**Packages Used or discussed**

Hmisc

MICE (Multivariate Imputation via Chained Equations)

Ggplot2

Corrplot

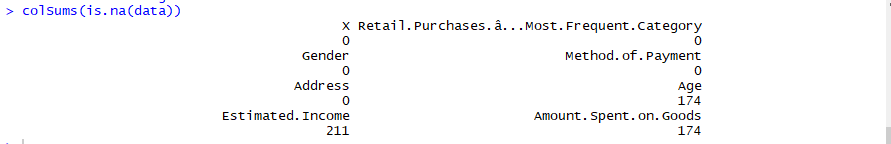
Nbclust <- NbClust package provides 30 indices for determining the number of clusters and proposes to user the best clustering scheme from the different results obtained by varying all combinations of number of clusters, distancemeasures, and clustering methods

Factoextra <- For using the elbow curve method as well silhouette method for clustering

Rpart <- For the decision trees

Stringr <- Used primarily for string manipulations

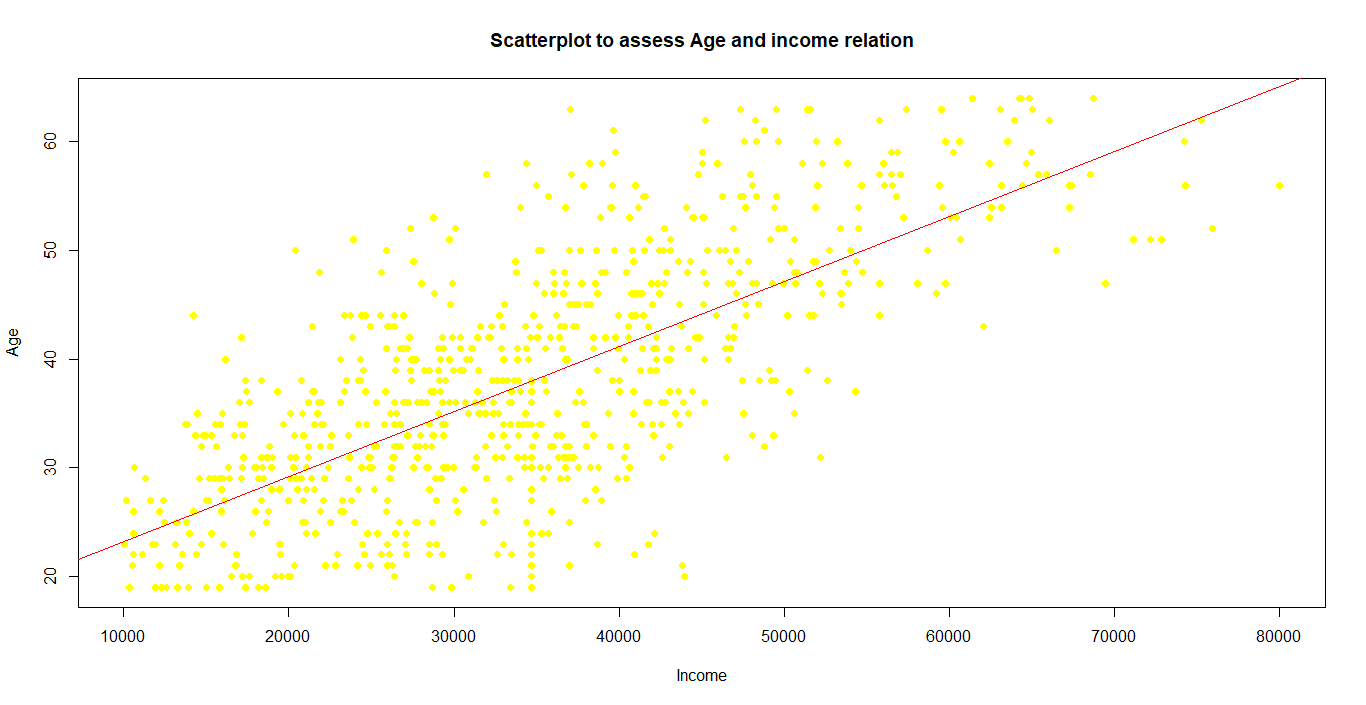
tm

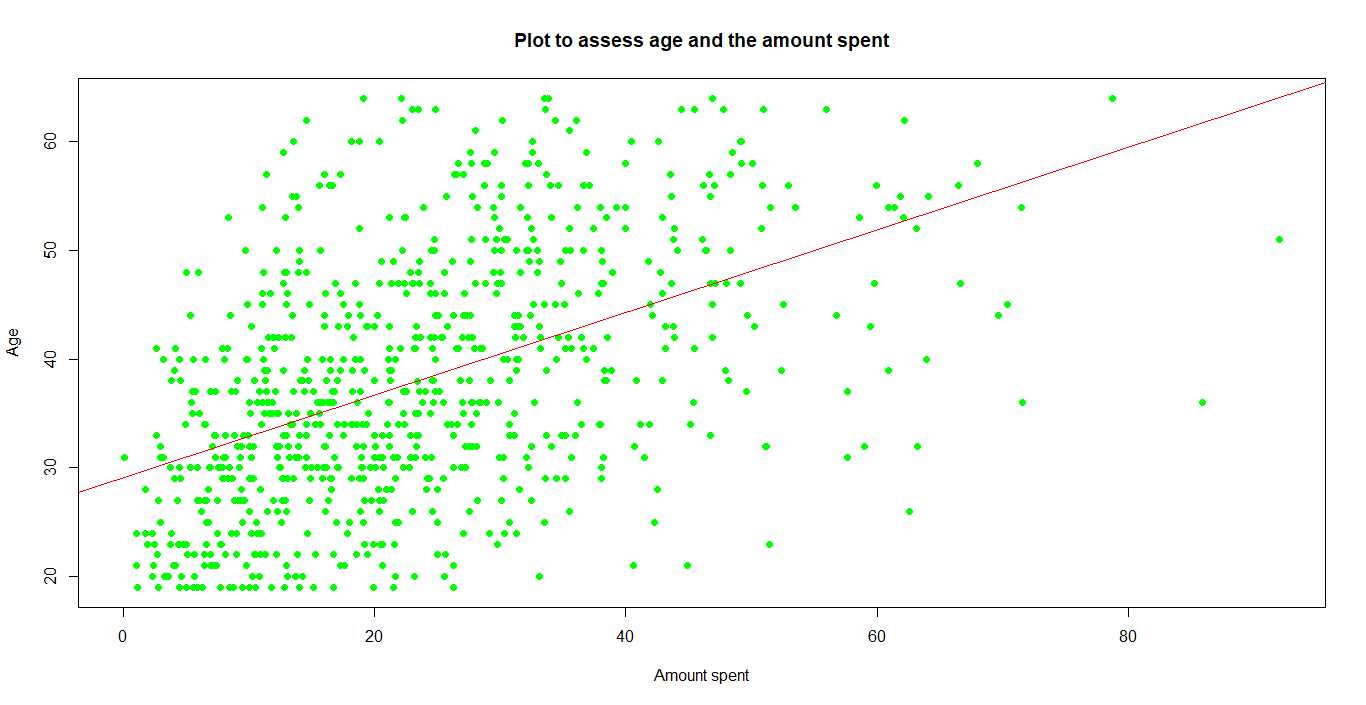


Initially checking for the missing values in the data set

**Replacing the missing values with the mean values**, reason being we can only impute missing values for continuous variables. Therefore, we need to only submit the columns with income, age, and amount spend to **MICE**. However, on a closer look at data, it seems that when one column is missing, all three columns are missing. Therefore, we have to drop all the rows with missing data or simply impute the missing data with the mean. The more advanced methods will not work.

The only other thing we can think to do is to covert the categorical variables (Retail.purchases, Gender and Method of payment) to dummy variables. Then these dummy values can be used as predictors of the missing data. However, this may not result in useful imputation.





