Assignment 1 – Visualization

2201STAT3990 Mateen Shaikh

Due: January 29 by 3:30pm

Instructions:

Complete the following assignment, answering all questions and providing plots when indicated. The total length of the assignment should not be more than 5 pages including plots. Use a standard 12pt font with no less than 1 inch margins. Submit your assignment as a PDF file on moodle, (other formats will be penalized). Also, submit a plain text file (.R or .py extensions should be converted to .txt) containing no more than 100 functional lines of code used to make the plots. (You should need *far fewer* than 100 lines to complete the assignment).

As always: there are no marks for style but there are deductions.

Motor Trend Road Tests

Data on 398 cars are provided, relating to fuel efficiency in the file on moodle, `autompg.ssv`. This is a reformatted version of the file from the UCI machine learning repository.

Ensure you set the working directory to the folder that contains the downloaded file.

Read this file in using the `read.csv` command and save it in some variable. Note that this is space-delineated, so use `sep=‘ ’` as an argument of the `read.csv` command. You can read the helpfile for reading a csv by typing

```
?read.csv
```

such as either the following

```
autompg = read.csv(file=file.choose(),sep=" ")
        #open a dialogue to the file

autompg = read.csv(file="autompg.ssv",sep=" ")
        #assume the file is in the same workding directory set by `setwd()'.
```

If you `View` the file, you will see that there are several missing values indicated by `NA`. Keep only the rows are have no missing values by using the `complete.cases` function. You should be left with 392 observations.

```
completeRows = complete.cases(autompg)
autompg = autompg[completeRows,]
dim(autompg)

## [1] 392   9
```

1. Provide a histogram of the distribution of the fuel efficiencies in miles per gallon (if necessary, look up the help from `?hist`). Appropriately label and title your histogram. Provide a simple one or two sentence cursory (superficial) summary on this histogram and to describe the distribution of data.

2. I am not American and do not understand MPG. Instead, I was this to be in L/100km. The conversion for a value x in mpg to the corresponding value y in L/100km is given by $y = \frac{62.14mi}{100km} \times \frac{3.784l}{1gallon} \times \frac{1}{x\frac{mi}{gallon}} = \frac{235}{x} \quad \frac{l}{100km}$. Create a variable containing fuel efficiency in L/100km. For example:

```
lPer100km <- 235/autompg$mpg
#or this line
lPer100km <- with(autompg, 235/mpg)

#optional: add this new variable to the data frame for easy access
autompg = cbind(lPer100km,autompg)
```

   Construct a boxplot and appropriately label/title it. Summarize this to describe the distribution of this transformed data.

3. The number of cylinders is a discrete variable. Use this to construct side-by-side (approximate) views of distributions of `lPer100km` on one axis and the number of cylinders. Comment on whether you notice a trend in terms of changes in location, spread, and skewness.

4. Repeat the above with miles per gallon and compare the relationship in comparison to the metric units.

5. Create a parallel-coordinate plot coloured in a blue-orange scale according to liters per 100km. You can use the following code:

```
blueOrange = blueOrange <-colorRampPalette(c("orange","blue"))
        #returns a customized function which takes parameters:
        # n = the number of (ordered) colours you want on the scale

theColours = blueOrange(length(lPer100km))
        #returns an actual collection of colours, for as many data points we have

coloursToPaint = theColours[rank(lPer100km)]
        #orders the colours according to the rank of the fuel efficiency
```

Remember to omit the name of the car (last variable of the dataframe), since character strings cannot be mapped to an ordinal scale.

6. Comment on which specific variables seem to be related to be directly related to fuel efficiency.

7. Ignore the colours (you can replot without colours if it helps). Which side-by-side variables seem to be **positively** correlated (high-values of one correspond to high-values of the other)? Also, which side-by-side variables seem to be **negatively** correlated (high-values of one correspond to low-values of the other)?

Cartograms

8. Consider a map like on slide 50. There are instances where such a map is useful. Imagine the map is distorted so that area of a riding is approximately represents the population in the riding.

   Consider the 2019 federal election from the United Kingdom.

   https://geographical.co.uk/places/mapping/item/2276-mapping-the-2017-general-election

9. (a) If someone not from the UK was watching news and did not know the geography of the United Kingdom and did not see borders between ridings, who would they guess won the election from the first graph: red (left-wing) or blue (right-wing)?

   (b) Looking at the third map, the borders have been distorted so that the area in a border is proportional to the population. Is the winner as obviously the same as the first graph?

10. In the third graph, think about why an area would have grown in the third graph and areas that would have shrunk in the third graph. Now, think what colour occurs more and which colour occurs less in the third graph. Make an inference relating to population density and which party the area voted. (The answer is a well-known phenomenon that is also true in Canada in left-right wing politics).

Population

Download the file `pop.csv` and read it in. Unlike autompg file, this one is comma-separated so omit the `sep=` parameter (default is the comma). This gives estimates of the world's population in last few thousand years.

11. Plot the first and second columns of this data set. You can add the parameter `type="l"` to connect the dots with a line, so that it is easier to read. Add appropriate labels to each axis and the title.

12. There is one, not-surprising trend to describe the world's population of humans over time. However, there seems to be a time (literally) that the trend was not true. Perform an internet search and explain where and why there is an exception to this trend.