Homework 1

Group BUAN635.501-1

2/8/2020

CLASS: "BUAN 6356"
GROUP MEMBERS: "Sai Raghavendra Sridhar(sxs180281), Shreya Tippannawar(sst190000), Smruti Viswanath Iyer(sxi180001), Piyush Dangwal(pxd142430),Shanshan Luo(sxl130330)"

Solutions :

a. Load the package "data.table":

```
if(!require("pacman")) install.packages("pacman")
```

```
## Loading required package: pacman
```

```
pacman::p_load(tidyverse, gplots, GGally, tinytex, data.table, reshape, knitr)
```

```
## Installing package into 'C:/Users/saira/Documents/R/win-library/3.6'
## (as 'lib' is unspecified)

## Warning: unable to access index for repository http://www.stats.ox.ac.uk/pub/RWin/bin/windows/contril
##   cannot open URL 'http://www.stats.ox.ac.uk/pub/RWin/bin/windows/contrib/3.6/PACKAGES'

## package 'tidyverse' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\saira\AppData\Local\Temp\RtmpEd3D1w\downloaded_packages

##
## tidyverse installed

## Warning in pacman::p_load(tidyverse, gplots, GGally, tinytex, data.table, : Failed to install/load:
## tidyverse
```

```
search()
```

```
##  [1] ".GlobalEnv"         "package:knitr"      "package:reshape"
##  [4] "package:data.table" "package:tinytex"    "package:GGally"
##  [7] "package:ggplot2"    "package:gplots"     "package:pacman"
## [10] "package:stats"      "package:graphics"   "package:grDevices"
## [13] "package:utils"      "package:datasets"   "package:methods"
## [16] "Autoloads"          "package:base"
```

b. Read in the data from "Utilities":

```
Utilities.dt <- read.csv("Utilities.csv")
```

Question 1: Compute the minimum, maximum, mean, median, and standard deviation for each of the numeric variables using data.table package. Which variable(s) has the largest variability? Explain your answer

```
Fixed_charge_vector <- c(min(Utilities.dt$Fixed_charge),max(Utilities.dt$Fixed_charge),
                         mean(Utilities.dt$Fixed_charge),median(Utilities.dt$Fixed_charge),
                         sd(Utilities.dt$Fixed_charge))

RoR_vector <- c(min(Utilities.dt$RoR), max(Utilities.dt$RoR),
                mean(Utilities.dt$RoR),median(Utilities.dt$RoR),
                sd(Utilities.dt$RoR))

Cost_vector <- c(min(Utilities.dt$Cost),max(Utilities.dt$Cost),
                 mean(Utilities.dt$Cost),median(Utilities.dt$Cost),
                 sd(Utilities.dt$Cost))

Load_factor_vector <- c(min(Utilities.dt$Load_factor), max(Utilities.dt$Load_factor),
                        mean(Utilities.dt$Load_factor),median(Utilities.dt$Load_factor),
                        sd(Utilities.dt$Load_factor))

Demand_growth_vector <- c(min(Utilities.dt$Demand_growth),max(Utilities.dt$Demand_growth),
                          mean(Utilities.dt$Demand_growth),median(Utilities.dt$Demand_growth),
                          sd(Utilities.dt$Demand_growth))

Sales_vector <- c(min(Utilities.dt$Sales), max(Utilities.dt$Sales),
                  mean(Utilities.dt$Sales),median(Utilities.dt$Sales),
                  sd(Utilities.dt$Sales))

Nuclear_vector <- c(min(Utilities.dt$Nuclear),max(Utilities.dt$Nuclear),
                    mean(Utilities.dt$Nuclear),median(Utilities.dt$Nuclear),
                    sd(Utilities.dt$Nuclear))

Fuel_cost_vector <- c(min(Utilities.dt$Fuel_Cost),max(Utilities.dt$Fuel_Cost),
                      mean(Utilities.dt$Fuel_Cost),median(Utilities.dt$Fuel_Cost),
                      sd(Utilities.dt$Fuel_Cost))

Comparison_df <- data.frame(Fixed_charge_vector,RoR_vector,Cost_vector,Load_factor_vector,
                            Demand_growth_vector,Sales_vector, Nuclear_vector,
                            Fuel_cost_vector )

row.names(Comparison_df) <- c("Minimum","Maximum","Mean","Median","Standard Deviation")

### Comparison table for Utilities variable

Comparison_df
```

