*Mini Project – Factor\_Hair\_Analysis*

*By Sakshi Jain*

*Table of Contents*

1. *Project Objective*
2. *Exploratory Data Analysis – Step by step approach*
   1. *Descriptive Statistics*
   2. *Missing Values*
   3. *Outliers*
   4. *Visual interpretation*
3. *Simple Linear Regression*
4. [*Evidence of Multicollinearity*](https://rstudio-pubs-static.s3.amazonaws.com/514452_3124107a4f444a81bb3fe549b58207b5.html#evidence-of-multicollinearity)
5. *Principal Component Analysis*
6. *Multiple Linear Regression*
7. *Model Validity: Predicted Satisfaction vs Actual Satisfaction with Conclusion*

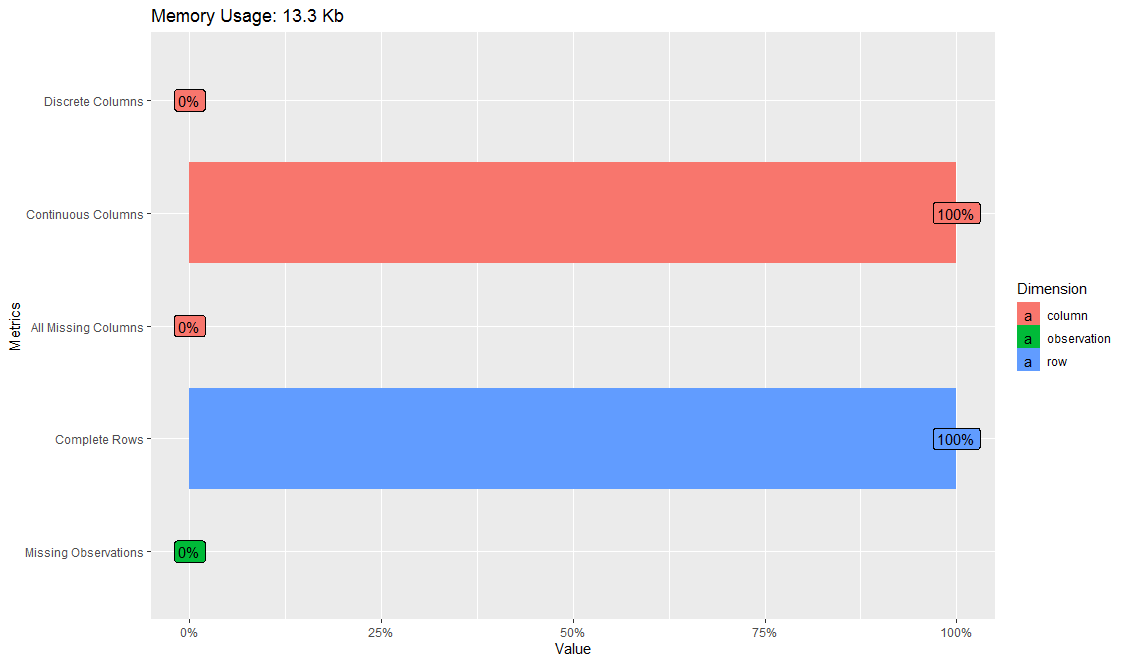
*Project Objective*

* *The objective of the report is to explore the data set “Factor-Hair-Revised.csv” in R and generate insights about the data set. This exploration report will consists of the following:*
* *Importing the dataset in R*
* *Understanding the structure of dataset*
* *Graphical exploration*
* *Advanced statistics*
* *Principal Component Analysis*
* *Insights from the dataset*

*Exploratory Data Analysis*

*Factor-Hair ==================================================================================*

* *I have used the below mentioned libraries to complete my project & also to conclude something relevant out of my project*
  + *library(ggplot2)*
  + *library(psych)*
  + *library(corrgram)*
  + *library(car)*
  + *library(corrplot)*
  + *library(nFactors)*
  + *library(dplyr)*
  + *library(DataExplorer)*
  + *library(kableExtra)*
* *The data set used has 100 rows & 13 columns. There were no missing values in the dataset. I have run the following graph to make it evident that there is no missing values present*

