BUAN 6356 - Homework 1

Group No.9 (Shubhi Kala, Spoorthi Thatipally, Hao-Yu Lin, Loc Nguyen, Tatsat Joshi)

2/10/2020

R Markdown file

Install and load necessary packages and check loading

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

Import Dataset

```
#Using fread for faster data reading and it returns data table by default
utilities.dt <- fread("Utilities.csv")

#Summary of the data table
summary(utilities.dt)
```

```
##    Company           Fixed_charge        RoR             Cost
##  Length:22          Min.   :0.750   Min.   : 6.40   Min.   : 96.0
##  Class :character   1st Qu.:1.042   1st Qu.: 9.20   1st Qu.:148.5
##  Mode  :character   Median :1.110   Median :11.05   Median :170.5
##                     Mean   :1.114   Mean   :10.74   Mean   :168.2
##                     3rd Qu.:1.190   3rd Qu.:12.35   3rd Qu.:195.8
##                     Max.   :1.490   Max.   :15.40   Max.   :252.0
##   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
```

Solution to Question 1

Calculation of Mean, Minimum, Maximum, Median, and Standard Deviation

```r
# Q1. Compute the minimum, maximum, mean, median, and standard deviation for each of the numeric variab

colNames <- c("Fixed_charge", "RoR", "Cost", "Load_factor", "Demand_growth", "Sales",
              "Nuclear", "Fuel_Cost")

# Finding the minimum, maximum, mean, median, and standard deviation for all the numeric variables
mean_dt <- utilities.dt[, lapply(.SD, mean), .SDcols = colNames]
min_dt <- utilities.dt[, lapply(.SD, min), .SDcols = colNames]
max_dt <- utilities.dt[, lapply(.SD, max), .SDcols = colNames]
median_dt <- utilities.dt[, lapply(.SD, median), .SDcols = colNames]
sd_dt <- utilities.dt[, lapply(.SD, sd), .SDcols = colNames]
cov_dt <- utilities.dt[, lapply(.SD, cv), .SDcols = colNames]

options(scipen = 999)

# Printing all the values in a data table
table <- data.table(rbind(mean_dt,min_dt,max_dt,median_dt, sd_dt, cov_dt))

cbind(c("Mean", "Minimum", "Maximum", "Median", "Standard Deviation", "Coefficient of Variance"), table)
```

```
##                          V1 Fixed_charge        RoR        Cost Load_factor
## 1:                     Mean    1.1140909 10.7363636 168.1818182 56.97727273
## 2:                  Minimum    0.7500000  6.4000000  96.0000000 49.80000000
## 3:                  Maximum    1.4900000 15.4000000 252.0000000 67.60000000
## 4:                   Median    1.1100000 11.0500000 170.5000000 56.35000000
## 5:       Standard Deviation    0.1845112  2.2440494  41.1913495  4.46114781
## 6: Coefficient of Variance    0.1656159  0.2090139   0.2449215  0.07829697
##    Demand_growth          Sales   Nuclear Fuel_Cost
## 1:      3.240909  8914.0454545 12.000000 1.1027273
## 2:     -2.200000  3300.0000000  0.000000 0.3090000
## 3:      9.200000 17441.0000000 50.200000 2.1160000
## 4:      3.000000  8024.0000000  0.000000 0.9600000
## 5:      3.118250  3549.9840305 16.791920 0.5560981
## 6:      0.962153     0.3982461  1.399327 0.5042934
```

Answer1 Inference:

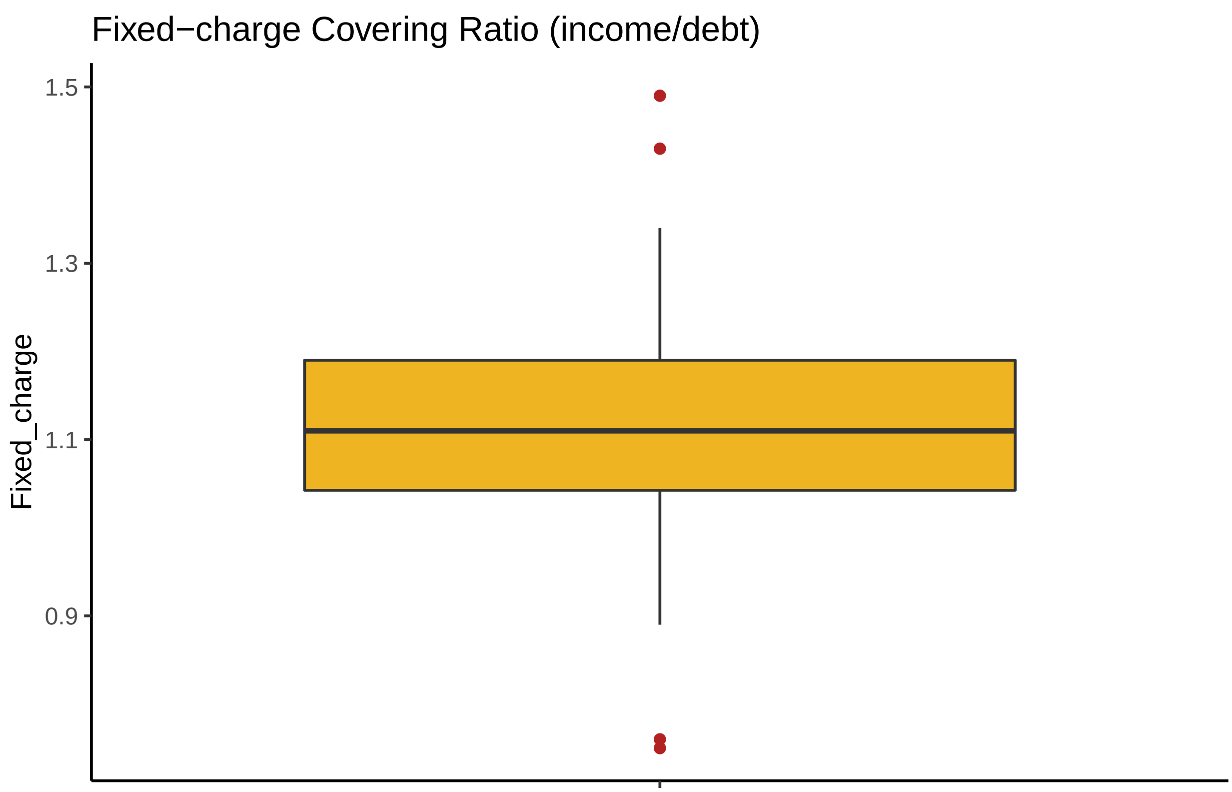
The measure of relative variability is Coefficient of Variation. Since the numerical variables have different units of measurement, we calculated coefficient of variation for the comparision.

