DATA 551 Project

Yi Tang, Chenxi Yang

14/03/2020

Introduction

Given three datasets about several aspects of three cars, including Saturn, Prius2007 and Prius2019. Regarding these datasets, we want to answer several questions using visualization by the package `ggplot2` in R.

For different issues, data is differently processed to statisfy the requirement of each question.

Are there any obvious recording errors in the files?

While briefly observe the dataset, we see four errors among all three datasets.

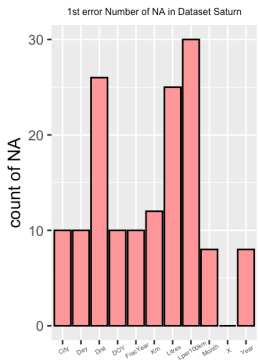
Firstly, the most obvious error for three datasets is that there are several `NA` s involved. Three bar charts show the number of `NA` in each column in the datasets provided. We observe that there are respectively 10, 5 and 0 columns containing `NA` terms in Saturn.csv, Prius2007.csv and Prius2019.csv.

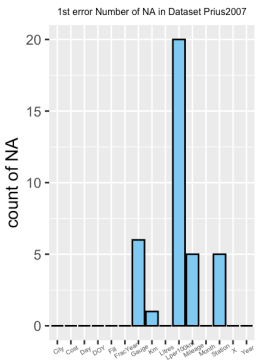
Secondly, the error is in the column named `Lper100km` in dataset Prius2007.csv. According to the boxplot, there are two outliers need to be focus on as one record is over 100 and the other one record is lower than 0, which are not normal.

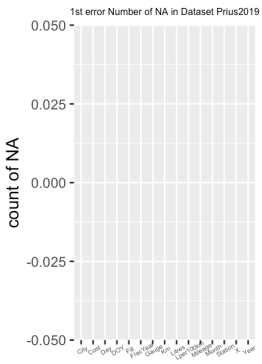
Thirdly, the error is in the column named `Gauge` in dataset Prius2019.csv. As we know the definition of `Gauge` is What the car reports as the number of 10ths of a tank at the time of the fill, which means all numbers in this column should be integers. However, we notice that there are two unnormal decimal points in the dataset. Therefore, we make a scatterplot to represent all observations and highlight the incorrect points in red color.

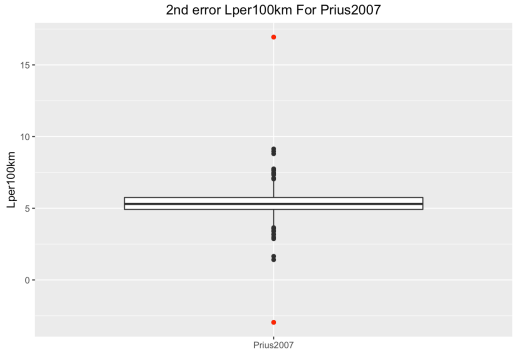
Lastlyly, for the column named `Station` in both Prius2007.csv and Prius2019.csv, there is one identical gas station that is labelled to two names `love` and `loves` , which are 5 and 3 observations respectively. In our later analysis, we convert `love` into `loves` to make the analysis more appropriate.

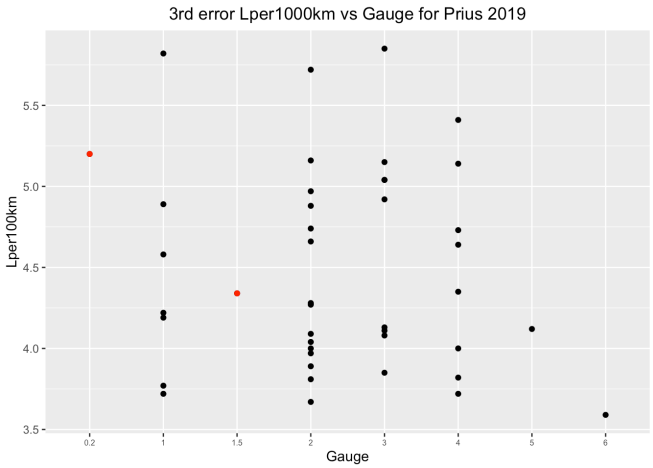
In order to continue further analysis, we remove the rows that have `NA` at each time using the columns and remove these unnormal records mentioned above and convert `love` to `loves` .

