**Factor Analysis**

**Introduction**:

Factor analysis is used to find latent variables or factors among observed variables. In other words, if your data contains many variables, you can use factor analysis to reduce the number of variables. Factor analysis groups variables with similar characteristics together. With factor analysis, you can produce a small number of factors from a large number of variables which is capable of explaining the observed variance in the larger number of variables.

There are three stages in factor analysis:

1. First, a correlation matrix is generated for all the variables. A correlation matrix is a rectangular array of the correlation coefficients of the variables with each other.
2. Second, factors are extracted from the correlation matrix based on the correlation coefficients of the variables.
3. Third, the factors are rotated in order to maximize the relationship between the variables and some of the factors.

**Problem Statement:**

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

**Table 1:**

|  |  |  |
| --- | --- | --- |
| 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. | | |

**Running the Factor Analysis Procedure**

1. **Descriptive Statistics**

**RCode:**

|  |
| --- |
| cereal1<-read.csv("cereal.csv", header = TRUE)  cereal<-cereal1[2:26]  summary(cereal) |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Observations | Minimum | Maximum | Mean | Std. deviation |
| Filling | 235 | 1.000 | 5.000 | 3.881 | 0.884 |
| Natural | 235 | 1.000 | 5.000 | 3.783 | 0.887 |
| Fibre | 235 | 1.000 | 5.000 | 3.528 | 0.997 |
| Sweet | 235 | 1.000 | 5.000 | 2.506 | 1.122 |
| Easy | 235 | 1.000 | 6.000 | 4.532 | 0.775 |
| Salt | 235 | 1.000 | 4.000 | 1.991 | 0.832 |
| Satisfying | 235 | 2.000 | 6.000 | 4.004 | 0.814 |
| Energy | 235 | 1.000 | 5.000 | 3.643 | 0.896 |
| Fun | 235 | 1.000 | 5.000 | 2.617 | 1.263 |
| Kids | 235 | 1.000 | 6.000 | 3.843 | 1.197 |
| Soggy | 235 | 1.000 | 5.000 | 2.255 | 1.200 |
| Economical | 235 | 1.000 | 5.000 | 3.217 | 1.125 |
| Health | 235 | 1.000 | 5.000 | 3.809 | 0.863 |
| Family | 235 | 1.000 | 6.000 | 3.877 | 1.112 |
| Calories | 235 | 1.000 | 5.000 | 2.702 | 0.990 |
| Plain | 235 | 1.000 | 5.000 | 2.268 | 1.090 |
| Crisp | 235 | 1.000 | 6.000 | 3.204 | 1.213 |
| Regular | 235 | 1.000 | 5.000 | 3.072 | 1.151 |
| Sugar | 235 | 1.000 | 5.000 | 2.145 | 1.044 |
| Fruit | 235 | 1.000 | 5.000 | 1.694 | 1.082 |
| Process | 235 | 1.000 | 6.000 | 2.936 | 1.144 |
| Quality | 235 | 1.000 | 5.000 | 3.694 | 0.910 |
| Treat | 235 | 1.000 | 6.000 | 2.630 | 1.255 |
| Boring | 235 | 1.000 | 5.000 | 1.830 | 0.946 |
| Nutritious | 235 | 1.000 | 5.000 | 3.664 | 0.888 |

**Interpretation**:

The first output from the analysis is a table of descriptive statistics for all the variables under investigation. Typically, the *mean*, *standard deviation*and *number of respondents* (N) who participated in the study are given. It is important to check N, missing data, or any variable going beyond expected values. Looking at the *maximum*, the variable having max 6 have to be replaced by 5 as the likert scale used is a 5 point.

**RCode:**

|  |
| --- |
| cereal[cereal==6] <- 5 |

1. **The Correlation matrix**

**RCode:**

|  |
| --- |
| corrpaste<-cor(cereal)  corrpaste |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variables | Filling | Natural | Fibre | Sweet | Easy | Salt | Satisfying | Energy | Fun | Kids | Soggy | Economical | Health | Family | Calories | Plain | Crisp | Regular | Sugar | Fruit | Process | Quality | Treat | Boring | Nutritious |
| Filling | **1** | **0.540** | **0.552** | **0.190** | **0.237** | -0.036 | **0.654** | **0.637** | **0.265** | **0.160** | -0.060 | 0.052 | **0.547** | **0.234** | 0.047 | **-0.251** | 0.127 | **0.420** | -0.079 | **0.261** | **-0.234** | **0.443** | **0.340** | **-0.178** | **0.526** |
| Natural | **0.540** | **1** | **0.652** | -0.091 | **0.231** | **-0.217** | **0.466** | **0.494** | 0.082 | 0.060 | 0.068 | 0.103 | **0.688** | 0.107 | **-0.162** | **-0.139** | 0.021 | **0.418** | **-0.317** | **0.300** | **-0.308** | **0.579** | **0.170** | **-0.218** | **0.651** |
| Fibre | **0.552** | **0.652** | **1** | -0.037 | **0.176** | **-0.175** | **0.415** | **0.504** | 0.063 | -0.093 | -0.042 | -0.034 | **0.684** | -0.008 | **-0.187** | -0.123 | 0.051 | **0.648** | **-0.226** | **0.293** | **-0.195** | **0.513** | **0.143** | -0.099 | **0.713** |
| Sweet | **0.190** | -0.091 | -0.037 | **1** | 0.125 | **0.444** | **0.180** | **0.185** | **0.327** | 0.119 | -0.084 | **-0.240** | -0.116 | 0.039 | **0.467** | **-0.290** | **0.260** | -0.025 | **0.648** | **0.347** | 0.115 | -0.078 | **0.375** | **-0.200** | -0.047 |
| Easy | **0.237** | **0.231** | **0.176** | 0.125 | **1** | 0.014 | **0.351** | **0.182** | **0.240** | **0.243** | -0.008 | 0.089 | **0.204** | **0.225** | -0.023 | 0.019 | **0.241** | 0.106 | -0.016 | 0.036 | -0.066 | **0.165** | **0.185** | **-0.170** | **0.204** |
| Salt | -0.036 | **-0.217** | **-0.175** | **0.444** | 0.014 | **1** | -0.013 | -0.067 | 0.033 | 0.024 | 0.024 | -0.126 | **-0.228** | -0.089 | **0.438** | 0.021 | 0.096 | **-0.165** | **0.592** | 0.026 | **0.298** | **-0.218** | 0.121 | 0.112 | **-0.160** |
| Satisfying | **0.654** | **0.466** | **0.415** | **0.180** | **0.351** | -0.013 | **1** | **0.603** | **0.348** | **0.303** | -0.013 | **0.212** | **0.522** | **0.345** | 0.005 | **-0.180** | **0.264** | **0.332** | -0.091 | **0.255** | **-0.187** | **0.472** | **0.370** | **-0.320** | **0.502** |
| Energy | **0.637** | **0.494** | **0.504** | **0.185** | **0.182** | -0.067 | **0.603** | **1** | **0.350** | **0.130** | -0.046 | 0.026 | **0.524** | **0.191** | 0.034 | **-0.256** | **0.249** | **0.386** | -0.086 | **0.274** | -0.104 | **0.457** | **0.324** | **-0.223** | **0.536** |
| Fun | **0.265** | 0.082 | 0.063 | **0.327** | **0.240** | 0.033 | **0.348** | **0.350** | **1** | **0.345** | -0.099 | 0.041 | 0.101 | **0.347** | 0.113 | **-0.322** | **0.399** | **0.137** | **0.165** | **0.251** | -0.009 | **0.225** | **0.585** | **-0.298** | **0.155** |
| Kids | **0.160** | 0.060 | -0.093 | 0.119 | **0.243** | 0.024 | **0.303** | **0.130** | **0.345** | **1** | 0.089 | **0.304** | -0.014 | **0.724** | 0.010 | 0.030 | **0.294** | -0.026 | -0.022 | **-0.234** | 0.014 | 0.112 | **0.276** | **-0.195** | 0.033 |
| Soggy | -0.060 | 0.068 | -0.042 | -0.084 | -0.008 | 0.024 | -0.013 | -0.046 | -0.099 | 0.089 | **1** | 0.117 | 0.006 | 0.083 | -0.080 | **0.346** | **-0.337** | **-0.137** | -0.094 | **-0.137** | 0.060 | -0.030 | **-0.256** | **0.227** | 0.033 |
| Economical | 0.052 | 0.103 | -0.034 | **-0.240** | 0.089 | -0.126 | **0.212** | 0.026 | 0.041 | **0.304** | 0.117 | **1** | **0.193** | **0.232** | **-0.210** | **0.231** | 0.091 | 0.080 | **-0.293** | **-0.338** | **-0.129** | **0.215** | -0.042 | -0.021 | **0.129** |
| Health | **0.547** | **0.688** | **0.684** | -0.116 | **0.204** | **-0.228** | **0.522** | **0.524** | 0.101 | -0.014 | 0.006 | **0.193** | **1** | 0.082 | **-0.307** | -0.100 | 0.082 | **0.543** | **-0.377** | **0.266** | **-0.293** | **0.686** | **0.215** | **-0.229** | **0.758** |
| Family | **0.234** | 0.107 | -0.008 | 0.039 | **0.225** | -0.089 | **0.345** | **0.191** | **0.347** | **0.724** | 0.083 | **0.232** | 0.082 | **1** | -0.066 | -0.028 | **0.279** | 0.044 | -0.062 | -0.126 | -0.027 | **0.237** | **0.291** | **-0.250** | 0.091 |
| Calories | 0.047 | **-0.162** | **-0.187** | **0.467** | -0.023 | **0.438** | 0.005 | 0.034 | 0.113 | 0.010 | -0.080 | **-0.210** | **-0.307** | -0.066 | **1** | -0.076 | **0.143** | **-0.165** | **0.526** | 0.126 | **0.271** | **-0.201** | **0.190** | -0.027 | **-0.226** |
| Plain | **-0.251** | **-0.139** | -0.123 | **-0.290** | 0.019 | 0.021 | **-0.180** | **-0.256** | **-0.322** | 0.030 | **0.346** | **0.231** | -0.100 | -0.028 | -0.076 | **1** | **-0.210** | -0.080 | **-0.147** | **-0.343** | 0.115 | **-0.227** | **-0.432** | **0.331** | **-0.145** |
| Crisp | 0.127 | 0.021 | 0.051 | **0.260** | **0.241** | 0.096 | **0.264** | **0.249** | **0.399** | **0.294** | **-0.337** | 0.091 | 0.082 | **0.279** | **0.143** | **-0.210** | **1** | **0.134** | **0.164** | 0.090 | 0.010 | **0.130** | **0.460** | **-0.326** | 0.103 |
| Regular | **0.420** | **0.418** | **0.648** | -0.025 | 0.106 | **-0.165** | **0.332** | **0.386** | **0.137** | -0.026 | **-0.137** | 0.080 | **0.543** | 0.044 | **-0.165** | -0.080 | **0.134** | **1** | -0.091 | **0.255** | **-0.150** | **0.441** | **0.168** | -0.095 | **0.568** |
| Sugar | -0.079 | **-0.317** | **-0.226** | **0.648** | -0.016 | **0.592** | -0.091 | -0.086 | **0.165** | -0.022 | -0.094 | **-0.293** | **-0.377** | -0.062 | **0.526** | **-0.147** | **0.164** | -0.091 | **1** | **0.145** | **0.366** | **-0.263** | **0.213** | -0.001 | **-0.275** |
| Fruit | **0.261** | **0.300** | **0.293** | **0.347** | 0.036 | 0.026 | **0.255** | **0.274** | **0.251** | **-0.234** | **-0.137** | **-0.338** | **0.266** | -0.126 | 0.126 | **-0.343** | 0.090 | **0.255** | **0.145** | **1** | **-0.142** | **0.165** | **0.314** | **-0.260** | **0.306** |
| Process | **-0.234** | **-0.308** | **-0.195** | 0.115 | -0.066 | **0.298** | **-0.187** | -0.104 | -0.009 | 0.014 | 0.060 | **-0.129** | **-0.293** | -0.027 | **0.271** | 0.115 | 0.010 | **-0.150** | **0.366** | **-0.142** | **1** | **-0.190** | 0.015 | **0.172** | **-0.286** |
| Quality | **0.443** | **0.579** | **0.513** | -0.078 | **0.165** | **-0.218** | **0.472** | **0.457** | **0.225** | 0.112 | -0.030 | **0.215** | **0.686** | **0.237** | **-0.201** | **-0.227** | **0.130** | **0.441** | **-0.263** | **0.165** | **-0.190** | **1** | **0.332** | **-0.284** | **0.660** |
| Treat | **0.340** | **0.170** | **0.143** | **0.375** | **0.185** | 0.121 | **0.370** | **0.324** | **0.585** | **0.276** | **-0.256** | -0.042 | **0.215** | **0.291** | **0.190** | **-0.432** | **0.460** | **0.168** | **0.213** | **0.314** | 0.015 | **0.332** | **1** | **-0.363** | **0.245** |
| Boring | **-0.178** | **-0.218** | -0.099 | **-0.200** | **-0.170** | 0.112 | **-0.320** | **-0.223** | **-0.298** | **-0.195** | **0.227** | -0.021 | **-0.229** | **-0.250** | -0.027 | **0.331** | **-0.326** | -0.095 | -0.001 | **-0.260** | **0.172** | **-0.284** | **-0.363** | **1** | **-0.170** |
| Nutritious | **0.526** | **0.651** | **0.713** | -0.047 | **0.204** | **-0.160** | **0.502** | **0.536** | **0.155** | 0.033 | 0.033 | **0.129** | **0.758** | 0.091 | **-0.226** | **-0.145** | 0.103 | **0.568** | **-0.275** | **0.306** | **-0.286** | **0.660** | **0.245** | **-0.170** | **1** |

