Homework 1

Group 5: Phylisha Martinez, Carolina Munoz, Vikrant Nakod, Hareesh Rajendran, Piyusha Kulkarni

2/8/2020

```
library(data.table)
library(ggplot2)
library(reshape)
```

```
##
## Attaching package: 'reshape'
```

```
## The following object is masked from 'package:data.table':
##
##     melt
```

```
Utilities.df <- read.csv("Utilities.csv")
Utilities.dt <- setDT(Utilities.df)
summary(Utilities.dt)
```

```
##          Company    Fixed_charge        RoR             Cost
##  Arizona     : 1   Min.   :0.750   Min.   : 6.40   Min.   : 96.0
##  Boston      : 1   1st Qu.:1.042   1st Qu.: 9.20   1st Qu.:148.5
##  Central     : 1   Median :1.110   Median :11.05   Median :170.5
##  Commonwealth: 1   Mean   :1.114   Mean   :10.74   Mean   :168.2
##  Florida     : 1   3rd Qu.:1.190   3rd Qu.:12.35   3rd Qu.:195.8
##  Hawaiian    : 1   Max.   :1.490   Max.   :15.40   Max.   :252.0
##  (Other)     :16
##   Load_factor    Demand_growth        Sales          Nuclear
##  Min.   :49.80   Min.   :-2.200   Min.   : 3300   Min.   : 0.0
##  1st Qu.:53.77   1st Qu.: 1.450   1st Qu.: 6458   1st Qu.: 0.0
##  Median :56.35   Median : 3.000   Median : 8024   Median : 0.0
##  Mean   :56.98   Mean   : 3.241   Mean   : 8914   Mean   :12.0
##  3rd Qu.:60.30   3rd Qu.: 5.350   3rd Qu.:10128   3rd Qu.:24.6
##  Max.   :67.60   Max.   : 9.200   Max.   :17441   Max.   :50.2
##
##    Fuel_Cost
##  Min.   :0.309
##  1st Qu.:0.630
##  Median :0.960
##  Mean   :1.103
##  3rd Qu.:1.516
##  Max.   :2.116
##
```

```
names(Utilities.dt)
```

```
## [1] "Company"       "Fixed_charge"  "RoR"           "Cost"
## [5] "Load_factor"   "Demand_growth" "Sales"         "Nuclear"
## [9] "Fuel_Cost"
```

```
str(Utilities.dt)
```

```
## Classes 'data.table' and 'data.frame':   22 obs. of  9 variables:
##  $ Company      : Factor w/ 22 levels "Arizona ","Boston ",..: 1 2 3 4 13 5 6 7 8 9 ...
##  $ Fixed_charge : num  1.06 0.89 1.43 1.02 1.49 1.32 1.22 1.1 1.34 1.12 ...
##  $ RoR          : num  9.2 10.3 15.4 11.2 8.8 13.5 12.2 9.2 13 12.4 ...
##  $ Cost         : int  151 202 113 168 192 111 175 245 168 197 ...
##  $ Load_factor  : num  54.4 57.9 53 56 51.2 60 67.6 57 60.4 53 ...
##  $ Demand_growth: num  1.6 2.2 3.4 0.3 1 -2.2 2.2 3.3 7.2 2.7 ...
##  $ Sales        : int  9077 5088 9212 6423 3300 11127 7642 13082 8406 6455 ...
##  $ Nuclear      : num  0 25.3 0 34.3 15.6 22.5 0 0 0 39.2 ...
##  $ Fuel_Cost    : num  0.628 1.555 1.058 0.7 2.044 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
class(Utilities.dt)
```

```
## [1] "data.table" "data.frame"
```

Question 1

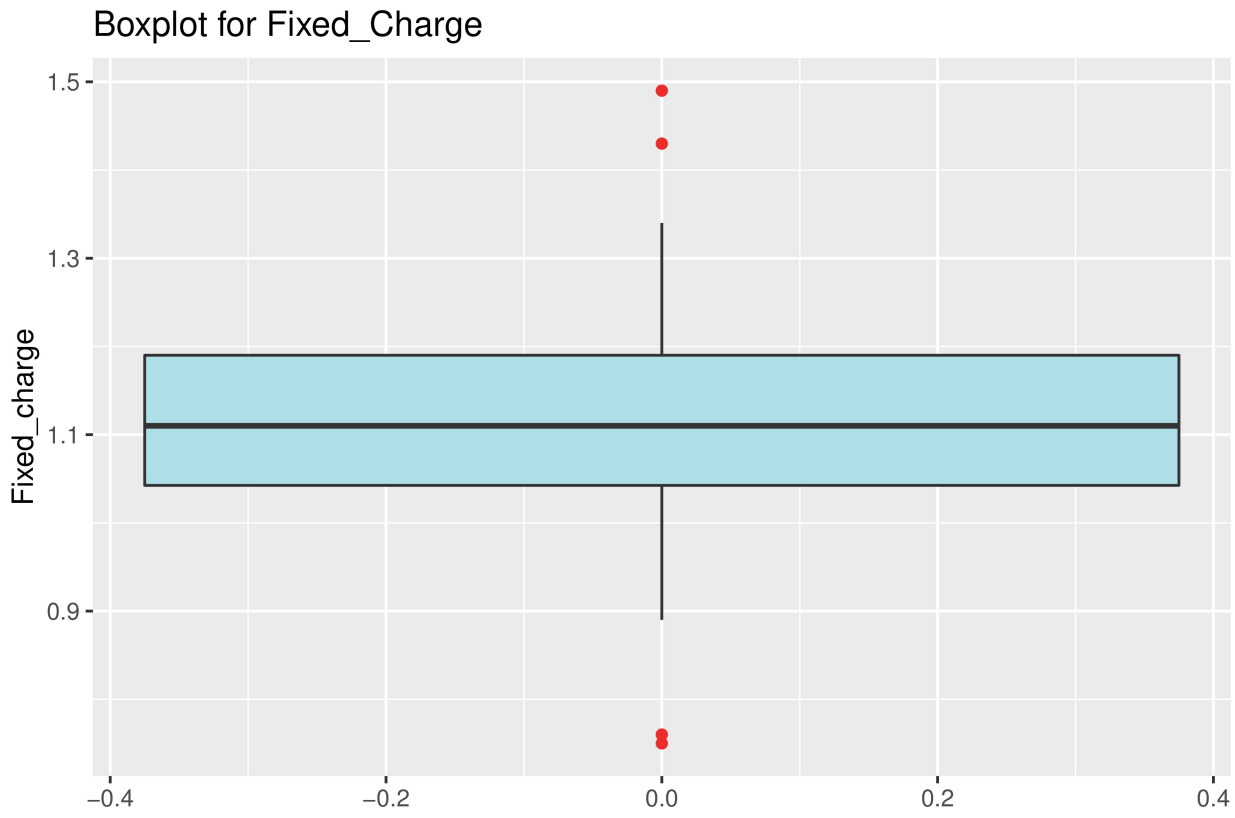
```
summary_stats <- function(x) {
   c(min = min(x), max = max(x), mean = mean(x), median = median(x), sd = sd(x))
}

Utilities.dt[,sapply(.SD, summary_stats), .SDcols = !"Company"]
```

```
##        Fixed_charge       RoR      Cost Load_factor Demand_growth     Sales
## min       0.7500000  6.400000  96.00000   49.800000     -2.200000  3300.000
## max       1.4900000 15.400000 252.00000   67.600000      9.200000 17441.000
## mean      1.1140909 10.736364 168.18182   56.977273      3.240909  8914.045
## median    1.1100000 11.050000 170.50000   56.350000      3.000000  8024.000
## sd        0.1845112  2.244049  41.19135    4.461148      3.118250  3549.984
##         Nuclear Fuel_Cost
## min     0.00000 0.3090000
## max    50.20000 2.1160000
## mean   12.00000 1.1027273
## median  0.00000 0.9600000
## sd     16.79192 0.5560981
```

