HW2

Yating Liao (7636428840)

2023-01-25

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

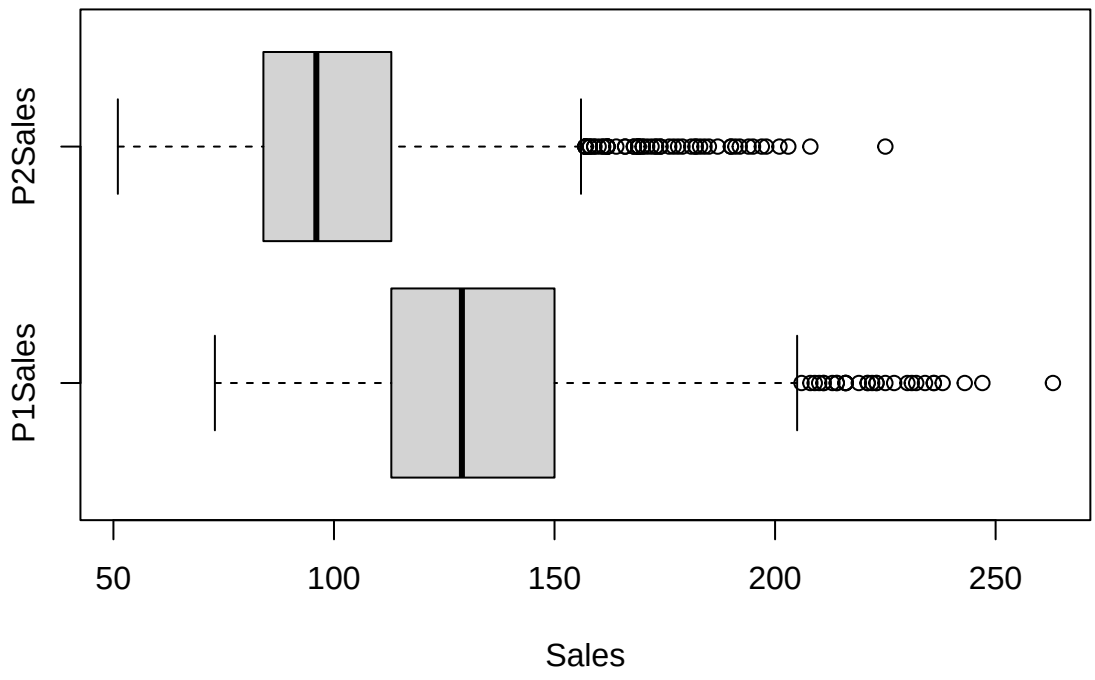
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
library(readr)
store <- read.csv("~/Graduate/ISE 535 Data Mining/store.csv")
#1.
summary(store)
```

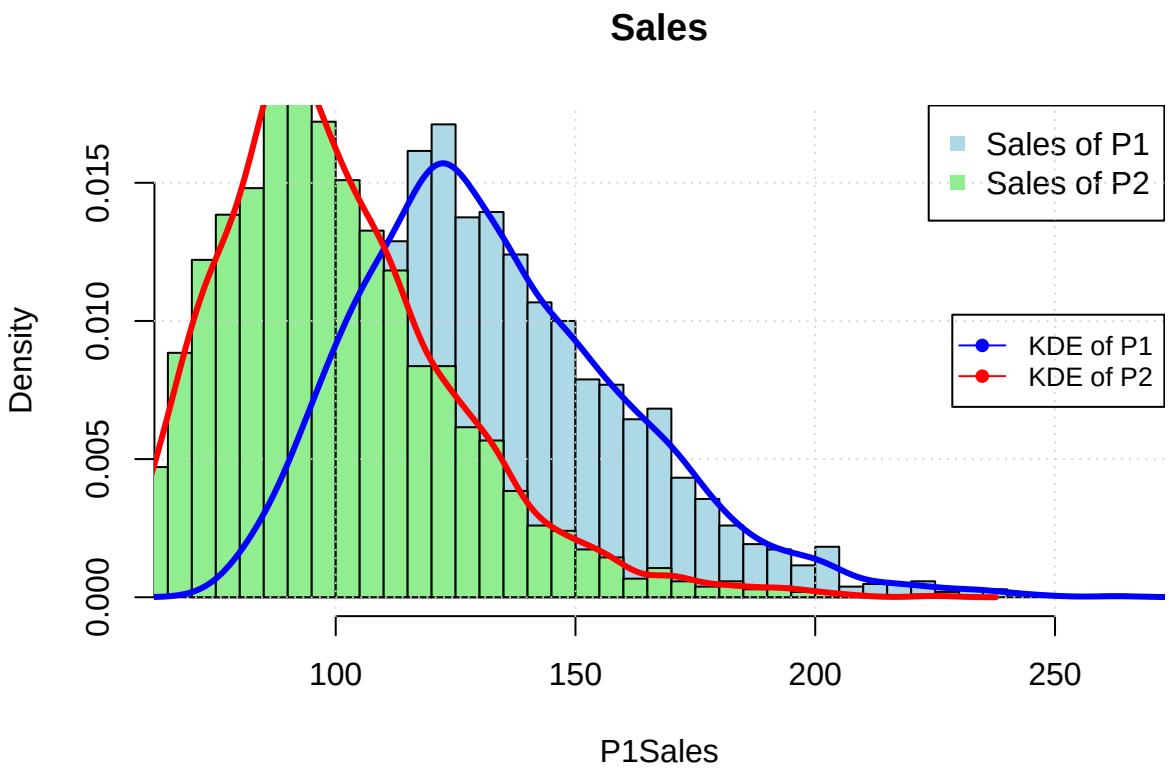
```
##     storeID           Year          Week          p1sales       p2sales
##  Min.   :101.0   Min.   :1.0   Min.   : 1.00   Min.   : 73   Min.   : 51.0
##  1st Qu.:105.8   1st Qu.:1.0   1st Qu.:13.75   1st Qu.:113   1st Qu.: 84.0
##  Median :110.5   Median :1.5   Median :26.50   Median :129   Median : 96.0
##  Mean   :110.5   Mean   :1.5   Mean   :26.50   Mean   :133   Mean   :100.2
##  3rd Qu.:115.2   3rd Qu.:2.0   3rd Qu.:39.25   3rd Qu.:150   3rd Qu.:113.0
##  Max.   :120.0   Max.   :2.0   Max.   :52.00   Max.   :263   Max.   :225.0
##     p1price         p2price         p1prom        p2prom
##  Min.   :2.190   Min.   :2.29   Min.   :0.0   Min.   :0.0000
##  1st Qu.:2.290   1st Qu.:2.49   1st Qu.:0.0   1st Qu.:0.0000
##  Median :2.490   Median :2.59   Median :0.0   Median :0.0000
##  Mean   :2.544   Mean   :2.70   Mean   :0.1   Mean   :0.1385
##  3rd Qu.:2.790   3rd Qu.:2.99   3rd Qu.:0.0   3rd Qu.:0.0000
##  Max.   :2.990   Max.   :3.19   Max.   :1.0   Max.   :1.0000
##    country
##  Length:2080
##  Class :character
##  Mode  :character
##
##
##
```