*Only part of correlation matrix included so font would not be too small to read*

**Interpretation**:

The next output from the analysis is the correlation coefficient. A correlation matrix is simply a rectangular array of numbers which gives the correlation coefficients between a single variable and every other variable in the investigation. The correlation coefficient between a variable and itself is always 1; hence the principal diagonal of the correlation matrix contains 1s. The correlation coefficients above and below the principal diagonal are the same.

The matrix shows how each of the 25 items is associated with each of the other 24. Note that some of the correlations are high (e.g., + or −.60 or greater) and some are low (i.e., near zero). Relatively high correlations indicate that two items are associated and will probably be grouped together by the factor analysis. Items with low correlations (e.g., ≤.20) usually will not have high loadings on the same factor.

1. **Kaiser-Meyer-Olkin (KMO) and Bartlett's Test**

**RCode:**

|  |
| --- |
| bartlettTest = cortest.bartlett(corrpaste, nrow(cereal))  print(bartlettTest)  kmoTest = KMO(corrpaste)  print(kmoTest) |

KMO and Bartlett's Test:

|  |  |
| --- | --- |
| Kaiser-Meyer-Olkin Measure of Sampling Adequacy.  Bartlett's Test of Sphericity Chi-Square  df  p.value | 0.855  2877.739  300  0 |

**Interpretation**:

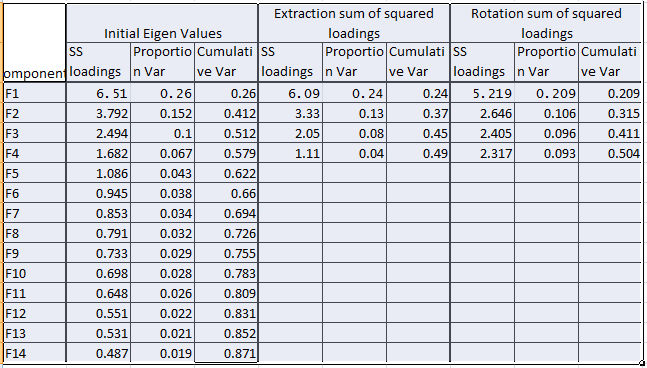
The next item from the output is the Kaiser-Meyer-Olkin (KMO) and Bartlett's test. The KMO measures the sampling adequacy which should be greater than 0.6 for a satisfactory factor analysis to proceed. Looking at the table below, the KMO measure is 0.85. From the same table, we can see that the Bartlett's test of sphericity is also significant. That is, its associated probability is less than 0.05. Thus the null hypothesis is rejected (The null hypothesis is that the correlation matrix is an identity matrix i.e. there is no scope for dimensionality reduction.). Thus, the dimensionality reduction is a possibility using PCA/FA.

1. **No. of Factors and Total Variance Explained**

**RCode:**

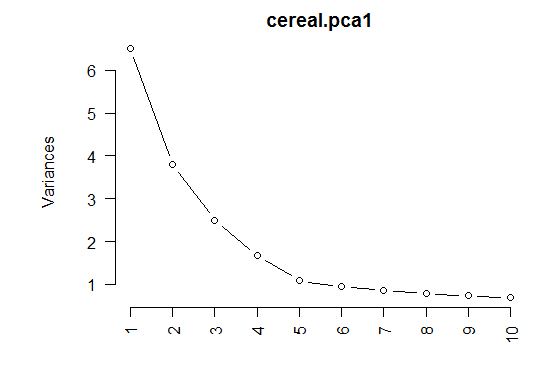
|  |
| --- |
| pc1 <- principal(cereal, nfactors = length(cereal), rotate = "none")  pc1 |

The next item shows all the factors extractable from the analysis along with their eigenvalues, the percent of variance attributable to each factor, and the cumulative variance of the factor and the previous factors. Notice that the first factor accounts for 26% of the variance, the second 15.2% and the third, fourth for 10%, 6.7% respectively and so on.



**Scree Plot**

The scree plot is a graph of the eigenvalues against all the factors. The graph is useful for determining how many factors to retain. The point of interest is where the curve starts to flatten. It can be seen that the curve begins to flatten between factors 4 and 5. Note also that factor 5 has an eigen value of around 1 and not much significant , so only 4 factors have been retained.



1. **Component (Factor) Matrix**

**RCode:**

|  |
| --- |
| pcal2<-fa(cereal,nfactors = 4,rotate="none", scores=TRUE, fm = "pa")  pcal2 |

