to factors
newFactorLabels <- factor(mtcars$am, labels = c("Automatic","Manual"))
mtcars$am <- newFactorLabels

ggplot() +
  geom_point(aes(x=mpg, y=wt, color=am),data=mtcars) +
  labs(x="Miles/(US) gallon", y="Weight (1000 lbs)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  ggtitle("The Scatterplot of MPG and Weight")
```



Question 2.

Work on the dataset "student-por.csv" on canvas, which contains student's achievement in a Portugue language course in secondary education of two Portuguese schools (Cortez and Silva, 2008). The data attributes include student grades, demographic, social and school related features and it was collected by using school reports and questionnaires. This is one of the available datasets. You may refer to this website for more information about the attributes and the data: https://archive.ics.uci.edu/ml/datasets/student+performance. Answer the following questions:

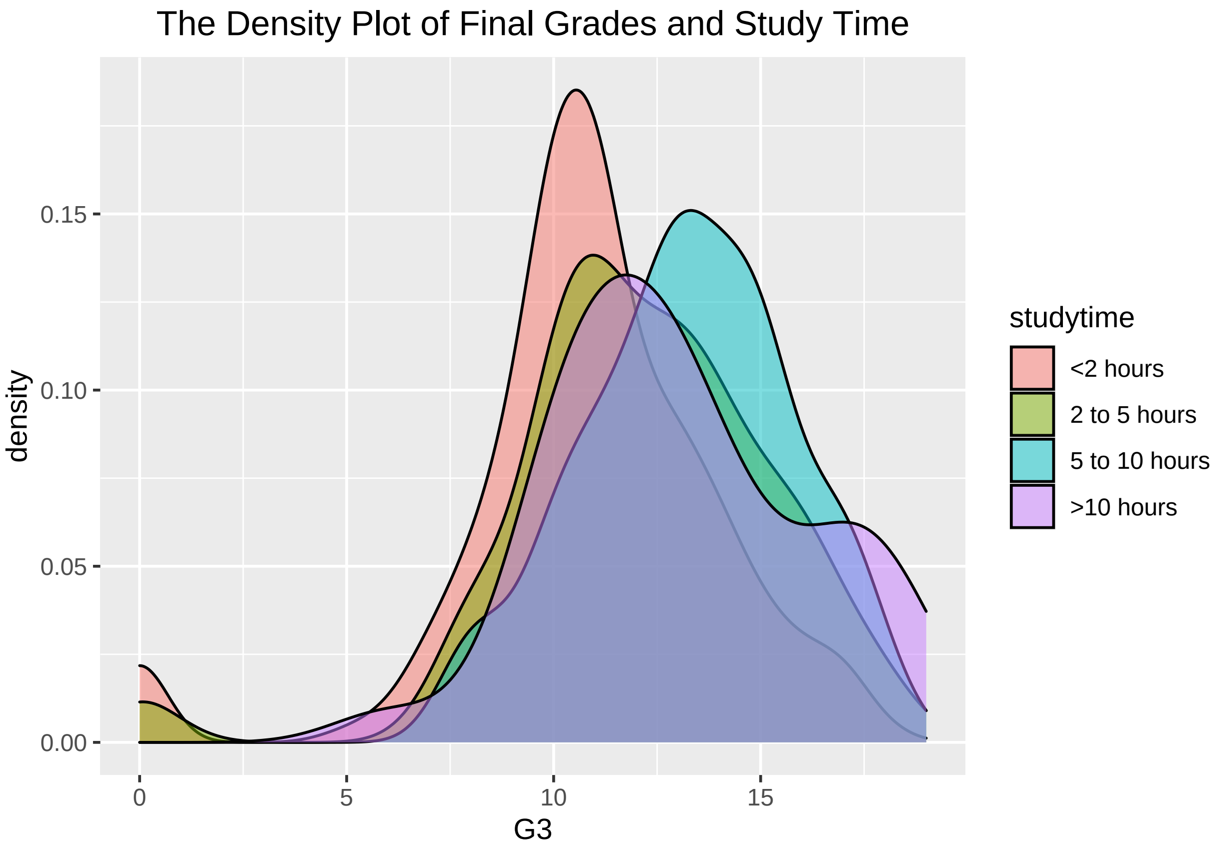
```
por <- read.table("student-por.csv",sep=";",header=TRUE)
```

(1) Students spend various time travel to school and their weekly study time are different. Based on your analysis, what kind of travel time and study time combination has the highest average final grade in the Portuguese language course?

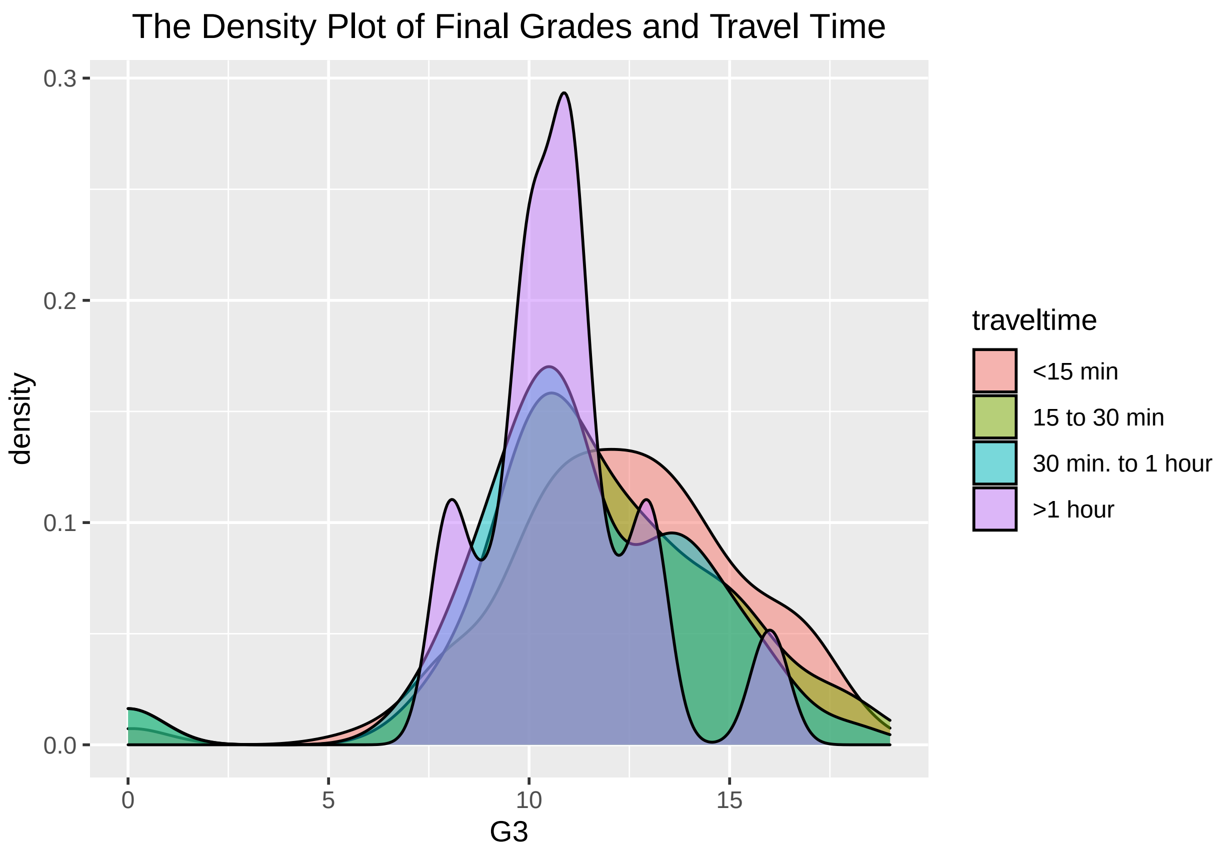
```
studyTimeFactorLabels <- factor(por$studytime, labels = c("<2 hours","2 to 5 hours", "5 to 10 hours", "
por$studytime <- studyTimeFactorLabels

