Practical Assessment

BC3409 AI in Accounting and Finance

Ng Chen Ee Kenneth

U1721316F

2 Nov. 20

Contents

Problem Statement

Airbnb allows its hosts to price their properties. There is a lack of indicators that allow hosts to compare similar listings in their neighbourhood. This has resulted in a suboptimal pricing. As a co-founder of a promising start-up, you intend to solve this problem using state of the art machine learning and natural language processing techniques to advise Airbnb on how to market their property for the best returns.

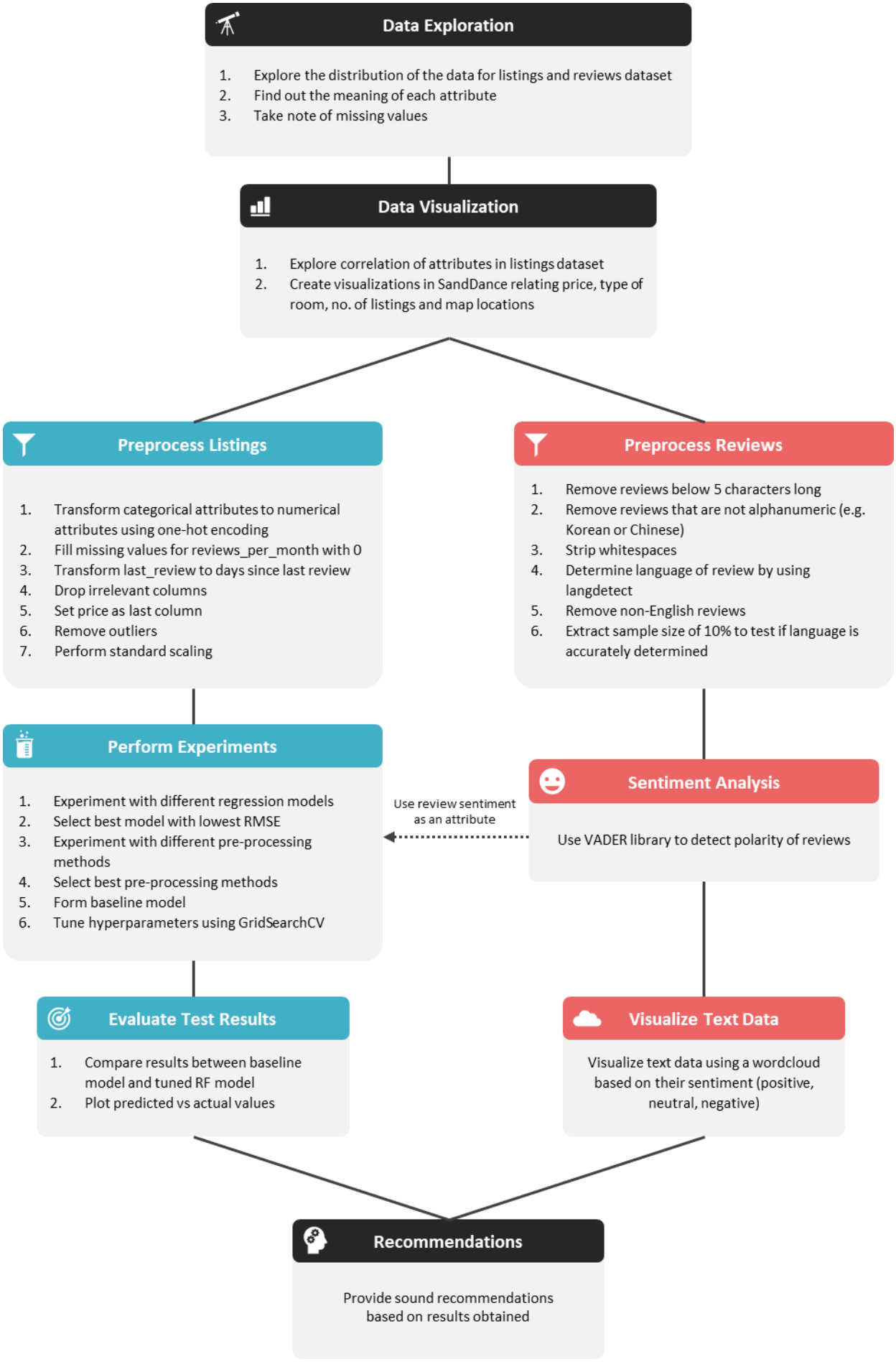
Overview

The next page shows a summary of the steps taken to analyse the datasets and predict prices of the listings. As the objective of the project is to determine an appropriate pricing for the listing using a machine learning model, we can make use of the listings dataset to predict the price of a listing based on the attributes of the listing. (location, no. of reviews, reviews per month, etc.)

Assumptions

It is difficult to determine the quality of the rooms based on just the location of the stay. Due to the nature of the dataset, I made a few assumptions which may not correctly reflect the optimal pricing of Airbnb rooms.

1. Listings are assumed to be price optimal
2. Hosts provide the same amenities for a specific room type
3. A uniform price is shared among listings for their additional fees
4. Listings cater to the same number of guests
5. Listings have the same number of rooms for a specific room type


Data Exploration
1. Explore the distribution of the data for listings and reviews dataset
2. Find out the meaning of each attribute
3. Take note of missing values
Data Visualization
1. Explore correlation of attributes in listings dataset
2. Create visualizations in SandDance relating price, type of room, no. of listings and map locations
Preprocess Listings
1. Transform categorical attributes to numerical attributes using one-hot encoding
2. Fill missing values for reviews_per_month with 0
3. Transform last_review to days since last review
4. Drop irrelevant columns
5. Set price as last column
6. Remove outliers
7. Perform standard scaling
Preprocess Reviews
1. Remove reviews below 5 characters long
2. Remove reviews that are not alphanumeric (e.g. Korean or Chinese)
3. Strip whitespaces
4. Determine language of review by using langdetect
5. Remove non-English reviews
6. Extract sample size of 10% to test if language is accurately determined
Perform Experiments
1. Experiment with different regression models
2. Select best model with lowest RMSE
3. Experiment with different pre-processing methods
4. Select best pre-processing methods
5. Form baseline model
6. Tune hyperparameters using GridSearchCV
Use review sentiment as an attribute
Sentiment Analysis
Use VADER library to detect polarity of reviews
Evaluate Test Results
1. Compare results between baseline model and tuned RF model
2. Plot predicted vs actual values
Visualize Text Data
Visualize text data using a wordcloud based on their sentiment (positive, neutral, negative)
Recommendations
Provide sound recommendations based on results obtained

Data Exploration

The exploration of the dataset is broken down into 2 segments.

