HW1_GROUP BUAN6356002_11

Navya Bulusu, Jeffrey Chang, Bhavana Chowdary Kandimalla, Naishal Kanubhai Thakkar, Rucha Nichkawade

9/17/2019

1. Load packages

```
if(!require("pacman")) install.packages("pacman")
pacman::p_load(forecast, tidyverse, gplots, GGally, mosaic,
               scales, mosaic, mapproj, mlbench, data.table,reshape)
search()
theme_set(theme_classic())
```

2.Import the utilities.csv into a data table package

```
utilities <- fread("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.

```
##Storing the numerical data
utilities_num <- utilities[,-1]
##calculating  minimum, maximum, mean, median, standard deviation
print(paste("Minimum"))
```

```
## [1] "Minimum"
```

```
Minimum = sapply(utilities_num,min, na.rm=TRUE)
Minimum
```

```
##  Fixed_charge           RoR          Cost   Load_factor Demand_growth
##         0.750         6.400        96.000        49.800        -2.200
##         Sales       Nuclear     Fuel_Cost
##      3300.000         0.000         0.309
```

```
print(paste("Maximum"))
```

```
## [1] "Maximum"
```

```
Maximum = sapply(utilities_num,max, na.rm=TRUE)
Maximum
```

```
##  Fixed_charge           RoR          Cost   Load_factor Demand_growth
##         1.490        15.400       252.000        67.600         9.200
##         Sales       Nuclear     Fuel_Cost
##     17441.000        50.200         2.116
```

```
print(paste("Mean"))
```

```
## [1] "Mean"
```

```
mean = sapply(utilities_num, mean, na.rm=TRUE)
mean
```

```
## Fixed_charge          RoR         Cost   Load_factor Demand_growth 
##      1.114091    10.736364   168.181818     56.977273      3.240909 
##         Sales      Nuclear    Fuel_Cost 
##   8914.045455    12.000000     1.102727
```

```r
print(paste("Median"))
```

```
## [1] "Median"
```

```r
Median = sapply(utilities_num, mean, na.rm=TRUE)
Median
```

```
## Fixed_charge          RoR         Cost   Load_factor Demand_growth 
##      1.114091    10.736364   168.181818     56.977273      3.240909 
##         Sales      Nuclear    Fuel_Cost 
##   8914.045455    12.000000     1.102727
```

```r
print(paste("Standard Deviation"))
```

```
## [1] "Standard Deviation"
```

```r
standard_deviation = sapply(utilities_num,sd, na.rm=TRUE)
standard_deviation
```