**

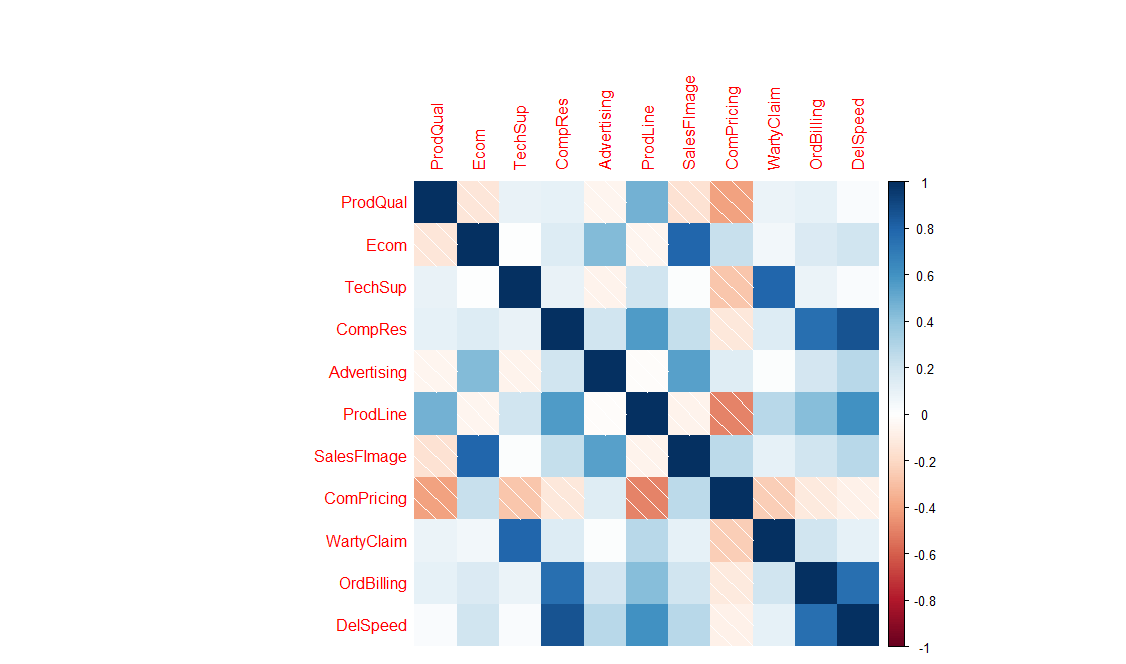
* *As this graph suggests missing values is 0%*
* *For checking whether there are any outliers i checked the structure of the data using the Str command in r. ID variable was of Integer type, rest all variables were of Numerical Category From the structure i could make out that there were not visible outliers present in the dataset. Next i ran the summary command. It shows the six point analysis of each variable present in the dataset i.e. Min Value, 1st Quartile, Mean, Median, 3rd quartile & Max value. By looking at the summary i concluded that most of the variables resemble normal distribution*
* *We also learned that first column named "ID" is just a column number & we won't be needing in the for process, so Created new data frame named hair1 with all data except the column "ID"*

*Simple Linear Regression Models*

* *Simple linear regression models for all 11 factors with Response variable - Satisfaction was done using lm function individually & combined*
* *The lm function gives 2 values – Intercept & Slope value. Slope is the independent variable that suggests that for any one unit of change in independent variable will change the value of dependent variable by one unit*
* *When we check the summary of the combined model, below mentioned insights was generated:*
  + *As in our model the adjusted R-squared: 0.7774, meaning that independent variables explain 78% of the variance of the dependent variable, only 3 variables are significant out of 11 independent variables.*
  + *The p-value of the F-statistic is less than 0.05(level of Significance), which means our model is significant. This means that, at least, one of the predictor variables is significantly related to the outcome variable.*
  + *Our model equation can be written as: Satisfaction = -0.66 + 0.37\*ProdQual -0.44\*Ecom + 0.033\*TechSup + 0.16\*CompRes -0.02\*Advertising + 0.14ProdLine + 0.80\*SalesFImage-0.038\*CompPricing -0.10\*WartyClaim + 0.14\*OrdBilling + 0.16\*DelSpeed*

