Homework 1

Nikhil Ambati (50495129)

Given: We have a data set which has 25 property values and 15 features of them. Our objective is to identify what features we include in building our model. Also identify the issues which we see in the dataset and address them.

- **Issues with data set:**
  Firstly, we have identified whether if we have any NA or missing values in the dataset. As per **ColSums()** function combined with the **is.na** we could find that we have missing values for **Property 22.8** for the **feature5**, **Property 16.4** for the **feature8** and **Property 22.8** for the **feature12.**

  Solution: We could eliminate these missing values by using **na.omit().**

- **Normalizing the data**

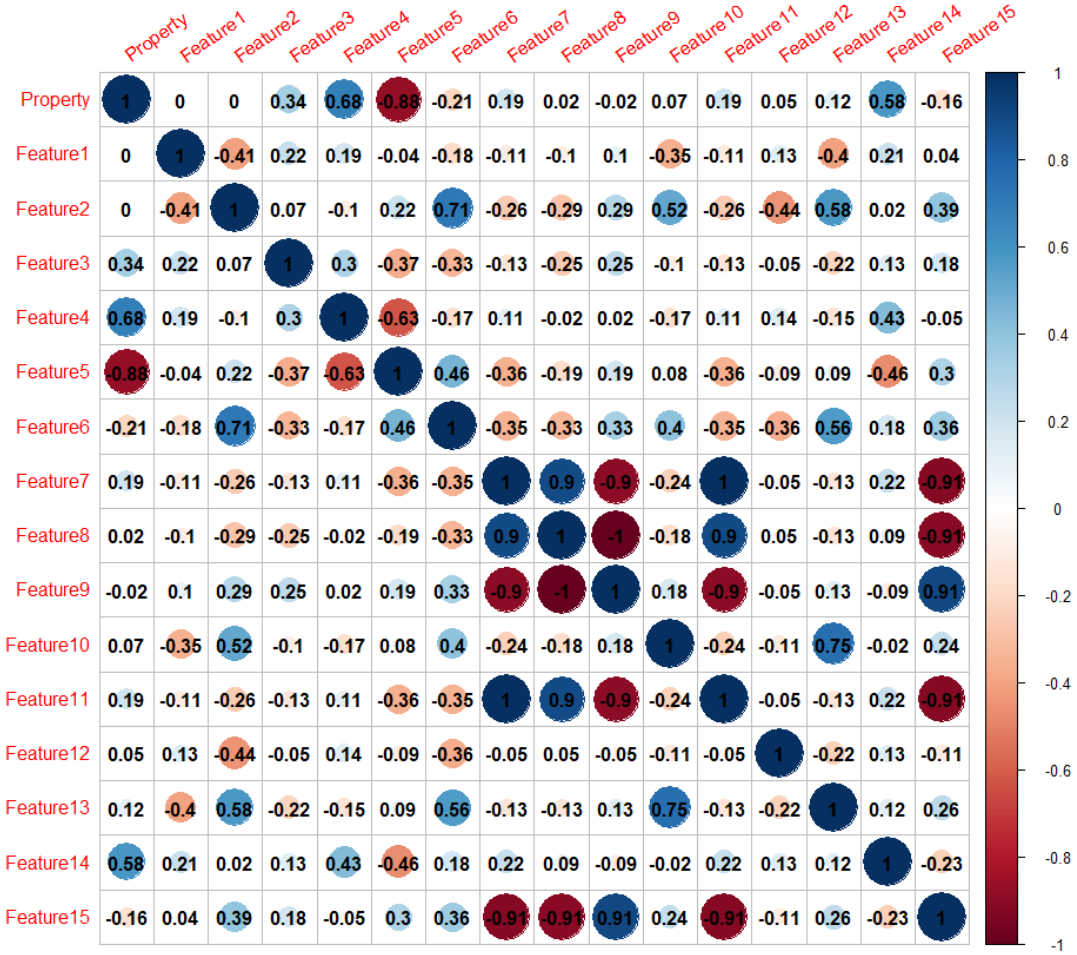
  Identify Non-Numeric values:
  Let us first identify if we have any non-numeric values in the dataset. We can do this by using function `is.numeric()` and in case if we have anything non-numeric it will return the columns. But in our case, all the values are numeric, so in return we got 0 columns.

  Using Scale Function:
  We could see that values are numeric but are in differently ranged, so we can use Scale() to transform the data so that it has a particular scale.

