**Advanced Statistics | Group Assignment Group 3 | BABI May 2019**

**Problem 1: Cereal characteristics**

***What is the problem at hand?***

We have ‘valuable’ customer survey information from 235 responses pertaining to 12 brands across 25 cereal characteristics. As elaborate and comprehensive the survey might be, it is however, intellectually challenging for the marketing decision makers, to arrive at decisions based on these large numbers of characteristics/ dimensions. Hence we have to arrive at less number of factors/dimensions, without losing out the value from all dimensions

***What do we know from the data?***

Note: Output and code in R have been represented with blue font, and the full code for this case is as embedded below



1. **Exploratory analysis of the data in R reveals the following:**

* Though the survey scores are on a 1-5 Likert scale, there are scores such as 6, which needs to be corrected to the nearest correct value such as 6 to 5
* Most of the characteristics imply a better product with a higher score, **except for Soggy, Sugar, Process and Boring**; hence their scores need to be reversed, for uniform interpretation across the scores. The column names were then and name updated with suffix 2
* The characteristics are not just large (25) in number but also display a variety of distributions and thus not entirely normal. Refer to output and plots below for a visual representation of the distribution. For example:
  + Easy has a high average score of 4.5, while Fruit has an average score of 1.7.
  + Satisfying has a low SD of 0.8 while Fun has a high SD of 1.3
  + Calories is symmetrical, with Skew of 0, while Easy is highly skewed with a Skew of -1.9

> mean<-colMeans(x=mydata2, na.rm = TRUE)

> print(mean,digits=2)

Filling Natural Fibre Sweet Easy Salt Satisfying Energy Fun Kids Soggy2

3.9 3.8 3.5 2.5 4.5 2.0 4.0 3.6 2.6 3.8 3.7

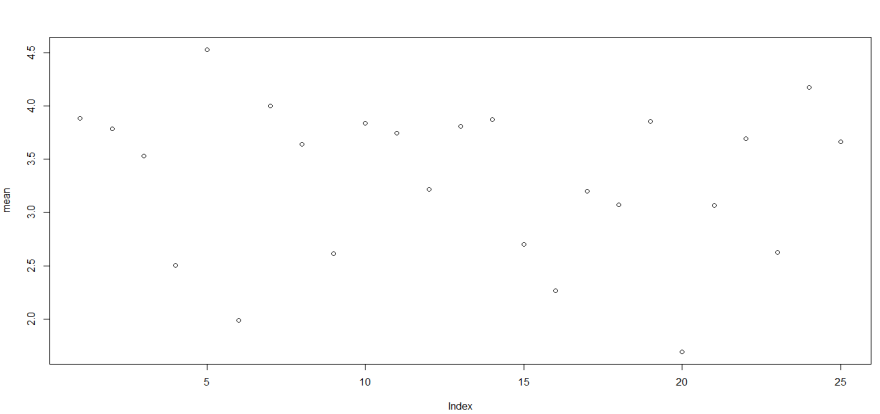
Economical Health Family Calories Plain Crisp Regular Sugar2 Fruit Process2 Quality

3.2 3.8 3.9 2.7 2.3 3.2 3.1 3.9 1.7 3.1 3.7

Treat Boring2 Nutritious

2.6 4.2 3.7

> plot(mean)



> stddev<-apply(mydata2,2,sd)

> print(stddev, digits = 2)

Filling Natural Fibre Sweet Easy Salt Satisfying Energy Fun Kids Soggy2

0.88 0.89 1.00 1.12 0.77 0.83 0.81 0.90 1.26 1.19 1.20

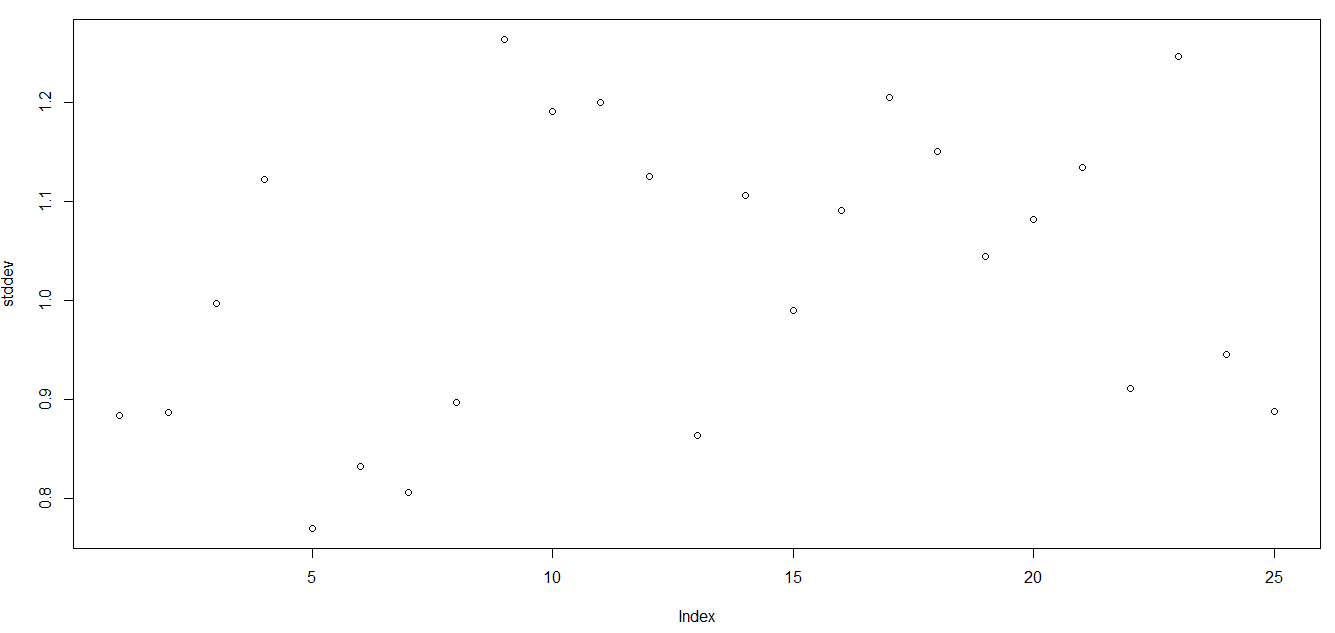
Economical Health Family Calories Plain Crisp Regular Sugar2 Fruit Process2 Quality

