Assignment A2: Estimation				
Student Name	Yash Mistry		Student No	218612723
Problem attempted	Complex Model 80-100%	Simple Model 40-79%	Student Id	Ydmistry@deakin.edu.au
Place "Yes" in one only	?	yes	*Do not attempt a complex model unless you can complete a simple model first!*	

Partial Submission	Exceptional	Very Good	Good	Acceptable	Improve	Unaccept.
Exec Problem	5	4	3	2	1	0
Data Exploration	10	8	6	4	2	0

Final Submission	Exceptional	Very Good	Good	Acceptable	Improve	Unaccept.
Exec Solution	5	4	3	2	1	0
Data Preparation	20	16	12	8	4	0
Model Development	20	16	12	8	4	0
Model Evaluation	20	16	12	8	4	0
Model Application	20	16	12	8	4	0

Brief Comments	Total
Read these notes These and the following notes are trying to help you! Read the rubric on how the report content is going to be assessed! Your partial submission may not be perfect but has to reflect a genuine effort. We expect your partial submission sections to be improved for the final submission. We will not look at your partial submission until we mark the final submission. We will assess the final submission and its mark stands. However, we will deduct marks if the quality of the partial submission is poor. Do not attempt a complex model unless you can complete a simple model first! If you cannot formulate a complex problem, you will not get extra points for other complex criteria. Use the font already used in the template, i.e. Arial 10 (and not MyTiniestFont 2). If any submission aspects could only be determined by running the process, the marks will be severely reduced. Note: If it is not in this report, it does not exist and does not get marked! So, we will not check your RapidMiner scripts to check anything that was missing from the report. Any part which carries points but is missing in the report gets zero marks. We expect consistency between the report and RapidMiner scripts, so... Note: Anything reported that cannot be substantiated by RapidMiner scripts will be marked as zero. It means that we will check the RapidMiner scripts when in doubt or even just curious.	0 to 100

Executive problem statement (one page)

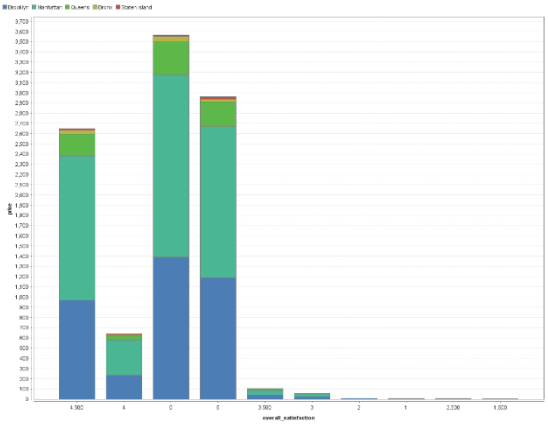
Aim
To figure out if there is any relation in overall experience and price in the Airbnb Database.

Simple Model
Airbnb approached us with their database to figure out some of their questions and wanted some insights with the help of database. In this analysis we must check if there is any trend related to the attribute's overall satisfaction and price of the Airbnb listings.

Airbnb has listed across 882000 listing in their database. After filtering and making sure of no duplicate value is present, we analyse the data and figure out if there is any relation between price of the listing and the satisfaction value. The data consists of different value and attributes. Attributes such as room id, geo location, min nights to stay, the number of occupants allowed etc. the data is quite large so for the prediction purposes we will have to rationalise the data into a smaller sample size data and get through with the ans. Airbnb listings have a satisfactory attribute which will help us in analysing the data more insightful.

The TOMSLEE data consists of all the binomial, polynomial and even date type of data to use from. In our business analytics solutioning we have to select and use the attributes we seem to be fit for solutioning.

To check if any relation is present regarding the price the listings and the satisfactory rating provided by the occupants. There can be a lot of factors that can be playing a role in overall satisfaction related data analysis.



From the different type of image and illustrations, we can see that there are a lot of missing information in the database. To get corrected we can remove any duplicate, or even remove all the value with no overall satisfaction values. In future we would try to put a specific value in the database for a much better prediction in the database.

