MIS772 – Predictive Analytics

T1 2023

Assignment 2 – Individual
Student name: Akilvish Paliwal
Student number: s222116915

Executive summary (1 page)

Greetings to the senior business manager of Airbnb! This report focuses on developing processes and models to solve business problems and derive insights from data related to properties and reviews from the customers. The three business problems targeted in this report are, how co related the sentiments of the customers are to the rating they provide for the property, which properties are located outside Melbourne CBD and what will their review score rating would be and lastly the top 3 most frequently co-occurring review score attributes that drop below zero. The first requirement posed by business was a process model in rapid miner to understand if review scores rating, i.e., overall customers review score of property, is related to the sentiments of the customers. Business is posed with a challenge to verify the connection between good and bad reviews with the actual review rating that customers give for the property. Identifying the sentiments of the customers (positive or negative) through the comments they gave for the property, and the overall customers review score of property paved a way to understand that they are positively related to each other. The relationship is not very strong but there is a significant relation. As the sentiments of the customers are on the positive side, better will be the overall customers review score of the property. The approach behind this problem was to look for an overall sentiment of the customer that is positive minus negative and comparing it with review scores rating. Business also requested for a predictive model to estimate the review score rating of all the properties located outside Melbourne CBD. The properties inside CBD have a lot of opportunities to be explored by the travellers but the properties outside CBD have dramatically less attention compared to those in CBD. Thus, the business wants to have a model that could actually predict how the customers would rate a property and help the business to analyse their next step and maybe change strategy to market it. Understanding the overall review score of a property opens up the doors for the business to either communicate with the owner to improve in certain areas or maybe help the owners to attract more tourists. For the underlying challenge, the given data is sufficient enough after appropriately making the data fit for the exploration. Using the relevant data, we could estimate review scores rating for seventy percent of the properties with a slight margin of doubt. We can also estimate the effect of each single variable on review scores rating if we make a unit increase in it, hence could help the business to pinpoint where the room for improvement is. As business is extremely concerned about the review scores of the properties, the last requirement was to identify the review scores variables that are most frequently co-occurring and dropping below zero. And as a specific requirement, business had posed to find out top 3 most frequently co-occurring review scores variables that are, (accuracy, checkin, cleanliness, communication, location, and value) of a property drops below 10. The research performed on the data to come up with a process to achieve that business goal confirmed that the communication, location and checkin are the variables. Airbnb clearly needs a business plan to target the specific review scores variables that are most frequently going below 10, thus on a common workaround, multiple properties could improve and thereby, improving customer satisfaction and customer retention. The reviews given by the customers are very important and cannot be neglected and to work on them, Airbnb clearly seeks out a definite plan of action. As per the analysis, improvement in communication from the host with customers, and checkin experience needs to be worked upon. And as far as the location is concerned, owner cannot change the location of the house, but Airbnb can definitely tie up with other recreational activity companies or travel providers to collab and improve overall experience of the tourists. Lack of cheerful communication from the host and checkin process can be improved on the individual level of each owner and Airbnb can support the owners in the process by providing guidelines and tips on how other property owners are doing it and what each specific type of guest expects. In the light of the report, the recommendations are to improve practices from the host and improve fun activities, travel options and facilities for the properties outside CBD alongside leveraging benefits from the owners that are on the top of the list and provide an authentic and unforgettable travel experience to the customers. This will definitely improve the customer retention by reducing customer churn and attract new customers and boost the travel market for the long run.

Data exploration, pattern discovery, and preparation (2 pages)

We extracted a listings sample of approximately 1000 properties and a reviews sample of approximately 100,000 reviews of those respective properties for a period of April 2011 to Feb 2021 from http://insideairbnb.com/get-the-data.html.

VARIABLES: The listings sample has 25 attributes. Id is an identifier of the property. Other *nominal attributes* including property name, property description, property neighbourhood review, property neighbourhood, property type, room type seem to be of less analytical significance. The *numerical attributes* like number of properties hosted by an owner, accommodates, bedrooms, price per night, reviews per month, review scores rating, review scores accuracy, review scores cleanliness, review scores checkin, review scores communication, review scores location, review scores value are highly important to our modelling. Also, latitude and longitude variables to locate the property. The *review_scores* variables that show the average review ratings of factors like cleanliness, communication, location, value, checkin and accuracy seem to have 36 records missing each. The reviews sample has 4 attributes and a lot of missing values. The listing id is similar to id of the property and review id is the identifier for the review. A date variable "date of review" and a nominal variable "comment" are important.