According to the above computation we find that Nuclear has the highest variability because it's Coefficient of variation is 1.399327 (Standard Deviation / Mean = 16.79192 / 12.00) followed by Demand_growth and Fuel_cost.

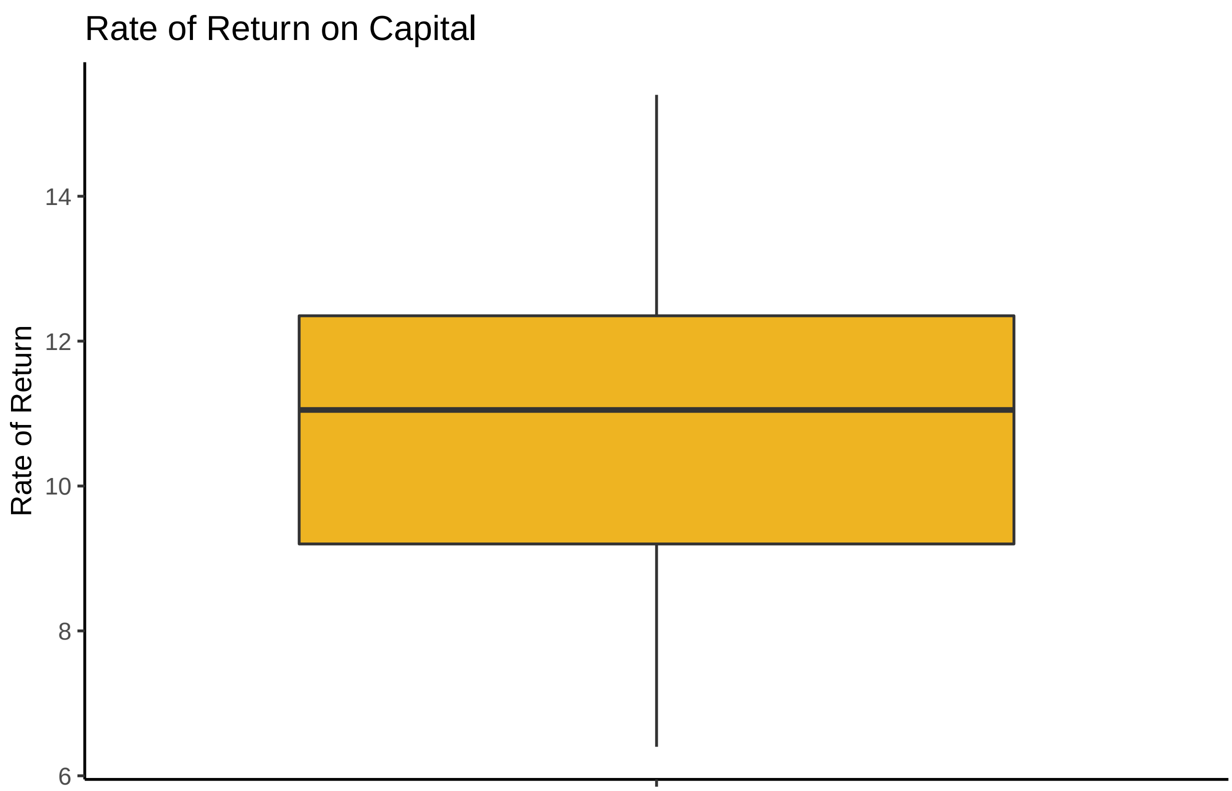
Solution to Question 2

Boxplots for the numerical variables

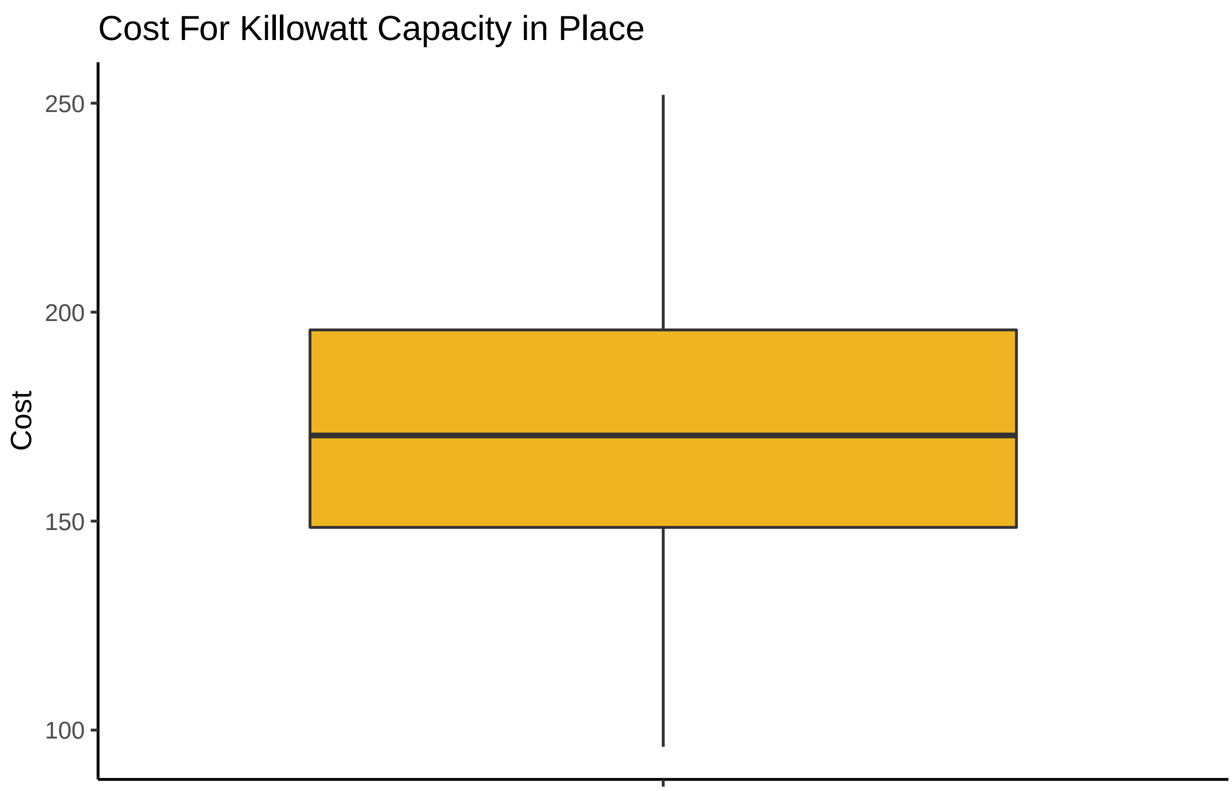
```
##Q2: Create boxplots for each of the numeric variables. Are there any extreme values for any of the va