1.12 0.86 1.11 0.99 1.09 1.20 1.15 1.04 1.08 1.13 0.91

Treat Boring2 Nutritious

1.25 0.95 0.89

> plot(stddev)



> ske<-apply(mydata2,2,skew)

> print(ske, digits=2)

Filling Natural Fibre Sweet Easy Salt Satisfying Energy Fun Kids Soggy2

-0.54 -0.63 -0.53 0.38 -1.83 0.41 -0.34 -0.41 0.48 -0.79 -0.76

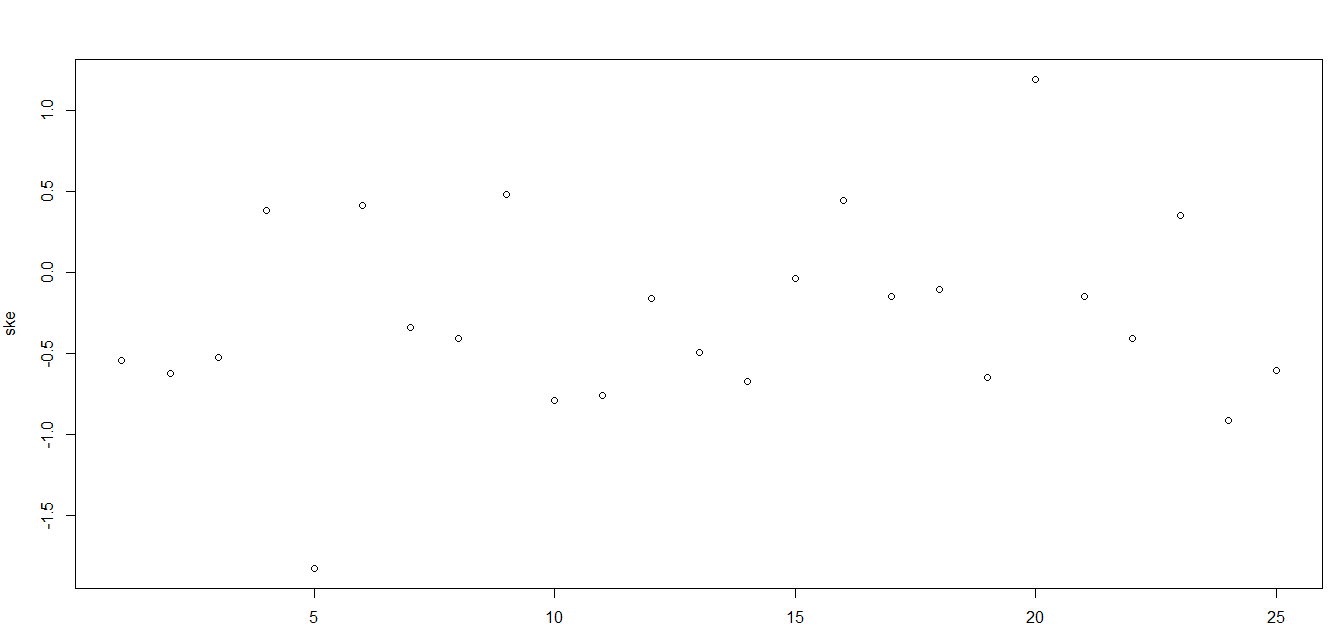
Economical Health Family Calories Plain Crisp Regular Sugar2 Fruit Process2 Quality

-0.16 -0.50 -0.67 -0.04 0.44 -0.15 -0.11 -0.65 1.19 -0.15 -0.41

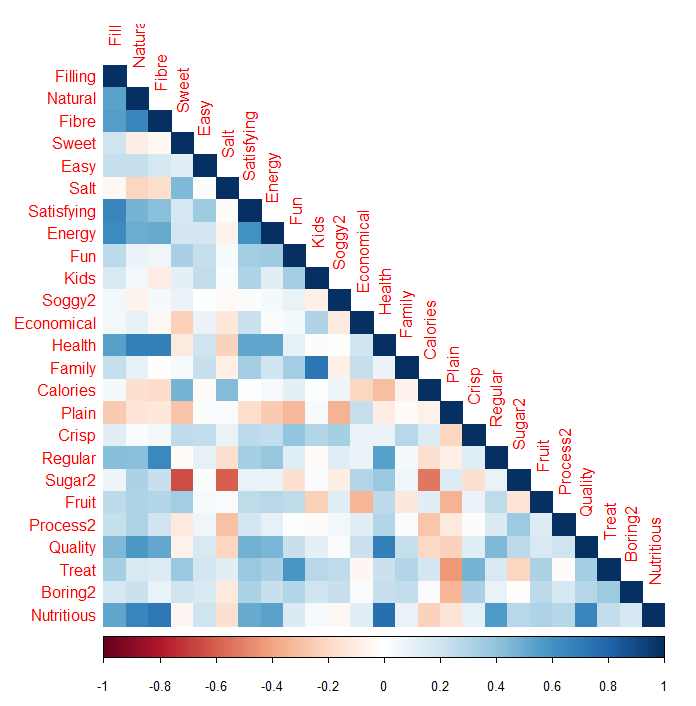
Treat Boring2 Nutritious

0.35 -0.92 -0.61

> plot(ske)



* The correlations of the characteristics are also spread across the spectrum with an absolute range of 0.8 to 0. For example Boring and Sugar have a correlation coefficient of almost zero, while Nutritious and Healthy are highly correlated with a coefficient of 0.8. Refer to graph below for a the correlation chart



1. **Reducing dimensionality: Factor Analysis/ PCA:**

Given our task of reducing the dimensions to a meaningful level, without losing much of the data value, a PCA was executed and the results are as below.