The table below shows the loadings of the all variables on the 4 factors extracted. The higher the absolute value of the loading, the more the factor contributes to the variable.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | F1 | F2 | F3 | F4 | F5 | Initial communality | Final communality | Specific variance |
| Filling | **0.723** | -0.109 | 0.060 | 0.193 | -0.130 | 0.635 | 0.592 | 0.408 |
| Natural | **0.734** | 0.235 | 0.112 | 0.109 | -0.140 | 0.636 | 0.638 | 0.362 |
| Fibre | **0.728** | 0.228 | 0.314 | 0.159 | 0.147 | 0.703 | 0.727 | 0.273 |
| Sweet | 0.081 | **-0.750** | 0.191 | 0.163 | -0.124 | 0.612 | 0.648 | 0.352 |
| Easy | **0.311** | -0.126 | -0.194 | 0.098 | 0.017 | 0.221 | 0.160 | 0.840 |
| Salt | -0.214 | **-0.500** | 0.141 | 0.404 | 0.112 | 0.446 | 0.491 | 0.509 |
| Satisfying | **0.722** | -0.168 | -0.158 | 0.170 | -0.110 | 0.614 | 0.615 | 0.385 |
| Energy | **0.697** | -0.137 | 0.059 | 0.124 | -0.058 | 0.572 | 0.527 | 0.473 |
| Fun | 0.383 | **-0.488** | -0.217 | -0.142 | -0.036 | 0.481 | 0.453 | 0.547 |
| Kids | 0.211 | -0.261 | **-0.766** | 0.130 | -0.050 | 0.632 | 0.719 | 0.281 |
| Soggy | -0.099 | 0.242 | -0.145 | **0.450** | -0.284 | 0.314 | 0.372 | 0.628 |
| Economical | 0.150 | 0.242 | **-0.468** | 0.100 | 0.158 | 0.372 | 0.334 | 0.666 |
| Health | **0.814** | 0.301 | 0.116 | 0.074 | 0.076 | 0.750 | 0.777 | 0.223 |
| Family | 0.303 | -0.197 | **-0.676** | 0.038 | -0.090 | 0.595 | 0.597 | 0.403 |
| Calories | -0.165 | **-0.564** | 0.162 | 0.212 | -0.024 | 0.417 | 0.417 | 0.583 |
| Plain | -0.308 | 0.370 | -0.207 | **0.418** | 0.125 | 0.408 | 0.465 | 0.535 |
| Crisp | 0.288 | **-0.457** | -0.228 | -0.212 | 0.328 | 0.404 | 0.496 | 0.504 |
| Regular | **0.593** | 0.125 | 0.192 | 0.063 | 0.294 | 0.521 | 0.494 | 0.506 |
| Sugar | -0.258 | **-0.738** | 0.246 | 0.258 | 0.126 | 0.664 | 0.754 | 0.246 |
| Fruit | 0.374 | -0.266 | **0.473** | -0.153 | -0.225 | 0.488 | 0.509 | 0.491 |
| Process | **-0.312** | -0.252 | 0.007 | 0.226 | 0.179 | 0.294 | 0.244 | 0.756 |
| Quality | **0.724** | 0.134 | -0.042 | -0.027 | 0.073 | 0.613 | 0.550 | 0.450 |
| Treat | 0.463 | **-0.562** | -0.084 | -0.203 | 0.063 | 0.555 | 0.583 | 0.417 |
| Boring | **-0.380** | 0.267 | 0.112 | 0.328 | 0.086 | 0.338 | 0.343 | 0.657 |
| Nutritious | **0.802** | 0.210 | 0.149 | 0.129 | 0.082 | 0.711 | 0.733 | 0.267 |

1. **Rotated Component (Factor) Matrix**

**RCode:**

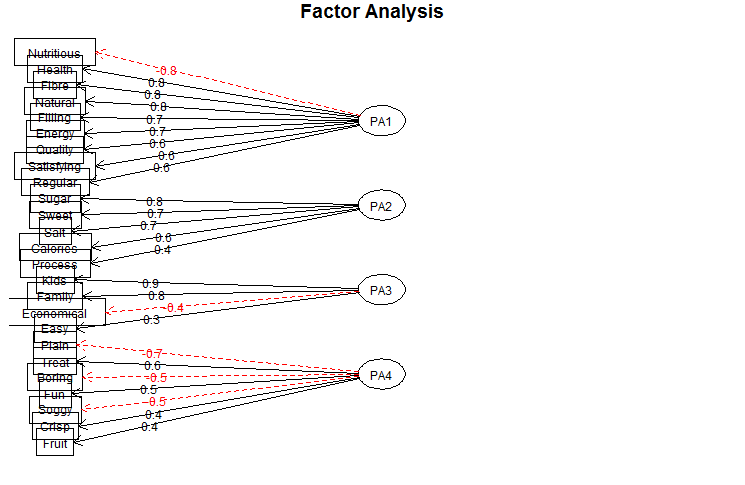
|  |
| --- |
| pcal3<-fa(cereal,nfactors=4,rotate="varimax", scores=TRUE, fm="pa")  pcal3 |

The idea of rotation is to reduce the number factors on which the variables under investigation have high loadings. Rotation does not actually change anything but makes the interpretation of the analysis easier. Looking at the table below, we can see that:

1. “*Filling, Natural, Fibre, Satisfying, Energy, Health, Regular, Quality and Nutritious”* are substantially loaded on Factor 1.
2. “*Sweet, Salt, Calories, Sugar and Process”* are substantially loaded on Factor 2.
3. “*Easy, Kids, Economical and Family”* are substantially loaded on Factor 3.
4. “*Fun, Soggy, Plain, Crisp, Fruit, Treat and Boring”* are substantially loaded on Factor (Component) 4.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | D1 | D2 | D3 | D4 |
| Filling | **0.710** | 0.094 | 0.199 | 0.150 |
| Natural | **0.756** | -0.207 | 0.052 | 0.032 |
| Fibre | **0.822** | -0.118 | -0.124 | 0.018 |
| Sweet | 0.073 | **0.713** | 0.066 | 0.339 |
| Easy | 0.237 | 0.050 | **0.309** | 0.072 |
| Salt | -0.095 | **0.680** | 0.009 | -0.082 |
| Satisfying | **0.633** | 0.069 | 0.409 | 0.176 |
| Energy | **0.661** | 0.083 | 0.189 | 0.209 |
| Fun | 0.170 | 0.185 | 0.402 | **0.477** |
| Kids | -0.021 | 0.033 | **0.845** | 0.028 |
| Soggy | 0.045 | 0.032 | 0.119 | **-0.524** |
| Economical | 0.068 | -0.284 | **0.423** | -0.214 |
| Health | **0.827** | -0.289 | 0.047 | 0.048 |
| Family | 0.066 | -0.056 | **0.756** | 0.103 |
| Calories | -0.114 | **0.624** | -0.013 | 0.118 |
| Plain | -0.152 | -0.071 | 0.082 | **-0.644** |
| Crisp | 0.065 | 0.134 | 0.367 | **0.481** |
| Regular | **0.623** | -0.102 | -0.038 | 0.092 |
| Sugar | -0.184 | **0.822** | -0.062 | 0.159 |
| Fruit | 0.384 | 0.195 | -0.289 | **0.435** |
| Process | -0.242 | **0.369** | 0.012 | -0.130 |
| Quality | **0.648** | -0.242 | 0.193 | 0.171 |
| Treat | 0.250 | 0.234 | 0.311 | **0.605** |
| Boring | -0.170 | 0.059 | -0.218 | **-0.506** |
| Nutritious | **0.830** | -0.177 | 0.046 | 0.056 |

Also, can be shown diagrammatically below:

**7.Adequacy of the Model**

* Goodness of fit:

1. Tucker Lewis Index of factoring reliability = 0.89 (indicates good reliability. Reliability is a value between 0 and 1 with a larger value indicating better reliability.)

* Residual Fit Statistics

1. RMSEA index = 0.067. <0.06 is excellent (ideal is less than 0.06 and not too high)
2. The root mean square of the residuals (RMSR) is 0.04. <0.06 is excellent.

Looking at the metrics, we can conclude that we have an acceptable model.

1. **Reliability of Factors**

**RCode:**

|  |
| --- |
| factor1 <- c(2,3,4,8,9,14,19,23,26)  factor2 <- c(5,7,16,20,22)  factor3 <- c(6,11,13,15)  factor4 <- c(10,12,17,18,21,24,25)  factor1alpha <- psych::alpha(cereal1[,factor1], check.keys = TRUE)  factor2alpha <- psych::alpha(cereal1[,factor2], check.keys = TRUE)  factor3alpha <- psych::alpha(cereal1[,factor3], check.keys = TRUE)  factor4alpha <- psych::alpha(cereal1[,factor4], check.keys = TRUE) |

factor1alpha$total$raw\_alpha: 0.912

factor2alpha$total$raw\_alpha: 0.772

factor3alpha$total$raw\_alpha: 0.654

factor4alpha$total$raw\_alpha: 0.754

As the alpha values are close to and greater than >0.7, the factors are reliable.

|  |  |  |
| --- | --- | --- |
| Factor1 | Nutritional Value | Filling, Natural, Fibre, Satisfying, Energy, Health, Regular, Quality and Nutritious |
| Factor2 | Fat Content | Sweet, Salt, Calories, Sugar and Process |
| Factor3 | Value for money | Easy, Kids, Economical and Family |
| Factor4 | Taste & Texture | Fun, Soggy, Plain, Crisp, Fruit, Treat and Boring |

1. **Creating Average Factor Scores grouped by the cereal**

**RCode:**

|  |
| --- |
| cereal1$factor1Score <- apply(cereal1[,factor1],1,mean)  cereal1$factor2Score <- apply(cereal1[,factor2],1,mean)  cereal1$factor3Score <- apply(cereal1[,factor3],1,mean)  cereal1$factor4Score <- apply(cereal1[,factor4],1,mean)  colnames(cereal1)[27:30] <-c("Nutritional Value", "Fat Content", "Value for money", "Taste & Texture")  aggregateCereal<-aggregate(cereal1[,27:30], list(cereal1[,1]), mean)  format(aggregateCereal, digits = 2) |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Brands | Nutritional Value | Fat Content | Value for money | Taste & Texture |
| 1 | AllBran | 3.9 | 2.2 | 3.4 | 2 |
| 2 | CMuesli | 4 | 2.8 | 3.8 | 2.5 |
| 3 | CornFlakes | 3.3 | 2.7 | 4.2 | 2.3 |
| 4 | JustRight | 3.6 | 2.7 | 3.5 | 2.4 |
| 5 | Komplete | 4 | 2.6 | 3 | 2.5 |
| 6 | NutriGrain | 3.4 | 3.1 | 4.1 | 2.5 |
| 7 | PMuesli | 4.1 | 2.9 | 3.6 | 2.6 |
| 8 | RiceBubbles | 2.9 | 2.2 | 4.3 | 2.4 |
| 9 | SpecialK | 3.5 | 2.3 | 4 | 2.3 |
| 10 | Sustain | 4.2 | 2.2 | 3.6 | 2.7 |
| 11 | Vitabrit | 3.9 | 1.9 | 4.1 | 2.2 |
| 12 | Weetabix | 3.9 | 2.1 | 3.9 | 2.1 |
|  |  |  |  |  |  |

# Problem Statement

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.

