Assignment 1

The Old Republic

2023-06-21

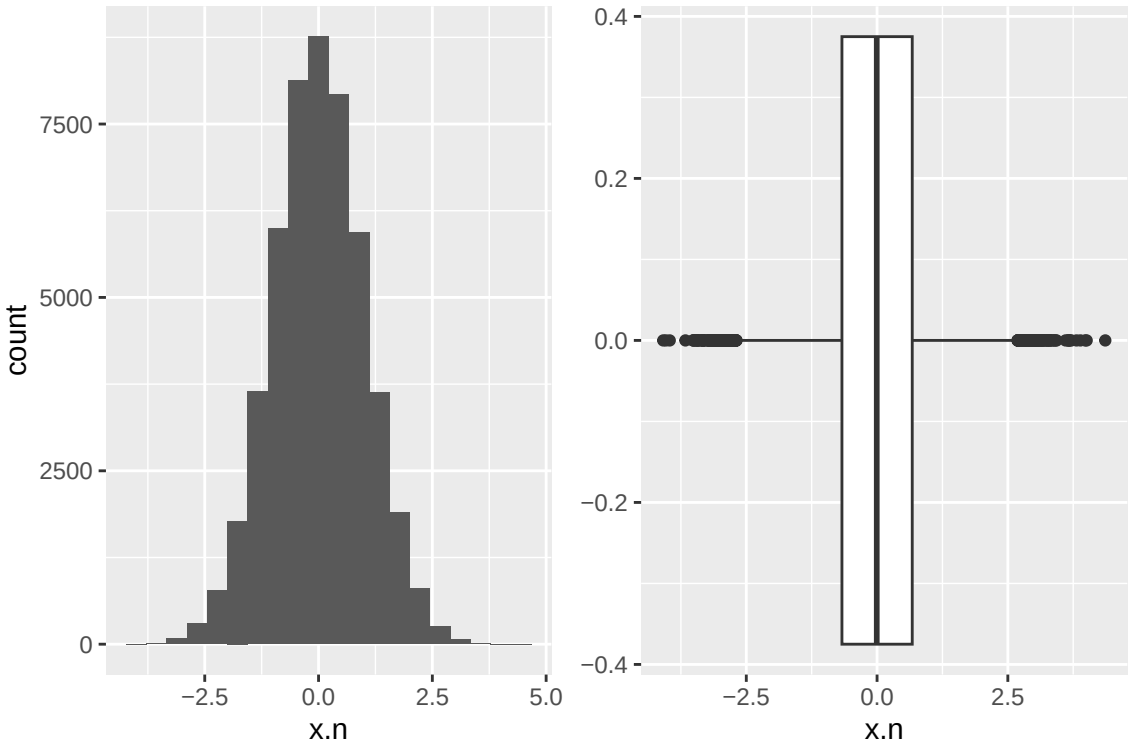
```
library(Pareto)
set.seed(100)
Data = data.frame(x.n = rnorm(50000), x.p = rPareto(50000, t=1, alpha=2))
summary(Data)
```

```
##       x.n                 x.p          
##  Min.   :-4.087893   Min.   :  1.000  
##  1st Qu.:-0.671144   1st Qu.:  1.154  
##  Median :-0.005919   Median :  1.412  
##  Mean   :-0.000208   Mean   :  1.994  
##  3rd Qu.: 0.672466   3rd Qu.:  1.992  
##  Max.   : 4.363243   Max.   :159.275
```

Question 1

1. Histogram and Boxplot of the Variable x.n

```
library(ggplot2)
library(grid)
library(gridExtra)
hist = ggplot(Data, aes(x= x.n)) + geom_histogram(bins = 20)

box = ggplot(Data, aes(x= x.n)) + geom_boxplot()
grid.arrange(hist, box, ncol = 2)
```



2. The sample mean, and standard deviation of x.n are $-2.0849558 \times 10^{-4}$ and 0.9989658 respectively, we see that these parameters are approximately the same as the standard normal distribution. In fact, as we increase the sample size to infinity, the mean and standard deviation will also approach 0 and 1.