```
#The min of weekly sales of P2 is 51.0, median is 96.0, mean is 100.2, max sales is 225.0.
P1Sales = store$p1sales
P2Sales = store$p2sales
country = as.factor(store$country)
#Density histogram of weekly sales of P2
hist(P2Sales,freq = F, col="grey95",main="")
#Kernel Density Estimate overlapping the histogram
lines(density(P2Sales),col="red",lwd = 2)
```

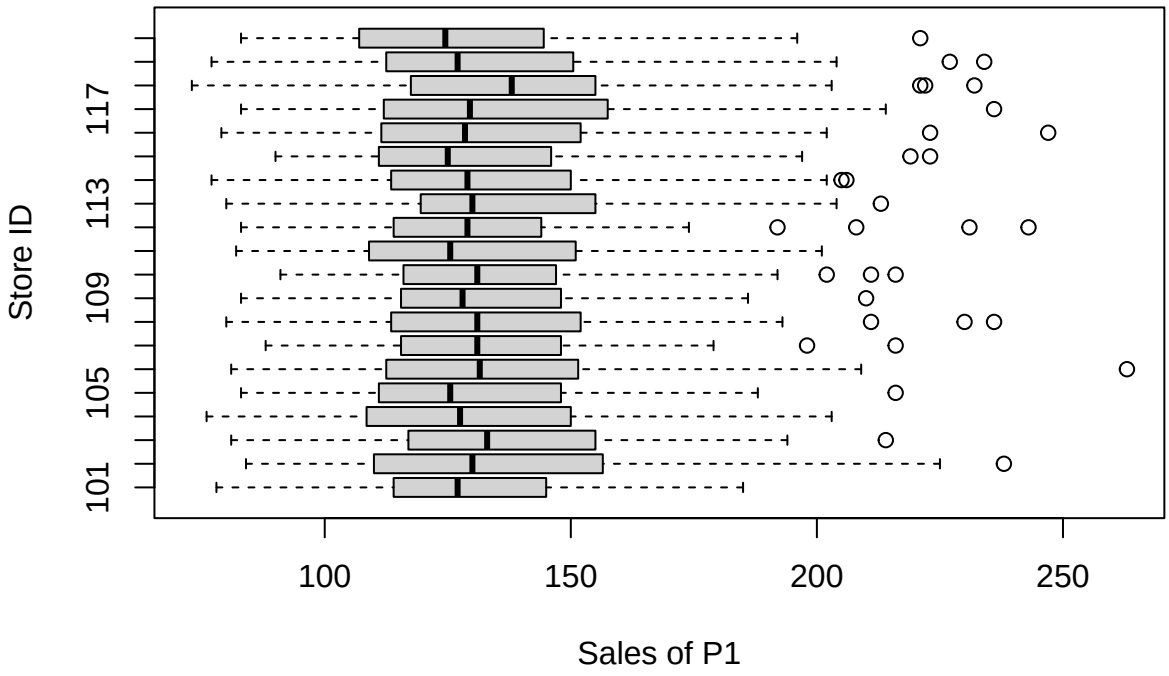

