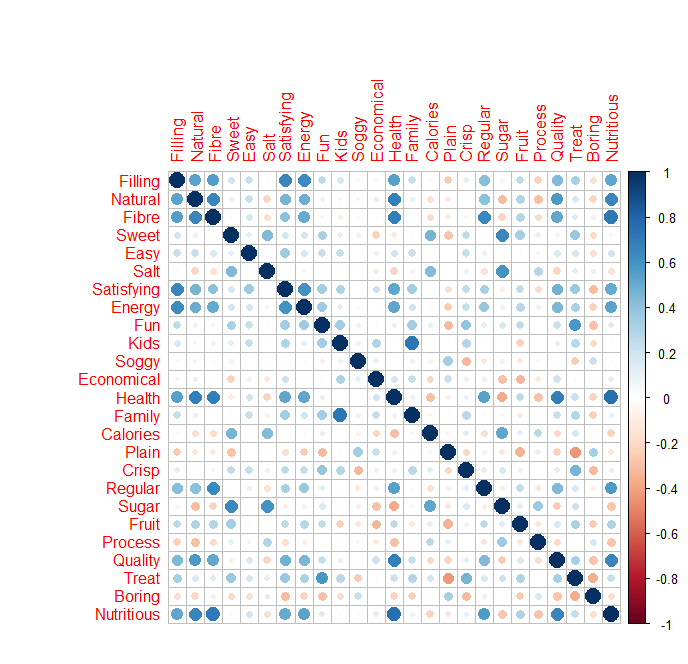
|  |
| --- |
| Photo displaying partial image of two pie charts on a canvas-textured page |
|  |
| |  |  |  | | --- | --- | --- | |  |  |  | |

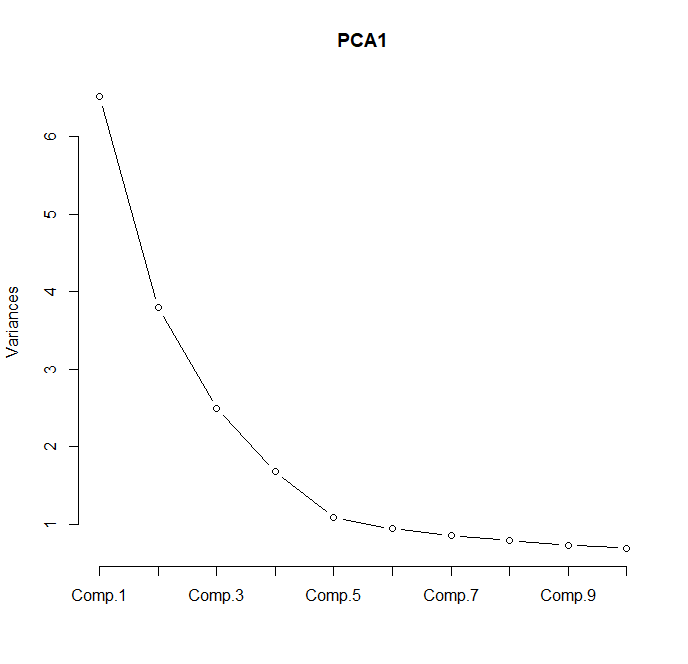
**Question – 1**

The summary statistics for the data shows that some of the variables in the dataset have a scale from 1-6 while almost most of the variables have been rated in the scale from 1-5.

**#Assumption:** We assume that the data might have been recorded incorrectly for these variables with maximum value as 6 and should instead have been 5. We have replaced all such values from 6 to 5