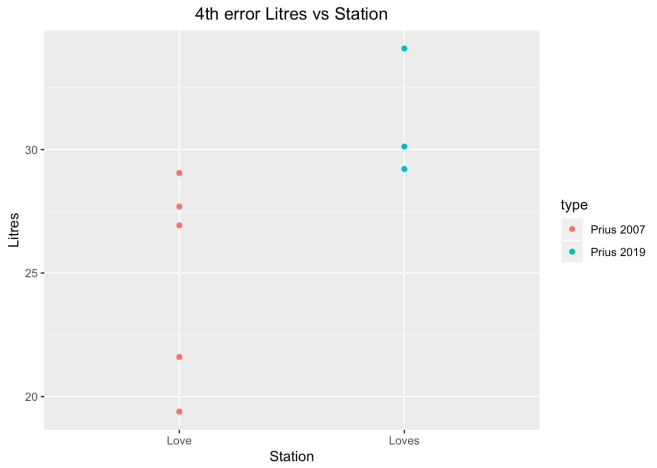






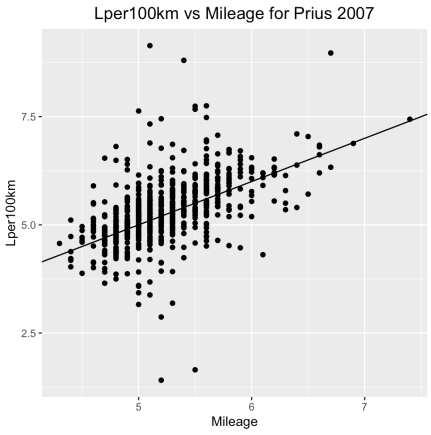


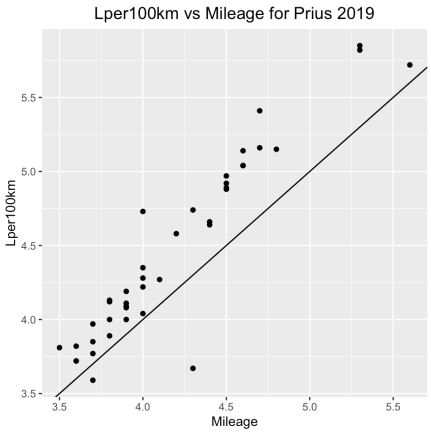




4 obvious recording errors

Is “Mileage” accurate, or should I continue to calculate it myself?

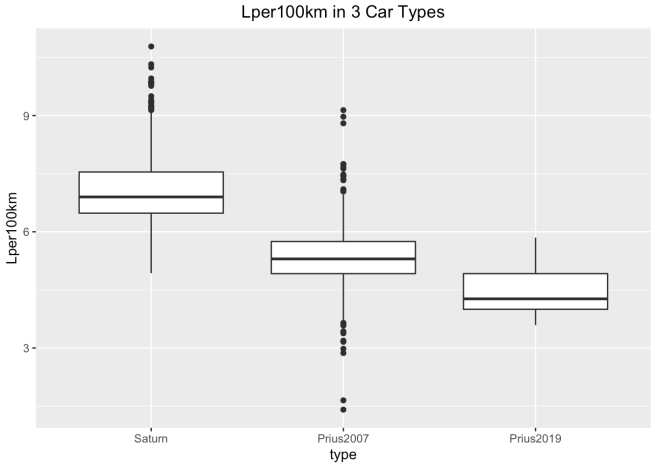




We should continue to calculate `mileage` ourselves. According to the scatterplot of mileage vs Lper100km, the points do not follow the regression line (y=x), which indicates at same situation, the number of mileage reported by the car is different to the number of litres per 100 kilometers calculated by ourselves. Therefore, we should calculate it manually.