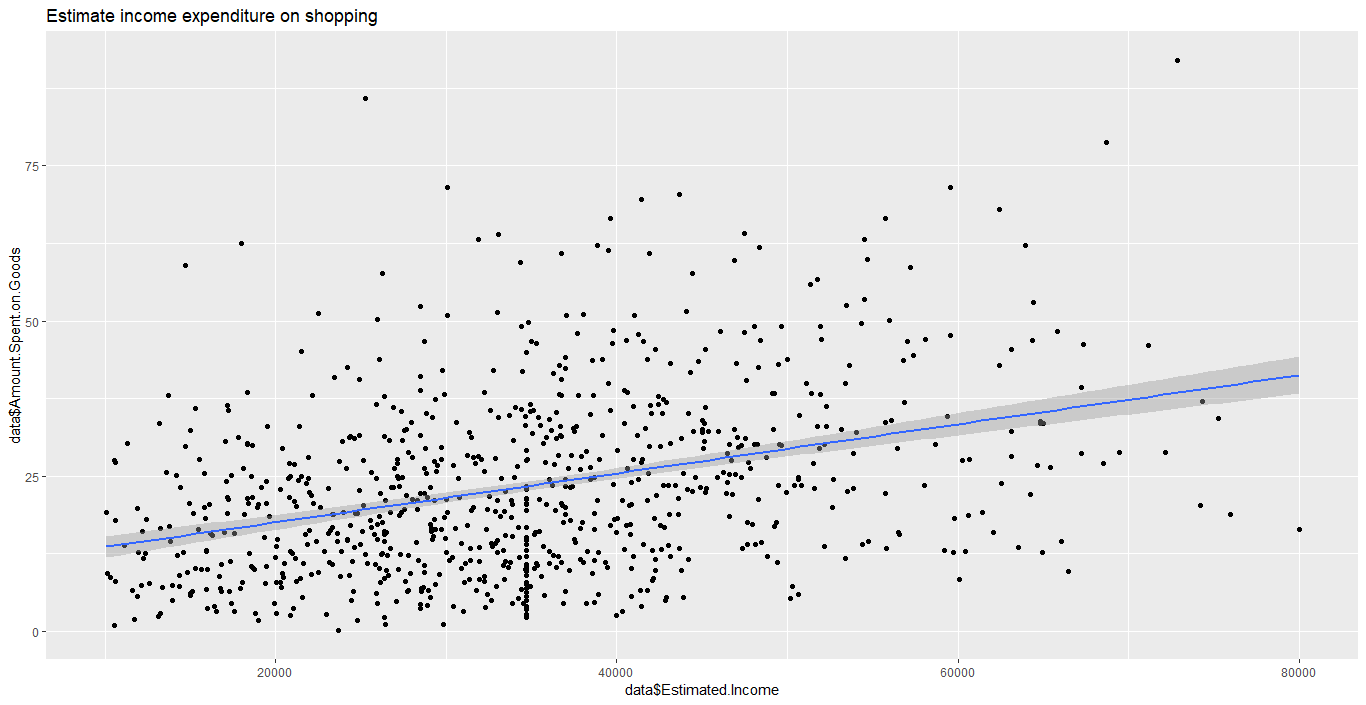


Fig: Relation between income and expenditure

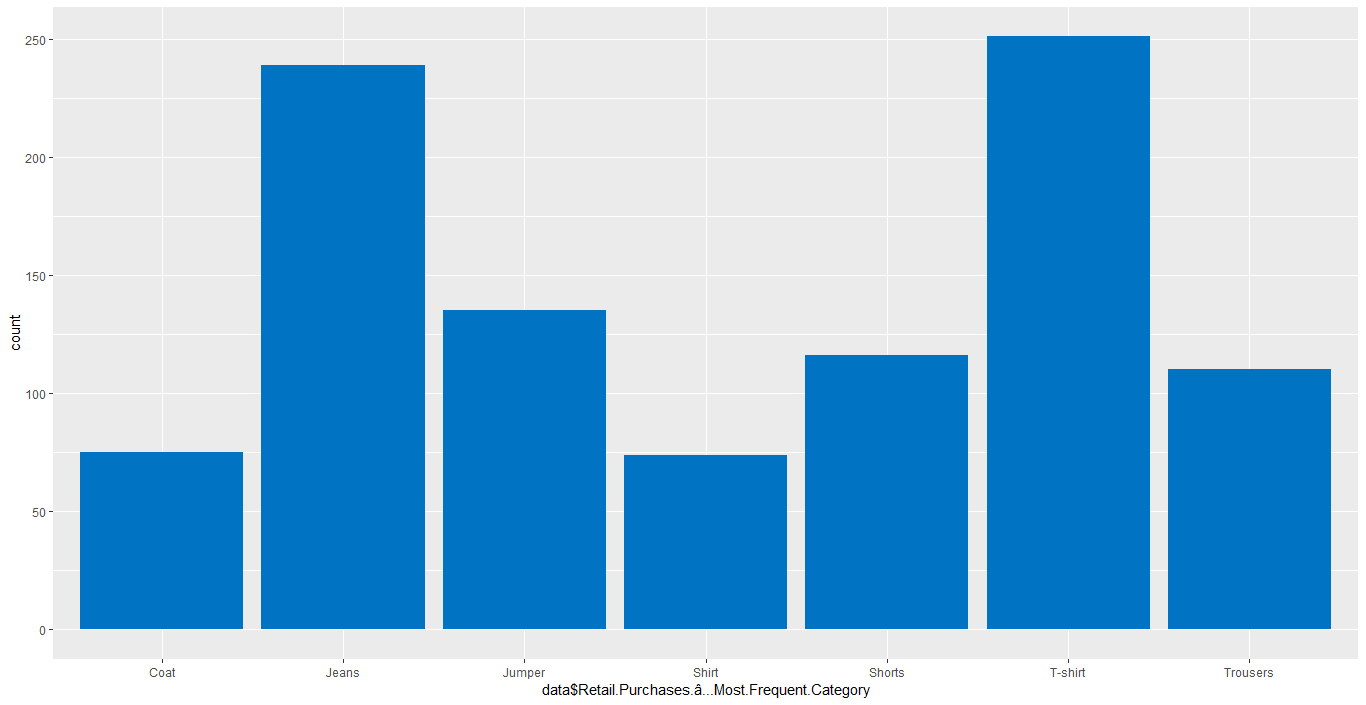


Fig: Statistics of different products purchased

* T-shirt is the most preferred product while shirt are least preferred across both the genders