Question 2

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

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

```
## 1       21          27 13495
## 2       21          27 16500
## 3       19          26 16500
## 4       24          30 13950
## 5       18          22 17450
## 6       19          25 15250
```

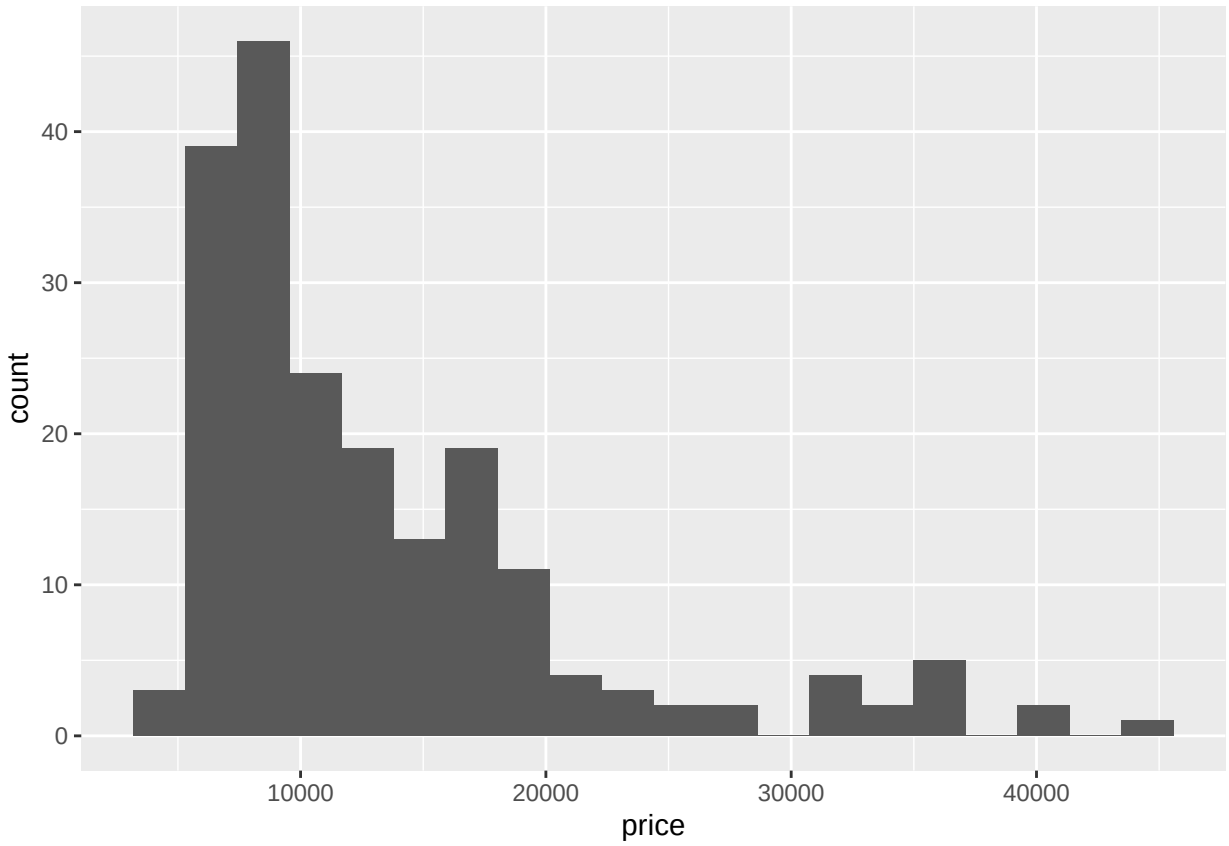
1.

```
library(tidyr)
Data = drop_na(Data, price, curb.weight, engine.size, horsepower, highway.mpg)
head(Data)
```