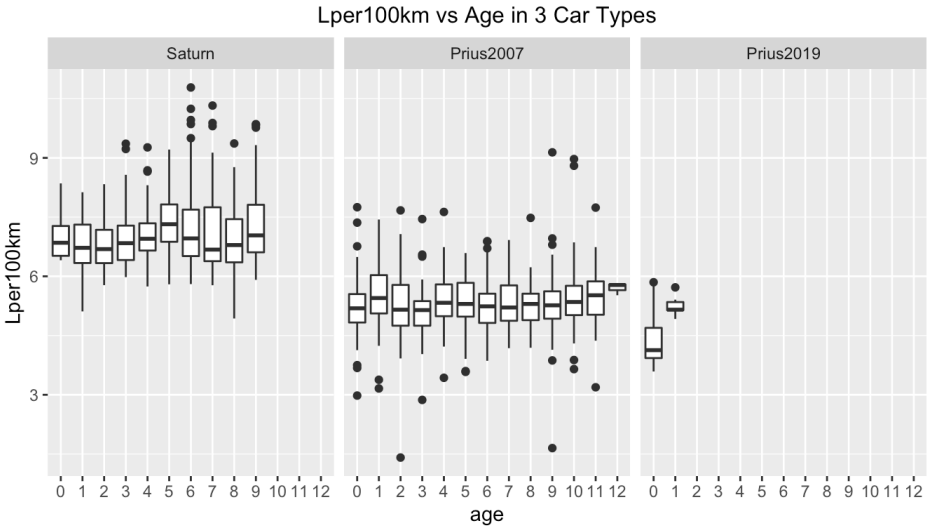
What variables appear to be related to fuel efficiency?

The relationship between fuel efficiency and the car



According to the boxplot, we observe that Saturn has the largest median value of litres per 100 kilometers, followed by Prius2007 and Prius2019. As more litres used per 100 kilometers indicates lower fuel efficiency, so in this case, we can conclude that the newest car has higher fuel efficiency.

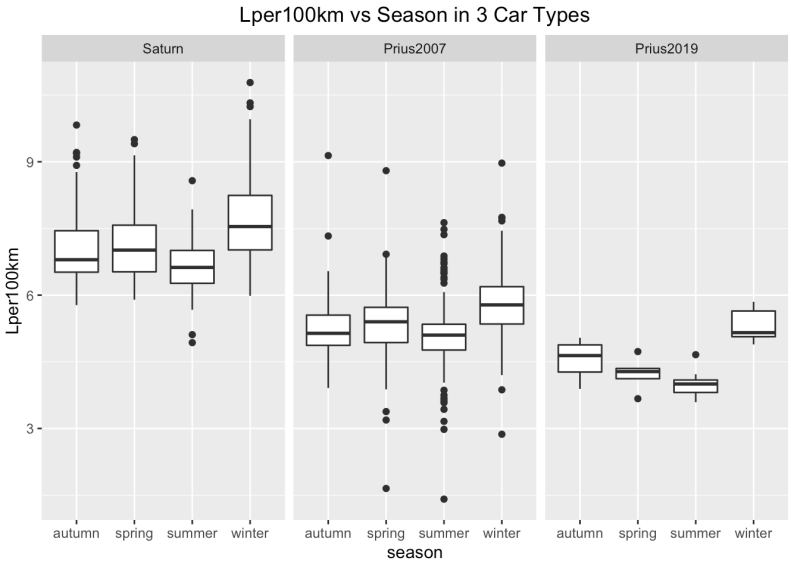
The relationship between fuel efficiency and the age of the car



According to the boxplots below, for Saturn and Prius2007, there is no significant change of the median of the fuel efficiency, which indicates the fuel efficiency is not largely affected by the age of the car. For the boxplot of Prius2019,

we see a relatively large increase of the median of the fuel efficiency from 0 year to 1 year. However, as there are only two records regarding to this car, the result may not be unrepresentative.