In the data base we can even see a lot of missing value and a lot of overall satisfaction rating of 5. Rating of 3 and lower are very scarce and even very less to even predict. From the data we can see that there seems of be a lot of data that is missing. A lot of the attributes that are missing are overall_satisfation and even reviews. From the data analysis we can even figure out that price and overall_satisfaction have a trend which result in a better rating of the listing overall. the data analysis showed me that the less the min stay of a listing and the better pricing of the Airbnb listing is a very good factor which plays in have a overall good satisfaction rating. The neighbourhood attributes played some role in the data analysis too. If suburb such as Staten island has a listing which is very expensive might not get any overall rating whatsoever. But if a suburb such as Brooklyn has a listing which has a good affordable prize will have a good rating, review, and a better overall satisfaction.

Data preparation (one page)

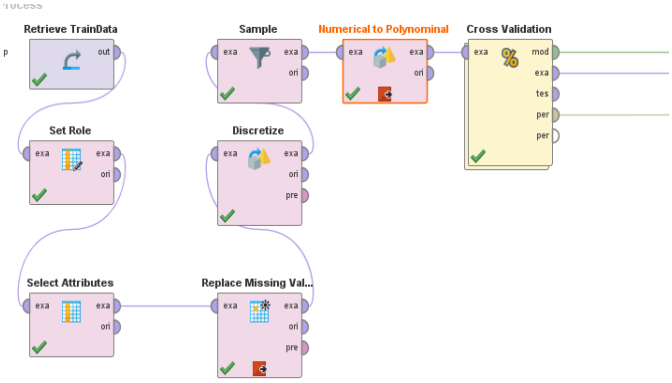
Aim
To demonstrate my understanding of the data analysis and my report of the Airbnb data.

Simple Model
In the part of data exploration, we have to check all the proper data exploration and data analysis part of the finding. We first need to label and use specific attributes into prediction.

Name	Type	Missing	Statistics
Label overall_satisfaction	Real	250359	Min 0 Max 5
room_id	Integer	0	Min 105 Max 17732388
host_id	Integer	280	Min 43 Max 120885348

As we can see from the following figure, there are a lot of missing value in the attributes such as review and even overall satisfaction. To solve that we have/ can replace the missing value with zero or even impute some random and appropriate values. We could also eliminate all the attributes which are essential but would not be correctly predicted. So just to solve the problem with as much data as possible we can impute a number which will not alter our data and its prediction.

From the following above illustrations, we can see the data analysis I have conducted. In my analysis I have taken the overall_satisfaction attributes as my label. From my analysis we can see that price and even min stay can be said to be playing a very important role in the overall satisfactory attributes. Price is another big attribute which plays a big role in it.



Row No.	overall_sati...	room_id	reviews	latitude	longitude	price	borough	neighborhood
1	5	9361	4	40.684	-73.964	low	Brooklyn	Clinton Hill
2	4	15796	19	40.682	-73.959	low	Brooklyn	Clinton Hill
3	5	18152	55	40.770	-73.951	very low	Manhattan	Upper East S...
4	5	22589	23	40.778	-73.985	medium	Manhattan	Upper West ...
5	5	44973	40	40.707	-73.953	low	Brooklyn	Williamsburg
6	1	45940	0	40.716	-73.959	low	Brooklyn	Williamsburg
7	5	60680	18	40.727	-73.980	low	Manhattan	East Village
8	5	64982	19	40.730	-73.982	medium	Manhattan	East Village
9	5	71010	2	40.835	-73.939	medium	Manhattan	Washington ...
10	5	72815	73	40.716	-73.996	high	Manhattan	Chinatown
11	5	82946	15	40.785	-73.979	very low	Manhattan	Upper West ...
12	5	83378	32	40.745	-74.002	medium high	Manhattan	Chelsea
13	5	83760	41	40.720	-73.997	low	Manhattan	Little Italy
14	5	84744	39	40.712	-73.965	low	Brooklyn	Williamsburg
15	1	85429	0	40.686	-73.959	very low	Brooklyn	Bedford-Stuyv...

The following illustration shows us the accuracy and even the kappa of our data analysis. We have a respectable accuracy of 88.96 and kappa just shy of 0.79. from this we can visualize all the attributes and even figure out the result we would like to seek. Which is does the price and overall satisfaction have any trend.