|  |  |
| --- | --- |
| 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

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Price** | **County** | **Size** | **Elevation** | **Sewer** | **Date** | **Flood** | **Distance** |
| 4.50 | 1 | 138.40 | 10 | 3000 | -103 | 0 | 0.30 |
| 10.60 | 1 | 52.00 | 4 | 0 | -103 | 0 | 2.50 |
| 1.70 | 0 | 16.10 | 0 | 2640 | -98 | 1 | 10.30 |
| 5.00 | 0 | 1695.20 | 1 | 3500 | -93 | 0 | 14.00 |
| 5.00 | 0 | 845.00 | 1 | 1000 | -92 | 1 | 14.00 |
| 3.30 | 1 | 6.90 | 2 | 10000 | -86 | 0 | 0.00 |
| 5.70 | 1 | 105.90 | 4 | 0 | -68 | 0 | 0.00 |
| 6.20 | 1 | 56.60 | 4 | 0 | -64 | 0 | 0.00 |
| 19.40 | 1 | 51.40 | 20 | 1300 | -63 | 0 | 1.20 |
| 3.20 | 1 | 22.10 | 0 | 6000 | -62 | 0 | 0.00 |
| 4.70 | 1 | 22.10 | 0 | 6000 | -61 | 0 | 0.00 |
| 6.90 | 1 | 27.70 | 3 | 4500 | -60 | 0 | 0.00 |
| 8.10 | 1 | 18.60 | 5 | 5000 | -59 | 0 | 0.50 |
| 11.60 | 1 | 69.90 | 8 | 0 | -59 | 0 | 4.40 |
| 19.30 | 1 | 145.70 | 10 | 0 | -59 | 0 | 4.20 |
| 11.70 | 1 | 77.20 | 9 | 0 | -59 | 0 | 4.50 |
| 13.30 | 1 | 26.20 | 8 | 0 | -59 | 0 | 4.70 |
| 15.10 | 1 | 102.30 | 6 | 0 | -59 | 0 | 4.90 |
| 12.40 | 1 | 49.50 | 11 | 0 | -59 | 0 | 4.60 |
| 15.30 | 1 | 12.20 | 8 | 0 | -59 | 0 | 5.00 |
| 12.20 | 0 | 320.60 | 0 | 4000 | -54 | 0 | 16.50 |
| 18.10 | 1 | 9.90 | 5 | 0 | -54 | 0 | 5.20 |
| 16.80 | 1 | 15.30 | 2 | 0 | -53 | 0 | 5.50 |
| 5.90 | 0 | 55.20 | 0 | 1320 | -49 | 1 | 11.90 |
| 4.00 | 0 | 116.20 | 2 | 900 | -45 | 1 | 5.50 |
| 37.20 | 0 | 15.00 | 5 | 0 | -39 | 0 | 7.20 |
| 18.20 | 0 | 23.40 | 5 | 4420 | -39 | 0 | 5.50 |
| 15.10 | 0 | 132.80 | 2 | 2640 | -35 | 0 | 10.20 |
| 22.90 | 0 | 12.00 | 5 | 3400 | -16 | 0 | 5.50 |
| 15.20 | 0 | 67.00 | 2 | 900 | -5 | 1 | 5.50 |
| 21.90 | 0 | 30.80 | 2 | 900 | -4 | 0 | 5.50 |

# **Data**

# Nature of Variables

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

* Dependent Variable is Price
* Independent Variables are county , size, Elevation, Sewer, Date, Flood. Distance

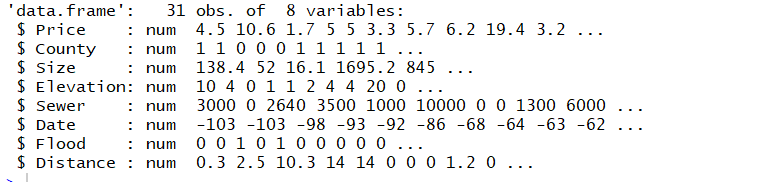
setwd("C:/Users/Rajeswari\_S/Documents/R\_Tutorials/Group\_Assignment")

library(xlsx)

data <- read.xlsx("Dataset\_LeslieSalt.xlsx",1)

attach(data)

str(data)



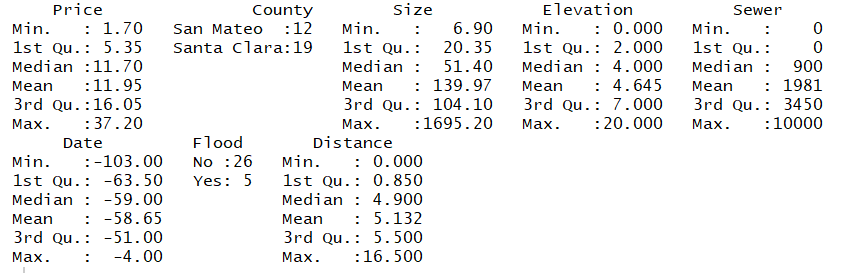
# **Feature Engineering**

Convert factor variables

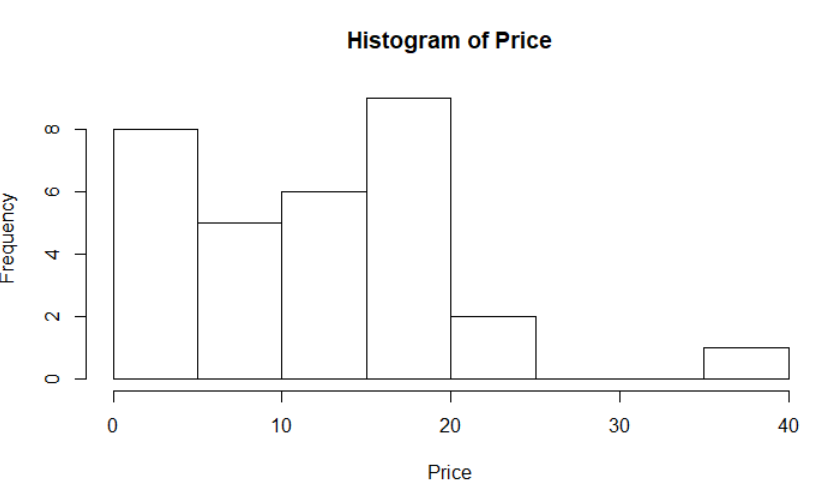
data$County <- factor(data$County, levels=c("0","1"), labels=c("San Mateo", "Santa Clara" ))

data$Flood <- factor(data$Flood, levels=c("0","1"), labels=c("No", "Yes" ))

summary(data)



hist(Price)

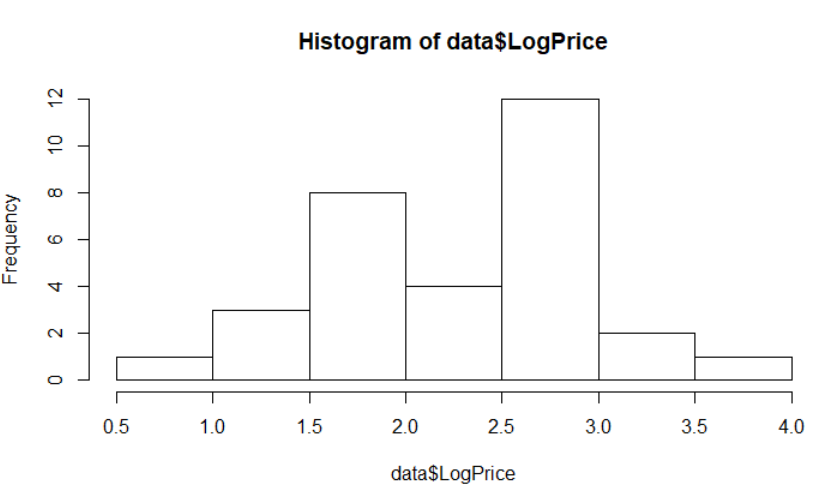


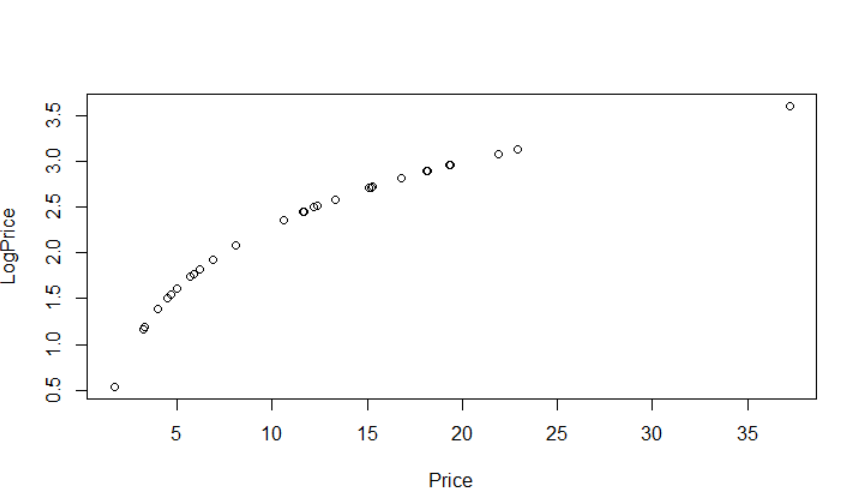
From the above histogram, Price (dependent variable) is right skewed. Such skewness is problematic, violates the assumptions of regression model.