The relationship between fuel efficiency and the season of the year

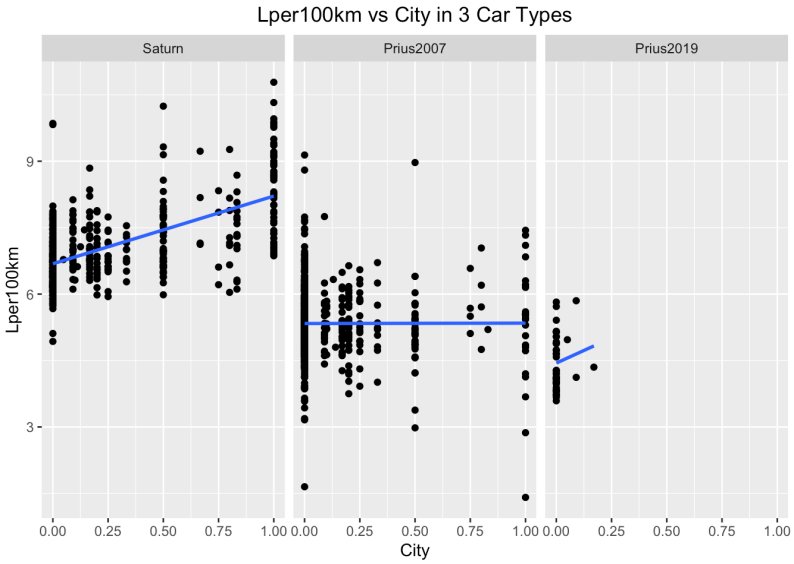


Regarding to the season, we divide 12 months into four seasons:

Spring: March, April and May
Summer: June, July and August
Autumn: September, October and November Winter: January, February and December

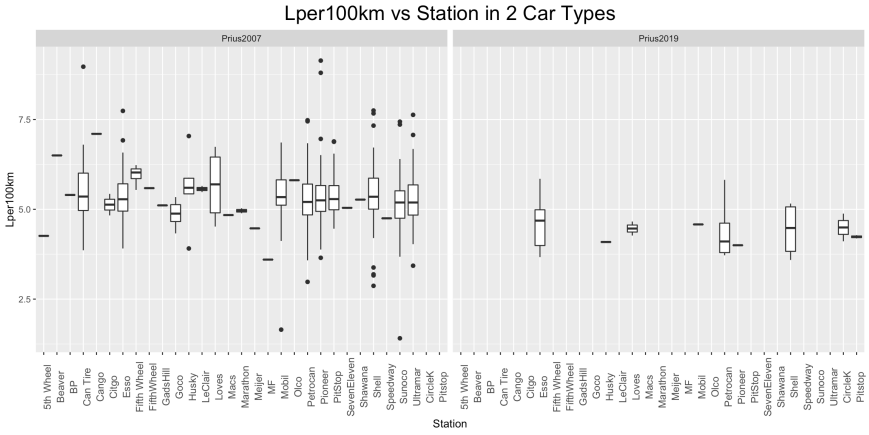
According to the boxplot below, there is no significant difference between the fuel efficiency in Spring, Summer and Autumn. However, the fuel efficiency in Winter is relatively higher than other three seasons, which may lead to the colder weather.

The relationship between fuel efficiency and the city driving



According the scatterplots and regression lines below, we observe that for Saturn and Prius2019, along with the increase of city driving, more fuels are used per 100 kilometers, which indicates to the lower fuel efficiency. However, for Prius2007, the regression line seems straight. For different value of city driving, the fuel efficiency does not have much change.

The relationship between fuel efficiency and the brand of gasoline



As the brand of gasoline only exists in two datasets: Prius2007.csv and Prius2019.csv, so we extract these two columns from the original datasets.