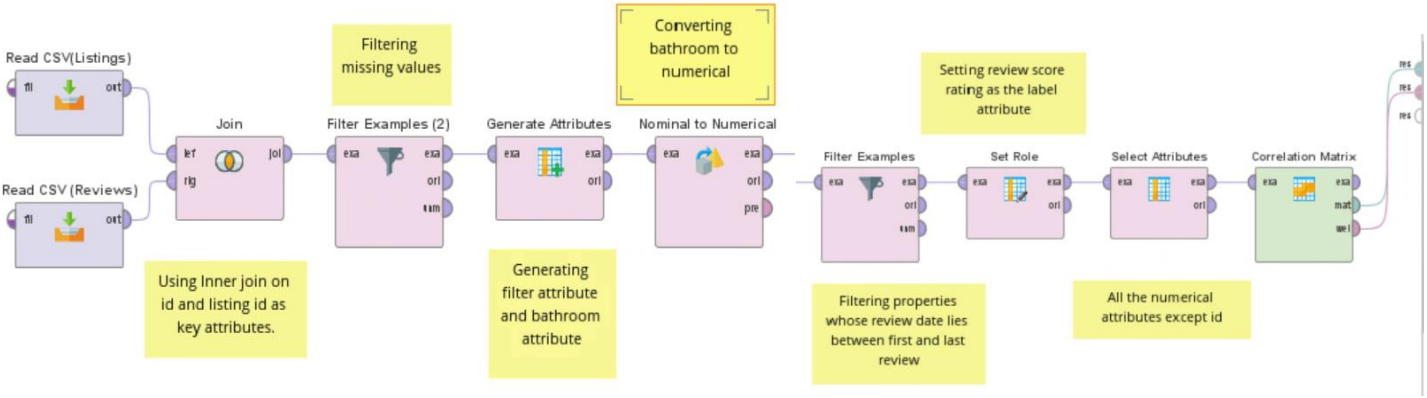
Correlation: There is a high correlation of 0.72 between bedrooms and accommodates which is obvious, more the number of bedrooms, more people can be accommodated. All other attributes have low to moderate correlation with each other. And checking the correlation of variables with the label attribute (review_scores_rating), review_scores_location, latitude, host total listing count, bathroom and longitude show high correlation of more than 0.8. Reviews per night show 1 correlation which looks a little odd.

Attribut...	bathroom	host_to...	latitude	longitude	accom...	bedroo...	price_p...	reviews...	review_...	review_...	revie
bathroom	1	0.001	0.009	0.001	0.030	0.075	0.012	0.004	0.006	0.007	0.001
host_tot...	0.001	1	0.001	0.008	0.023	0.022	0.003	0.015	0.000	0.001	0.012
latitude	0.009	0.001	1	0.045	0.003	0.008	0.006	0.001	0.000	0.004	0.000
longitude	0.001	0.008	0.045	1	0.005	0.018	0.069	0.000	0.018	0.022	0.015
accomm...	0.030	0.023	0.003	0.005	1	0.723	0.292	0.001	0.018	0.048	0.023
bedrooms	0.075	0.022	0.008	0.018	0.723	1	0.332	0.004	0.002	0.011	0.005
price_pe...	0.012	0.003	0.006	0.069	0.292	0.332	1	0.003	0.000	0.001	0.001
reviews_...	0.004	0.015	0.001	0.000	0.001	0.004	0.003	1	0.001	0.001	0.000
review_s...	0.006	0.000	0.000	0.018	0.018	0.002	0.000	0.001	1	0.370	0.158
review_s...	0.007	0.001	0.004	0.022	0.048	0.011	0.001	0.001	0.370	1	0.131
review_s...	0.001	0.012	0.000	0.015	0.023	0.005	0.001	0.000	0.158	0.131	1
review_s...	0.002	0.001	0.000	0.007	0.003	0.000	0.001	0.000	0.184	0.121	0.254
review_s...	0.003	0.000	0.001	0.001	0.001	0.001	0.005	0.002	0.013	0.009	0.000