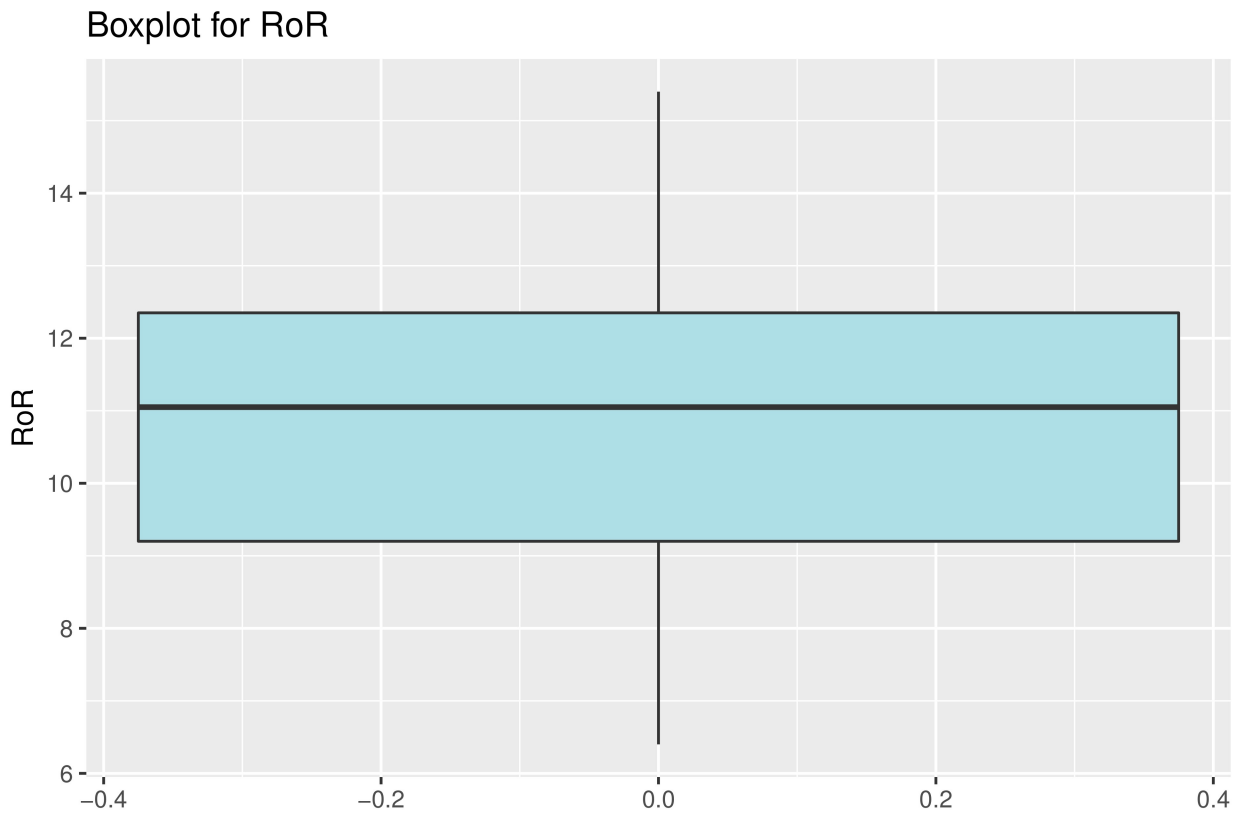
Answer 1- A good measure of variance in data is examining the standard deviation for each variable. The variable with the largest variablity is Sales.After comparing all standard deviations, it is evident sales has the largest standard deviation and consequently the largest variability.