*Evidence of Multicollinearity*

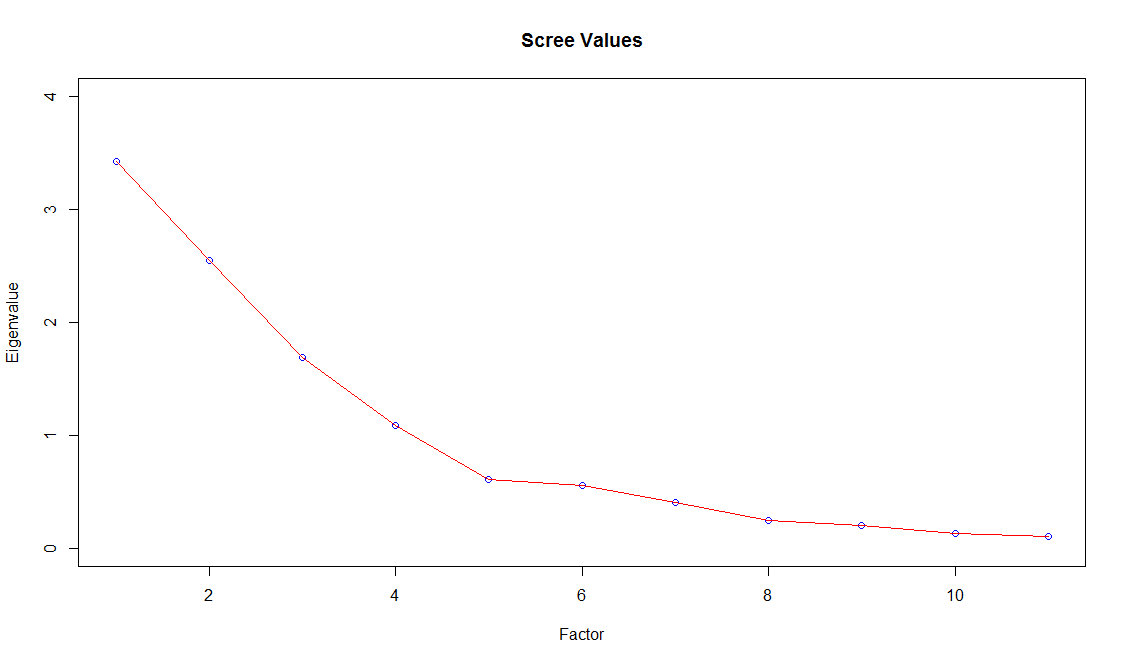
* *We next run correlation command to check the presence of Multicollinearity in the dataset. We tried plotting it using corrplot command.*