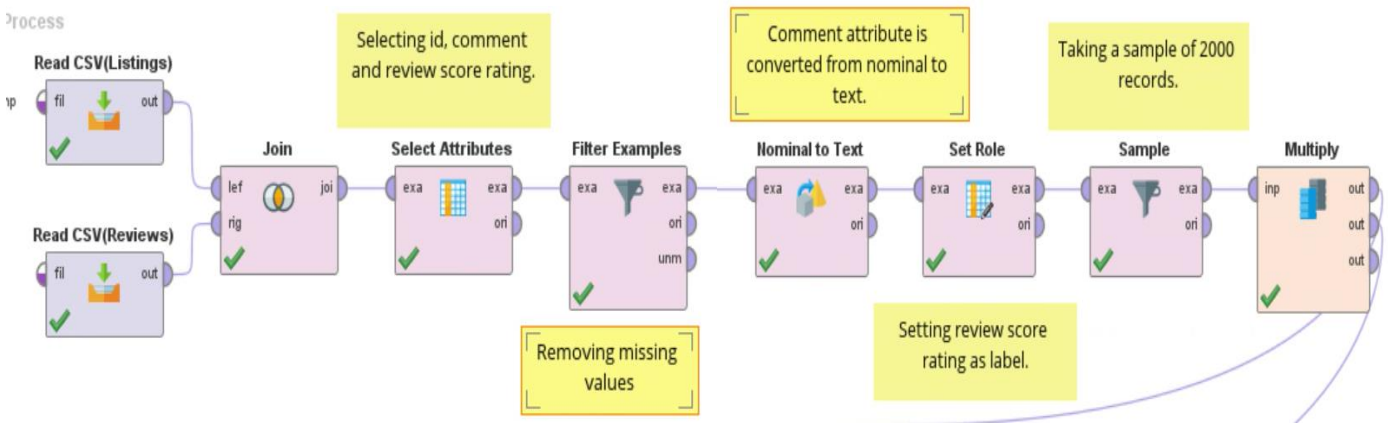
attribute	weight
reviews_per_month	1
review_scores_location	0.989
latitude	0.966
host_total_listings_count	0.956
bathroom	0.893
longitude	0.844
review_scores_communication	0.416
price_per_night	0.409
review_scores_checkin	0.392
review_scores_value	0.246
review_scores_cleanliness	0.207
review_scores_accuracy	0.088
accommodates	0.015

DATA CLEANING AND TRANSFORMATION:

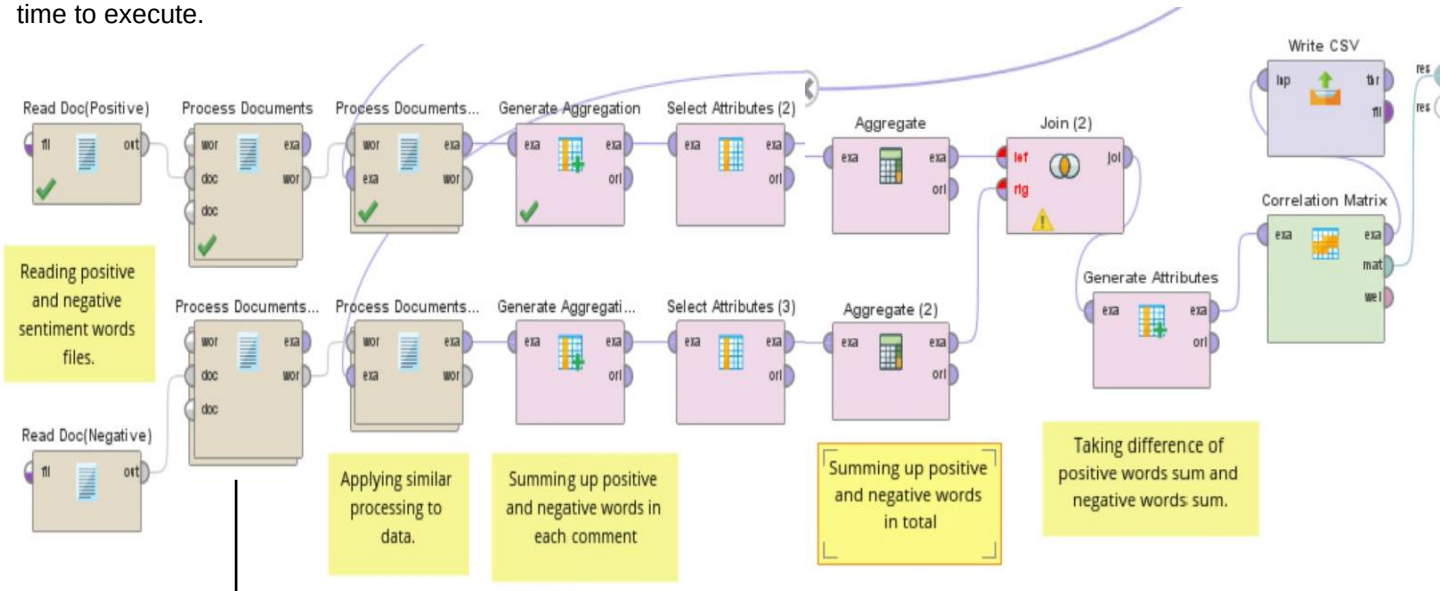
After joining the two CSVs using inner join on id and listing id, we removed missing values from the data. The variable date of review should lie between date of first review and last review, which is true for all the reviews. The attribute "bathroom text" shows number of bathrooms but is nominal and after converting it to numerical, it is evident that mostly properties have 1 bath. So, assigning 1 for properties with 1 bath and 0 for others using unique integers.