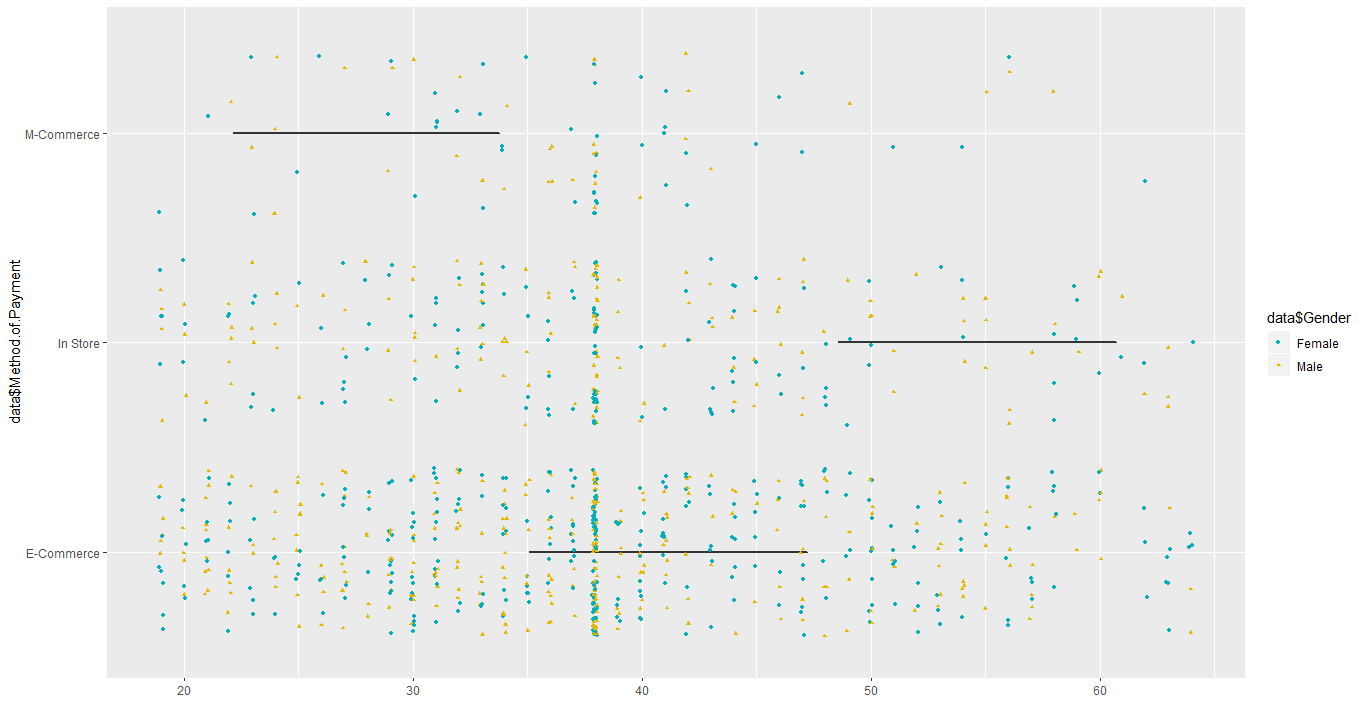


Fig: #Age, method of payment and gender wise distribution

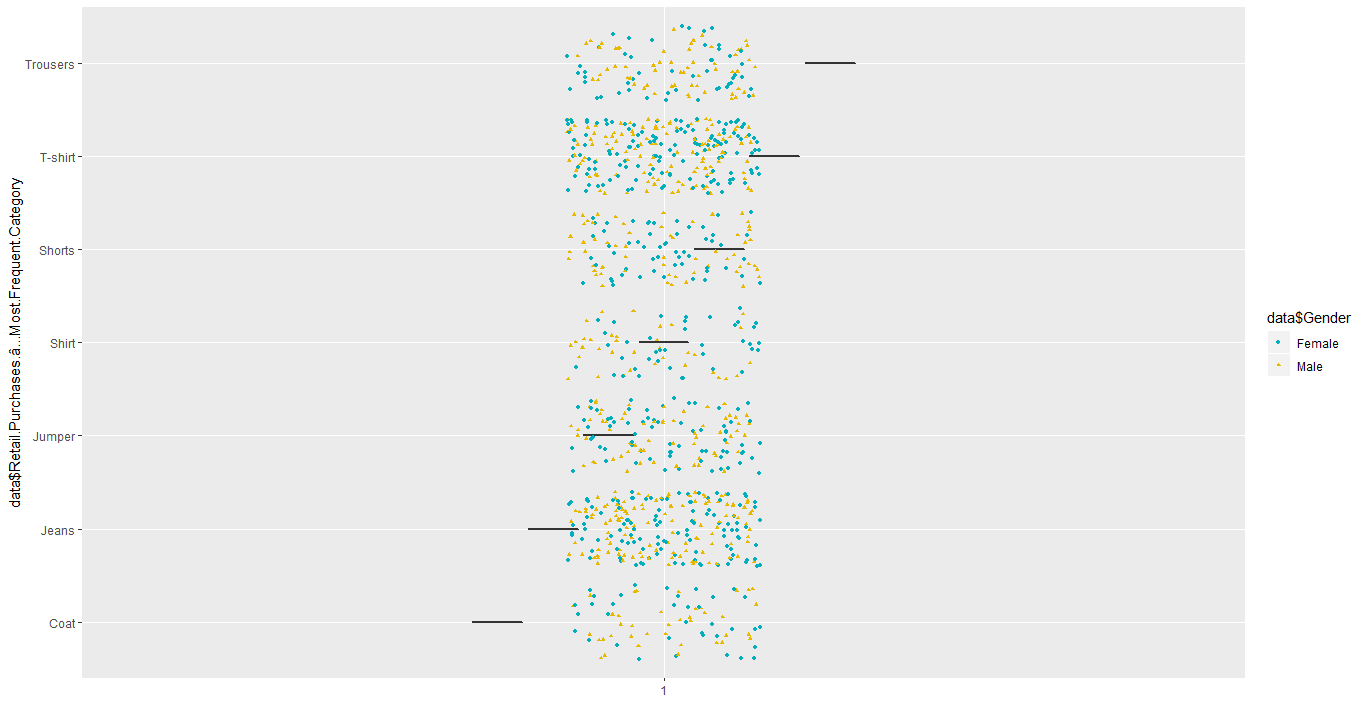


Fig: # Preference of products among gender

* Females prefer Jeans, T-shirt, coat and trousers while Male prefer shorts and jeans