A correlation matrix is built for all the 25 variables in the dataset. The plot shows that there are multiple variables correlated to each other and hence it would be effective to run a dimensionality reduction technique: Factor Analysis to try and club the variables.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Comp.1 | Comp.2 | Comp.3 | Comp.4 | Comp.5 | Comp.6 | Comp.7 | Comp.8 | Comp.9 | Comp.10 |
| Standard deviation | 2.55 | 1.95 | 1.58 | 1.3 | 1.04 | 0.97 | 0.92 | 0.89 | 0.86 | 0.84 |
| Proportion of Variance | 0.26 | 0.15 | 0.1 | 0.07 | 0.04 | 0.04 | 0.03 | 0.03 | 0.03 | 0.03 |
| Cumulative Proportion | 0.26 | 0.41 | 0.51 | 0.58 | 0.62 | 0.66 | 0.69 | 0.73 | 0.76 | 0.78 |
|  | *Comp.11* | *Comp.12* | *Comp.13* | *Comp.14* | *Comp.15* | *Comp.16* | *Comp.17* | *Comp.18* | *Comp.19* | *Comp.20* |
| Standard deviation | 0.81 | 0.74 | 0.73 | 0.7 | 0.65 | 0.62 | 0.6 | 0.6 | 0.55 | 0.52 |
| Proportion of Variance | 0.03 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.01 | 0.01 | 0.01 | 0.01 |
| Cumulative Proportion | 0.81 | 0.83 | 0.85 | 0.87 | 0.89 | 0.9 | 0.92 | 0.93 | 0.95 | 0.96 |
|  | *Comp.21* | *Comp.22* | *Comp.23* | *Comp.24* | *Comp.25* |  |  |  |  |  |
| Standard deviation | 0.51 | 0.49 | 0.47 | 0.45 | 0.41 |  |  |  |  |  |
| Proportion of Variance | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |  |  |  |  |  |
| Cumulative Proportion | 0.97 | 0.98 | 0.99 | 0.99 | 1 |  |  |  |  |  |

The above table shows the Components after running PCA. Based on the above table, 4 components have a significant effect on cumulative proportion and the rest components add only about 4% at every step.

Let’s do a Scree Plot to confirm the no. of Factors to be taken before running the Factor Analysis

The Scree Plot shows that 4 factors are ideal for performing Factor Analysis.

Adding the 5th factor only adds 4% of the variance explained to the overall model.

**Running Factor Analysis with 4 factors and using Varimax, Promax and Oblimin rotation, the ideal set of variables with meaningful labels could be formed by using Varimax/Oblimin rotation**

Loadings:

Factor1 Factor2 Factor3 Factor4

q1.cerealFilling 0.693 0.179 0.195

q1.cerealNatural 0.749 -0.215

q1.cerealFibre 0.814 -0.114 -0.108

q1.cerealSweet 0.715 0.358

q1.cerealEasy 0.248 0.290

q1.cerealSalt 0.691

q1.cerealSatisfying 0.618 0.213 0.370

q1.cerealEnergy 0.652 0.239 0.167

q1.cerealFun 0.152 0.160 0.534 0.371

q1.cerealKids 0.876

q1.cerealSoggy -0.455 0.135

q1.cerealEconomical -0.271 -0.205 0.391

q1.cerealHealth 0.831 -0.287

q1.cerealFamily 0.126 0.793

q1.cerealCalories -0.123 0.613 0.134

q1.cerealPlain -0.138 -0.647

q1.cerealCrisp 0.141 0.441 0.313

q1.cerealRegular 0.634

q1.cerealSugar -0.181 0.821 0.166

q1.cerealFruit 0.364 0.174 0.438 -0.284

q1.cerealProcess -0.236 0.367 -0.122

q1.cerealQuality 0.663 -0.247 0.190 0.165

q1.cerealTreat 0.250 0.206 0.632 0.277

q1.cerealBoring -0.169 -0.509 -0.198

q1.cerealNutritious 0.834 -0.174

Factor1 Factor2 Factor3 Factor4

SS loadings 5.202 2.629 2.400 2.347

Proportion Var 0.208 0.105 0.096 0.094

Cumulative Var 0.208 0.313 0.409 0.503

Based on the above output the following variables can be grouped into the following factors. The labels of the factors can be referred as

|  |  |  |  |
| --- | --- | --- | --- |
| *Nutritious\_Value* | *Taste* | *Perception* | *Demographics* |
| Filling | Sweet | Fun | Kids |
| Natural | Salt | Soggy | Economical |
| Fibre | Calorie | Plain | Family |
| Satisfying | Sugar | Crisp | Easy |
| Energy | Process | Fruit |  |
| Health |  | Treat |  |
| Quality |  | Boring |  |
| Regular |  |  |  |
| Nutritious |  |  |  |

**Averaging the old labels in each factor and aggregating them to the brand to see the final mean rating for the new labels**

**Brand Preference:**

From the table on the right, the highest average ratings for the 4 labels is as below:

**Nutrition Value (***Komplete, PMuesli & Sustain)*

**Taste (***CMuesli, PMuseli & NutriGain***)**

**Perception (***Komplete, PMuesli, Sustain***)**

**Demographics (***Vitabrit, Cornflakes, RiceBubbles***)**

We can clearly see that PMuesli is the most preferred brand with highest rating among the three labels out of four.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cereals | Nutrition\_Value | Taste | Perception | Demographics |
| AllBran | 3.91 | 2.20 | 1.99 | 3.35 |
| CMuesli | 3.97 | 2.78 | 2.48 | 3.81 |
| CornFlakes | 3.34 | 2.69 | 2.34 | 4.23 |
| JustRight | 3.63 | 2.71 | 2.41 | 3.48 |
| Komplete | 3.98 | 2.59 | 2.53 | 3.00 |
| NutriGrain | 3.41 | 3.10 | 2.47 | 4.10 |
| PMuesli | 4.07 | 2.86 | 2.63 | 3.58 |
| RiceBubbles | 2.93 | 2.19 | 2.44 | 4.31 |
| SpecialK | 3.47 | 2.34 | 2.25 | 3.96 |
| Sustain | 4.18 | 2.17 | 2.65 | 3.65 |
| Vitabrit | 3.94 | 1.92 | 2.23 | 4.12 |
| Weetabix | 3.87 | 2.08 | 2.11 | 3.92 |

**Question – 2**

**Nature of each variable?**

Price, Size, Elevation, Sewer, Date & Distance are continuous variables whereas Country and Flood are the two Categorical variables with dummy binary values used for converting to continuous variables

Here, Price has been taken as the Dependent (or response variable) and other variables as possible independent (or predictor variables)

**Check whether variables require transformation individually**

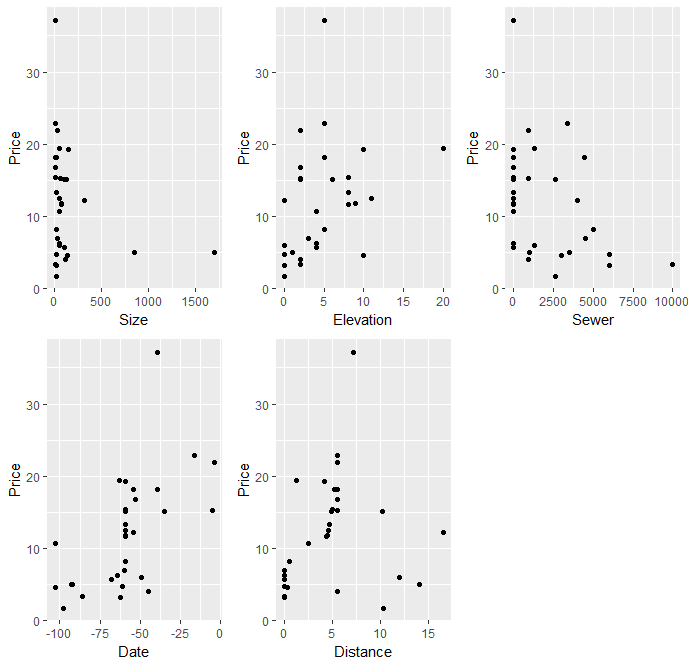
There are multiple variables which require individual transformation

**Price -** Price is Right skewed and hence log transformation applied to better the result of linear model

**Size -** Size because again its heavily skewed and hence log transformation again applied to make the distribution normal

Other variables also show skewed distribution but due to large no. of 0 values, transformation cannot be applied as it will make the variables value difficult to interpret

**Setup a regression equation, run the model and discuss the results**

The following steps have been performed and final regression equation set accordingly

***Check for Linearity of independent variables w.r.t independent variable price***

From the scatter plot of the independent variables w.r.t to Price, it can be seen that data is scattered and have a somewhat linear pattern except for Size and Sewer.