Question 2

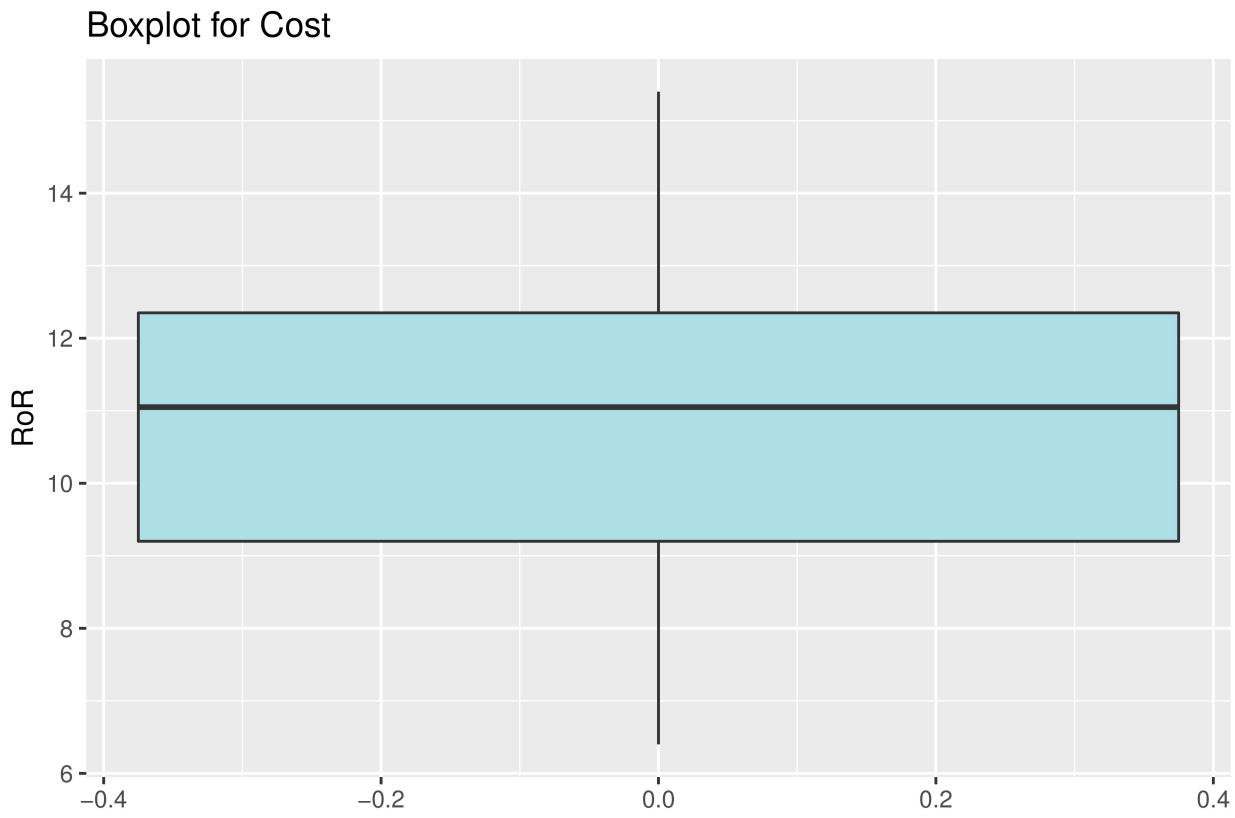
```
#Fixed_Charge
ggplot(Utilities.dt) +
  geom_boxplot(aes(y=Fixed_charge),
               fill = "powderblue", outlier.color = "firebrick2") +
  xlab("") + ggtitle("Boxplot for Fixed_Charge")
```



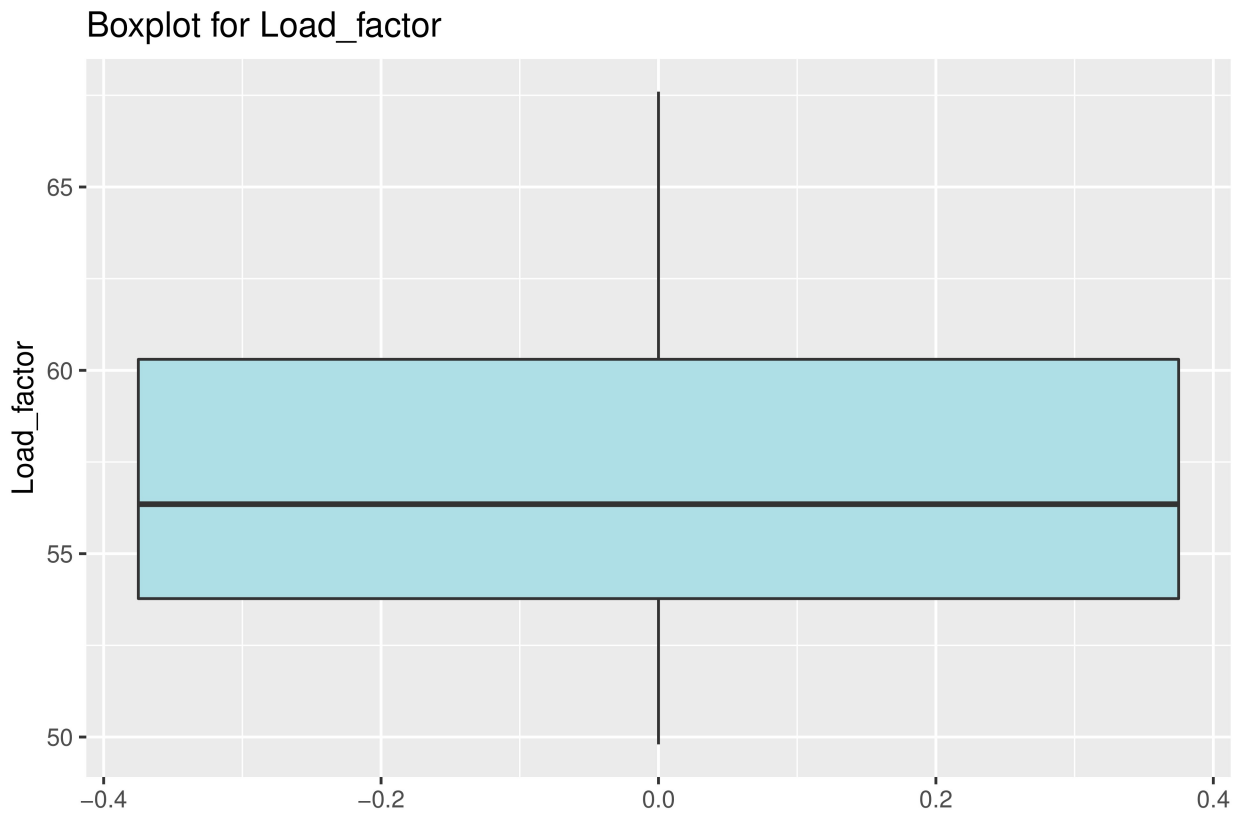

```
#RoR
ggplot(Utilities.dt) +
  geom_boxplot(aes(y=RoR),
               fill = "powderblue", outlier.color = "firebrick2") +
  xlab("") + ggtitle("Boxplot for RoR")
```