> ev = eigen(cor(mydata2))

> EigenValue=ev$values

> EigenValue

[1] 6.5104814 3.7921753 2.4942279 1.6821942 1.0856935 0.9450867 0.8532528 0.7910547 0.7326378 0.6977062 0.6481540 0.5507242

[13] 0.5314532 0.4874731 0.4168149 0.3869282 0.3640988 0.3608730 0.3061363 0.2755866 0.2628312 0.2428432 0.2183801 0.1986326

[25] 0.1645601

Based on below Eigen values, a Scree plot was plotted

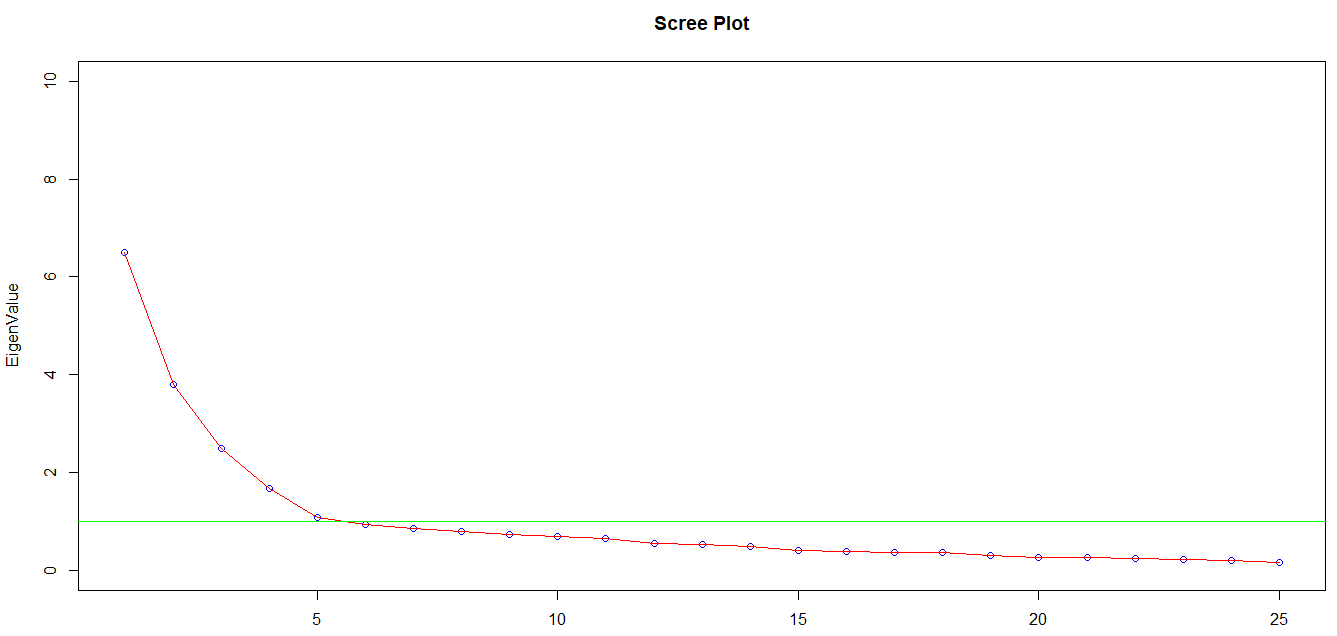
Factor<-c(1:25)

> Scree=data.frame(Factor,EigenValue)

> plot(Scree,main="Scree Plot", col="Blue",ylim=c(0,10)) # plot scree plot

> lines(Scree,col="Red")

> abline(h=1, col="Green")



Based on the above Scree plot and Kaiser Normalization rule (Eigen value>1), we settle with 5 factors. From the profiles we also see that these 5 factors are able to explain ~97% of the data, with decreasing level of importance as indicated by their ‘proportion explained’ scores. See below output

Principal Components Analysis

Call: principal(r = mydata, nfactors = 5, rotate = "none")

Standardized loadings (pattern matrix) based upon correlation matrix

PC1 PC2 PC3 PC4 PC5 h2 u2 com

Filling 0.747 0.100 -0.072 0.228 -0.111 0.638 0.362 1.29

Natural 0.750 -0.256 -0.131 0.131 -0.144 0.683 0.317 1.45

Fibre 0.732 -0.240 -0.332 0.179 0.162 0.761 0.239 1.91

Sweet 0.089 0.776 -0.184 0.185 -0.161 0.704 0.296 1.36

Easy 0.347 0.142 0.270 0.157 0.007 0.238 0.762 2.72

Salt -0.223 0.545 -0.136 0.484 0.132 0.617 0.383 2.60

Satisfying 0.745 0.160 0.170 0.198 -0.105 0.660 0.340 1.40

Energy 0.728 0.135 -0.071 0.170 -0.030 0.583 0.417 1.21

Fun 0.411 0.526 0.256 -0.146 -0.082 0.539 0.461 2.64

Kids 0.218 0.251 0.786 0.109 -0.086 0.748 0.252 1.44

Soggy2 0.110 0.276 -0.179 -0.578 0.499 0.704 0.296 2.74

Economical 0.160 -0.286 0.577 0.108 0.247 0.513 0.487 2.15

Health 0.812 -0.314 -0.125 0.088 0.078 0.788 0.212 1.39

Family 0.317 0.193 0.726 0.024 -0.143 0.687 0.313 1.62

Calories -0.171 0.630 -0.174 0.280 -0.009 0.536 0.464 1.73

Plain -0.329 -0.404 0.249 0.485 0.149 0.591 0.409 3.57

Crisp 0.309 0.490 0.269 -0.240 0.419 0.641 0.359 3.87

Regular 0.620 -0.145 -0.224 0.090 0.397 0.621 0.379 2.20

Sugar2 0.254 -0.747 0.225 -0.261 -0.099 0.751 0.249 1.75

Fruit 0.394 0.287 -0.540 -0.144 -0.294 0.636 0.364 3.27

Process2 0.341 -0.301 -0.006 -0.341 -0.353 0.448 0.552 3.95

Quality 0.752 -0.155 0.037 -0.013 0.091 0.599 0.401 1.12

Treat 0.485 0.588 0.094 -0.195 0.062 0.632 0.368 2.26

Boring2 0.414 0.296 0.133 -0.433 -0.164 0.491 0.509 3.29

Nutritious 0.807 -0.226 -0.161 0.148 0.071 0.754 0.246 1.33

PC1 PC2 PC3 PC4 PC5

SS loadings 6.510 3.792 2.494 1.682 1.086

Proportion Var 0.260 0.152 0.100 0.067 0.043

Cumulative Var 0.260 0.412 0.512 0.579 0.623

Proportion Explained 0.418 0.244 0.160 0.108 0.070

Cumulative Proportion 0.418 0.662 0.822 0.930 1.000

Mean item complexity = 2.2

Test of the hypothesis that 5 components are sufficient.

