STA490 Mini EDA Case Study

Introduction

Previous studies have shown that multi-tasking seemed to be the hardest in the early morning, which implies that the time of day can greatly affect human's reaction time. Therefore, this exploratory data analysis aimed to investigate whether there is a diurnal pattern to one's reaction time, in other words, whether there is a relationship between the time of day and one's reaction time.

The data for this research was collected from 40 students, each conducted 4 reaction time tests in a day at different times. The first measurement was recorded within one hour of waking up; the second measurement was recorded 4-6 hours after the first measurement; the third measurement was recorded 4-6 hours after the last measurement; and the last measurement was recorded within one hour of going to bed. The data also included other variables to record:

- `Stimulant`, which indicates whether the student has had anything with a stimulant effect.
- `Fatigue`, which indicates the degree of fatigue the student was feeling, on a scale of 1-7.
- `Hunger`, which indicates the degree of hunger the student was feeling, on a scale of 1-10.
- `ill`, which indicates whether the student got sick.

Data cleaning

There are two parts in my data cleaning section: 1.Data checking and 2. Missing data.

I first did some basic checking on data and noticed that there were some missing values and incorrect-formatted data, so then I worked on some of the missing data in order to move on to the rest of my analysis. More detailed steps and explanations are shown below.

1. Data checking

After some basic checkings on the data, I found that:

- The numbers in `student` and measurements in `order` are consecutive and have no missing values.
- `rt` does not contain any unusual values that seem physically impossible for a reaction time.
- `rt` has some missing data. Specifically, student 4, 21, 32 each has one `rt` data missing, and student 35 has all four `rt` data missing.
- `fatigue` and `hunger` also have some missing data.
- Data in `stimulant` and `ill` is not well recorded/formatted, containing some mix case (combination of the lower and upper case) character and some other incorrect form of data (response other than yes/no).

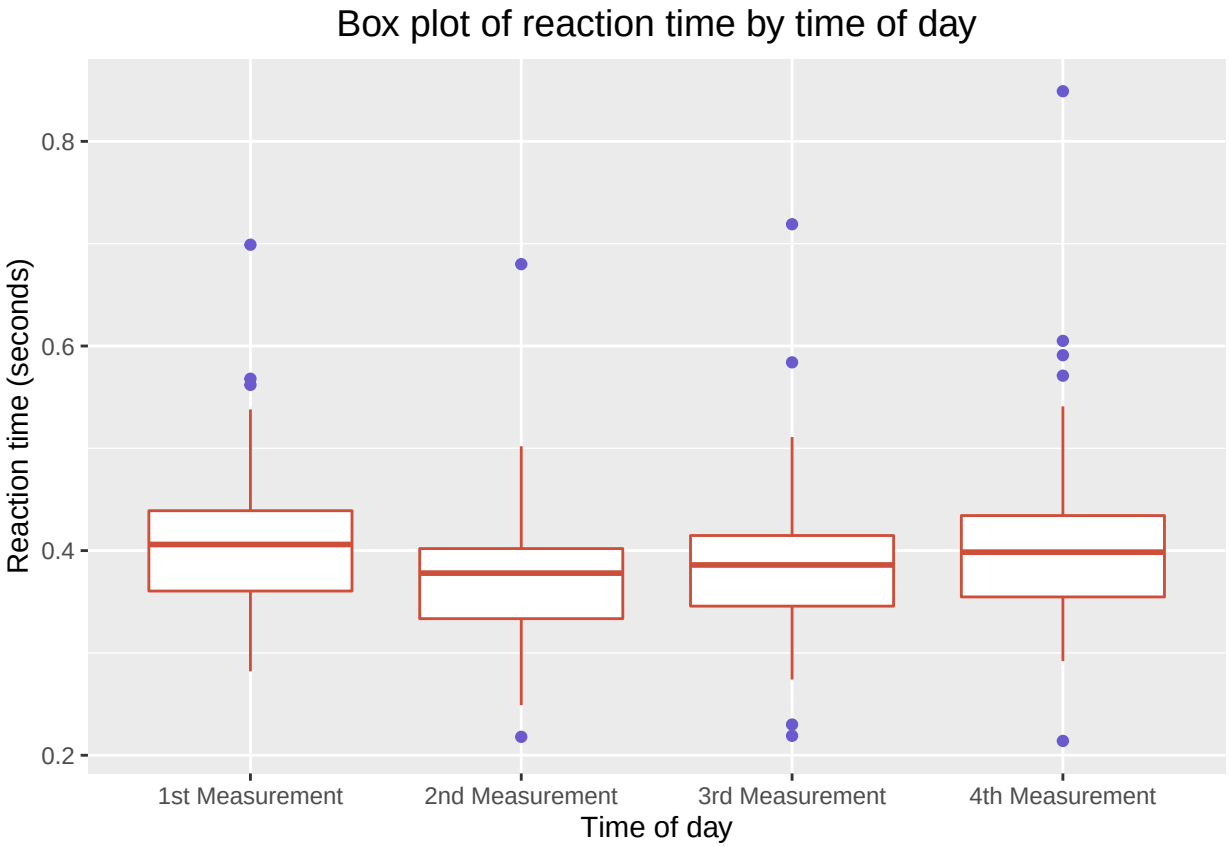
Since the research goal was to visualize the relationship between the time of day and one's reaction time, using `student`, `order` and `rt` as variables in the later analysis, I decided to leave the missing data in `fatigue` and `hunger` and to deal with the missing data in `rt` only at this stage. Similarly, I did not deal with the not well recorded/formatted data in `stimulant` and `ill` just yet, consideing our focus in the later analysis.