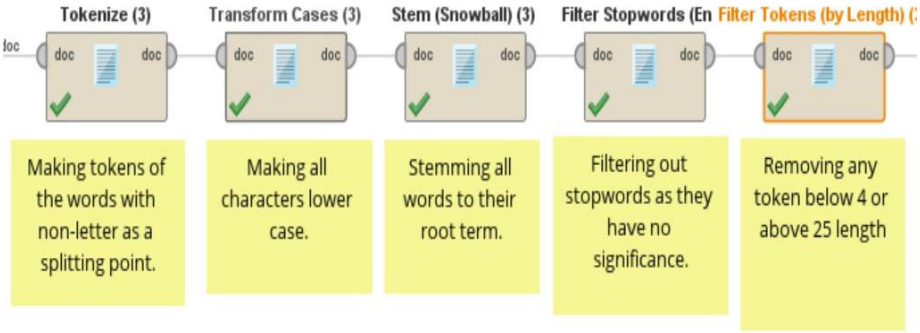
Sentiment Analysis:



For performing the sentiment analysis, after joining the two sample data files, we select id, comment, and review_scores_rating attributes. We choose them because we need id to identify records, we use comment attribute to check positive or negative sentiment and compare it with review_scores_rating. We need to convert comment to text for further process. After setting review score rating as label, we take a sample of 2000 records so that process takes less time to execute.



All the process document operators have similar process flow as fig. 1.1.



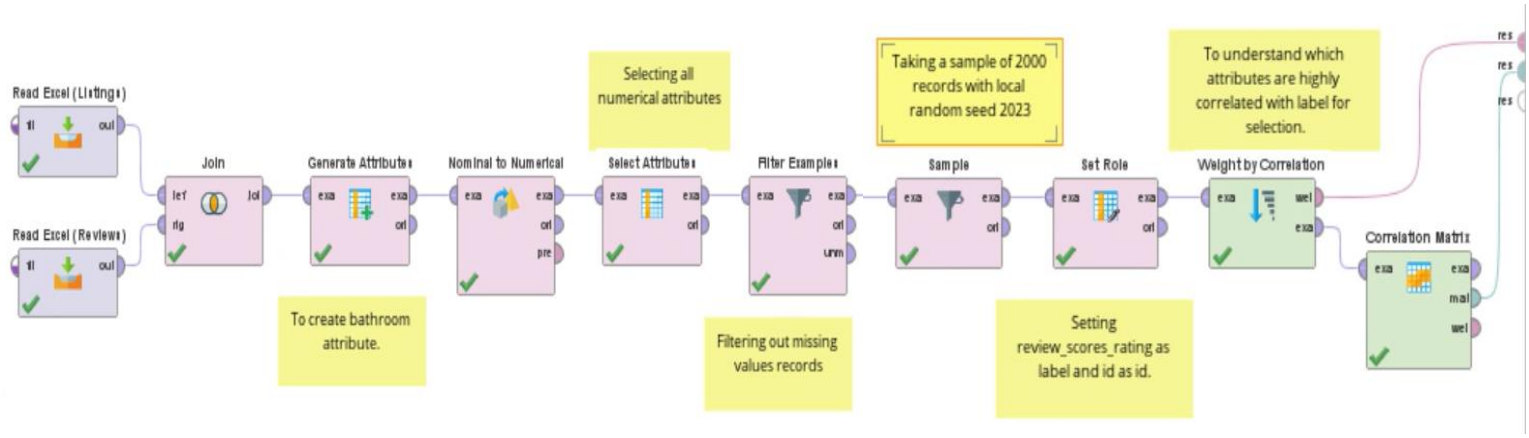
We tokenize the words from positive and negative words dictionary, change them to lower case, stem them to their root word, filter out the stop words and keep words in word length of 4 to 25 words. After processing dictionaries, we process dataset similarly and check word count of positive and negative words in the whole document. Taking the difference of positive and negative words would give us the raw sentiment score.

Fig. 1.1

After looking at the results of correlation matrix, we inferred that there is a significant positive correlation of 0.217 between raw sentiment score and review_scores_rating. This means, higher the raw sentiment, higher the review_scores_rating. Correlation is not strong but lies between weak to moderate range. We save the output of the process in a csv file for future use.

Predictive modelling (2 pages)

Attribute Selection:



For the selection of attributes for modelling, we check the correlation of attributes with the label attribute that is review_scores_rating and across each other as well. We create the attribute for number of baths as per transformation, select all numerical attributes, eliminate missing values, and take a sample of 2000 records with local random seed of 2023.

Looking at the results, we inferred that review scores cleanliness, review scores accuracy, review scores value, review scores checkin and review scores communication have moderate to high correlation with the label and review scores location, longitude, accommodates have low correlation. All other variables not so significant.

And bedrooms and accommodates have very high correlation (0.85), so, we can remove bedrooms from selection. Accommodates also eliminates price per night (0.56). There is a moderate correlation between review scores accuracy, review scores cleanliness, review scores checkin, review scores communication, and review scores value.

