Homework_1

Shraddha Hemant Kadam (sxk190069@utdallas.edu)

06/16/2020

Loading Packages

```
pacman::p_load(data.table, gplots)
pacman::p_load(esquisse, forecast, tidyverse, gplots, GGally, gganimate,
               mosaic, scales, mosaic, mapproj, mlbench, data.table)
pacman::p_load(tidyverse, reshape, gplots, ggmap,
               mlbench, data.table, factoextra)
theme_set(theme_classic())
library(ggplot2)
```

Import data

```
utilities.df <- fread("~/Utilities.csv")
names(utilities.df)

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

view(utilities.df)
colnames(utilities.df) <- tolower(colnames(utilities.df))
```

Question:1

```
utilities.dt <- setDT(utilities.df)
summary(utilities.dt[,2:9])

##   fixed_charge        ror             cost         load_factor
##  Min.   :0.750   Min.   : 6.40   Min.   : 96.0   Min.   :49.80
##  1st Qu.:1.042   1st Qu.: 9.20   1st Qu.:148.5   1st Qu.:53.77
##  Median :1.110   Median :11.05   Median :170.5   Median :56.35
##  Mean   :1.114   Mean   :10.74   Mean   :168.2   Mean   :56.98
##  3rd Qu.:1.190   3rd Qu.:12.35   3rd Qu.:195.8   3rd Qu.:60.30
##  Max.   :1.490   Max.   :15.40   Max.   :252.0   Max.   :67.60
##  demand_growth        sales          nuclear         fuel_cost
##  Min.   :-2.200   Min.   : 3300   Min.   : 0.0   Min.   :0.309
##  1st Qu.: 1.450   1st Qu.: 6458   1st Qu.: 0.0   1st Qu.:0.630
##  Median : 3.000   Median : 8024   Median : 0.0   Median :0.960
##  Mean   : 3.241   Mean   : 8914   Mean   :12.0   Mean   :1.103
##  3rd Qu.: 5.350   3rd Qu.:10128   3rd Qu.:24.6   3rd Qu.:1.516
##  Max.   : 9.200   Max.   :17441   Max.   :50.2   Max.   :2.116

sd_num <- apply(utilities.dt[,2:9],2,sd)

sd_num
```

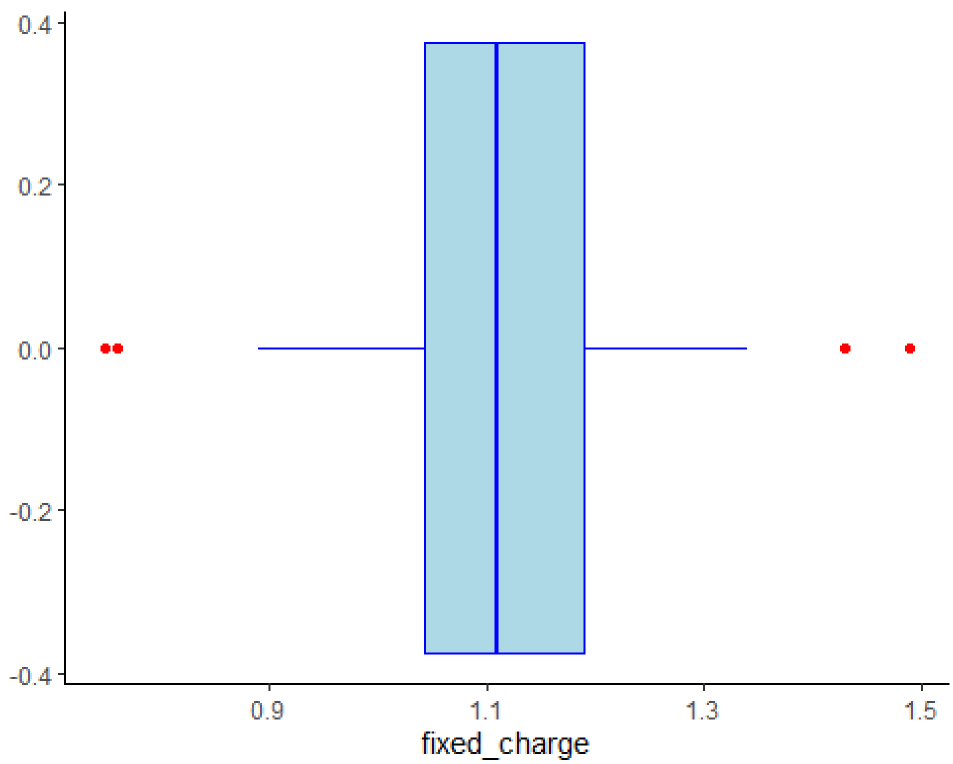
```
##  fixed_charge           ror          cost   load_factor demand_growth
##     0.1845112     2.2440494    41.1913495     4.4611478     3.1182503
##         sales       nuclear     fuel_cost
##  3549.9840305    16.7919198     0.5560981
```

```
"The variable 'Sales' and 'Cost' has the largest variability.
The Standard Deviation is a measure of how spread out numbers are. And, since
sales and cost has a standard deviation of 3549.9840305 and 41.1913495
respectively, we can figure out that they have the largest variability."
```