1. Exploring the distribution of the data
2. Attribute Information (Discover the meaning of attributes)

Exploring Distribution

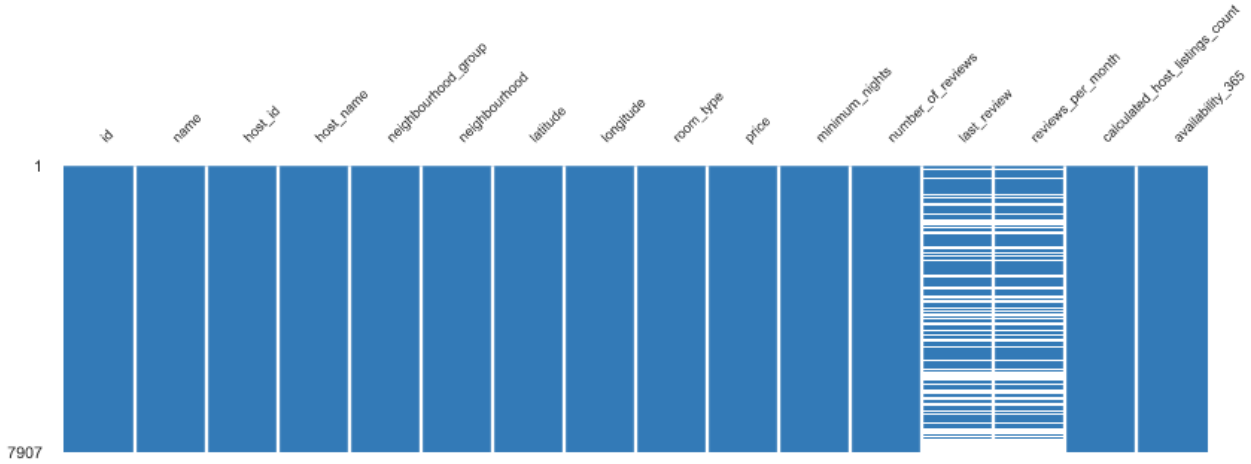


Figure 1: Columns with missing values in listings dataset

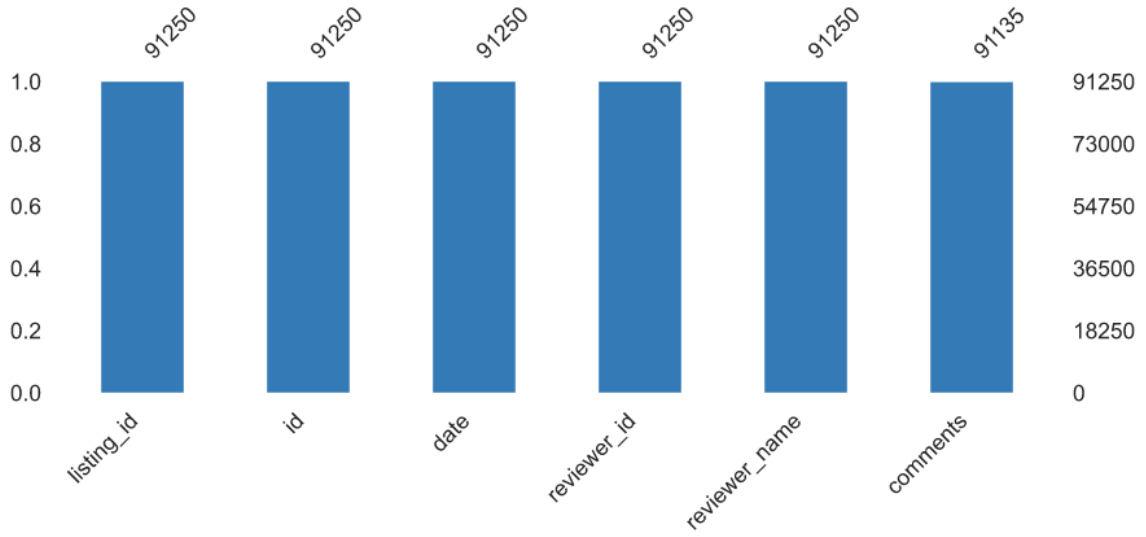


Figure 2: Columns with missing values in reviews dataset

From Figure 1, we can observe that there are multiple missing values (34.9%) in the last_review and reviews_per_month attributes. The last review attribute is tied to reviews per month. This is probably due to the fact that there are 0 reviews for the listing. There are also a few missing values (0.1%) for the comments attribute in the reviews dataset as seen in Figure 2. These instances with missing values will either need to be replaced with an appropriate value or removed from the dataset.

Attribute Information

COLUMN NAME	DESCRIPTION
Listing Dataset	
id	id of listing
name	Name of listing
host_id	id of host (can be repeated as 1 host can have many listings)
host_name	Name of host
neighbourhood_group	5 regions (Central, West, East, Northeast, North)
neighbourhood	43 distinct neighbourhoods (e.g. Kallang)
latitude	X position on the map
longitude	Y position on the map
room_type	3 types of room (Average Price: Entire home > Private > Shared)
price	Price of room (range from 0 to 10000)
minimum_nights	Minimum number of nights guest has to stay
number_of_reviews	Number of reviews for this listing
reviews_per_month	No. of reviews posted on listing per month
calculated_host_listings_count	Actual number of host listings (metric to measure host experience)
availability_365	No. of days in a year it was made available at that point in time
Reviews Dataset	
listing_id	id of listing
id	Unique id of review
date	Date of review
reviewer_id	Unique id of the reviewer
reviewer_name	Name of the reviewer
comments	Comment written by reviewer

The above information was gathered by analysing the distribution of the columns using Pandas Profiling.

Files: 1. Pandas Reports -> create_report.py, listing_report.html, reviews_report.html

