ADVANCE STATISTICS GROUP ASSIGNMENT

Karpagam Sivaprakasam

Nikhil Kumar

Sahil Mattoo

Manoharan A

Question 1

Problem Statement - 1

As part of a study of consumer consideration of ready-to-eat cereals sponsored by Kellogg Australia, Roberts and Lattin (1991) surveyed consumers regarding their perceptions of their favorite brands of cereals. Each respondent was asked to evaluate three preferred brands on each of 25 different attributes. Respondents used a five point likert scale to indicate the extent to which each brand possessed the given attribute.
For the purpose of this assignment, a subset of the data collected by Roberts and Lattin reflecting the evaluations of the 12 most frequently cirted cereal brands in the sample (in the original study, a total of 40 different brands were evaluated by 121 respondents, but the majority of brands were rated by only a small number of consumers). The 25 attributes and 12 brands are listed below

Cereal Brand	Attributes 1-12	Attributes 13-25
All Bran	Filling	Family
Cerola Muesli	Natural	Calories
Just Right	Fibre	Plain
Kellogg's corn falkes	Sweet	Crisp
Komplete	Easy	Regular
Nutrigrain	Salt	Sugar
Purina Muesli	Satisfying	Fruit
Rice Bubbles	Energy	Process
Special K	Fun	Quality
Sustain Kids	Treat	
Vitabrit	Soggy	Boring
Weetbix	Economical	Nutritious
Health		

In total 116 respondents provided 235 observations of the 12 selected brands. How do you characterize the consideration behavior of the 12 selected brands? Analyze and interpret your results using factor analysis.

Approach

Interpretation of the Problem Statement

Kelloggs wants to study the cereal market in order to understand how different attributes are behaving and which attributes are similar in nature.

Approach

1. Exploratory data analysis
2. Study correlation among the variables to confirm if correlation exists
3. Perform dimension reduction technique using Principal Component Analysis to prioritize number of components
4. Categorize and name the components based on the their homogeneity
5. Perform Factor Analysis with the prioritized components
6. Derive scores and interpret the results

1. Exploratory Data Analysis

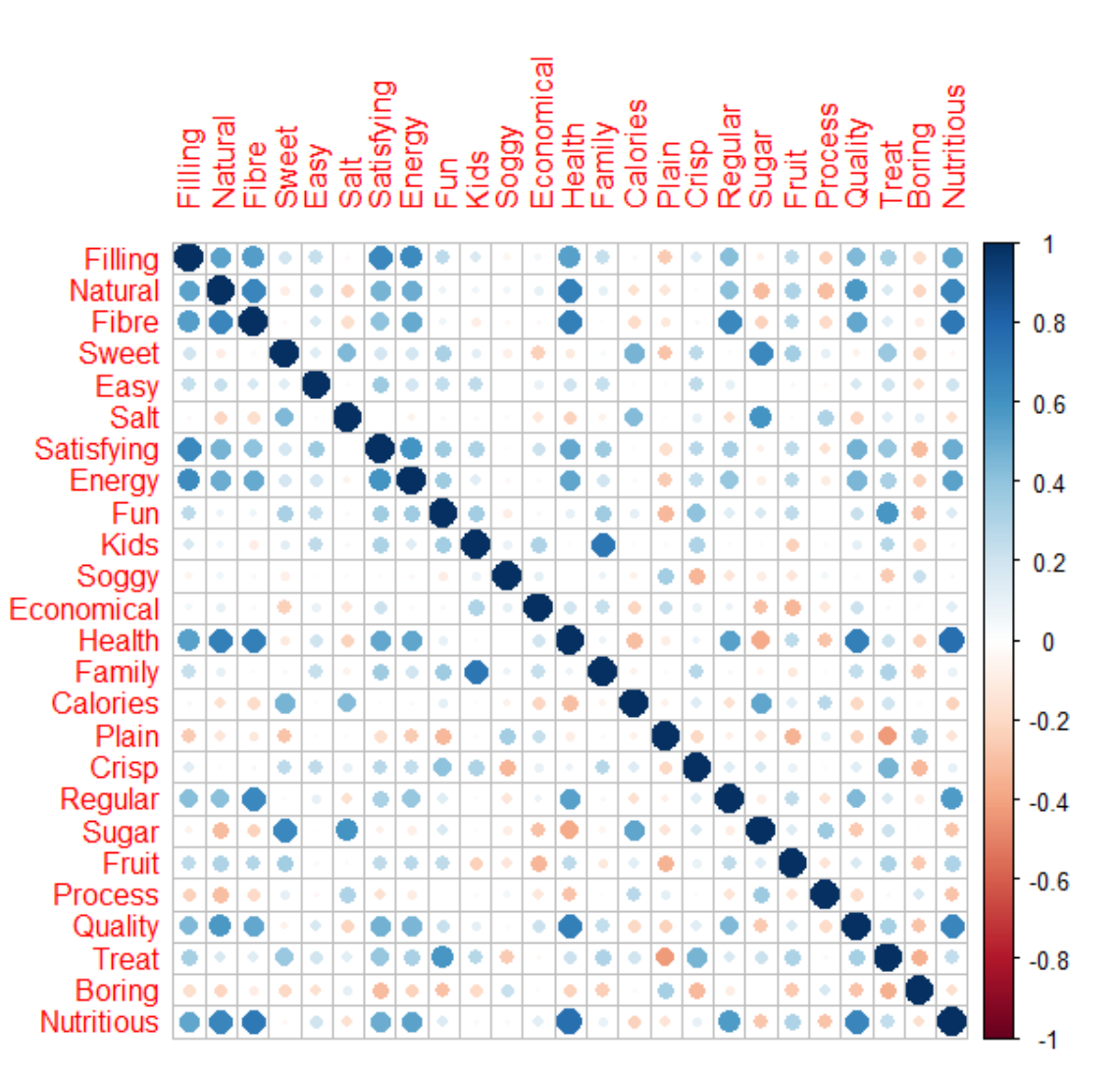
```
> summary(cereal)
       Cereals      Filling          Natural          Fibre           Sweet            Easy            Salt      
 CornFlakes :27   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000  
 Weetabix   :27   1st Qu.:3.000   1st Qu.:3.000   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:4.000   1st Qu.:1.000  
 Vitabrit   :25   Median :4.000   Median :4.000   Median :4.000   Median :2.000   Median :5.000   Median :2.000  
 NutriGrain :24   Mean   :3.881   Mean   :3.783   Mean   :3.528   Mean   :2.506   Mean   :4.532   Mean   :1.991  
 SpecialK   :23   3rd Qu.:4.500   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:3.000   3rd Qu.:5.000   3rd Qu.:3.000  
 RiceBubbles:21   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :6.000   Max.   :4.000  
 (Other)    :88                                                                                                  
   Satisfying        Energy           Fun             Kids           Soggy         Economical        Health     
 Min.   :2.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:3.000   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:3.000   1st Qu.:1.000   1st Qu.:3.000   1st Qu.:3.000  
 Median :4.000   Median :4.000   Median :2.000   Median :4.000   Median :2.000   Median :3.000   Median :4.000  
 Mean   :4.004   Mean   :3.643   Mean   :2.617   Mean   :3.843   Mean   :2.255   Mean   :3.217   Mean   :3.809  
 3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:3.000   3rd Qu.:5.000   3rd Qu.:3.000   3rd Qu.:4.000   3rd Qu.:4.000  
 Max.   :6.000   Max.   :5.000   Max.   :5.000   Max.   :6.000   Max.   :5.000   Max.   :5.000   Max.   :5.000  

     Family          Calories         Plain           Crisp          Regular          Sugar           Fruit      
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000  
 Median :4.000   Median :3.000   Median :2.000   Median :3.000   Median :3.000   Median :2.000   Median :1.000  
 Mean   :3.877   Mean   :2.702   Mean   :2.268   Mean   :3.204   Mean   :3.072   Mean   :2.145   Mean   :1.694  
 3rd Qu.:5.000   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:3.000   3rd Qu.:3.000  
 Max.   :6.000   Max.   :5.000   Max.   :5.000   Max.   :6.000   Max.   :5.000   Max.   :5.000   Max.   :5.000  

     Process         Quality          Treat           Boring        Nutritious    
 Min.   :1.000   Min.   :1.000   Min.   :1.00   Min.   :1.00   Min.   :1.000  
 1st Qu.:2.000   1st Qu.:3.000   1st Qu.:2.00   1st Qu.:1.00   1st Qu.:3.000  
 Median :3.000   Median :4.000   Median :3.00   Median :2.00   Median :4.000  
 Mean   :2.936   Mean   :3.694   Mean   :2.63   Mean   :1.83   Mean   :3.664  
 3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:3.00   3rd Qu.:2.00   3rd Qu.:4.000  
 Max.   :6.000   Max.   :5.000   Max.   :6.00   Max.   :5.00   Max.   :5.000  
```

- As per the survey, the scale used is Likert scale of 1-5, however there were responses with a rating of 6.
- Considered the rating of 6 as 5 in our study

Exploratory Data Analysis After Data Cleaning

```
> summary(data2)
       Cereals        Filling          Natural          Fibre           Sweet            Easy            Salt      
 CornFlakes :27   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000  
 Weetabix   :27   1st Qu.:3.000   1st Qu.:3.000   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:4.000   1st Qu.:1.000  
 Vitabrit   :25   Median :4.000   Median :4.000   Median :4.000   Median :2.000   Median :5.000   Median :2.000  
 NutriGrain :24   Mean   :3.881   Mean   :3.783   Mean   :3.528   Mean   :2.506   Mean   :4.528   Mean   :1.991  
 SpecialK   :23   3rd Qu.:4.500   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:3.000   3rd Qu.:5.000   3rd Qu.:3.000  
 RiceBubbles:21   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :4.000  
 (Other)    :88                                                                                                  
   Satisfying        Energy           Fun             Kids           Soggy         Economical        Health     
 Min.   :2     Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:3     1st Qu.:3.000   1st Qu.:2.000   1st Qu.:3.000   1st Qu.:1.000   1st Qu.:3.000   1st Qu.:3.000  
 Median :4     Median :4.000   Median :2.000   Median :4.000   Median :2.000   Median :3.000   Median :4.000  
 Mean   :4     Mean   :3.643   Mean   :2.617   Mean   :3.838   Mean   :2.255   Mean   :3.217   Mean   :3.809  
 3rd Qu.:5     3rd Qu.:4.000   3rd Qu.:3.000   3rd Qu.:5.000   3rd Qu.:3.000   3rd Qu.:4.000   3rd Qu.:4.000  
 Max.   :5     Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000  

     Family          Calories         Plain           Crisp          Regular          Sugar           Fruit      
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.0   Min.   :1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:2.0   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000  
 Median :4.000   Median :3.000   Median :2.000   Median :3.0   Median :3.000   Median :2.000   Median :1.000  
 Mean   :3.872   Mean   :2.702   Mean   :2.268   Mean   :3.2   Mean   :3.072   Mean   :2.145   Mean   :1.694  
 3rd Qu.:5.000   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:4.0   3rd Qu.:4.000   3rd Qu.:3.000   3rd Qu.:3.000  
 Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.0   Max.   :5.000   Max.   :5.000   Max.   :5.000  

     Process         Quality          Treat          Boring        Nutritious    
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.00   Min.   :1.000  
 1st Qu.:2.000   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.00   1st Qu.:3.000  
 Median :3.000   Median :4.000   Median :3.000   Median :2.00   Median :4.000  
 Mean   :2.932   Mean   :3.694   Mean   :2.626   Mean   :1.83   Mean   :3.664  
 3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:3.000   3rd Qu.:2.00   3rd Qu.:4.000  
 Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.00   Max.   :5.000  
```