```
## Fixed_charge          RoR         Cost   Load_factor Demand_growth 
##     0.1845112    2.2440494   41.1913495     4.4611478     3.1182503 
##         Sales      Nuclear    Fuel_Cost 
##  3549.9840305   16.7919198    0.5560981
```

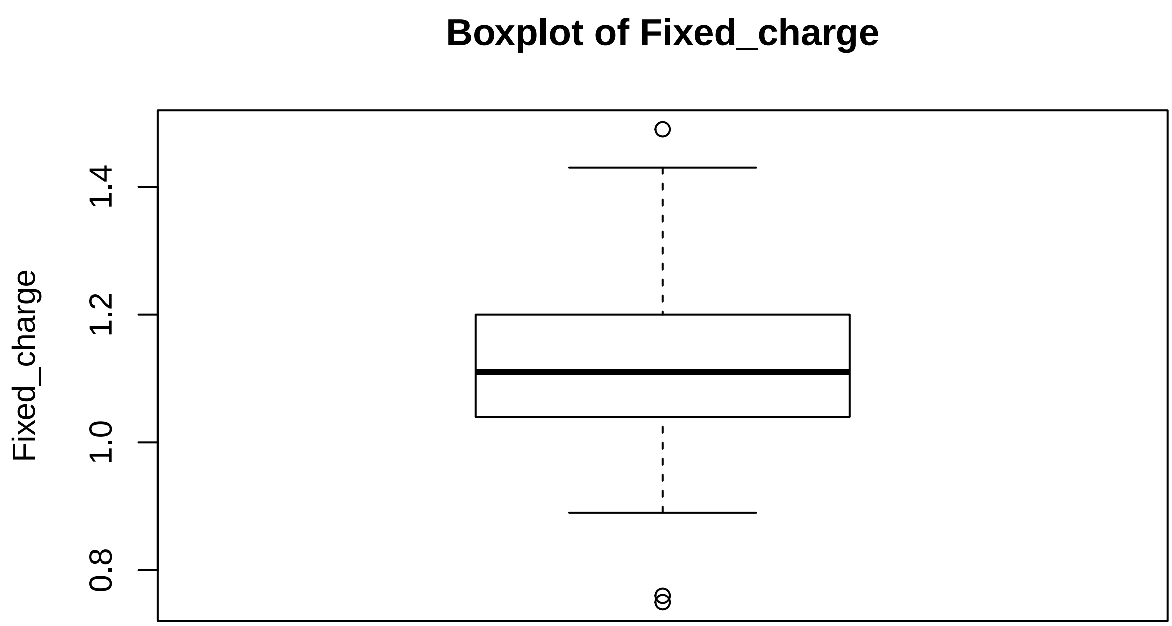
Explanation:

The “Sales” variable has the largest variability. Variability of the variable is measured by its standard deviation, hence “Sales” has the highest standard deviation. Since the unit for sales is ‘kilowatthour use per year’ there is a large variation in data for this variable.

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.

```
##boxplots for each of the numeric variables
boxplot(utilities_num$Fixed_charge, ylab='Fixed_charge', main='Boxplot of Fixed_charge')
```



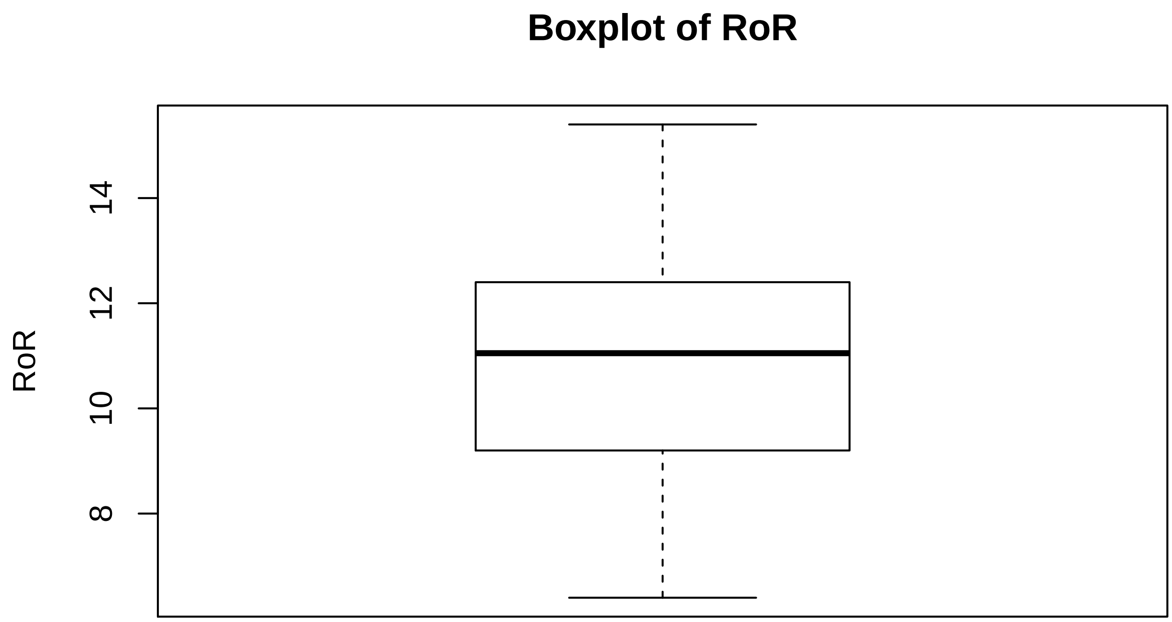

