* **Project Title:-**

Housing Trends Prediction

* **Personal Information:-**

**Name** : Saad Azhar

**Email :** [sazhar2@masonlive.gmu.edu](mailto:sazhar2@masonlive.gmu.edu), [saadazhar1@gmail.com](mailto:saadazhar1@gmail.com)

* **Introduction and Problem Formulation:-**

**Problem:** Prediction of housing trends, prices and rents in a certain area.

The purpose of project is to provide customers with a system which would enable them to know about the prices and rents of the houses in future. Moreover another purpose of this project is to provide customers with houses which have similar characteristics while buying or renting. This feature would allow them to have multiple options to choose from. For course project I am dealing more specifically with house rent prediction.

* **Dataset Description:-**

I had Zillow house dataset which I used for rent and price prediction. In this dataset I had a total of 35 attributes. Out of which some were numerical and some were textual attributes. As I was dealing with numerical attributes in this project so I straight away removed majority of the textual attributes. Then, from the remaining numerical attributes around 7-8 were of no use for me. So I removed them too. Finally I got 13 attributes with which I decided to proceed with the project. These final 13 attributes are as follows:-

1. ID
2. Address
3. Zip Code
4. City
5. Bedrooms
6. Bathrooms
7. Estimated Price
8. Area Space
9. Year Built
10. Estimated Rent
11. Mortgage
12. Price Per Square Feet
13. Status of House(Condo/Town House/ Full house)

**Pre-Processing and Partitioning of data set:** My data set had a lot of missing values. To fill up those numerical missing values I used the mean technique to impute them. For textual missing values I straight away removed those records and for outliers I removed them too because they were affecting my resulting accuracy.

Moreover, my data set included information (status) for 3 types of houses which were Town Houses, Condos and Full houses. So I decided to separate the data set into 3 parts with respect to each house category. It enabled me to build separate models on three separate data sets and it ultimately helped me in increasing the accuracy of the model/algorithm from around 85% to above 90% in each cluster . After that I made clusters on these 3 separate datasets and divided each cluster into training and validation set with a ratio of 60% training data and 40% validation data. I had no test data set for this project because it was a case of unsupervised learning. At the end I had 3 separate models and separate results for each cluster of Townhouses, Condos and Full houses.

* **Project Milestones:-**

1. **Metadata Extraction and Imputation:**
   * 1. **Extracting metadata:**

I was provided with a data set for which there was no need to extract metadata as it was already preprocessed till metadata extraction point. I started with the imputation step in this project.

* + 1. **Dealing with missing data:**

There were a lot of numerical missing values in the data set. For that I used mean value imputations. For textual missing values I straight away removed those entries. After that I used Min-Max Normalization to normalize the data. This enabled me to further improve model’s accuracy and better clustering. I used Z-score normalization too but I achieved better accuracy using Min-Max Normalization so I decided to stick with it.

* + 1. **Dealing with outliers:**

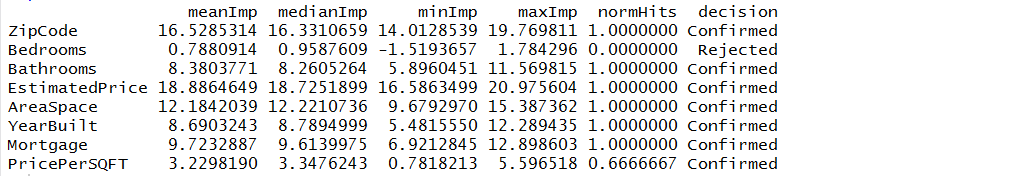
In case of outliers I straight away removed those entries as it causes very absurd and variable end results. Moreover outliers have a big effect on accuracy too. So I decided altogether to get rid of them.

* **Summarization and Visualization:-**

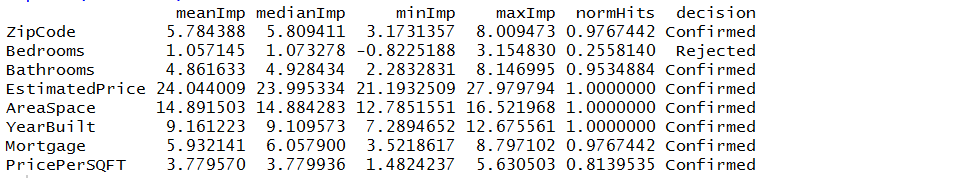
1. **Statistical summary:**
2. **Feature Selection:**

Following are the statistics of feature selection. For feature selection I am using a method known as boruta package in R. I applied this algorithm only on the numerical attributes.

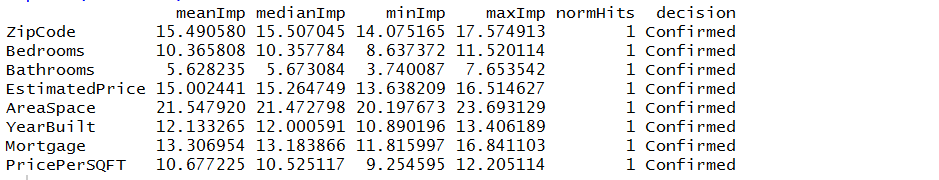
Feature Importance of Townhouses:



Feature Importance of Houses:



Feature Importance of Condos:

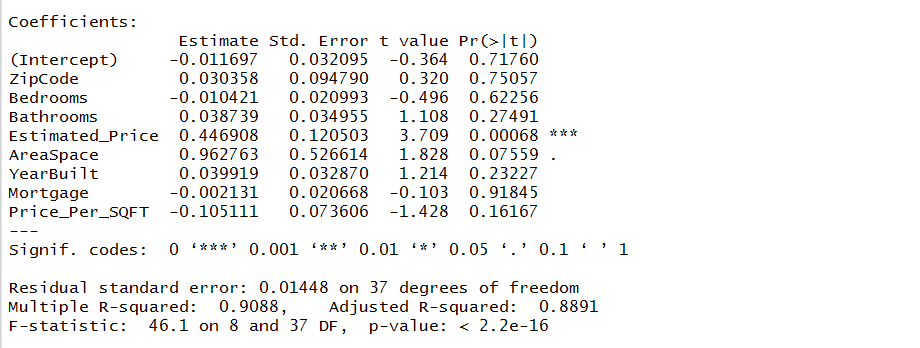


1. **Attributes P-Values:**

**Attributes P-Value for Houses:**

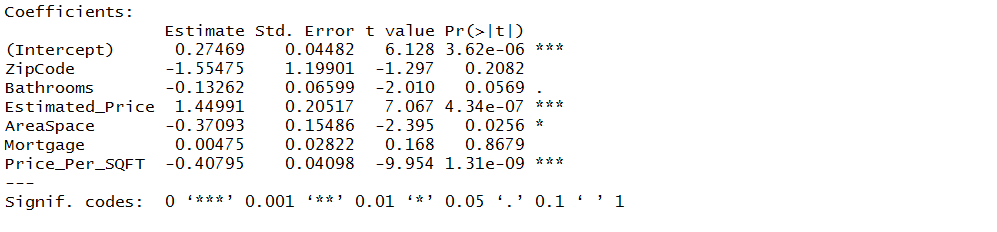
Following are the P-values for houses which tells us the significant variables used in model building. According to the following screenshot Zip Code, Bed rooms and Mortgage Value acts as insignificant in the regression model.

Yi6



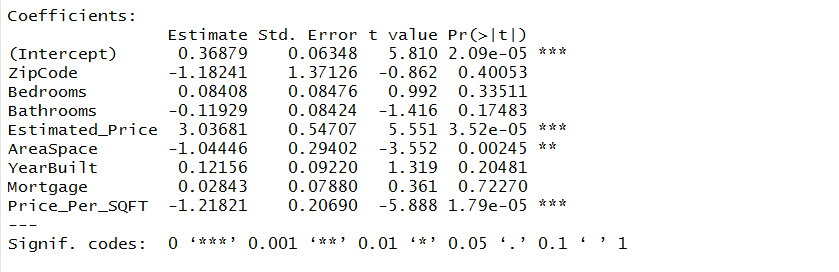
**Attributes P-Value for Town Houses:**

Following are the P-values for houses which tells us the significant variables used in model building. According to the following screenshot only Mortgage Value acts as insignificant in the regression model.



**Attributes P-Value for Condos:**

Following are the P-values for houses which tells us the significant variables used in model building. According to the following screenshot only Mortgage Value acts as insignificant in the regression model.



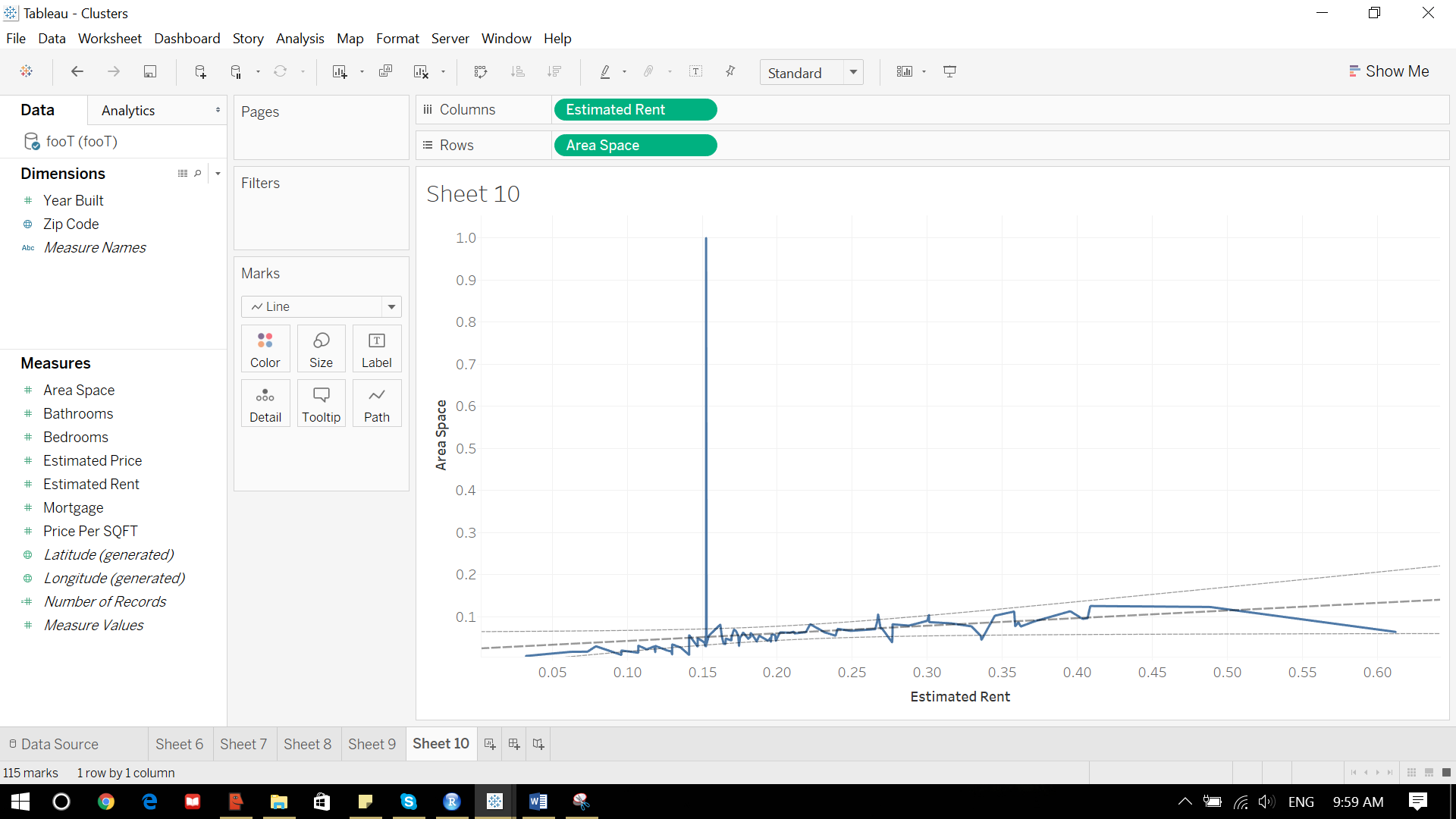
1. **Fitness and Accuracy statistical summary:-**

Fitness and Accuracy statistical summary is explained in detail in Model Section (Section 4). Kindly refer section 4 for model statistics.