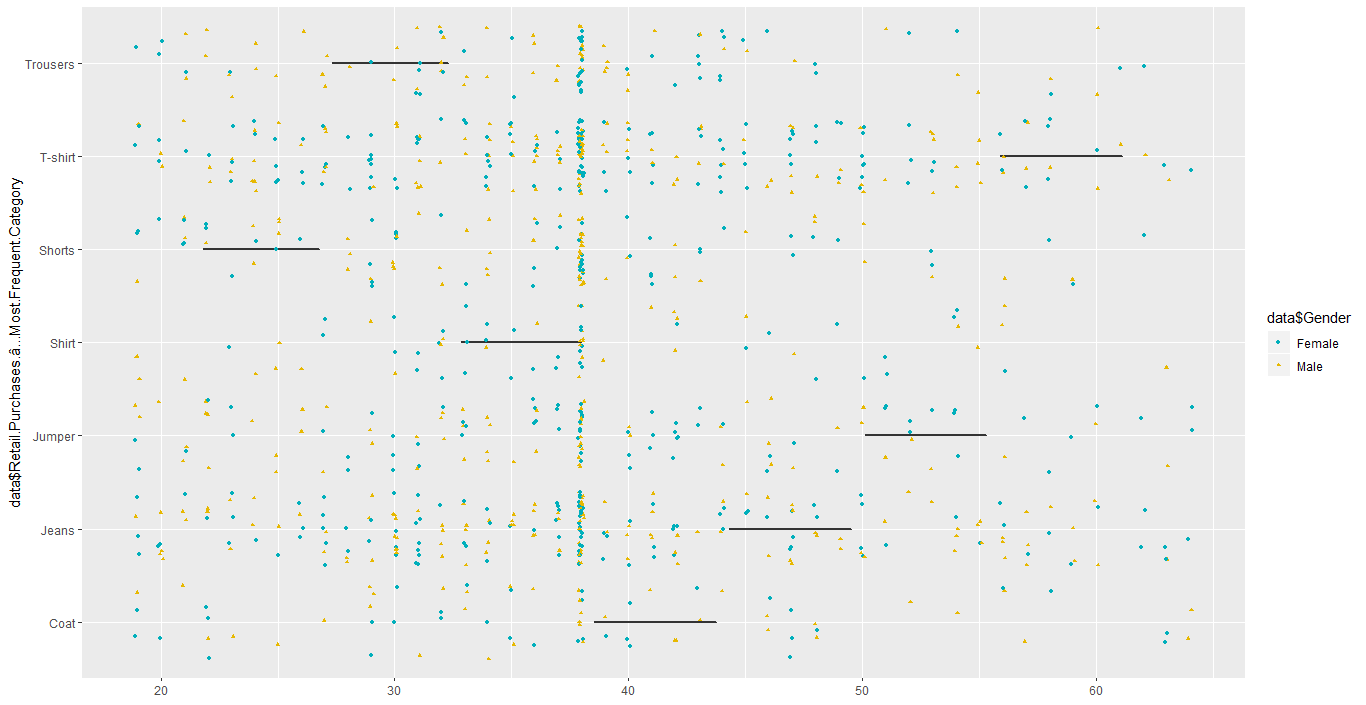


Fig: #Age and product choice

- There is a heavy concentration at Jeans and T-shirt, the preferred product choice among the population

- The older population especially female prefer to wear formal attire like coat and trousers