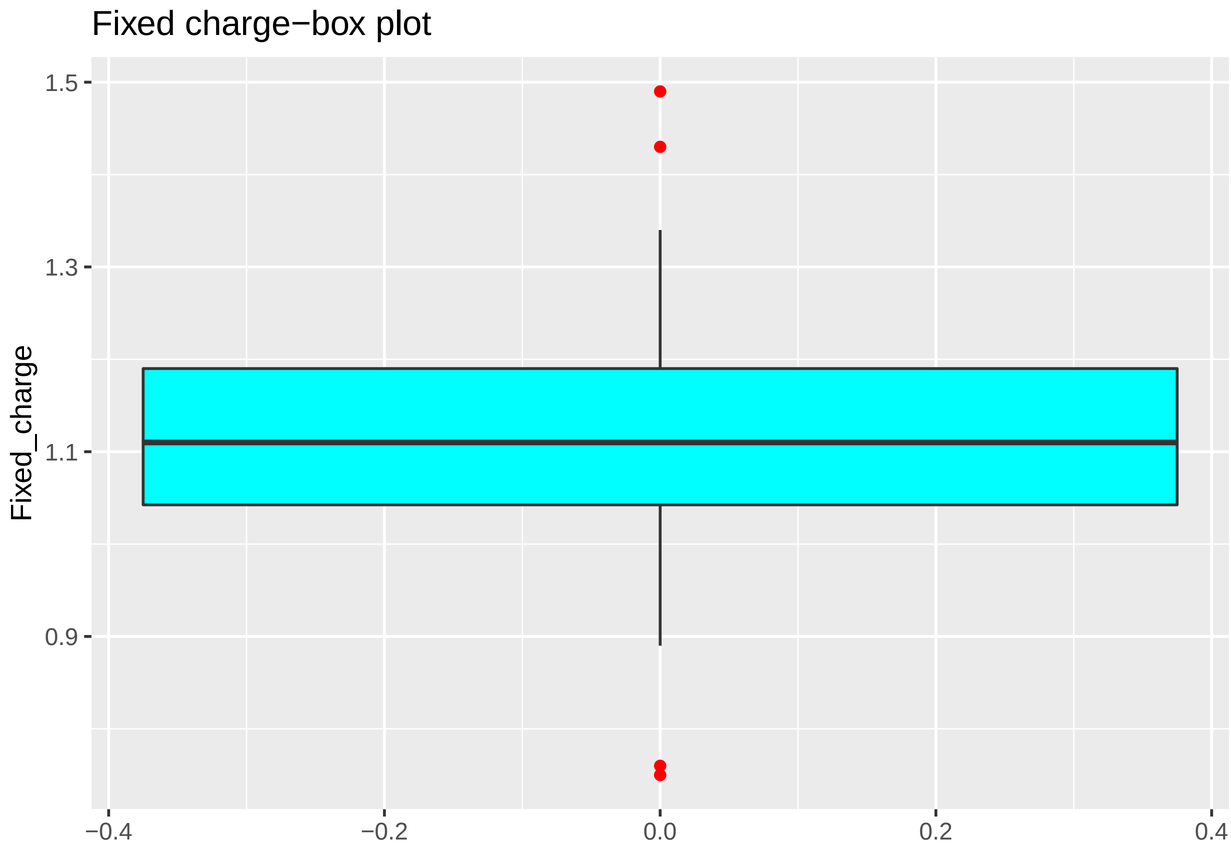
```
##                     Fixed_charge_vector RoR_vector Cost_vector
```

```
## Minimum                      0.7500000   6.400000    96.00000
## Maximum                      1.4900000  15.400000   252.00000
## Mean                         1.1140909  10.736364   168.18182
## Median                       1.1100000  11.050000   170.50000
## Standard Deviation           0.1845112   2.244049    41.19135
##                    Load_factor_vector Demand_growth_vector Sales_vector
## Minimum                     49.800000            -2.200000     3300.000
## Maximum                     67.600000             9.200000    17441.000
## Mean                        56.977273             3.240909     8914.045
## Median                      56.350000             3.000000     8024.000
## Standard Deviation           4.461148             3.118250     3549.984
##                    Nuclear_vector Fuel_cost_vector
## Minimum                   0.00000        0.3090000
## Maximum                  50.20000        2.1160000
## Mean                     12.00000        1.1027273
## Median                    0.00000        0.9600000
## Standard Deviation       16.79192        0.5560981
```