To reduce skewness, transform the dependent variable to log form

data$LogPrice=log(Price)

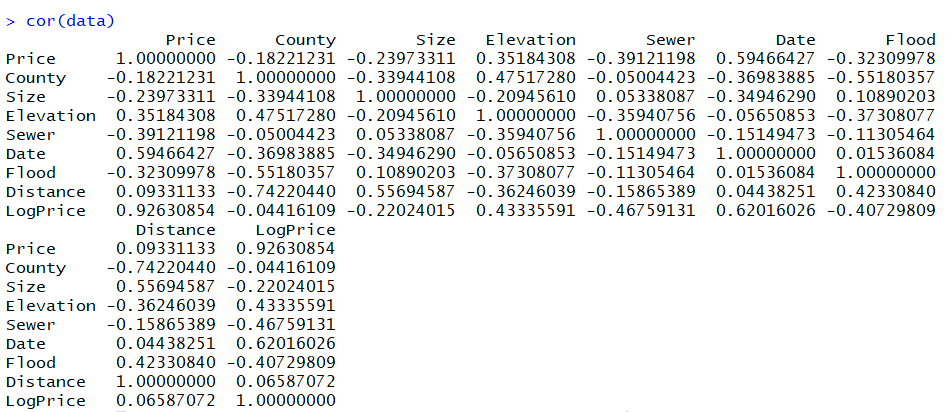
hist(data$LogPrice)





## **Correlation among variables**

cor(data)



1. From the above correlation matrix, elevation, sewer, flood and date are highly correlated with the LogPrice.

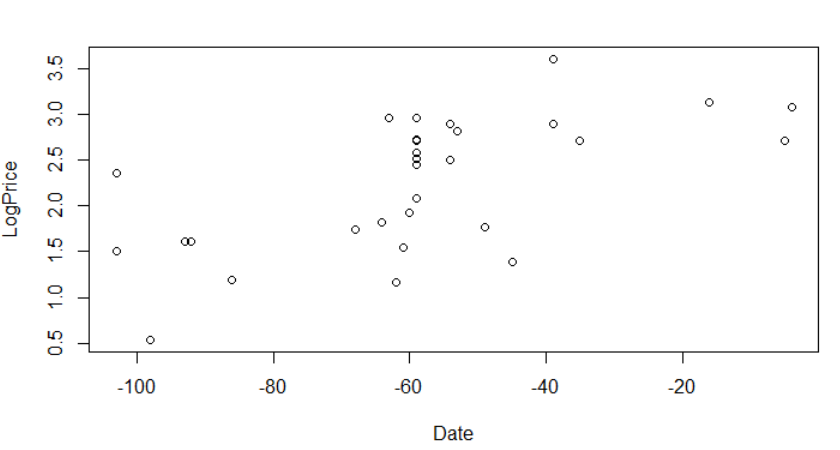
2. There is positive correlation between Elevation and county. Properties in Santa Clara county were at higher elevation than properties in San Mateo

3. Distance is having negative correlation towards elevation and positive correlation towards flood.

## **Plot Diagrams**

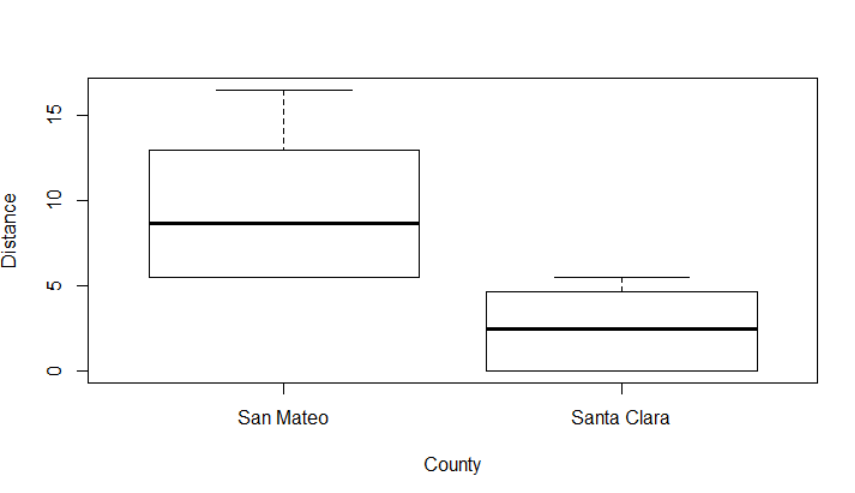
### **Price Vs Date**

plot(LogPrice~Date,data=data)



Above diagram shows that, price increases as time increases. Properties which are bought recently, bought with high price. Price increases overtime. Hence Date having significant impact on price.

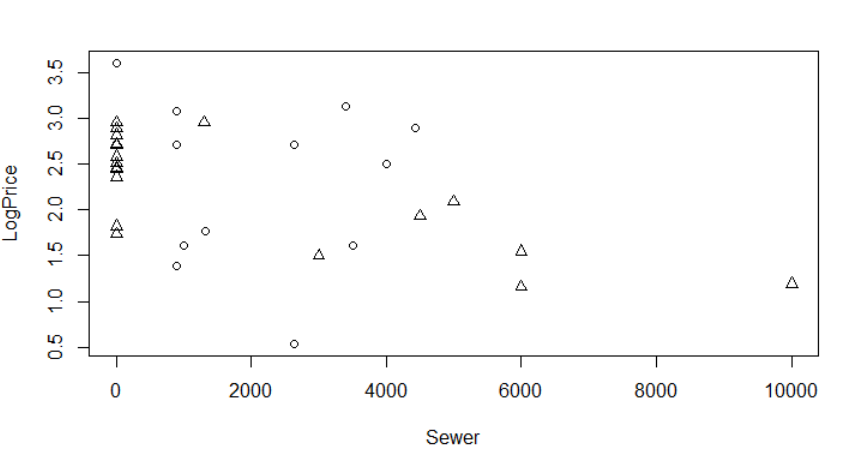
### **Distance Vs County**

plot(Distance~County,data=data)

Above box plot shows shat Leslie property is very near to County ‘Santa Clara’.

### **Price Vs Sewer Vs County**

plot(LogPrice~Sewer,data=data,pch=(1:2)[County])



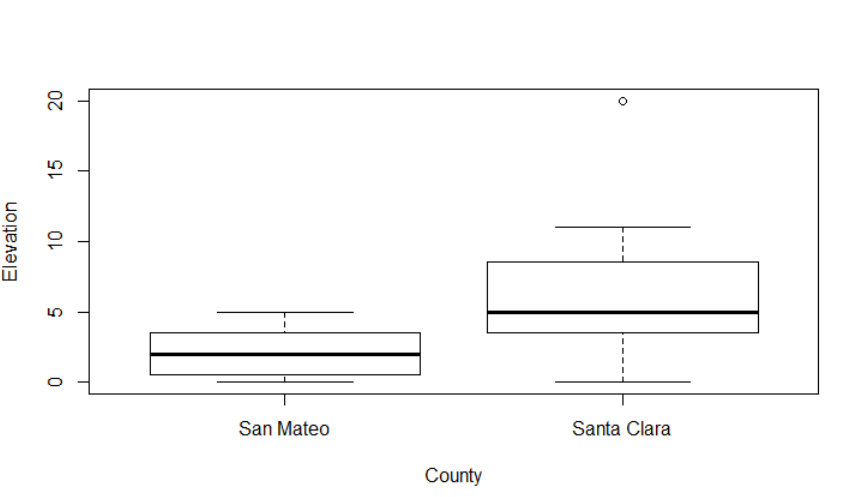
From above scatter diagram, distance to Sewer connection doesn’t have much importance. Property for which is having sewer connection 4000 feet distance, is also priced equivalent to property with sewer connection less than 500 feet distance. However. if sewer connection is more than 5000 feet distance, though they property resides in highly elevated area (no flooding) , it is priced low. We may include this variable or not depends on how much it improves the model.

### **Price Vs Size Vs County**

From the above scatter diagram, size doesn’t have much importance on pricing. All most 90% of properties are is same size. So, including size in our regression model will not give significant improvement in the model.

### **County Vs Elevation**

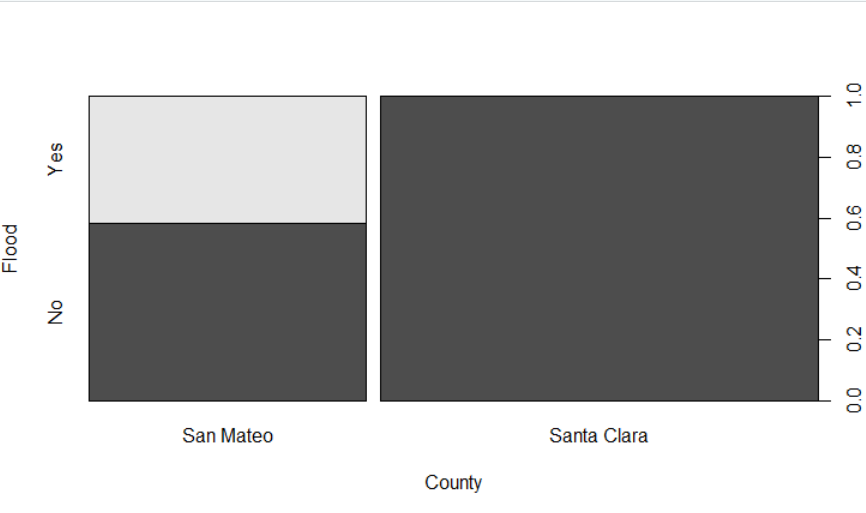
plot(Elevation~County,data=data)



Above diagram depicts that, ‘Santa Clara’ county is highly elevated compare to ‘San Mateo’. Leslie property is near to Santa Clara.

### **County Vs Flood**

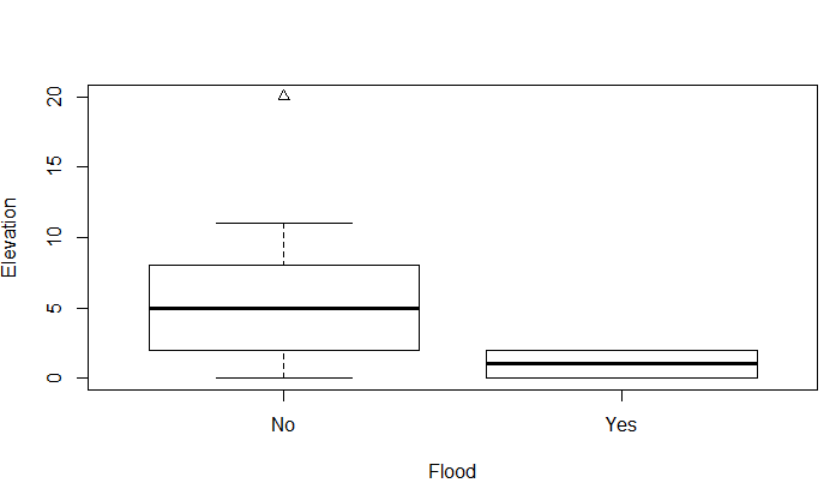
plot(Flood~County,data=data)



Elevation of Santa Clara is high, so property in ‘Santa-Clara’ is not subject to flooding compare to ‘San Mateo’.