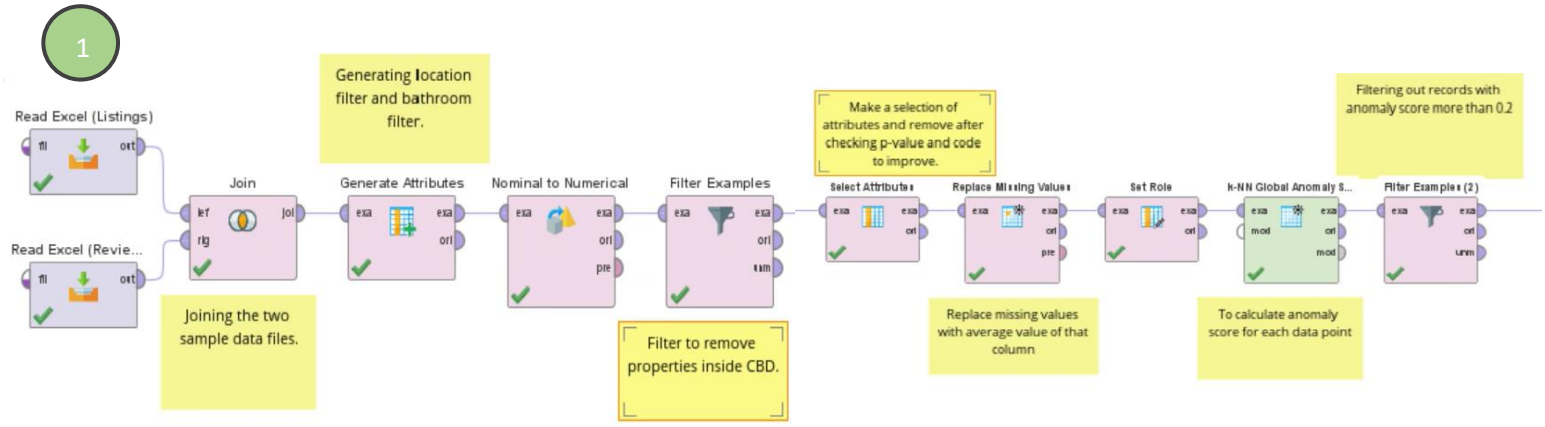
Attributes	accommodat...	price_per_night	review_scores_accuracy	review_scores_cleanliness
bathroom	-0.178	-0.073	-0.037	-0.057
host_total_listings_count	0.090	-0.020	-0.125	-0.074
latitude	0.107	0.085	0.012	0.074
longitude	0.047	0.267	0.160	0.196
accommodates	1	0.564	-0.094	-0.215
bedrooms	0.855	0.580	-0.021	-0.083
price_per_night	0.564	1	0.040	0.003
reviews_per_month	-0.032	-0.015	0.020	0.031
review_scores_accuracy	-0.094	0.040	1	0.577
review_scores_cleanliness	-0.215	0.003	0.577	1
review_scores_checkin	-0.156	0.019	0.450	0.371
review_scores_communication	-0.012	0.044	0.510	0.342
review_scores_location	0.027	0.071	0.196	0.171

attribute	weight ↓
review_scores_cleanliness	0.683
review_scores_accuracy	0.664
review_scores_value	0.618
review_scores_checkin	0.517
review_scores_communication	0.451
review_scores_location	0.276
longitude	0.231
accommodates	0.197
host_total_listings_count	0.105
bedrooms	0.098
reviews_per_month	0.073
price_per_night	0.073
bathroom	0.069

Since the model is regression model we need to make sure that the model performance is free from multi-collinearity and thus use trial and error to select attributes.

Model Generation for predicting review_scores_rating: After joining the sample data files, we generate a “location” variable, if (longitude > 144.9 && longitude < 145.06 && latitude > -37.95 && latitude < -37.75,"Inside CBD", "Outside CBD") as per the given latitudes and longitudes and remove properties inside CBD. As per the data transformation, we select attributes as predictors and check their p-value and code to remove the attributes that are not significant for the model. Since, we have only 33,000 records after filtering outside CBD properties, we *replace missing values* with average of their columns. In this model, we *use k-NN global anomaly score* to identify outliers. We calculate residuals using *calculate residual* operators. And by observing the residual plot, we remove the records having anomaly score higher than 0.2, thus making the residual plot normal and fully utilising the k-NN global anomaly score operator for the model.