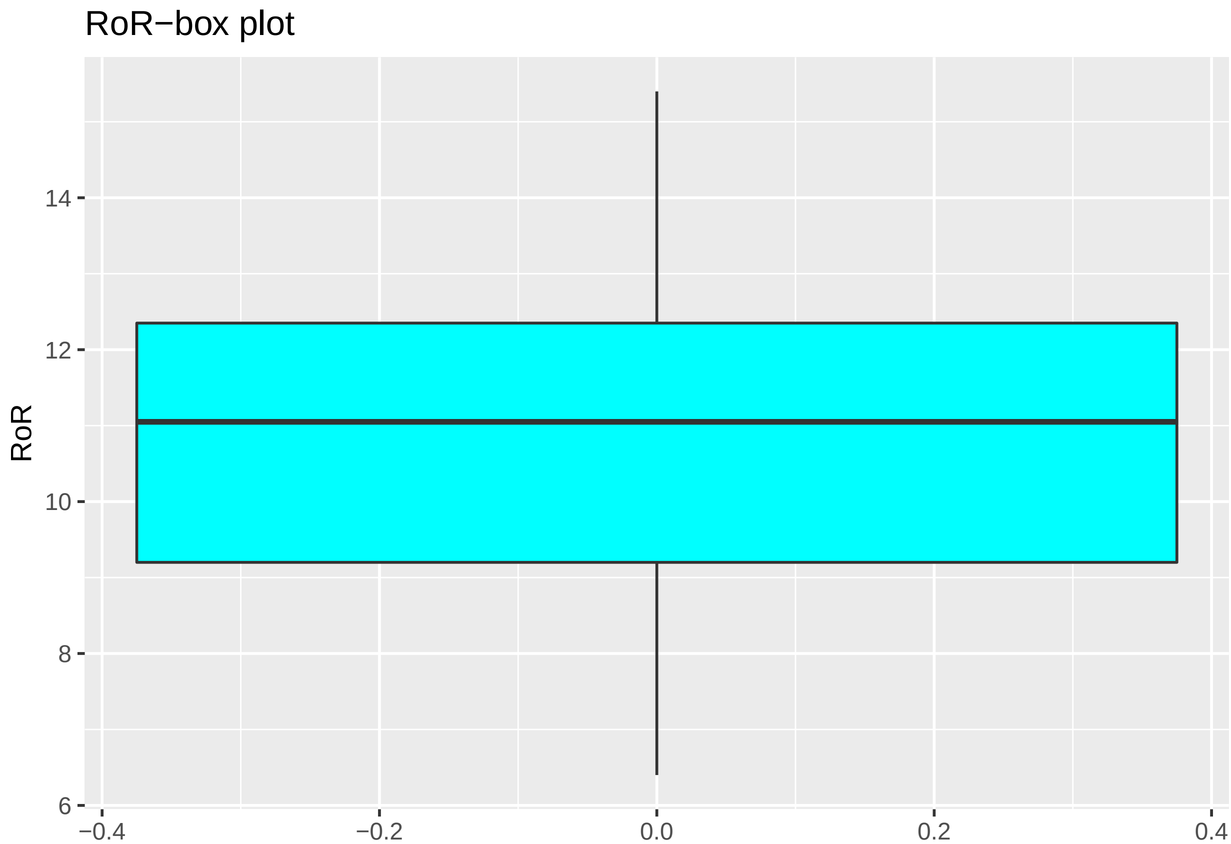
Answer 1: From above comparison dataframe we get to know that sales_vector has highest SD.From that we can infer that sales_vector has largest variability.

Question 2: Create boxplots for each of the numeric variables. Are there any extreme values for any of the variables? Which ones? Explain your answer