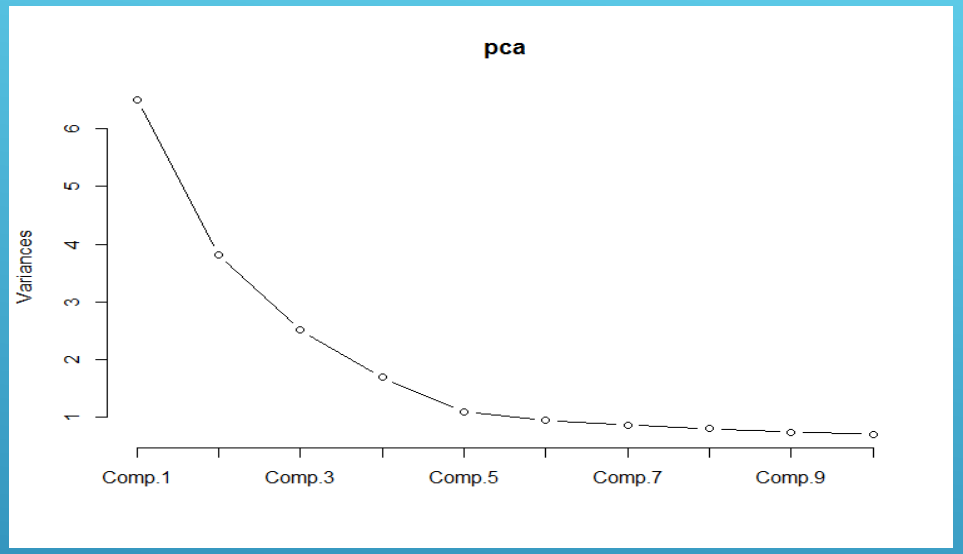
2. Correlation Matrix

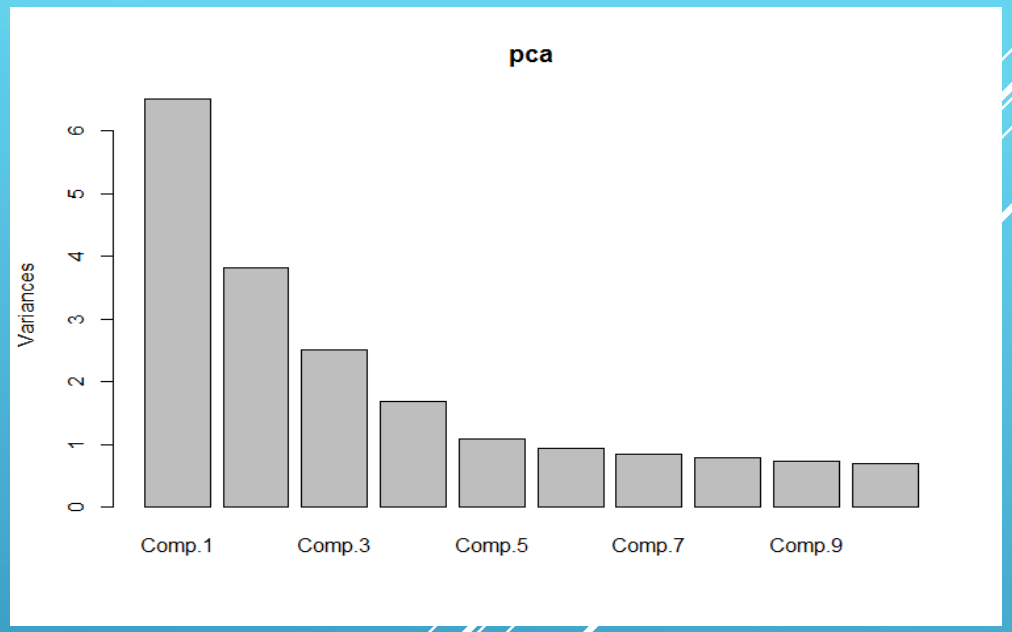


- Correlation matrix is done to check the correlation among the variables.
- We see medium and strong correlation among few variables

  Examples
  - Filling has medium correlation with Natural and Fibre while strong correlation Satisfying and Energy
  - Health has medium correlation with Filling and Health while strong correlation Quality and Nutritious
- Since there is medium and strong correlation among the variables, there is a scope to reduce the number variables by performing Principal Component Analysis (PCA)

3. Principal Component Analysis






```
> summary(PCA)
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6    Comp.7    Comp.8    Comp.9
Standard deviation     2.5515645 1.9473508 1.57931249 1.29699429 1.04196617 0.97215571 0.92371682 0.88941254 0.85594266
Proportion of Variance 0.2604193 0.1516870 0.09976912 0.06728777 0.04342774 0.03780347 0.03413011 0.03164219 0.02930551
Cumulative Proportion  0.2604193 0.4121063 0.51187538 0.57916315 0.62259089 0.66039436 0.69452447 0.72616666 0.75547217
                         Comp.10   Comp.11   Comp.12   Comp.13   Comp.14   Comp.15   Comp.16   Comp.17   Comp.18
Standard deviation     0.83528812 0.80508009 0.74210795 0.72900837 0.69819277 0.6456121 0.62203549 0.60340597 0.60072703
Proportion of Variance 0.02790825 0.02592616 0.02202897 0.02125813 0.01949893 0.0166726 0.01547713 0.01456395 0.01443492
Cumulative Proportion  0.78338042 0.80930658 0.83133555 0.85259368 0.87209260 0.8887652 0.90424233 0.91880628 0.93324119
                         Comp.19   Comp.20   Comp.21   Comp.22   Comp.23   Comp.24   Comp.25
Standard deviation     0.55329582 0.52496345 0.51267066 0.492791255 0.467311607 0.445682156 0.405660102
Proportion of Variance 0.01224545 0.01102346 0.01051325 0.009713729 0.008735206 0.007945303 0.006582405
Cumulative Proportion  0.94548664 0.95651011 0.96702336 0.976737086 0.985472292 0.993417595 1.000000000
>
```

- As per the Scree Plot, 4 components are sufficient to describe the data set, but as per Kaiser rule, 5th component is also a qualifier.
- However if cumulative variance is considered to be greater than 75%, then 9 components need to be evaluated.
- The number of factors that can best describe the data can be concluded post PCA & FA

PCA Code and Output

PCA

```
PCAval <- princomp(~PCAmat, scores = TRUE, cor=TRUE)
summary(PCAval)
loadings(PCAval)
Factors <- loadings(PCAval)
Factors
With Rotation
Rot_var <- varimax(loadings(PCAval))
Rot_Pro <- promax(loadings(PCAval))
```

- PCA analysis without rotation doesn't provide any satisfactory results which can help us in making the inference.
- Rotation Varimax and Promax too results in the similar kind of output.
- Based upon the results of PCA, we would run FA with component value ranging 4 to 9

```
Loadings:
                  Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10 Comp.11
PCAmatEasy                                      -0.64866 -0.49238
PCAmatCalories
PCAmatEnergy
PCAmatFibre
PCAmatHealth
PCAmatFilling
PCAmatNatural
PCAmatSweet
PCAmatSalt
PCAmatSatisfying
PCAmatFun
PCAmatKids                -0.49248
PCAmatSoggy                       0.44218 0.48808
PCAmatEconomical                                  0.43902
PCAmatFamily              -0.45527
PCAmatPlain                                                                       0.45643
PCAmatCrisp                       -0.40022
PCAmatRegular
PCAmatSugar
PCAmatFruit
PCAmatProcess                                   0.42457                 -0.47243
PCAmatQuality
PCAmatTreat
PCAmatBoring
PCAmatNutritious
```

4. Factor Analysis – Without Rotation

```
FAmat <- as.matrix(cerealdata[, 2:26])
FAmat
FAFactor <- factanal(~FAmat, 5, rotation="none")
FAFactor
```

```
Uniquenesses:
   FAmatFilling     FAmatNatural      FAmatFibre      FAmatSweet       FAmatEasy       FAmatSalt FAmatSatisfying
          0.283            0.389           0.311           0.361           0.838           0.513           0.373
    FAmatEnergy        FAmatFun       FAmatKids      FAmatSoggy FAmatEconomical     FAmatHealth     FAmatFamily
          0.432            0.523           0.240           0.775           0.705           0.213           0.348
 FAmatCalories      FAmatPlain      FAmatCrisp    FAmatRegular      FAmatSugar      FAmatFruit    FAmatProcess
          0.578            0.547           0.638           0.552           0.203           0.561           0.759
   FAmatQuality      FAmatTreat     FAmatBoring FAmatNutritious
          0.389            0.386           0.674           0.242
```

```
Loadings:
                 Factor1 Factor2 Factor3 Factor4 Factor5
FAmatFilling      0.720   0.234           0.143  -0.344
FAmatNatural      0.763  -0.123
FAmatFibre        0.756  -0.127  -0.281   0.137
FAmatSweet               0.749  -0.265
FAmatEasy         0.292   0.205   0.168
FAmatSalt        -0.253   0.495  -0.222   0.345
FAmatSatisfying   0.687   0.305   0.129          -0.188
FAmatEnergy       0.683   0.229                  -0.206
FAmatFun          0.291   0.537   0.179  -0.267
FAmatKids         0.168   0.390   0.749   0.133
FAmatSoggy               -0.173   0.148   0.412
FAmatEconomical   0.182  -0.160   0.449   0.142   0.118
FAmatHealth       0.853  -0.185  -0.105           0.112
FAmatFamily       0.263   0.334   0.685
FAmatCalories    -0.233   0.533  -0.220   0.133  -0.132
FAmatPlain       -0.233  -0.344   0.200   0.481
FAmatCrisp        0.201   0.462   0.164  -0.228   0.171
FAmatRegular      0.605          -0.194           0.192
FAmatSugar       -0.350   0.709  -0.341   0.185   0.145
FAmatFruit        0.326   0.243  -0.455  -0.246
FAmatProcess     -0.327   0.241           0.200   0.185
FAmatQuality      0.747                           0.211
FAmatTreat        0.371   0.595          -0.322   0.127
FAmatBoring      -0.314  -0.281  -0.102   0.371
FAmatNutritious   0.833          -0.138           0.169
```

Factor 1 - Filling, Natural, Fibre, Satisfying , Energy, Health, regular, quality and Nutritious
Factor 2 - sweet, salt, fun, calories, crisp, sugar, treat
Factor 3 - Kids, family, economical, fruit
Factor 4 - Soggy, plain

- Since we don't find homogeneity within the groups to come out with a common label, we need to perform Factor Analysis with rotation using Varimax

4. Factor Analysis – With Rotation using Varimax

VARIMAX

Loadings:

	Factor1	Factor2	Factor3	Factor4	Factor5
FAmatFilling	0.647		0.190	0.144	0.487
FAmatNatural	0.731	-0.215			0.153
FAmatFibre	0.816				
FAmatSweet		0.696		0.351	0.166
FAmatEasy	0.230		0.307		
FAmatSalt		0.689			
FAmatSatisfying	0.570		0.387	0.199	0.333
FAmatEnergy	0.611		0.168	0.225	0.339
FAmatFun	0.125	0.155	0.377	0.538	
FAmatKids			0.867		
FAmatSoggy			0.130	-0.454	
FAmatEconomical		-0.258	0.409	-0.197	-0.110
FAmatHealth	0.840	-0.271			
FAmatFamily			0.794	0.122	
FAmatCalories	-0.155	0.592		0.122	0.179
FAmatPlain	-0.115			-0.638	-0.150
FAmatCrisp		0.157	0.335	0.459	
FAmatRegular	0.657				
FAmatSugar	-0.177	0.852		0.170	
FAmatFruit	0.341	0.161	-0.284	0.439	0.152
FAmatProcess	-0.214	0.387		-0.101	-0.184
FAmatQuality	0.681	-0.222	0.200	0.218	-0.102
FAmatTreat	0.234	0.216	0.299	0.650	
FAmatBoring	-0.150		-0.198	-0.508	
FAmatNutritious	0.849	-0.154			