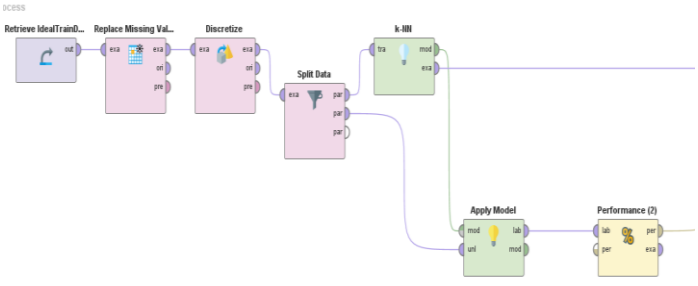
kappa: 0.783 +/- 0.007 (micro average: 0.783)

	true 1	true 2	true 3	true 4	true 5	tr
pred. 1	3454	2	7	31	135	0
pred. 2	0	0	0	1	0	0
pred. 3	1	0	0	1	8	0

accuracy: 88.76% +/- 0.36% (micro average: 88.76%)

	true 1	true 2	true 3	true 4
pred. 1	3454	2	7	31
pred. 2	0	0	0	1
pred. 3	1	0	0	1
pred. 4	12	3	2	21

Extension:

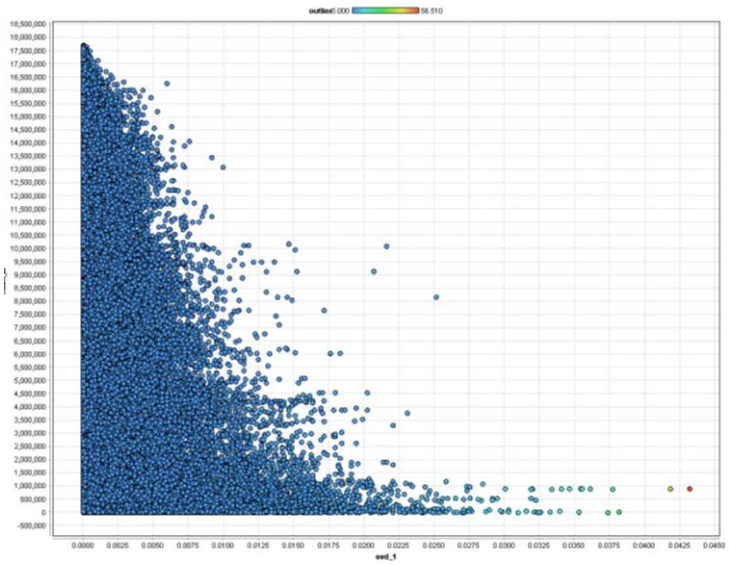


With a lot of more process and refining we can see that the best model to use would be K-NN or Gradient which will depend on the sample size too. The K-NN model will take up more time than gradient trees.

Executive solution statement (one page)

Aim

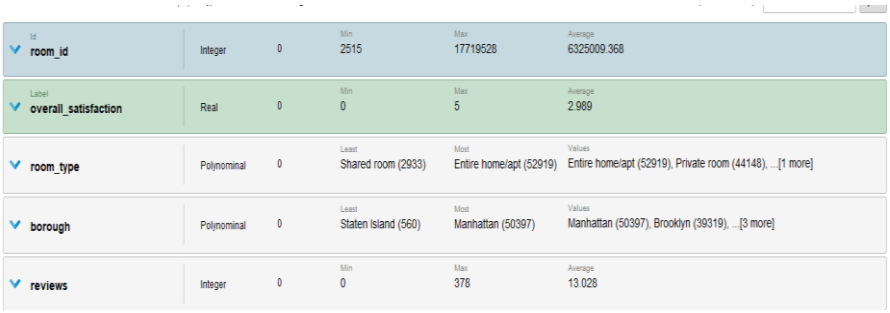
What kind of properties receive which kind of reviews in our given data set



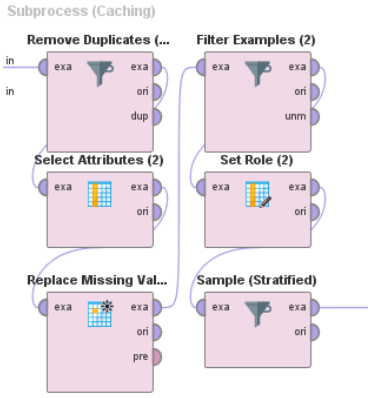
Simple Model

In this part of our business model problem, we must find the different review to the different types of properties. As we can see from the illustration on the left, their seem to be some outlier on the far right below in green and yellow. We have to even check for different room types, which is Shared room, Private Room or Entire House.