The root mean square of the residuals (RMSR) is 0.052

with the empirical chi square 377.674 with prob < 2.65e-15

Fit based upon off diagonal values = 0.966

Though the above output was good enough, we also explored the possibility of better scores via a rotated profile, and indeed it was better as indicated below in the table and graph

Principal Components Analysis

Call: principal(r = mydata, nfactors = 5, rotate = "varimax")

Standardized loadings (pattern matrix) based upon correlation matrix

RC1 RC2 RC3 RC4 RC5 h2 u2 com

Filling 0.716 0.088 0.246 0.224 -0.080 0.638 0.362 1.51

Natural 0.759 -0.227 0.076 0.157 -0.157 0.683 0.317 1.39

Fibre 0.854 -0.093 -0.140 0.021 0.054 0.761 0.239 1.09

Sweet 0.030 0.703 0.160 0.427 0.012 0.704 0.296 1.78

Easy 0.261 0.075 0.405 -0.015 0.010 0.238 0.762 1.80

Salt -0.076 0.774 0.002 -0.107 -0.035 0.617 0.383 1.06

Satisfying 0.630 0.058 0.473 0.183 -0.040 0.660 0.340 2.08

Energy 0.686 0.097 0.237 0.217 0.022 0.583 0.417 1.50

Fun 0.142 0.178 0.533 0.389 0.229 0.539 0.461 2.70

Kids -0.037 0.010 0.859 -0.093 -0.017 0.748 0.252 1.03

Soggy2 -0.021 -0.018 -0.139 0.223 0.796 0.704 0.296 1.22

Economical 0.109 -0.291 0.433 -0.473 0.074 0.513 0.487 2.85

Health 0.841 -0.276 0.048 0.039 0.029 0.788 0.212 1.23

Family 0.039 -0.090 0.823 0.006 -0.019 0.687 0.313 1.03

Calories -0.123 0.706 0.021 0.148 -0.001 0.536 0.464 1.15

Plain -0.104 -0.025 -0.013 -0.695 -0.311 0.591 0.409 1.44

Crisp 0.097 0.180 0.421 0.096 0.642 0.641 0.359 2.03

Regular 0.716 -0.047 -0.091 -0.095 0.298 0.621 0.379 1.42

Sugar2 0.193 -0.825 0.029 -0.140 -0.116 0.751 0.249 1.21

Fruit 0.349 0.169 -0.214 0.664 -0.015 0.636 0.364 1.91

Process2 0.189 -0.522 0.035 0.349 -0.128 0.448 0.552 2.21

Quality 0.686 -0.241 0.213 0.083 0.134 0.599 0.401 1.58

Treat 0.244 0.246 0.413 0.432 0.393 0.632 0.368 4.20

Boring2 0.111 -0.139 0.343 0.527 0.254 0.491 0.509 2.53

Nutritious 0.849 -0.167 0.049 0.061 0.016 0.754 0.246 1.10

RC1 RC2 RC3 RC4 RC5

SS loadings 5.509 3.062 3.033 2.364 1.596

Proportion Var 0.220 0.122 0.121 0.095 0.064

Cumulative Var 0.220 0.343 0.464 0.559 0.623

Proportion Explained 0.354 0.197 0.195 0.152 0.103

Cumulative Proportion 0.354 0.551 0.746 0.897 1.000

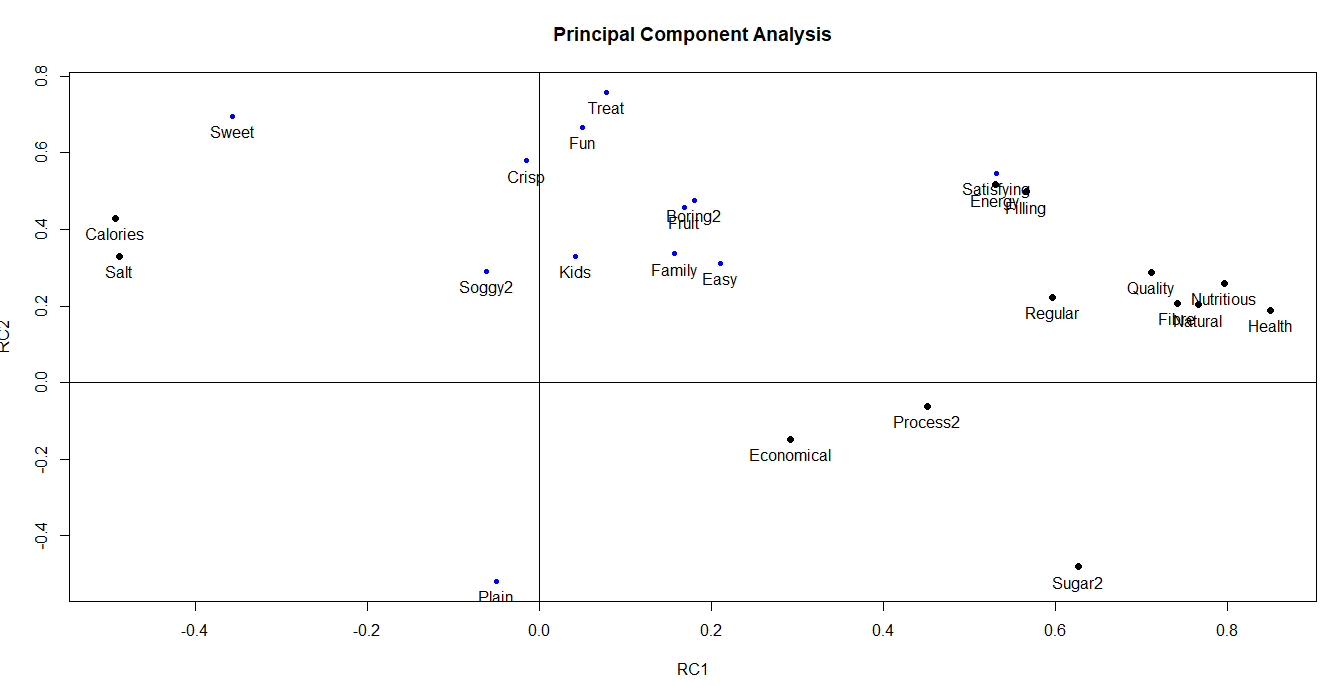
Mean item complexity = 1.7

Test of the hypothesis that 5 components are sufficient.