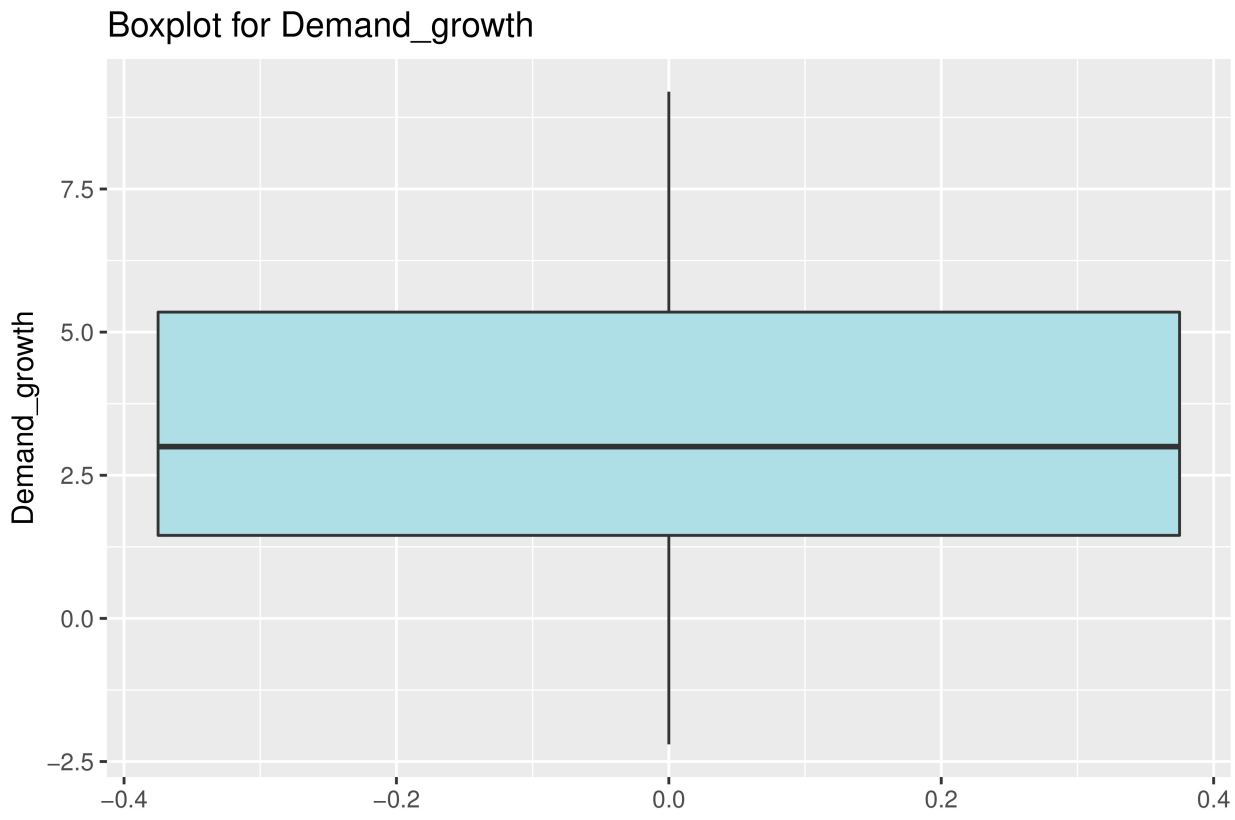

```
#Cost

ggplot(Utilities.dt) +
  geom_boxplot(aes(y=RoR),
               fill = "powderblue", outlier.color = "firebrick2") +
  xlab("") + ggtitle("Boxplot for Cost")
```