**

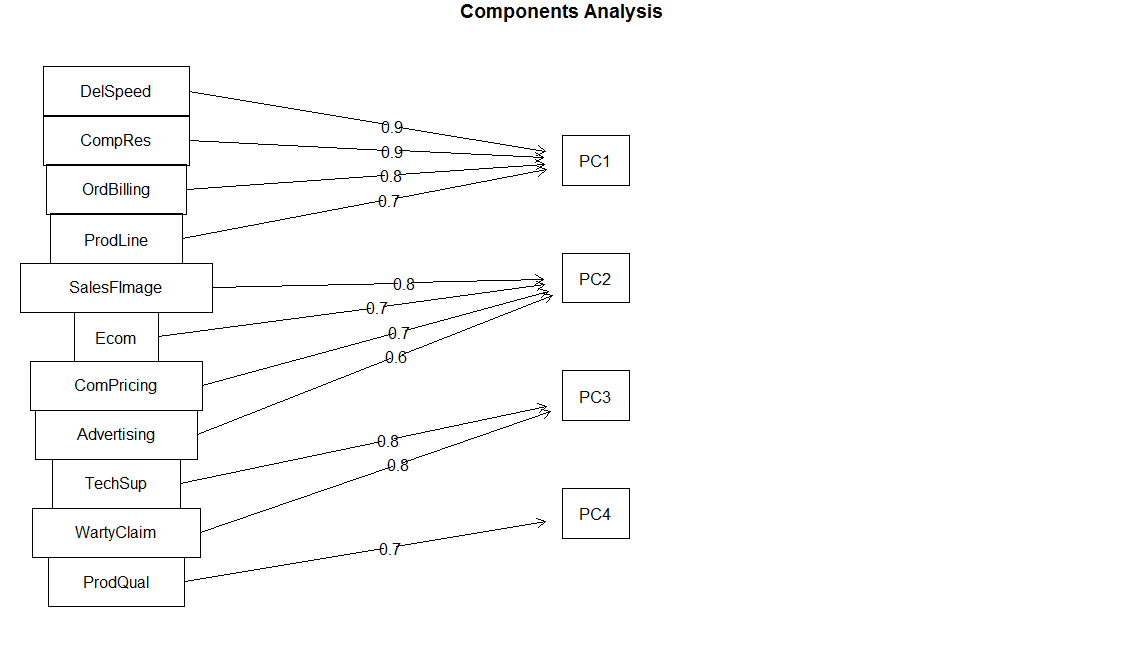
* *The above mentioned graphs suggest the following:*
  + CompRes and DelSpeed are highly correlated
  + OrdBilling and CompRes are highly correlated
  + WartyClaim and TechSupport are highly correlated
  + CompRes and OrdBilling are highly correlated
  + OrdBilling and DelSpeed are highly correlated
  + Ecom and SalesFImage are highly correlated
* *Assumptions of Regression: Variables are independent of each other-Multicollinearity shouldn’t be there*
* *For evidence of Multicollinearity amongst the variables of hair dataset****Variance Inflation Factors (VIF) concept was used****. Any variable having value of VIF > 4 suggests presence of Multicollinearity amongst predictor variables*
* *VIF for delivery speed & Comp Res was found to be greater than 4. We can infer that variables DelSpeed and CompRes are a cause of concern which clearly justifies the presence of Multicollinearity in the data.* We can safely assume that there is a high degree of Multicollinearity between the independent variables
* *Remedial Measures: Two of the most commonly used methods to deal with Multicollinearity in the model are the following.*
  + *Remove some of the highly correlated variables using VIF or stepwise algorithms.*
  + *Perform an analysis design like principal component analysis (PCA)/ Factor Analysis on the correlated variables*