# Boxplot for Fixed_charge:
ggplot(utilities.dt) +
  geom_boxplot(aes(x = "", y = Fixed_charge),fill = "goldenrod2", outlier.color = "firebrick") +
  ylab("Fixed_charge") + xlab("") + ggtitle("Fixed-charge Covering Ratio (income/debt)")
```



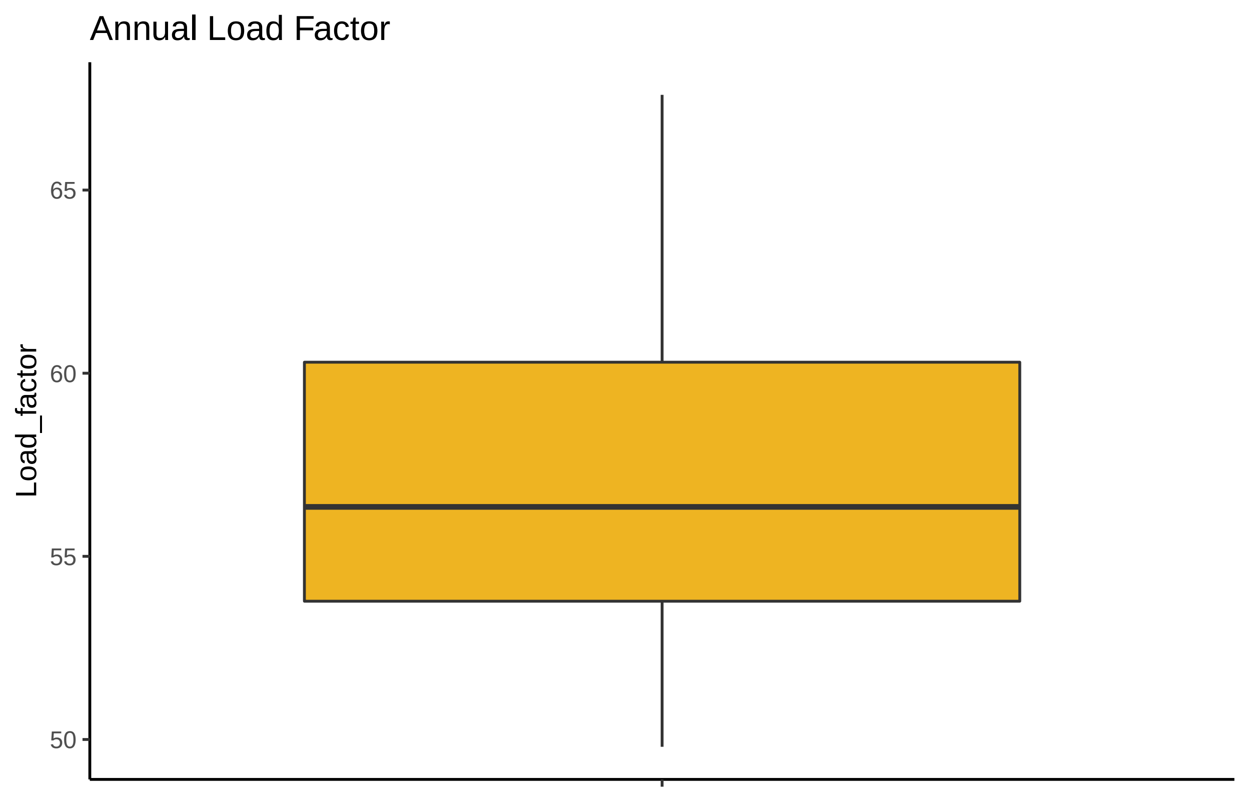

```
# Boxplot for RoR:
ggplot(utilities.dt) +
  geom_boxplot(aes(x = "", y = RoR), fill = "goldenrod2", outlier.color = "firebrick") +
  ylab("Rate of Return") + xlab("") +ggtitle("Rate of Return on Capital")
```



