Assignment 1: Questions

Introduction to Data Science

6/5/2023

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 23, 2023, 11.55pm**.

This assignment will also teach you some useful R commands. Figures should be made using the package ggplot. Pay attention to the layout of the plot.

The sample mean and standard deviation

Question 1

If we have a data set given with n observations denoted by $x_1, x_2, \ldots, x_n$. Then we can always determine the sample mean, denoted by $\hat{\mu}$ and sample standard deviation $\hat{\sigma}$. Moreover, these statistics will always be finite since we have a finite sample.

However, we have to be careful with blindly using the sample mean and sample standard deviation to summarize a data set. In this assignment, we will show situations where the sample mean and standard deviation do not provide useful information about the data set. Moreover, using the sample mean and sample standard deviation can be dangerous, since it does not reflect the true nature of the data.

Run the following code.

```
library(Pareto) # if necessary, first install the package. Use install.pacakges("Pareto")
set.seed(100)
Data=data.frame(x.n=rnorm(50000),x.p=rPareto(50000,t=1,alpha=2))
```

Assume the observations in the data frame `Data` represent observations of two variables that you have to investigate.

1. Use `ggplot` to make a histogram and a boxplot of the variable `x.n`. The `gridExtra` package contains the `grid.arrange` function which is convenient to organize multiple plots.
2. Determine the sample mean and sample standard deviation of the variable `x.n`. Is this what you would expect given the data generation process?
3. Explain how the sample mean and standard deviation that you calculated in the previous question can be used to summarize the variable. In particular, can the mean be used to predict new observations?
4. Consider the following statement: *'The mean and the standard deviation of the observations of the variable `x.p` cannot be used to summarize the data. Moreover, the mean is a bad predictor for new observations because it neglects possible very extreme realizations.'* Provide an analysis to support this statement. Make useful plots and tables.
   Tip: Start by determining the mean and standard deviation of the data set. Make a histogram and boxplot. You can use the function `filter` to determine a subset of a data frame.

Question 2

Load the following data set, containing information about the cars of the policyholders of a car insurer. The first variable, called symboling, indicates the risk class of the driver, where a higher number indicates a higher risk. The variable normalized losses correspond with the yearly loss the insurer incured for this driver. The variable highway.mpg indicates the miles per gallon, the car can drive on a highway. Note that this number is larger for economic cars, whereas a small number indicates a car that needs a lot of fuel. The variable curb weight is the weight of the vehicle without any passengers or items in it except for the standard equipment that comes with it. This is the weight of your vehicle when it's not being used and resting on a flat surface. Lastly, the variable price corresponds with the price of the car. In this data science study, the variable price is the target variable, i.e. the variable we would like to predict.

```
library("ggplot2")
Data = read.csv("Car_data.csv", na.strings=c("?"))
head(Data)
```

```
##   symboling normalized.losses        make fuel.type aspiration num.of.doors
## 1         3                NA alfa-romero       gas        std          two
## 2         3                NA alfa-romero       gas        std          two
## 3         1                NA alfa-romero       gas        std          two
## 4         2               164        audi       gas        std         four
## 5         2               164        audi       gas        std         four
## 6         2                NA        audi       gas        std          two
##    body.style drive.wheels engine.location wheel.base length width height
## 1 convertible          rwd           front       88.6  168.8  64.1   48.8
## 2 convertible          rwd           front       88.6  168.8  64.1   48.8
## 3   hatchback          rwd           front       94.5  171.2  65.5   52.4
## 4       sedan          fwd           front       99.8  176.6  66.2   54.3
## 5       sedan          4wd           front       99.4  176.6  66.4   54.3
## 6       sedan          fwd           front       99.8  177.3  66.3   53.1
##   curb.weight engine.type num.of.cylinders engine.size fuel.system bore stroke
## 1        2548        dohc             four         130        mpfi 3.47   2.68
## 2        2548        dohc             four         130        mpfi 3.47   2.68
## 3        2823        ohcv              six         152        mpfi 2.68   3.47
## 4        2337         ohc             four         109        mpfi 3.19   3.40
## 5        2824         ohc             five         136        mpfi 3.19   3.40
## 6        2507         ohc             five         136        mpfi 3.19   3.40
##   compression.ratio horsepower peak.rpm city.mpg highway.mpg price
## 1               9.0        111     5000       21          27 13495
## 2               9.0        111     5000       21          27 16500
## 3               9.0        154     5000       19          26 16500
## 4              10.0        102     5500       24          30 13950
## 5               8.0        115     5500       18          22 17450
## 6               8.5        110     5500       19          25 15250
```

1. Remove all lines where there are missing values for the variable price.
2. Make a histogram for the variable price.
3. Consider the variables curb.weight, engine.size, horsepower and highway.mpg. Investigate the relation between each of these variables and price of the car. What are your conclusions?
4. Perform a principal component analysis using the four variables curb.weight, engine.size, horsepower and highway.mpg. You can use the function `prcomp`. Can you given an interpretation of the first three principal components?
5. Investigate the relation of each of the principal components with the variable price. What do you observe if you compare with question 2.3?