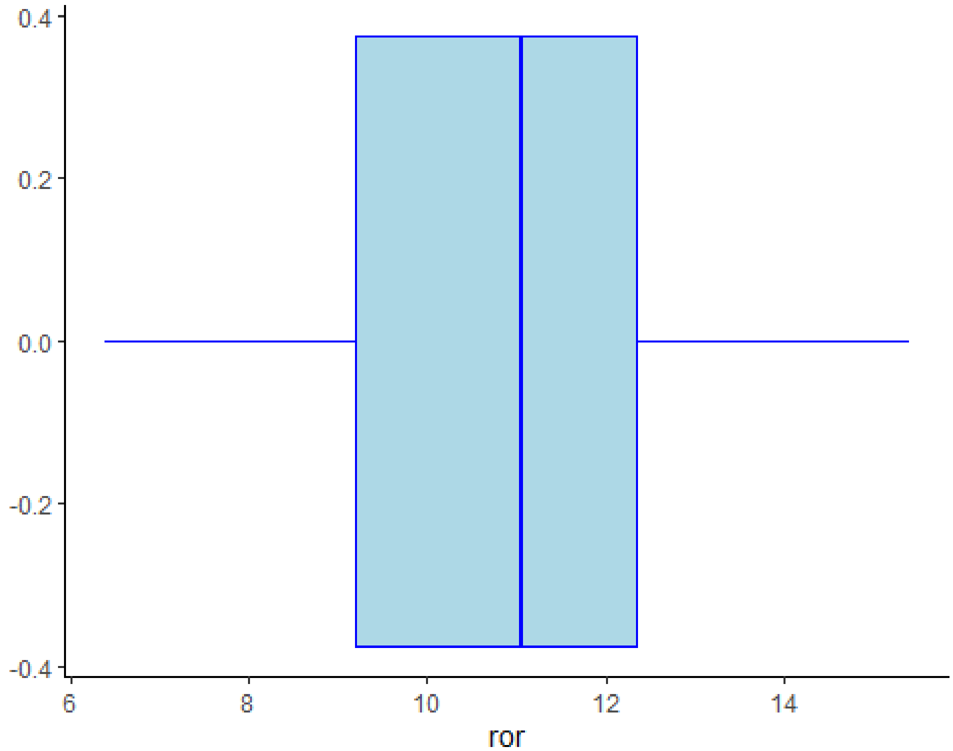
```
## [1] "The variable 'Sales' and 'Cost' has the largest variability.\nThe
Standard Deviation is a measure of how spread out numbers are. And, since
sales and cost has a standard deviation of 3549.9840305 and 41.1913495
respectively, we can figure out that they have the largest variability."
```