1. **Histogram/bar/Line charts:-**

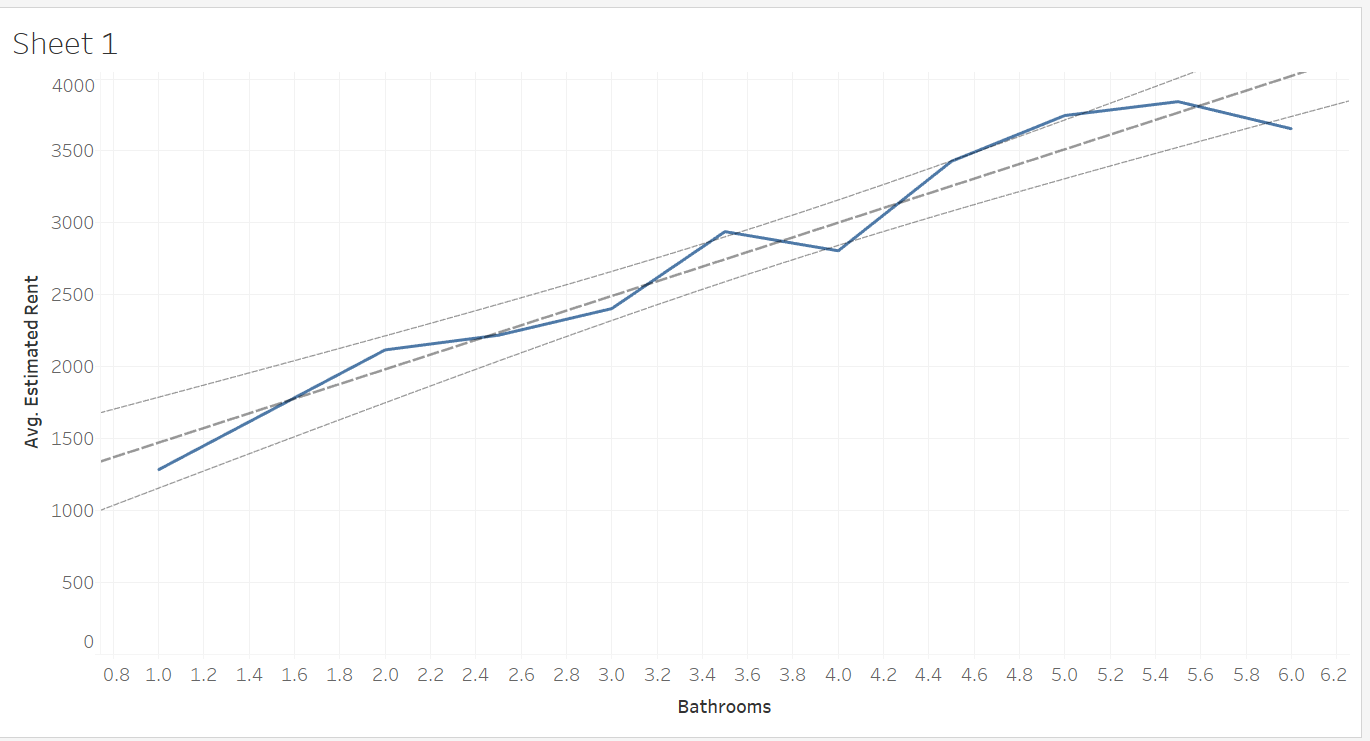
**Area Space Vs Estimated Rent:**

The following visualization chart is comparison in between Area Space and Estimated Rent among houses. In this cluster the rent increases as area space increases generally but there is one irregular instant where area space is maximum but rent is not maximum. This means that it can be an outlier. For instance after data normalization, if the area space is equal to 1 then it means that area is maximum and at this point the rent is even below the average rent range. Technically, with maximum area the rent should be close to the maximum range according to the trend shown in the graph. As this is not the case for this one particular instant so this huge irregular instant can be characterized as an outlier. Unique (one of) instant. Apart from this instant there is some sort gradual increase symmetry in the cluster.



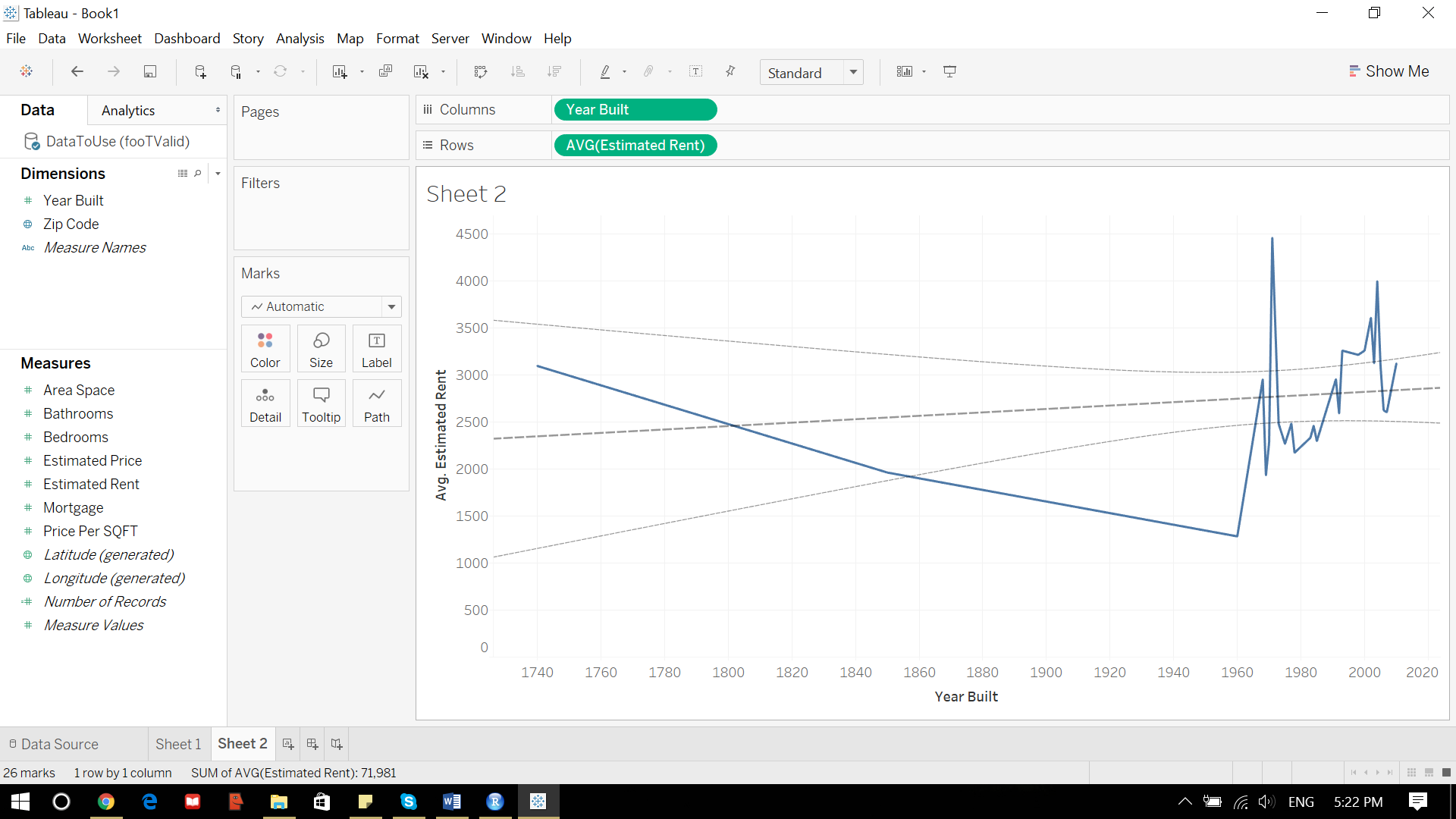
**Estimated Rent Vs Bathrooms:**

In the following screen shot it is again very clear that rent is increasing gradually with increasing number of bathrooms among houses. The peak average rent is at 5.5 bathrooms which is 3843$.