```
#2.
#Two Boxplots on the same chart
boxplot(list(P1Sales= P1Sales,P2Sales = P2Sales),horizontal = T,xlab = "Sales")
```




```
#Two Different-color overlapping histograms
hist(P1Sales,breaks = 30,freq = F,col="lightblue",main="Sales")
hist(P2Sales,breaks = 30,freq = F,col="lightgreen",main="",add=TRUE)
#add legend
label = c("Sales of P1","Sales of P2")
color = c("lightblue","lightgreen")
char = c(15,15)
legend("topright",label,pch = char, col = color)
grid()
# From the histograms, we know that the sale of Product 1 is more than that of Product 2.

#Two different-color overlapping Kernel density estimates
lines(density(P1Sales),col="blue",pch = 18, lwd = 3)
lines(density(P2Sales),col="red",pch = 19,lwd = 3)
label2 = c("KDE of P1","KDE of P2")
color2 = c("blue","red")
legend("right",label2,col = color2, lty = 1,pch = 19,cex = 0.8)
```




```
#3.
boxplot(P1Sales~store$storeID,horizontal = T,xlab = "Sales of P1",ylab = "Store ID")
```



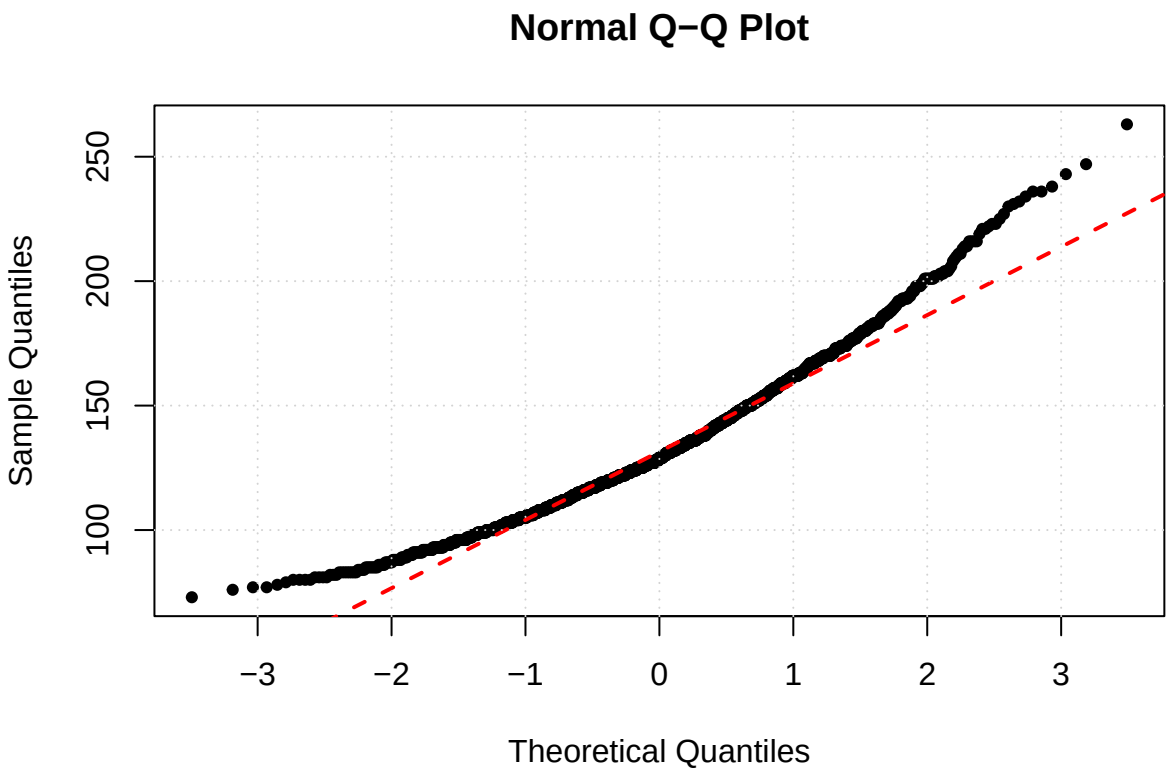