Question:2

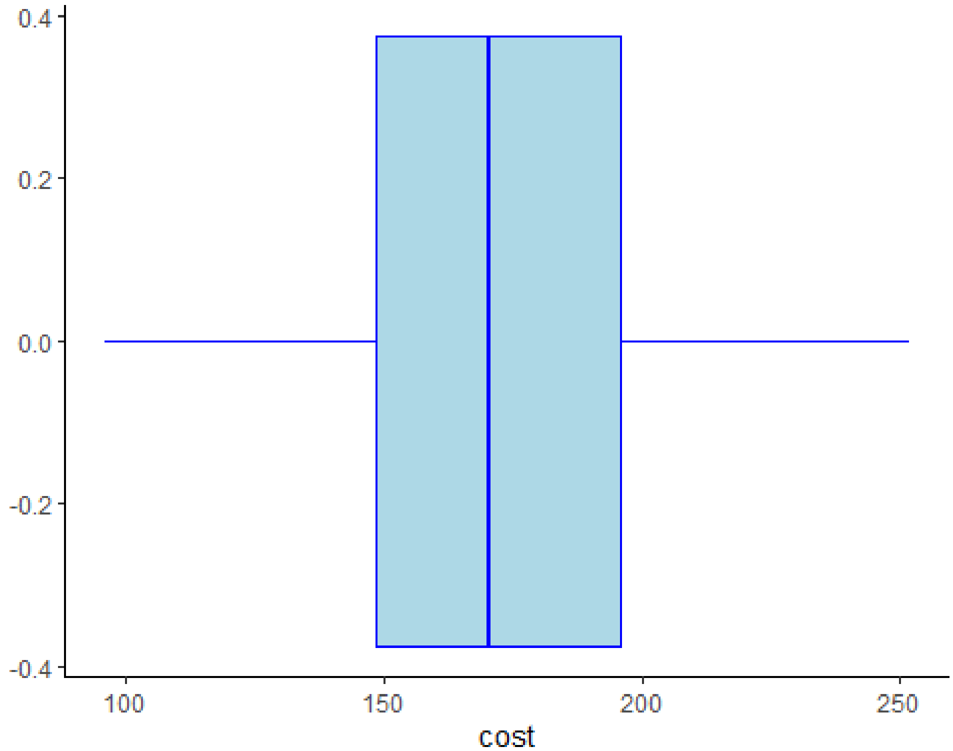
```
ggplot(data = utilities.dt) +
  geom_boxplot(mapping = aes(y = fixed_charge), color= "blue",
fill="lightblue", outlier.colour = "red")+
  coord_flip()
```



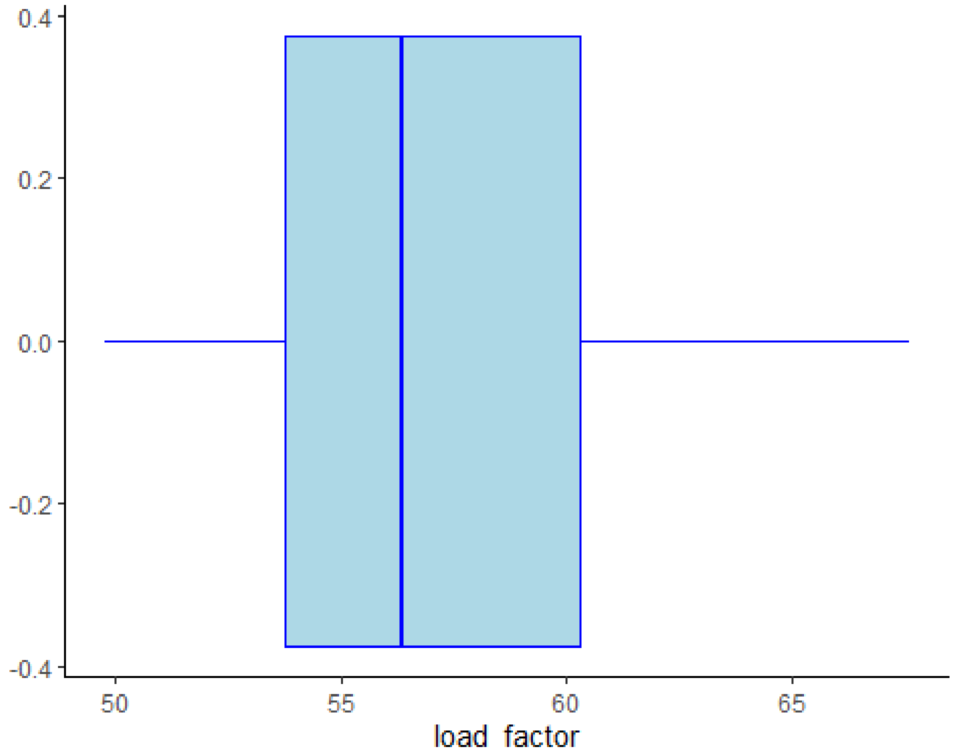

```
ggplot(data = utilities.dt) +
  geom_boxplot(mapping = aes(y = ror), color= "blue", fill="lightblue",
outlier.colour = "red")+
  coord_flip()
```