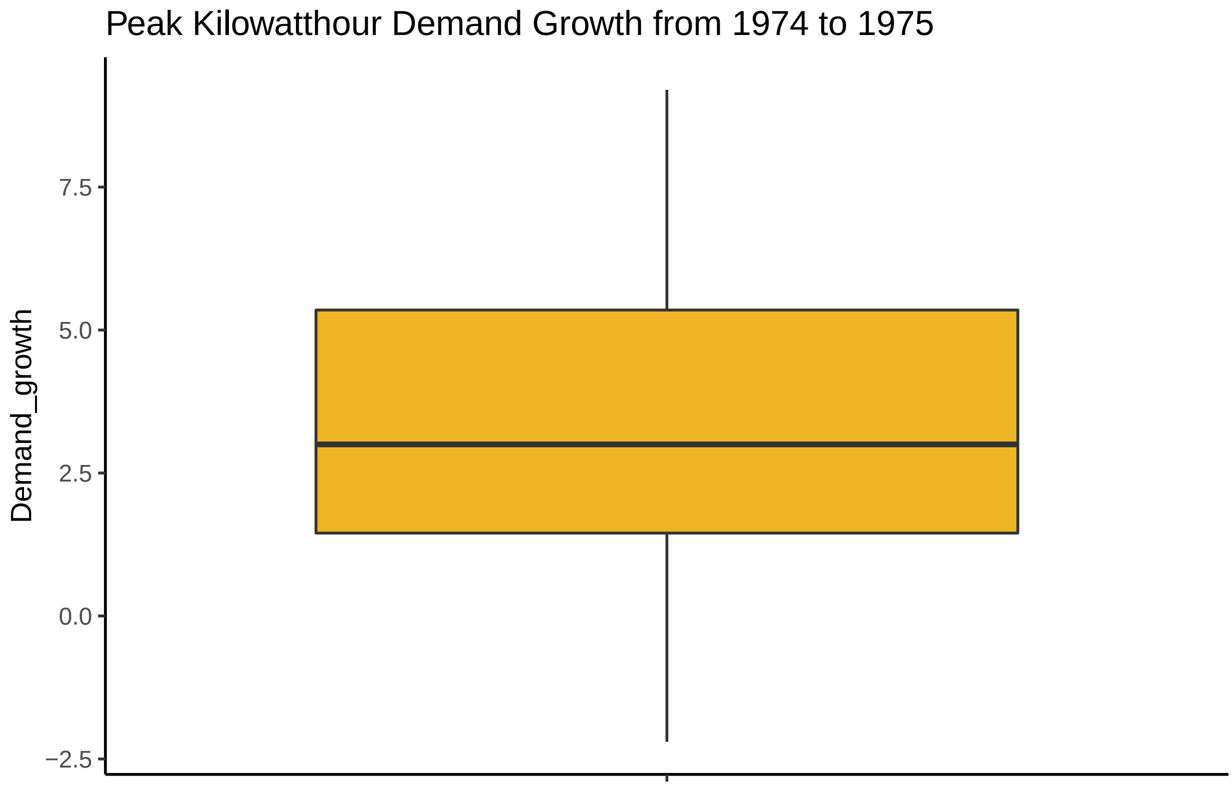

```r
# Boxplot for Cost:
ggplot(utilities.dt) +
  geom_boxplot(aes(x = "", y = Cost), fill = "goldenrod2", outlier.color = "firebrick") +
  ylab("Cost") + xlab("") + ggtitle("Cost For Killowatt Capacity in Place")
```