```
print(paste("The outliers of Fixed_charge value"))
```

```
## [1] "The outliers of Fixed_charge value"
```

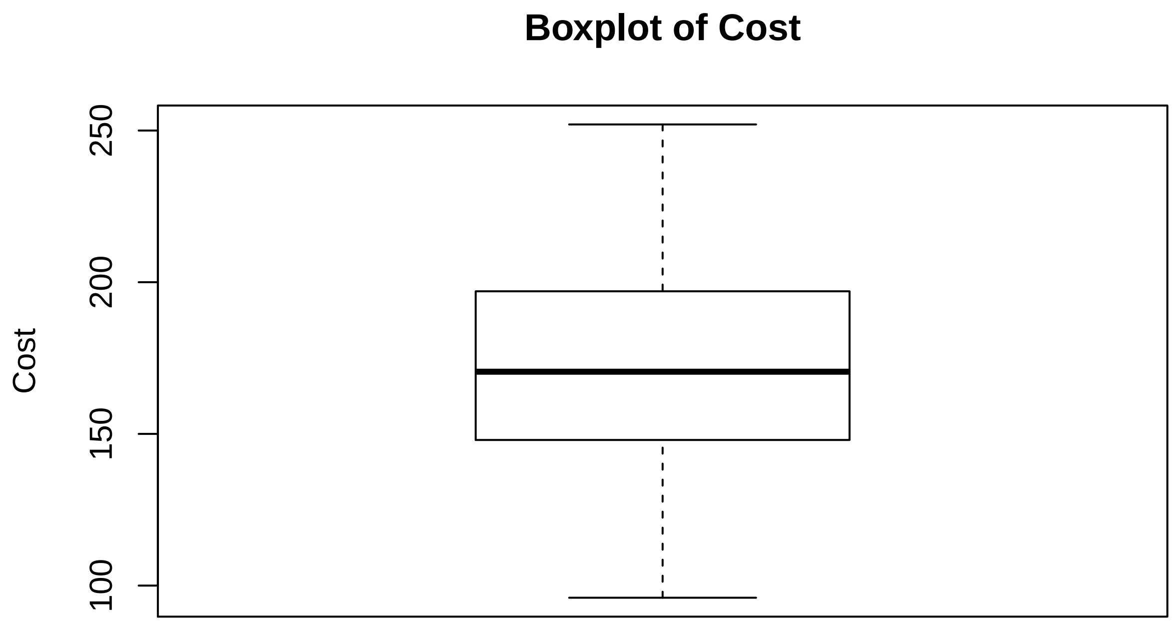
```
q1_Fixed_charge=1.042
q2_Fixed_charge=1.110
q3_Fixed_charge=1.190
iqr_Fixed_charge=q3_Fixed_charge-q1_Fixed_charge

for (value in utilities$Fixed_charge) {
  if (value>(q3_Fixed_charge+1.5*iqr_Fixed_charge) | value <(q1_Fixed_charge-1.5*iqr_Fixed_charge)) {
    print(value)}
}
```