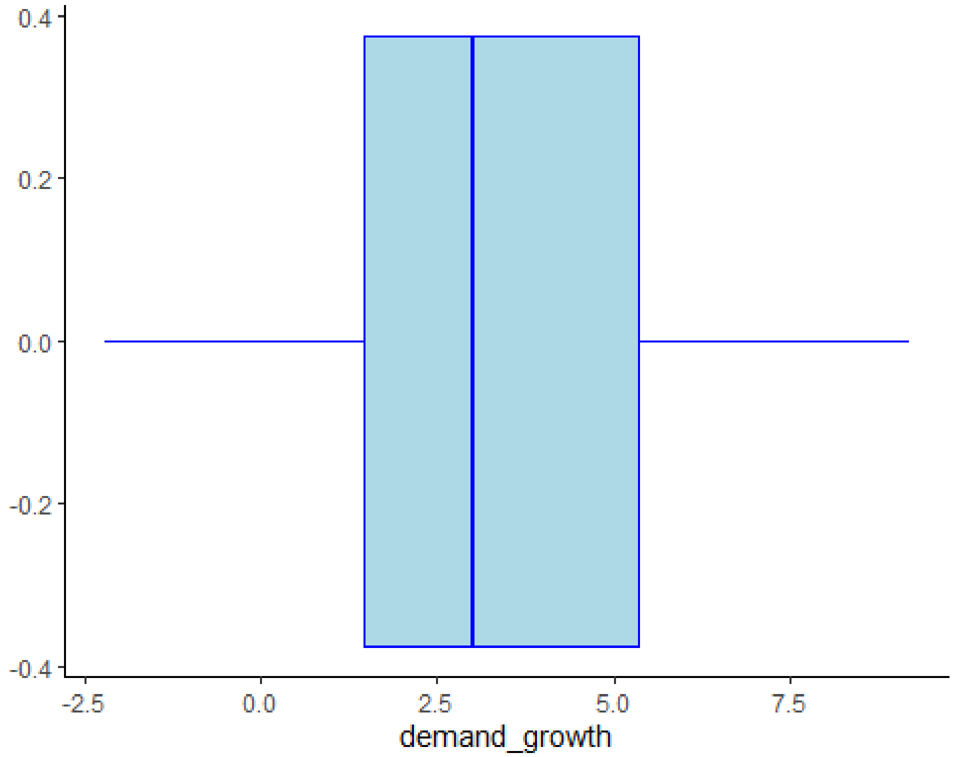

```
ggplot(data = utilities.dt) +
  geom_boxplot(mapping = aes(y = cost), color= "blue", fill="lightblue",
outlier.colour = "red")+
  coord_flip()
```