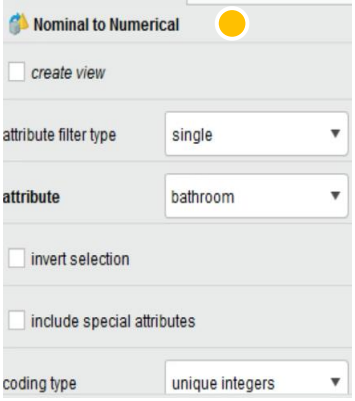
Rather than split validation, we use *cross validation* because it is a better technique to check the model performance. And for modelling, we use *linear regression model* since we are trying to estimate a numerical variable (review_scores_rating) with other numerical variables.

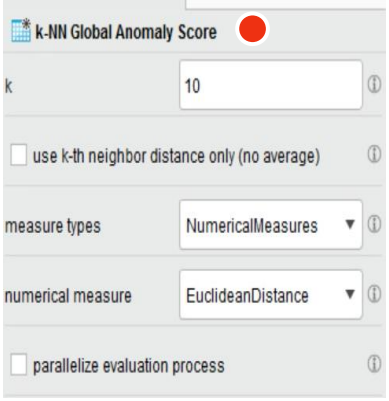


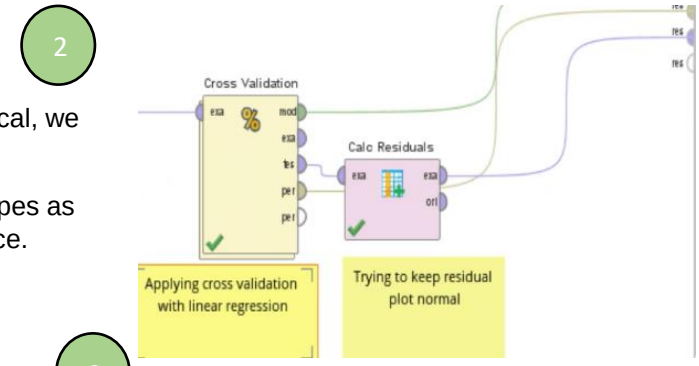
Important parameters:

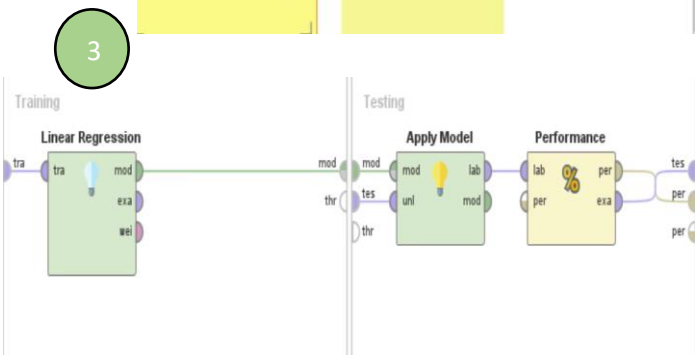
- For converting the newly generated attribute bathroom to numerical, we use *unique integers* as coding type.
- For k-NN global anomaly score operator, keep k=10, measure types as numerical measures and numerical measure as Euclidian distance.

For calculating residuals, used following expression, residuals = review_scores_rating-[prediction(review_scores_rating)].

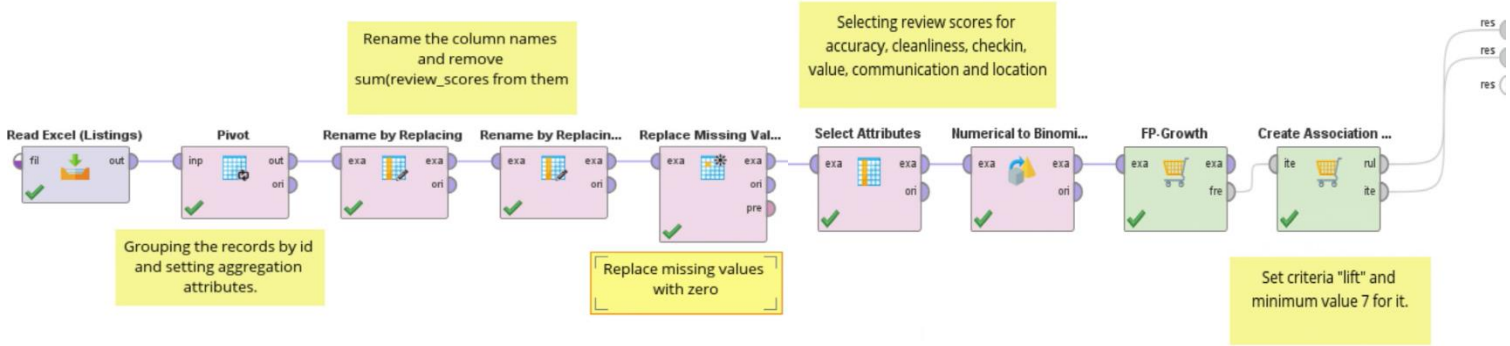








Process for top 3 most frequently co-occurring review score attributes that drop below 10:



In order to predict the top three most frequently co-occurring review score attributes that drop below 10, we use association. Using *pivot operator* to group the listings of the property as per the id and taking all the review scores variables as aggregation attributes (accuracy, checkin, cleanliness, communication, location, and value). Using *rename by replacing* to remove unnecessary terms from column headings. After *replacing missing* values with zero, selecting review scores, convert all attributes from *numerical to binominal* and *apply FP Growth association*. FP growth works best for any minimum support threshold (high or low) and works efficient for huge datasets with large number of unique items. And we use create association rules operator to create association rules with lift as the criterion and minimum value of 7.