Data Visualization

Regions

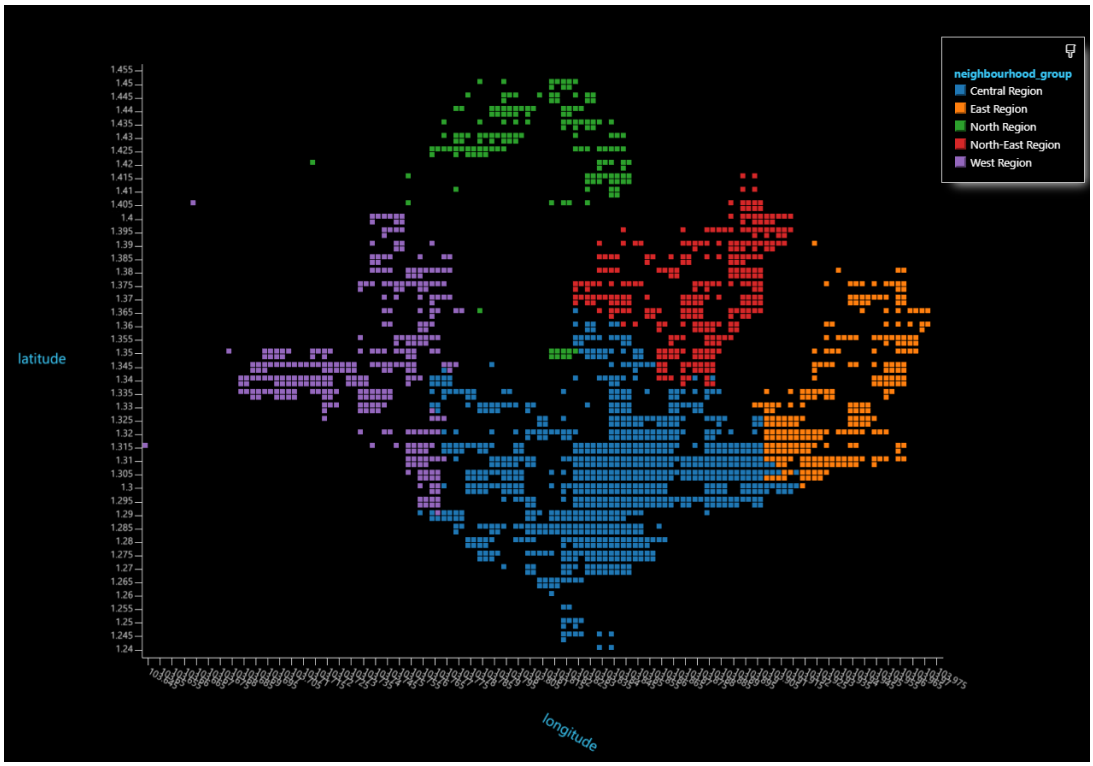


Figure 3: Neighbourhood groups of listings

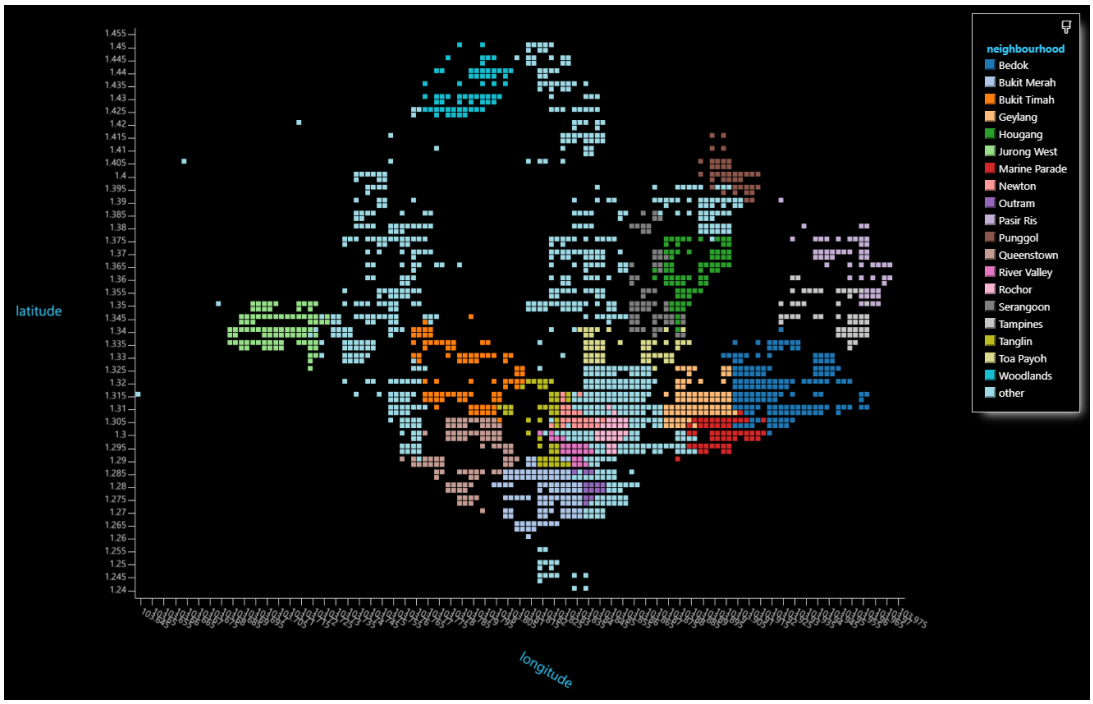


Figure 4: Neighbourhood of listings

Figure 3 and 4 shows that the Airbnb listings are segregated into 5 different regions and 43 different neighbourhoods.

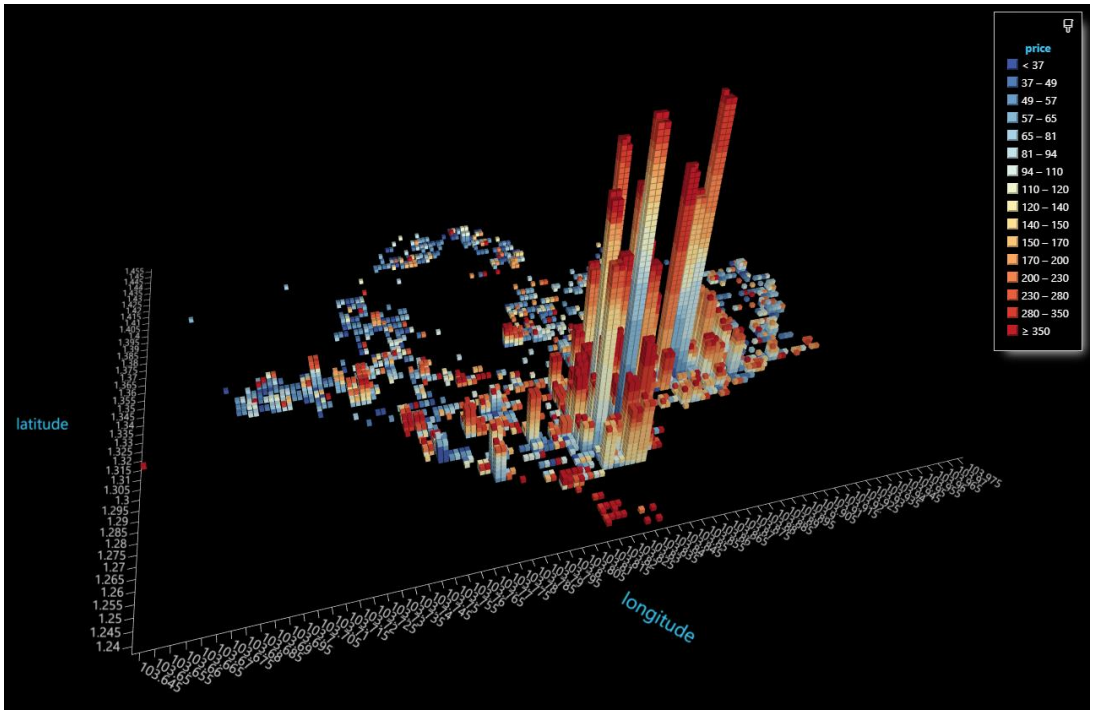


Figure 5: Visualization of the number of listings and their prices

In Figure 5, we can see that most of the listings are concentrated around the central region of Singapore. Additionally, the prices of the listings are generally lower for other regions.

Number of Listings

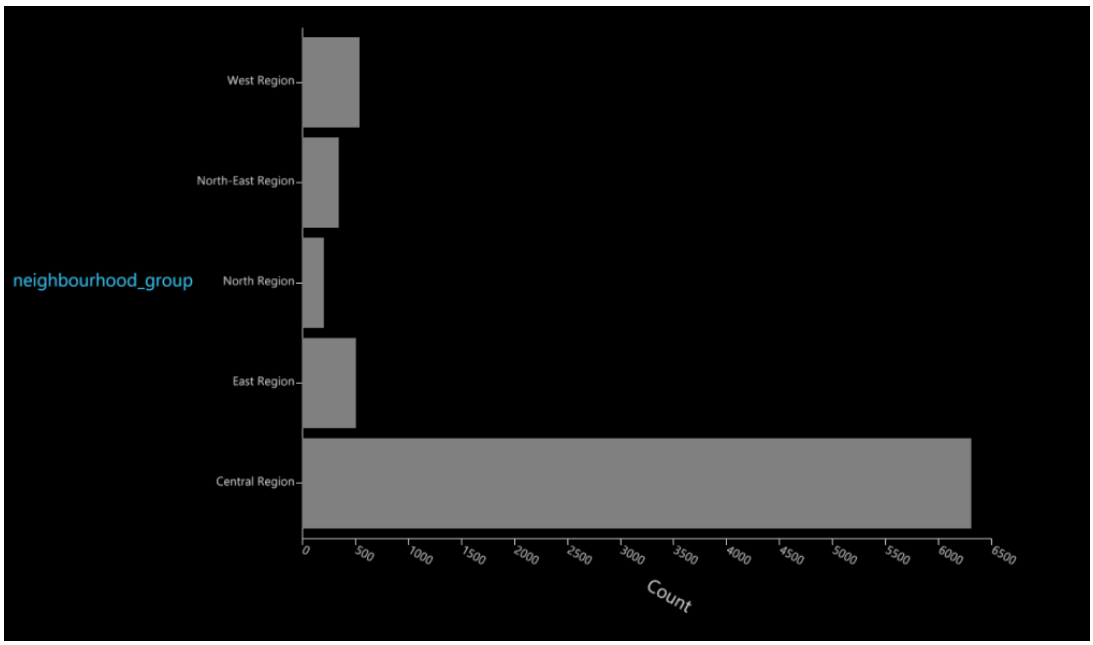


Figure 6: Number of listings based on region

More than 80% of the total number of listings are in the Central Region as we can see from the figure above. North Region has the lowest number of listings at 2.7%.

NO.	NEIGHBOURHOOD	NEIGHBOURHOOD GROUP	COUNT
1	Kallang	Central Region	**1043**
2	Geylang	Central Region	994
3	Novena	Central Region	537
4	Rochor	Central Region	536
5	Outram	Central Region	477
…			
39	Sungei Kadut	North Region	4
40	Western Water Catchment	West Region	3
41	Mandai	North Region	1
42	Marina South	Central Region	1
43	Tuas	West Region	1

The top 5 listings are all found in the Central Region as expected, taking up 45.7% of all listings. The neighbourhoods with the least number of listings are found in the outskirts of Singapore such as Tuas or Marina South where it is uncommon to find both hosts and guests.