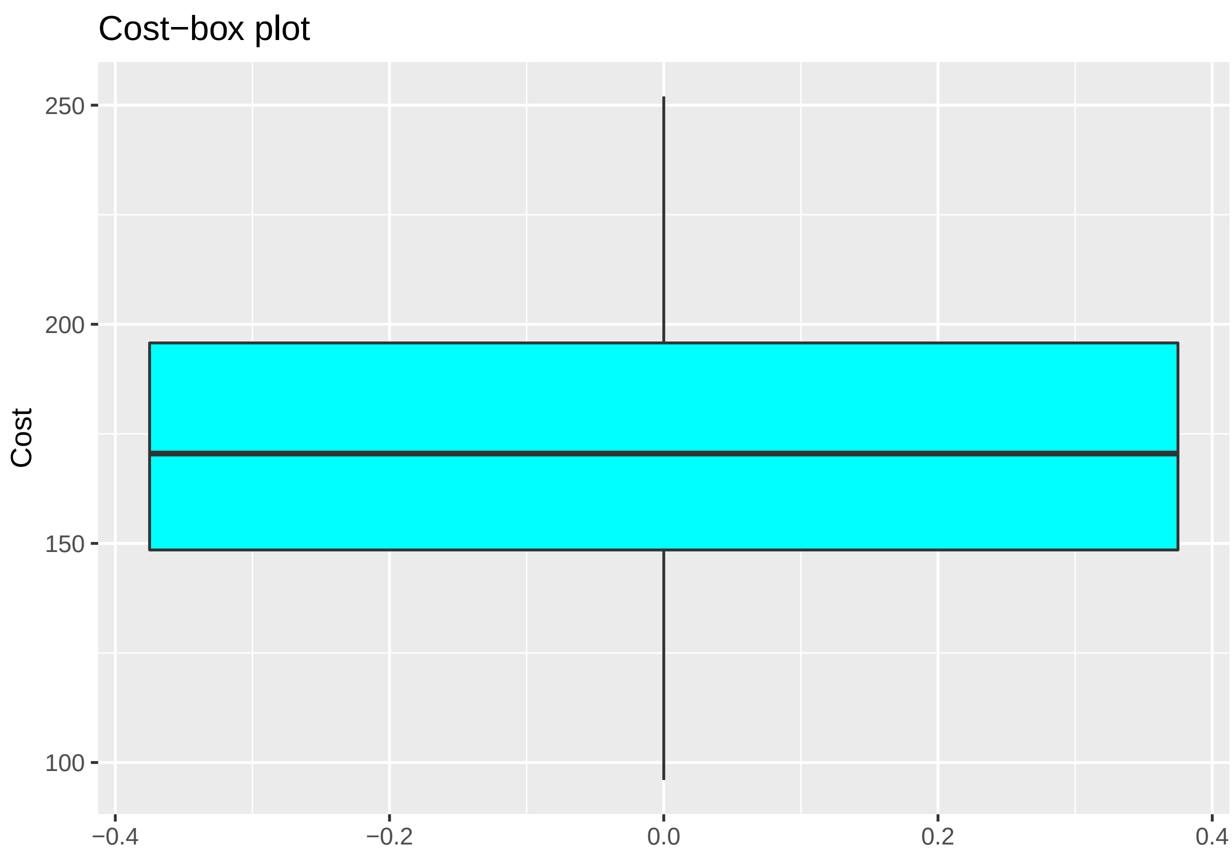
```r
###Fixed_charge_Box plot
ggplot(Utilities.dt) +
  geom_boxplot(aes(y= Fixed_charge),fill = "cyan", outlier.color = "red")+
  ggtitle("Fixed charge-box plot")
```



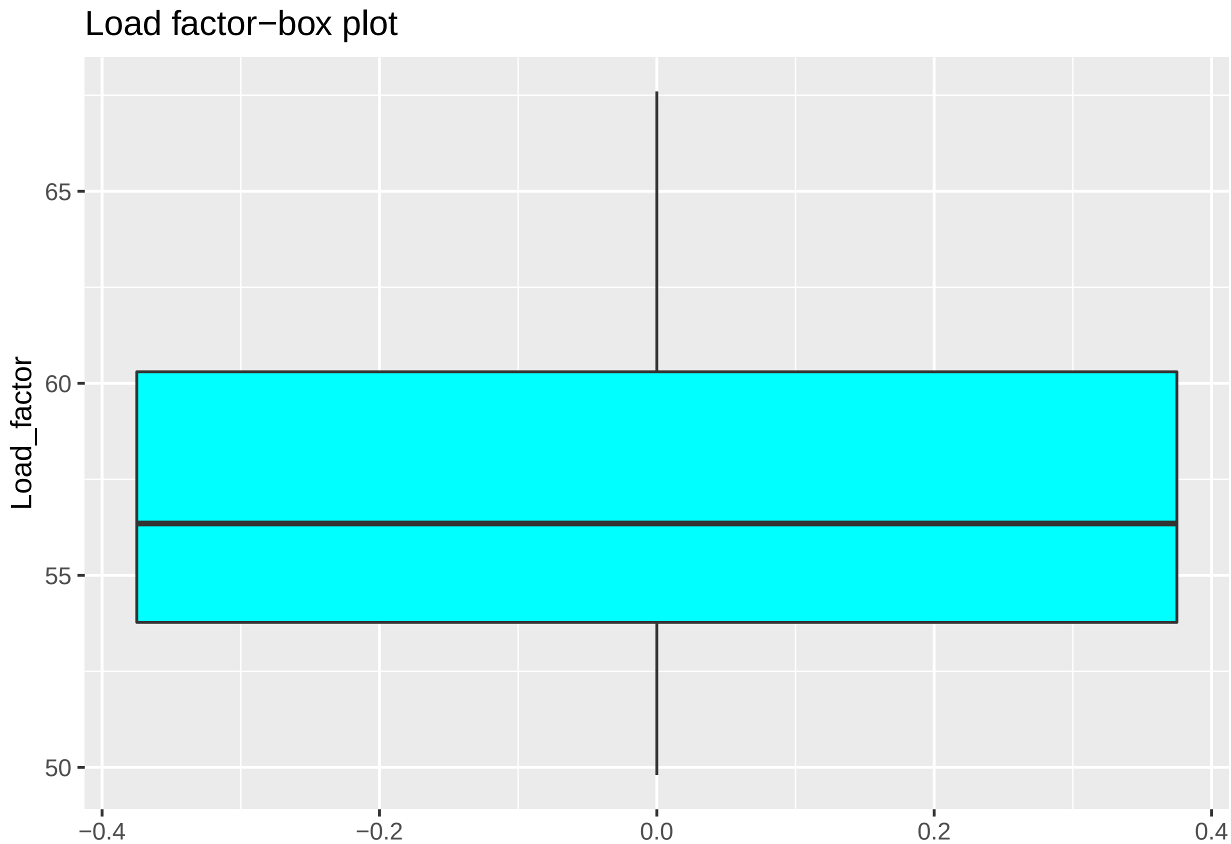

```r
###RoR_Box plot
ggplot(Utilities.dt) +
  geom_boxplot(aes(y= RoR),fill = "cyan", outlier.color = "red")+
  ggtitle("RoR-box plot")
```




```r
###Cost Box Plot
ggplot(Utilities.dt) +
  geom_boxplot(aes(y= Cost),fill = "cyan", outlier.color = "red")+
  ggtitle("Cost-box plot")
```