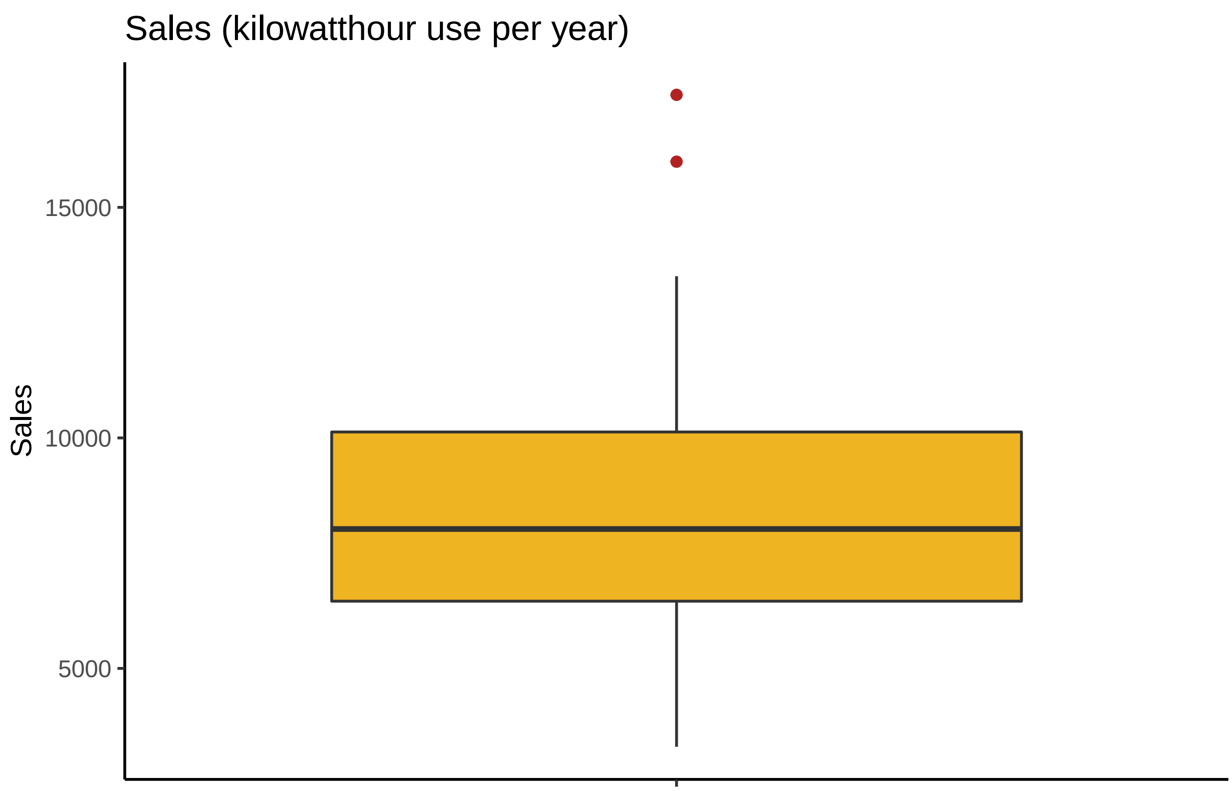

```r
# Boxplot for Load_factor:
ggplot(utilities.dt) +
  geom_boxplot(aes(x = "", y = Load_factor), fill = "goldenrod2", outlier.color = "firebrick") +
  ylab("Load_factor") + xlab("") + ggtitle("Annual Load Factor")
```



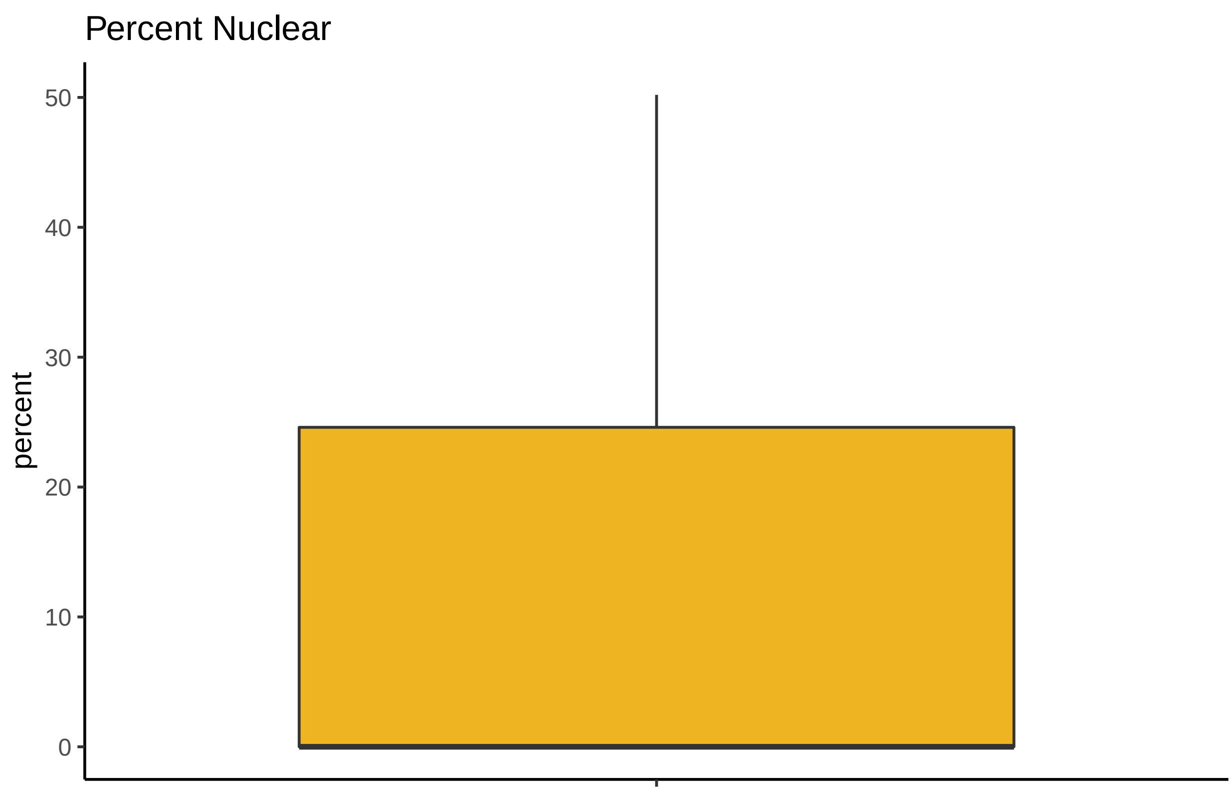

```r
# Boxplot for Demand_growth:
ggplot(utilities.dt) +
  geom_boxplot(aes(x = "", y = Demand_growth), fill = "goldenrod2", outlier.color = "firebrick") +
  ylab("Demand_growth") + xlab("") +ggtitle("Peak Kilowatthour Demand Growth from 1974 to 1975")
```