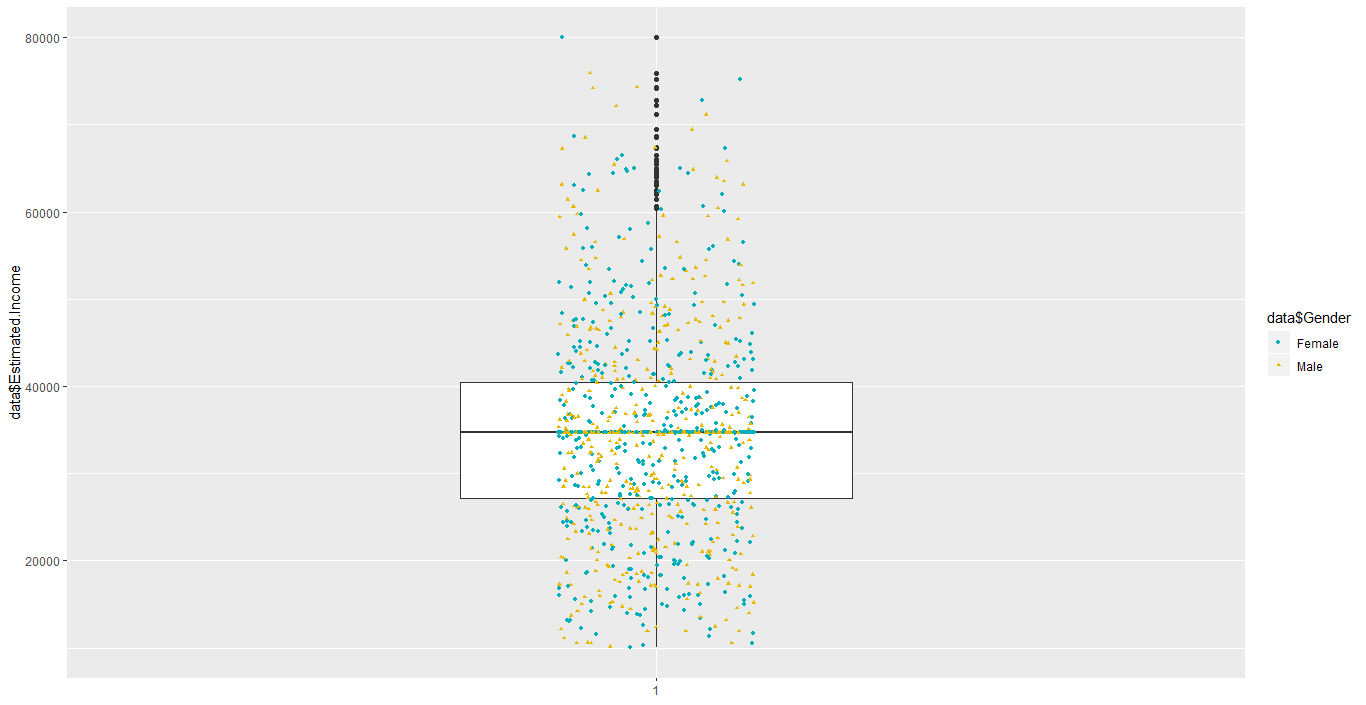


Fig: # Income and Gender-wise distribution

* Majority of the people are in the income slab of $30,000 -$40,000
* There is a gender wise wage gap in higher income slab of $80,000 and above. Males are earning more than females.

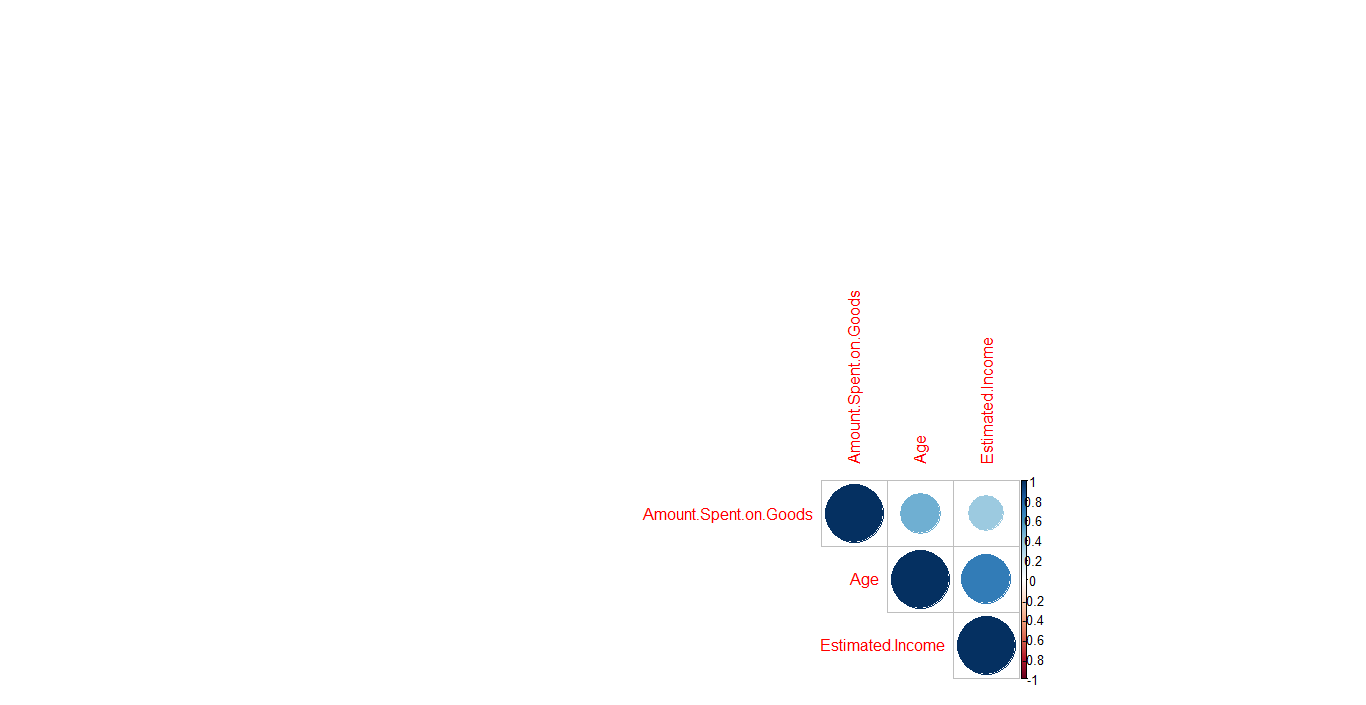


Fig: #correlation plot between the continuous variables in the dataset

* The amount spent is directly related to the age
* The age is directly linked to the income