2. Missing data

Student 35 was taken out directly, since there was no measurement for `rt` at all, but our research goal was to investigate the daily pattern to reaction time. For the rest of the missing data, I decided to remove the observations missing `rt` values, since the 4 observations consist of only 2.5% of our overall dataset. I did not chose to replace them with mean/median, because replacing would possibly obscure the trend of our data, which is one of the most important and interesting things to look into.

Preliminary insights

A botxplot was used to visualized the relationship between the time of day and one's reaction time. Since a boxplot is much more compact than a histogram and able to fit more on one plot, we can more easily compare the distributions of reaction time during different time of the day, as shown below:



From the box plot, we can see that the median reaction time are very close among 4 measurements during the day, which is around 0.4 seconds, with the second and third measurement slightly below 0.4. The interquartile range (Q3-Q1) are also similar among 4 measurements, which indicates a similar variability of the four distributions. There are some outliers from each of the four measurements.

These findings suggest that on average, the reaction time of a student does not vary too much with different time of day. In other words, the reaction time of a student on average is very similar, no matter what time the student conducted the reaction time test. There might be some relatively long and relatively short reaction time measurements as indicated by the outliers among different times of a day, however, the fist and fourth measurement tend to have slightly more relatively long reaction time measurements than others.

Next steps

Now we have visually examined the relationship between time of day and one's reaction time, we can move on to some next steps. In the next steps, I would like to work on some further investigations:

- Clean the missing data in `fatigue` and `hunger`, either remove or impute.
- Clean and preparing text data in `stimulant` and `ill` by converting all letters to lower or upper case 'yes/no' and converting particular words to 'yes/no'.
- Investigate the relationships between `rt` and all variables `order`, `stimulant`, `fatigue`, `hunger` and `ill`, by fitting a GLM model.
- Investigate if there is any interaction between these variables which would affect one's reaction time, by summarizing the model and statistics above. Since intuitively, high degree of `fatigue`, and `hunger` or a positive response to `ill` at the same time might have a significant effect on one's reaction time.

In the future analysis, we might encounter potential challenges as well:

- If we choose to remove all the missing data, our sample size might become too small to draw a conclusion, considering we had 160 measurements only.
- If we choose to impute the missing data, however, might potentially change the overall trend of our dataset.
- The cleaning process of `stimulant` and `ill` might be challenging, since there is not only mix-case character but also multiple text-format responses.