The root mean square of the residuals (RMSR) is 0.052

with the empirical chi square 377.674 with prob < 2.65e-15

Fit based upon off diagonal values = 0.966



1. **Output interpretation, factor grouping and naming:**

Based on the above loadings, we see that most characteristics (23 of 25), except for **Easy and Treat** are well explained by one of the five factors. We assume a correlation of **0.45,** to be reasonably good to conclude significance. See color coded table below for characteristics and their factor loadings

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **RC1** | **RC2** | **RC3** | **RC4** | **RC5** | h2 | u2 | Com |
| Filling | 0.72 | 0.09 | 0.25 | 0.22 | -0.08 | 0.64 | 0.36 | 1.5 |
| Natural | 0.76 | -0.23 | 0.08 | 0.16 | -0.16 | 0.68 | 0.32 | 1.4 |
| Fibre | 0.85 | -0.09 | -0.14 | 0.02 | 0.05 | 0.76 | 0.24 | 1.1 |
| Satisfying | 0.63 | 0.06 | 0.47 | 0.18 | -0.04 | 0.66 | 0.34 | 2.1 |
| Energy | 0.69 | 0.1 | 0.24 | 0.22 | 0.02 | 0.58 | 0.42 | 1.5 |
| Health | 0.84 | -0.28 | 0.05 | 0.04 | 0.03 | 0.79 | 0.21 | 1.2 |
| Regular | 0.72 | -0.05 | -0.09 | -0.1 | 0.3 | 0.62 | 0.38 | 1.4 |
| Quality | 0.69 | -0.24 | 0.21 | 0.08 | 0.13 | 0.6 | 0.4 | 1.6 |
| Nutritious | 0.85 | -0.17 | 0.05 | 0.06 | 0.02 | 0.75 | 0.25 | 1.1 |
| Sweet | 0.03 | 0.7 | 0.16 | 0.43 | 0.01 | 0.7 | 0.3 | 1.8 |
| Salt | -0.08 | 0.77 | 0 | -0.11 | -0.03 | 0.62 | 0.38 | 1.1 |
| Calories | -0.12 | 0.71 | 0.02 | 0.15 | 0 | 0.54 | 0.46 | 1.2 |
| Sugar2 | 0.19 | -0.82 | 0.03 | -0.14 | -0.12 | 0.75 | 0.25 | 1.2 |
| Process2 | 0.19 | -0.52 | 0.03 | 0.35 | -0.13 | 0.45 | 0.55 | 2.2 |
| Fun | 0.14 | 0.18 | 0.53 | 0.39 | 0.23 | 0.54 | 0.46 | 2.7 |
| Kids | -0.04 | 0.01 | 0.86 | -0.09 | -0.02 | 0.75 | 0.25 | 1 |
| Family | 0.04 | -0.09 | 0.82 | 0.01 | -0.02 | 0.69 | 0.31 | 1 |
| Economical | 0.11 | -0.29 | 0.43 | -0.47 | 0.07 | 0.51 | 0.49 | 2.8 |
| Plain | -0.1 | -0.02 | -0.01 | -0.69 | -0.31 | 0.59 | 0.41 | 1.4 |
| Fruit | 0.35 | 0.17 | -0.21 | 0.66 | -0.01 | 0.64 | 0.36 | 1.9 |
| Boring2 | 0.11 | -0.14 | 0.34 | 0.53 | 0.25 | 0.49 | 0.51 | 2.5 |
| Soggy2 | -0.02 | -0.02 | -0.14 | 0.22 | 0.8 | 0.7 | 0.3 | 1.2 |
| Crisp | 0.1 | 0.18 | 0.42 | 0.1 | 0.64 | 0.64 | 0.36 | 2 |
| Easy | 0.26 | 0.08 | 0.41 | -0.01 | 0.01 | 0.24 | 0.76 | 1.8 |
| Treat | 0.24 | 0.25 | 0.41 | 0.43 | 0.39 | 0.63 | 0.37 | 4.2 |

In addition, we also see that the characteristics explained by each component, clusters quite well into similar traits, as below. Namely- Healthy, Tasty, Kids, Staple food, Texture

|  |  |  |
| --- | --- | --- |
| **Principal component** | **Characteristics explained** | **Characteristics grouped under** |
| RC1 | Filling, Natural, Fibre, Satisfying, Energy, Health, Regular, Quality, Nutritious | **Healthy** |
| RC2 | Sweet, Salt, Calories, Sugar2, Process2 | **Tasty** |
| RC3 | Fun, Kids, Family | **Kids** |
| RC4 | Economical, Plain, Fruit, Boring2 | **Staple food** |
| RC5 | Soggy2, Crisp | **Texture** |

Now that the characteristics have been reduced, we could also use the scores to use them for additional data analysis. The scores are as in the embedded document (double-click to open)

>Rotate$scores

>View(Rotate$scores)



For example, we could use the scores to arrive at an aggregated view on the brand performance/gaps across the grouped characteristic. From our aggregation we see that PMuesli is perceived the best overall across grouped characteristic, while Weetabix is perceived as the worst

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Average Scores (standardized)**  **(Red fill indicates lowest score; Green fill indicates highest score)** | | | | | |
| **Brand** | **RC1**  **Healthy** | **RC2**  **Tasty** | **RC3**  **Kids** | **RC4**  **Staple food** | **RC5**  **Texture** | **Across grouped characteristics** |
| **PMuesli** | 0.62 | 0.63 | (0.31) | 0.77 | 0.15 | 1.85 |
| **CMuesli** | 0.41 | 0.41 | (0.11) | 1.09 | (0.12) | 1.69 |
| **Sustain** | 0.64 | (0.28) | (0.20) | 1.09 | 0.19 | 1.43 |
| **NutriGrain** | (0.44) | 0.83 | 0.60 | 0.02 | 0.38 | 1.39 |
| **Komplete** | 0.50 | 0.22 | (0.98) | 1.10 | (0.15) | 0.68 |
| **JustRight** | (0.05) | 0.29 | (0.44) | 0.70 | (0.01) | 0.50 |
| **CornFlakes** | (0.57) | 0.13 | 0.60 | (0.19) | 0.15 | 0.11 |
| **SpecialK** | (0.31) | (0.20) | 0.16 | (0.33) | 0.36 | (0.33) |
| **RiceBubbles** | (1.20) | (0.51) | 0.70 | (0.17) | 0.23 | (0.96) |
| **AllBran** | 0.50 | (0.23) | (1.05) | (0.77) | 0.52 | (1.03) |
| **Vitabrit** | 0.45 | (0.53) | 0.08 | (0.91) | (0.59) | (1.50) |
| **Weetabix** | 0.33 | (0.44) | (0.18) | (0.65) | (0.75) | (1.70) |

**Problem 2: Estimating Leslie Salt Plot Price**

***What is the problem at hand?***

We have data pertaining to land prices (Price, County, Size, Elevation, Sewer, Date, Flood, and Distance) in a region and we’re expected to arrive at a reasonable price estimate to enable purchase decisions on the property