```
## [1] 1.43
## [1] 1.49
## [1] 0.75
## [1] 0.76
```

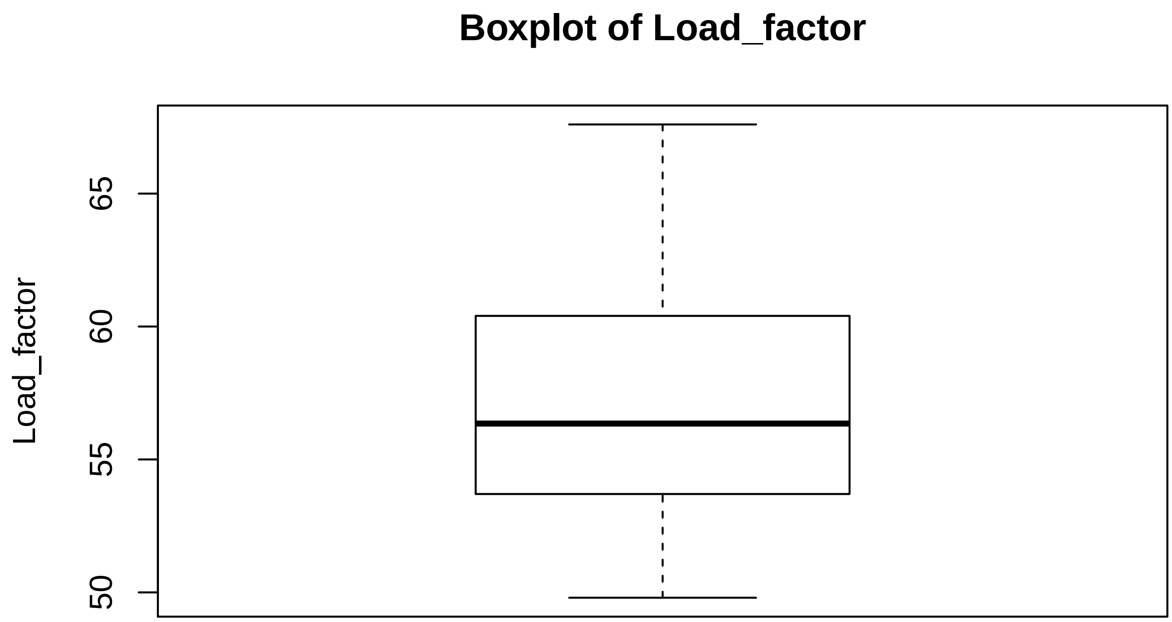
```
boxplot(utilities_num$RoR,ylab='RoR', main='Boxplot of RoR')
```



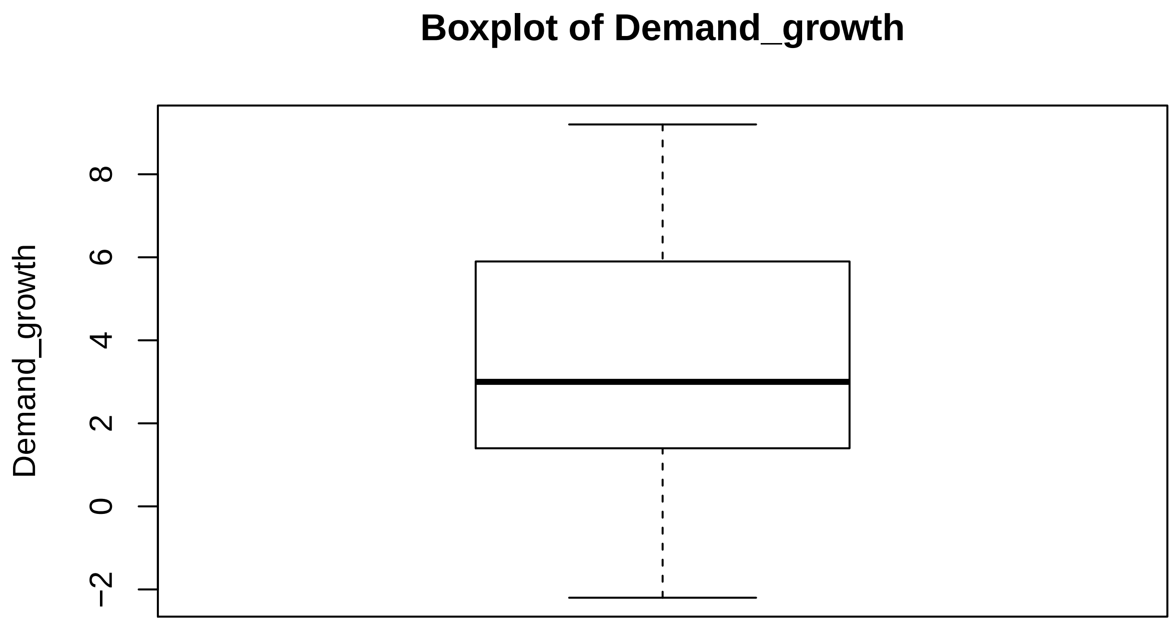