First we remove all the duplicates from the dataset. After that we figure out what are the attributes that can play a big role in satisfaction related to different types of house types/room types. For doing that we can use the subprocess parameter. From the shown model process on the left we can see that we have used K-NN global anomaly (extension to be gotten from the marketplace). And by using the SVD process, we can see that the outlier. When we get the outlier we can



room_id	Integer	0	Min 2515	Max 17719528	Average 6325009.368
overall_satisfaction	Real	0	Min 0	Max 5	Average 2.989
room_type	Polynominal	0	Least Shared room (2933)	Most Entire home/apt (52919)	Values Entire home/apt (52919), Private room (44148), ...[1 more]
borough	Polynominal	0	Least Staten Island (560)	Most Manhattan (50397)	Values Manhattan (50397), Brooklyn (39319), ...[3 more]
reviews	Integer	0	Min 0	Max 378	Average 13.028



From the adjacent figure we can see that we must remove the duplicates and select every attribute that can influence the satisfaction meter for different room type. For dealing with the missing values and even get a sample data for a better data processing and no hiccups during the performance.

After doing all the pre-processing we can check our report with cross referencing and even make a model with different cluster/ data processing models to get the best result for our business problems.

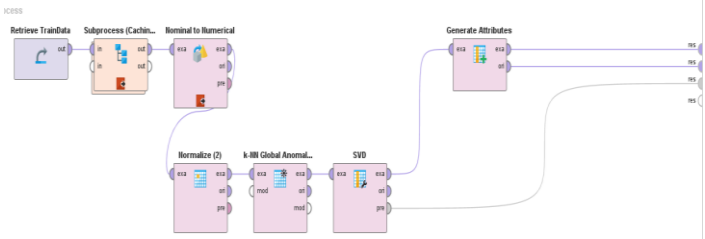
Data Preparation and Exploration (first page of a two page section)

Aim

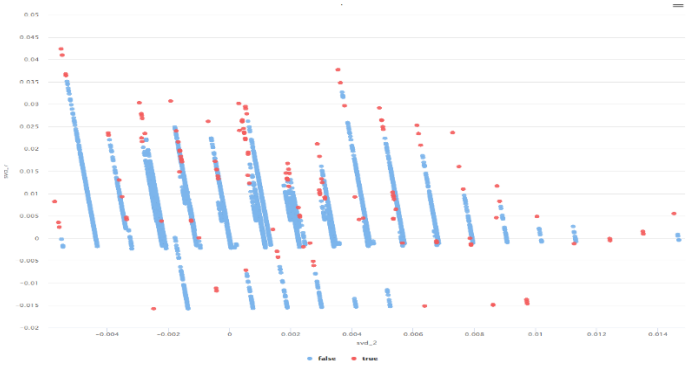
The Aim on this data exploration is to find the best processes into finding all the outlier and even check if the outlier are true or false.

Simple Model

Using Clustering models and SVD method to figure out all the outlier and their score into getting a better result to find the business problem solution. We even must normalize all the value that are not suitable or work with different models. Such as the borough and room type are polynomial attributes. So, to change that we must use the nominal to polynomial value parameter. From the data, we cleared it of all the duplicate values and even sort out all the attributes we will really need to find the business-related solutions.

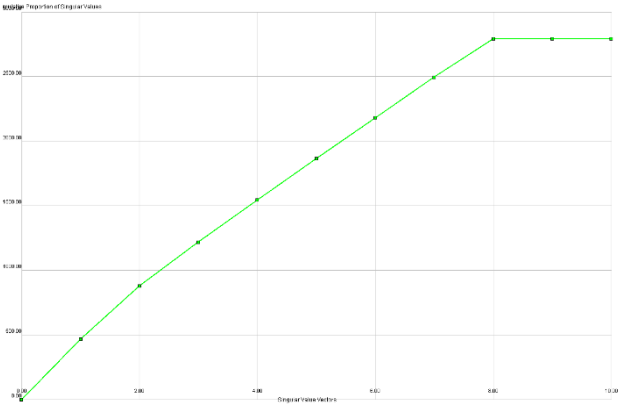


At first we just use Knn cluster and even SVD directly into our business data. To see any outlier to check for any relation between the attributes ew have sleected. We have discretise the borugh attribute and even room type to check the satisfaction level in the dat aregarding the room type and even the borugh.