Factor 1 - Filling, Natural, Fibre, Satisfying , Energy, Health, regular, quality and Nutritious
Factor 2 - sweet, salt, calories, sugar
Factor 3 - Kids, economical, family,
Factor 4 – Fun, Soggy, plain, crisp, Fruit, Treat, Boring

- Since we don't find homogeneity within the factors to come out with a common label, we need to perform Factor Analysis with rotation using Promax

4. Factor Analysis – With Rotation using Promax

Loadings:

	Factor1	Factor2	Factor3	Factor4	Factor5
FAmatFilling	0.647		0.190	0.144	0.487
FAmatNatural	0.731	-0.215			0.153
FAmatFibre	0.816				
FAmatSweet		0.696		0.351	0.166
FAmatEasy	0.230		0.307		
FAmatSalt		0.689			
FAmatSatisfying	0.570		0.387	0.199	0.333
FAmatEnergy	0.611		0.168	0.225	0.339
FAmatFun	0.125	0.155	0.377	0.538	
FAmatKids			0.867		
FAmatSoggy			0.130	-0.454	
FAmatEconomical		-0.258	0.409	-0.197	-0.110
FAmatHealth	0.840	-0.271			
FAmatFamily			0.794	0.122	
FAmatCalories	-0.155	0.592		0.122	0.179
FAmatPlain	-0.115			-0.638	-0.150
FAmatCrisp		0.157	0.335	0.459	
FAmatRegular	0.657				
FAmatSugar	-0.177	0.852		0.170	
FAmatFruit	0.341	0.161	-0.284	0.439	0.152
FAmatProcess	-0.214	0.387		-0.101	-0.184
FAmatQuality	0.681	-0.222	0.200	0.218	-0.102
FAmatTreat	0.234	0.216	0.299	0.650	
FAmatBoring	-0.150		-0.198	-0.508	
FAmatNutritious	0.849	-0.154			

Factor 1 - Filling, Natural, Fibre, Satisfying , Energy, Health, regular, quality and Nutritious
Factor 2 - sweet, salt, calories, sugar, process
Factor 3 - Kids, economical, family,
Factor 4 – Fun, Soggy, plain, crisp, Fruit, Treat, Boring

We see homogeneity within the factorss after running FA with rotation using Promax

Labelling of the factors

Factor 1 – Nutrition Value
Factor 2 – Taste
Factor 3 – User Type
Factor 4 - Excitement

4. Factor Analysis – With Rotation using Promax (with cut-off 0.4)

```
Loadings:
                Factor1   Factor2   Factor3   Factor4   Factor5
FAmatNatural     0.653439
FAmatFibre       0.837381
FAmatHealth      0.838064
FAmatRegular     0.733334
FAmatQuality     0.680093
FAmatNutritious  0.886926
FAmatFun                   0.552021
FAmatSoggy                -0.541388
FAmatPlain                -0.682207
FAmatCrisp                 0.502230
FAmatTreat                 0.674285
FAmatBoring               -0.560899
FAmatSweet                           0.626955
FAmatSalt                            0.741083
FAmatCalories                        0.507131
FAmatSugar                           0.869987
FAmatKids                                      0.881979
FAmatFamily                                    0.788466
FAmatFilling     0.400957                                 0.625343
FAmatEasy
FAmatSatisfying                                          0.434479
FAmatEnergy      0.416390                                 0.429693
FAmatEconomical                                0.422162
FAmatFruit                 0.405592
FAmatProcess                         0.450030
```

FA Analysis - Cut off

print(FAFactor3, digits=6, cutoff= .4 , sort= TRUE)

Factor 1 - Filling, Natural, Fibre, Satisfying , Energy, Health, regular, quality and Nutritious
Factor 2 - sweet, salt, calories, sugar, process
Factor 3 - Kids, economical, family,
Factor 4 – Fun, Soggy, plain, crisp, Fruit, Treat, Boring

Labelling

Factor 1 – Nutrition Value
Factor 2 – Taste
Factor 3 – User Type
Factor 4 - Excitement

4. Factors (Components) with Scoring

Scoring without Factor Labels

	Factor1	Factor2	Factor3	Factor4
FAmatFilling	0.71981240	0.23418809	-0.06925688	0.14282066
FAmatNatural	0.76265744	-0.12325892	-0.07813895	0.07535841
FAmatFibre	0.75594968	-0.12722622	-0.28133342	0.13694656
FAmatSweet	-0.02266113	0.74922663	-0.26543144	0.05403151
FAmatEasy	0.29153976	0.20508282	0.16767988	0.08293966
FAmatSalt	-0.25305078	0.49502325	-0.22223829	0.34471033
FAmatSatisfying	0.68715120	0.30523634	0.12893705	0.09743554
FAmatEnergy	0.68289915	0.22908256	-0.07414822	0.04243887
FAmatFun	0.29123276	0.53717677	0.17856122	-0.26673581
FAmatKids	0.16760434	0.38984611	0.74881743	0.13303639
FAmatSoggy	-0.03960017	-0.17305840	0.14840864	0.41190659
FAmatEconomical	0.18234491	-0.16032096	0.44891677	0.14231222
FAmatHealth	0.85293568	-0.18482783	-0.10474494	0.04226010
FAmatFamily	0.26281173	0.33378109	0.68509738	0.04228800
FAmatCalories	-0.23270198	0.53341273	-0.21991628	0.13297924
FAmatPlain	-0.23346424	-0.34437612	0.20027545	0.48122621
FAmatCrisp	0.20093500	0.46183879	0.16444451	-0.22822149
FAmatRegular	0.60531739	-0.03368073	-0.19411175	0.07885232
FAmatSugar	-0.34962693	0.70899751	-0.34115233	0.18514336
FAmatFruit	0.32588885	0.24276458	-0.45507355	-0.24575519
FAmatProcess	-0.32688542	0.24071983	-0.04045862	0.20001976
FAmatQuality	0.74721539	-0.02648325	0.04173641	-0.07177571
FAmatTreat	0.37146753	0.59479704	0.04589608	-0.32228519
FAmatBoring	-0.31441606	-0.28091509	-0.10186355	0.37071395
FAmatNutritious	0.83338818	-0.09031295	-0.13772235	0.08883345

Scoring with Factor Labels

	Nutrition Value	Taste	User Type	Excitment
FAmatFilling	0.71981240	0.23418809	-0.06925688	0.14282066
FAmatNatural	0.76265744	-0.12325892	-0.07813895	0.07535841
FAmatFibre	0.75594968	-0.12722622	-0.28133342	0.13694656
FAmatSweet	-0.02266113	0.74922663	-0.26543144	0.05403151
FAmatEasy	0.29153976	0.20508282	0.16767988	0.08293966
FAmatSalt	-0.25305078	0.49502325	-0.22223829	0.34471033
FAmatSatisfying	0.68715120	0.30523634	0.12893705	0.09743554
FAmatEnergy	0.68289915	0.22908256	-0.07414822	0.04243887
FAmatFun	0.29123276	0.53717677	0.17856122	-0.26673581
FAmatKids	0.16760434	0.38984611	0.74881743	0.13303639
FAmatSoggy	-0.03960017	-0.17305840	0.14840864	0.41190659
FAmatEconomical	0.18234491	-0.16032096	0.44891677	0.14231222
FAmatHealth	0.85293568	-0.18482783	-0.10474494	0.04226010
FAmatFamily	0.26281173	0.33378109	0.68509738	0.04228800
FAmatCalories	-0.23270198	0.53341273	-0.21991628	0.13297924
FAmatPlain	-0.23346424	-0.34437612	0.20027545	0.48122621
FAmatCrisp	0.20093500	0.46183879	0.16444451	-0.22822149
FAmatRegular	0.60531739	-0.03368073	-0.19411175	0.07885232
FAmatSugar	-0.34962693	0.70899751	-0.34115233	0.18514336
FAmatFruit	0.32588885	0.24276458	-0.45507355	-0.24575519
FAmatProcess	-0.32688542	0.24071983	-0.04045862	0.20001976
FAmatQuality	0.74721539	-0.02648325	0.04173641	-0.07177571
FAmatTreat	0.37146753	0.59479704	0.04589608	-0.32228519
FAmatBoring	-0.31441606	-0.28091509	-0.10186355	0.37071395
FAmatNutritious	0.83338818	-0.09031295	-0.13772235	0.08883345

Question 2

Problem Statement - 2

Leslie Salt Data Set
In 1968, the city of Mountain View, California, began the necessary legal proceedings to acquire a parcel of land owned by the Leslie Sal Company. The Leslie property contained 246.8 acres and was located right on the San Francisco Bay. The land had been used for salt evaporation and had an elevation of exactly sea level. However, the property was diked so that the waters from the bay park were kept out. The city of Mountain View intended to fill the property and use it for a city park.
Ultimately, it fell into the courts to determine a fair market value for the property. Appraisers were hired, but what made the processes difficult was that there were few sales of byland property and none of them corresponded exactly to the characteristics of the Leslie property. The experts involved decided to build a regression model to better understand the factors that might influence market valuation. They collected data on 31 byland properties that were sold during the previous 10 years. In addition to the transaction price for each property, they collected data oina large number of other factors, including size, time of sale, elevation, location, and access to sewers. A listing of these data, including only those variables deemed relevant for this exercise. A description of the variables is provided below.

Problem Statement – 2 Contd.

Variable name	Description
Price	Sales price in $000 per acre
County	San Mateo=0, Santa Clara =1
Size	Size of the property in acres
Elevation	Average Elevation in foot above sea level
Sewer	Distance (in feet) to nearest sewer connection
Date	Date of sale counting backward from current time (in months)
Flood	Subject to flooding by tidal action =1; otherwise =0
Distance	Distance in miles from Leslie Property (in almost all cases, this is toward San Francisco

Discuss and Answer the following questions:

1. What is the nature of each of the variables? Which variable is dependent variable and what are the independent variables in the model?
2. Check whether the variables require any transformation individually
3. Set up a regression equation, run the model and discuss your results

1. What is the nature of each of the variables? Which variable is dependent variable and what are the independent variables in the model?

