STA404/504 HW3

Ethan Gutknecht

08/21/2022

Learning Objectives:

- Basic visualizations 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.

The items below will be considered for grading:

- The plots are correct, with professionalism.
- Axis labels and titles are correct and complete.
- Units are clearly labeled.
- Proper grammar in the write-up.
- The discussion and the story told is interesting and appropriate.

Question 1.

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

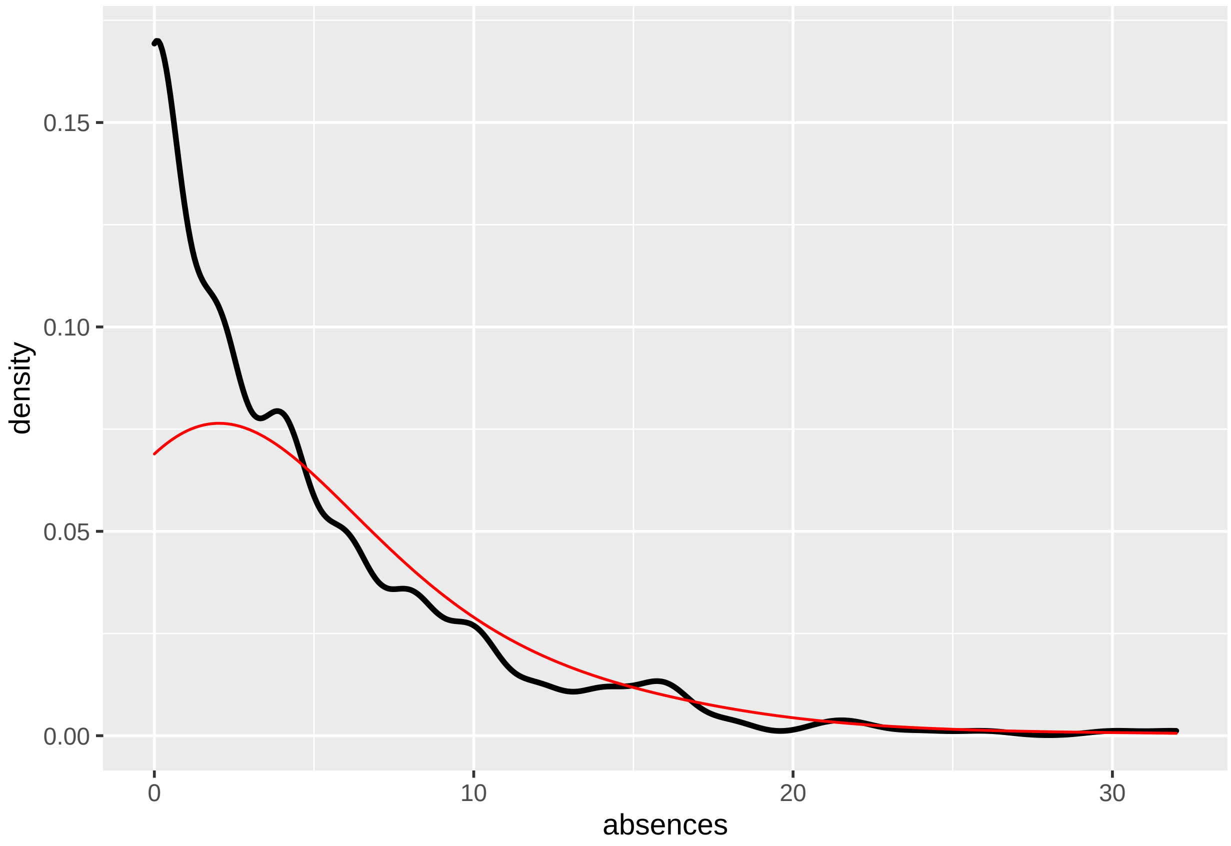
Answer the following questions focus on only the students in the school **"Gabriel Pereira"**. (**Hint:** among many other options, pipe function with filter will be useful.)

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

gabrielPereira <- filter(allData, school == "GP")
```

1. Create a plot that shows the distribution of the number of school absences. Provide some discussions.