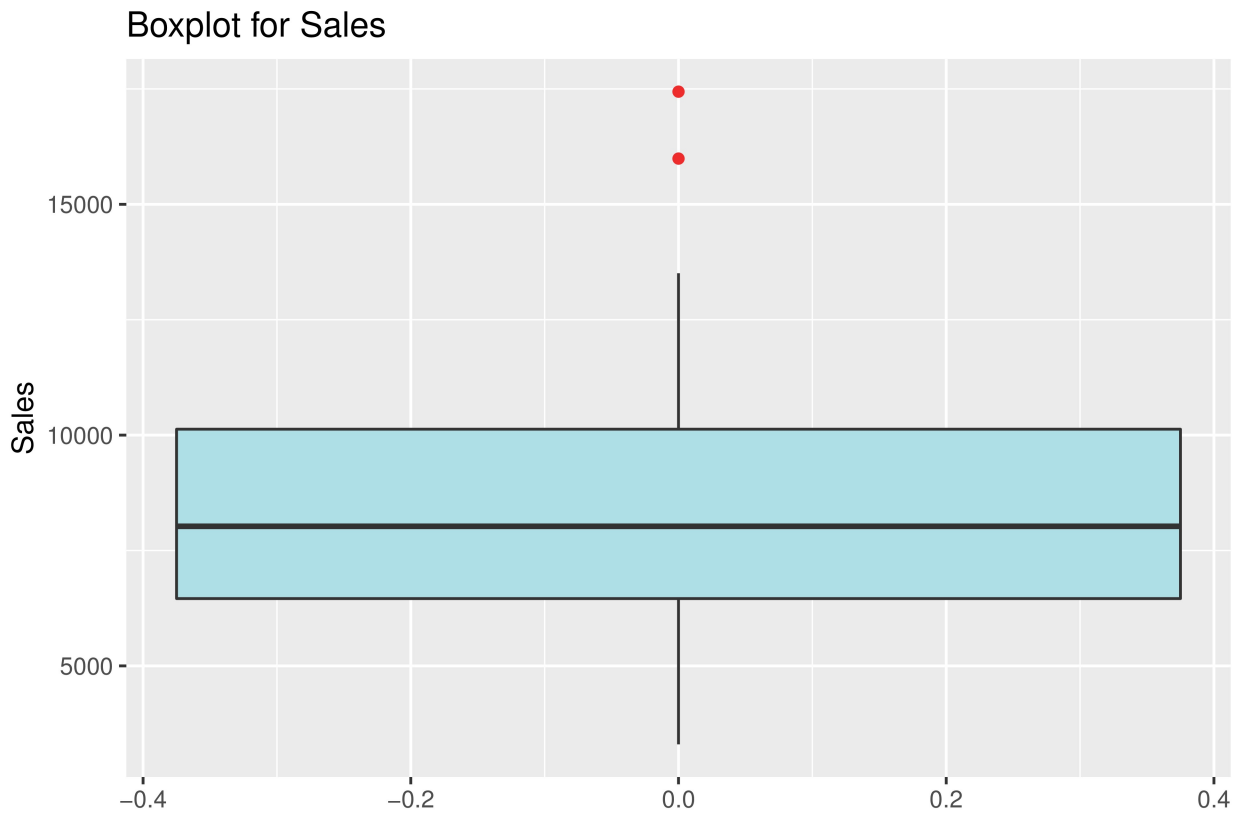

```
#Load_factor
ggplot(Utilities.dt) +
  geom_boxplot(aes(y=Load_factor),
               fill = "powderblue", outlier.color = "firebrick2") +
  xlab("") + ggtitle("Boxplot for Load_factor")
```



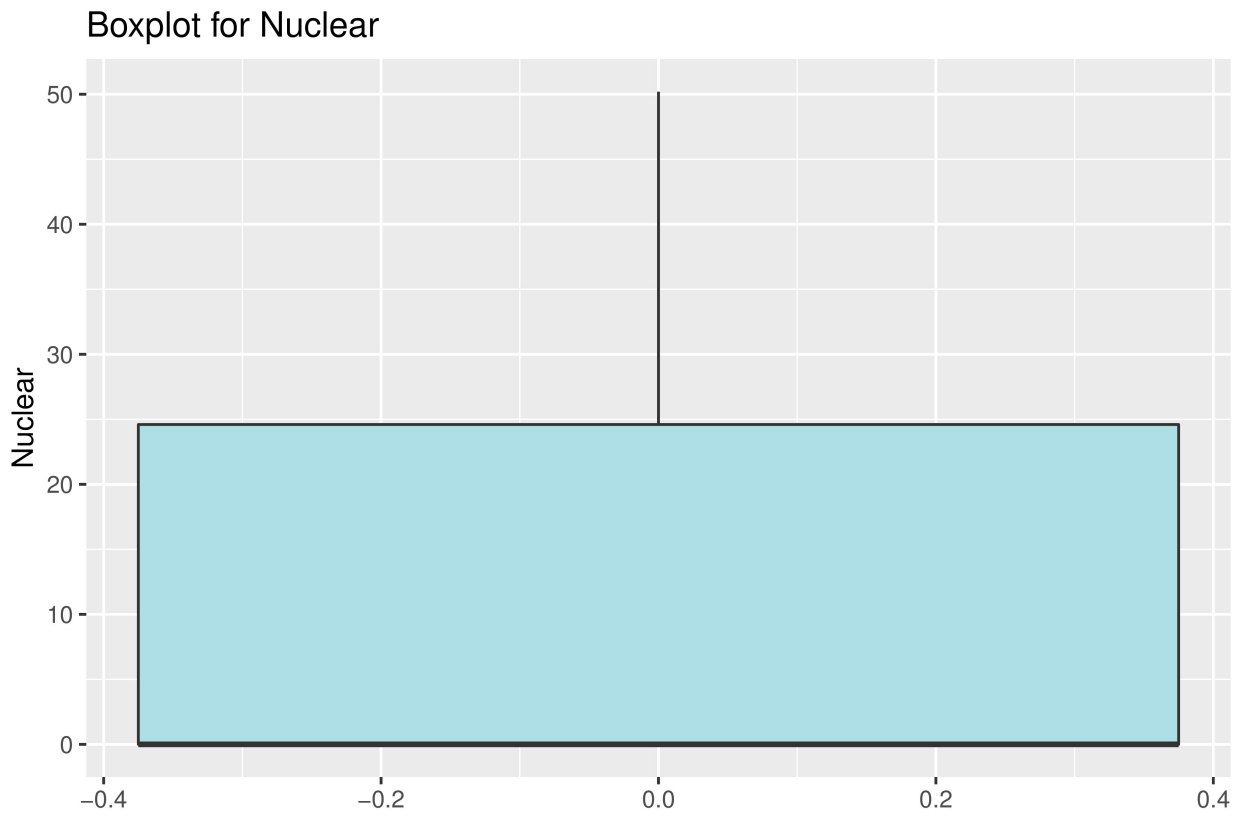

```
#Demand_growth
ggplot(Utilities.dt) +
  geom_boxplot(aes(y=Demand_growth),
               fill = "powderblue", outlier.color = "firebrick2") +
  xlab("") + ggtitle("Boxplot for Demand_growth")
```