Model evaluation and improvement (2 pages)

Predictive Model Evaluation:

Model	Squared Correlation	Root mean squared error
Model 1	0.691 +/- 0.014	1.357 +/- 0.014
Model 2	0.705 +/- 0.014	1.326 +/- 0.018
Model 3	0.709 +/- 0.018	1.234 +/- 0.016

For model 1, we developed a linear regression model with cross validation and k-NN global anomaly outlier detection. As per the model output, 69% of variation in review_scores_rating can be explained by the selected attributes/predictors.

RMSE, that is, the square root of the average squared difference between the predicted and actual values for this model is 1.358.

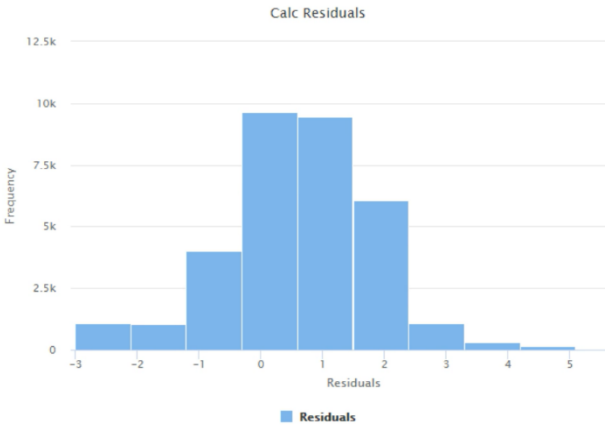


Fig. 3.1

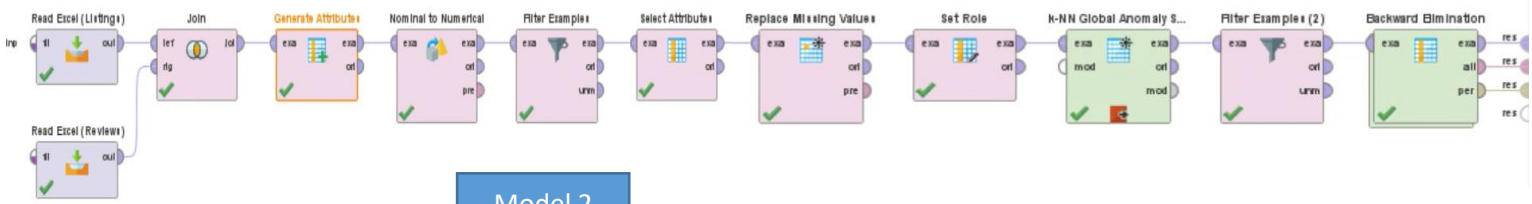
Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value
bathroom	-0.569	0.018	-0.108	0.971	-31.075	0
host_total_listi...	-0.061	0.003	-0.072	0.992	-22.147	0
longitude	0.699	0.032	0.080	0.947	21.971	0
accommodates	-0.253	0.004	-0.222	1.000	-61.286	0
price_per_night	0.013	0.000	0.361	1.000	93.156	0
review_scores...	1.691	0.044	0.160	0.581	38.100	0
review_scores...	1.876	0.023	0.307	0.763	81.822	0
review_scores...	3.238	0.058	0.188	0.883	55.810	0
review_scores...	2.136	0.024	0.338	0.815	90.693	0
(Intercept)	-93.030	4.603	?	?	-20.211	0

Fig. 3.2

Figure 3.1 is the plot of residuals for the model 1 and the fig. 3.2 is the set of attributes with their coefficients for model 1.

Interpretation: Keeping the plot of residuals as normal as possible with mean at zero indicates that the residuals are randomly scattered around zero and have a symmetric distribution and thus suggests model's assumptions are met and it has good fit for the data.