```r
boxplot(utilities_num$Cost,ylab='Cost', main='Boxplot of Cost')
```



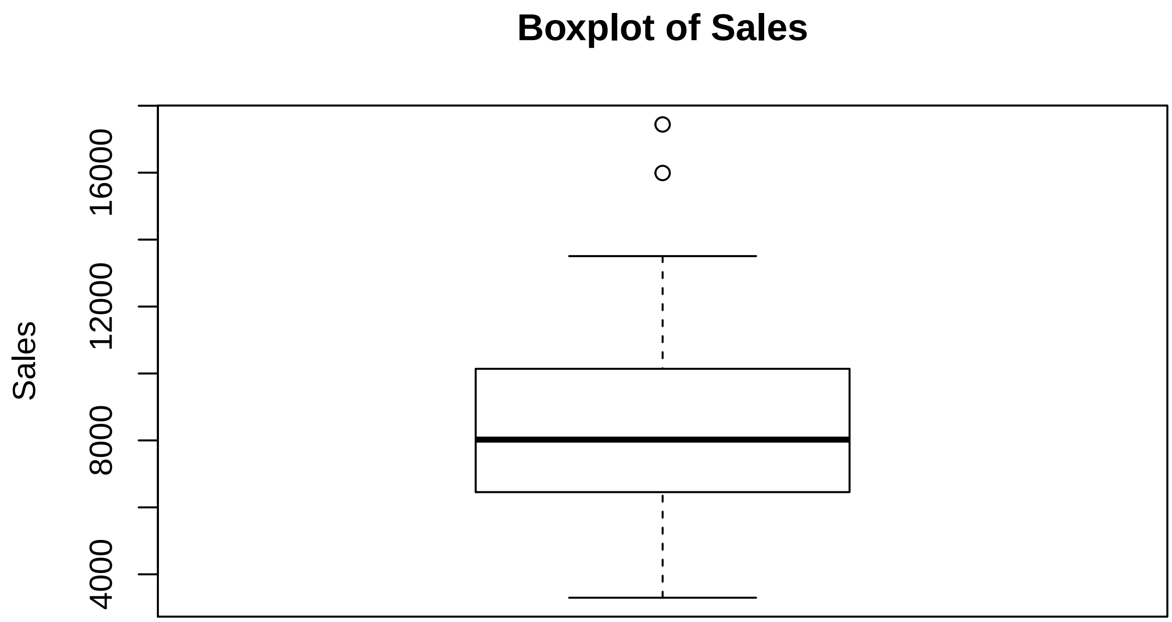

```
boxplot(utilities_num$Load_factor,ylab='Load_factor', main='Boxplot of Load_factor')
```




```r
boxplot(utilities_num$Demand_growth,ylab='Demand_growth', main='Boxplot of Demand_growth')
```




