Assignment Introduction to Data science 2024

Introduction

Solve the questions below and report your solutions and findings using RMarkdown. The final pdf should be submitted via Canvas. The deadline for this assignment is **June 24, 2024, 11.59pm**.

This assigment will ask you for an exploratory data analysis of a data set. Figures should be made using the package ggplot. Pay attention to the layout of the plot. It is important that you always comment on the results you generate and provide conclusions and interpretations if possible.

Load the `mpg` data set which is coming together with the ggplot package:

```
library(ggplot2)
summary(mpg)
```

```
##  manufacturer          model               displ            year     
##  Length:234         Length:234         Min.   :1.600   Min.   :1999  
##  Class :character   Class :character   1st Qu.:2.400   1st Qu.:1999  
##  Mode  :character   Mode  :character   Median :3.300   Median :2004  
##                                        Mean   :3.472   Mean   :2004  
##                                        3rd Qu.:4.600   3rd Qu.:2008  
##                                        Max.   :7.000   Max.   :2008  
##       cyl           trans               drv                 cty       
##  Min.   :4.000   Length:234         Length:234         Min.   : 9.00  
##  1st Qu.:4.000   Class :character   Class :character   1st Qu.:14.00  
##  Median :6.000   Mode  :character   Mode  :character   Median :17.00  
##  Mean   :5.889                                         Mean   :16.86  
##  3rd Qu.:8.000                                         3rd Qu.:19.00  
##  Max.   :8.000                                         Max.   :35.00  
##       hwy             fl               class          
##  Min.   :12.00   Length:234         Length:234        
##  1st Qu.:18.00   Class :character   Class :character  
##  Median :24.00   Mode  :character   Mode  :character  
##  Mean   :23.44                                        
##  3rd Qu.:27.00                                        
##  Max.   :44.00                                        
```

```
Below you find an overview of the different variables:
```

- `manufacturer`: manufacturer of the car.
- `model`: Model of the car.
- `displ`: engine displacement in liters.
- `year`: year of manufacturing
- `cyl`: number of cylinders
- `trans`: type of transmission
- `drv`: drive type. f=front wheel, r=rear wheel, 4=4 wheel
- `cty`: city mileage in miles per gallon

- `hwy`: highway mileage in miles per gallon. This means that more economic cars have higher values for `hwy`.
- `fl`: fuel type
- `class`: Vehicle class (e.g. SUV, minivan, etc.)

  The variable of interest is `hwy`, the mileage on the highway. The other variables are the predictor variables which can potentially be used to predict the highway mileage of a car.

You are asked to perform an exploratory data analysis of this data set. Below, you find some ideas to get started.

Q1. (10 points) Investigate the variable `hwy` by using a histogram and a boxplot. Determine the measures for the center (mean and median) and the spread (standard deviation and IQR). Use the boxplot to indicate possible outliers. Colour outliers in red by using `+geom_boxplot(outlier.shape = 2, outlier.colour =` red")".

Q2. (10 points) Determine which observations are outliers using the IQR method and the Z scores.

Q3. (2 points) Make a pie chart of the variable `class` using the function `pie`.

Q4. (10 points) Investigate the `hwy` in relation with the variable `class`. Make a separate boxplot for each level of `class`. Indicate outliers in red in the boxplots. Interpret the boxplots to understand the effect on the mileage of a car. Use the function `aggregate` to make a table with the mean, median, standard deviation and IQR (in the different rows) of each vehicle class (in the columns). The table should have column and row names. Why do we find more outliers if we consider separate vehicle classes compared to the entire data set (see question 1)?

Q5. (5 points) The variable `displ` is a continuous variable. However, we can also make boxplots in function of a continuous variable. For this, you add the following code: `+geom_boxplot(varwidth = FALSE, aes(group = cut_width(displ, 1)))`. If you change `cut_width(displ, 1)` to `cut_width(displ, 0.5)`, what do you see?

Q6. (5 points) Make a histogram of the variable `hwy`, but now add also the year of manufacturing in the figure. Calculate the mean for `hwy` for cars manufactured in 1999 and in 2008.

Q7. (8 points) Investigate in a similar way the relation between `hwy` and the vehicle class by constructing a histogram and calculating the mean for `hwy` for the different vehicle classes. Also create a scatterplot between the variables `cty` and `hwy`. Indicate the vehicle class of the observations in the plot.