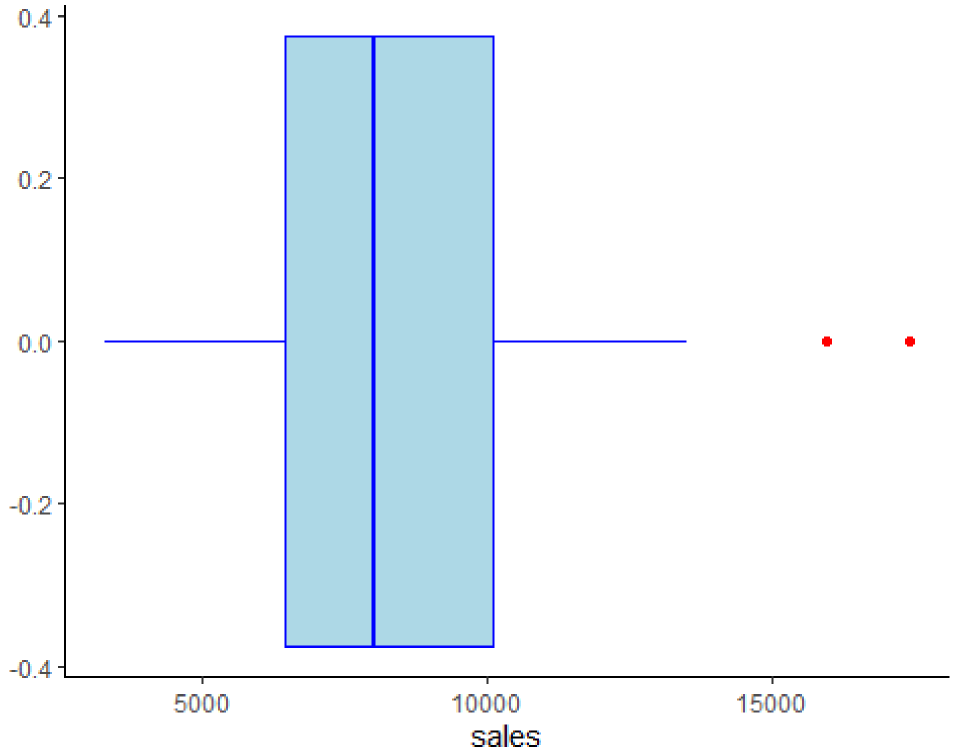

```
ggplot(data = utilities.dt) +
  geom_boxplot(mapping = aes(y = load_factor), color= "blue",
fill="lightblue", outlier.colour = "red")+
  coord_flip()
```



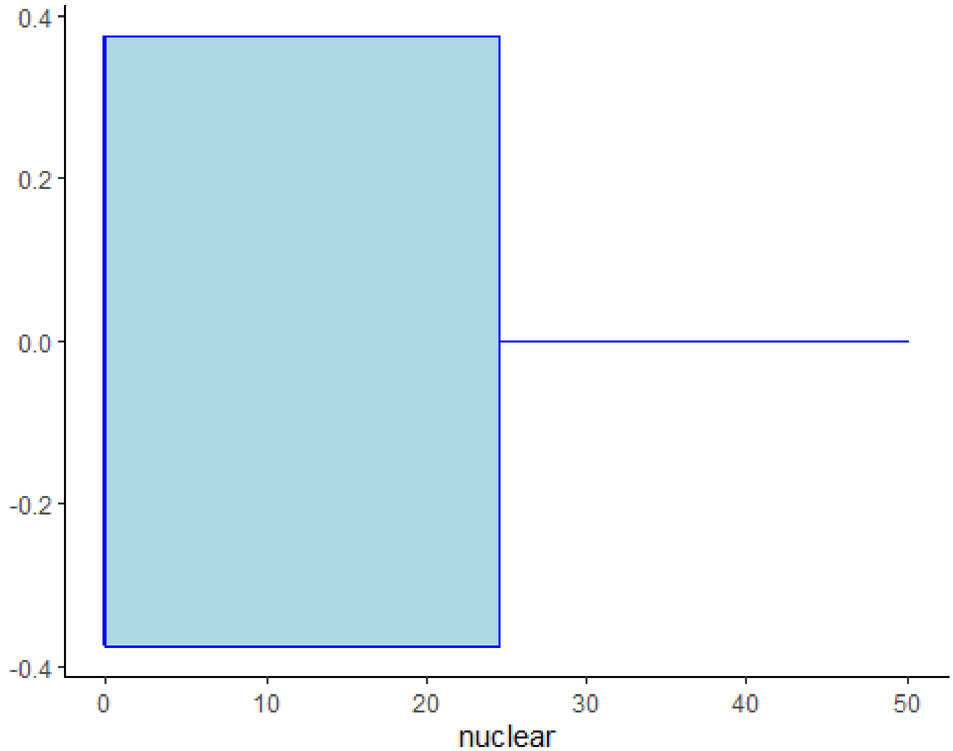

```
ggplot(data = utilities.dt) +
  geom_boxplot(mapping = aes(y = demand_growth), color= "blue",
fill="lightblue", outlier.colour = "red")+
  coord_flip()
```