Nature of Variable and classification into dependent and independent

Variable Name	Description	Nature of Variable	Dependent / Independent
Price	Sales price in $000 per acre	Continuous	Dependent
Country	San Mateo=0, Santa Clara =1	Categorical	Independent
Size	Size of the property in acres	Continuous	Independent
Elevation	Average Elevation in foot above sea level	Continuous	Independent
Sewer	Distance (in feet) to nearest sewer connection	Continuous	Independent
Date	Date of sale counting backward from current time (in months)	Continuous	Independent
Flood	Subject to flooding by tidal action =1; otherwise =0	Categorical	Independent
Distance	Distance in miles from Leslie Property (in almost all cases, this is toward San Francisco	Continuous	Independent

Assumptions for Linear regression

#	Assumption	Validation stage	Method used
1	There must be a **linear relationship** between the outcome variable and the independent variables.	Before running regression	Scatter plot /correlation co-efficient
2	**Multivariate Normality** –Multiple regression assumes that the residuals are normally distributed.	After running regression as data on residual available only after running regression	Q-Q plot / Shapiro test
3	**No Multicollinearity** —Multiple regression assumes that the independent variables are not highly correlated with each other.	After running regression	Using Variance Inflation Factor (VIF)
4	**Homoscedasticity** –This assumption states that the variance of error terms are similar across the values of the independent variables	After running regression	Standardized residuals Vs Predicted values Plot

The above assumptions will be validated at the appropriate stages of the regression exercise as we progress

2. Check whether the variables require any transformation individually

Testing Normality of Variables using Shapiro Wilks test

Shapiro Wilks test is done to check the normality of both dependent and independent variables

shapiro.test(saltdata$Price)
data: saltdata$Price
W = 0.90607, p-value = **0.01025**

shapiro.test(saltdata$Size)
data: saltdata$Size
W = 0.40531, p-value = **4.108e-10**

shapiro.test(saltdata$Elevation)
data: saltdata$Elevation
W = 0.85914, p-value **= 0.000798**

shapiro.test(saltdata$Sewer)
data: saltdata$Sewer
W = 0.80027, p-value = **5.221e-05**

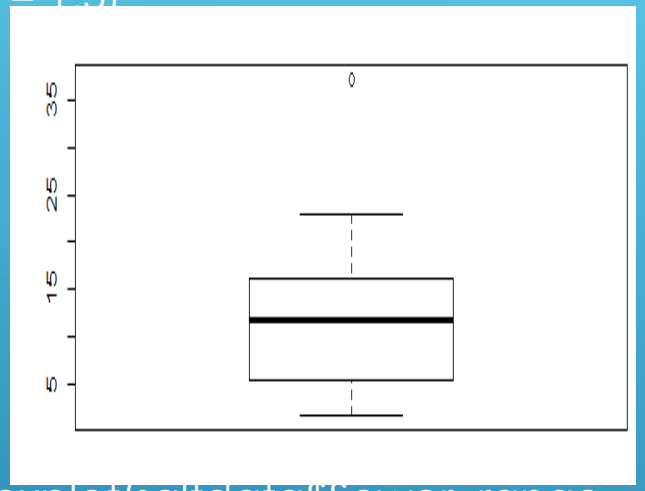
shapiro.test(saltdata$Date)
data: saltdata$Date
W = 0.91472, p-value **= 0.01714**

shapiro.test(saltdata$Distance)
data: saltdata$Distance
W = 0.87773, p-value **= 0.002096**

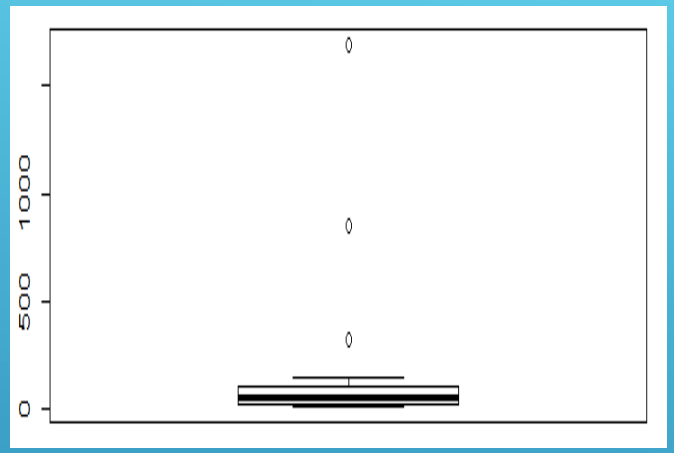
As the p value is < 0.05, none of the above variables follow normal distribution.

Checking for presence of outliers

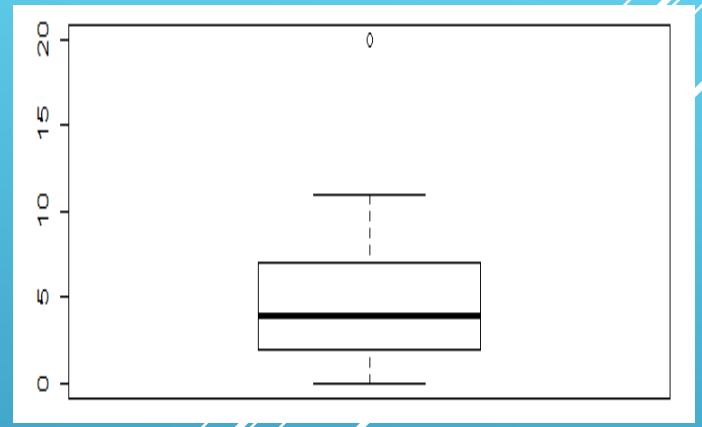
boxplot(saltdata$Price, range = 1.5)



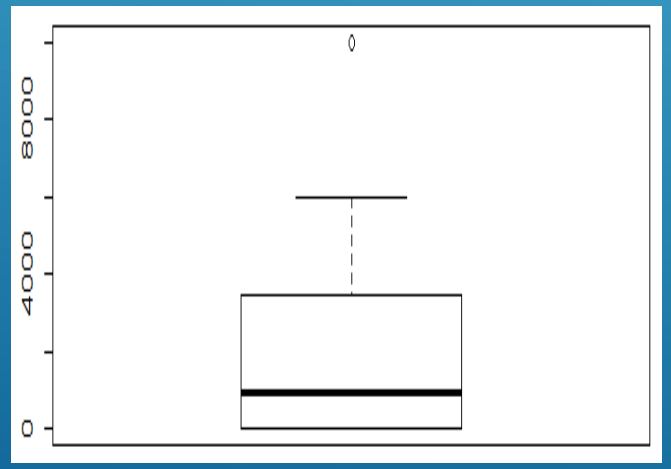
boxplot(saltdata$Size, range = 1.5)



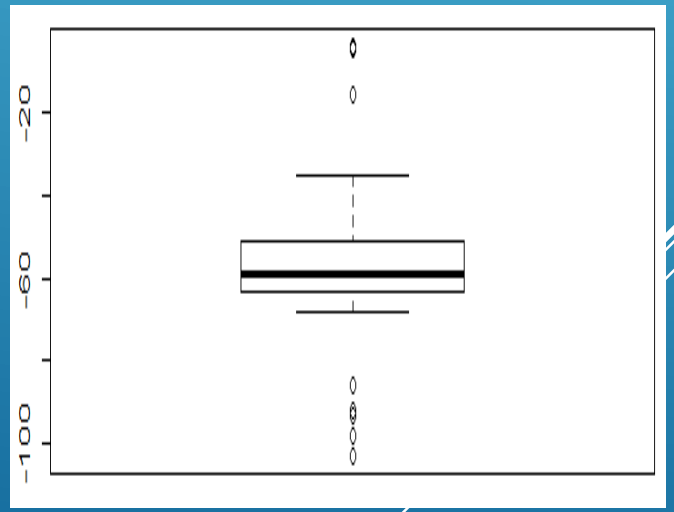
boxplot(saltdata$Elevation, range = 1.5)



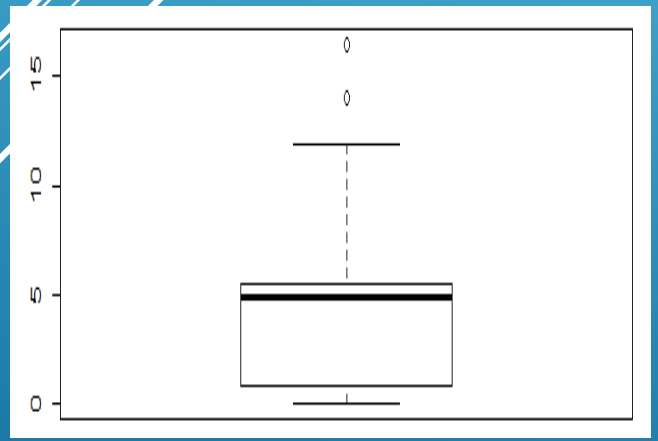
boxplot(saltdata$Sewer, range = 1.5)



boxplot(saltdata$Date,range = 1.5)



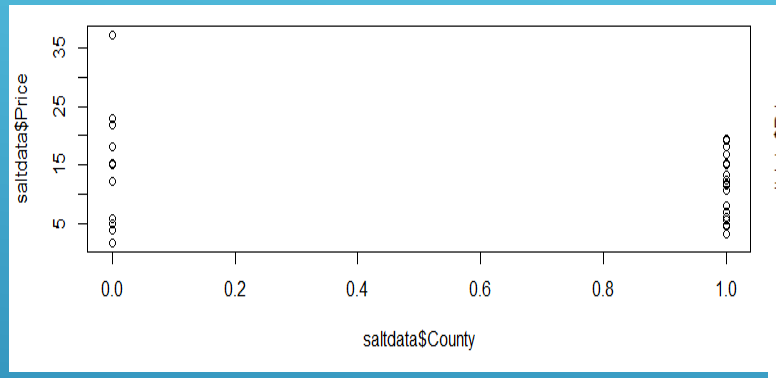
boxplot(saltdata$Distance,range = 1.5)

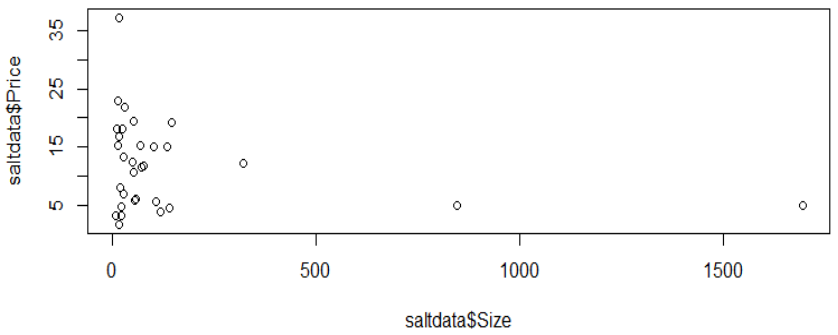


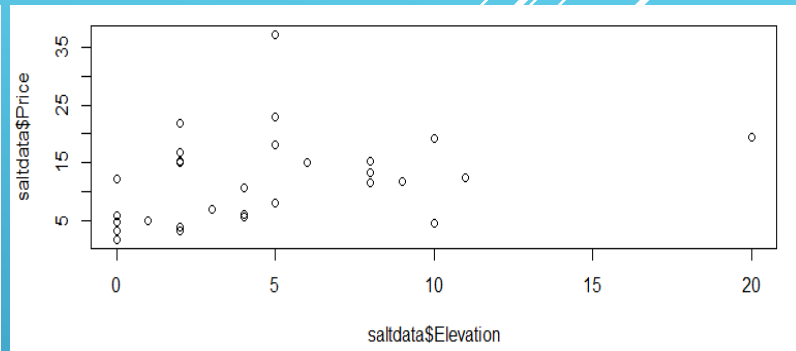
- Box plot shows all the continuous data variables are having many outliers, which are very far away from min/max points
- One of the ways to remediate non-normal data is to remove outliers. However, removing outliers will make