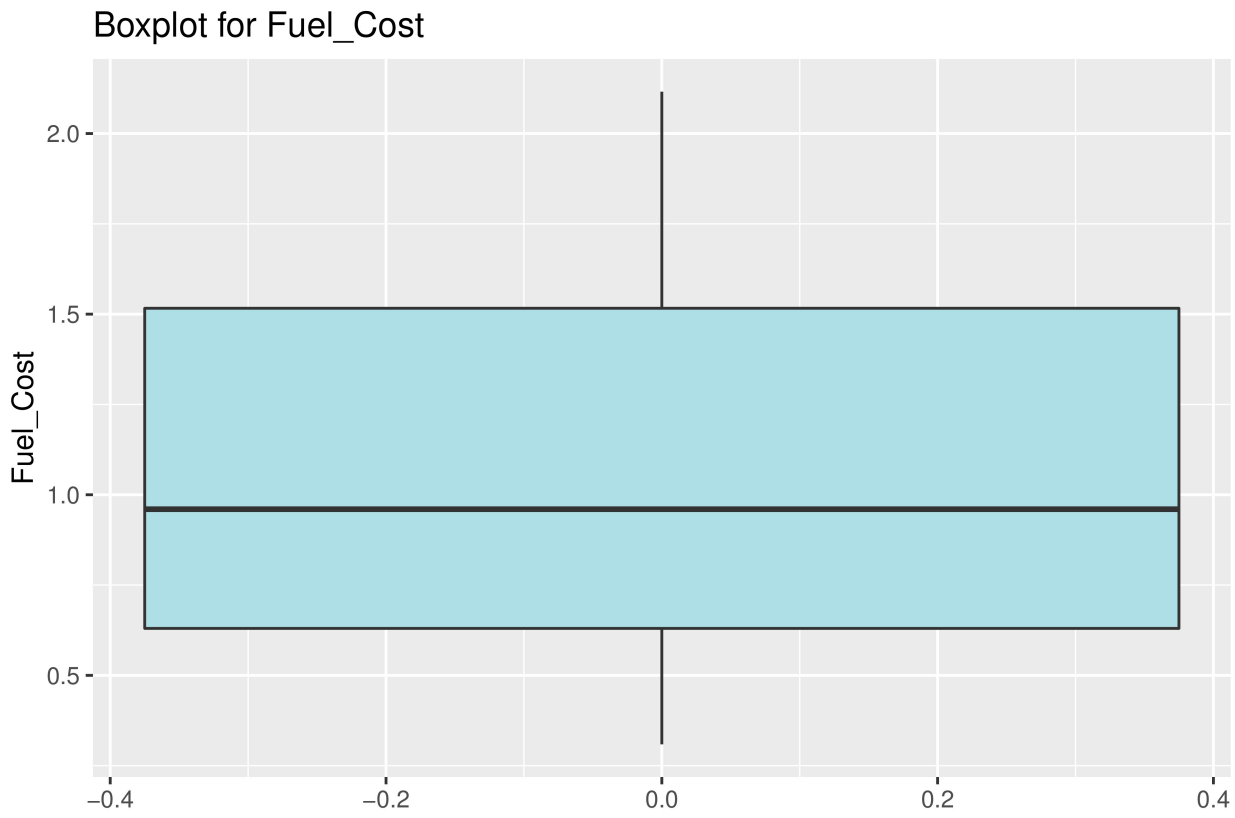

```
#Sales
ggplot(Utilities.dt) +
  geom_boxplot(aes(y=Sales),
               fill = "powderblue", outlier.color = "firebrick2") +
  xlab("") + ggtitle("Boxplot for Sales")
```




```
#Nuclear
ggplot(Utilities.dt) +
  geom_boxplot(aes(y=Nuclear),
               fill = "powderblue", outlier.color = "firebrick2") +
  xlab("") + ggtitle("Boxplot for Nuclear")
```




```
#Fuel_Cost
ggplot(Utilities.dt) +
  geom_boxplot(aes(y=Fuel_Cost),
               fill = "powderblue", outlier.color = "firebrick2") +
  xlab("") + ggtitle("Boxplot for Fuel_Cost")
```



Answer 2- An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.Following is a statistical definition of an outlier for X: X < Q1 - 1.5 IQR or X # Q3 + 1.5 IQR where IQR is the difference between the 75th and 25th percentiles. Yes, there are extreme values for two variables.They are Fixed_charge and Sales. The boxplot for Fixed_charge,has a total of 4 outliers, 2 on the higher side and 2 on the lower side. In the boxplot for Sales,the 2 outliers are only on the higher side.