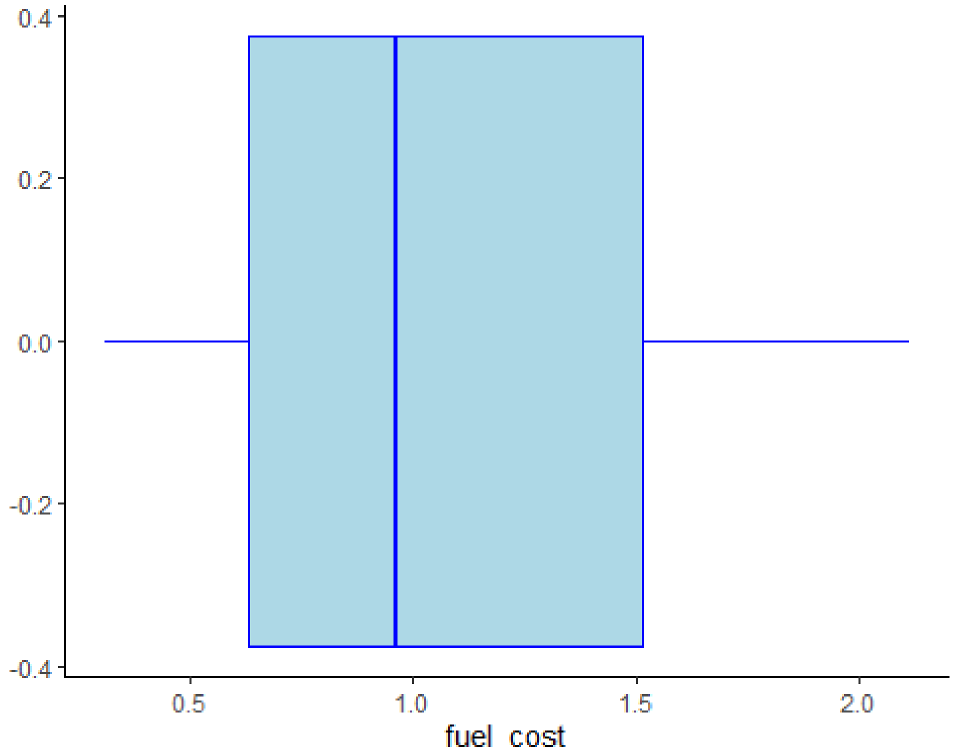

```
ggplot(data = utilities.dt) +
  geom_boxplot(mapping = aes(y = sales), color= "blue", fill="lightblue",
outlier.colour = "red")+
  coord_flip()
```




```
ggplot(data = utilities.dt) +
  geom_boxplot(mapping = aes(y = nuclear), color= "blue", fill="lightblue",
outlier.colour = "red")+
  coord_flip()
```




```
ggplot(data = utilities.dt) +
  geom_boxplot(mapping = aes(y = fuel_cost), color= "blue", fill="lightblue",
outlier.colour = "red")+
  coord_flip()
```