Variable Selection:

1) **Using Correlation**

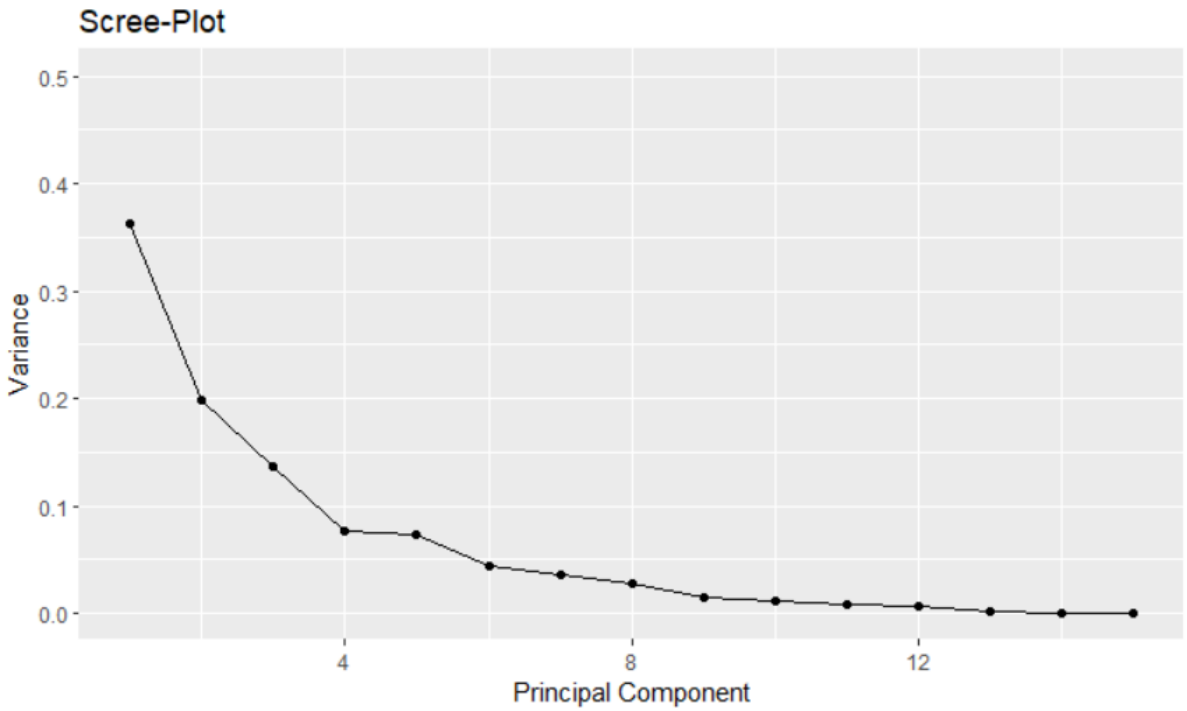
   Firstly, we will use correlation technique to select the variables. The below graph is plotted by using **corrplot()** method.



Conclusion:

1) **Feature 5** has negatively correlated with property and it has low correlation with other Features so we can consider this. Hence it adds important information to the model.
2) **Feature 6** and feature 2 has positive correlation, so we can consider Feature 6. Hence making it a viable option.
3) **Feature 7** and Feature 8 are positively correlated and also Feature 8 and Feature 11 are also correlated so, we could choose Feature 7 as substitute to others. Addition to this Feature 7 has low positive correlation to the property which will add fresh information.
4) **Feature 9** and Feature 15 are highly positively correlated, so we can consider Feature 9.
5) **Feature 10** and Feature 13 has positive correlation, thus making It a fresh addition to the information, so we can consider Feature 10.
6) **Feature 11** is negatively correlated. We could select Feature 11. Hence making it a good feature to keep in the dataset.
7) **Feature 12** has very low correlation to every other feature so we can consider this. It a fresh addition to the information.

2) **Scree Plot:**



8) We could find from the graph that **Feature 1, Feature 2, Feature 3 and Feature 4** has majority of variation. So, these Features are most significant since they contain majority of the information.
9) From **Feature 12 to Feature 15**, we could see that line is almost flat which mean there's no variation and information is very less. So we can eliminate these Features.

Conclusion:

As per the observation made from methods Correlation and Scree-Plot, I have selected features **Feature 1, Feature 2, Feature 3, Feature 4, Feature 5, Feature 6, Feature 7, Feature 9, Feature 10, Feature 11** and eliminated the following **Feature 8, Feature 12 to 15.** Hence the selected features will give majority of information and give add fresh information.