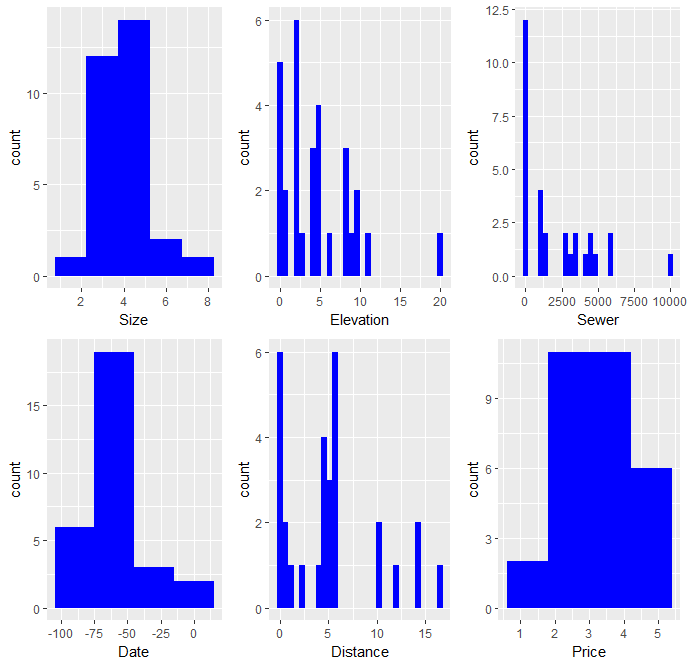
Also, there can be seen a value which is an outlier at Price value of 37.2. This value as evident from the scatter plot is an outlier and must be removed.

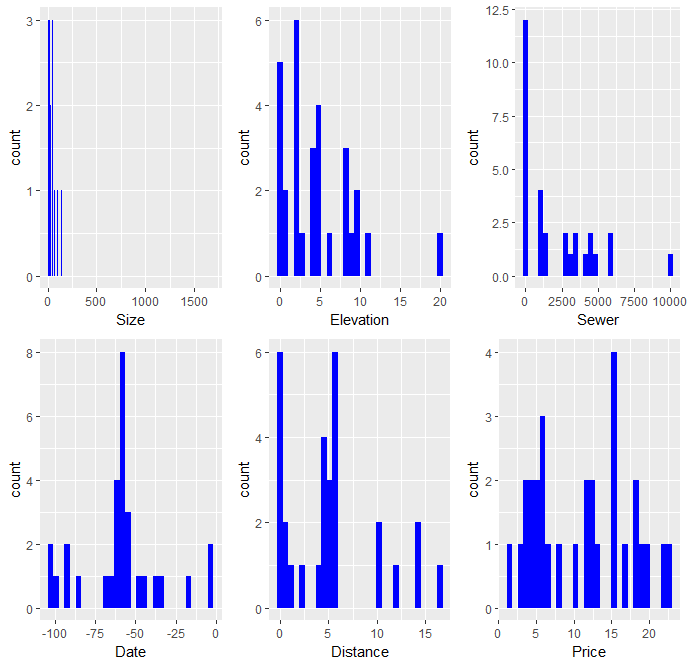
***Check for Normality***

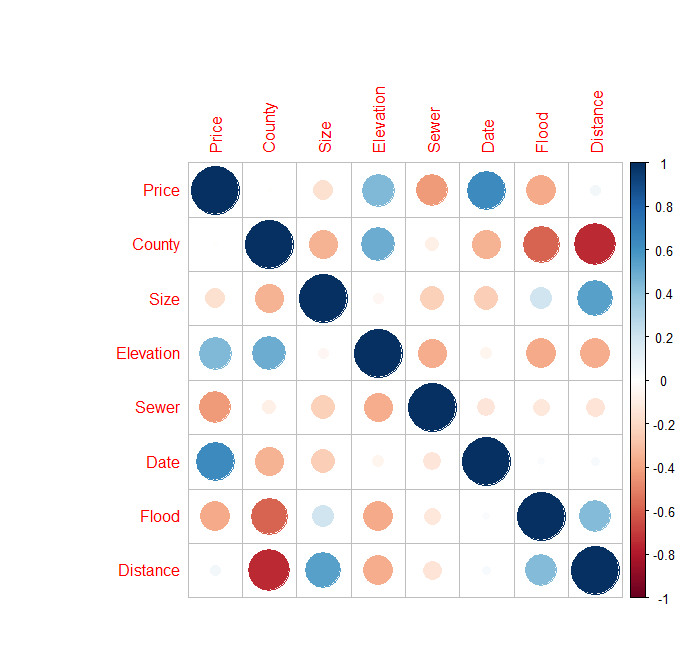
For Normality assumption, all the independent continuous variables are plotted in a histogram to see the distribution

From the Histograms, it is evident that Date is normally distributed and all the other variables are right skewed or do not follow a normal distribution

**Transforming Price and Size (log transformation) to make it normal. Sewer and Distance are not normal. But any transformation applied to them will make the variables less informative as 0 values would remain 0 (Shapiro Test gives “p” value greater than .05)**

*** Before Transformation*** ***After Transformation***

******

***Multi Collinearity***

We do a corrplot for all the variables in the dataset.