The figure shown below are the result of the singular Value Data we used to find the outlier. From the image shown in the left we can see the outlier in red colour and the data in blue. To do that we have used generate attribute with the function of if any outlier in the clustering will be reflected in different colour.

Component	Singular Value	Proportion of Singular Values	Cumulative Singular Values	Cumulative Proportion of Singular Values
SVD 1	466.966	0.167	466.966	0.167
SVD 2	413.265	0.148	880.232	0.315
SVD 3	339.751	0.122	1219.983	0.437
SVD 4	326.571	0.117	1546.554	0.554
SVD 5	319.663	0.114	1866.216	0.668
SVD 6	316.680	0.113	2182.897	0.781
SVD 7	314.761	0.113	2497.658	0.894
SVD 8	295.837	0.106	2793.495	1.000
SVD 9	0.000	0.000	2793.495	1.000
SVD 10	0.000	0.000	2793.495	1.000

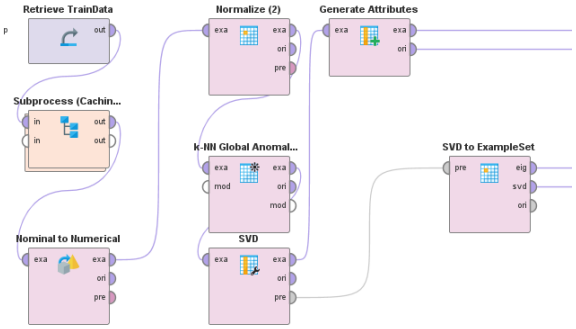


Furthermore, we must upgrade the model for testing it in the training and testing data to get the result for our business problems. We will develop the data model into a much better and much higher efficient result. While we can use the same data set to get much better result we can even remove or even add some more attributes to check for any more relation between each other.

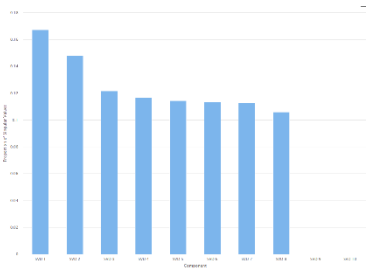
Data Preparation and Exploration (Second page of a two-page section)

Continuation of the previous page.

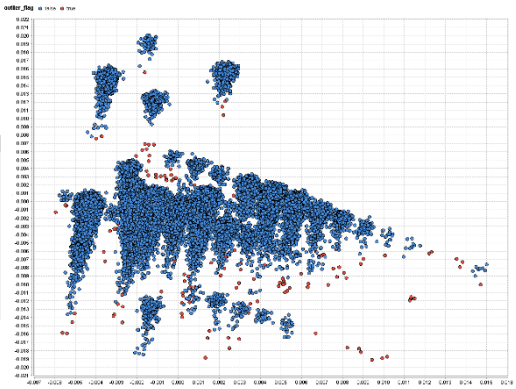
After doing all the process, we further developed our process into finding a better result for the outlier and mainly the business problems we have in our hands. The first outlier model was sufficient but needed more refining. So for second model we used the old one and refined it more. The below image show the stats of the outlier and the Overall_satisfaction attribute. We even have to normalize all the value that are not suitable or work with different models. Such as the borough and room type are polynomial attributes. So, to change that we have to use the nominal to polynomial value parameter.

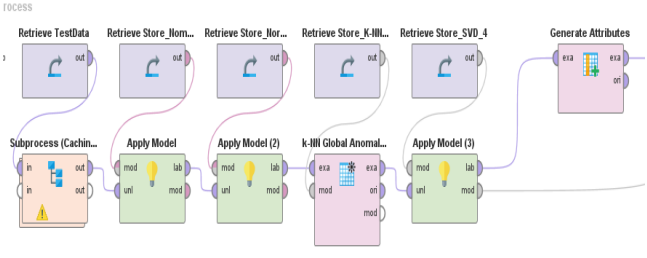


The refined output from the model is shown below the Singular Value and Cumulative Singular value is recorded usingthe SVD TO EXAMPLESET parameter and used for a better refined result. Image shown below can be used to check the outlier. As you can see the outlier are flaged and shown in red colour while all the usefull value is in red. From our model we can see that most of the enitre house and even private rooms recevive a lot of satisfaction level. Shared room on the other hand does not have a overall satisfaction level. Something about sharing the room with someoen else makes the experience of the property go down.