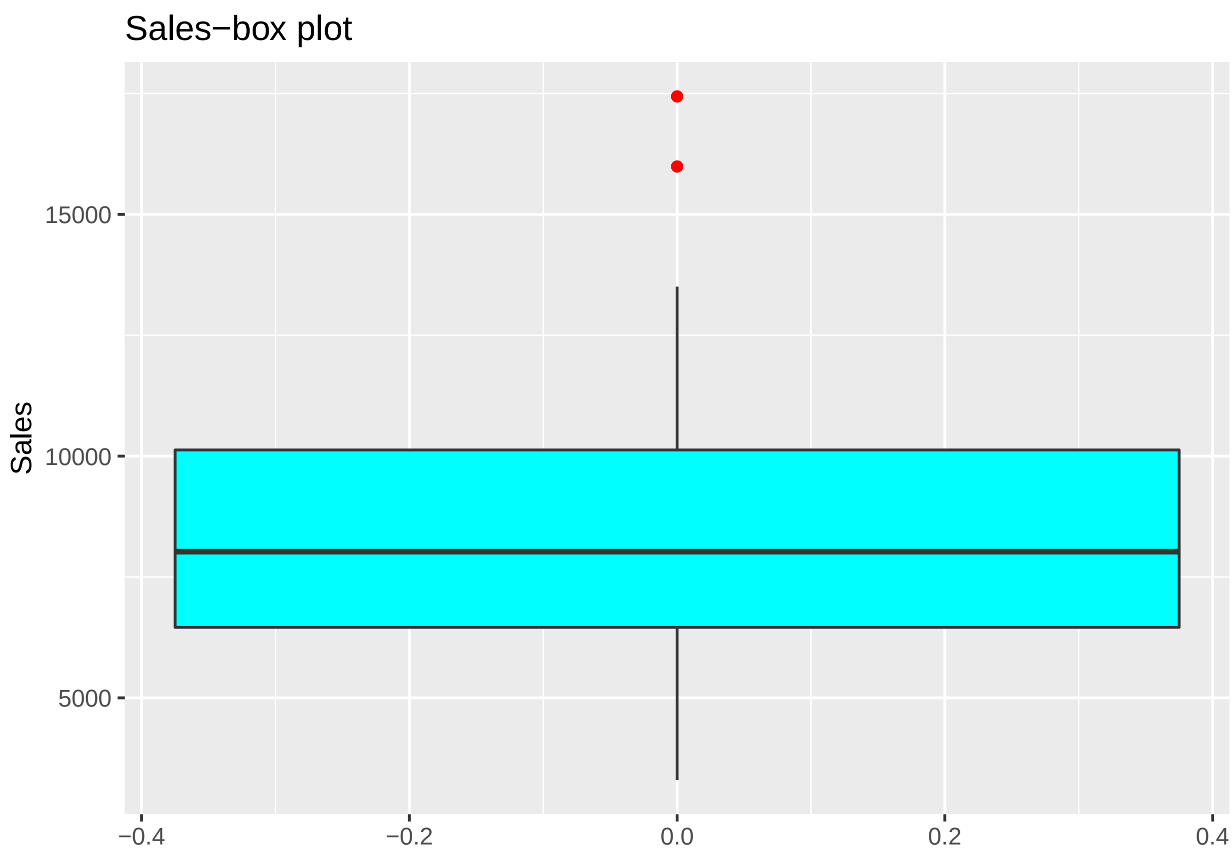



```
###Load_factor Box Plot
ggplot(Utilities.dt) +
  geom_boxplot(aes(y= Load_factor),fill = "cyan", outlier.color = "red")+
  ggtitle("Load factor-box plot")
```