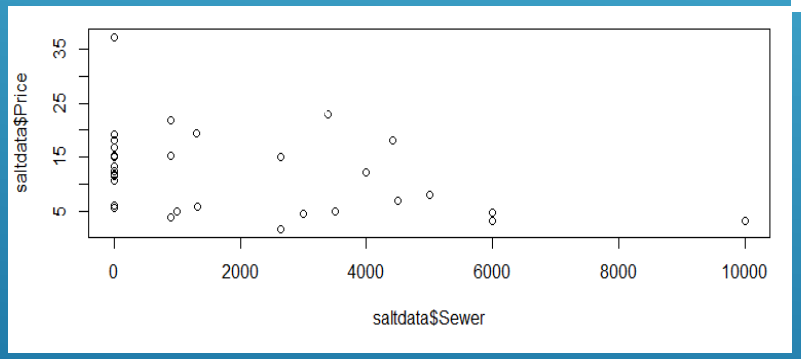
Check for linear relationship

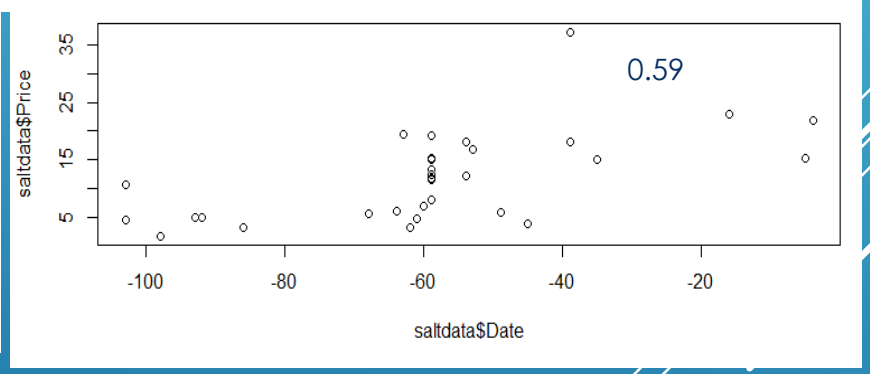
One of the assumptions for linear regression: There must be a **linear relationship** between the outcome variable and the independent variables. Checked the correlation between each independent variable with the dependent variable

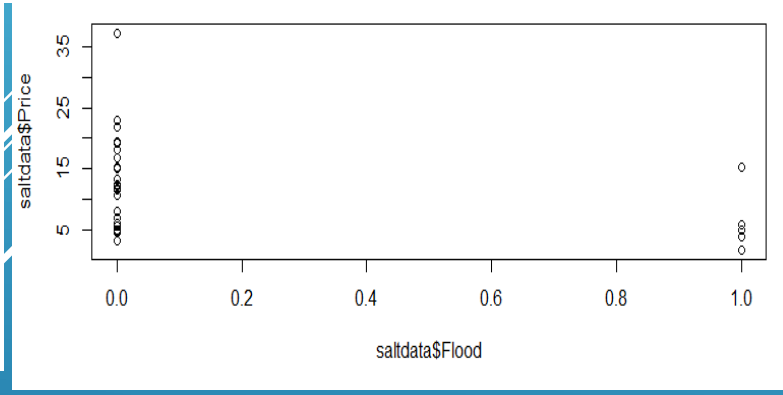


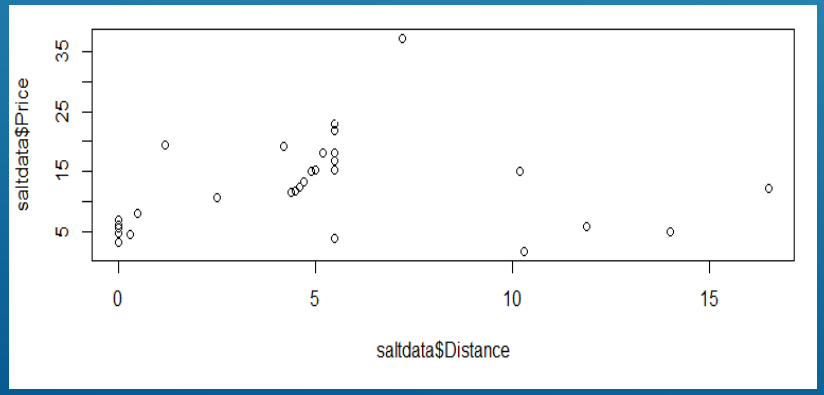












- Only 'Date' has medium positive correlation (coefficient : 0.59) while all other variables have very weak correlation

Summary of normality and linearity tests

- Shapiro Wilks test shows that both dependent and independent variables do not follow normal distribution
- Box plot shows all the continuous data variables are having many outliers, which are very far away from min/max points
- Correlation plots show the independent variables do not show linear relationship with the dependent variable

Hence there is a need to transform the data.
Both Dependent and Independent variables need to transformed.

3. Set up a regression equation, run the model and discuss your results

Regression without transforming the data

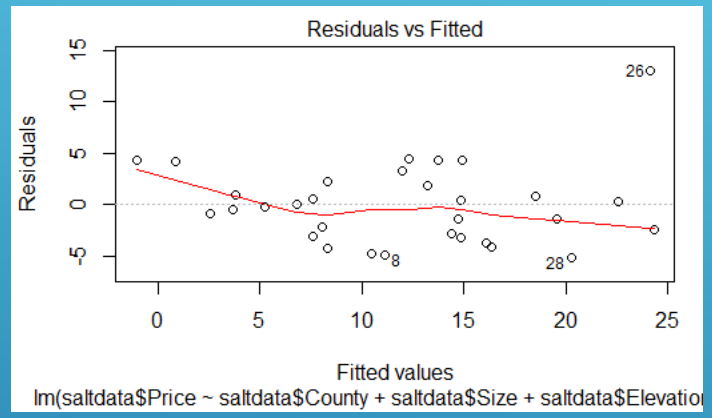
Regression is done considering all the independent variables without transforming the data, just to understand the outcome as to compare the outcome after transforming the data

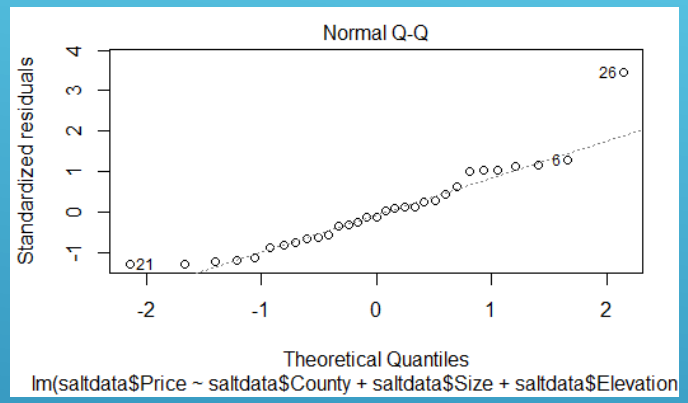
```
Residuals:
   Min     1Q Median     3Q    Max
-5.169 -2.957 -0.256  2.070 13.031

Coefficients:
                  Estimate    Std. Error    t value    Pr(>|t|)
(Intercept)      2.364e+01    3.829e+00      6.174    2.68e-06 ***
saltdata$County -8.789e+00    3.652e+00     -2.407    0.024532 *
saltdata$Size   -6.043e-03    3.501e-03     -1.726    0.097702 .
saltdata$Elevation 5.193e-01  2.386e-01      2.177    0.040030 *
saltdata$Sewer  -9.573e-04    4.169e-04     -2.296    0.031126 *
saltdata$Date    8.508e-02    4.865e-02      1.749    0.093646 .
saltdata$Flood  -1.202e+01    2.989e+00     -4.020    0.000536 ***
saltdata$Distance 1.858e-01   3.395e-01      0.547    0.589386
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.431 on 23 degrees of freedom
Multiple R-squared:  0.747, Adjusted R-squared:  0.67
F-statistic: 9.703 on 7 and 23 DF,  p-value: 1.351e-05
```

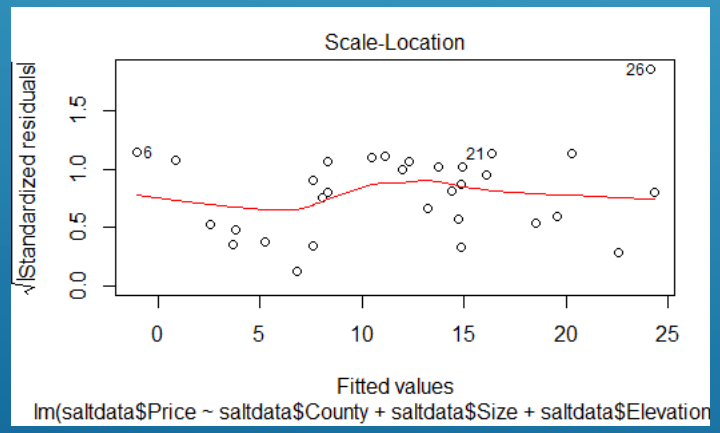
- Rsqr is 74.7% while Rsqr adjusted is 67%. The huge difference between the Rsqr and Rsqr adjusted shows that the Rsqr is inflated by multicolinearity among the independent variables
- Intercept, County, Elevation, Sewer and Flood are the significant variables

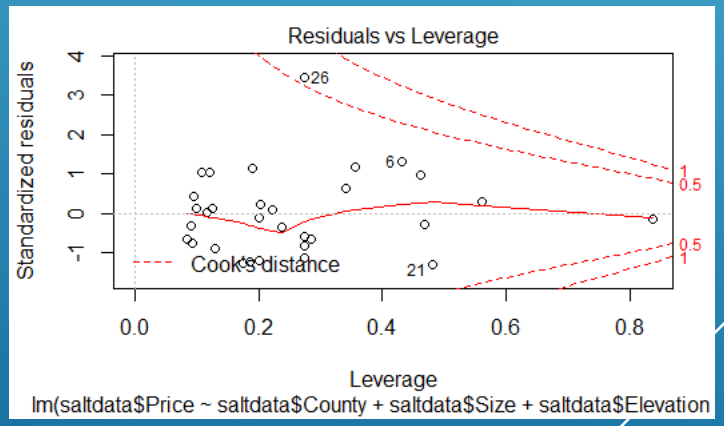
Regression without transforming the data
Validation of Assumptions – Residual Plots

Residual Plots









Variance Inflation Factor(VIF)

```
vif(mlr_wo_t)
  saltdata$County       saltdata$Size      saltdata$Elevation
saltdata$Sewer
        4.995597            2.003925            1.649759
1.635122

 saltdata$Date      saltdata$Flood      saltdata$Distance
 2.174889           1.907942            3.623612
```

- From correlation plots, it was found that there is **no linear relationship** between the outcome variable and the independent variables
- The Q-Q Plot shows many outliers indicating that the residuals are **not normally** distributed.
- Though acceptable VIF is <10, the recommended VIF is <5. We see County has VIF about 5. So, there is **multicolinearity** among the independent variables
- The residuals Vs Fitted line shows a pattern. So, it **fails** to meet the **Homoscedasticity** rule

All assumptions of linier regression have failed when we used the data without transformation