Files: 2. Preprocessing -> Listings -> eda.ipynb

Prices

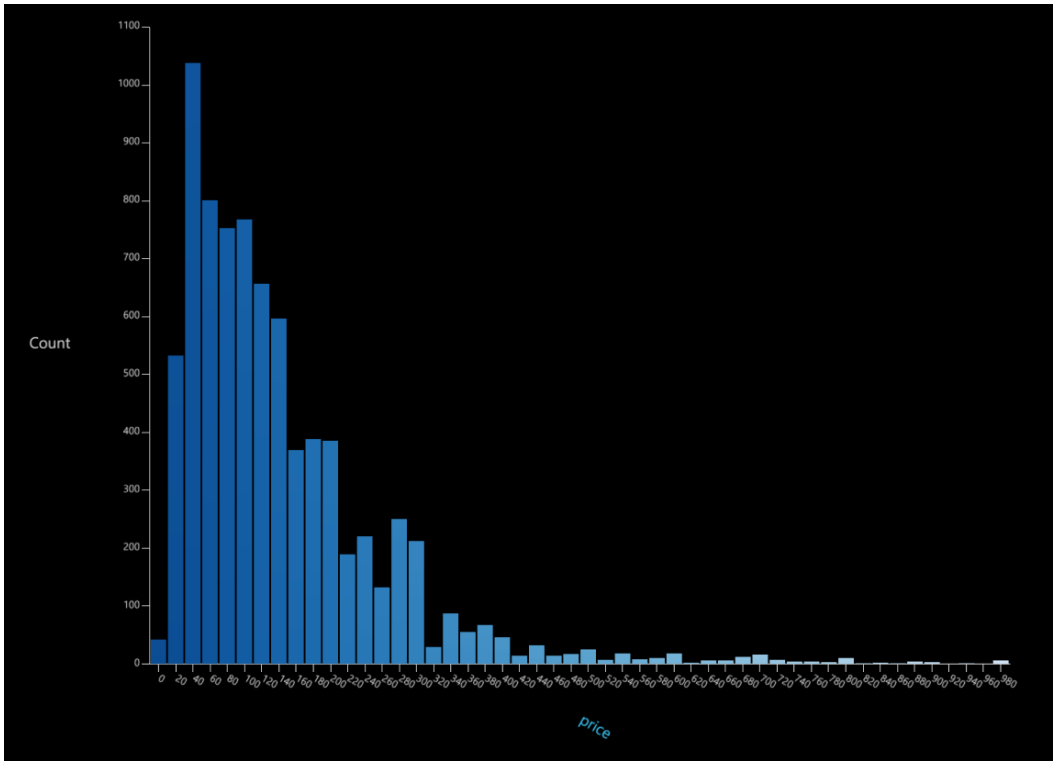


Figure 7: Distribution of price of listings

The above figure shows the distribution of the price of the listings. Most price listings are between the range of 22 to 400 per night. Based on the dataset, the price listing of an Airbnb room can be dependent on different factors such the type of room (Entire home, private or shared) and the location.

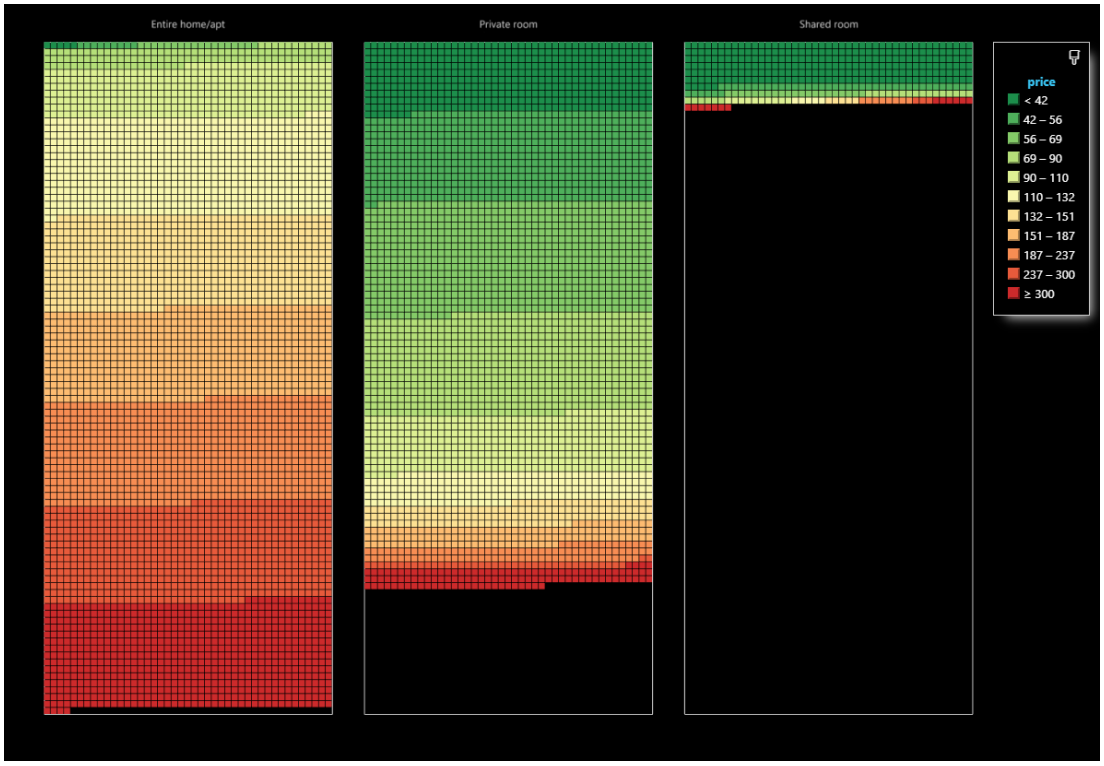


Figure 8: Price of listing based on type of room

Airbnb listings are categorized into 3 home types [1]

- **Entire place**: Guests have the whole place to themselves. This usually includes a bedroom, a bathroom, and a kitchen. Hosts should note in the description if they'll be on the property (e.g. "Host occupies ground floor of the home")
- **Private room**: Guests have their own private room for sleeping. Other areas could be shared.
- **Shared room**: Guests sleep in a bedroom or a common area that could be shared with others.

From Figure 8, we can deduce the shared room prices listed on Airbnb is usually cheaper than the private or the entire apartment. Private rooms are slightly more expensive in general as compared to shared room and renting an entire apartment usually costs more than $90 per night. Entire apartment listings take up more than half of all listings (52.2%) while private and shared room take up 42.75% and 4.98% respectively.

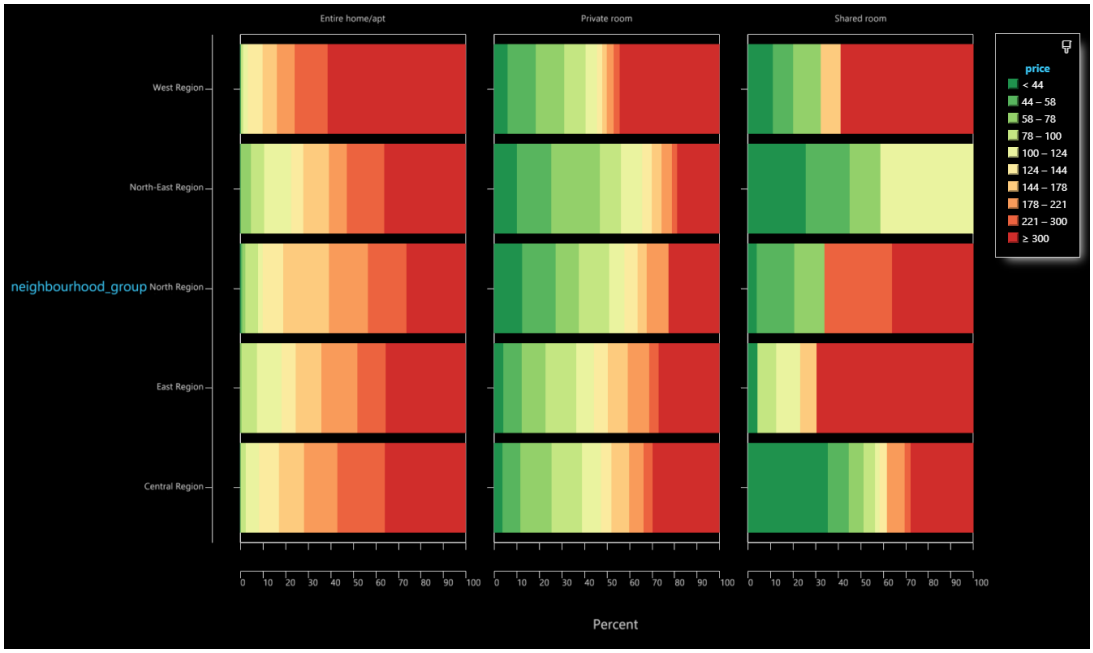


Figure 9: Price of listing based on type of room faceted with room type

In the visualization above, we split the home types into their respective regions. The stacked bar chart depicts the percentage of listings having a similar price listing. What we can observe is that the prices of the shared room tend to be lower in the North-East region (average: 55) as well as the Central Region. (average: 59.19)