### **Elevation Vs Flood**

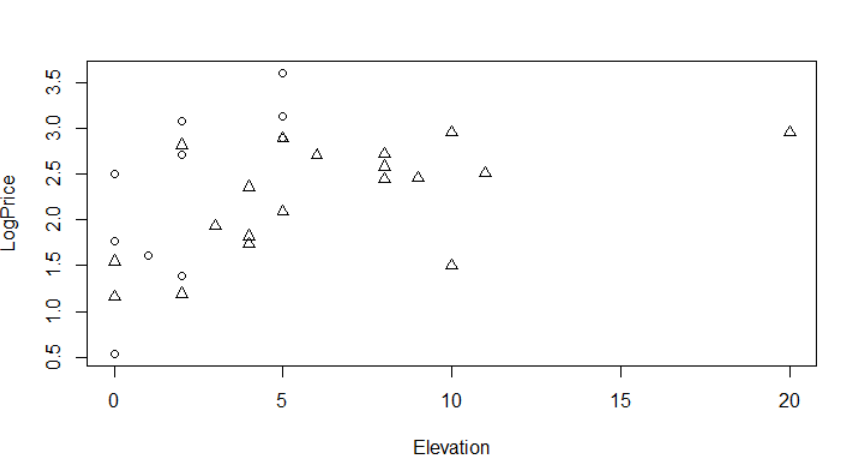
plot(Elevation~Flood,data=data,pch=(1:2)[County])



Highly elevated property is not subject to flooding.

### **Price Vs Elevation Vs County**

plot(LogPrice~Elevation+County,data=data,pch=(1:2)[County])



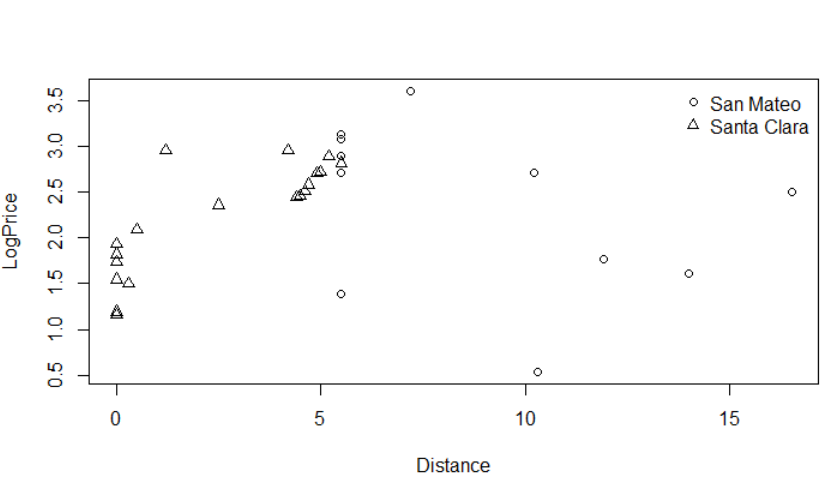
Above diagram shows that there is greater percentage of increase in price for each additional foot of elevation. However, slope varies based on county. There is ‘x’ percentage increase in the for 1 feet increase in elevation level if the property in ‘Santa Clara’. However, ‘y’ percentage increase in the for 1 feet increase in elevation level if the property in ‘San Mateo’

Variation in price is not only effect on elevation alone. It is combination of elevation and county.

### **Price Vs Distance Vs County**

plot(LogPrice~Distance,data=data,pch=(1:2)[County])

legend("topright",legend=c("San Mateo","Santa Clara"),pch=1:2,bty="n")



Though above graph & correlation matrix shows that distance is not having much impact on price, it appears to play role. Due to pairwise correlation between Elevation & Flood (Price Vs Elevation, Price Vs Flood), impact of distance variable on price is reduced. It would be better to consider distance variable. We can remove this variable from the model, if it is providing the improvement in the model.

# 

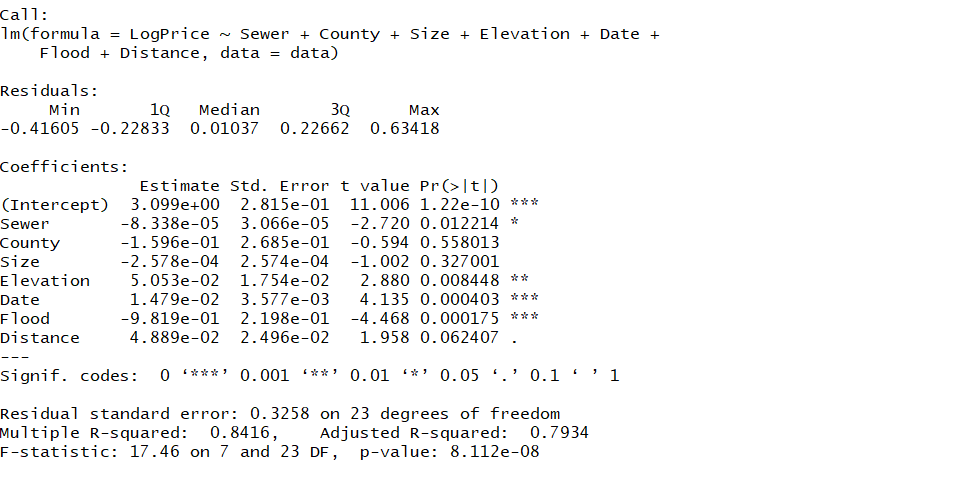
# **Regression Models**

## 

## **Full Model**

**fullmodel=lm(LogPrice ~ Sewer + County + Size + Elevation + Date + Flood + Distance, data = data)**

**summary(fullmodel)**

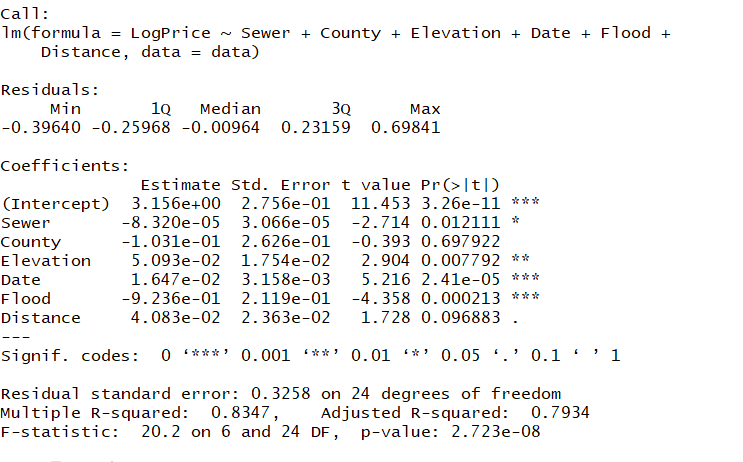


P-value is less than 0.05 however adjusted R square (0.79) is less than Multiple R Square (0.84). As size doesn’t have much importance with respect to p -value & explanation from above scatter diagrams, build model without size

## **Model without Size**

**model1=3=lm(LogPrice ~ Sewer + County + Elevation + Date + Flood + Distance , data = data)**

**summary(model3)**

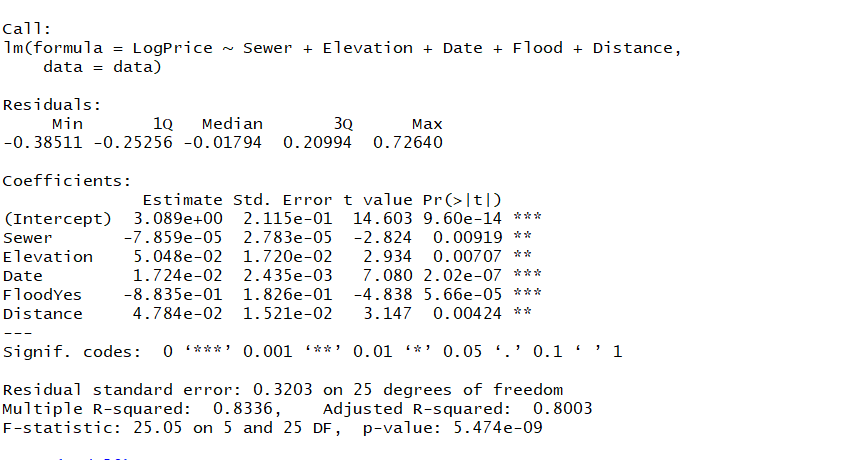


## **Model without Size & County**

**R-Code**

**model9=lm(LogPrice~Sewer+Elevation+Date+Flood+Distance, data=data)**

**summary(model9)**



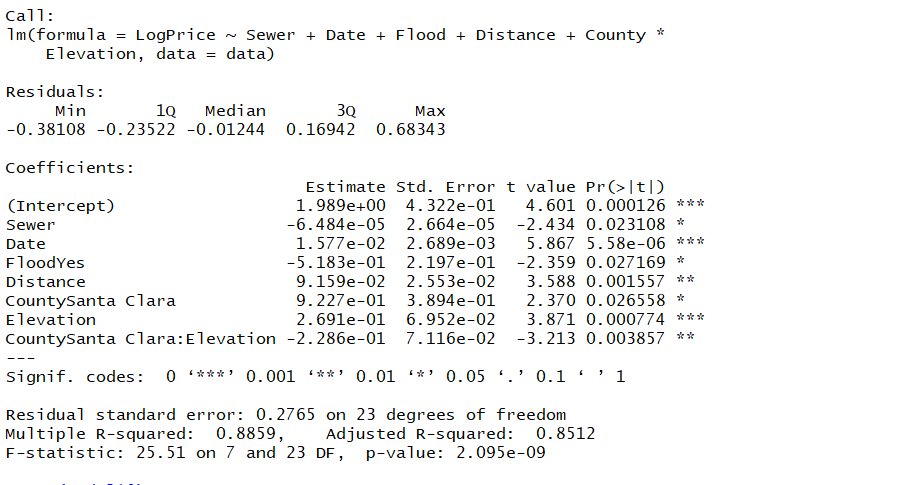
After removing county, we have improvement in Adjusted R square and P value.