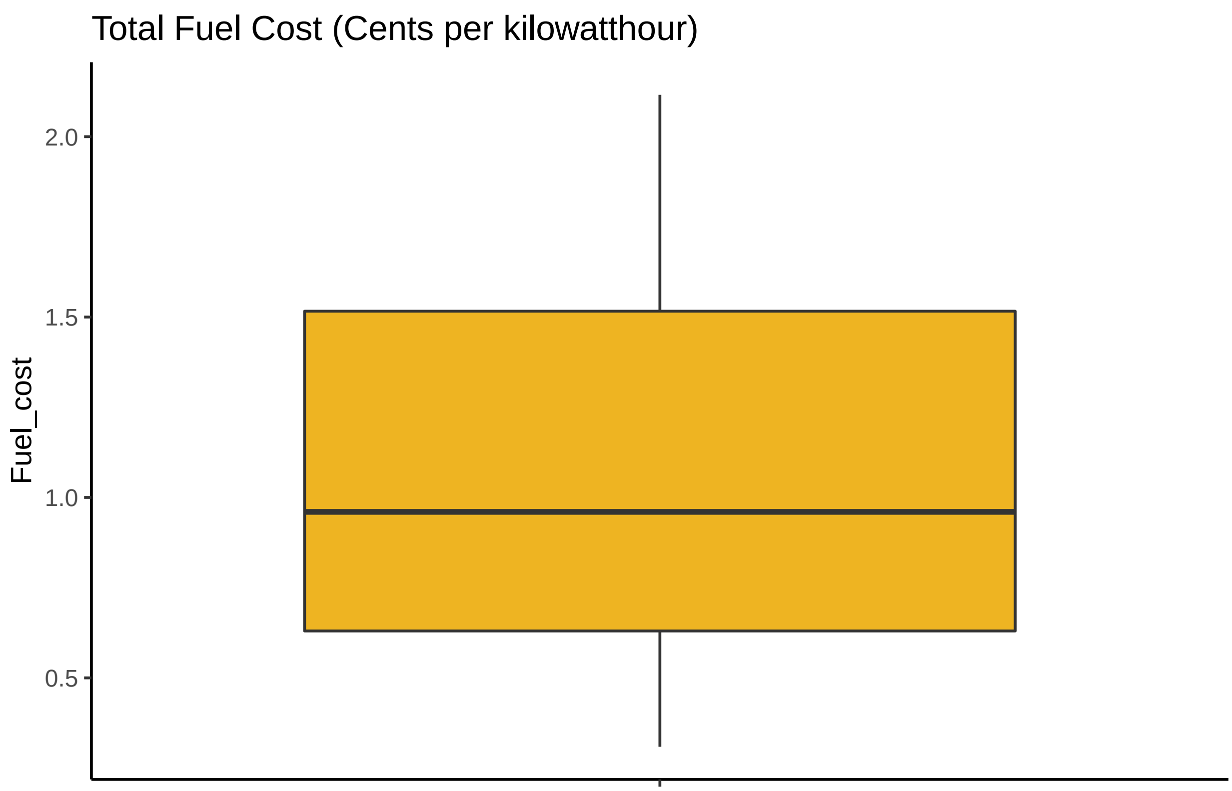

```
# Boxplot for Sales:
ggplot(utilities.dt) +
  geom_boxplot(aes(x = "", y = Sales), fill = "goldenrod2", outlier.color = "firebrick") + ylab("Sales")
  xlab("") + ggtitle("Sales (kilowatthour use per year)")
```




```r
# Boxplot for Nuclear:
ggplot(utilities.dt) +
  geom_boxplot(aes(x = "", y = Nuclear), fill = "goldenrod2",
               outlier.color = "firebrick") + ylab("percent") + xlab("") + ggtitle("Percent Nuclear")
```




```
# Boxplot for Fuel Cost:
ggplot(utilities.dt) +
  geom_boxplot(aes(x = "", y = Fuel_Cost), fill = "goldenrod2",
               outlier.color = "firebrick") + ylab("Fuel_cost") + xlab("") + ggtitle("Total Fuel Cost (
```



Answer2 Inference:

From the boxplot, it can be inferred that:

"Fixed_Charge" and "Sales" are two variables that have extreme values, since they have outliers that extend beyond 1.5 times the inter-quartile range.