*Principal Component Analysis*

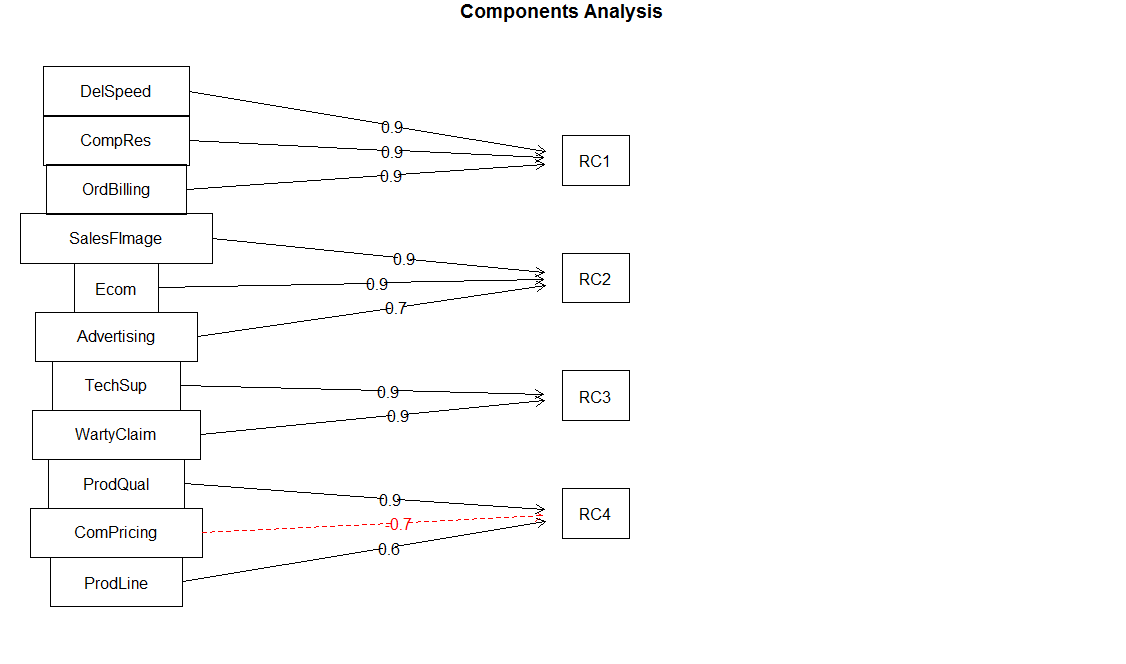
* + *First we conducted a bartlett sphericity test to check whether Principal Component Analysis can be done on the predictor variables of the dataset*
  + *After running the test, we infer that since the p value for the test is quite less significance level of alpha = 0.001 so we reject the null hypothesis Ho (that PCA cannot be conducted implying that there is no correlation amongst the predictor variables)*
  + *Next we need to understand how many factors we need to explain the maximum variation in the dataset*
  + *One way to determine the number of factors or components in a data matrix or a correlation matrix is to examine the “scree” plot of the successive eigenvalues. Sharp breaks in the plot suggest the appropriate number of components or factors extract.*
  + *The scree plot graphs the Eigenvalue against each factor. We can see from the graph that after factor 4 there is a sharp change in the curvature of the scree plot. This shows that after factor 4 the total variance accounts for smaller amounts. Selection of factors from the scree plot can be based on:*
  + *Kaiser-Guttman normalization rule says that we should choose all factors with an eigenvalue greater than 1.*
  + *Bend elbow rule*

**

* + *So as per the elbow or Kaiser-Guttman normalization rule, we are good to go ahead with 4 factors*
  + *Now we will use only 4 factors. We will use orthogonal rotation (varimax) because in orthogonal rotation the rotated factors will remain uncorrelated whereas in oblique rotation the resulting factors will be correlated*
  + *Rotation = “None”: Factor 1 accounts for 31% of the variance; Factor 2 accounts for 23% of the variance; Factor 3 accounts for 15% of the variance; Factor 4 accounts for 10% of the variance. All the 4 factors together explain for 79% of the variance in performance*