However as explained in Plot diagram section, relationship between county and elevation having some impact on price. Hence lets build the model by adding County\* Elevation and removing Elevation

## **Final Model**

**model10=lm(LogPrice~Sewer+Date+Flood+Distance+County\*Elevation, data=data)**

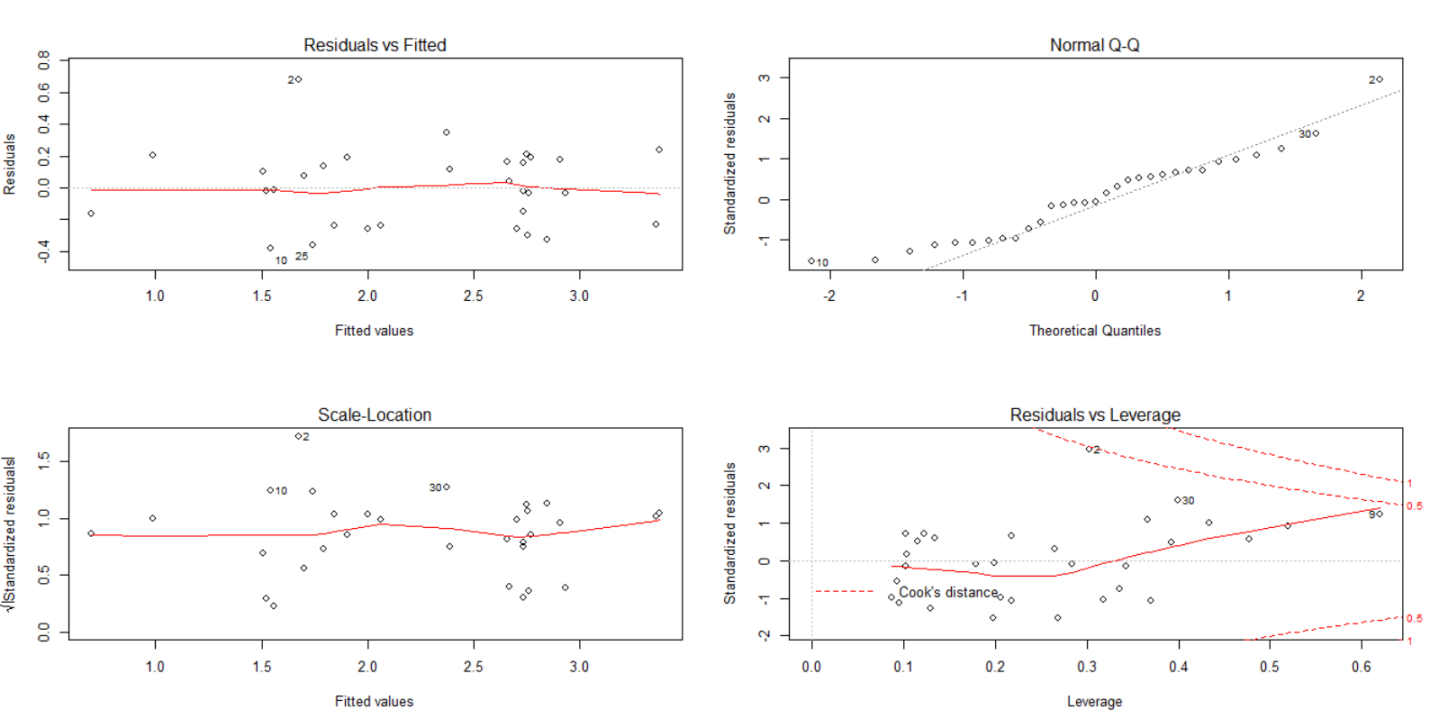
**summary(model10)**



After adding elevation and County relationship, we have significant improvement in the model.

## **Residual Diagrams**

par(mfrow = c(2, 2))

plot(model10)

# **Predict the price of Leslie Salt property**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Price** | **County** | **Size** | **Elevation** | **Sewer** | **Date** | **Flood** | **Distance** |
| ? | 1 | 0.646 | 0 | 0 | 0 | 0 | 0 |

County – In record 11, 12, distance mentioned as 0. Hence those property are nearby property Leslie. Both belongs to county 1. Hence, we mentioned Leslie property also ‘1’

Size – 0.646 (given in the problem)

Elevation –0 (property at sea leve)

Sewer -0 (no data provided. Hence 0)

Date – assume that property may sold in next 3 months

Flood- 0 (property is diked)

Distance – 0 (distance with respect to Leslie property)

### **If property sold now**

**leslie\_salt <- data.frame(0,"Santa Clara",246.8,0,0,0,"No",0,log(0))**

**colnames(leslie\_salt) <- c("Price", "County", "Size", "Elevation", "Sewer", "Date", "Flood", "Distance","LogPrice")**

**data<-rbind(data,leslie\_salt)**

**leslie\_salt\_price <- predict(model10, newdata = data[32,])**

**price=exp(leslie\_salt\_price)**

**price**

**Answer:**

|  |
| --- |
| 18.38024  **The price of Leslie Salt property is $18340/acre.** **If property sold in another 3 months** **leslie\_salt <- data.frame(0,"Santa Clara",246.8,0,0,3,"No",0,log(0))**  **colnames(leslie\_salt) <- c("Price", "County", "Size", "Elevation", "Sewer", "Date", "Flood", "Distance",**  **"LogPrice")**  **data<-rbind(data,leslie\_salt)**  **leslie\_salt\_price <- predict(model10, newdata = data[33,])**  **price=exp(leslie\_salt\_price)**  **price**  **Answer:**  19.27097  **The price of Leslie Salt property is $19271/acre.** **If property sold in another 6 months** **leslie\_salt <- data.frame(0,"Santa Clara",246.8,0,0,6,"No",0,log(0))**  **colnames(leslie\_salt) <- c("Price", "County", "Size", "Elevation", "Sewer", "Date", "Flood", "Distance",**  **"LogPrice")**  **data<-rbind(data,leslie\_salt)**  **leslie\_salt\_price <- predict(model10, newdata = data[34,])**  **price=exp(leslie\_salt\_price)**  **price**  **Answer:**  20.20487  **The price of Leslie Salt property is $20204/acre.** |
|  |
| |  | | --- | |  | |

# **Appendix -A – Acceptance Criteria for Regression model**

The most common metrics to look at while selecting the model are

| **STATISTIC** | **CRITERION** |
| --- | --- |
| R-Squared | Higher the better *(> 0.70)* |
| Adj R-Squared | Higher the better |
| F-Statistic | Higher the better |
| Std. Error | Closer to zero the better |
| t-statistic | Should be greater 1.96 for p-value to be less than 0.05 |
| AIC | Lower the better |
| BIC | Lower the better |
| Mallows cp | Should be close to the number of predictors in model |
| MAPE (Mean absolute percentage error) | Lower the better |
| MSE (Mean squared error) | Lower the better |
| Min\_Max Accuracy => mean(min(actual, predicted)/max(actual, predicted)) | Higher the better |

# **APPENDIX -B- Regression Models**





**Problem Statement:**

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

**Solution:**

After going through the data, to check if the linear relationship exists or not we perform Multiple regression analysis on the variables Provided. Following is the MLR Equation we are testing:

X1 = Alpha +X2(Beta2)+X3(Beta3)+ X4(Beta4) +x5(Beta5) +X6(Beta6) +Error.

Following are the values observed once Regression is performed using Excel.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT | |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |  |  |
| Multiple R | 0.996584 |  |  |  |  |  |  |  |
| R Square | 0.993179 |  |  |  |  |  |  |  |
| Adjusted R Square | 0.991556 |  |  |  |  |  |  |  |
| Standard Error | 17.64924 |  |  |  |  |  |  |  |
| Observations | 27 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 5 | 952538.9 | 190507.8 | 611.5904 | 5.4E-22 |  |  |  |
| Residual | 21 | 6541.41 | 311.4957 |  |  |  |  |  |
| Total | 26 | 959080.4 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | -18.8594 | 30.15023 | -0.62551 | 0.538372 | -81.5602 | 43.84142 | -81.5602 | 43.84142 |
| X2 | 16.20157 | 3.544437 | 4.570986 | 0.000166 | 8.830513 | 23.57263 | 8.830513 | 23.57263 |
| X3 | 0.174635 | 0.057606 | 3.031541 | 0.006347 | 0.054837 | 0.294434 | 0.054837 | 0.294434 |
| X4 | 11.52627 | 2.532103 | 4.552053 | 0.000174 | 6.260472 | 16.79207 | 6.260472 | 16.79207 |
| X5 | 13.58031 | 1.770457 | 7.670514 | 1.61E-07 | 9.898447 | 17.26218 | 9.898447 | 17.26218 |
| X6 | -5.31097 | 1.705427 | -3.11416 | 0.005249 | -8.8576 | -1.76434 | -8.8576 | -1.76434 |

On Analysing, we see that the Adjusted R Square Value is 99.15. We can also see that the Standard error for all the coefficients are high. Hence there may be a multi collinearity existing between the Predictor or Independent Variables. In other words, the predictor variables are dependent on eat other.

To analyse the correlation, perform a correlation analysis using Excel on the independent or predictor variables.