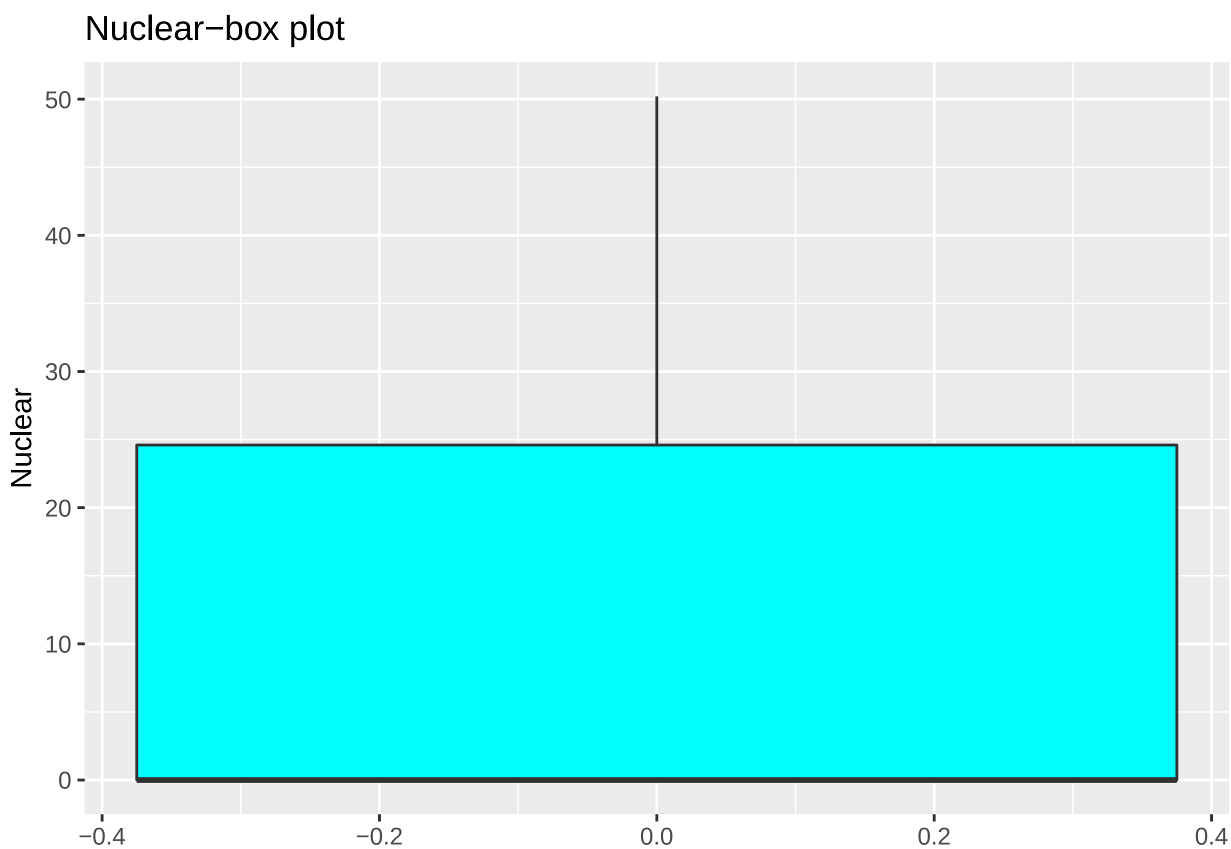



```r
###Demand_growth Box Plot
ggplot(Utilities.dt) +
  geom_boxplot(aes(y= Demand_growth),fill = "cyan", outlier.color = "red")+
  ggtitle("Demand growth-box plot")
```