Files: 2. Preprocessing -> Listings -> eda.ipynb

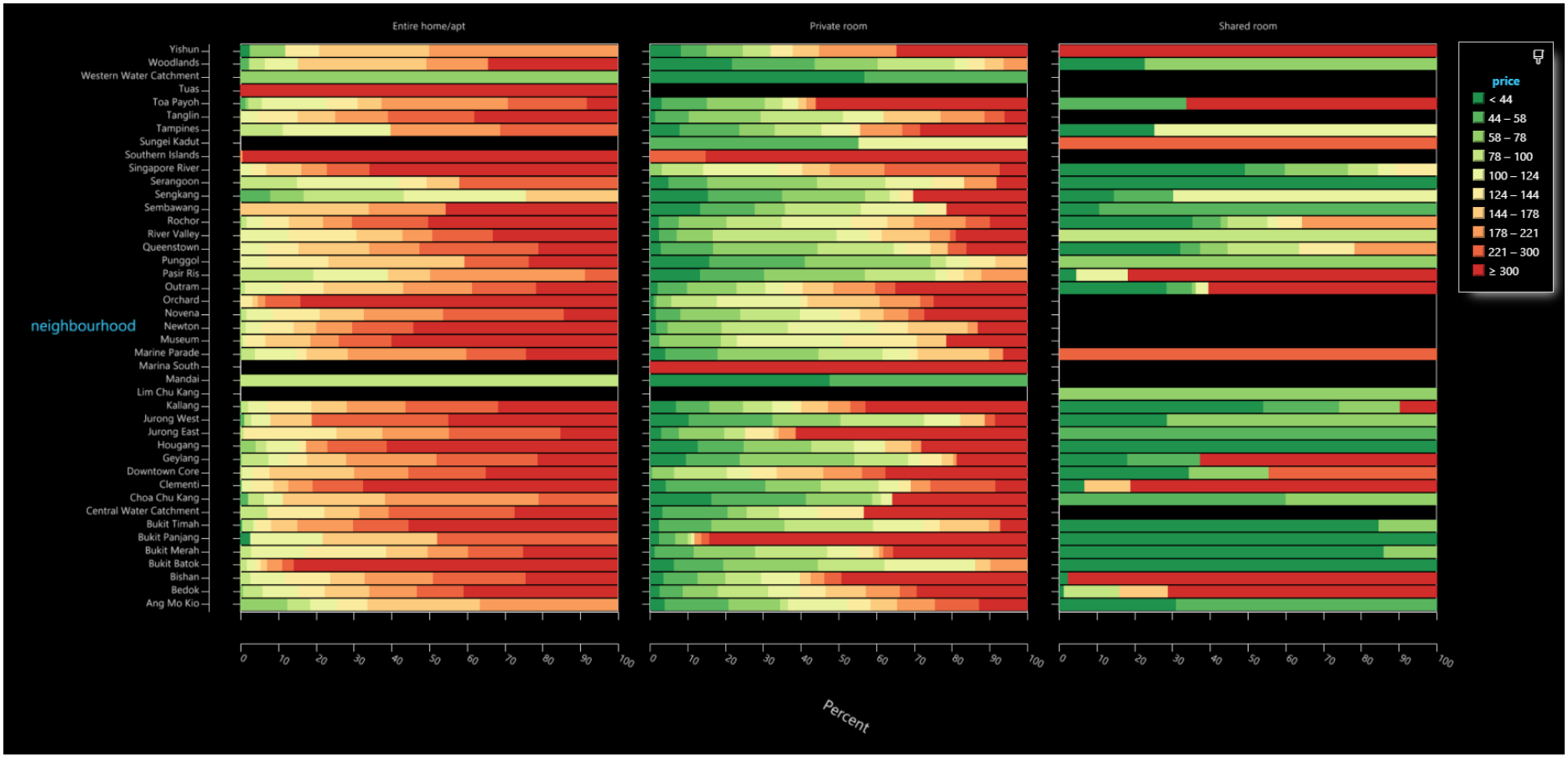


Figure 10: Prices of each neighbourhood faceted with the room type

Referencing Figure 10, we can see that the prices of entire apartment tend to vary widely based on location. The locations furthest away from the central region of Singapore tend to fall on opposite ends of the price range. There is not enough information here to make an inference as the price of a room can be affected by other factors such as the available amenities, whether the host is a superhost or the restrictions that the host has set in place.

Data Preprocessing

Preprocess Reviews

After inspecting the data, it was discovered that the comments left on a listing may not be limited to just English. There are people who leave reviews in other languages such as German, Chinese, Japanese and Korean. As such, to facilitate the analysis of the text data, we have to remove the reviews that are not in English.

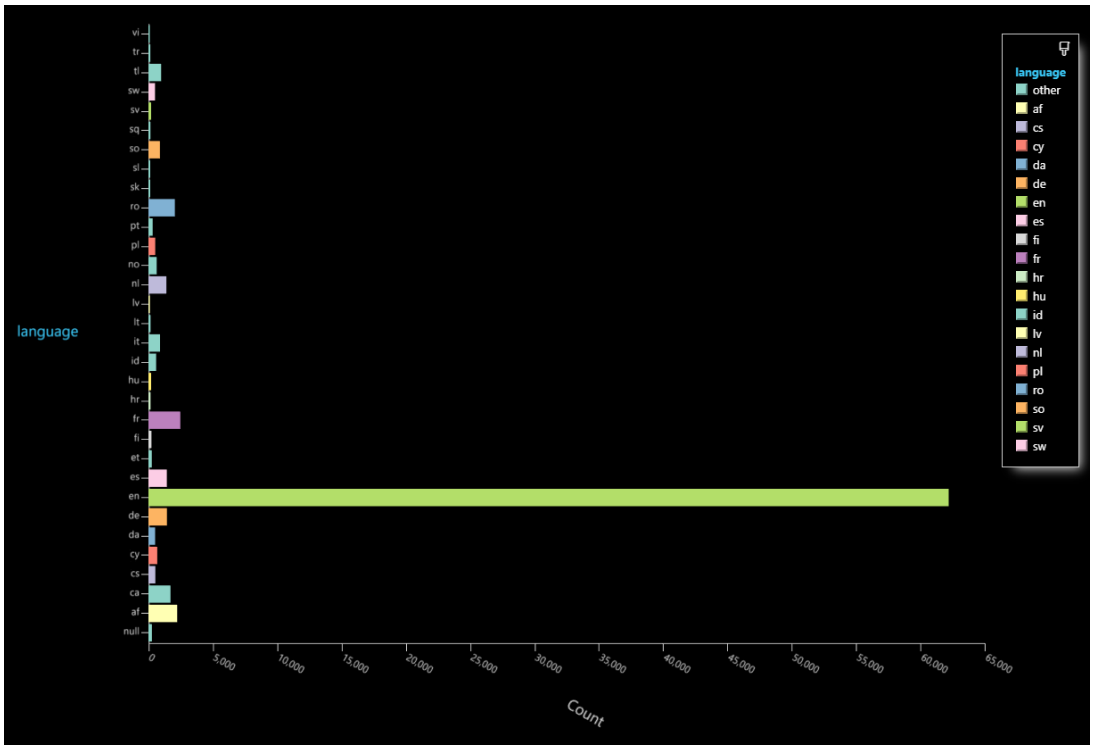


Figure 11: Variety of languages used in reviews

Steps to cleaning the text data

1. Remove reviews that are too short (lesser than 5 characters). These reviews will likely be unhelpful in determining the sentiment of the reviews
2. Remove all characters that are non-alphanumeric
3. Strip trailing or leading whitespaces
4. Lowercase characters
5. Remove empty strings
6. Take the first 15 characters of the comments and use langdetect library to detect the language of the comment. This is done because it takes hours for the langdetect library to scan through the corpus. By reducing the word count, we effective reduce the time needed to get the language of the comment.
7. Filter out comments that are not in English
8. Evaluate results by taking a sample of the dataset (10%)