Following results are observed.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *X2* | *X3* | *X4* | *X5* | *X6* |
| X2 | 1 |  |  |  |  |
| X3 | 0.843616 | 1 |  |  |  |
| X4 | 0.748587 | 0.906231 | 1 |  |  |
| X5 | 0.838023 | 0.863917 | 0.795434 | 1 |  |
| X6 | -0.76574 | -0.80738 | -0.84128 | -0.86959 | 1 |

We can see from above correlation matrix that Correlation among all the variables is high (Greater than 0.6). Hence we conclude that Multi collinearity exists between the variables .In other words it is difficult to explain the Effect of each Predictor variable (X2-X6) on the Response Variable(X1).

To Check the effect or importance of each predictor variable on X1, we perform PCA .

**PCA using R:**

[Workspace loaded from ~/.RData]

> setwd("C:/Users/Babloo/Desktop/GA\_PCA")

> library(readxl)

> Dataset\_All\_Greens\_Franchise <- read\_excel("C:/Users/Babloo/Desktop/GA\_PCA/Dataset\_All Greens Franchise.xls")

> View(Dataset\_All\_Greens\_Franchise)

Using Apply Function to get the mean and Std Dev of all the variables

apply(data, 2, mean)

X1 X2 X3 X4 X5 X6

286.574074 3.325926 387.481481 8.100000 9.692593 7.740741

> apply(data, 2, var)

X1 X2 X3 X4 X5 X6

36887.705840 4.044302 36545.105413 14.246923 26.419943 23.968661

We can see from above that the variance of all the variables are different.

it is important to standardize the variables to have mean zero and standard derivation one before

Performing PCA using PRCOMP() Function

pr.out = prcomp(data[,-1], scale = T)

> pr.out

> pr.out$rotation

PC1 PC2 PC3 PC4 PC5

X2 0.4346055 0.72351531 -0.2173219 0.4731253 -0.1287140

X3 0.4587593 -0.07708152 -0.4766292 -0.3421379 0.6628458

X4 0.4451881 -0.58485406 -0.3612434 0.1704407 -0.5479110

X5 0.4529839 0.22908597 0.3987150 -0.6786263 -0.3504538

X6 -0.4441522 0.27577053 -0.6603979 -0.4117164 -0.3479134

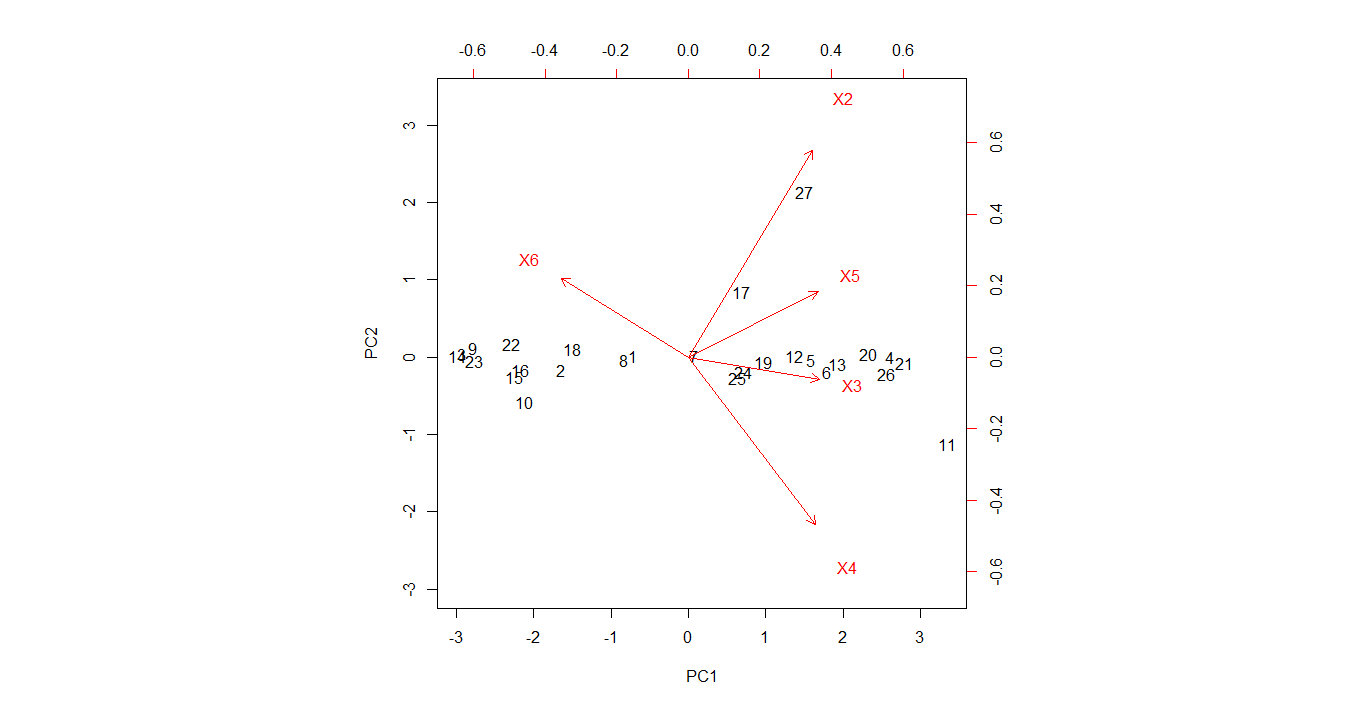
We see that there are four distinct principal components

This is to be expected because there are in general min(n-1, p) informative principal components

in a data set with n observations and p variables.

Plotting the Principle components using Biplot()

biplot(pr.out, scale = 0)



We notice that PC1 corresponds to measure of X3,X5 and X6 where as PC2 corresponds to X2 and X4. Hence these 2 components are sufficient to explain the variance between variables.

Computing the Variance for the principle components

> pr.out$sdev

[1] 2.0768570 0.5277129 0.4797178 0.3520671 0.2326019

> pr.var

[1] 4.31333506 0.27848093 0.23012914 0.12395124 0.05410364

To compute the proportion of variance explained by each principal component, we simply divide the variance explained by each principal component by the total variance explained by all the principal components.

> pve = pr.var/sum(pr.var)

> pve

[1] 0.86266701 0.05569619 0.04602583 0.02479025 0.01082073

> pve\*100

[1] 86.266701 5.569619 4.602583 2.479025 1.082073

We see that the first principal component explains 86.2% of the variance in the data

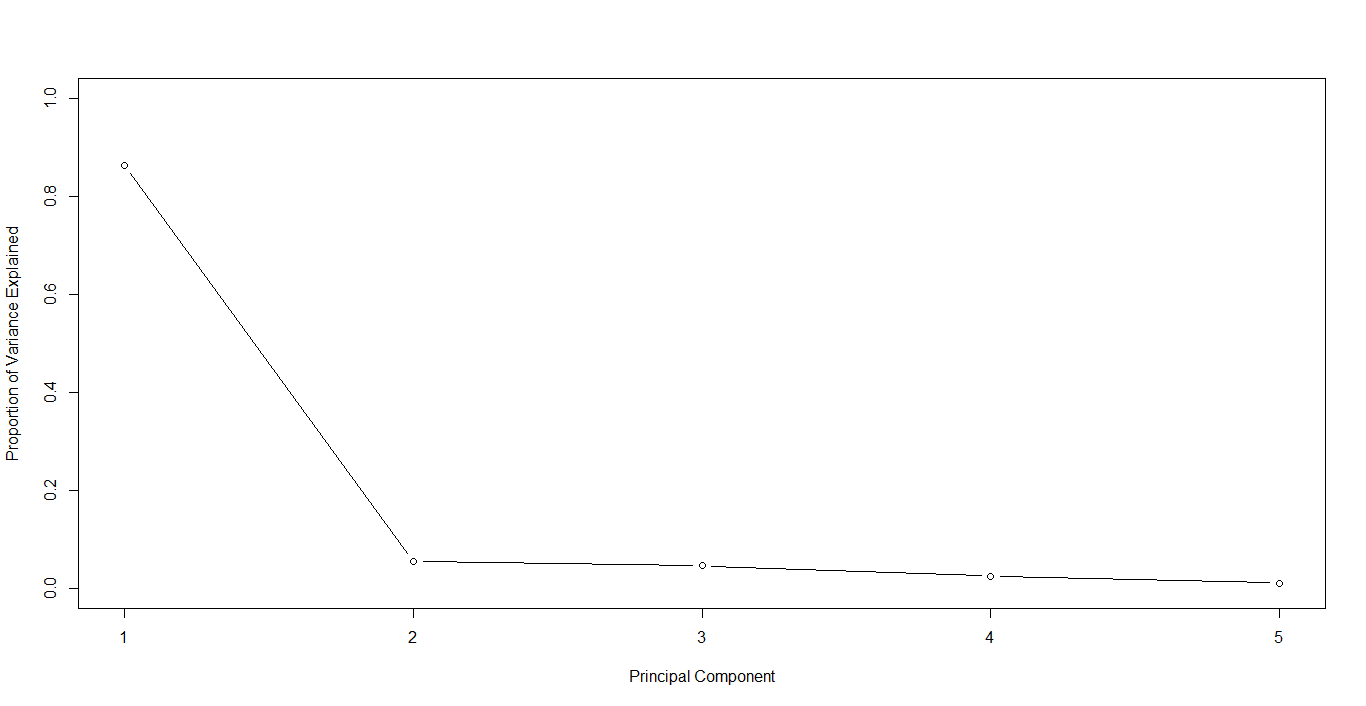
Second principal component explains 5.56% of the variance

Third principal component explains 4.6% of the variance

Fourth explains 2.47%

Fifth explains 1.08%.

plot the PVE explained by each component using Scree Plots:



**Conclusion:**

From the above PCA Analysis, we conclude that the first component (PC1) explains 86% of the variance i.e predictor variables X3,X5 and X6 (Inventory, Size of sales district and no of stores in the District) explains or drives the Annual net sales of the product mentioned.

Since the variance of other components are less than 6%, We can continue building the Regression model using only the First Component.