travelTimeFactorLabels <- factor(por$traveltime, labels = c("<15 min","15 to 30 min", "30 min. to 1 hou
por$traveltime <- travelTimeFactorLabels

ggplot() +
  geom_density(aes(x=G3,fill=studytime), data=por,alpha=0.5) +
  theme(plot.title = element_text(hjust = 0.5)) +
  ggtitle("The Density Plot of Final Grades and Study Time")
```



```
ggplot() +
  geom_density(aes(x=G3,fill=traveltime), data=por,alpha=0.5) +
  theme(plot.title = element_text(hjust = 0.5)) +
  ggtitle("The Density Plot of Final Grades and Travel Time")
```



The travel time of <15 minutes and study time of >10 hours should give you the highest final grade.

(2) For students who falls into the travel time and study time combination group above, what is their median second period grade for Portuguese language course?

```
filteredDataQ2 <- filter(por, studytime == ">10 hours" & traveltime == "<15 min" )

medianG2Grade <- filteredDataQ2 %>% summarise(q2.median=median(G2))
medianG2Grade
```

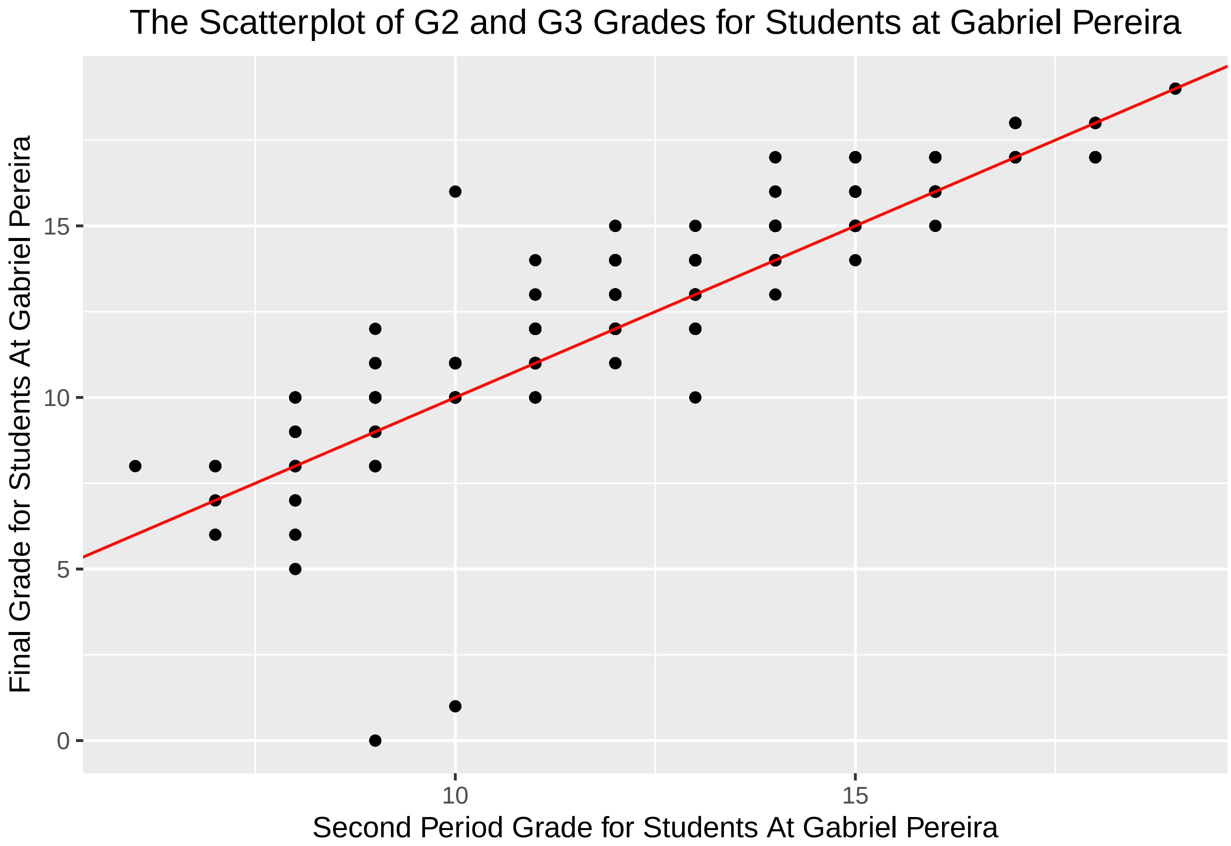
```
##   q2.median
## 1        12
```

The median grade for students that travel <15 minutes and study >10 hours have a median grade of 12.

(3) Focus only on the students in the school "Gabriel Pereira", create a graph that shows the relationship between the students first period grade and their final grade for the Portuguese language course. Write a few sentences discussion that reflects your opinion about such relationship. Make sure to use some knowledge we learned in class to accurately reflect the relationship, the aesthetic settings, background color and theme, title (please center the title), labels, are clear and appropriate.

```
filteredDataQ3 <- filter(por, school == "GP")

ggplot() +
  geom_point(aes(x=G2, y=G3),data=filteredDataQ3) +
  labs(x="Second Period Grade for Students At Gabriel Pereira", y="Final Grade for Students At Gabriel
  theme(plot.title = element_text(hjust = 0.5)) +
  ggtitle("The Scatterplot of G2 and G3 Grades for Students at Gabriel Pereira") +
  geom_abline(slope=1, intercept=0, color="red")
```



```
# reference for "y=x" line:
# https://stackoverflow.com/questions/20436549/adding-x-y-line-to-hexplot-in-ggplot2
```

Most of the students that ended with their final grade being remotely similar to their second quarter grade. We can see a small increase of students grade towards the end of the semester since a lot of points are above the "y = x" line.