Notebooks: 2. Preprocessing -> Reviews -> preprocess_reviews.ipynb

Data Files: reviews.csv, reviews_cleaned.csv, reviews_cleaned_sentiment.csv, sentiment_rating.ipynb

Manual Inspection

	comments	length	compound	English
5	2 UBahnstationen und Busstationen in nchster Umgebung perfekte LageFantastische Aussicht 25 Stock ber SingapurKommunikation mit Eva ging sehr einfach und schnell	173	0	No
7	Yuan hat sich sofort gemeldet und schnell auf alle Fragen geantwortet Die Unterkunft war fr den Preis echt in Ordnung und sehr sauber Sie war sehr lieb und entgegenkommend Immer wieder gerne	195	-0.9136	No
4	Kingsley is heel vriendelijk en behulpzaam Snelle reactie en denkt met je mee over het reizen met het ov De kamer is prima We zijn nog even samen naar de supermarkt gegaan waar we een leuke rondleiding hebben gekregen	223	0	No
3	Rosey a t accueillante elle nous a attendue jusqu 2h du matin car notre avion atterrissait tard elle nous a donn des conseils pour conomiser un peu l appartement est bien situa 10 min d un centre commercial et dune station MRT mais l appartement ntait pas propre Beaucoup de longs cheveux noirs sur le sol et colls dans la salle de bain beaucoup de bazar au sol la salle de bain est extrmement rudimentaire mais surtout l odeur dans les toilettes et l salle de bains taient extrmement dsagreable cause de pastilles d urinoir poses partout y compris dans le lavabo a peut convenir des voyageurs petit budget mais pas pour tout le monde	672	0.7906	No
	Da Catherine mich gebeten hat meine Bewertung auf Deutsch zu schreiben komme ich dem gerne nachWir hatten eine tolle Zeit in ihrem Apartment Die Kommunikation mit ihr war schon vor unserer Ankunft sehr unkompliziert Uns stand vor Ort genug Platz und ein groer Schrank mit Ablageflche zur Verfgung Das Zimmer ist ansonsten sehr hell und gemtlich und kann mit Vorhngen abgedunkelt werden AC hat problemlos funktioniert genau so wie das InternetDa wir das Zimmer gemietet hatten um selbst ein Apartment in Singapur zu finden haben wir uns riesig ber			

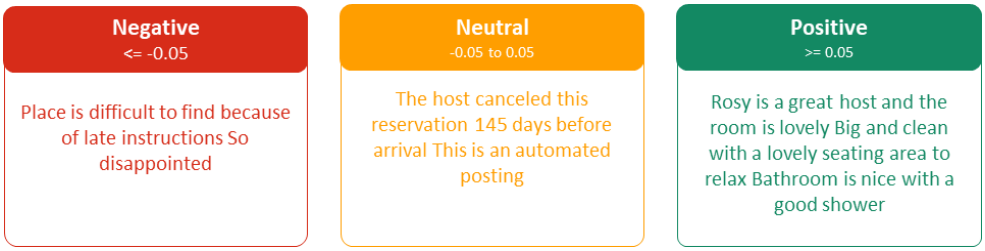
Figure 12: Mistakes made during language detection

Upon picking a sample size of 10% of the cleaned dataset (6214 instances out of 62133). It was discovered that the language detect library is able to correctly classify 99.903% of the comments. There were only 6 errors found in the sample dataset and part of it can be seen in Figure 12. The reasons why it was misclassified can be due to the following:

1. First term being a name of an individual or a number
2. The comment was half in English and half in non-English
3. Spelling errors causing the comment to be misclassified as another language

Files: 4. Misc -> sample_sentiment.csv

Sentiment Analysis



Next, we perform sentiment analysis on the text data by using the VADER ((Valence Aware Dictionary and Sentiment Reasoner) library in Python. VADER is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. It measures the compound score of a length of text by summing the valence scores of each word in the lexicon, then normalizing it to be between -1 (most extreme negative) and +1 (most extreme positive). An example of the classification done using VADER is shown above.

Visualizing Text Data

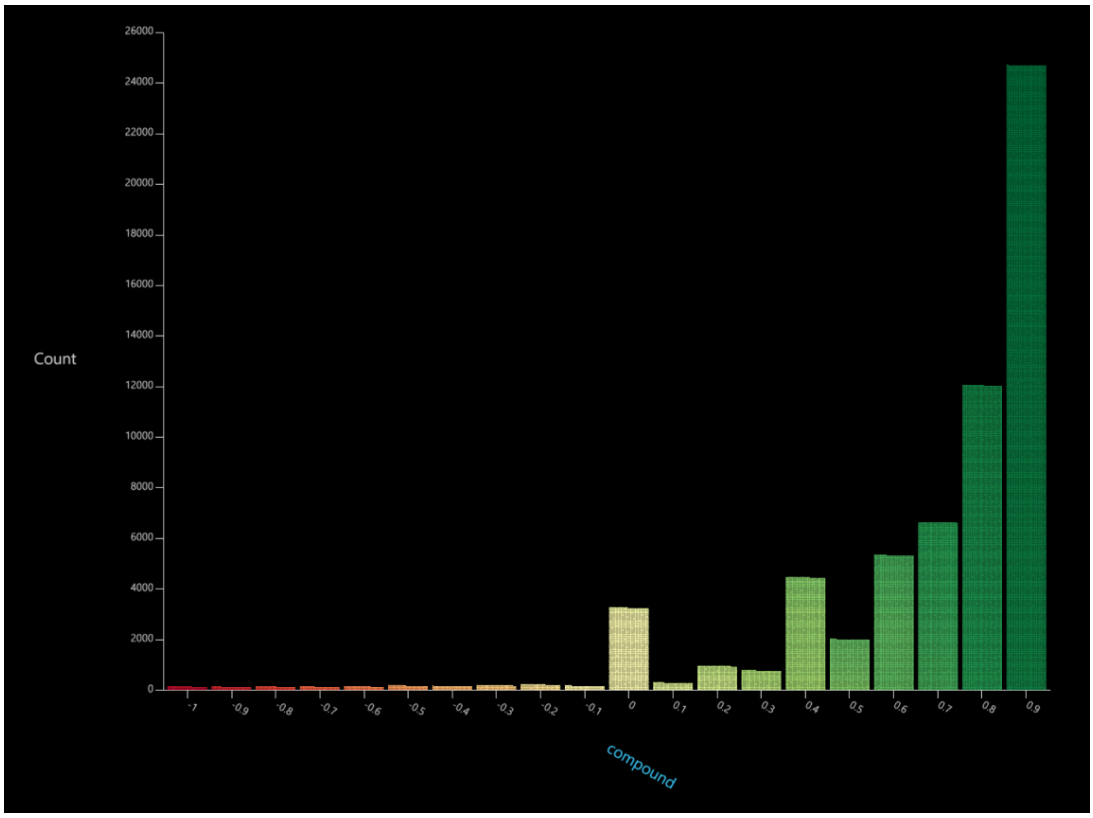


Figure 13: Sentiment of reviews from -1 (most negative) to +1 (most positive)

From Figure 13, we can see that there are an overwhelming number of positive reviews left by guests. Only a small percentage (2.61%) of reviews are negative.