Box Cox Transformation – Identifying optimal ƛ for transformation

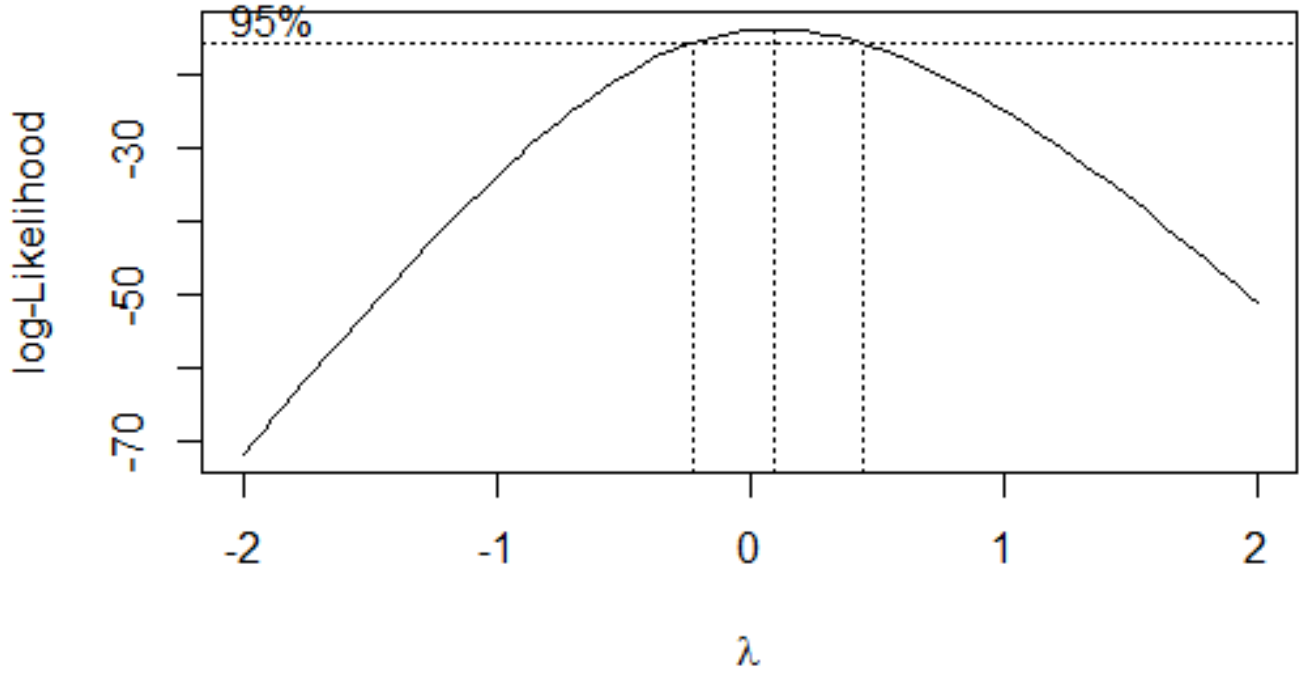
Formula for Box Cox Transformation

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0; \\ \log y, & \text{if } \lambda = 0. \end{cases}$$

Generally used ƛ Values

λ	Transformation
2	$W_i = Y_i^2$
0.5	$W_i = \sqrt{Y_i}$
0	$W_i = \mathrm{Ln}(Y_i)$
−0.5	$W_i \frac{1}{\sqrt{Y_i}}$
−1	$W_i = \frac{1}{Y_i}$

Box cox transformation is done for the regression model to find out the optimal ƛ value



The optimal ƛ value is almost zero. Hence using natural log will give the best transformation

Regression after transforming the data

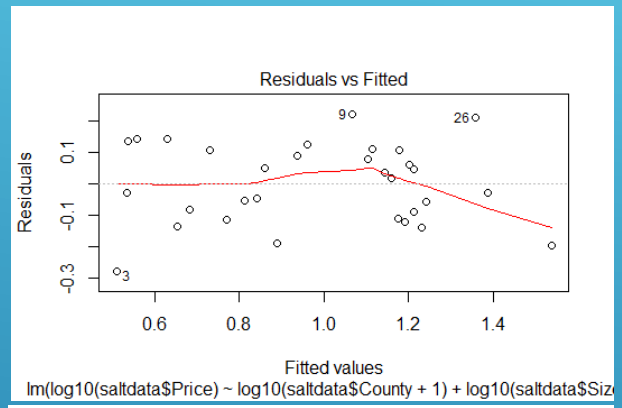
Regression is done considering all the independent variables after transforming the data using log

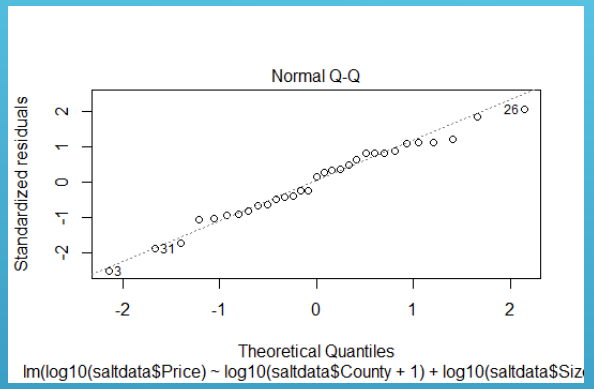
```
Residuals:
     Min       1Q   Median       3Q      Max
-0.27871 -0.10032  0.01879  0.10793  0.22147

Coefficients:
                                Estimate   Std. Error   t value     Pr(>|t|)
(Intercept)                     1.55687    0.21500      7.241       2.27e-07 ***
log10(saltdata$County + 1)     -0.19052    0.42519     -0.448       0.658294
log10(saltdata$Size)           -0.06631    0.05509     -1.204       0.240921
log10(saltdata$Elevation + 1)   0.32144    0.09573      3.358       0.002723 **
log10(saltdata$Sewer + 1)      -0.03041    0.02350     -1.294       0.208534
log10(saltdata$Date * (-1))    -0.42491    0.10426     -4.075       0.000466 ***
log10(saltdata$Flood + 1)      -1.22814    0.29417     -4.175       0.000364 ***
log10(saltdata$Distance + 1)    0.33442    0.11963      2.795       0.010279 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1449 on 23 degrees of freedom
Multiple R-squared:  0.8339,      Adjusted R-squared:  0.7834
F-statistic:  16.5 on 7 and 23 DF,  p-value: 1.369e-07
```

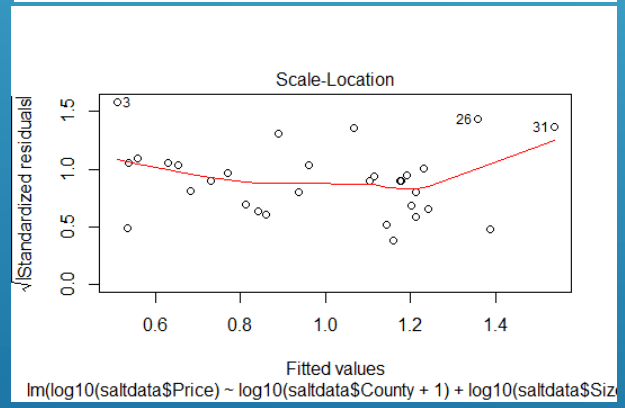
- Rsqr is 83.4% while Rsqr adjusted is 78.3%. The huge difference between the Rsqr and Rsqr adjusted shows that the Rsqr is inflated by multi-coliniarity among the independent variables
- Intercept, Elevation, Date, Flood and Distance are the significant variables

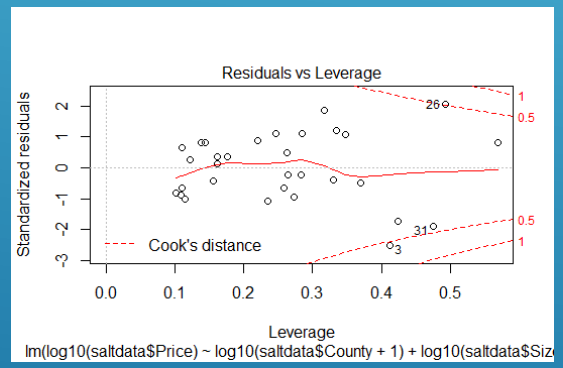
Regression after transforming the data
Validation of Assumptions – Residual Plots

Residual Plots









Variance Inflation Factor(VIF)

```
vif(mlr_t)
  log10(saltdata$County + 1)        log10(saltdata$Size)
log10(saltdata$Elevation + 1)
                    5.739676                    1.303105                    1.738083
   log10(saltdata$Sewer + 1)   log10(saltdata$Date * (-1))
log10(saltdata$Flood + 1)
                    2.378274                    1.597023                    1.566473
log10(saltdata$Distance + 1)
                    3.346551
```

- The Q-Q Plot shows many outlliers indicating that the residuals are **not normally** distributed.
- Though acceptable VIF is <10, the recommended VIF is <5. We see County has VIF about 6. So, there is **multicollinearity** among the independent variables
- The residuals Vs Fitted line shows a pattern. So, it **fails** to meet the **Homoscedasticity** rule

All assumptions of linear regression have failed even after transforming the data

Regression after transforming the data and addressing VIF

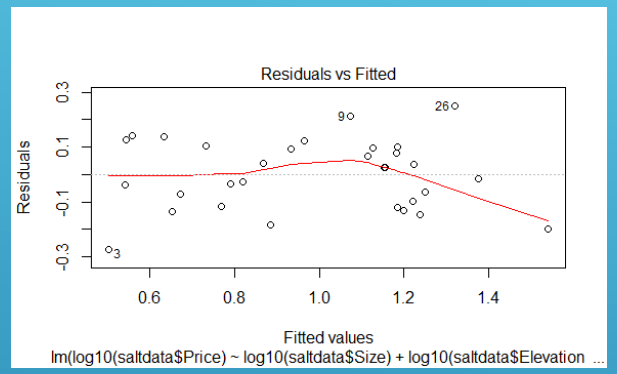
Regression is done considering all the independent variables, except County (due to high VIF) after transforming the data using log

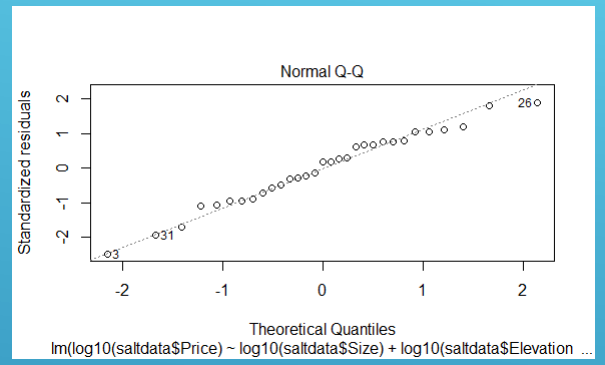
```
Residuals:
     Min       1Q   Median       3Q      Max
-0.27360 -0.10591  0.02420  0.09881  0.25062

Coefficients:
                                 Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)                       1.52212    0.19716     7.720    5.89e-08 ***
log10(saltdata$Size)             -0.06297    0.05366    -1.173    0.252141
log10(saltdata$Elevation + 1)     0.31283    0.09221     3.393    0.002400 **
log10(saltdata$Sewer + 1)        -0.02427    0.01878    -1.293    0.208438
log10(saltdata$Date * (-1))      -0.44836    0.08866    -5.057    3.60e-05 ***
log10(saltdata$Flood + 1)        -1.19633    0.28068    -4.262    0.000271 ***
log10(saltdata$Distance + 1)      0.37404    0.07923     4.721    8.44e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1425 on 24 degrees of freedom
Multiple R-squared:  0.8325,      Adjusted R-squared:  0.7906
F-statistic: 19.88 on 6 and 24 DF,  p-value: 3.179e-08
```