```
ggplot() +
  geom_density(aes(x=absences), data=gabrielPereira,adjust=0.7, size=1.0) +
  geom_density(aes(x=absences),color="red", data=gabrielPereira,adjust=3, size=0.5)
```



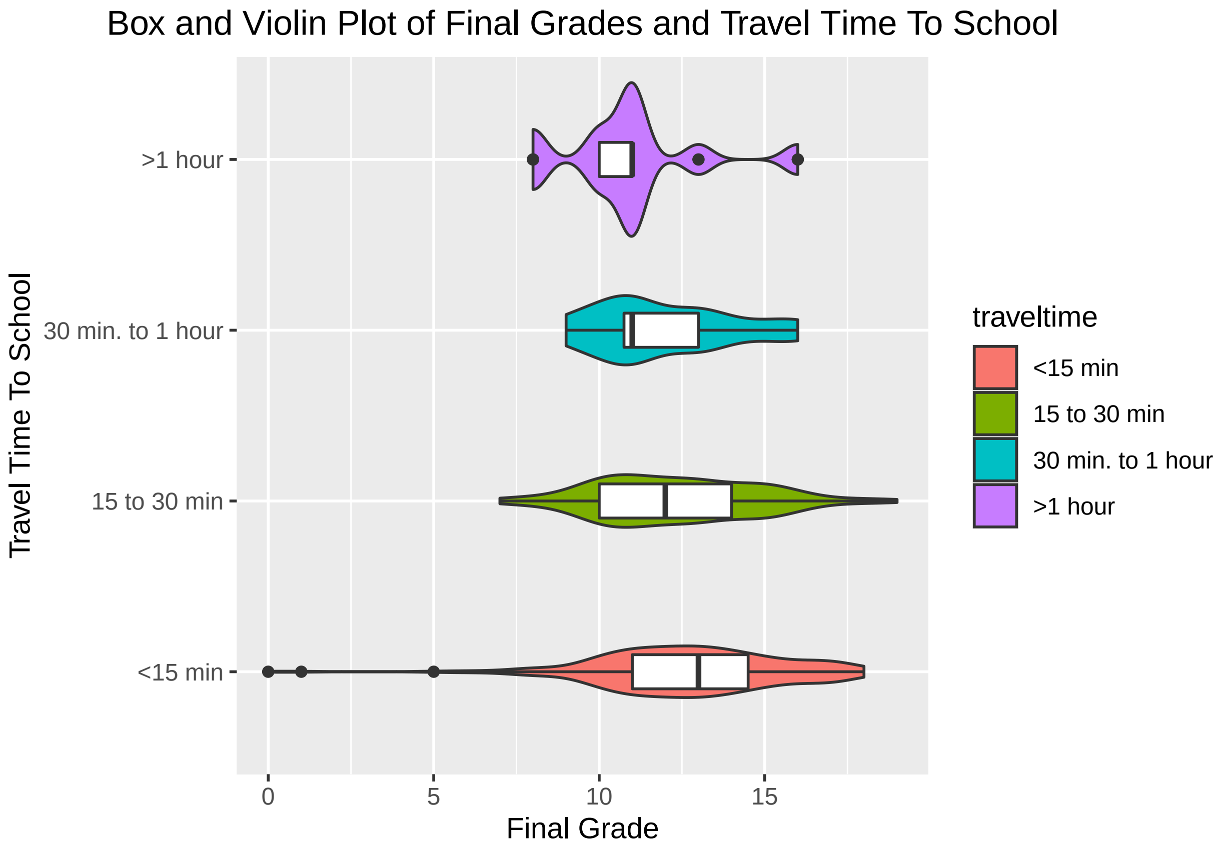
As we can see, majority of the people that went to Gabriel Pereira high school has less than ten absences. There were some students that has more than ten absences but that makes up a small proportion of the students.

2. How does the final grade differ by students' home to school travel time? Use a plot to help tell a story of your finding. Describe in a few sentences what this display tells. Why did you choose this type of display?

Please make sure the label you use for travel time is reader friendly, do not use 1,2,3,4, etc., to represent the students' home to school travel time. (Hint: among many other options, pipe function with mutate will be helpful.)

```
# Change the label of travel times to be accurate
travelTimeFactorLabels <- factor(gabrielPereira$traveltime, labels = c("<15 min","15 to 30 min", "30 mi
gabrielPereira$traveltime <- travelTimeFactorLabels