Tomas	The host canceled this reservation 11 days before arrival This is an automated posting	88	en	0	0	1	0
Michelle	The host canceled this reservation 16 days before arrival This is an automated posting	88	en	0	0	1	0
Tiang	The host canceled this reservation 23 days before arrival This is an automated	88	en	0	0	1	0
é"¦	The host canceled this reservation 28 days before arrival This is an automated	88	en	0	0	1	0
Legend	The host canceled this reservation 12 days before arrival This is an automated posting	88	en	0	0	1	0
ä¼Š	The host canceled this reservation 13 days before arrival This is an automated posting	88	en	0	0	1	0
S Vijay	The host canceled this reservation 20 days before arrival This is an automated	88	en	0	0	1	0
Honey	The host canceled this reservation 11 days before arrival This is an automated posting	88	en	0	0	1	0

Figure 14: Automated reservation cancellation

There are also a number of neutral reviews found to be automated review cancellation done by Airbnb when the sample data is manually inspected.

Files: 4. Misc -> sample_sentiment.csv

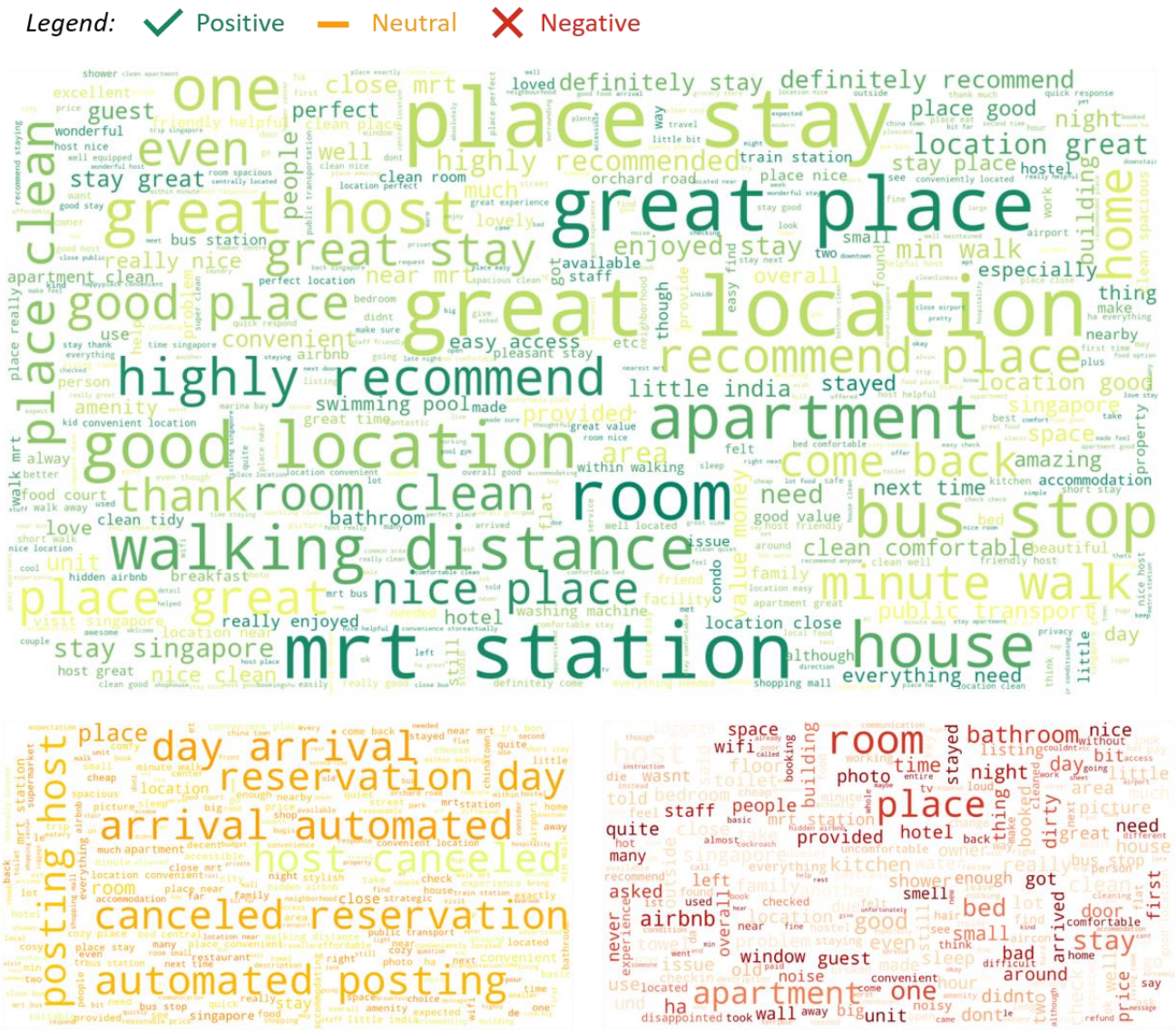


Figure 15: Word cloud of positive, neutral and negative reviews

The word cloud is generated by lemmatizing the terms and removing the stop words and non-alphanumeric characters. We can infer that most guests that leave reviews mentioned that the location was a great stay due to the distance to the MRT station or bus stop. Several other key words include having a clean place, friendly host and great accommodation. On the other hand, the negative reviews mentioned issues related to the bathroom, noise and smell.

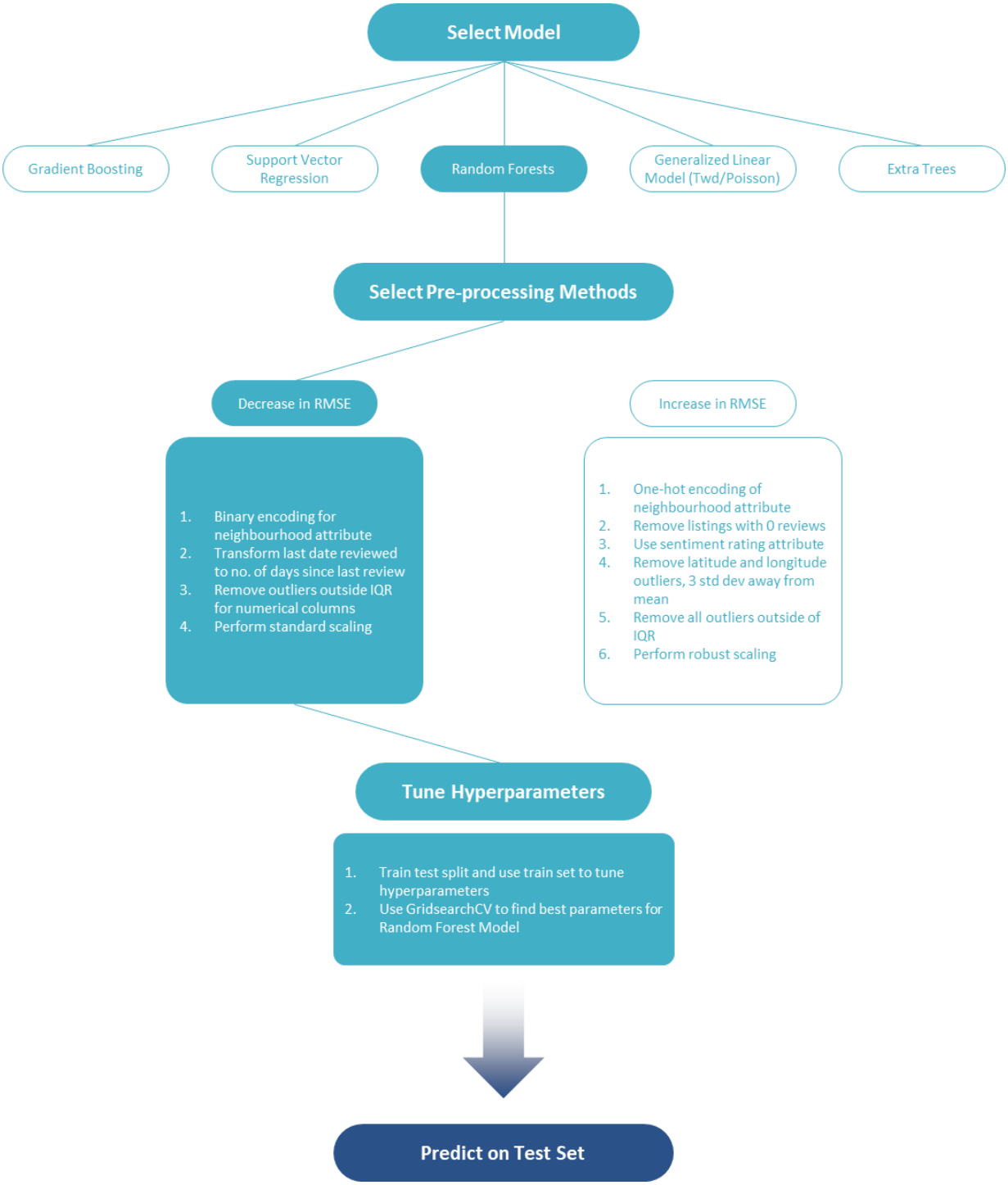
Notebooks: 5. Wordcloud -> wordcloud.ipynb, wordcloud_functions.py