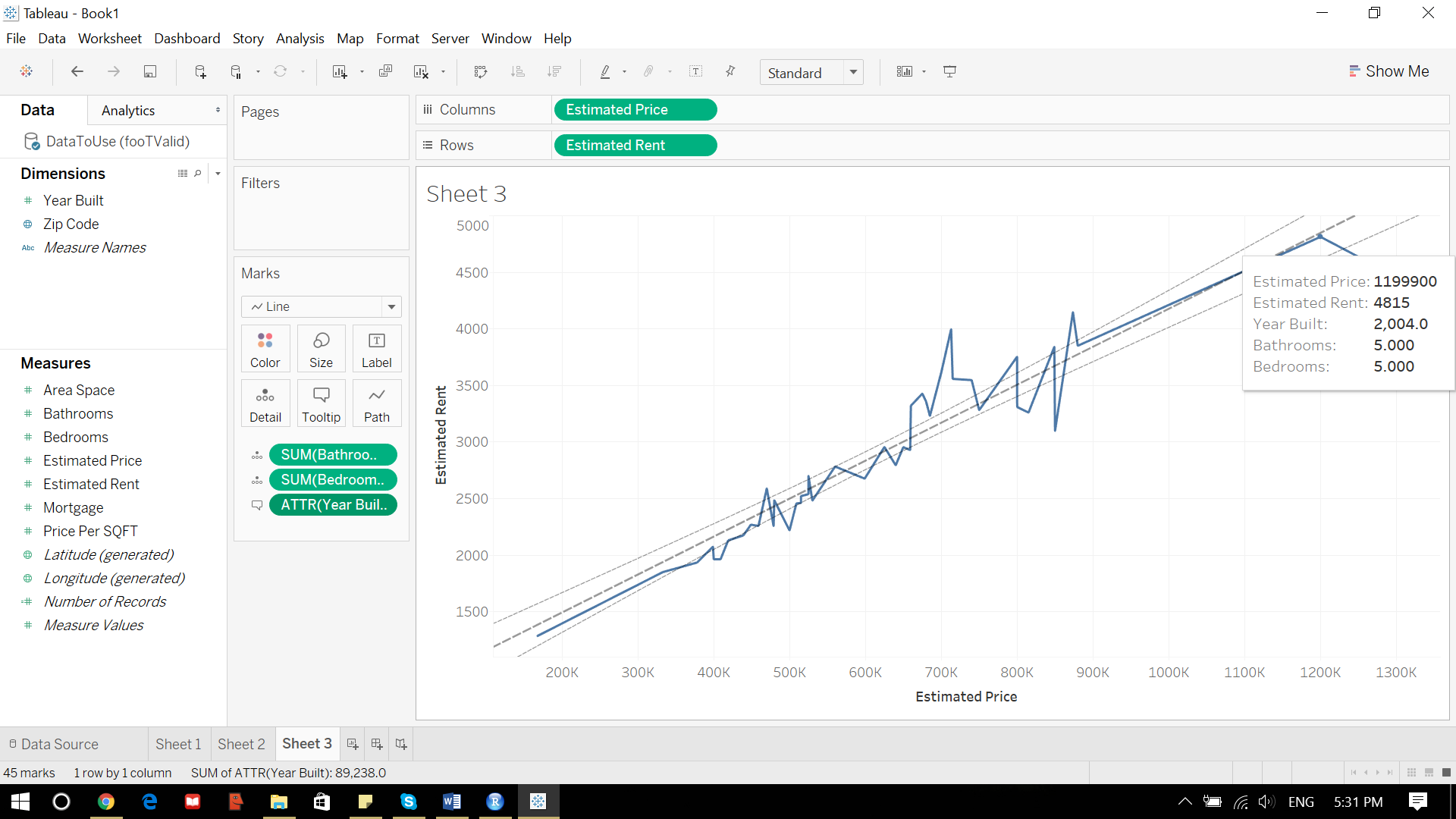
**Year Built Vs Estimated Rent:-**

In the following screen shot you can analyze that rent is decreasing in very old houses but it picks up after 1960 and then there is a rapid increase followed by variable decrease and increase. So the peak rent is after mid aged houses(relatively new houses).



**Estimated Rent Vs Estimated Price:-**

In the following screenshot if you see here maximum rent is at almost peak price with year 2004. Which means the house is relatively new. So it is depicting that houses which are relatively new are also expensive. Moreover there is a lot of variation in the data. Variable increase starts after 400K $ price till almost 1200K $ price.

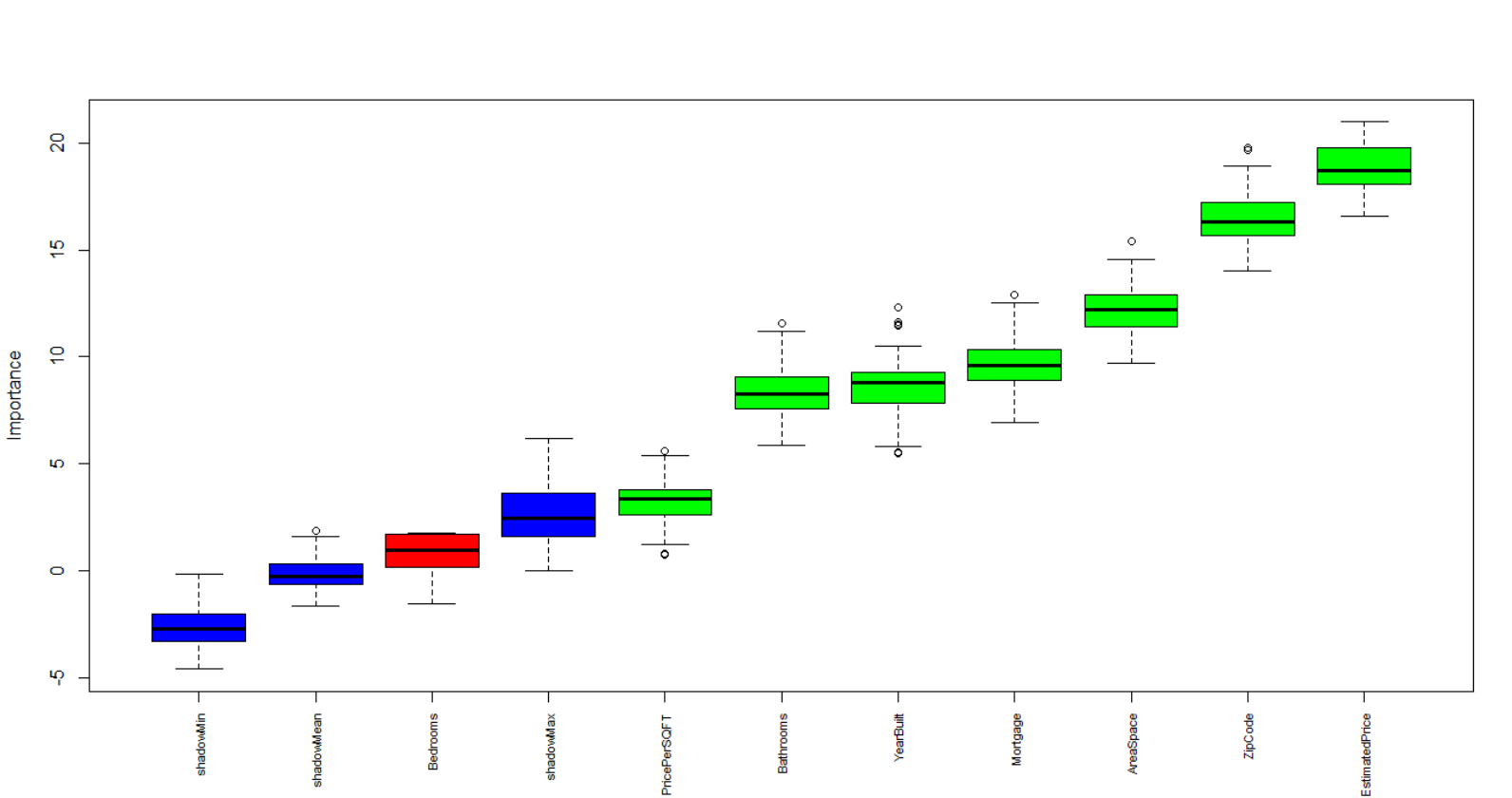


1. **Other plots:-**

Following are the plots for feature importance which I created from the output of feature selection boruta package. Working of boruta package, I’ll be explaining in the next step. These plots just depict the importance of all features.

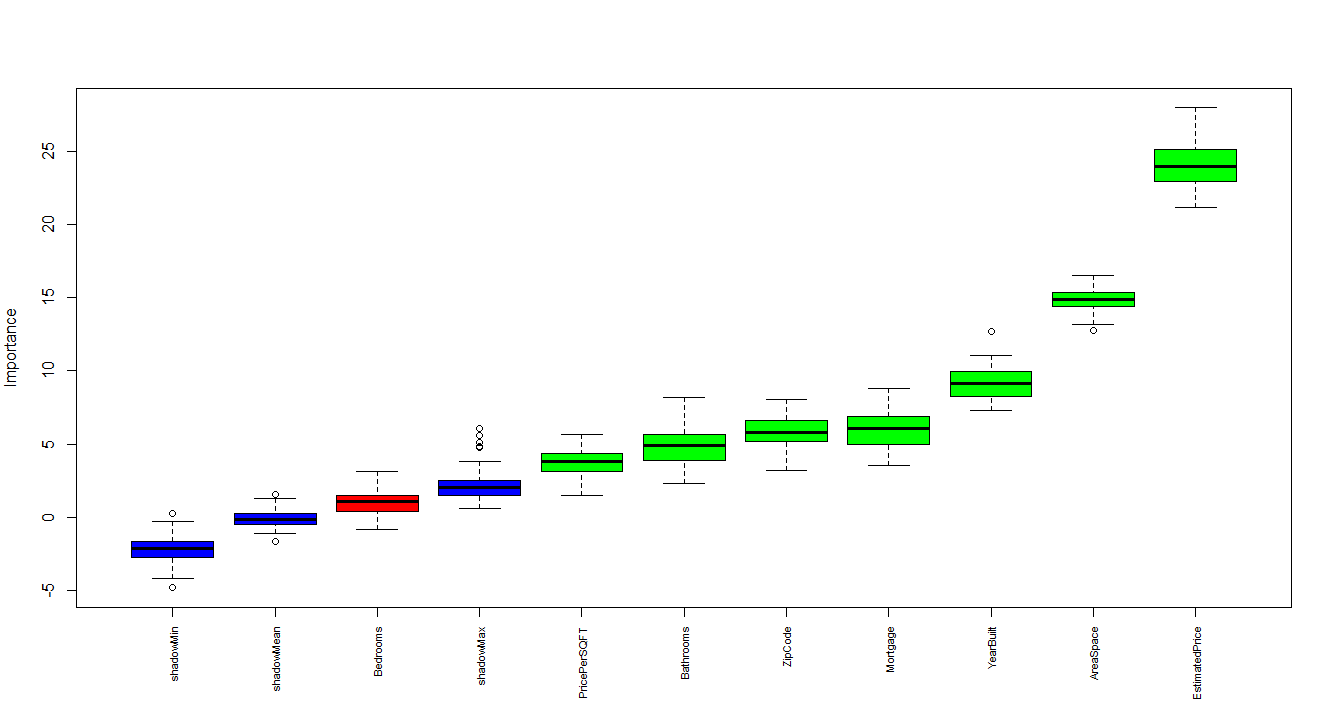
**Importance Plot for Townhouses:**

Feature with the most importance is being represented by top right green box plot which is estimated price in this case and so on. Green color represent all important and confirmed features. Red color represents the rejected features which in this case is bedroom and blue color represent the mean, min, max of shadow values which are used to calculate feature importance in boruta package.