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

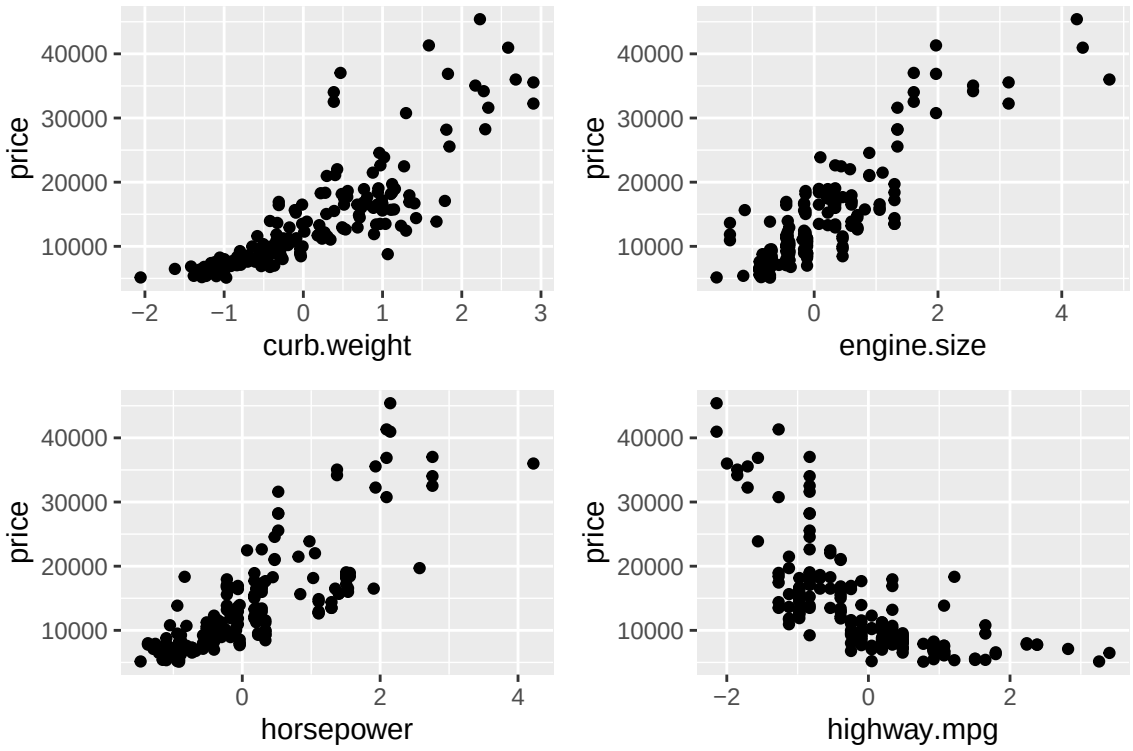
2. Histogram of price

```
ggplot(Data, aes(x= price)) + geom_histogram(bins = 20)
```



3. EDA graphs

```
Data$curb.weight = (Data$curb.weight - mean(Data$curb.weight))/sd(Data$curb.weight)
Data$engine.size = (Data$engine.size - mean(Data$engine.size))/sd(Data$engine.size)
Data$horsepower= (Data$horsepower - mean(Data$horsepower))/sd(Data$horsepower)
Data$highway.mpg= (Data$highway.mpg - mean(Data$highway.mpg))/sd(Data$highway.mpg)
weight = ggplot(Data, aes(x = curb.weight, y = price)) + geom_point()
size = ggplot(Data, aes(x = engine.size, y = price)) + geom_point()
hp = ggplot(Data, aes(x = horsepower, y = price)) + geom_point()
mpg = ggplot(Data, aes(x = highway.mpg, y = price)) + geom_point()
grid.arrange(weight,size,hp, mpg, ncol = 2)
```



We can see from the above plots that, curb.weight, engine.size and horsepower are directly proportional to price. Also, we see that highway.mpg is inversely proportional to price. 4.