It is evident that few independent variables are correlated to each other and hence there is a little multi collinearity in the dataset.

From the Corrplot it is visible that Elevation, Sewer, Date and Flood are the variables that have correlation with the Price.

Therefore, we will go ahead and use the 4 variables in our regression model

Interpretation: Strangely, Elevation is not that significant in the model**. This might be due to Elevation & Sewer being negatively correlated to each other (Multicollinearity effect on the model)**

Call:

lm(formula = Price ~ Elevation + Sewer + Date + Flood, data = ga2)

Residuals:

Min 1Q Median 3Q Max

-0.61017 -0.21423 0.07185 0.19590 0.58466

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.325e+00 2.048e-01 16.233 8.69e-15 \*\*\*

Elevation 3.541e-02 1.778e-02 1.992 0.05742 .

Sewer -9.171e-05 2.972e-05 -3.086 0.00491 \*\*

Date 1.641e-02 2.678e-03 6.130 2.08e-06 \*\*\*

Flood -6.833e-01 1.940e-01 -3.522 0.00167 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3498 on 25 degrees of freedom

Multiple R-squared: 0.7736, Adjusted R-squared: 0.7374

F-statistic: 21.35 on 4 and 25 DF, p-value: 9.212e-08

Running again, a second model using Forward selection method:

Call:

lm(formula = Price ~ Date + Elevation + Flood + Distance, data = ga2)

Residuals:

Min 1Q Median 3Q Max

-0.50614 -0.19582 -0.08102 0.15147 0.98724

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.769235 0.200611 13.804 3.37e-13 \*\*\*

Date 0.017741 0.002535 6.998 2.46e-07 \*\*\*

Elevation 0.073767 0.015679 4.705 8.01e-05 \*\*\*

Flood -0.728982 0.189006 -3.857 0.000715 \*\*\*

Distance 0.054654 0.015531 3.519 0.001683 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3361 on 25 degrees of freedom

Multiple R-squared: 0.7909, Adjusted R-squared: 0.7575

F-statistic: 23.64 on 4 and 25 DF, p-value: 3.473e-08

We see that all the variables are significant in explaining the variance in Price. Also, we have increased are Rsquare adjusted to 76% in this model. Now let’s go ahead and check for the assumptions:

**The AIC value of the two models give less value for the second model and hence the second model is the better option**

vif(model2) #Multi collinearit is absent

Date Elevation Flood Distance

1.003678 1.237829 1.317267 1.308392

> dwtest(model2)#auto correlation is absent

Durbin-Watson test

data: model2

DW = 2.2326, p-value = 0.629

alternative hypothesis: true autocorrelation is greater than 0

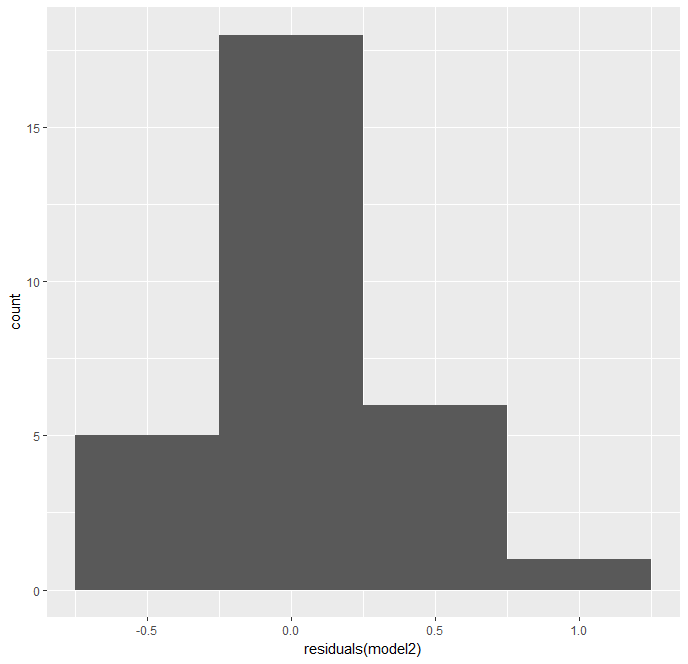
> gqtest(model2)#heteroscadsticity is absent