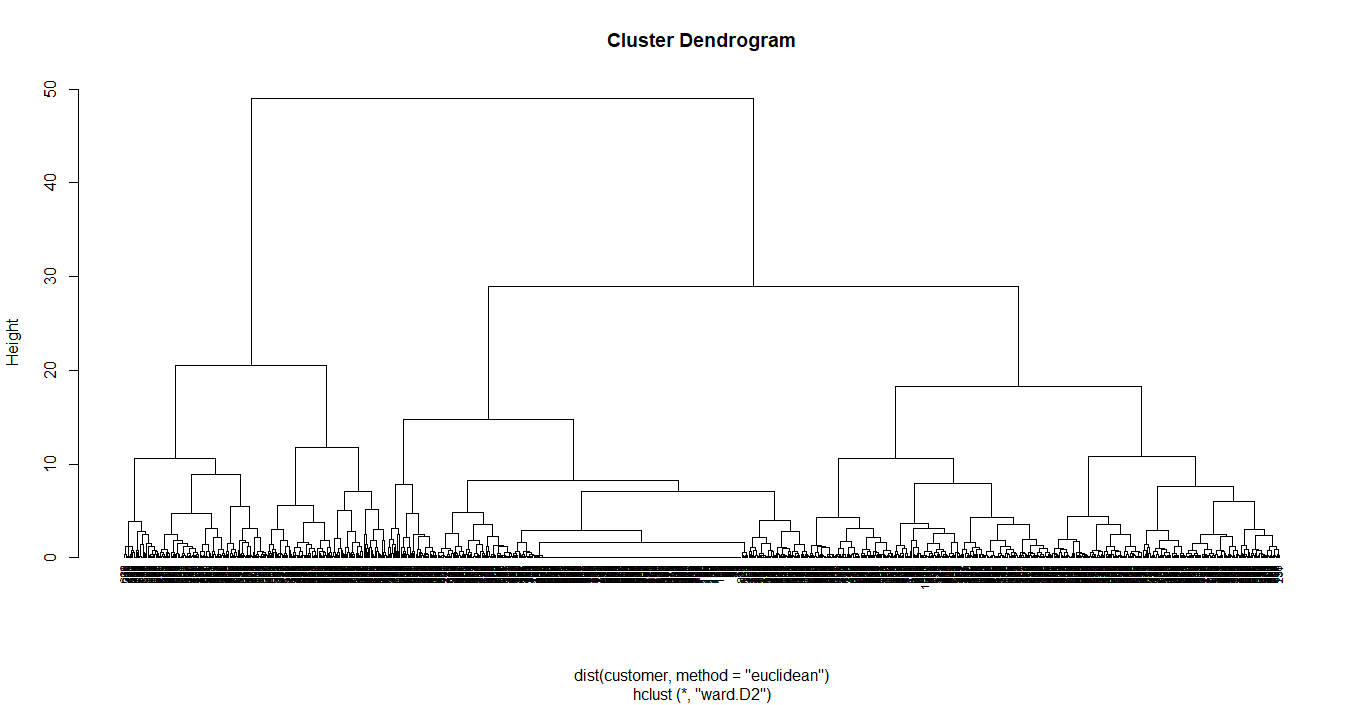


Fig: Clustered denrogram

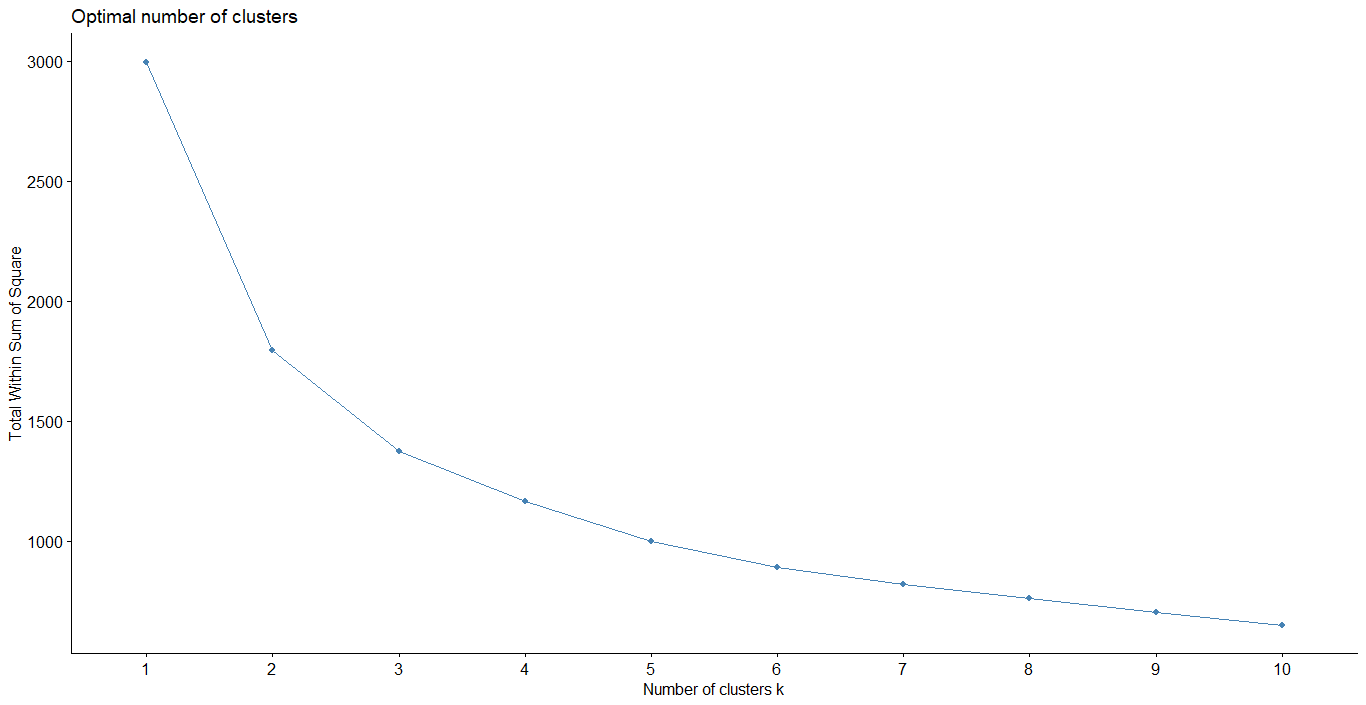


Fig: # of clusters defined using the elbow curve

If we see the elbow curve the clusters can be 2,3 or 4. Thus silhouette method gives us the best possible cluster formation.

The silhouette method. This method determines how well each object lies within its cluster. A high average silhouette width indicates a good clustering.

Silhouette refers to a method of interpretation and validation of consistency within clusters of data. The technique provides a succinct graphical representation of how well each object has been classified.