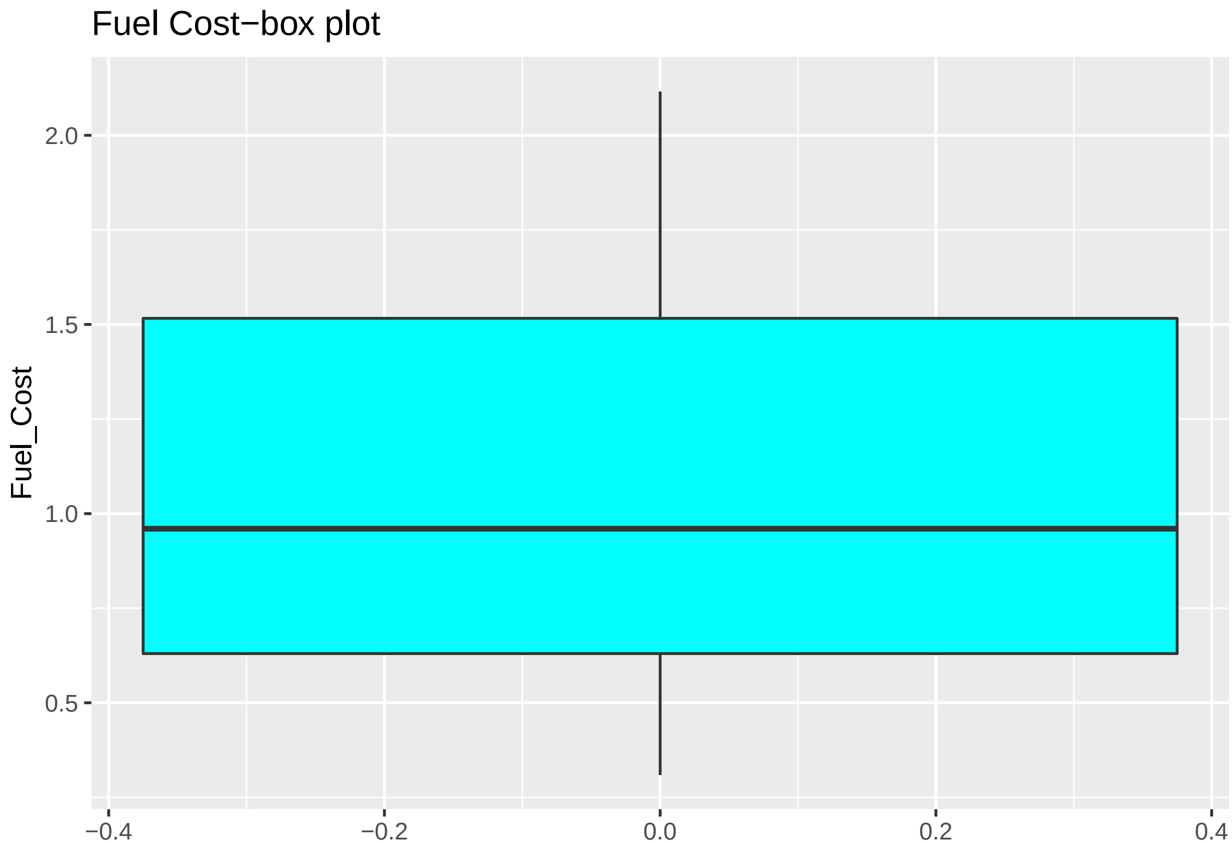

```r
###Sales Box Plot
ggplot(Utilities.dt) +
  geom_boxplot(aes(y= Sales),fill = "cyan", outlier.color = "red")+
  ggtitle("Sales-box plot")
```




```
###Nuclear Box Plot
ggplot(Utilities.dt) +
  geom_boxplot(aes(y= Nuclear),fill = "cyan", outlier.color = "red")+
  ggtitle("Nuclear-box plot")
```