Goldfeld-Quandt test

data: model2

GQ = 0.40434, df1 = 10, df2 = 10, p-value = 0.9153

alternative hypothesis: variance increases from segment 1 to 2



VIF value of less than 1 for all the variables indicate that multi collinearity is absent in the model

Durbin-Watson Test: P-Value indicates that Auto-correlation is absent

Goldfeld Quandt Test: P-Value indicates that Heteroscadasticity is absent.

Hence, we conclude that all the assumptions hold true for the Linear Regression and the model is a good predictor of Price.

**The final Linear Regression Equation is**

***Price = 2.77 + 0.02(Date) + 0.07(Elevation) -0.73(Flood) + 0.05(Distance)***

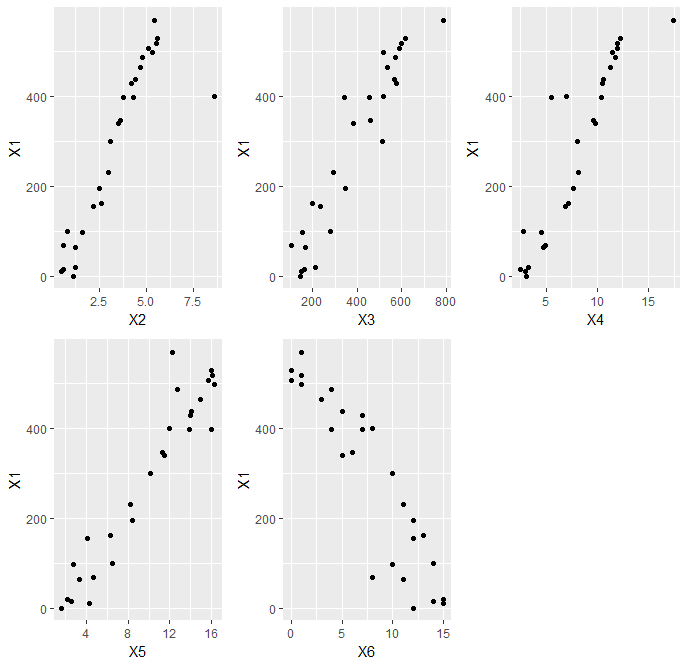
This equation can be further broken down into 2

When it’s not a flood plain (Flood = 0)

***Price = 2.77 + 0.02(Date) + 0.07(Elevation) -0.73(Flood) + 0.05(Distance)***

When the plain is flood (Flood = 1)

***Price = 2.04 + 0.02(Date) + 0.07(Elevation) + 0.05(Distance)***

**Question – 3**

We will begin the question by starting with the assumptions

***Linear Relationship with Dependent Variable (X1)***

The scatterplot of independent variable with dependent variable shows that all the five variables X2, X3, X4, X5 & X6 have a linear relationship with X1