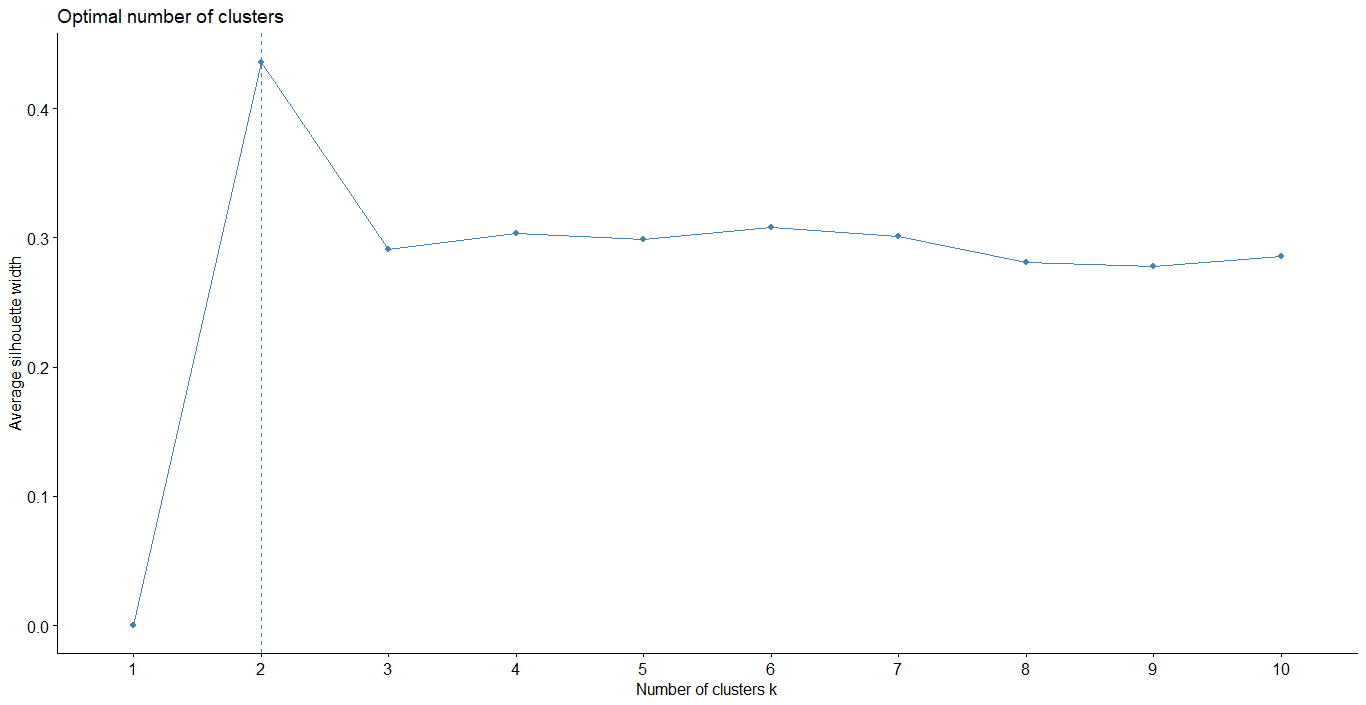


Fig: # of clusters (verification using another method called silhouette)

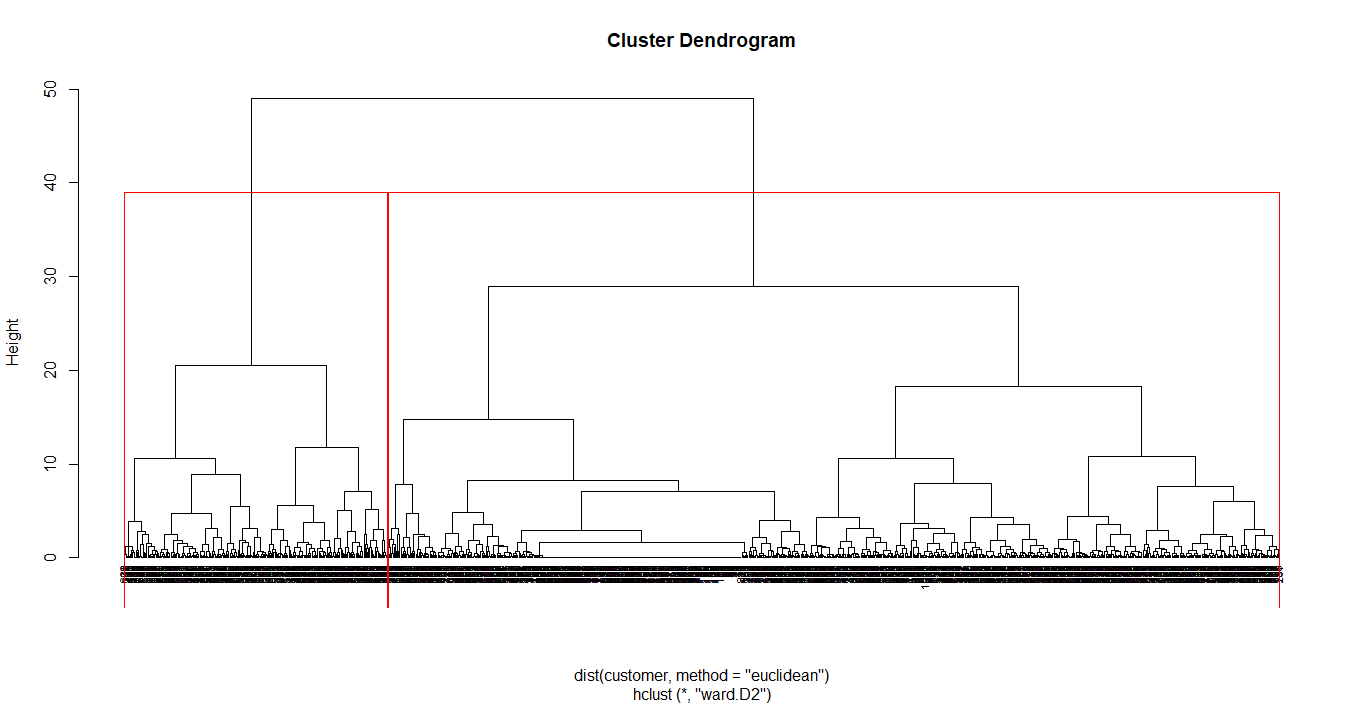
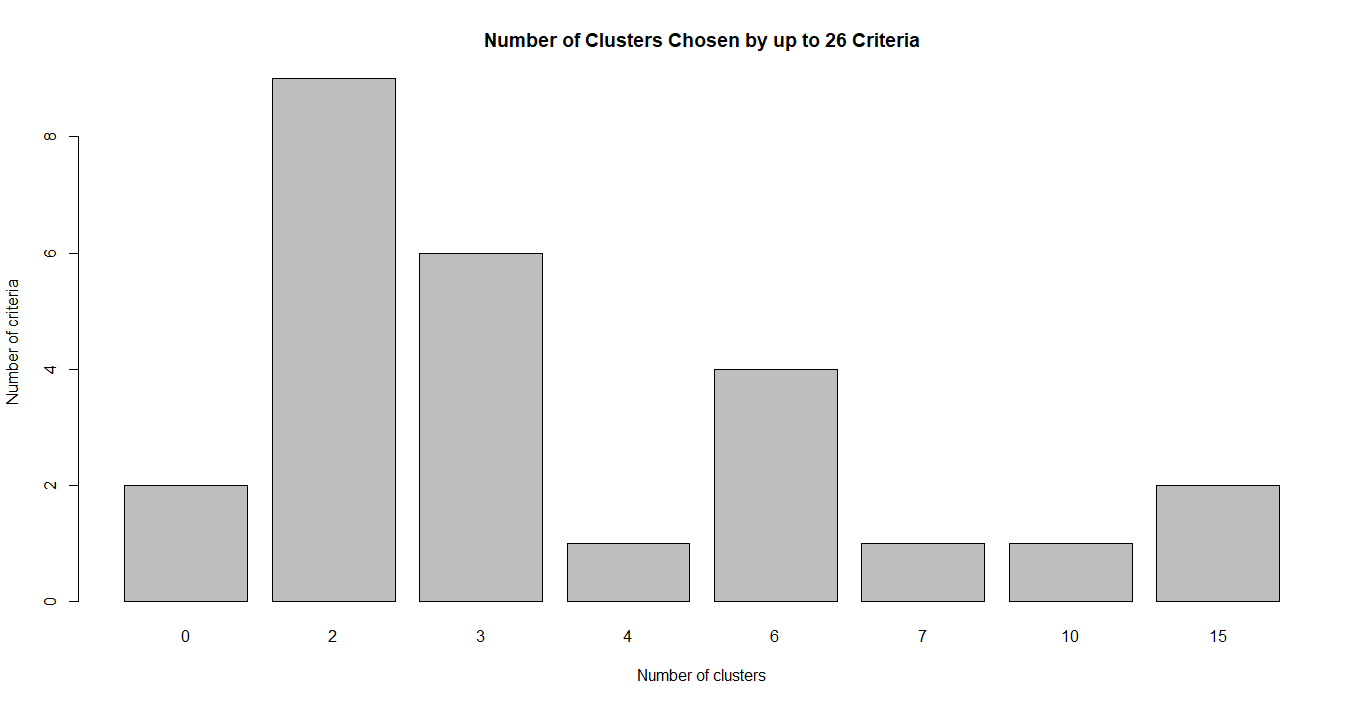