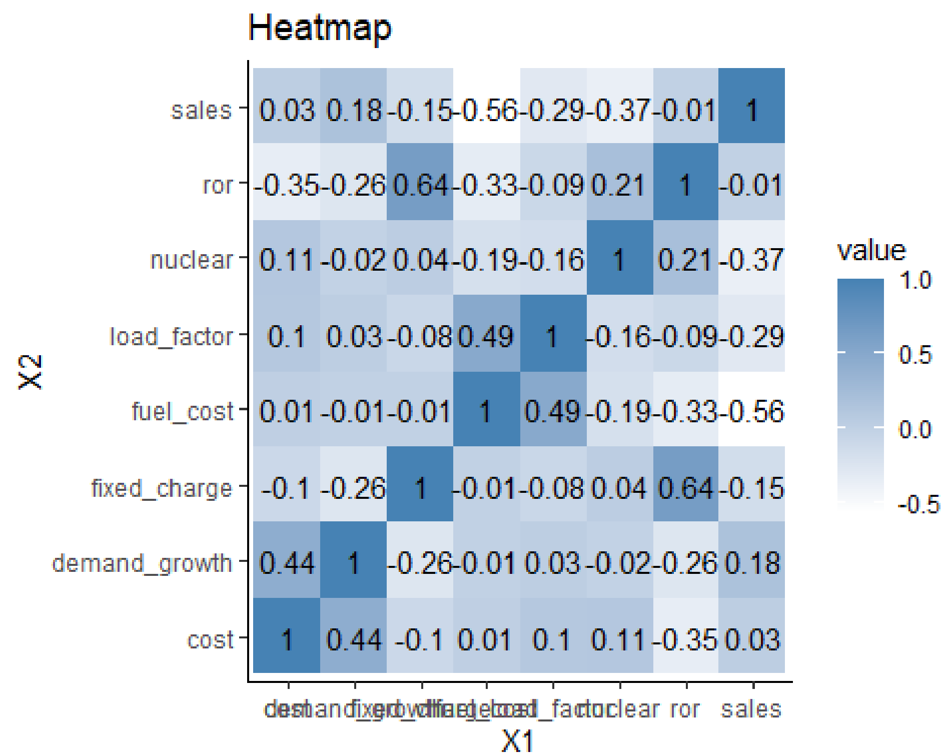



```
"Yes, there are extreme values for two variables named.
The two variables that have extreme values are 'Fixed_charge' and 'Sales'.
We can easily identify the extreme values in their respective boxplots as
they appear to be the outliers in 'Red' color."
```

```
## [1] "Yes, there are extreme values for two variables named.\nThe two
variables that have extreme values are 'Fixed_charge' and 'Sales'.\nWe can
easily identify the extreme values in their respective boxplots as they
appear to be the outliers in 'Red' color."
```

Question:3

```
cor.mat <- round(cor(utilities.dt[,-c("company")]),2)
melted.cor.mat <- melt(cor.mat)
ggplot(melted.cor.mat, aes(x = X1, y = X2, fill = value)) +
  scale_fill_gradient(low="white", high="steelblue") +
  geom_tile() +
  geom_text(aes(x = X1, y = X2, label = value)) +
  ggtitle("Heatmap")
```




```
"From heatmap we get to see the correlation between all variables such that
the darker value represents that the correlation is high and the lighter
value represents that the correlation is low.
From the heatmap, above we can see that the variables 'ROR' and
'FIXED_CHARGE' have the highest positive correlation of '0.64' and the
numbers above and below the diagonal are symmetrical."
```

```
## [1] "From heatmap we get to see the correlation between all variables such
that the darker value represents that the correlation is high and the lighter
value represents that the correlation is low.\nFrom the heatmap, above we can
see that the variables 'ROR' and 'FIXED_CHARGE' have the highest positive
correlation of '0.64' and the numbers above and below the diagonal are
symmetrical."
```