***Multivariate Normality***

shapiro.test(ga3$X2) #Normal

Shapiro-Wilk normality test

data: ga3$X2 # P value suggest that we do not reject the

W = 0.94296, p-value = 0.1441 NULL Hypothesis and hence we conclude that

X2 is Normal

> shapiro.test(ga3$X3) #Normal

Shapiro-Wilk normality test

data: ga3$X3 # P value suggest that we do not reject the

W = 0.929, p-value = 0.06533 NULL Hypothesis and hence we conclude that

X3 is Normal

> shapiro.test(ga3$X4) #Normal

Shapiro-Wilk normality test

data: ga3$X4 # P value suggest that we do not reject the

W = 0.94587, p-value = 0.17 NULL Hypothesis and hence we conclude that

X4 is Normal

> shapiro.test(ga3$X5) #Not Normal

Shapiro-Wilk normality test

data: ga3$X5 # P value suggest that we do reject the

W = 0.90048, p-value = 0.01372 NULL Hypothesis and hence we conclude that

X5 is Not Normal

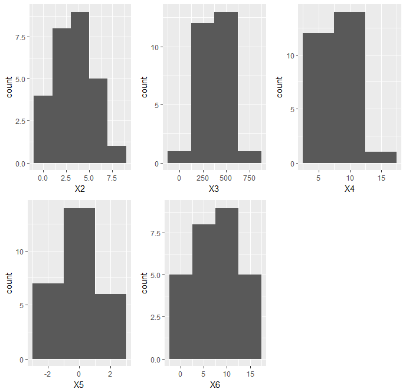
> shapiro.test(ga3$X6) #Normal

Shapiro-Wilk normality test

# P value suggest that we do not reject the

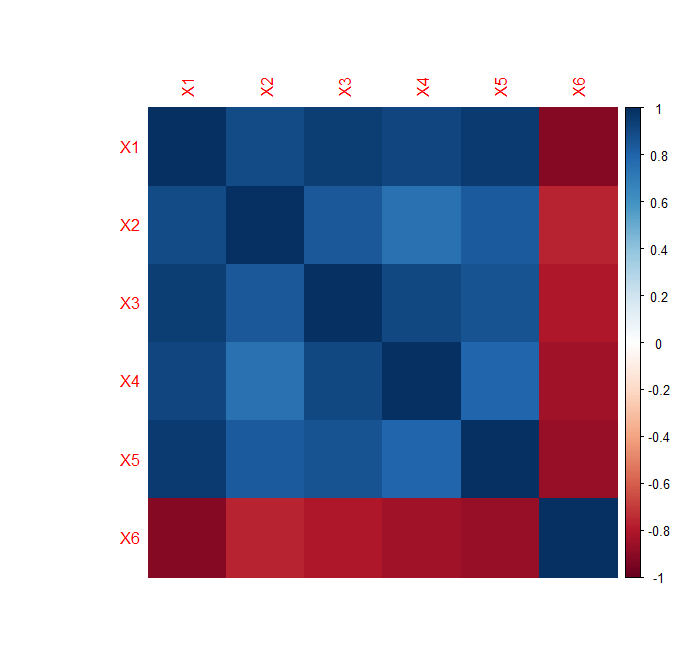
data: ga3$X6 NULL Hypothesis and hence we conclude that

W = 0.93593, p-value = 0.09668 X6 is Normal

**Transforming X5 using scale() function and plotting Histograms for all the variables**

The Histogram on the right after transforming shows that all the variables now have a Normal Distribution

**Multi Collinearity Assumption**



The plot shows that all the variables are highly correlated with each other, signifying high multicollinearity effect.

**Now, we start with building the Linear Regression model using backward selection method.**

Call:

lm(formula = X1 ~ ., data = ga3)

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) 112.76903 24.91496 4.526 0.000185 \*\*\*

X2 16.20157 3.54444 4.571 0.000166 \*\*\*

X3 0.17464 0.05761 3.032 0.006347 \*\*

X4 11.52627 2.53210 4.552 0.000174 \*\*\*

X5 69.80326 9.10021 7.671 1.61e-07 \*\*\*

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

A very high value of Rsquare and Rsquare adjusted show that there might be problem with the model. We will check this by running the assumptions test on the model.

**VIF to check for Multicollinearity effect**

X2 X3 X4 X5 X6

4.240914 10.122480 7.624391 6.912318 5.818768

We will now remove X3 from the model as it has the highest VIF factor in the model and run the model again

Call:

lm(formula = X1 ~ X2 + X4 + X5 + X6, data = ga3)

Residuals:

Min 1Q Median 3Q Max

-30.422 -12.858 -6.477 16.160 45.255

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 112.452 29.186 3.853 0.000863 \*\*\*

X2 20.444 3.815 5.359 2.22e-05 \*\*\*

X4 16.966 2.093 8.107 4.73e-08 \*\*\*

X5 80.560 9.817 8.206 3.86e-08 \*\*\*

X6 -4.043 1.937 -2.088 0.048629 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.68 on 22 degrees of freedom