"Fixed_Charge" variable has extreme values, Nevada and San Diego are the outliers which means there debt is greater than the income. NY and Central are also the outliers for them income is greater than the debt.

In the 'Sales', Nevada and Puget are the two companies which have high energy usage as compared to other companies.

Solution to Question 3

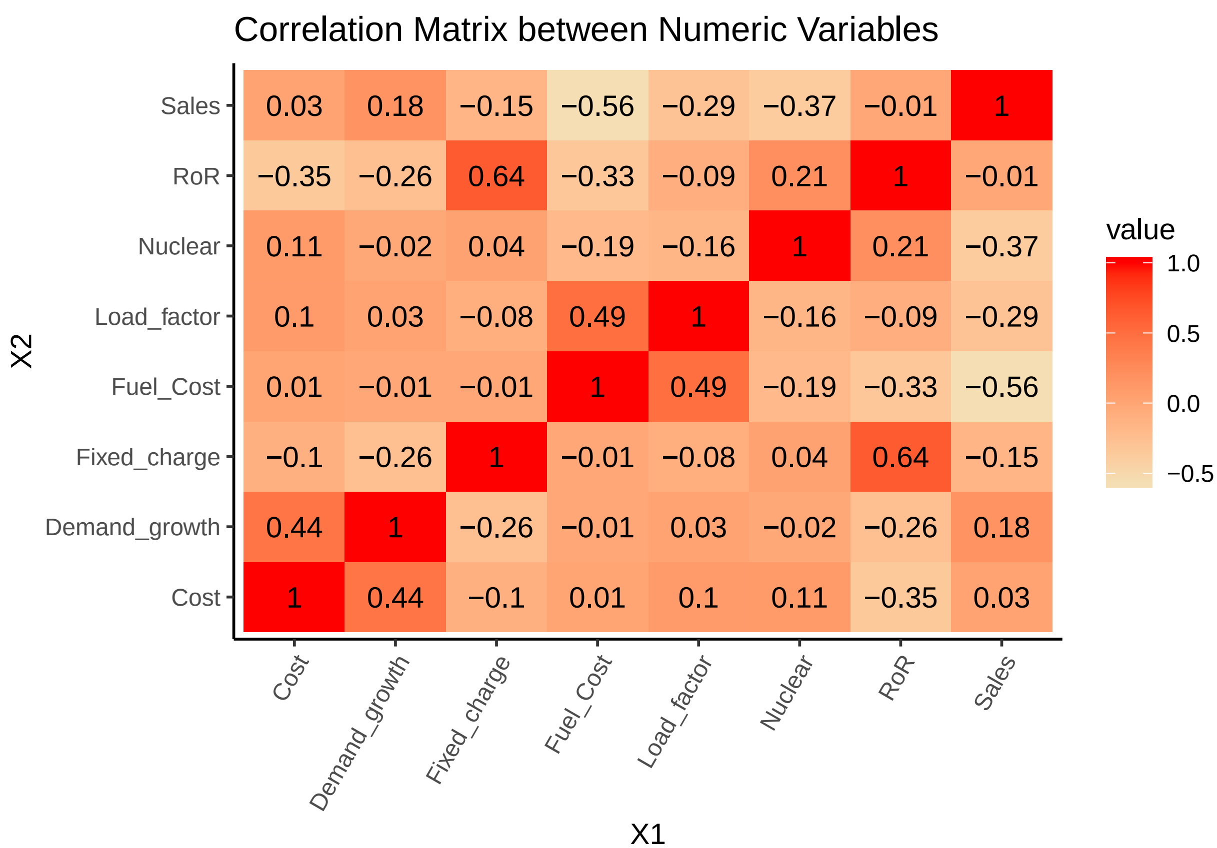
Heatmap for the numerical variables

```
# Q3. Create a heatmap for the numeric variables.  Discuss any interesting trend you see in this chart.

#Create Correlation Matrix
cor.mat <- round(cor(utilities.dt[,!c("Company")]),2)
head(cor.mat)
```

```
##                Fixed_charge   RoR  Cost Load_factor Demand_growth Sales Nuclear
```

```
## Fixed_charge          1.00  0.64 -0.10         -0.08          -0.26 -0.15    0.04
## RoR                   0.64  1.00 -0.35         -0.09          -0.26 -0.01    0.21
## Cost                 -0.10 -0.35  1.00          0.10           0.44  0.03    0.11
## Load_factor          -0.08 -0.09  0.10          1.00           0.03 -0.29   -0.16
## Demand_growth        -0.26 -0.26  0.44          0.03           1.00  0.18   -0.02
## Sales                -0.15 -0.01  0.03         -0.29           0.18  1.00   -0.37
##                Fuel_Cost
## Fixed_charge       -0.01
## RoR                -0.33
## Cost                0.01
## Load_factor         0.49
## Demand_growth      -0.01
## Sales              -0.56
```

```r
#Melt data to bring the correlation values in two axis
melted.cor.mat <- melt(cor.mat)

ggplot(melted.cor.mat, aes(x = X1, y = X2, fill = value)) +
  scale_fill_gradient(low = "wheat", high = "red") +
  geom_tile() +
  geom_text(aes(x = X1, y = X2, label = value)) +
  theme(axis.text.x = element_text(angle = 60, hjust = 1)) +  #writing x labels at an angle to increase

  ggtitle("Correlation Matrix between Numeric Variables")
```