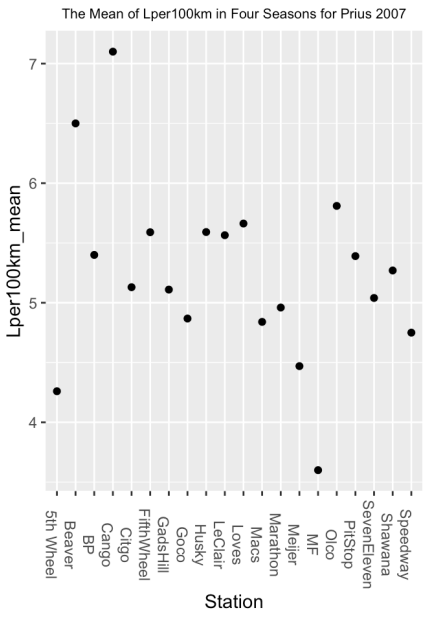
According to the boxplot below, we observe that for different type of car, different brand of gasoline perform better.

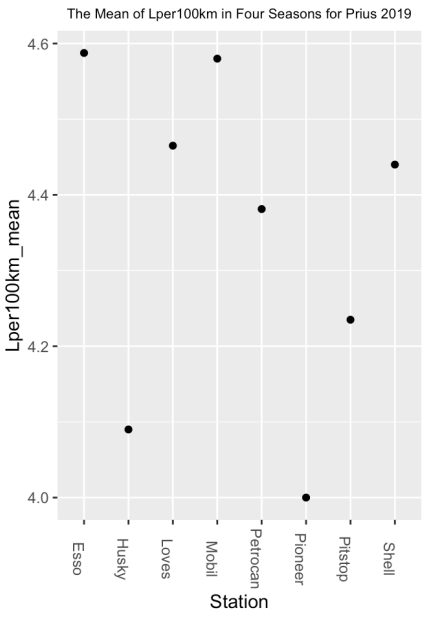
For Prius2007, get rid of the influence of outliers, Loves seems to have lower fuel efficiency as its median of litres per 100 kilometers is larger than other brands, while MF seems to have higher fuel efficiency.

For Prius2019, all nine brand seem to have relatively identical fuel efficiency.

If we calculate the mean of litres per 100 kilometers for different brands, we get the same result for Prius2009 that Loves seems to have lower fuel efficiency, while MF seems to have higher fuel efficiency than other brands. For Prius2017, Esso seems to have lower fuel efficiency, while Pioneer seems to have higher fuel efficiency than other brands. However, the difference is within 0.6 litres per 100 kilometers.

If the “Gauge” says I have 2/10 of a tank of gas, how much is really there?