***What do we know from the data?***

Note: Output and code in R have been represented with blue font, and the full code for this case is as embedded below



1. **Exploratory analysis of the data in R reveals the following:**
2. We have the Sale price per acreage (dependent variable) of plots, along with information on parameters such as (independent variables) County, Size, Elevation, Sewer, Date, Flood, and Distance. **We noticed that the dataset contained some factors as numbers, hence reassigned them as factors**

#Binary into Factor : County and Flood

Dataset\_LeslieSalt$County <- as.factor(Dataset\_LeslieSalt$County)

levels(Dataset\_LeslieSalt$County) <- c("San Mateo", "Santa Clara")

Dataset\_LeslieSalt$Flood <- as.factor(Dataset\_LeslieSalt$Flood)

levels(Dataset\_LeslieSalt$Flood) <- c("No", "Yes")

str(Dataset\_LeslieSalt)

> str(Dataset\_LeslieSalt)

'data.frame': 30 obs. of 8 variables:

$ Price : num 4.5 10.6 1.7 5 5 3.3 5.7 6.2 19.4 3.2 ...

$ County : Factor w/ 2 levels "San Mateo","Santa Clara": 2 2 1 1 1 2 2 2 2 2 ...

$ Size : num 138.4 52 16.1 1695.2 845 ...

$ Elevation: int 10 4 0 1 1 2 4 4 20 0 ...

$ Sewer : int 3000 0 2640 3500 1000 10000 0 0 1300 6000 ...

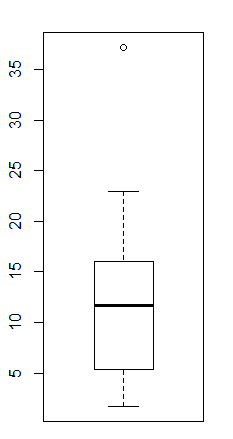
$ Date : int -103 -103 -98 -93 -92 -86 -68 -64 -63 -62 ...

$ Flood : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 1 1 1 1 1 ...

$ Distance : num 0.3 2.5 10.3 14 14 0 0 0 1.2 0 ...

1. On analyzing the Price variable we noticed that there was an outlier, which has been removed in for the analysis. See below code and box-plot for visual representation of the outlier

Pricebox<-boxplot (LeslieSalt$Price)



1. On removing the outlier, we have computed the sample statistics on Price and noticed that it has the a mean of 11.11 and an SD of 6.233, indicating high variation

LeslieSalt.clean<- LeslieSalt[-26,]

Pricebox.clean<-boxplot(LeslieSalt.clean$Price)

PriceMean.clean<-mean(LeslieSalt.clean$Price)

PriceSD.clean<-sd(LeslieSalt.clean$Price)

PriceMean.clean

[1] 11.11

PriceSD.clean

[1] 6.2333

In and regression analysis it is important to understand correlation amongst the variables for 2 reasons.

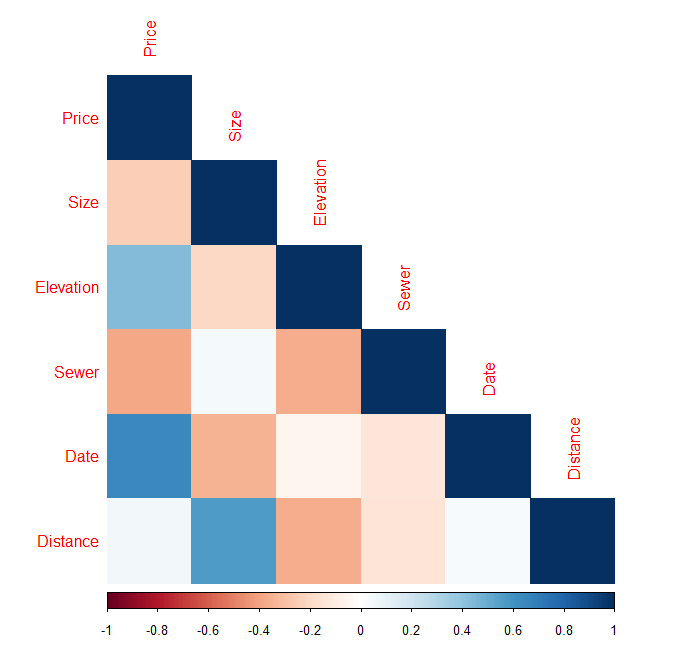
1. Check for correlation between the dependent variable and the independent variables
2. Check for correlations between the independent variable to avoid a multi-collinearity situation.

On plotting the correlations, as indicated below, we noticed that the dependent variable had high-correlation with the independent variables, which is desired, but we also noticed that some of the independent variables were also correlated within themselves, which is something we should be aware of to address in the model building

library(dplyr)

LeslieSalt.clean.matrix <- as.matrix(dplyr::select\_if(LeslieSalt.clean, is.numeric))

corrplot(cor(LeslieSalt.clean.matrix), method = "color", type="lower")



* Price has a positive correlation with Elevation and Date, while a negative correlation with Sewer and a minimal correlation with Size and Distance
* Distance is also seen to be considerably correlated with size

1. **Multiple Linear Regressions- discussion and finalization**

For arriving at the final model for making predictions, we had deployed a systematic approach to maximize the model explainability without over-fitting. We have tried several iterations including study of interaction effects between discrete variables and log transformation of Price. A sample of models evaluated has been discussed below. Model red-flags have been highlighted with red fill and highlights in green fill

Model 1:

We had initiated our analysis with a model incorporating all variables and had obtained a reasonable model, however several variables were insignificant and the residuals were not normally distributed too. Since we had the issue of multi-collinearity amongst the variables, this was expected. Hence we reject this model. Below is the code and output

model1<-lm(Price ~County + Size + Elevation+ Sewer+ Date+ Flood+Distance, data = LeslieSalt.clean)

summary(model1)

shapiro.test(model1$residuals)

Call:

lm(formula = Price ~ County + Size + Elevation + Sewer + Date +

Flood + Distance, data = LeslieSalt.clean)

Residuals:

Min 1Q Median 3Q Max

-3.7059 -2.6043 -0.3876 2.2315 4.7774

Coefficients:

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

(Intercept) 18.6267827 2.9067195 6.408 1.9e-06 \*\*\*

CountySanta Clara -2.6365930 2.8842949 -0.914 0.37056

Size -0.0034320 0.0025420 -1.350 0.19070

Elevation 0.5407713 0.1693998 3.192 0.00421 \*\*

Sewer -0.0005078 0.0003100 -1.638 0.11563

Date 0.1279277 0.0356334 3.590 0.00163 \*\*

FloodYes -7.8400025 2.2885764 -3.426 0.00242 \*\*

Distance 0.4097406 0.2453188 1.670 0.10904

---

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

Residual standard error: 3.145 on 22 degrees of freedom

Multiple R-squared: 0.8069, Adjusted R-squared: 0.7454

F-statistic: 13.13 on 7 and 22 DF, p-value: 1.493e-06

Shapiro-Wilk normality test

data: model1$residuals

W = 0.92974, p-value = 0.04834

Model 2:

Since County, Sewer, Size and Distance have turned out insignificant we’ve tried to create a model without these variables. Although the model has a better normal distribution of the residuals, we notice that the explainability (adjusted Rsq) and significance of the earlier significant variable Flood has dropped. Hence we reject this model

model2<-lm(Price ~Elevation+ Date+ Flood, data = LeslieSalt.clean)

summary(model2)

shapiro.test(model2$residuals)

Call:

lm(formula = Price ~ Elevation + Date + Flood, data = LeslieSalt.clean)

Residuals:

Min 1Q Median 3Q Max

-5.5172 -2.8233 -0.2048 2.6765 6.6460

Coefficients:

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

(Intercept) 19.2331 2.0181 9.530 5.72e-10 \*\*\*

Elevation 0.5477 0.1698 3.226 0.00338 \*\*

Date 0.1696 0.0283 5.994 2.50e-06 \*\*\*

FloodYes -3.6172 1.9813 -1.826 0.07941 .

---

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

Residual standard error: 3.752 on 26 degrees of freedom

Multiple R-squared: 0.6751, Adjusted R-squared: 0.6376

F-statistic: 18.01 on 3 and 26 DF, p-value: 1.57e-06

> shapiro.test(model2$residuals)

Shapiro-Wilk normality test

data: model2$residuals

W = 0.95135, p-value = 0.1837

Model 3 [**Final Model – used for Prediction**]:

Since Sewer and county are relatively away from distance we add them back into the model. The results we notice also confirm our thinking, as the adjusted Rsq (0.73) and the Shapiro test of normality for residuals(W=0.96), both show improved results. Hence we choose this model for making our final predictions

model3<-lm(Price ~Elevation+ Date+ Flood+County + Sewer, data = LeslieSalt.clean)

summary(model3)

shapiro.test(model3$residuals)

> model3<-lm(Price ~Elevation+ Date+ Flood+County + Sewer, data = LeslieSalt.clean)

> summary(model3)

Call:

lm(formula = Price ~ Elevation + Date + Flood + County + Sewer,

data = LeslieSalt.clean)

Residuals:

Min 1Q Median 3Q Max

-5.0186 -2.2651 -0.3114 2.1549 5.1596

Coefficients:

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

(Intercept) 22.0187525 1.9634490 11.214 5.01e-11 \*\*\*

Elevation 0.5086667 0.1726287 2.947 0.00704 \*\*

Date 0.1308357 0.0276699 4.728 8.28e-05 \*\*\*

FloodYes -7.6795702 2.1524916 -3.568 0.00156 \*\*

CountySanta Clara -4.4613706 1.8189990 -2.453 0.02183 \*

Sewer -0.0006846 0.0002789 -2.455 0.02173 \*

---

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

Residual standard error: 3.252 on 24 degrees of freedom

Multiple R-squared: 0.7747, Adjusted R-squared: 0.7278

F-statistic: 16.51 on 5 and 24 DF, p-value: 4.372e-07

> shapiro.test(model3$residuals)

Shapiro-Wilk normality test

data: model3$residuals

W = 0.96517, p-value = 0.4168

We arrive at the below confidence intervals for the variable co-efficients

> confint(model3)

2.5 % 97.5 %

(Intercept) 17.96639283 26.0711121049

Elevation 0.15237850 0.8649548336

Date 0.07372776 0.1879435442

FloodYes -12.12209456 -3.2370458608

CountySanta Clara -8.21560011 -0.7071411037

Sewer -0.00126015 -0.0001090038

Model 4 (**Optional – Part of Research**):

In addition to the above final model, we had also explored the possibility of further improving the fit, via log transformation to y-variable, Price. We indeed had better adjusted Rsq (0.76) and comparable Shapiro test values, but we’ve decided to stick to the earlier model due to 1) simplicity 2) lower risk of over-fitting and 3) better significance for County variable

model4<-lm(log(Price) ~Elevation+ Date+ Flood+County + Sewer, data = LeslieSalt.clean)

summary(model4)

shapiro.test(model4$residuals)

> model4<-lm(log(Price) ~Elevation+ Date+ Flood+County + Sewer, data = LeslieSalt.clean)

> summary(model4)

Call:

lm(formula = log(Price) ~ Elevation + Date + Flood + County +

Sewer, data = LeslieSalt.clean)

Residuals:

Min 1Q Median 3Q Max

-0.54690 -0.21040 0.01803 0.23982 0.56446

Coefficients:

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

(Intercept) 3.419e+00 2.005e-01 17.056 6.42e-15 \*\*\*

Elevation 4.525e-02 1.763e-02 2.567 0.016920 \*

Date 1.403e-02 2.825e-03 4.965 4.54e-05 \*\*\*

FloodYes -9.153e-01 2.198e-01 -4.164 0.000347 \*\*\*

CountySanta Clara -3.592e-01 1.857e-01 -1.934 0.065025 .