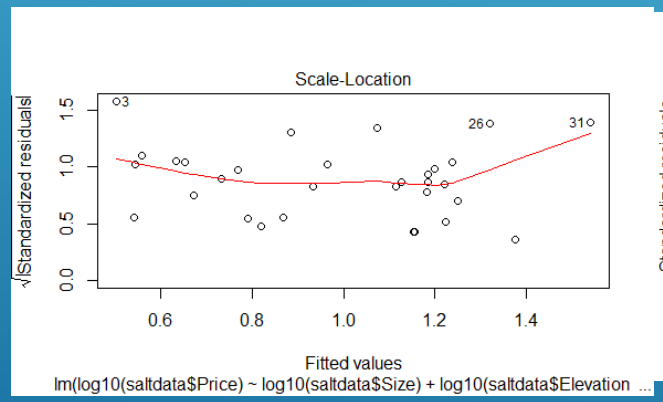
- Rsqr is 83.4% while Rsqr adjusted is 79.1%. The low difference between the Rsqr and Rsqr adjusted shows that the there is **no multicollinearity** among the independent variables
- Intercept, Elevation, Date, Flood and Distance are the significant variables

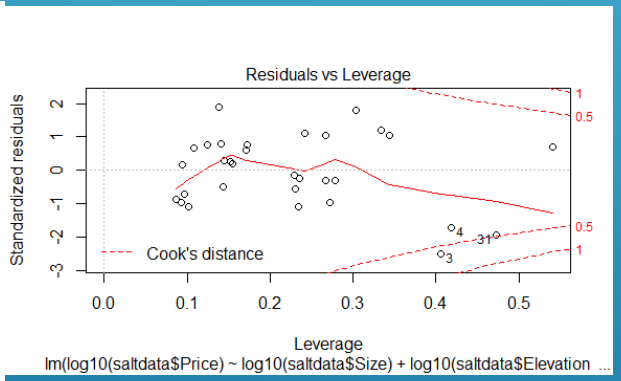
Regression after transforming the data and addressing VIF
Validation of Assumptions – Residual Plots

Residual Plots









Variance Inflation Factor(VIF)

```
vif(mlr_t2)
      log10(saltdata$Size) log10(saltdata$Elevation + 1)
log10(saltdata$Sewer + 1)
                  1.279203                     1.668075                     1.570477
 log10(saltdata$Date * (-1))     log10(saltdata$Flood + 1)
log10(saltdata$Distance + 1)
                  1.194569                     1.475219                     1.518253
```

- The Q-Q Plot shows many outliers indicating that the residuals are **not normally** distributed.
- VIF is less than 5. So, there is **no multicollinearity** among the independent variables
- The residuals Vs Fitted line shows a pattern, it **fails** for the **Homoscedasticity**

The assumptions normality and homoscedacity have failed even after transforming the data and addressing multicollinearity

Regression – Final outcome and Conclusion

- Intercept, Elevation, Date, Flood and Distance are the significant variables with p value <0.05
- Strong model with Rsqr is 83.4% while Rsqr adjusted is 79.1%.
- Regression Equation is

Log (Price) = 1.52 + 0.31 * Log (Elevation) – 0.45 * log (Date*(-1)) – 1.2 * log (Flood+1) + 0.4 * log (Distance +1)

- The **model can be used** with large set of samples (observations) and removing outliers so that the normality and homoscedascity are taken care

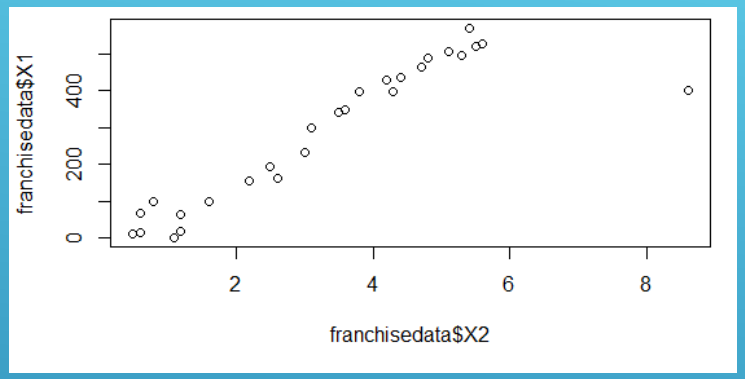
Question 3

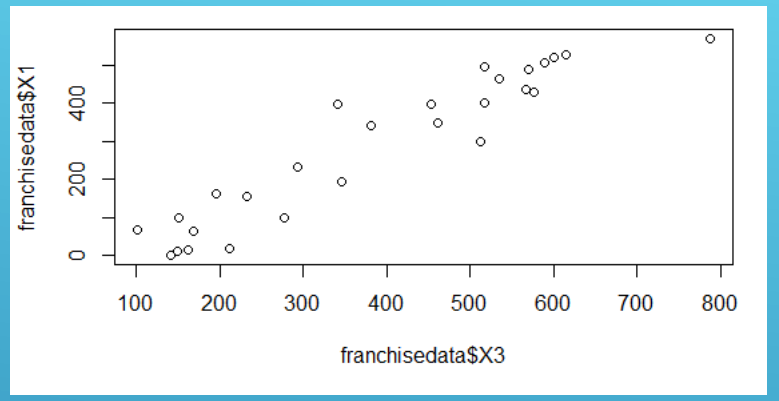
Problem Statement - 3

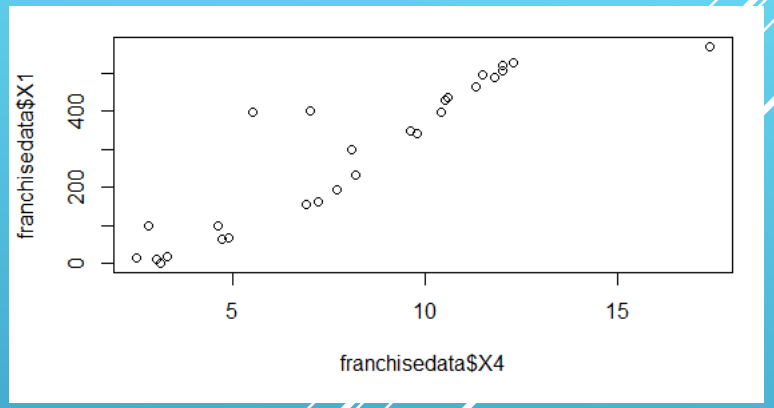
- All Greens Franchise Explain the importance of X2, X3, X4, X5, X6 on Annual Net Sales, X1.
- The data (X1, X2, X3, X4, X5, X6) are for each franchise store.

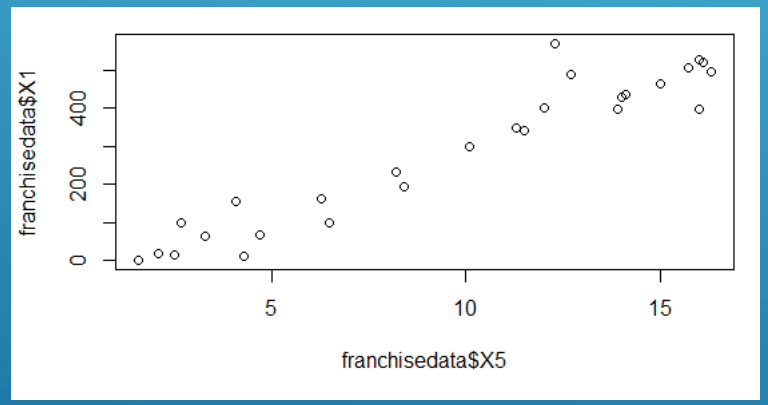
- X1 = annual net sales/$1000
- X2 = number sq. ft./1000
- X3 = inventory/$1000
- X4 = amount spent on advertising/$1000
- X5 = size of sales district/1000 families X6 = number of competing stores in district

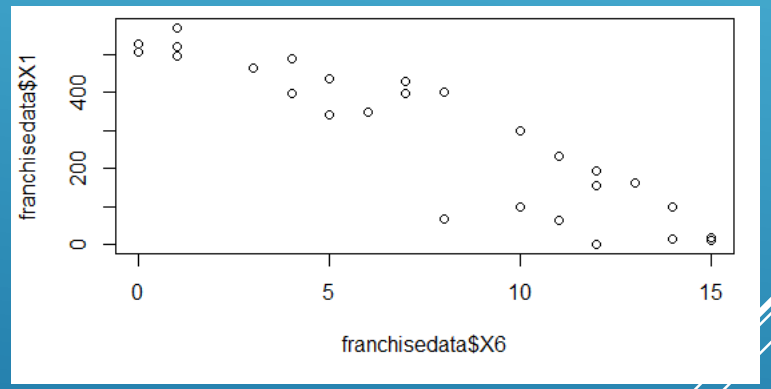
Check for linear relationship











Correlation Coefficients

	X1	X2	X3	X4	X5	X6
X1	1.00	0.89	0.95	0.91	0.95	-0.91
X2	0.89	1.00	0.84	0.75	0.84	-0.77
X3	0.95	0.84	1.00	0.91	0.86	-0.81
X4	0.91	0.75	0.91	1.00	0.80	-0.84
X5	0.95	0.84	0.86	0.80	1.00	-0.87
X6	-0.91	- 0.77	-0.81	-0.84	-0.87	1.00

- X2 to X5 have strong positive correlation with X1
- X6 has strong negative correlation with X1
- There appears strong correlation among all the independent variables. Will be validated later by checking VIF

Regression Model

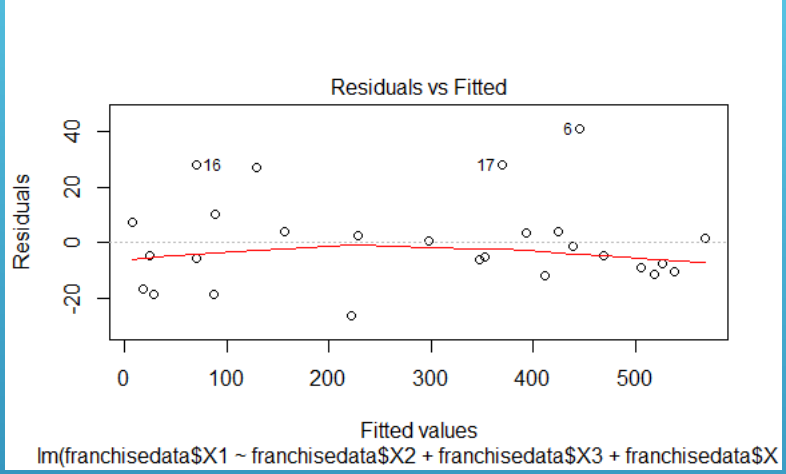
```
Call:
lm(formula = franchisedata$X1 ~ franchisedata$X2 + franchisedata$X3 +
    franchisedata$X4 + franchisedata$X5 + franchisedata$X6)

Residuals:
    Min      1Q  Median      3Q     Max
-26.338  -9.699  -4.496   4.040  41.139

Coefficients:
                  Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)       -18.85941   30.15023   -0.626   0.538372
franchisedata$X2   16.20157    3.54444    4.571   0.000166 ***
franchisedata$X3    0.17464    0.05761    3.032   0.006347 **
franchisedata$X4   11.52627    2.53210    4.552   0.000174 ***
franchisedata$X5   13.58031    1.77046    7.671   1.61e-07 ***
franchisedata$X6   -5.31097    1.70543   -3.114   0.005249 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.65 on 21 degrees of freedom
Multiple R-squared:  0.9932,    Adjusted R-squared:  0.9916
F-statistic: 611.6 on 5 and 21 DF,  p-value: < 2.2e-16
```