# Create plot
ggplot() +
  geom_violin(aes(x=G3,y =traveltime, fill=traveltime),data=gabrielPereira) +
  geom_boxplot(aes(x=G3, y = traveltime), width=0.2, data = gabrielPereira) +
  labs(x="Final Grade", y= "Travel Time To School") +
  ggtitle("Box and Violin Plot of Final Grades and Travel Time To School") +
  theme(plot.title = element_text(hjust = 0.5))
```

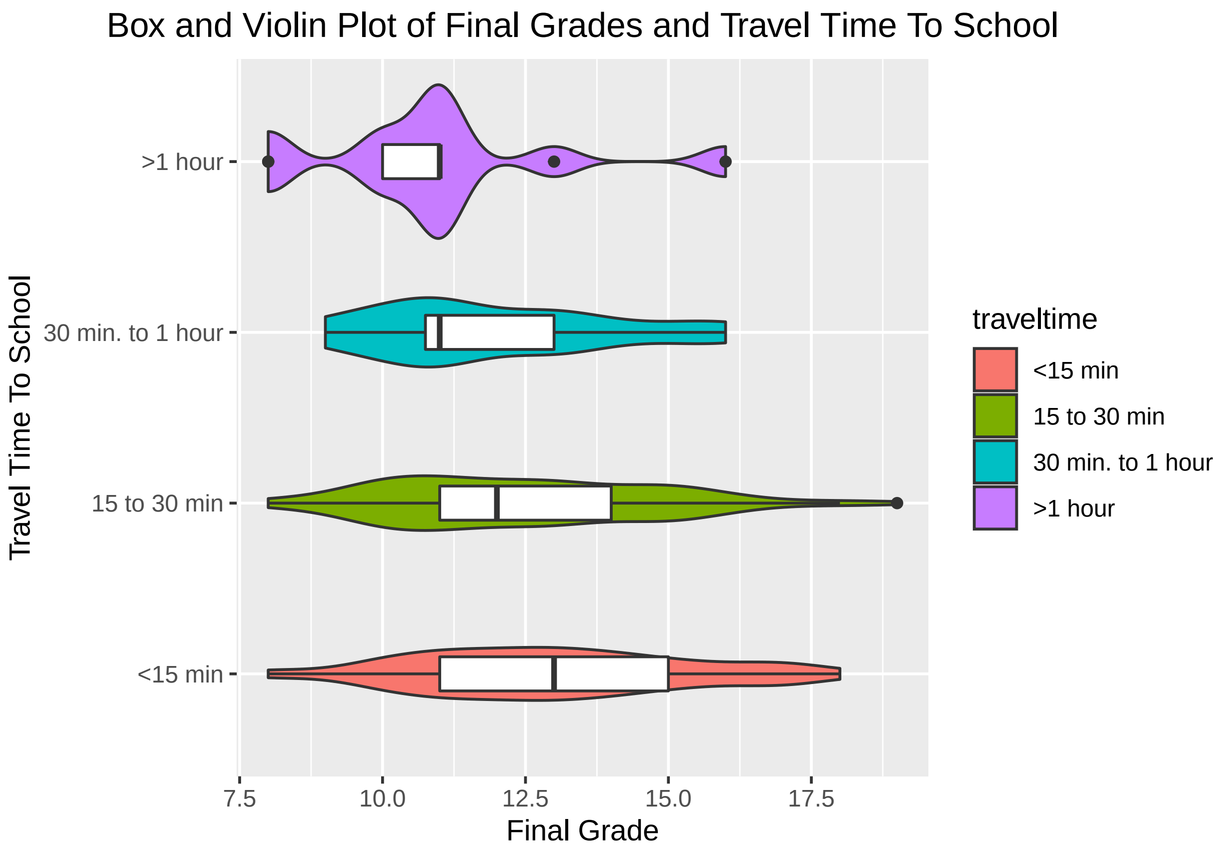


As you can see from the plot, the median of the final grade goes up as the travel time goes down. People that travel less than 15 minutes to school have the best median final grade out of everyone. I chose this plot because I knew it would give me a great visual of the distribution of the grades. I decided to add the box plot over top of the violin plot because it allows us to compare the median grades of all of the different travel times.

3. An important process in data visualization is to critique exiting visuals and improve them. Now take a look at the plot you created in question 2, are there anything you want to edit or improve? Provide a new graph that satisfies your need. (Hint: you may think about the design of the graph, the information you want to show in the graph, etc. to improve your graph in question 2.)