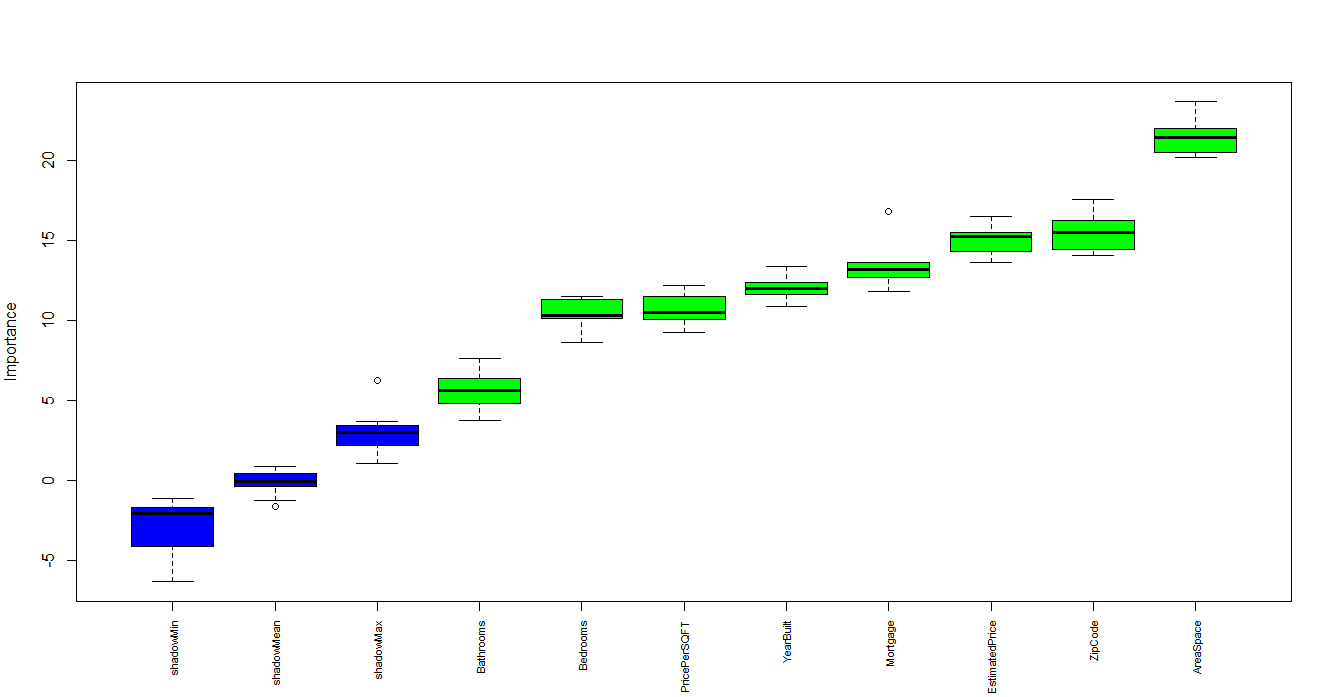
**Importance Plot for Houses:**

Feature with the most importance is being represented by top right green box plot which is estimated price in this case and so on. Green color represent all important and confirmed features. Red color represents the rejected features which in this case is bedroom and blue color represent the mean, min, max of shadow values which are used to calculate feature importance in boruta package.



**Importance Plot for Condos:**

Feature with the most importance is being represented by top right green box plot which is Area Space in this case and so on. Green color represent all important and confirmed features. Blue color represent the mean, min, max of shadow values which are used to calculate feature importance in boruta package.



1. **Clustering**

**Important attributes/features:**

Important attributes/features which I extracted were on the basis of Boruta feature importance package. The importance visual plots are given in the previous section. In this section I am going to explain the working of Boruta Package. Its working is as follows:-

1. The First step involves adding randomness to the given data set by creating shuffled copies of all features. Shuffled copies are also known as shadow features.
2. Second step is about training a random forest algorithm on the extended data set and applies a feature importance measure. Mean Decrease Accuracy is used to evaluate the importance of each feature where higher means more important.
3. Third step is about checking whether a real feature has a higher importance than the best of its shadow features at every iteration and constantly removes features which are deemed highly unimportant.
4. At the end, the algorithm stops either when all features gets confirmed or rejected.

By applying boruta package on all numerical attributes of my dataset, almost all of the attributes were confirmed just Bedroom was rejected as it’s clearly visualized in the feature importance plot in the previous section.

**Grouping objects based on certain attributes/features:-**

For grouping and clustering I just used 4 important features which were Estimated Price, Year Built, Area Space and bathrooms for better classification. Moreover, I used two different methods two combine houses with similar characteristics which are as follows:-

1. K-Means Algorithm
2. EM Algorithm

Ultimately I decided to go with EM algorithm for clusters generation after analysis of results and issues. The reasons are as follows:

1. **K-Means Algorithm:**

The issue with K-Means algorithm is that it’s very difficult to calculate exact number of clusters which could be sufficient for a specific data set and as my data could be increasing in future so it becomes very difficult to calculate the number of clusters each time. The issue is there is no predefined method to calculate number of clusters.

1. **EM Algorithm:**

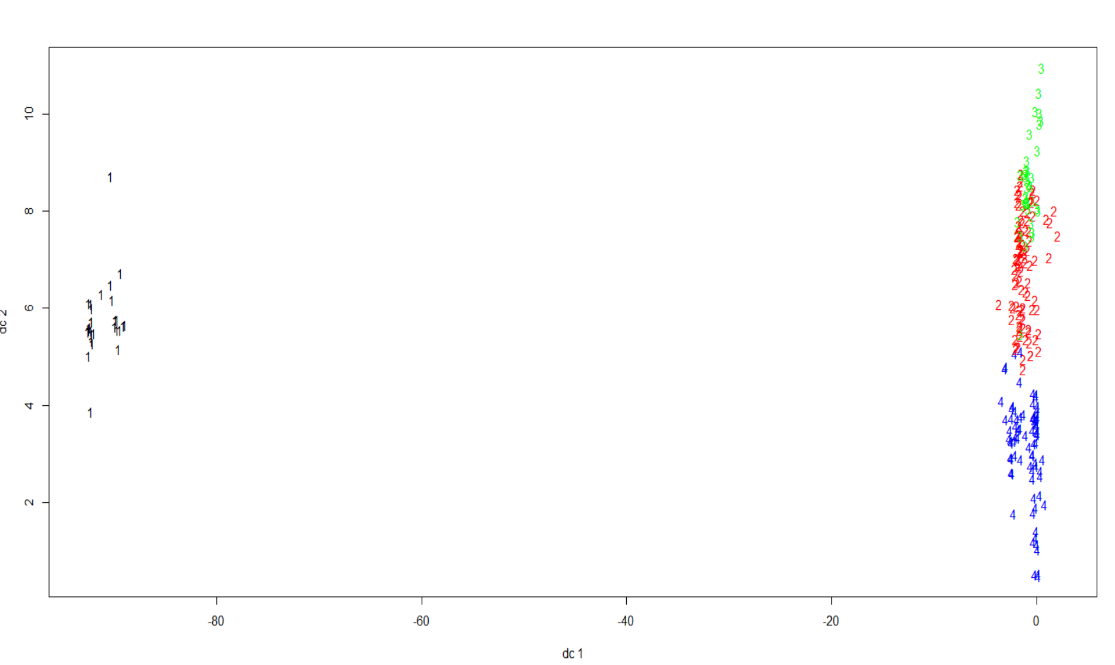
The benefit of EM algorithm is that it dynamically creates number of clusters according to the data set. The advantage achieved due to this feature is that if the size of data increases then the algorithm automatically calculates new number of clusters by itself and divides the data according to that new number.

This advantage of EM Cluster ultimately helped me in choosing this algorithm. Results and accuracy of model achieved from both the clustering algorithms were more or less the same. May be K-Means was clustering the data a little better but I needed to find a tradeoff between both algorithms and EM Cluster were helping me more in that cause.

**Cluster plots:-**

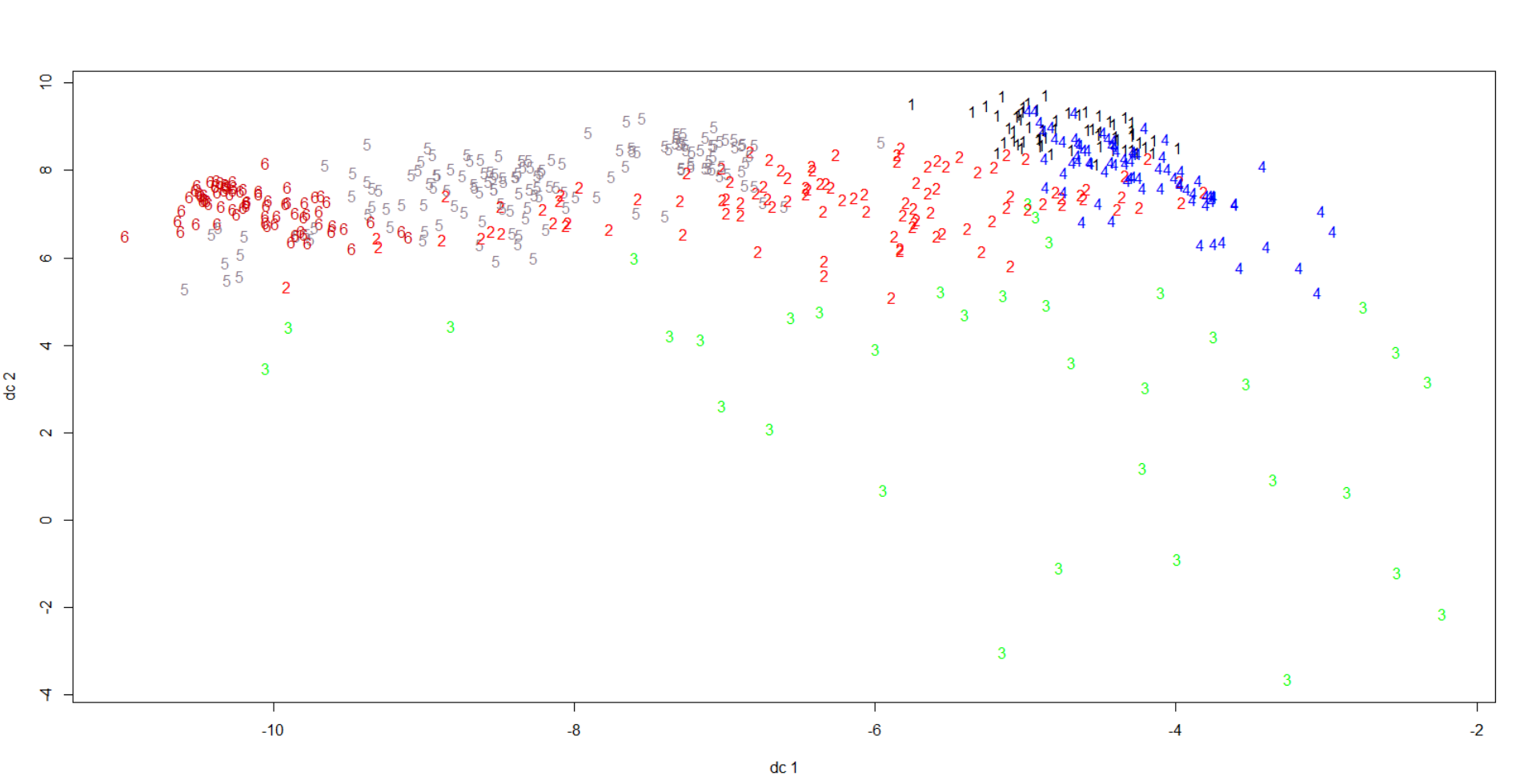
**EM Cluster plots for Townhouses:**

4 cluster were generated for townhouses according to this plot. These number of cluster will vary according to the data set