**

* + *After Varimax Rotation:*

**

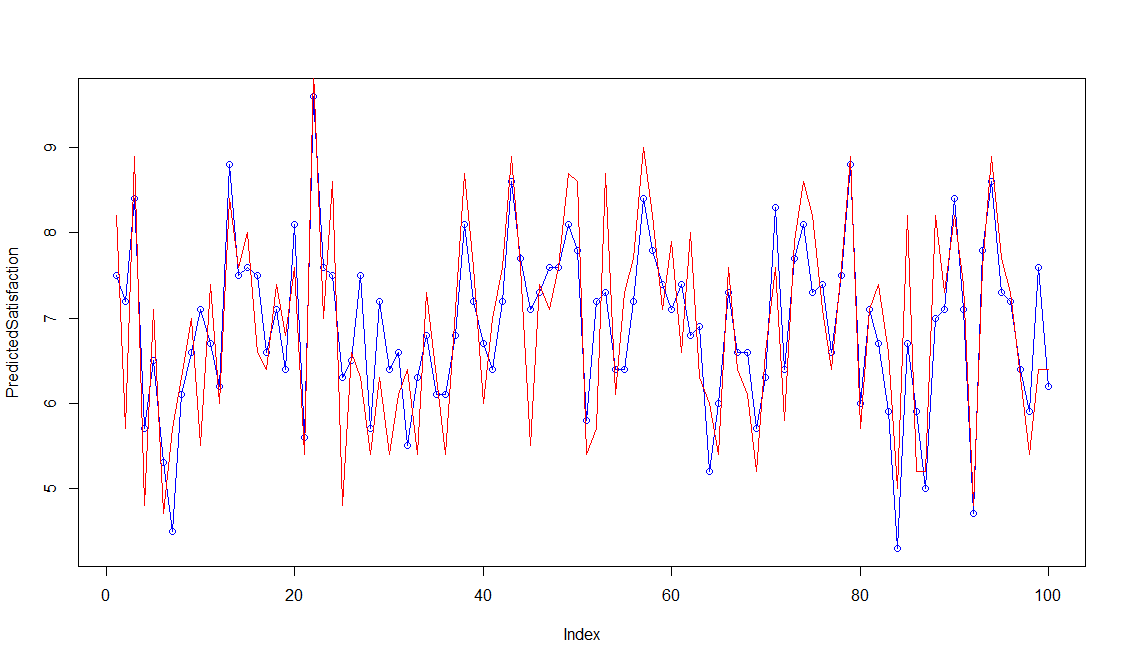
* + *The red dotted line means that Competitive Pricing marginally falls under the RC4 bucket and the loading are negative*
  + *Components Meaningful Names Column Name*
  + *RC1 Buying Experience Buyepr*
  + *RC2 Brand Recognition Brand*
  + *RC3 After Sales Service AfSSr*
  + *RC4 Product Prodt*
  + *Explanation:*
    - *RC1 - Buying Experience explains about variables affecting Complaint resolution, Order and Billing and delivery speed to customers*
    - *RC2 - Brand recognition handles Ecommerce, image of Sales force , Advertising which is face of the product*
    - *RC3 - After Sales Service gives information about Technical support, and Warranty and claims if there is any problem to customer after he has bought the item*
    - *RC4 - Product talks about the qualities of product like its varieties and types, prices its quality i.e all tangible aspects about the very existence of company*
  + *Scores of the PCA were converted into a data frame combined with Satisfaction variable for further analysis. Colnames names were assigned as mentioned above*

*Multiple Linear Regression*

* + *We now run a multiple Regression analysis on the newly created data frame, with Satisfaction as the dependent variable*
  + *Looking at the p values values of Coefficients like Intercept (constant beta-naught) we see that it is significant, so it definitely not zero and contributes to regression model*
  + *Similarly predictor variables like Buying experience, Brand Recognition and Product have significant betas implying that Response variable Satisfaction is linearly associated with them*
  + *After sales service is the only variable which has some high p-value implying that its beta coefficient may not be contributing that significantly to the model or may be zero*
  + *All together Adj-R^2 explains that these predictors explains the 64.6 % of the variability in the dataset which is still good enough*
  + *Overall p-value of Model given by F-statistic gives evidence against the null-hypothesis. Model is significantly valid at this point*
  + *We can conclude that Satisfaction is dependent only on 4 factors. We can converge the 11 variables into 4 factors*

*Predicted Satisfaction vs. Actual Satisfaction*

* + *Using the newly built multiple regression model new Satisfaction scores were predicted to check the validity of the model. New data frame was formed to have columns as 1. IDs, 2. Satisfaction ratings 3. Buying Experience 4. Brand Recognition 5. After Sales service 6. predicted satisfaction (from multiple linear model)*
  + *Plot analysis revealed that our new MLR Regression model is quite good and close to actual Satisfaction scores Blue dots represent Predicted Satisfaction ratings Red dots represent Actual satisfaction scores derived from multiple linear regression model*

**

* + *Nutshell we can say that the " Satisfaction" ratings of hair product depends very highly on the overall Buying experience of the Customer i.e. how quickly his product is delivered, its billed and if there are complaints are resolved in shortest possible time*
  + *Brand Recognition or products advertising comes in second in mind. Product itself comes in third in order to satisfy the customer (though statistically it variance explanation capacity has been ranked 4th) and after sales contribution in fourth*