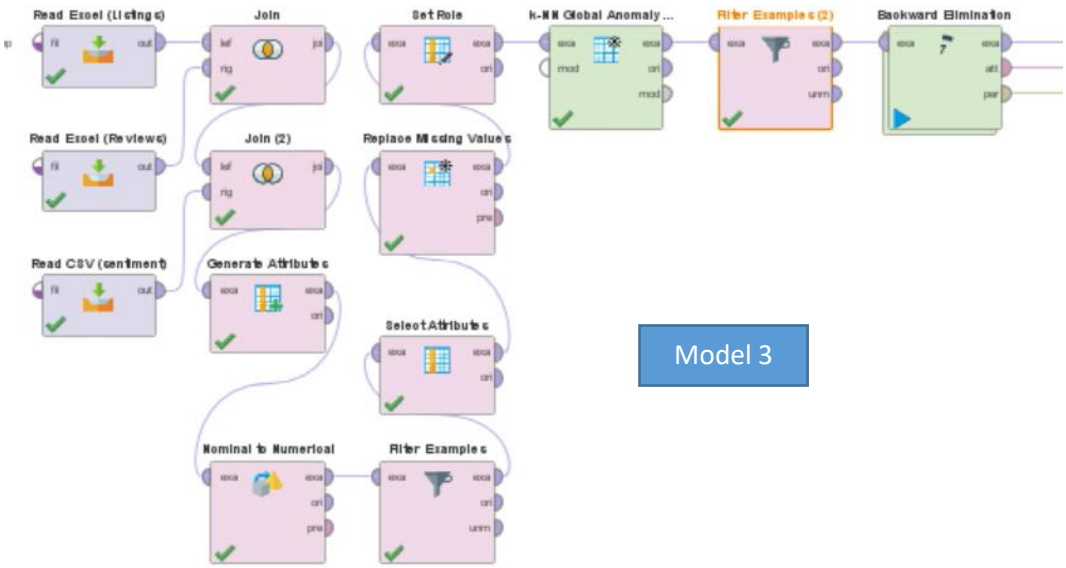
Attributes with p-value less than alpha means they are significant for the model. Significance of coefficients lie in the fact that for example, every one unit increase in longitude, the review_scores_rating increases by 0.699. Higher the tolerance, more independently the predictor contributes to the model.



Model 2

For model 2, we used the *backward elimination* as a feature selection technique for rather than manual selecting the attributes, we let machine do it for us. Backward elimination technique selects all attributes and then removes them one by one depending on their impact on the model performance. Model performance improved as 70.5% of variation in review_scores_rating can be explained by the selected attributes/predictors with RMSE of 1.32.

Attributes that are to be taken for the model are bathroom, host total listings count, latitude, longitude, accommodates, price per night, reviews per month, review scores accuracy, review scores cleanliness, review scores checkin, review scores communication, review scores location, review scores value.
We removed any record having outlier greater than 0.2 for this model as well as per previous model, since the outlier values went from 0 to 13.6 approx.



Model 3

In the third model, along with backward elimination, we also used the output of sentiment analysis to improve the performance. Performance is nearly similar to the previous model (70.9%) but RMSE is reduced for this model that is 1.23.

The outliers were between 0 to 3.5 and most of them were zero, so filtered out some records with outliers above 2 to avoid poor model performance.

Attributes that are to be considered are bathroom, host total listings count, latitude, longitude, accommodates, price per night, reviews per month, review scores accuracy, review scores checkin, review scores cleanliness, review scores communication, review scores location, review scores value, and difference (raw sentiment score).

Top 3 most frequently co-occurring review score attributes that drop below 10:

The results pointed out that there are 63 number of sets with item numbers varying from 1 to 6. And we are focused on 3 itemset. We choose lift criterion for the association rules, and we set lift=7 as lower limit. Maximum lift is 8.364. Checking the premises and conclusion, we can look for the most frequently co-occurring attributes that drop below 10. That is, after filtering out the rules below 7, we now calculate the frequency of occurrence of attributes, and the top 3 most occurring ones are the result.

No.	Premises	Conclusion	Lift
28	cleanliness, value, accuracy, communicati...	checkin, location	8.364
44	checkin, location	cleanliness, value, acc...	8.364
26	cleanliness, accuracy, communication	checkin, location	8.285
27	cleanliness, accuracy, communication	value, checkin, location	8.285
42	checkin, location	cleanliness, accuracy, c...	8.285
45	value, checkin, location	cleanliness, accuracy, c...	8.285
56	cleanliness, checkin, location	value, accuracy, comm...	8.272
54	cleanliness, checkin, location	accuracy, communication	8.126
57	cleanliness, value, checkin, location	accuracy, communication	8.126
50	checkin, location	value, accuracy, comm...	8.110
49	checkin, location	accuracy, communication	7.967
51	value, checkin, location	accuracy, communication	7.967
40	checkin, location	cleanliness, value, com...	7.772
46	accuracy, checkin, location	cleanliness, value, com...	7.772

Premises are the conditions or patterns that must be present in order for the rule to be applicable and conclusion are items that are expected to be found when the premises are present.

Thus, the top 3 most frequently co-occurring review score attributes that drop below 10 are *communication, location and checkin*.