****

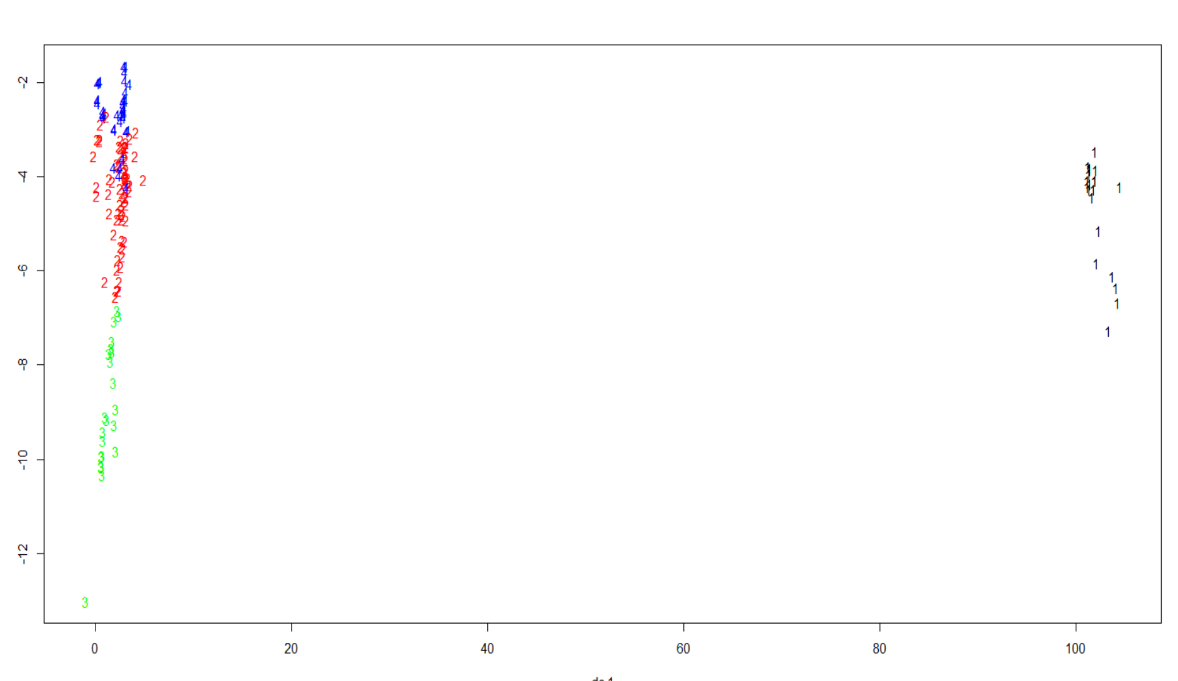
**EM Cluster plots for Houses:**

6 clusters were generated for houses according to this plot. These number of cluster will vary according to the data set.



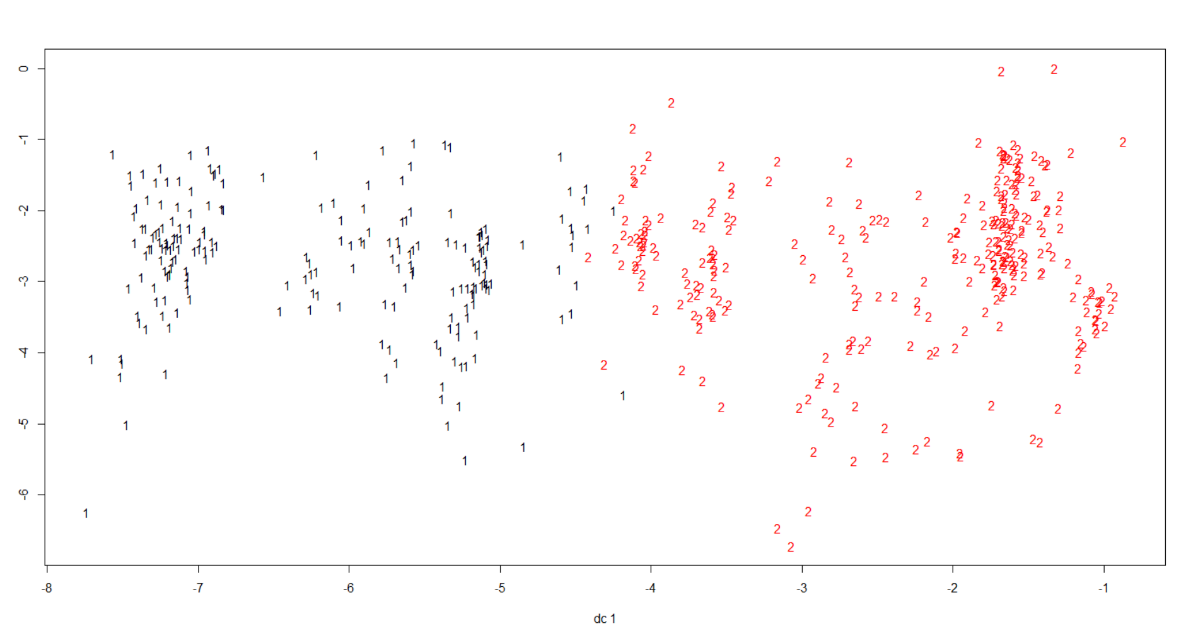
**EM Cluster plots for Condos:**

4 clusters were generated for condos according to this plot. These number of cluster will vary according to the data set.

****

**K-Means Cluster Plot for Houses:**

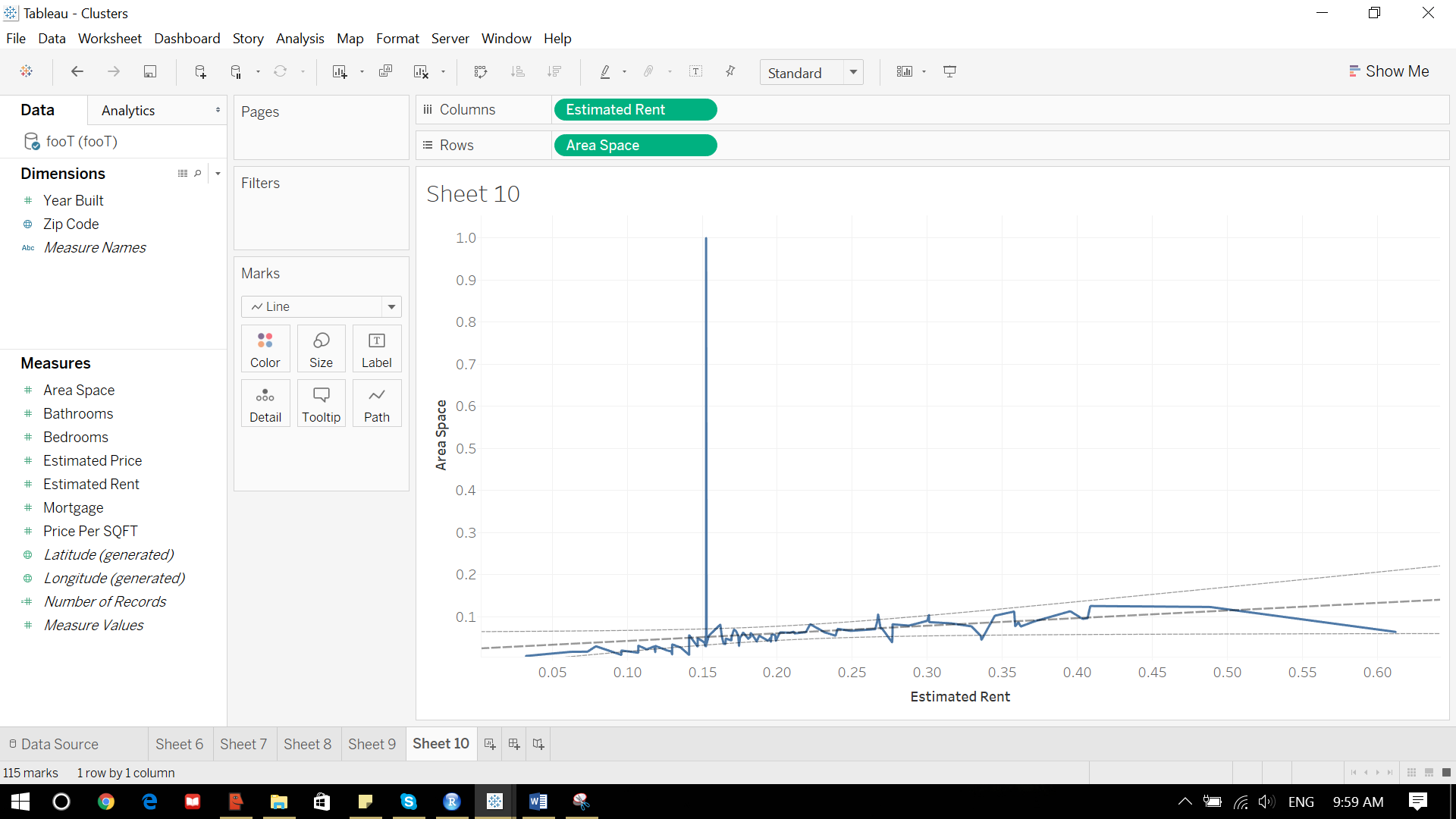
If you see the following plot of K-Means Clusters of Houses and compare it with the above plot of EM- Clusters of Houses then its pretty clear that K-Means is classifying the data in a much better way but the only issue arises when the size of data increases. It cannot computer value of K (clusters) dynamically.

****

* **Analytics:-**
  1. **Outliers:-**

**Area Space Vs Estimated Rent:**

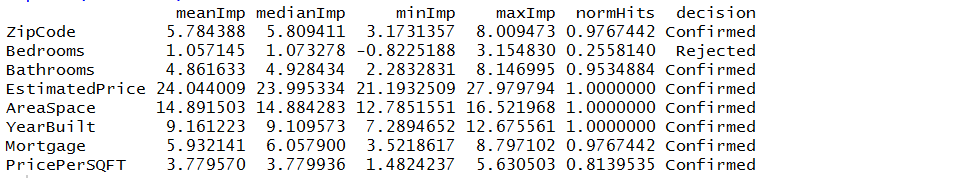
The following visualization chart is comparison in between Area Space and Estimated Rent among houses. In this cluster the rent increases as area space increases generally but there is one irregular instant where area space is maximum but rent is not maximum. This means that it can be an outlier. For instance after data normalization, if the area space is equal to 1 then it means that area is maximum and at this point the rent is even below the average rent range. Technically, with maximum area the rent should be close to the maximum range according to the trend shown in the graph. As this is not the case for this one particular instant so this huge irregular instant can be characterized as an outlier. Unique (one of) instant. Apart from this instant there is some sort gradual increase symmetry in the cluster.

