HW 2

Due: Tuesday January 30, 2pm

Instructions

- Produce your assignment as a RMarkdown document rendered to pdf (knit to pdf).
- Also submit your Rmd file (it will not be graded but we want it for reference purposes).
- Show all the code (use `echo=TRUE` as option in R chunks) as well as the results.
- 100 total points (the TA will determine the points per exercise and upload a revised version later).
- See Syllabus for HW policies.

Exercise 1:

- Make sure you have the library ggplot2 installed. If not, install it by running the uncommented code below:

```
## install.packages("ggplot2")
```

- Run the code below to access and prepare the data:

```
## Load the library
library(ggplot2)
## Activate the mpg data.frame provided by ggplot2
data(mpg)
## New versions of ggplot provide a tibble and have all character variables, while
## older were data frames and had factors. Transform manufacturer to factor
## to show how most data.frames treat character variables.
mpg <- as.data.frame(mpg)
mpg$manufacturer <- factor(mpg$manufacturer)
```

1. Inspect the structure of mpg data.frame. Note the kind of data provided.
2. Run the summary function to learn more about the variables.
3. Get a subset of the data.frame including all cars that are not `ford` nor `audi`.
4. Determine if the manufacturer variable (that is a factor) in your subset has or not dropped the now removed manufacurers audi and ford.
5. Devise a strategy to assure that the above factor has dropped the levels that have no elements
6. Further subset the data making sure that only front drive midsize cars model 2008 with at least 20 highway or city miles per gallon are included.
7. Determine how many cars per manufacturer satisfy these constraints. Only show manufacturers with at least one vehicle.
8. Only show the manufacturer(s) with more cars (Note: your solution should also contemplate the possibility of a tie for the first place.)

Exercise 2:

- Make your own quantile-quantile plot. You will have to compare the quantiles of the theoretical distribution (Normal with mean equal the sample mean of the data and sd equal the sample sd of the data), with your sorted data (see help of `qnorm` for parameters).
- Hint: to find the theoretical quantiles, you will have to create a vector of probabilities with the same number of elements as your data, excluding the extremes 0 and 1.
- Test it on the following data:

```r
set.seed(123)
r <- rnorm(1000)
```

Exercise 3:

- Adapt the quantile-quantile procedure developed above to compare to a Gamma distribution the following data.

```r
set.seed(123)
r <- rgamma(1000, shape = 1)
```

Exercise 4:

- Using `faithful` dataset, do a stem and leaf plot (`stem`), a histogram to which you will add a `rug`, and a boxplot, using in all cases the variable `eruptions`, and check using a test if it can be concluded or not that the eruptions are normally distributed.

Exercise 5:

After running the provided code for simple regression:

```r
library(UsingR)
data(father.son)

## Perform a linear model analysis (regression)
lmfs <- lm(sheight ~ fheight, data = father.son)
slmfs <- summary(lmfs)
anlmfs <- anova(lmfs)
```

1. Extract the coefficients of the regression line from `lmfs`, and add the regression line (in red) to the scatterplot. Hint: use function `abline` (see help)
2. Plot the father heights on x and the residuals in y. Add a horizontal line at 0.
3. Extract the `fstatistic` from `slmfs` and perform the calculation to obtain the p-value of the anova of the regression (i.e., the line that says: `F-statistic: 361.2 on 1 and 1076 DF,  p-value: < 2.2e-16`. (Hint: use the function `pf` that gives you the cumulative probability of an F distribution).
4. Extract the p-value directly from `anlmfs`.