```
#Store 102 has the largest weekly sales of product P1 (escept for outliers).

#4.
library(e1071)
skewness(P1Sales) # 0.739,it is right-tailed.
```

```
## [1] 0.73935
```

```
kurtosis(P1Sales) #0.656501,it is >0, so it is heavy-tailed and thicker than the normal tail.
```

```
## [1] 0.656501
```

```
# Normal Q-Q plot
qqnorm(P1Sales,pch = 19, cex = 0.7)
qqline(P1Sales,lty = 2, col = "red",lwd = 2)
grid()
```



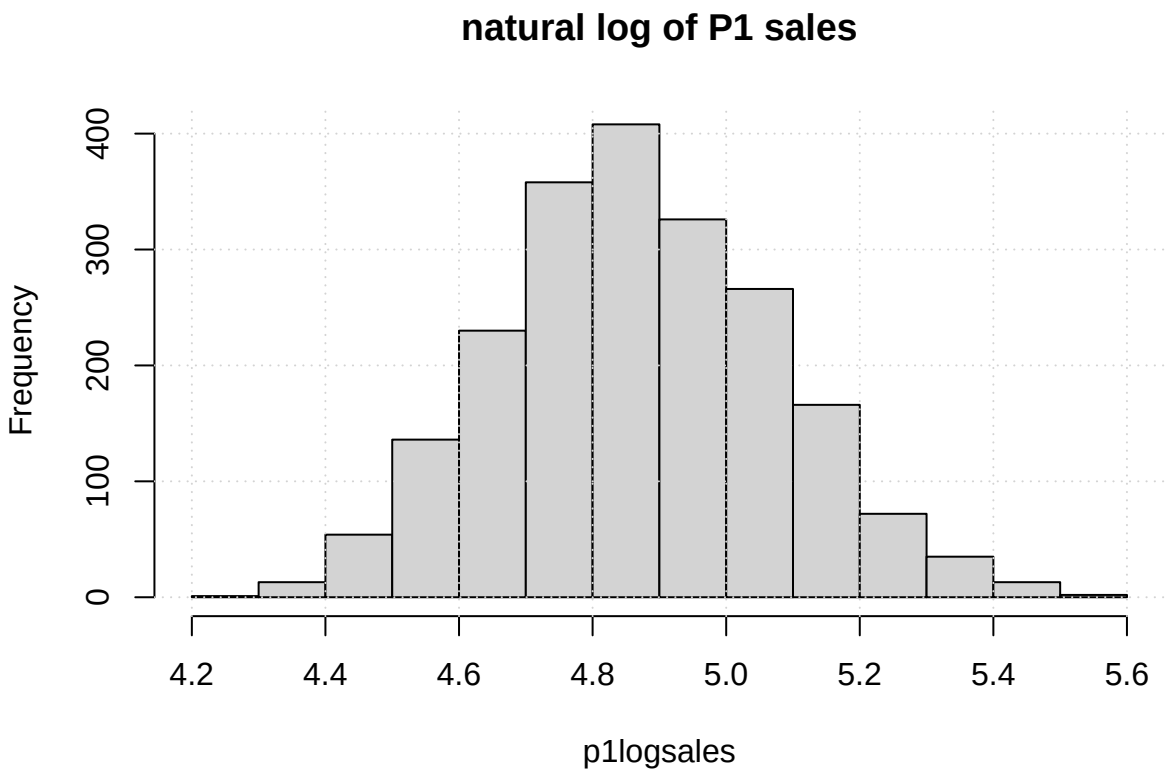