```
boxplot(utilities_num$Sales,ylab='Sales', main='Boxplot of Sales')
```



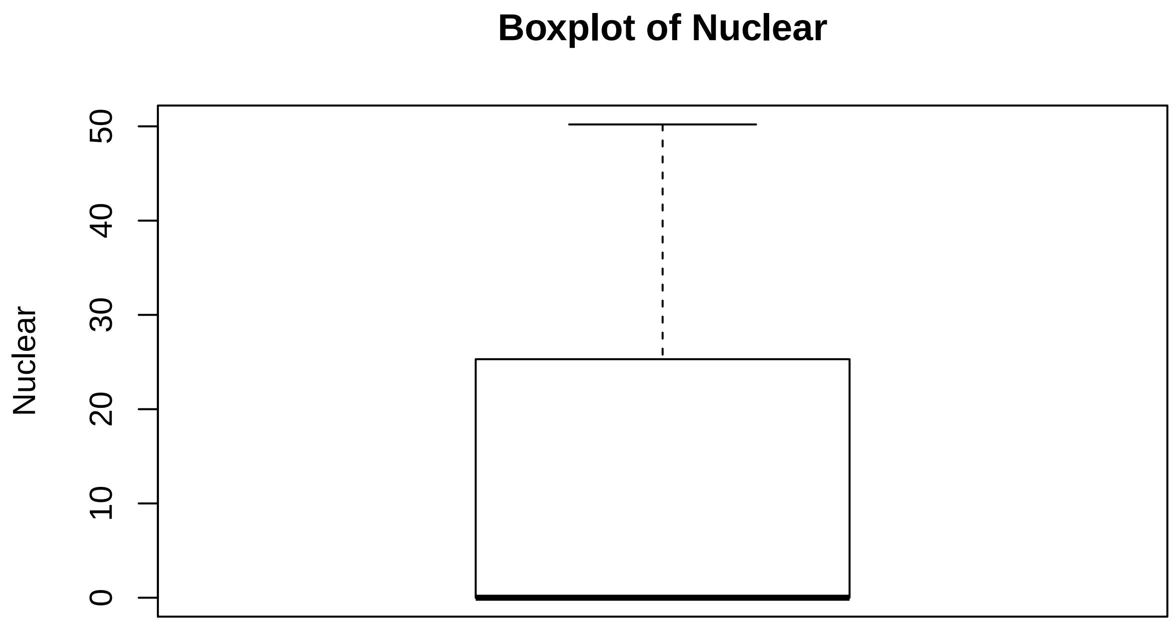

```r
print(paste("The outliers of Sales value"))
```

```
## [1] "The outliers of Sales value"
```

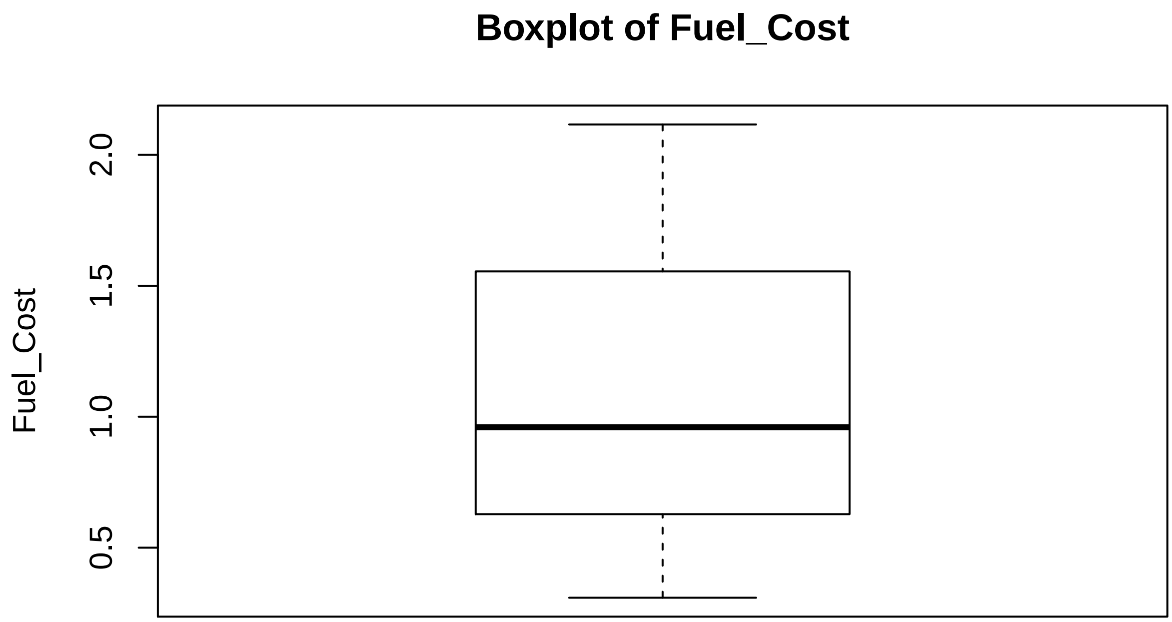
```r
q1_Sales=6458
q2_Sales=8024
q3_Sales=10128
iqr_Sales=q3_Sales-q1_Sales