Answer 3 Inference:

1. Heatmap is ued to analyse the correlation between the numerical variables. Correlation values ranges from -1 to 1.
2. The larger the number, the darker the color and the higher the correlation between two variables.
3. Diagonals are one because each variable is correlating to itself so it's a perfect correlation.
4. Here, 'ROR' and 'Fixed_charge' have high positive coorelation with a score of 0.64. So if the income increases and debt decreases it will boost up the Rate of Return and vice versa.
5. Also, the 'Load_factor' and 'Fuel_cost' has positive correlation of 0.49. It means with the increase in the Load factor, cost of the fuel will go up and vice versa.
6. On the other hand, 'Sales' and 'Fuel_cost' share a negative correlation of -0.56, so if the fuel cost increase then it would affect the sale.
7. 'Sales' also share a negative correlation of -0.37 with 'Nuclear', Also, the 'RoR' (Rate of return on capital) is negatively related to 'Cost' (cost in place).

Solution to Question 4

PCA using unscaled numerical variables

```r
# Q4 - Run principal component analysis using unscaled numeric variables in the dataset.  How do you in

pcs <- prcomp(na.omit(utilities.dt[,-c(1)]))
summary(pcs)
```

```
## 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
```

```r
pcs$rot
```

```
##                           PC1           PC2           PC3           PC4
## Fixed_charge   0.000007883140 -0.0004460932  0.0001146357 -0.0057978329
## RoR            0.000006081397 -0.0186257078  0.0412535878  0.0292444838
## Cost          -0.000324772402  0.9974928360 -0.0566502956 -0.0179103135
## Load_factor    0.000361835694  0.0111104272 -0.0964680806  0.9930009368
## Demand_growth -0.000154961568  0.0326730808 -0.0038575008  0.0544730799
## Sales         -0.999998303626 -0.0002209801  0.0017377455  0.0005270008
## Nuclear        0.001767631750  0.0589056695  0.9927317841  0.0949073699
## Fuel_Cost      0.000087804700  0.0001659524 -0.0157634569  0.0276496391
##                         PC5           PC6           PC7            PC8
## Fixed_charge   0.0198566131 -0.0583722527 -0.10029904246  0.993028030847
```

```
## RoR             0.2028309717 -0.9735822744 -0.05984233394 -0.067171657522
## Cost            0.0355836487 -0.0144563569 -0.00099867226 -0.001312104206
## Load_factor     0.0495177973  0.0333700701  0.02930751956  0.009745356549
## Demand_growth -0.9768581322 -0.2038187556  0.00889879033  0.008784362997
## Sales           0.0001471164  0.0001237088 -0.00009721241  0.000005226863
## Nuclear        -0.0057261758  0.0430954352 -0.01043774713  0.002059460566
## Fuel_Cost      -0.0215054038  0.0633116915 -0.99262829126 -0.095943724902
```

Answer 4 Inference:

Principal Component Analysis allows us to better visualize the variation present in the dataset with many variables.

So in this question we obtained the PCA for all the 8 numerical variables. Following insights can be drawn from the result,

1. From the summary of PCA it is implied that the new variable PC1 accounts for ~ 99.8 percent of the variation. The first principal component itself explained up to 0.9998 of the overall data variance and amount of variation decreases as we go from left to right.
2. Only one feature (PC1) can be used instead of all 8 numerical features to make predictions.
3. After applying the rotation, the weights for all components are generated.
4. From the rotational matrix, upon comparing the absolute values in PC1, it was identified that the "Sales" is dominant and has a highest contribution (9.999983e-01) and "Fuel_cost" gave the second highest contribution (8.780470e-05) to PC1.

Solution to Question 5

PCA using scaled numerical variables

```r
# Q5. Next, run principal component model after scaling the numeric variables.  Did the results/interpr

pcs.cor <- prcomp(na.omit(utilities.dt[,-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
```

```r
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
```

Answer 5 Inference

Running the PCA over all the variables with T scaling changed the output in the following ways:

1. From the summary, we can now imply that 7 PCs contribute to ~98% of the variation unlike the above where only PC1 was contributing to ~99% of information. Now 7 features are required to capture 0.97883 percent of the overall data instead of 1.
2. Also, on reading the rotational matrix values, “RoR” gave the highest contribution of 0.5711, and “Fixed Charge” contributed the second highest to PC1. “Sales”, on the other hand, had the least absolute value. Such deviation in the result is because Sales has the units in kilowatthour which is a different scale of measurement as compared to other variables. That’s why PCA output was severely affected in Q4.
3. PC2 was contributing 0.99998 before scaling, but after scaling it is capturing 0.2375 and PC7 is now contributing to 0.97883 which is the highest.
4. Hence, we can say that scaling is crucial when the magnitude of certain variables dominates the asssociation between the variables. (as we see the case of ‘Sales’). Unless all the variables are measured in the same scale, it’s recommended to normalise the data prior to PCA.