Multiple R-squared: 0.9902, Adjusted R-squared: 0.9884

F-statistic: 555.4 on 4 and 22 DF, p-value: < 2.2e-16

> vif(model2) #Multi collinearity check

X2 X4 X5 X6

3.579850 3.795323 5.861520 5.468943

> dwtest(model2) #Autocorrelation is absent

Durbin-Watson test

data: model2

DW = 2.1491, p-value = 0.595

alternative hypothesis: true autocorrelation is greater than 0

> gqtest(model2) #Homoscedasticity check

Goldfeld-Quandt test

data: model2

GQ = 0.4358, df1 = 9, df2 = 8, p-value = 0.8811

alternative hypothesis: variance increases from segment 1 to 2

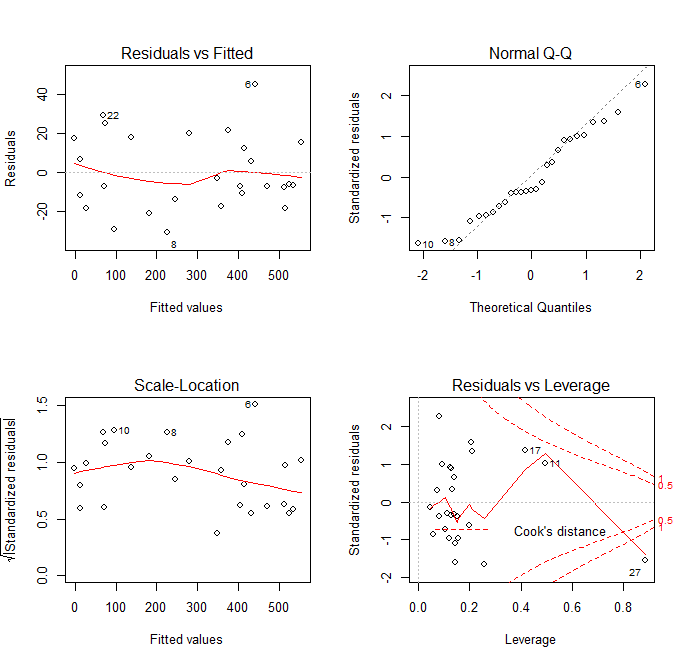
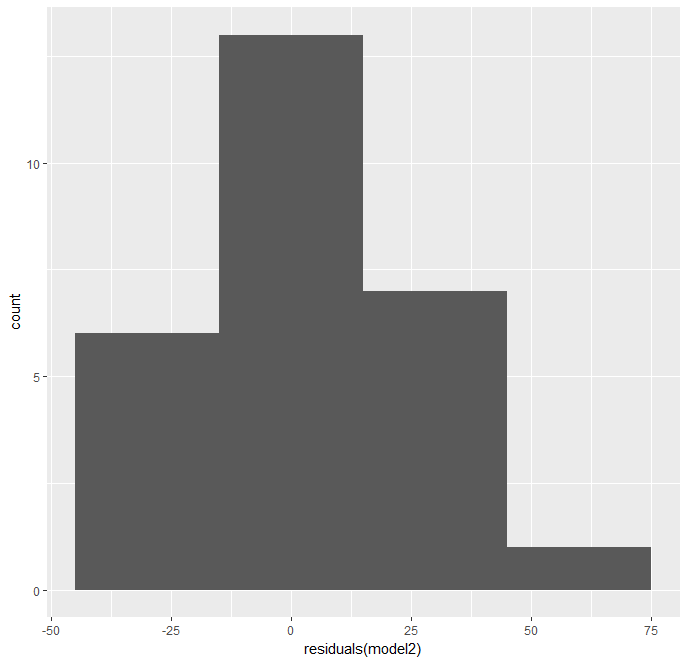
Running the diagnostics on the second model, although VIF values are still more than 5, we have still managed to considerably lower the VIF within acceptable ranges

***Auto correlation assumption***

The Durbin Watson test shows that p-value is greater than .05 and hence we do not reject the NULL hypothesis and hence we conclude that Autocorrelation is absent

**Homoscedasticity Check**

The Goldfeld Quandt Test p-value also indicates the absence of heteroscedasticity.

**Normality Check of Residuals**

The residuals vs. Fitted plot shows that the residuals are scattered all over the place and does not have pattern

The QQ Norm Plot and the Histogram also signifies that the residuals are normally distributed.

The final regression equation is:

***Annual Sales = 112.45 + 20.44(No. of Sq. Ft.) + 17(Adv. Spend) + 80.5(District Size) – 0.04(No. of Competing Stores)***