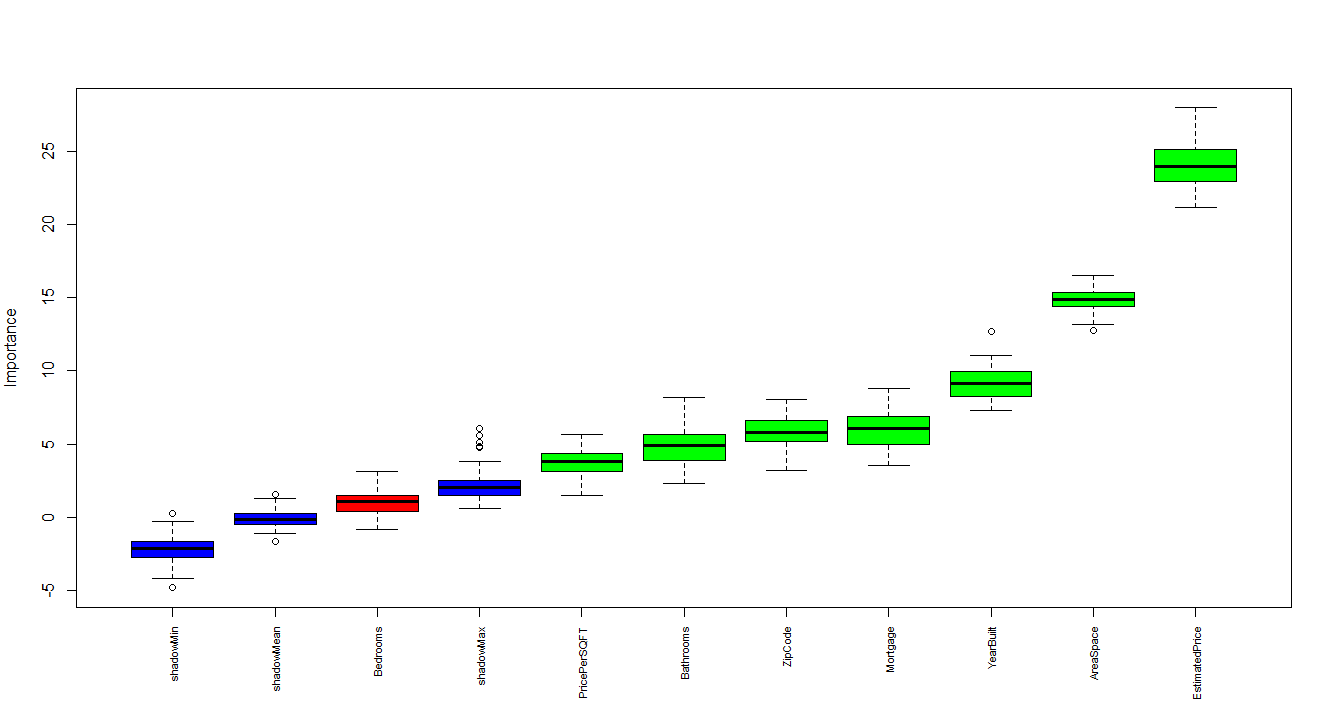


* 1. **Analysis of feature importance plots:-**

**Importance Plot for Houses:**

In this plot feature with the most importance is being represented by top right green box plot which is estimated price in this case and so on. Green color of box plots represent all important and confirmed features. Red color represents the rejected features which in this case is bedroom and blue color represent the mean, min, max of shadow values which are used to calculate feature importance in boruta package. This plot is exactly correlated to the following statistics of houses deduced from Boruta Package.





* **Model Development:-**

1. **Algorithms:-**

The Algorithms which I used to predict House rent are as follows:-

* 1. Multiple Regression
  2. Backward Multiple Regression

I am achieving accuracy of above 90% using both these algorithm and in almost all the clusters of full house, townhouse and Condos Dataset. Following are the attached results of full house, townhouse and Condos Datasets using these two algorithms with detailed analysis on their accuracies and errors.

**Results:-**

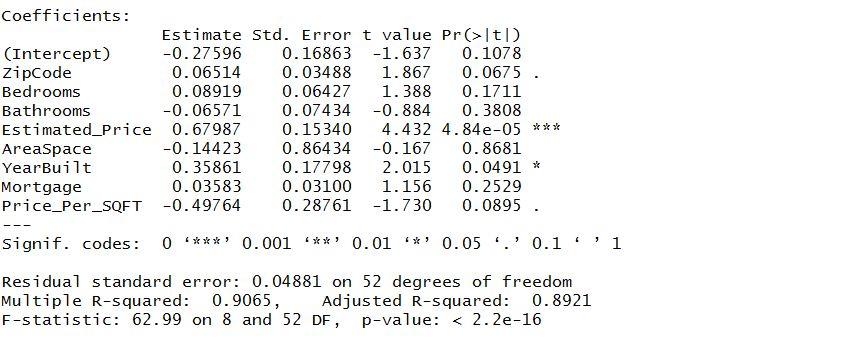
I am including results for some of the clusters. As there are around 14 clusters in total and to calculate them for two different algorithms makes 28 outputs. It would be difficult to analyze all of them. That’s why I am going with some of the clusters.

**Full Houses:-**

**Multiple Regression:-**

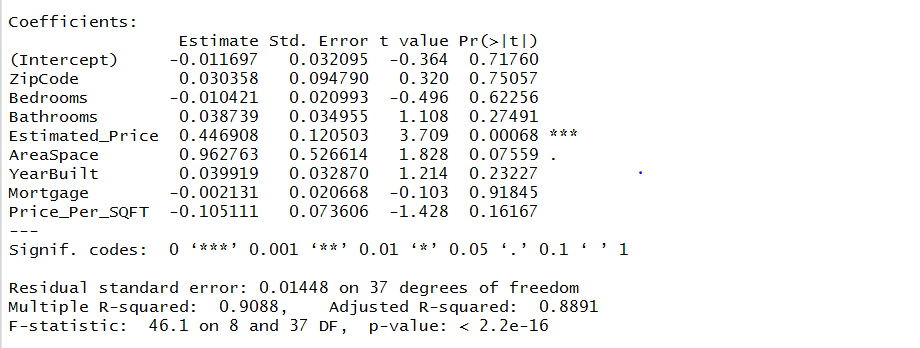
**Cluster 1:**

Accuracy of this cluster is 90.65% which means that model is predicting rents correctly with more than 90 percent of accuracy and its Residual error is 0.04881 which is very less. It means that this cluster’s prediction results are very reliable and good.



**Cluster 2:**

Accuracy of this cluster is 90.88% which means that model is predicting rents correctly with more than 90 percent of accuracy and its Residual error is 0.0144 which is very less. It means that this cluster’s prediction results are very reliable and good.

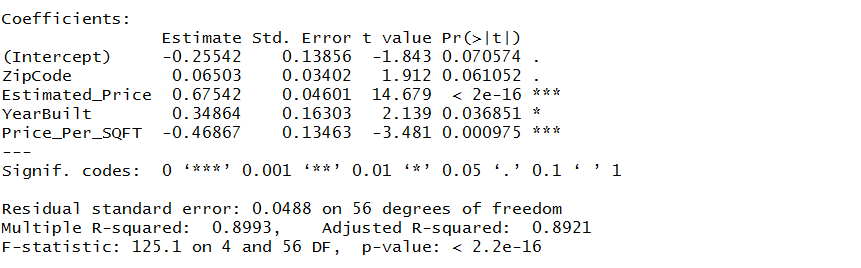


If you compare these two clusters. Then cluster 2 is producing more accurate results than cluster 1. The reason is R- Squared accuracy is more and residual error is less in cluster 2.

**Backward Multiple Regression:-**

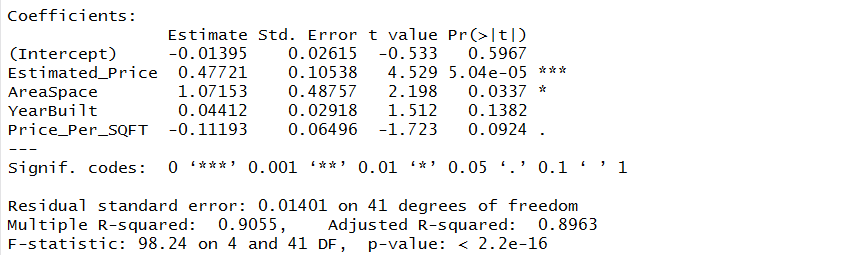
**Cluster 1**

Accuracy of this cluster is 89.93% which means that model is predicting rents correctly with more than 89 percent of accuracy and its Residual error is 0.0488 which is very less. It means that this cluster’s prediction results are very reliable and good.



**Cluster 2:-**

Accuracy of this cluster is 90.55% which means that model is predicting rents correctly with more than 90 percent of accuracy and its Residual error is 0.01401 which is very less. It means that this cluster’s prediction results are very reliable and good.



If you compare these two clusters. Then cluster 2 is producing more accurate results than cluster 1. The reason is R- Squared accuracy is more and residual error is less in cluster 2.

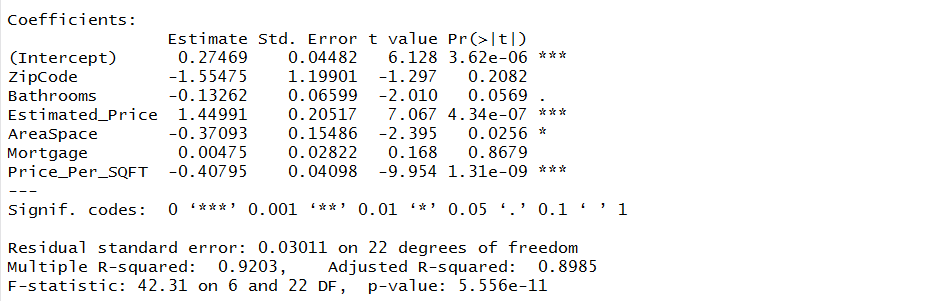
**Moreover, Multiple Regression is predicting rents in full houses with more accuracy than Backward Multiple Regression because it has more accuracy and generally less error.**

**Townhouses:-**

**Multiple Regression:-**

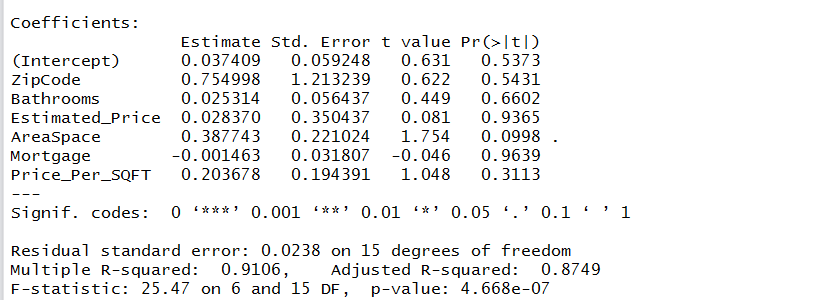
**Cluster 1:**

Accuracy of this cluster is 92.03% which means that model is predicting rents correctly with more than 90 percent of accuracy and its Residual error is 0.03011 which is very less. It means that this cluster’s prediction results are very reliable and good.