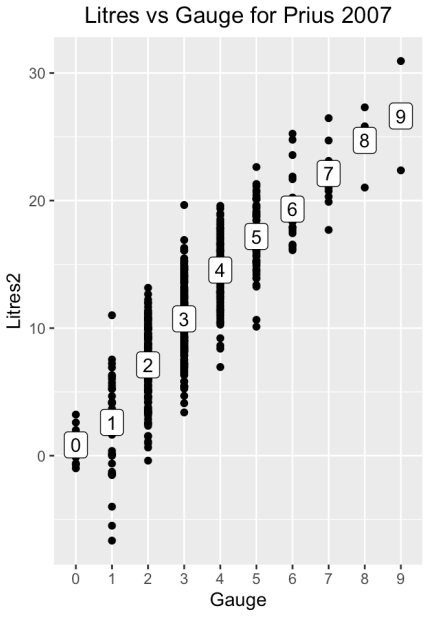


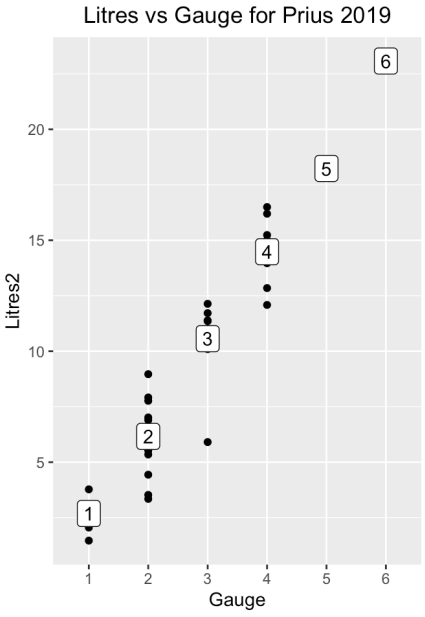


We extract the `Gauge` column from the Prius2007.csv and Prius2019.csv.

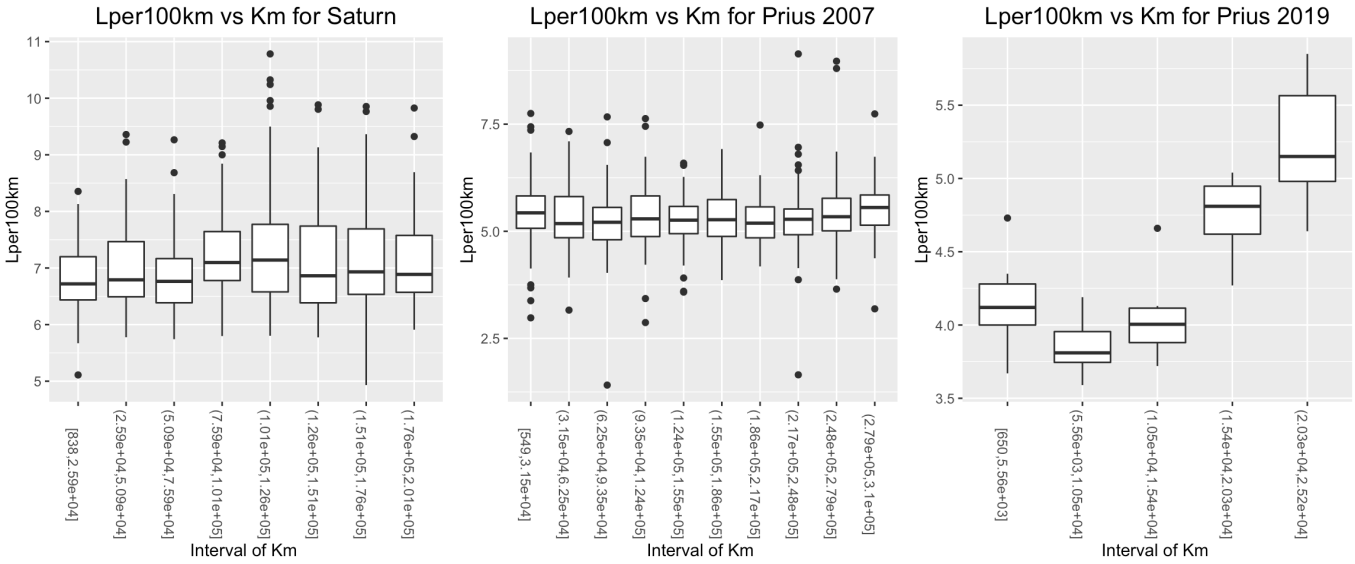
All labels on the plot indicate the mean litres of oil in the tank at that stage.

If we focus on 2/10 of a tank of gas, for Prius2007, there is nearly 7 (7.1181600) litres there, while for Prius2019, there is nearly 6 (6.160955) litres there.





Does the fuel efficiency change along with the increase of kilometers?



We divide the total number of kilometers into several intervals for each car.