Question:4

```
str(utilities.df)
```

```
## Classes 'data.table' and 'data.frame':   22 obs. of  9 variables:
##  $ company      : chr  "Arizona" "Boston" "Central" "Commonwealth" ...
##  $ 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>
```

```r
"From the above summary, we know that we have to exclude the categorical data
that is present in the 'company' variable and perform principal component
analysis on numerical data only."
```

```
## [1] "From the above summary, we know that we have to exclude the
categorical data that is present in the 'company' variable and perform
principal component analysis on numerical data only."
```

```r
pcs <- prcomp(na.omit(utilities.df[,-c("company")]))
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
"From the above importance of components, we can see that 8 principal
components have been generated from PC1 to PC8 capturing a lot more
information than it actually had by showing '0.9998' proportion of variation
for PC1.
For instance, if we are satisfied with 90% criteria then we can use only one
or two principal component instead of 8 which will lead to savings cost in
terms of  time and effort needed to run this model. "
```

```
## [1] "From the above importance of components, we can see that 8 principal
components have been generated from PC1 to PC8 capturing a lot more
information than it actually had by showing '0.9998' proportion of variation
for PC1. \nFor instance, if we are satisfied with 90% criteria then we can
use only one or two principal component instead of 8 which will lead to
savings cost in terms of  time and effort needed to run this model. "
```

```r
pcs$rot # rotation matrix
```

```
##                         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
```

```
"By analyzing the weighted averages from the above results, in PC1, the
highest absolute weight is of 'Sales' followed by 'Fixed_charge' and 'Ror'."
```

```
## [1] "By analyzing the weighted averages from the above results, in PC1,
the highest absolute weight is of 'Sales' followed by 'Fixed_charge' and
'Ror'."
```

```
scores <- pcs$x
head(scores, 5)
```

```
##            PC1        PC2        PC3        PC4        PC5        PC6
## [1,] -162.9706 -17.935305 -10.457202 -3.4516417  0.6541231  1.4687225
## [2,] 3826.0520  35.346864   4.528675 -0.4991446  1.5241699  0.3204650
## [3,] -297.9588 -55.942181  -7.692747 -3.8009945 -1.2502127 -4.4094109
## [4,] 2491.0808   1.567099  17.950500 -0.3207562  2.4246537  0.7508999
## [5,] 5614.0329  25.109768  -7.059965 -8.9352191  1.4984451  1.3031240
##             PC7         PC8
## [1,]  0.6050278  0.05418517
## [2,] -0.1832258 -0.27374180
## [3,] -0.2301224  0.01661941
## [4,]  0.3362190 -0.08815268
## [5,] -0.5610272  0.28387954
```

Question:5

```
pcs.cor <- prcomp(na.omit(utilities.df[,-c("company")]), 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
```

```
"After scaling the numerical variables, from the above importance of
components we can see that 8 principal components have been generated from
PC1 to PC8 capturing a lot more information than it actually had by showing
'0.2716' proportion of variation for PC1.
For instance, if we are satisfied with 90% criteria then we can use only
principal component from PC1 to PC6 instead of all 8 principal components
which will lead to savings cost in terms of  time and effort needed to run
this model. "
```

```
## [1] "After scaling the numerical variables, from the above importance of
components we can see that 8 principal components have been generated from
PC1 to PC8 capturing a lot more information than it actually had by showing
'0.2716' proportion of variation for PC1. \nFor instance, if we are satisfied
with 90% criteria then we can use only principal component from PC1 to PC6
instead of all 8 principal components which will lead to savings cost in
terms of  time and effort needed to run this model. "
```

```
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
```

```
"By analyzing the weighted averages from the above results, we can see that
the interpretation of the results has changed since we performed scaling on
the numerical values.
In PC1, the highest absolute weight has changed from 'Sales' followed by
'Fixed_charge' and 'Ror' to 'ROR' followed by 'FIXED_CHARGE'
It is noticeable that Sales is no longer predominant as it was before."
```

```
## [1] "By analyzing the weighted averages from the above results, we can see
that the interpretation of the results has changed since we performed scaling
on the numerical values. \nIn PC1, the highest absolute weight has changed
from 'Sales' followed by 'Fixed_charge' and 'Ror' to 'ROR' followed by
'FIXED_CHARGE'\nIt is noticeable that Sales is no longer predominant as it
was before."
```

```
scores1 <- pcs$x
head(scores1, 5)
```

```
##              PC1        PC2        PC3        PC4        PC5        PC6
## [1,] -162.9706 -17.935305 -10.457202 -3.4516417  0.6541231  1.4687225
## [2,] 3826.0520  35.346864   4.528675 -0.4991446  1.5241699  0.3204650
## [3,] -297.9588 -55.942181  -7.692747 -3.8009945 -1.2502127 -4.4094109
## [4,] 2491.0808   1.567099  17.950500 -0.3207562  2.4246537  0.7508999
## [5,] 5614.0329  25.109768  -7.059965 -8.9352191  1.4984451  1.3031240
##               PC7         PC8
## [1,]  0.6050278  0.05418517
## [2,] -0.1832258 -0.27374180
## [3,] -0.2301224  0.01661941
## [4,]  0.3362190 -0.08815268
## [5,] -0.5610272  0.28387954
```