Sewer -9.915e-05 2.848e-05 -3.482 0.001926 \*\*

---

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

Residual standard error: 0.3321 on 24 degrees of freedom

Multiple R-squared: 0.8041, Adjusted R-squared: 0.7633

F-statistic: 19.7 on 5 and 24 DF, p-value: 8.585e-08

> shapiro.test(model4$residuals)

Shapiro-Wilk normality test

data: model4$residuals

W = 0.97435, p-value = 0.6636

Equation for multiple linear regression equation which is as per the **final model (Model3)** is

**Price** = 17.96 + 0.15(**Elevation**) + 0.07(**Date**) -12.12 (**FloodYes**) -8.2 (**CountySanta Clara**) – 0.001(**Sewer**)

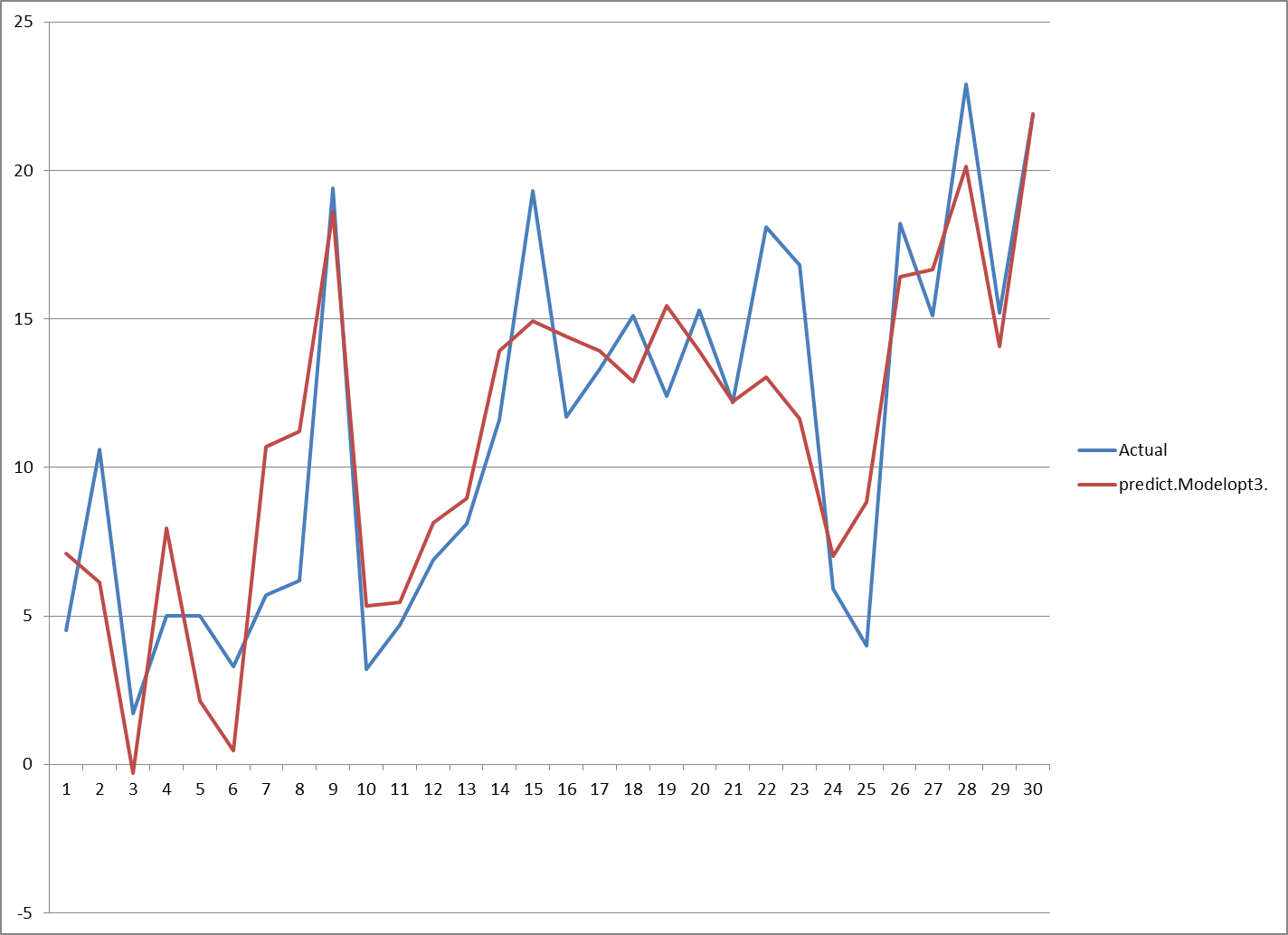
Key inferences:

1. By default Price per acre will be $17960
2. Per 1 Foot Elevation Change will increase the Price per acre by $150 (Keeping other variable constant)
3. Within one month Price per acre will increase by $70 (Keeping other variable constant)
4. Area which is having risk of Flood decrease the price per acre by $12120 (Keeping other variable constant)
5. Moving from San Mateo to Santa Clara will decrease the price per acre to $8200(Keeping other variable constant)
6. Increase in 1 foot distance from the nearest sewer decrease the property price per acre by $1 [very minimal impact though](Keeping other variable constant)

Summary table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Approach** | **Adj-Rsq** | **Shapiro’s test** | **Conclusions** |
| 1 | All variables | 0.7454 | W = 0.92974  p-value = 0.04834 | Some variables are significant, but residuals are not normal. Hence model unfit |
| 2 | Include only significant variables from previous model- Elevation+ Date+ Flood | 0.6376 | W = 0.95135  p-value = 0.1837 | Variable co-efficient are not are significant, and residuals are not normal. Hence model unfit |
| 3 **(Final)** | Include possibly significant variables from correlation plot- Elevation+ Date+ Flood+County + Sewer | 0.7278 | W = 0.96517  p-value = 0.4168 | Variable co-efficient are significant, and residuals are normal. **Hence Final model chosen for prediction**  **Note: this model have taken the variables in account which is justifying the 72.8% of Price impact due to predictor variables** |
| 4 (Optional ) | Log transformation to Price, with variables from Model 3 | 0.7633 | W = 0.97435  p-value = 0.6636 | Variable co-efficients are significant, but for County and residuals are normal. Model could serve as an alternate option |

**Backtracking plot** of the final chosen model reveals good prediction capability for the model

****

1. **Final Price prediction**

Based on the **model3 (Final)**, and our values for below criteria,

* County=Santa Clara (as it is mentioned city of mountain view, we have taken Santa Clara)
* Size=248.1(given in case)
* Elevation=0 (since property is at Mean Sea Level)
* Sewer=0 (since property is dyked, low chance for flooding, hence need for sewer is negligible)
* Date=3 (**we assume sale to happen in about 3 months**)
* Flood=0 (since the property is dyked, we can consider this a non-flooded
* Distance=0 (since this is the base-line for distance)

> newdata<-data.frame(County="Santa Clara", Size=248.1, Elevation=0, Sewer=0, Date=3, Flood="No", Distance=0)

> prediction3<-predict(model3,newdata, interval="confidence")

> prediction3

fit lwr upr

1 17.94989 13.01417 22.8856

We arrive at the final price as **$17950 per acre** (from $13104 to $22886 at 95% confidence level); hence the final total price of the 246.8 acres is estimated to be **$4.43 mn** (from $3.23 mn to $5.65 mn at 95% confidence level