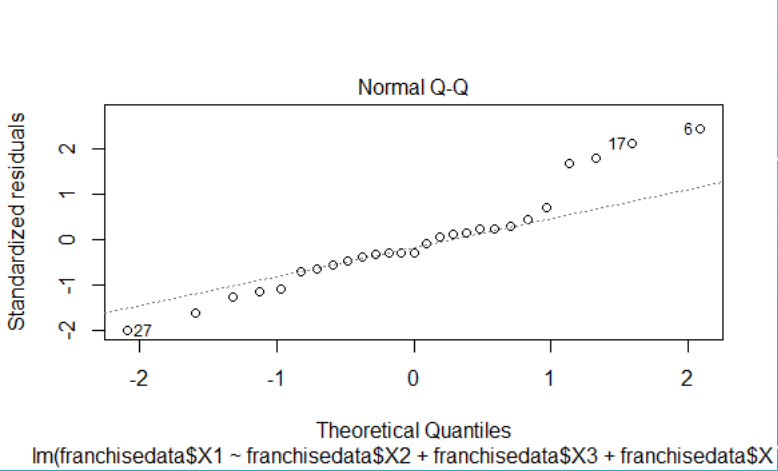
franchisedata$X2	franchisedata$X3	franchisedata$X4	franchisedata$X5	franchisedata$X6
4.240914	**10.122480**	7.624391	6.912318	5.818768

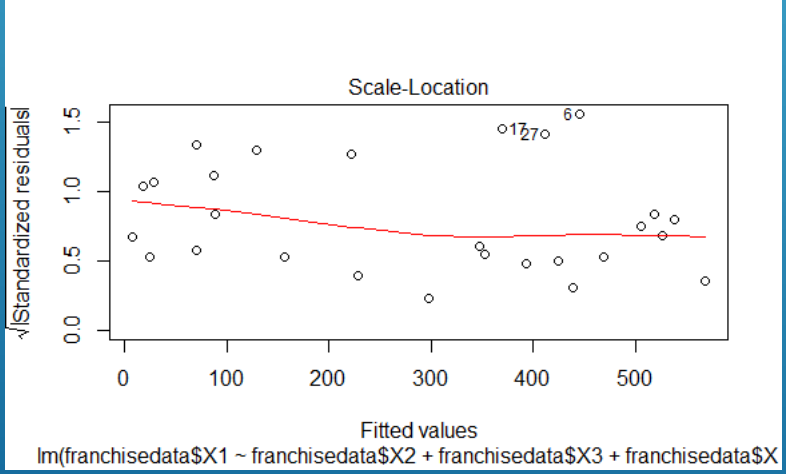
X1 = annual net sales/$1000
X2 = number sq. ft./1000
X3 = inventory/$1000
X4 = amount spent on advertising/$1000
X5 = size of sales district/1000 families
X6 = number of competing stores in district

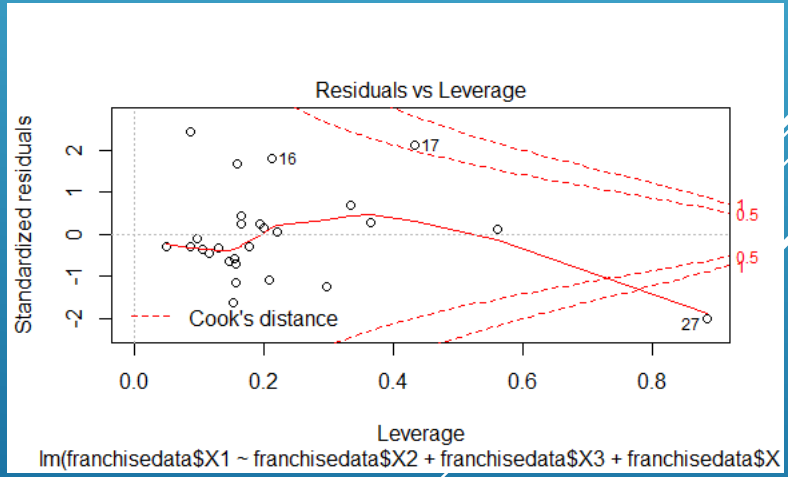
- VIF for Inventory (X3) is >10
- If we consider inventory depends on the sq.ft area, considering any one of these variables will be good enough
- Since VIF for inventory (X3) is >10, remove inventory from the model

Validating Assumptions – Residual Plots









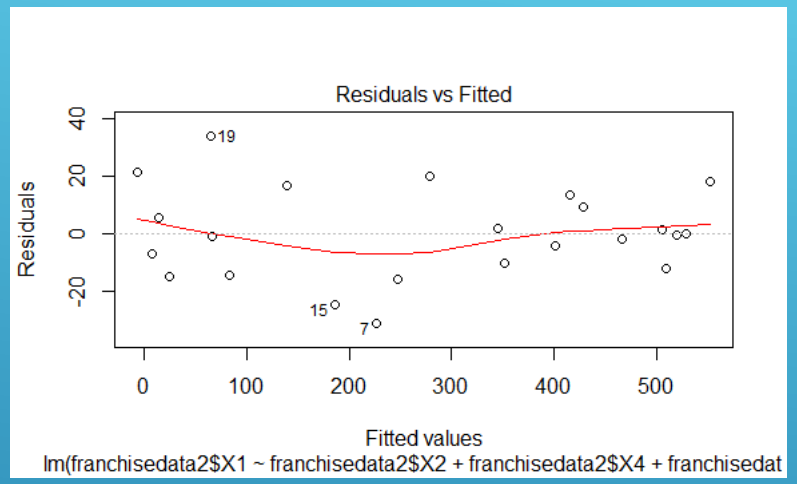
- The Q-Q Plot shows many outliers indicating that the residuals are **not normally** distributed.
- The residuals Vs Fitted line shows a pattern, it **fails** for the **Homoscedasticity**
- There are many outliers – 6th, 16th, 17thand 27th observations. Running a model after removing outliers may give better results

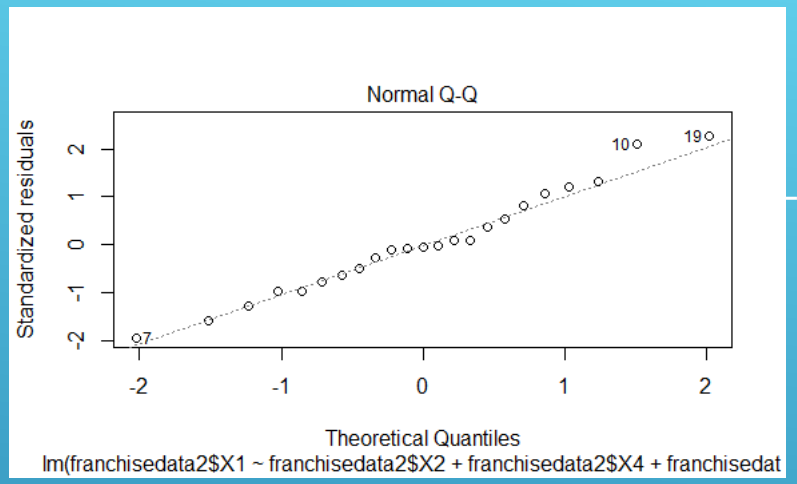
Regression Model after addressing VIF and removing outliers

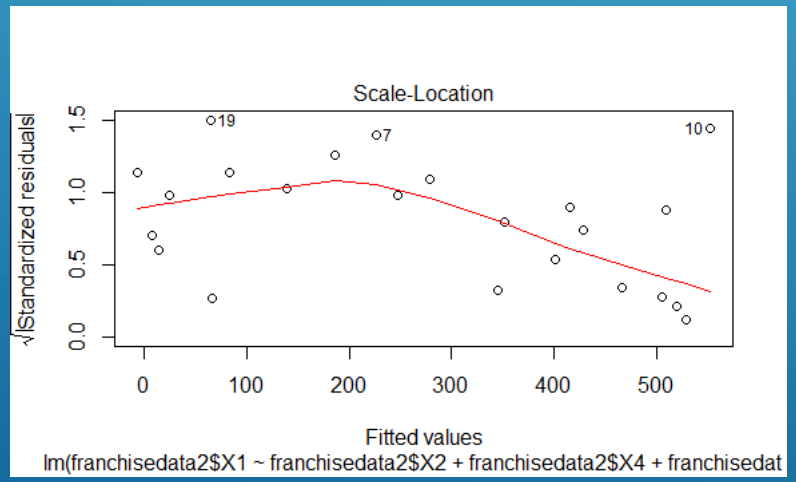
```
Call:
lm(formula = franchisedata2$X1 ~ franchisedata2$X2 + franchisedata2$X4 +
    franchisedata2$X5 + franchisedata2$X6)

Residuals:
    Min      1Q  Median      3Q     Max
-31.354 -11.314  -0.733  11.137  33.870

Coefficients:
                    Estimate   Std. Error  t value   Pr(>|t|)
(Intercept)         -70.078     32.400      -2.163    0.0443 *
franchisedata2$X2   25.380      9.987       2.541     0.0205 *
franchisedata2$X4   17.228      3.250       5.300     4.87e-05 ***
franchisedata2$X5   15.268      2.248       6.791     2.32e-06 ***
franchisedata2$X6   -2.327      1.787       -1.302    0.2092
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.5 on 18 degrees of freedom
Multiple R-squared: 0.9936,      Adjusted R-squared:  0.9921
F-statistic: 693.7 on 4 and 18 DF,  p-value: < 2.2e-16
```

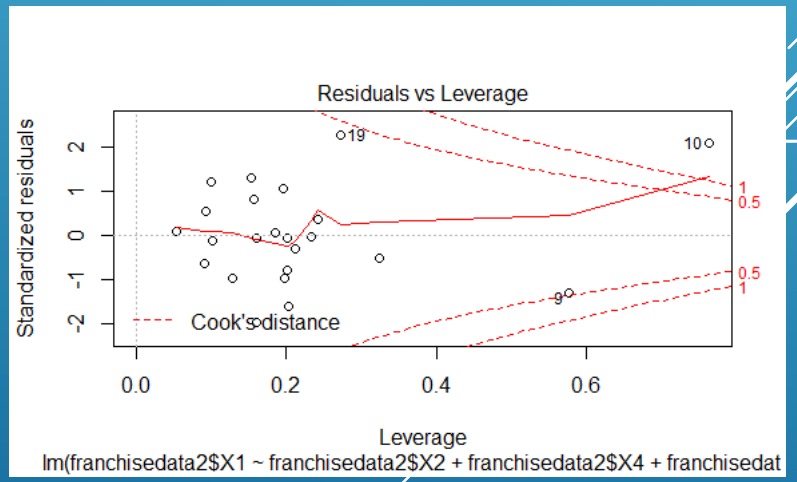
- Rsqr is 99.4% while Rsqr adjusted is 99.2%, and p value <0.05, indicating that the model is significant
- Intercept, Sq ft (X2), advertising spend (X4) and Size of Sales district (X5) are the significant factors

Regression Model after addressing VIF and removing outliers









- The Q-Q Plot shows almost all the residuals are along the fitted line, though there are few outliers
- The residuals Vs Fitted line shows a pattern, it **fails** for the **Homoscedasticity**
- There are many outliers – 6th, 16th, 17thand 27th observations. Running a model after removing outliers may give better results

Regression – Final outcome and Conclusion

- Intercept, Sq ft (X2), advertising spend (X4) and Size of Sales district (X5) are the significant factors

- Strong model with Rsqr is 99.4% while Rsqr adjusted is 99.2%.

- Regression Equation is

Annual net sales (X1) = -70 + 25.4 * Sq ft (X2) +17.2 * advertising spend (X4) + 15.3*Size of Sales district (X5)

- The **model can be used** with large set of samples (observations) and removing outliers so that the normality and homoscedascity are taken care

Thank You