for (value6 in utilities$Sales) {
  if (value6>(q3_Sales+1.5*iqr_Sales) | value6 <(q1_Sales-1.5*iqr_Sales)) {
    print(value6)}
}
```

```
## [1] 17441
## [1] 15991
```

```r
boxplot(utilities_num$Nuclear,ylab='Nuclear',main='Boxplot of Nuclear')
```




```r
boxplot(utilities_num$Fuel_Cost,ylab='Fuel_Cost',main='Boxplot of Fuel_Cost')
```



Explanation:

Yes. There are extreme values for Sales and Fixed Charge Variables.

The extreme values for sales are : 15991 for Puget company 17441 for Nevada company

The data points of sales for Puget and Nevada company differ significantly from other observations which also explains that there is a largest variation within sales.

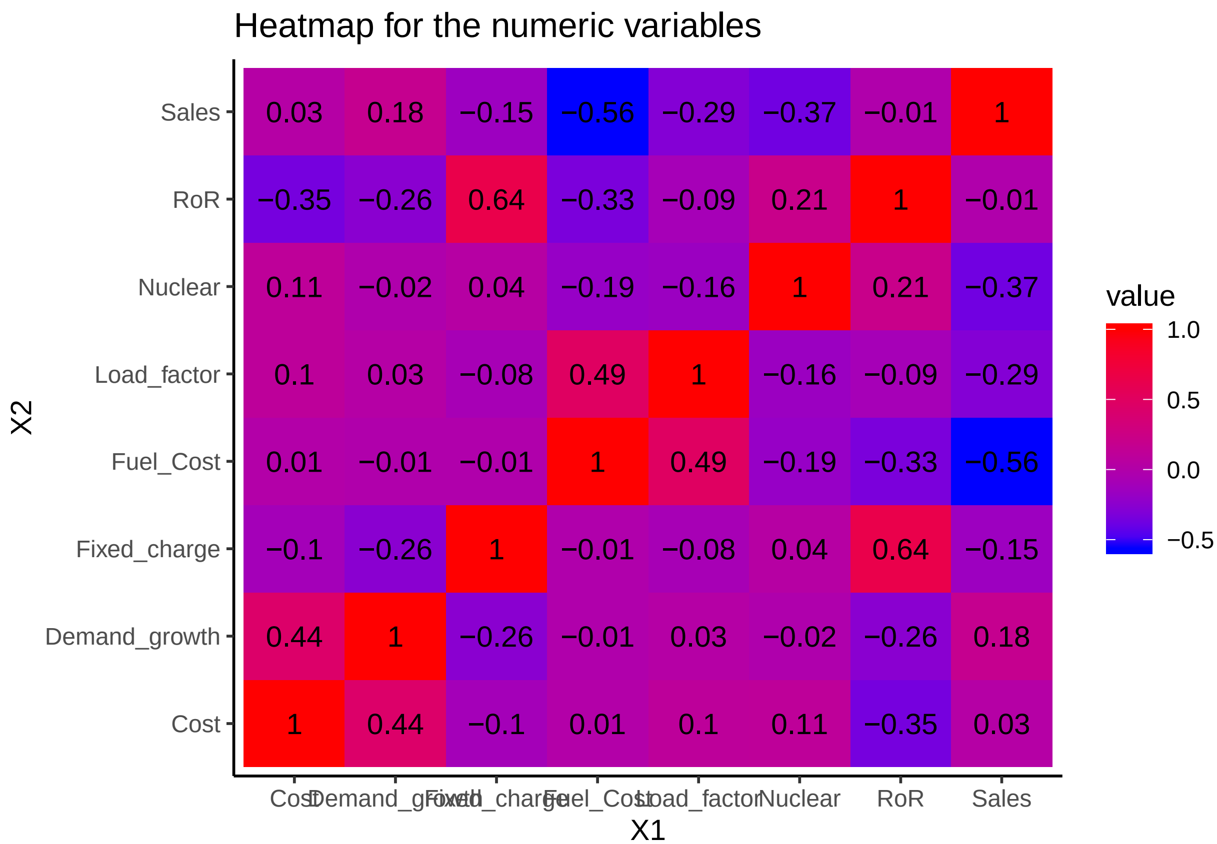
The extreme values for Fixed charge is: 1.43 for Central company 1.49 offered by the NY company 0.76 for San Diego 0.75 for Nevada

The data point of San Diego & Nevada for fixed charge 0.76 & 0.75 which it significantly lower than other observations and the data points for central company and NY company are 1.43 and 1.49 respectively which are significantly higher than other observations.

Question 3:

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