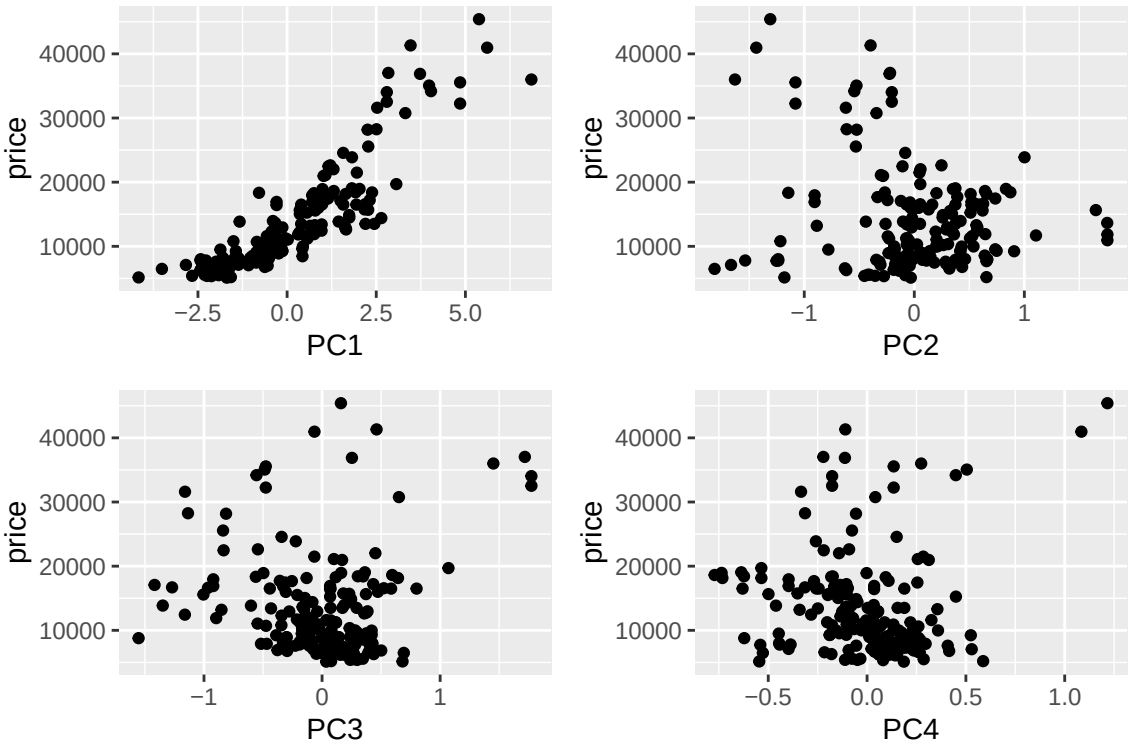
```
Data_stripped = Data[,c("curb.weight", "engine.size", "horsepower", "highway.mpg", "price")]
Data.PCA = prcomp(Data_stripped[, 1:4], center=FALSE)

Data.PCA
```

```
## Standard deviations (1, .., p=4):
## [1] 1.8317619 0.5712544 0.4886130 0.2820890
##
## Rotation (n x k) = (4 x 4):
##                     PC1         PC2        PC3        PC4
## curb.weight   0.5073415 -0.18580359 -0.6571533 -0.5255770
## engine.size   0.4999754 -0.63617369  0.1031988  0.5784960
## horsepower    0.5045853  0.09956758  0.7262252 -0.4561544
## highway.mpg -0.4878759 -0.74219024  0.1734832 -0.4254813
```

- We see from the components of PC1, that all the components are equally significant, and that highway.mpg has an inverse relation to the other components. So, there exists relation between the 'power' of a car, and the mileage.
- We see from PC2, that it depends strongly on mileage and engine.size.
- PC3, depends on horsepower and weight, and might depict a negative relation between horsepower and the weight of a car.

5.

```
pc1 = ggplot(Data, aes(x = Data.PCA$x[,'PC1'], y = price)) + geom_point() + labs(x='PC1')
pc2 = ggplot(Data, aes(x = Data.PCA$x[,'PC2'], y = price)) + geom_point()+ labs(x='PC2')
pc3 = ggplot(Data, aes(x = Data.PCA$x[,'PC3'], y = price)) + geom_point()+ labs(x='PC3')
pc4 = ggplot(Data, aes(x = Data.PCA$x[,'PC4'], y = price)) + geom_point()+ labs(x='PC4')
grid.arrange(pc1, pc2, pc3, pc4, ncol = 2, nrow= 2)
```



We see a strong positive relation between PC1 and price, and no relation between PC2, PC3, PC4 and price respectively. This is in constrast with question 2.3, where there exist relations between each of the variables and the price. This is partly because the principal components are composed of the variables observed in 2.3, and so, the relations are altered based on the loading vectors.