**Cluster 2:**

Accuracy of this cluster is 91.06% which means that model is predicting rents correctly with more than 90 percent of accuracy and its Residual error is 0.0238 which is very less. It means that this cluster’s prediction results are very reliable and good.

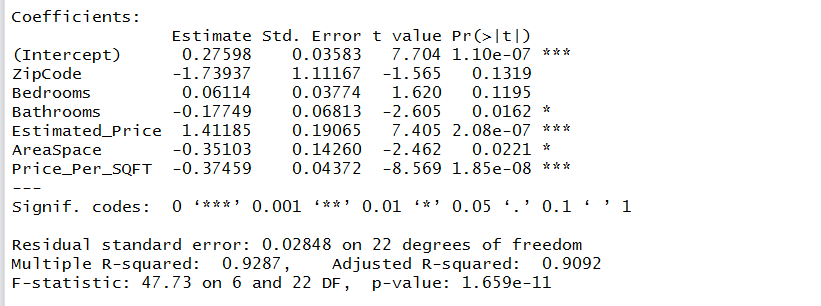


If you compare these two clusters. Then cluster 1 is producing more accurate results than cluster 1. The reason is R- Squared accuracy is more in cluster 2.

**Backward Multiple Regression:-**

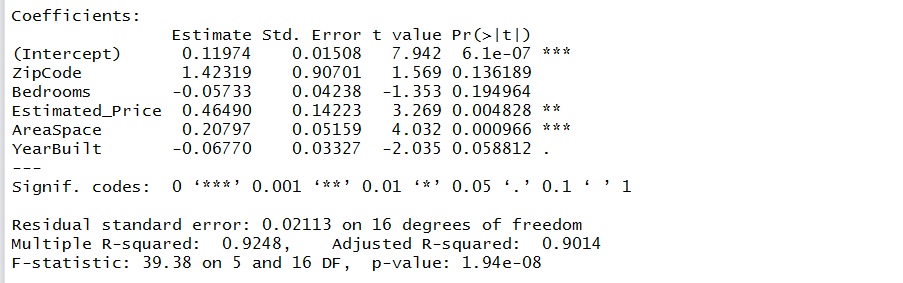
**Cluster 1**

Accuracy of this cluster is 92.87% which means that model is predicting rents correctly with more than 90 percent of accuracy and its Residual error is 0.02848 which is very less. It means that this cluster’s prediction results are very reliable and good.



**Cluster 2:-**

Accuracy of this cluster is 92.48% which means that model is predicting rents correctly with more than 90 percent of accuracy and its Residual error is 0.02113 which is very less. It means that this cluster’s prediction results are very reliable and good.



If you compare these two clusters. Then cluster 1 is producing more accurate results than cluster 1. The reason is R- Squared accuracy is more in cluster 2.

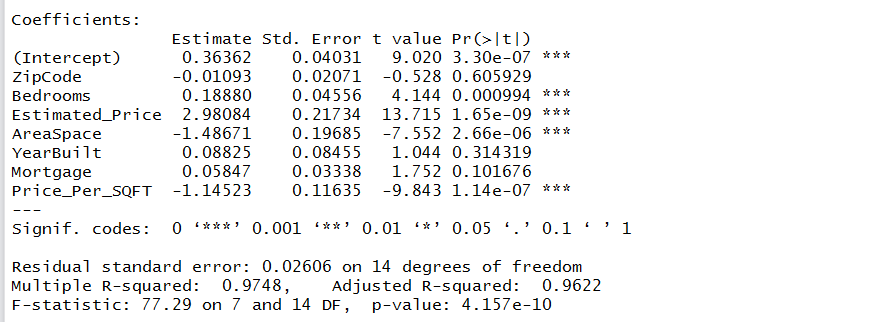
**Moreover, Backward Multiple Regression is predicting rents in Townhouses with more accuracy than Multiple Regression because it has more accuracy and generally less error.**

**Condos:-**

**Multiple Regression:-**

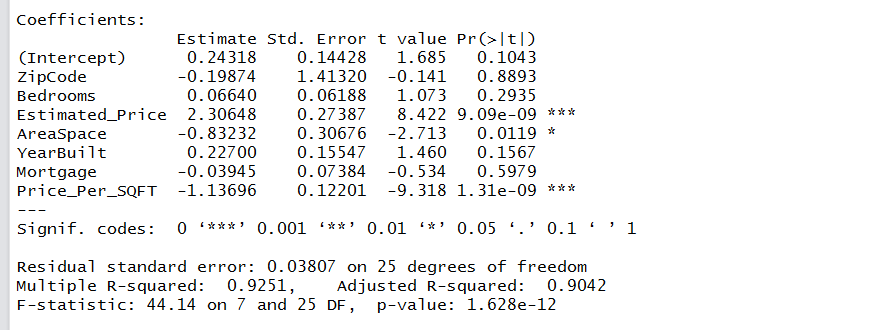
**Cluster 1:**

Accuracy of this cluster is 97.48% which means that model is predicting rents correctly with more than 90 percent of accuracy and its Residual error is 0.02606 which is very less. It means that this cluster’s prediction results are very reliable and good.



**Cluster 2:**

Accuracy of this cluster is 92.51% which means that model is predicting rents correctly with more than 90 percent of accuracy and its Residual error is 0.03807 which is very less. It means that this cluster’s prediction results are very reliable and good.

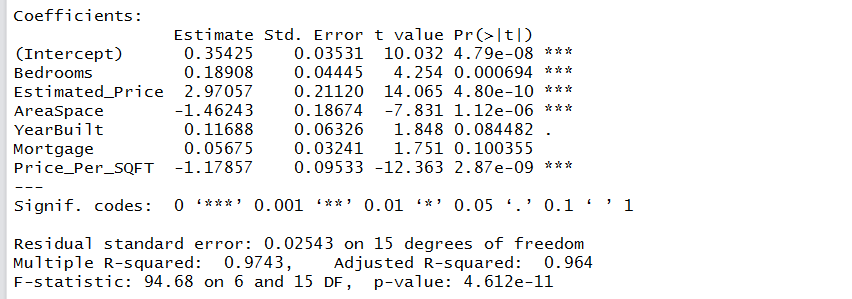


If you compare these two clusters. Then cluster 1 is producing more accurate results than cluster 1. The reason is R- Squared accuracy is more and residual error is less in cluster 2.

**Backward Multiple Regression:-**

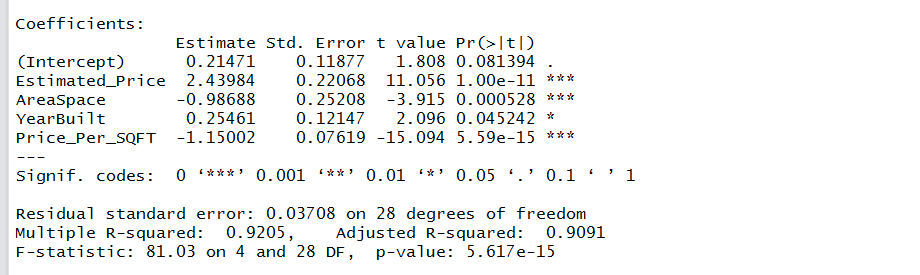
**Cluster 1**

Accuracy of this cluster is 97.43% which means that model is predicting rents correctly with more than 90 percent of accuracy and its Residual error is 0.0254which is very less. It means that this cluster’s prediction results are very reliable and good.



**Cluster 2:-**

Accuracy of this cluster is 92.05% which means that model is predicting rents correctly with more than 90 percent of accuracy and its Residual error is 0.03708 which is very less. It means that this cluster’s prediction results are very reliable and good.



If you compare these two clusters. Then cluster 1 is producing more accurate results than cluster 1. The reason is R- Squared accuracy is more and residual error is less in cluster 2.

**Moreover, Multiple Regression is predicting rents in Condos with more accuracy than Backward Multiple Regression because it has more accuracy and generally less error.**

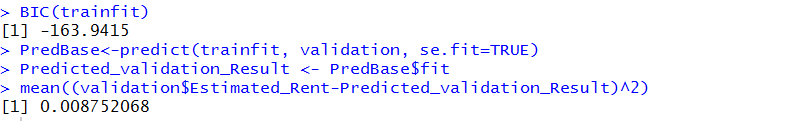
1. **Measure of the loss function:-**

**BIC and Mean Squared Error:-**

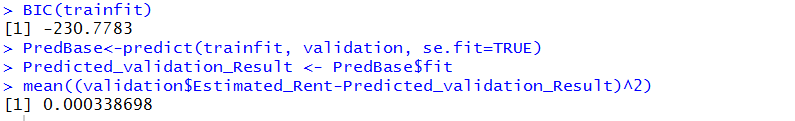
**Full Houses:-**

**Multiple Regression:-**

**Cluster 1:**



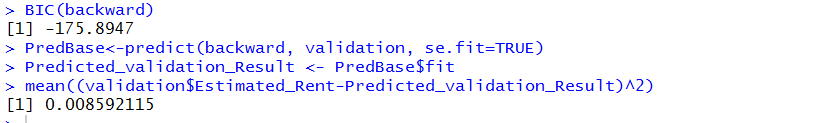
**Cluster 2:**



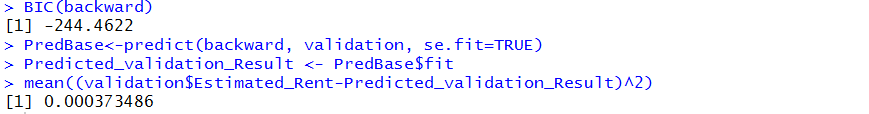
If you compare these two clusters. Then cluster 2 is producing more accurate results than cluster 1. The reason is Mean Squared Error and BIC value in cluster 2 is less than cluster 1.

**Backward Multiple Regression:-**

**Cluster 1**



**Cluster 2:-**

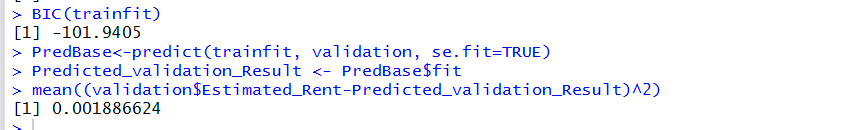


If you compare these two clusters. Then cluster 2 is producing more accurate results than cluster 1. The reason is Mean Squared Error and BIC value in cluster 2 is less than cluster 1.