```
##a heatmap for the numeric variables
cor.mat <- round(cor(utilities_num),2)
melted.cor.mat <- melt(cor.mat)
ggplot(melted.cor.mat, aes(x = X1, y = X2, fill = value)) +
scale_fill_gradient(low="blue", high="Red") +
geom_tile() +
geom_text(aes(x = X1, y = X2, label = value)) +
ggtitle("Heatmap for the numeric variables")
```



Explanation:

The interesting trend observed here is:

The Variables Fixed_charge and RoR have relatively strong positive correlation coefficient of 0.64,It can be inferred that if there is High RoR on Capital there is increase in the Fixed Charges

The Variables Load_Factor and Fuel cost have relatively positive correlation coefficient of 0.49. Load Factor refers to the ratio of total billing energy(KWh) to possible total energy used. The fuel cost depends on the operating schedule. If the operating schedule is high, the power generated will be high thus increasing the load factor. Therefore, as the load factor increases, the fuel cost will also increase.

The variable Sales and Fuel_cost have relatively negative correlation coefficient of -0.56. If the Fuel_cost increases, the price will increase and thus the sales will be less.

Question 4:

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

```r
### PCA on 8 variables
pcs8 <- prcomp(na.omit(utilities [,-1]))
summary(pcs8)
```

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

```r
pcs8$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
```

Explanation:

Principal Component (PC) 1 accounts for more than 99% of the total variation of the original 8 factors. Thus, we can catch 99.98% of the variability's information with PC1. This is because of the (kilowatt_hour_use/year) dimension of sales. Additionally, to catch 100% of the information we only need 3 PC instead of the 8 original dimensions. Also, if the variables aren't normalized we cannot use them for model building.

Question 5:

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

```
### PCA using Normalized variables
pcs.cor <- prcomp(na.omit(utilities [,-1]), scale. = T)
summary(pcs.cor)
```

```
## 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
```

```
pcs.cor$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
```

Explanation:

Yes, the results change. Before Normalizing the Data, Principal Component(PC) 1 gives a 99.98% cumulative performance while after scaling we have to consider the weights up to PC6 to about 95% cumulative performance result from the given data. Since the variables are measured in different units, the variability is unclear and hence cannot be compared. Hence, we normalize the data to eliminate those differences. As the variables are scaled the dominating variables in PC1 are changed from 'Sales' to 'RoR' and 'Fixed Charge'. The dominating variables in PC2 are Sales and Fuel_Cost and similarly, we can find dominating variables for all other components. After normalization of variables, the variation can be justified by 6 variables (PC1~PC6) which are still 2 variables less than in the former.