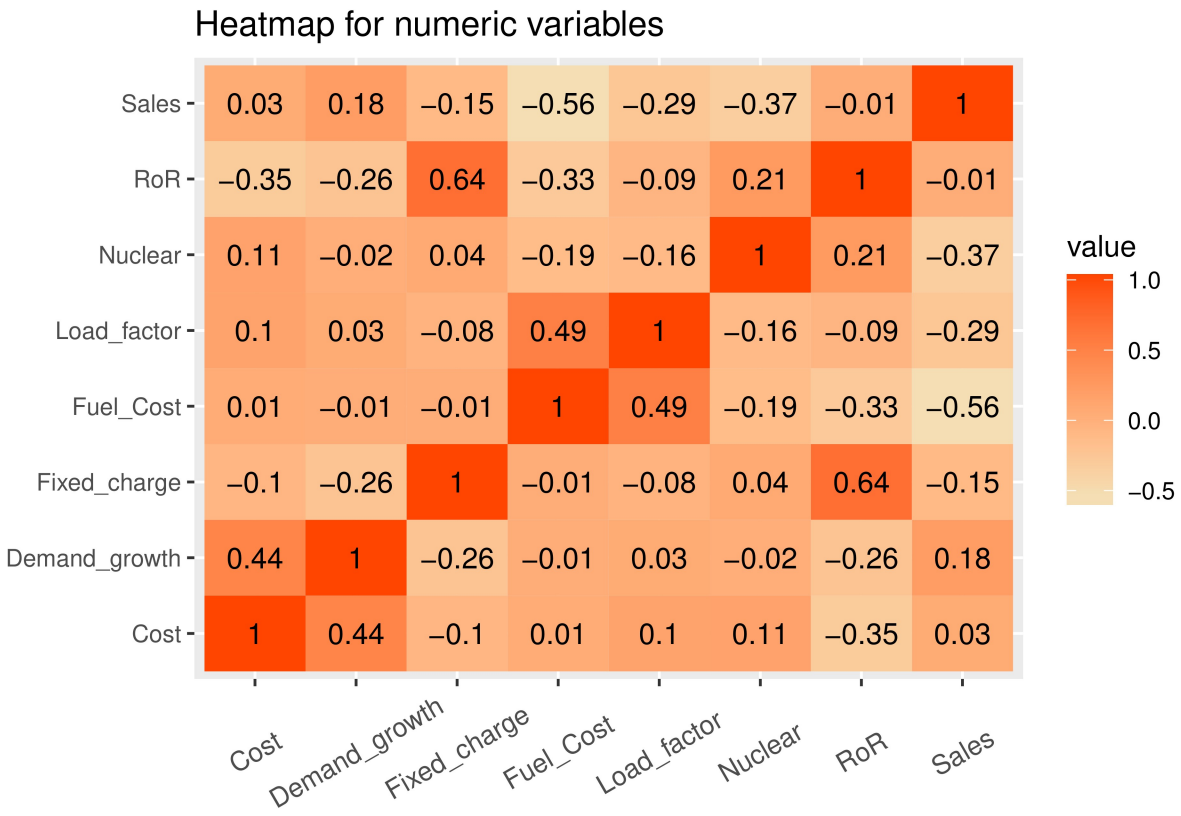
```
cor.mat <- round(cor(Utilities.dt[,!c("Company")]),2) # rounded correlation matrix
melted.cor.mat <- melt(cor.mat)
melted.cor.mat
```

```
##                X1            X2 value
## 1   Fixed_charge  Fixed_charge  1.00
## 2             RoR  Fixed_charge  0.64
## 3            Cost  Fixed_charge -0.10
## 4     Load_factor  Fixed_charge -0.08
## 5  Demand_growth  Fixed_charge -0.26
## 6           Sales  Fixed_charge -0.15
## 7         Nuclear  Fixed_charge  0.04
## 8       Fuel_Cost  Fixed_charge -0.01
## 9   Fixed_charge           RoR  0.64
## 10            RoR           RoR  1.00
## 11           Cost           RoR -0.35
## 12   Load_factor           RoR -0.09
## 13 Demand_growth           RoR -0.26
## 14          Sales           RoR -0.01
## 15        Nuclear           RoR  0.21
```

```
## 16     Fuel_Cost           RoR -0.33
## 17  Fixed_charge          Cost -0.10
## 18           RoR          Cost -0.35
## 19          Cost          Cost  1.00
## 20   Load_factor          Cost  0.10
## 21 Demand_growth          Cost  0.44
## 22         Sales          Cost  0.03
## 23       Nuclear          Cost  0.11
## 24     Fuel_Cost          Cost  0.01
## 25  Fixed_charge   Load_factor -0.08
## 26           RoR   Load_factor -0.09
## 27          Cost   Load_factor  0.10
## 28   Load_factor   Load_factor  1.00
## 29 Demand_growth   Load_factor  0.03
## 30         Sales   Load_factor -0.29
## 31       Nuclear   Load_factor -0.16
## 32     Fuel_Cost   Load_factor  0.49
## 33  Fixed_charge Demand_growth -0.26
## 34           RoR Demand_growth -0.26
## 35          Cost Demand_growth  0.44
## 36   Load_factor Demand_growth  0.03
## 37 Demand_growth Demand_growth  1.00
## 38         Sales Demand_growth  0.18
## 39       Nuclear Demand_growth -0.02
## 40     Fuel_Cost Demand_growth -0.01
## 41  Fixed_charge         Sales -0.15
## 42           RoR         Sales -0.01
## 43          Cost         Sales  0.03
## 44   Load_factor         Sales -0.29
## 45 Demand_growth         Sales  0.18
## 46         Sales         Sales  1.00
## 47       Nuclear         Sales -0.37
## 48     Fuel_Cost         Sales -0.56
## 49  Fixed_charge       Nuclear  0.04
## 50           RoR       Nuclear  0.21
## 51          Cost       Nuclear  0.11
## 52   Load_factor       Nuclear -0.16
## 53 Demand_growth       Nuclear -0.02
## 54         Sales       Nuclear -0.37
## 55       Nuclear       Nuclear  1.00
## 56     Fuel_Cost       Nuclear -0.19
## 57  Fixed_charge     Fuel_Cost -0.01
## 58           RoR     Fuel_Cost -0.33
## 59          Cost     Fuel_Cost  0.01
## 60   Load_factor     Fuel_Cost  0.49
## 61 Demand_growth     Fuel_Cost -0.01
## 62         Sales     Fuel_Cost -0.56
## 63       Nuclear     Fuel_Cost -0.19
## 64     Fuel_Cost     Fuel_Cost  1.00
```

```
ggplot(melted.cor.mat, aes(x = X1, y = X2, fill = value)) +
  scale_fill_gradient(low="wheat", high="orangered") +
  geom_tile() +
  geom_text(aes(x = X1, y = X2, label = value)) +
```

```
  ggtitle("Heatmap for numeric variables")+xlab("")+ylab("")+
  theme(axis.text.x =element_text(angle=30,size=10, vjust =0.5))
```