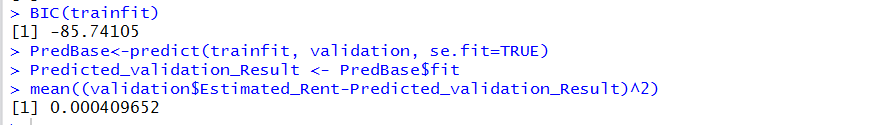
**Townhouses:-**

**Multiple Regression:-**

**Cluster 1:**



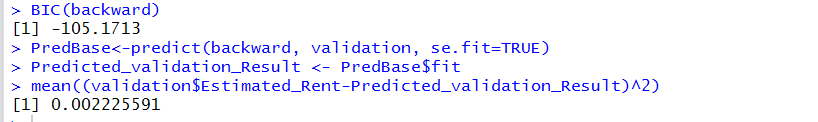
**Cluster 2:**



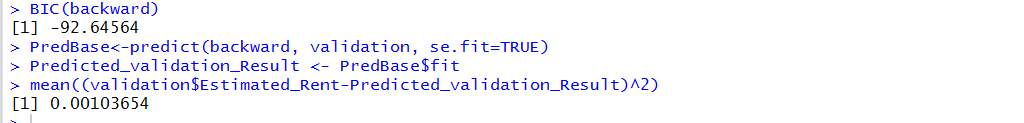
If you compare these two clusters. Then cluster 2 is producing more accurate results than cluster 1. The reason is Mean Squared Error value in cluster 2 is less than cluster 1.

**Backward Multiple Regression:-**

**Cluster 1**



**Cluster 2:-**

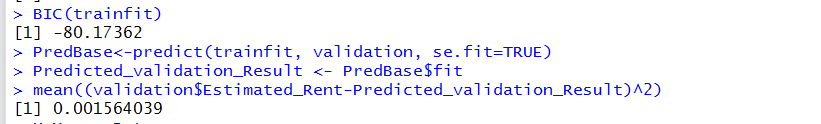


If you compare these two clusters. Then cluster 2 is producing more accurate results than cluster 1. The reason is Mean Squared Error value in cluster 2 is less than cluster 1.

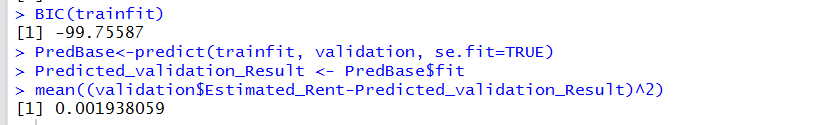
**Condos:-**

**Multiple Regression:-**

**Cluster 1:**



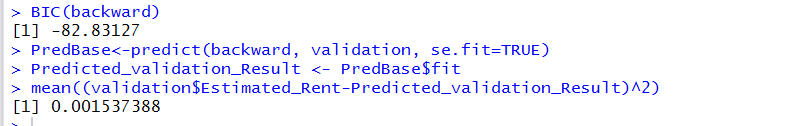
**Cluster 2:**



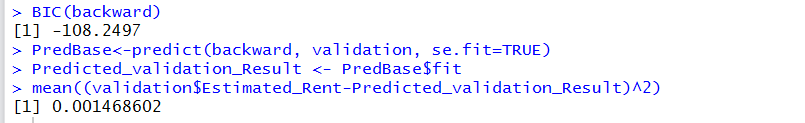
If you compare these two clusters. Then cluster 1 is producing more accurate results than cluster 2. The reason is Mean Squared Error value in cluster 2 is less than cluster 1.

**Backward Multiple Regression:-**

**Cluster 1**



**Cluster 2:-**



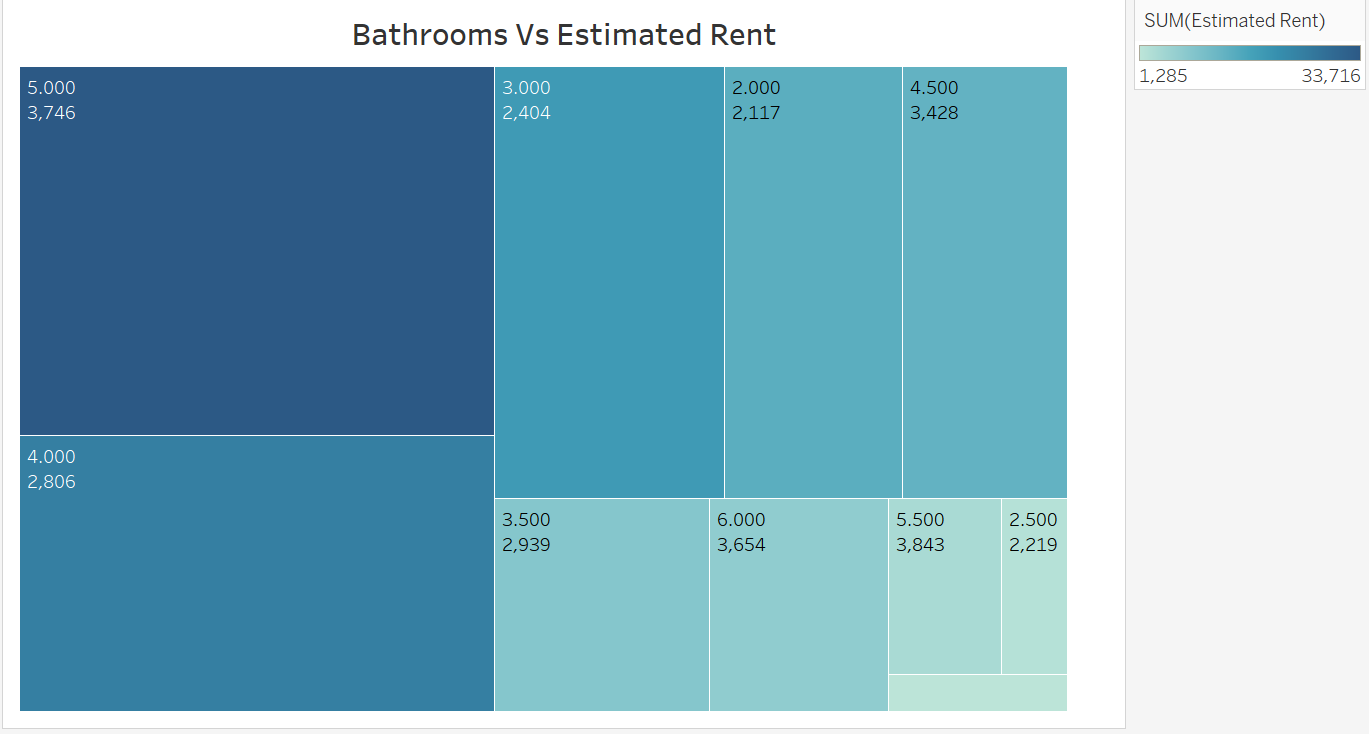
If you compare these two clusters. Then cluster 2 is producing more accurate results than cluster 1. The reason is Mean Squared Error and BIC value in cluster 2 is less than cluster 1.

1. **Overall Optimization:-**

For overall optimization of the system I am using following techniques:-

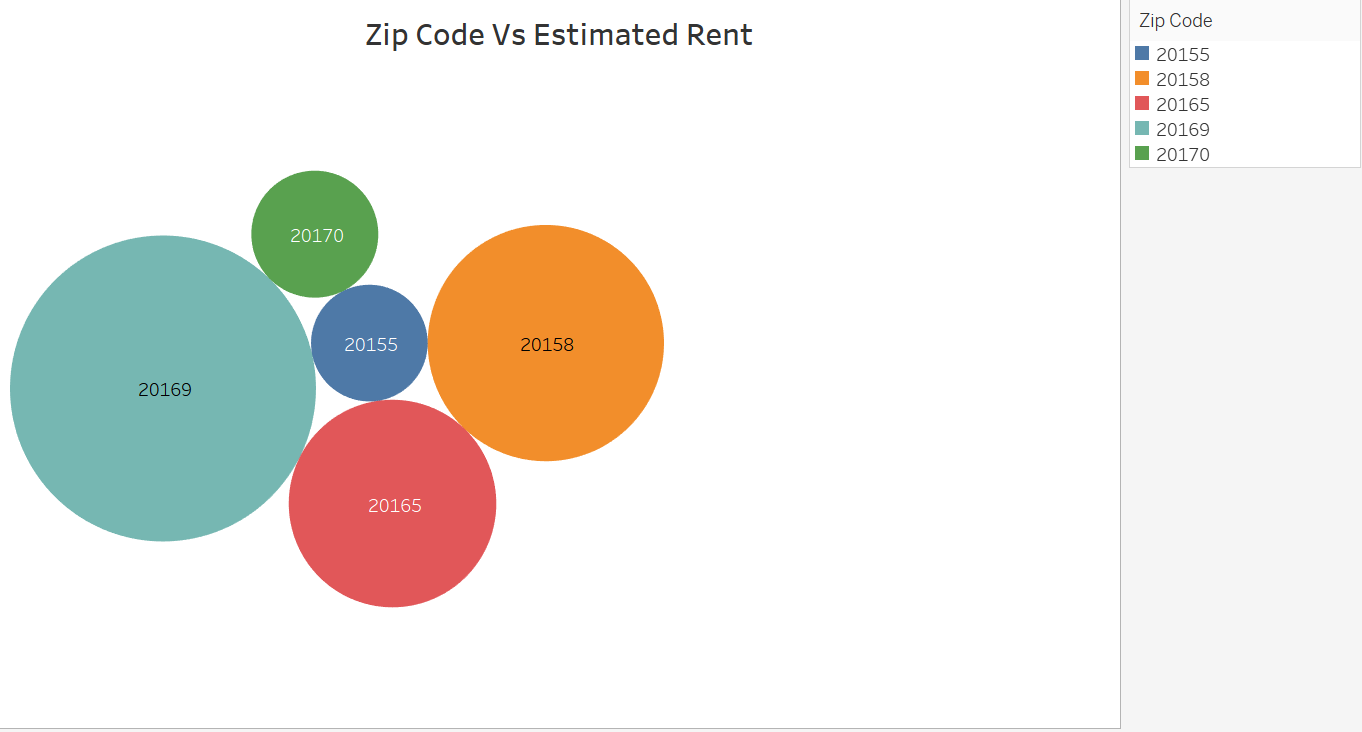
1. Clusters are made using just four important attributes which are estimated price, Area Space, Year built and Bathrooms. This helps in characterizing the houses with similar characteristics together and ultimately results in achieving very consistent accuracies with less error rate.
2. Division of one large Data set into 3 small full house, townhouse and condo data sets. This technique is also used to optimize results.
3. Moreover, I am running two different models on each cluster separately of three different data sets which are Full House, Townhouse and Condo Datasets. This is a way to further improve accuracy.
4. **Visualize your model:-**
5. **Bathrooms Vs Estimated Rent:-**

This is the visualization chart of the predicted values. This clearly tells us that Houses with 5 washrooms would be having maximum rent in future. The darker the blue color, more would be the rent. The upper labels are the washrooms and lower labels are the average rent in those number of washrooms.



1. **Zip Code Vs Estimated Rent:-**

This visualization chart of Predicted Rents clearly gives us information on Zip Code with maximum average rents of houses. The bigger the circle, more would be the rent. The Zip Code is labelled on the chart.



* **References:-**

**Tools Used :** Excel, R-Studio, Tableau

**Programming Language :**  R programming Language

**Book :** R for every one (for learning R)

**Links : https://cran.r-project.org/web/packages/EMCluster/EMCluster.pdf**