Fig: 2 clusters being formed based on the empirical study(please see the code)



I have tried to implement the recommender system using logistic regression (making gender as the dummy variable and based on the outcomes deciding on which product shall be recommended). To verify I have also used decision trees. However, on looking at the data in Excel using Pivot Tables. There is very little difference among the possible predictors with type of purchase. Therefore, it is unlikely that a useful recommender system can be developed.  I tried logistic regression and decision trees - neither worked well in terms of the accuracy.

In the logistic regression model, there are lot of important statistical parameters involved like fisher score iterations that are the number of times the model had to be run to get convergence. It’s an iterative process. The "Null deviance" is the sum of deviances for the logistic model with only the constant term. The "Residual deviance" is the sum of the deviances for the model with predictors involved.

Please see the appended code for the triage.

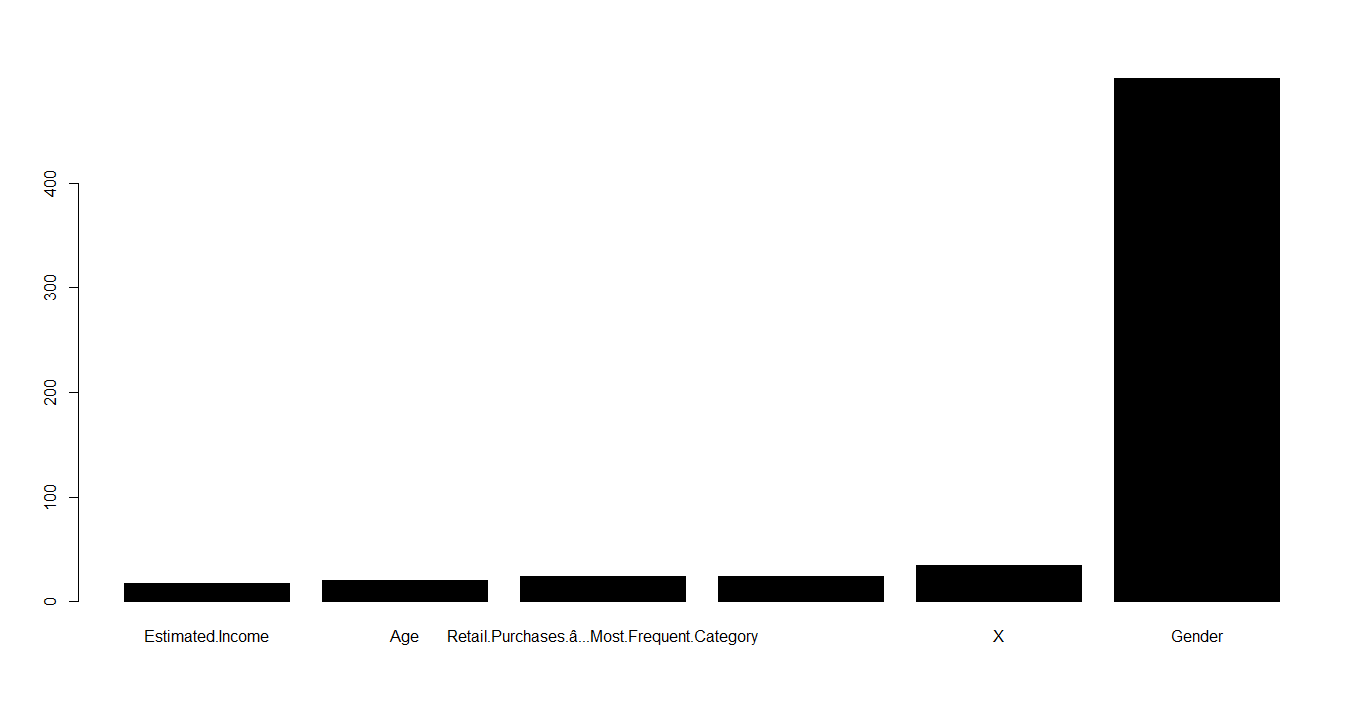


Fig: Variable importance using decision trees

**# Location based analytics**

For the purpose of location based analytics, I have made the intervals of income, age as well as Expenditure

* Residents of Michael Ville have used E-commerce websites more as compared to residents of other cities
* Generally, the residents of the cities of UK are not much comfortable in using mobile application for buying products. Henceforth, the data is scattered between 1-4 users.
* Talking about John Ville, it is female dominated city whereas Robert Ville is a male dominated city based on the dataset given
* People in Amande Ville prefer shopping as they tend to buy lot of products, **with jeans as their preferred product**
* The population seems to be more tilted towards buying jeans and T-shirt as their more preferred product
* The younger population seems to dominate the shopping trend, **as the age analysis shows that maximum number of people are in the age bracket of 15-45 years. Hence, it can be rightly inferred that people with this age slab would prefer more buying Jeans and T-shirt as compared to other line of products**
* Kimberly valley has the maximum aged people in the age group of 60-75 years
* Majority of the people are in the income range of $30,000-$40,000