Row No.	Component	Singular Value	Proportion of Singular Va...	Cumulative Singular Values	Cumulative Proportion of Singul...
1	SVD 1	466.966	0.167	466.966	0.167
2	SVD 2	413.265	0.148	880.232	0.315
3	SVD 3	339.751	0.122	1219.983	0.437
4	SVD 4	326.571	0.117	1546.554	0.554
5	SVD 5	319.663	0.114	1866.216	0.668
6	SVD 6	316.680	0.113	2182.897	0.781
7	SVD 7	314.761	0.113	2497.658	0.894
8	SVD 8	295.837	0.106	2793.495	1.000
9	SVD 9	0.000	0.000	2793.495	1.000
10	SVD 10	0.000	0.000	2793.495	1





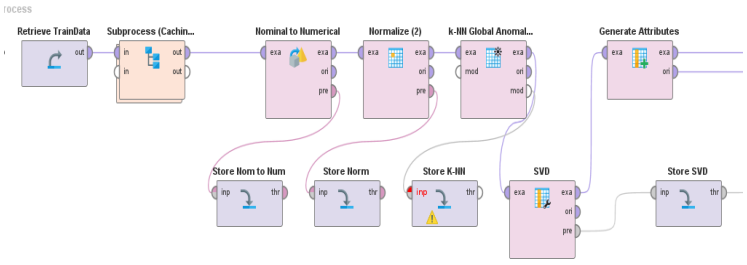


Figure 6.1 TrainData

Fig6.2 TestData

We then applied the train model data process (Fig6.1) and applied it to the test data (6.2) and we got the same result. After carefully saving all the model procedure and steps into detailed repositories, we tested the same model used in training data to test data. We got the same result which means that our model is of a good and reliable process and can somewhat be used for clustering predictions and variations.

Model Development (one page)

Aim

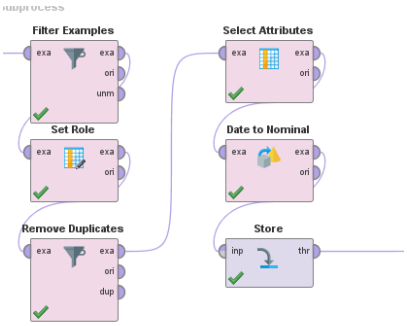
If any rental Is new in our data listing, we must figure out the satisfaction level of any new listing of any properties.

Simple Model

To figure out a way to find the best value of satisfaction for any new listing we have to compare 2 or even more models and find the best model to use for getting a better result and the best model for prediction. In this business. In my business model we must make sure that all the new date attribute is used in this for the solutioning. After using the date attributes, we must assess the data and find out more of the data to be used.

Row No.	room_id	last_modified	room_type	borough	neighborhood	reviews	overall_sati...
1	105	12/05/2017 08:56	Private room	Manhattan	Hell's Kitchen	39	5
2	2515	12/05/2017 18:11	Private room	Manhattan	Harlem	95	4.500
3	3046	12/05/2017 19:47	Entire home/apt	Manhattan	Harlem	10	4.500
4	3101	12/05/2017 00:57	Private room	Manhattan	Harlem	11	4.500
5	3102	12/05/2017 01:03	Private room	Manhattan	Harlem	23	5
6	3153	12/05/2017 05:13	Private room	Queens	Astoria	33	5
7	4573	12/05/2017 19:01	Entire home/apt	Brooklyn	Crown Heights	3	4
8	5030	12/05/2017 14:19	Entire home/apt	Manhattan	Upper East S...	90	4.500
9	5044	12/05/2017 15:32	Entire home/apt	Manhattan	East Village	17	5
10	5053	12/05/2017 16:16	Entire home/apt	Manhattan	Upper East S...	129	4.500
11	5054	12/05/2017 16:21	Entire home/apt	Manhattan	Hell's Kitchen	44	4.500
12	5079	12/05/2017 18:31	Entire home/apt	Manhattan	Harlem	149	5
13	5099	12/05/2017 20:14	Entire home/apt	Manhattan	Murray Hill	31	4.500
14	5121	11/05/2017 09:11	Private room	Brooklyn	Bedford-Stuyv...	36	4.500
15	5135	11/05/2017 23:45	Private room	Manhattan	Harlem	176	4.500
16	5136	11/05/2017 23:50	Entire home/apt	Brooklyn	Sunset Park	1	5

From the above illustration we can see the data type we have to use for the review for all the new listings. We've to use the date attribute and make it into text or polynomial for the model to process.