```
#natural log
p1logsales = log(P1Sales)
skewness(p1logsales) #0.16,roughly normal(symmetrical)shape
```

```
## [1] 0.1601015
```

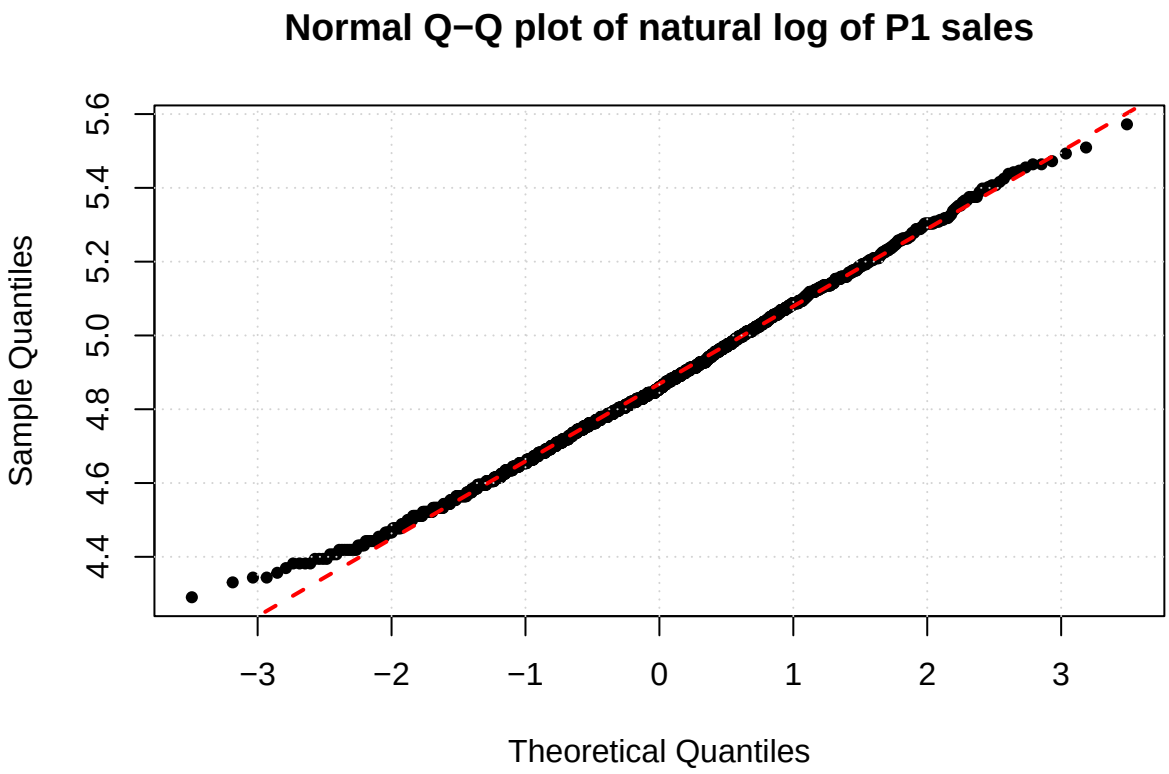
```
kurtosis(p1logsales) #-0.203, very cliffy
```

```
## [1] -0.2026867
```

```
hist(p1logsales,main="natural log of P1 sales")
grid()
```




```
#Q-Q plot of natual log of P1 sales.
qqnorm(p1logsales,pch = 19, cex = 0.7,main="Normal Q-Q plot of natural log of P1 sales")
qqline(p1logsales,lty = 2, col = "red",lwd = 2)
grid()
```




```
#The normal distribution very fit the log of P1 Sales.

#5(1)
p1sales_sum = aggregate(P1Sales~country,data = store,sum)
p1sales_sum
```

```
##   country P1Sales
## 1      AU   14544
## 2      BR   27836
## 3      CN   27381
## 4      DE   68876
## 5      GB   40986
## 6      JP   55381
## 7      US   41737
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.