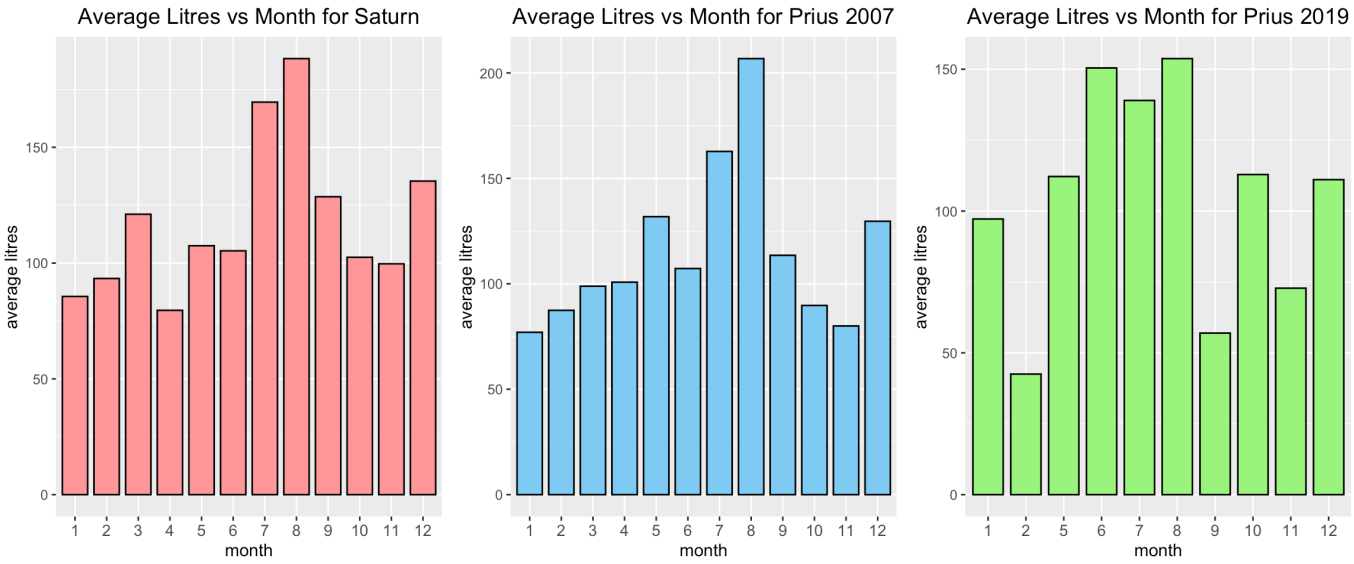
For Saturn, we divide the total number of kilometers into 8 intervals. We observe that the fuel efficiency generally shows a downward trend as the median of litres per 100 kilometers increases over time. However, there are two times of the increase of the fuel efficiency after 50900km and 176000km. Our guess for this result is that the car may have two maintainences at those kilometers. After each maintainence, the fuel efficiency would increase for a period.

For Prius2007, we divide the total number of kilometers into 10 intervals. We observe that the fuel efficiency does not have a significant change over time.

For Prius2019, we divide the total number of kilometers into 5 intervals. We observe that its fuel efficiency is high at the beginning around 500km and gradually lower after 2000km. Our guess is that the brandnew car may have high fuel efficiency. The longer it has been used, the lower the fuel efficiency.

The relationship between month and the fuel consumption

Regarding to the three bar charts, we can observe the relationship between month and the fuel consumption. Genereally speaking, for all three cars, the fuel consumption is high in July and August and relatively low in January. Overall, we can figure that the newer car has higher



fuel efficiency.

Conclusion

In this project, we analyze 6 questions, including the obvious recording errors in the files, the accuracy of “mileage”, variables related to fuel efficiency, real values of “Gauge”, the fuel efficiency changes along with the increase of kilometers and the relationship between month and the fuel consumption.

To sum up, we find that the car with newest technology tend to have higher fuel efficiency than other older cars. However, along with the increase of cumulative kilometers, the fuel efficiency will gradually decrease. In addition, the car type, the age of car, the season and the brand of gasoline may have an impact on the fuel efficiency, but the relationship varies with each car. Last, based on the result about the fuel consumption, we conclude that these three cars are more used in summer time than in winter time.