```
q3Data <- filter(gabrielPereira, G3 > 7);

# Create plot
ggplot() +
  geom_violin(aes(x=G3,y =traveltime, fill=traveltime),data=q3Data) +
  geom_boxplot(aes(x=G3, y = traveltime), width=0.2, data = q3Data) +
  labs(x="Final Grade", y= "Travel Time To School") +
  ggtitle("Box and Violin Plot of Final Grades and Travel Time To School") +
  theme(plot.title = element_text(hjust = 0.5))
```

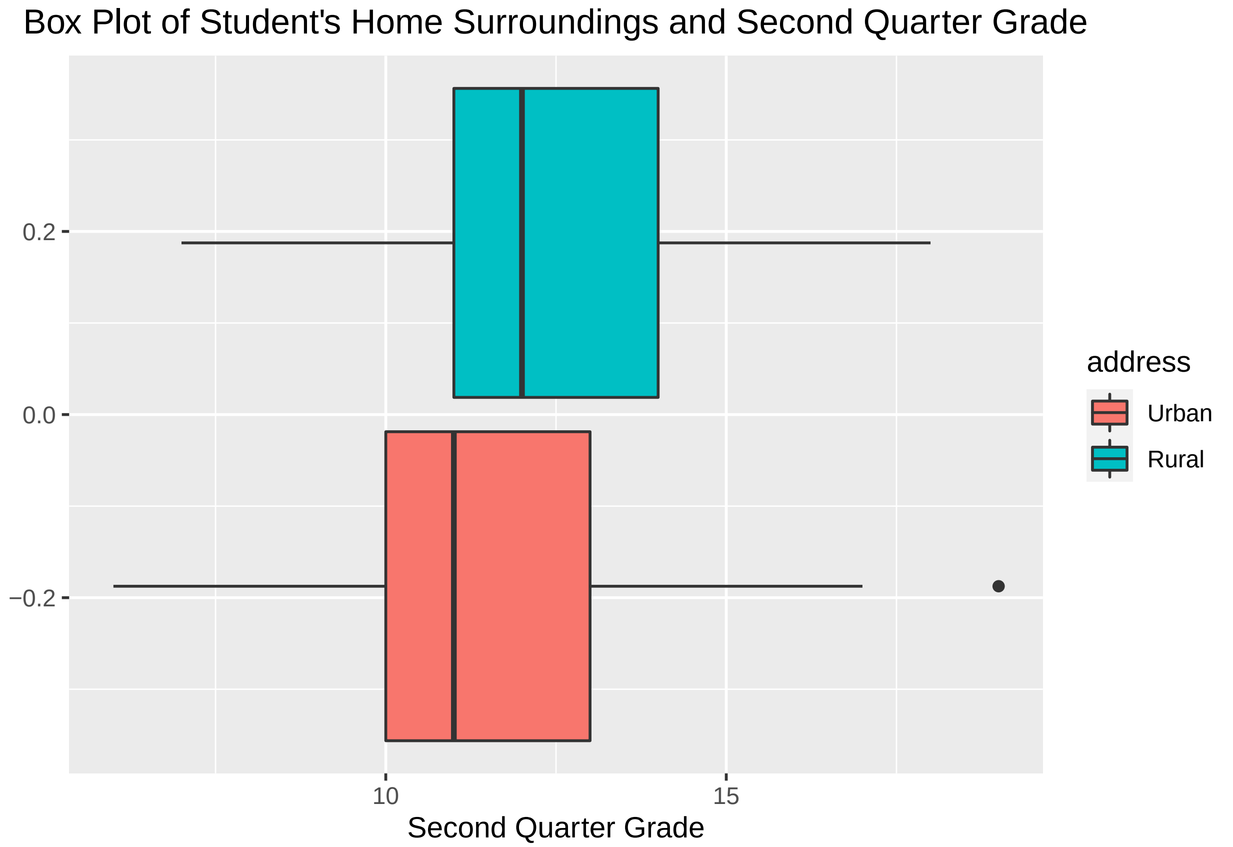


I decided to filter the data from question two to only show data of final grades above 7.5. The travel time of <15 had some outlines that made the distribution smaller than it should have been. Now you can see each distribution easier while not affecting the medians of the final grades.

4. Create an additional plot that tells an interesting story about this data. Describe the plot in a few sentences. Feel free to choose the variables and the type of plot by yourself. Just keep in mind that the variables used in the plot, the type of this plot, and the associated-story should be different from those already been discussed in questions 1 through 3, and the question 2 (3) in homework 2.

```r
# Change the label of travel times to be accurate
addressLabel <- factor(gabrielPereira$address, labels = c("Urban","Rural"))
gabrielPereira$address <- addressLabel;

ggplot() +
  geom_boxplot(aes(x=G2, fill=address),data = gabrielPereira) +
  labs(x="Second Quarter Grade") +
  ggtitle("Box Plot of Student's Home Surroundings and Second Quarter Grade") +
  theme(plot.title = element_text(hjust = 0.5))
```



Based on the information from the plot, we can see that students that live in Rural areas and go to Gabriel Pereira high school have a higher median and distribution location of second quarter grade than students who live in a urban area.