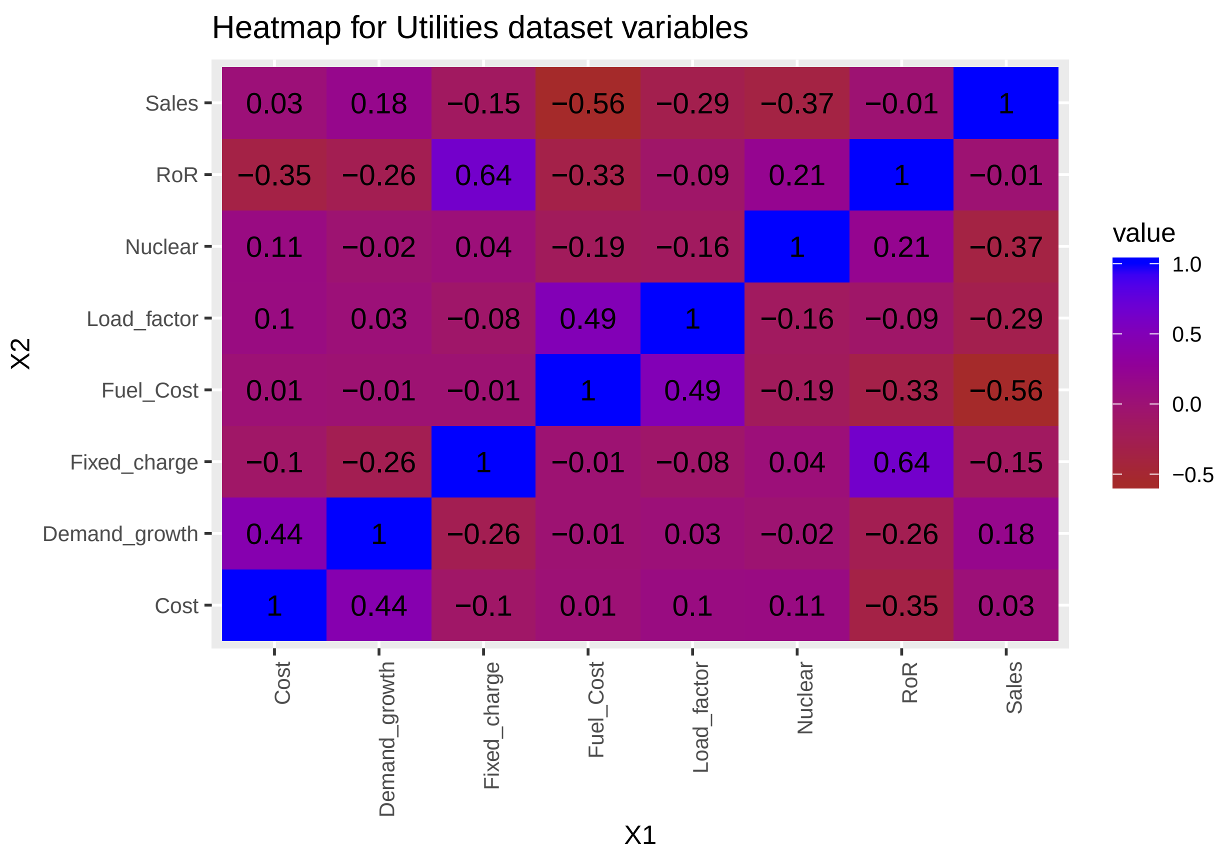

```
###Fuel_cost
ggplot(Utilities.dt) +
  geom_boxplot(aes(y= Fuel_Cost),fill = "cyan", outlier.color = "red")+
  ggtitle("Fuel Cost-box plot")
```



Interpretation of Solution 2: Yes, from the box plot of the 8 variables we can infer that there are extreme values. Fixed charge and Sales variable has 4 and 2 values as outliers respectively. There are extreme values for both the variables. The values are nearly 1.5 times interquartile range.

Question 3: Create a heatmap for the numeric variables. Discuss any interesting trend you see in this chart

```r
correlation_matrix <- round(cor(Utilities.dt[,-c(1)]),2)
melted_correlation_matrix <- melt(correlation_matrix)
ggplot(melted_correlation_matrix,aes(x=X1,y=X2,fill = value))+
  scale_fill_gradient(low = "brown",high = "blue")+
  geom_tile()+
  geom_text(aes(x=X1,y=X2,label = value))+
  theme(text = element_text(size = 10), axis.text.x = element_text(angle = 90,hjust = 1))+
  ggtitle("Heatmap for Utilities dataset variables")
```



Answer 3 : There is extreme positive correlation between RoR and Fixed_Charge as we can see from the figure with correlation coefficient of 0.64. Second comes RoR and Fixed_Charge with correlation coefficient of 0.49.The third most is between Demand_growth and Cost variables with a correlation coefficient of 0.44. The rise and fall between these variables are closely associated

Lowest correlation coefficient is between Sales and Fuel_Cost which is -0.56. The second most is between Nuclear and Sales variables with a correlation coefficient of -0.37. The third one is between RoR and Cost variables with a correlation coefficient of -0.35. Rise and fall goes hand in hand between these variables.

Question 4: Run principal component analysis using unscaled numeric variables in the dataset. How do you interpret the results from this model?

```
pcs_u <- prcomp(Utilities.dt[,-c(1)])
pcs_u$rot
```

```
##                        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
```

```r
summary(pcs_u)
```

```
## 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
```

Answer 4: From PCA for unscaled numeric variables we can infer PC1 value suits model accuracy as it accounts for 99.98% of total variance as proportion of variance is 0.9998. It is also the main contributor variance as it has SD 3549.9901. Sales component from the dataset contributes to the effect on principal components as its variation is very high as seen.

Question 5: Next, run principal component model after scaling the numeric variables. Did the results/interpretations change? How so? Explain your answers.

```r
Pca_s <- prcomp(Utilities.dt[,-c(1)], scale. = T)
Pca_s$rot
```

```
##                       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
```

```
summary(Pca_s)
```

```
## 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
```

Answer 5: We can see that PC1 has highest standard deviation when compared with other PCs with Standard deviation of 1.4741. In case if we need model to capture 95.176% of variance we will shift from PC1 to PC6 or if we want to capture 97.883% of variance we will shift from PC1 to PC7. ROR has positive influence on the PCs. Second positive influence comes from Fixed_charge. As all these are scaled, as a result sales doesnt have high variation or influence as in the previous case.