These are all the subprocess that we have used for the efficiency of the process. We have elimminated any duplicates and even all the values which are not present to work on. Keeping the room_id as id and Dat (last modified)as label we can work out the process into getting the desired ans.

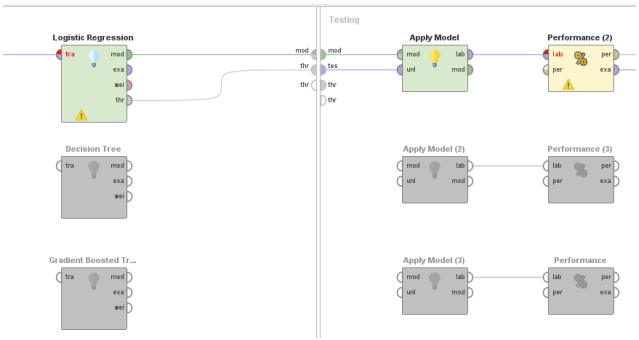
Model Evaluation and Optimisation (first page of a two-page section)

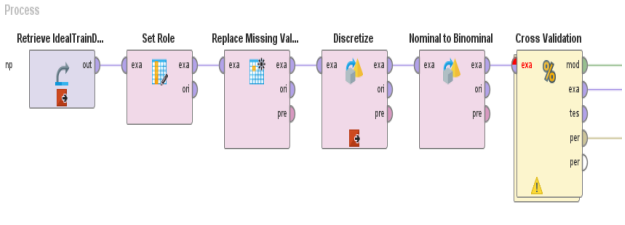
Aim

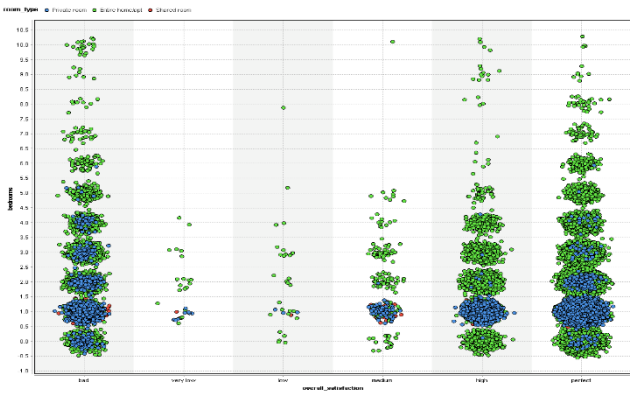
In this section of the report , we will look into different result that we can get from various prediction models and compare each of the models to each other for a better and sufficient result for our own data. In this section of model building we have used, gradient boosted trees, decision tress and even logistic regression.

Simple Model

In this part of the report we will see the model building process for RMEA and MAE comparision I have used Decision tree, gradient boosted trees and even logistic regression for comparision. With that build in the cross validation parameter. We can see the model building and its performance on the results tab for a bette rlook at the business problem solution.







While Gradient boosted trees has given the best result for the model building process, we can see the comparison from various model building parameters and make use of the best model for our data.

From our result we have seen that K-NN and decision tree were not able to give the best result when it comes to see the satisfaction review to any new listings. Gradient boosted trees on the other hand were able to give a better result when compared to other model building parameters.

Model Application (one page)

Aim
In this section we will display all the result and compare it to each other and maximize the data result for our business problem solution.

Simple Model
Apply the best model to new data and investigate results. All results must be analysed, interpreted, and reported. A statement is to be included with justification as to what degree the model advice can be trusted.

After checking with the training data our new model to check whether the model is giving us the suitable result , we can apply the same for the test data as well. From applying the train model to the test data we see some similar results as well. Its suitable to say that the train data to find the satisfaction level in any new listing found in the Tomslee NYC data can be predicted. From the data model we see most of the new listing have a average rating altogether. It can be in creased by having a lot of review and if it is the popular suburbs such as Brooklyn and quuens.

We can see that most of the better result for finding the result is from gradient boosted trees. From the result we found we can see that all the new listings are on the average satisfaction level with very less review. For the model building we see that review play a very important role in overall satisfaction. Bedrooms more than four does not have a good overall score more than three. The borough attributes and even the overall_satisfaction attribute plays a very important role in the listings and its review.