Answer 3- In the above heatmap, the variables ROR and Fixed_charge have the highest correlation which is 0.64. Hence, we can say that the Rate on Return increases or decreases when the fixed_charge of the company increases or decreases. Also, the variables Sales and Fuel_Cost have the strongest negative correlation which is -0.56. Hence we can say that as the Fuel_cost of a company increases the sales go down and vice versa.

```
pca_noscale <- prcomp(na.omit(Utilities.dt[,-1]))
summary(pca_noscale)
```

```
## Importance of components:
##                             PC1      PC2      PC3   PC4   PC5   PC6    PC7
## Standard deviation     3549.9901 41.26913 15.49215 4.001 2.783 1.977 0.3501
## Proportion of Variance    0.9998  0.00014  0.00002 0.000 0.000 0.000 0.0000
## Cumulative Proportion     0.9998  0.99998  1.00000 1.000 1.000 1.000 1.0000
##                            PC8
## Standard deviation     0.1224
## Proportion of Variance 0.0000
## Cumulative Proportion  1.0000
```

```
pca_noscale$rotation
```

```
##                          PC1           PC2           PC3           PC4
```

```
## Fixed_charge   7.883140e-06 -0.0004460932  0.0001146357 -0.0057978329
## RoR            6.081397e-06 -0.0186257078  0.0412535878  0.0292444838
## Cost          -3.247724e-04  0.9974928360 -0.0566502956 -0.0179103135
## Load_factor    3.618357e-04  0.0111104272 -0.0964680806  0.9930009368
## Demand_growth -1.549616e-04  0.0326730808 -0.0038575008  0.0544730799
## Sales         -9.999983e-01 -0.0002209801  0.0017377455  0.0005270008
## Nuclear        1.767632e-03  0.0589056695  0.9927317841  0.0949073699
## Fuel_Cost      8.780470e-05  0.0001659524 -0.0157634569  0.0276496391
##                         PC5           PC6           PC7           PC8
## Fixed_charge   0.0198566131 -0.0583722527 -1.002990e-01  9.930280e-01
## RoR            0.2028309717 -0.9735822744 -5.984233e-02 -6.717166e-02
## Cost           0.0355836487 -0.0144563569 -9.986723e-04 -1.312104e-03
## Load_factor    0.0495177973  0.0333700701  2.930752e-02  9.745357e-03
## Demand_growth -0.9768581322 -0.2038187556  8.898790e-03  8.784363e-03
## Sales          0.0001471164  0.0001237088 -9.721241e-05  5.226863e-06
## Nuclear       -0.0057261758  0.0430954352 -1.043775e-02  2.059461e-03
## Fuel_Cost     -0.0215054038  0.0633116915 -9.926283e-01 -9.594372e-02
```

Answer 4- Here, we have performed PCA without scaling the variables. While some insight can possibly be drawn from it, none of the results could be considered 'significant' given the wide range of values between each variable. PCA requires the set of input variables to have similar scales of measurement. As we have not scaled the variables in the above model each variable has a different unit which resulted in 99.98 variance in the Principal component 1.For instance, the variables ROR and Sales may have different units and we cannot compare these. Therefore, in order to find the correct results we need to scale these variables and then perform the PCA.

```
pca_scaled <- prcomp(Utilities.dt[,!c("Company")],scale. = T)
summary(pca_scaled)
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     1.4741 1.3785 1.1504 0.9984 0.80562 0.75608 0.46530
## Proportion of Variance 0.2716 0.2375 0.1654 0.1246 0.08113 0.07146 0.02706
## Cumulative Proportion  0.2716 0.5091 0.6746 0.7992 0.88031 0.95176 0.97883
##                           PC8
## Standard deviation     0.41157
## Proportion of Variance 0.02117
## Cumulative Proportion  1.00000
```

```
pca_scaled$rotation
```

```
##                     PC1         PC2         PC3         PC4        PC5
## Fixed_charge   0.44554526 -0.23217669  0.06712849 -0.55549758  0.4008403
## RoR            0.57119021 -0.10053490  0.07123367 -0.33209594 -0.3359424
## Cost          -0.34869054  0.16130192  0.46733094 -0.40908380  0.2685680
## Load_factor   -0.28890116 -0.40918419 -0.14259793 -0.33373941 -0.6800711
## Demand_growth -0.35536100  0.28293270  0.28146360 -0.39139699 -0.1626375
## Sales          0.05383343  0.60309487 -0.33199086 -0.19086550 -0.1319721
## Nuclear        0.16797023 -0.08536118  0.73768406  0.33348714 -0.2496462
## Fuel_Cost     -0.33584032 -0.53988503 -0.13442354 -0.03960132  0.2926660
##                       PC6         PC7         PC8
## Fixed_charge  -0.00654016  0.20578234 -0.48107955
```

```
## RoR           -0.13326000 -0.15026737  0.62855128
## Cost           0.53750238 -0.11762875  0.30294347
## Load_factor    0.29890373  0.06429342 -0.24781930
## Demand_growth -0.71916993 -0.05155339 -0.12223012
## Sales          0.14953365  0.66050223  0.10339649
## Nuclear        0.02644086  0.48879175 -0.08466572
## Fuel_Cost     -0.25235278  0.48914707  0.43300956
```

Answer 5- The interpretation did change from the unscaled PCS. Because, unlike on the unscaled where a single PCS variable will give you 99.9% information, we need to have 6 PCS variables to get 95% information. Unlike the unscaled PCS1 value where sales has the high influence, the scaled PCS1 value have influence of ROR. Because we consider only the absolute value. It is because each column has a different units and PCS is created by joining all the information from every column of the original table. That's why when we scale it and generate the PCS, we get a more reliable Principal Component Value.