Files: reviews_cleaned_sentiment.csv

Modelling

Overview

Below summarizes the process of deriving the model with the best hyperparameters to get the best performing results. We first select the best performing model based on the RMSE score. Next, we experiment with various pre-processing methods and choose methods that effectively reduce the RMSE score. After the pre-processing methods have been chosen, we tune the hyperparameters on the training set by using a 'coarse to fine' approach, selecting the best parameters. The final test would then be to use these hyperparameters on the test set to derive the test RMSE score.



Selection of Model

Name (6 visualized)	Notes	MAE	MSE	RMSE ▲
RF	dataset_1	37.454	3388.511	58.211
Extra Trees	dataset_1	37.737	3643.629	60.362
GB	dataset_1	46.248	4389.386	66.252
GLM (Poisson)	dataset_1	52.335	5382.103	73.363
GLM (Twd)	dataset_1	58.88	6088.416	78.028
SVR	dataset_1	57.11	7546.291	86.869

Figure 16: Test results of different models

Random Forest is selected as the best performing model as it has the lowest RMSE score among the 6 models used. Random Forests work by training many decision trees on random subsets of the features, then averaging out their predictions.

Advantages and disadvantages of using the Random Forest Model

Advantages

1. It can be trained using a small sample of data and provide good performance
2. It can be used to solve both regression and classification problems
3. It has methods to balance errors in datasets where classes are imbalanced
4. Random Forest are usually quite robust to outliers

Disadvantages

1. A large number of trees can make the algorithm too slow and ineffective for real-time predictions
2. It does not provide good interpretability unlike decision trees

Link to model results

Selection of Pre-processing Methods

Model Name	Dataset Name	One-hot encoding	Binary Encoding	Remove listings with 0 reviews	Last day review	Sentiment Rating	Remove lat and long outliers	Remove all outliers	Remove outliers (numerical cols only)	Use robust scaling	RMSE
Baseline Model	Dataset_1	Yes	No	No	No	No	No	No	No	No	58.211
RF 2	Dataset_2	No	Yes	No	No	No	No	No	No	No	58.017
RF 3	Dataset_3	No	Yes	Yes	No	No	No	No	No	No	65.288
RF 4	Dataset_4	No	Yes	No	Yes	No	No	No	No	No	58.003
RF 5	Dataset_5	No	Yes	No	Yes	Yes	No	No	No	No	58.366
RF 6	Dataset_6	No	Yes	No	Yes	No	Yes	No	No	No	58.285
RF 7	Dataset_7	No	Yes	No	Yes	No	No	Yes	No	No	51.695

RF 8	Dataset_8	No	Yes	No	Yes	No	No	No	Yes	Yes	49.608
RF 9	Dataset_8	No	Yes	No	Yes	No	No	No	Yes	No	49.145

In the above table, each of the RF (random forest) model is compared to the baseline model. If there is a decrease in RMSE, the data pre-processing method is used. Details on how the pre-processing methods are derived can be found in the ipynb files.

Creation of sentiment attribute: 2. Preprocessing -> Reviews -> sentiment_rating.ipynb

Creation of Datasets: 2. Preprocessing -> Listings -> preprocess_listings.ipynb,

Testing of Preprocessing Methods: 3. Prediction -> preprocess_test.py, dataset_n.csv

Data files: listings.csv, reviews_rating.csv

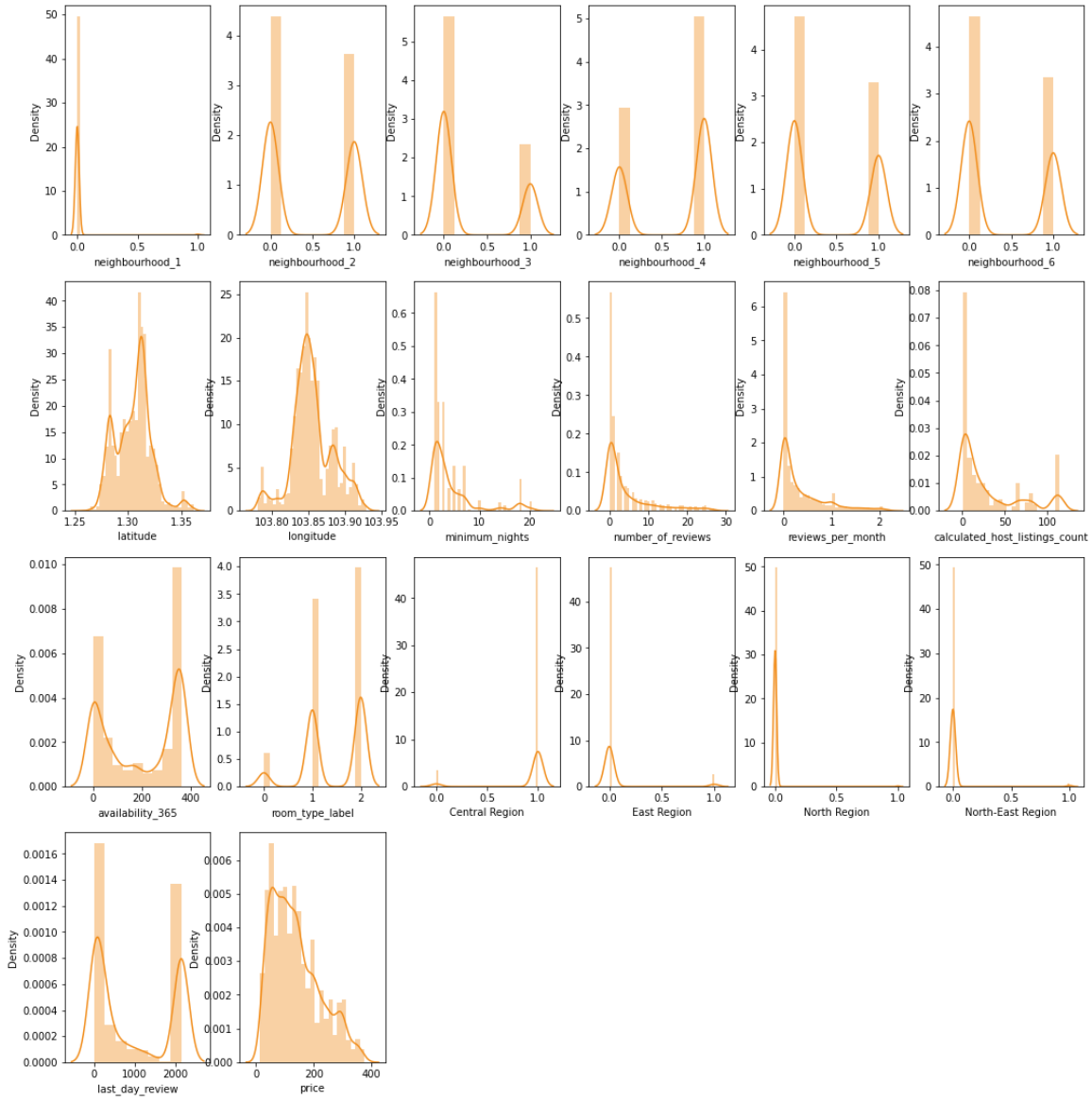


Figure 17: Distribution of attributes plotted using KDE after removal of outliers

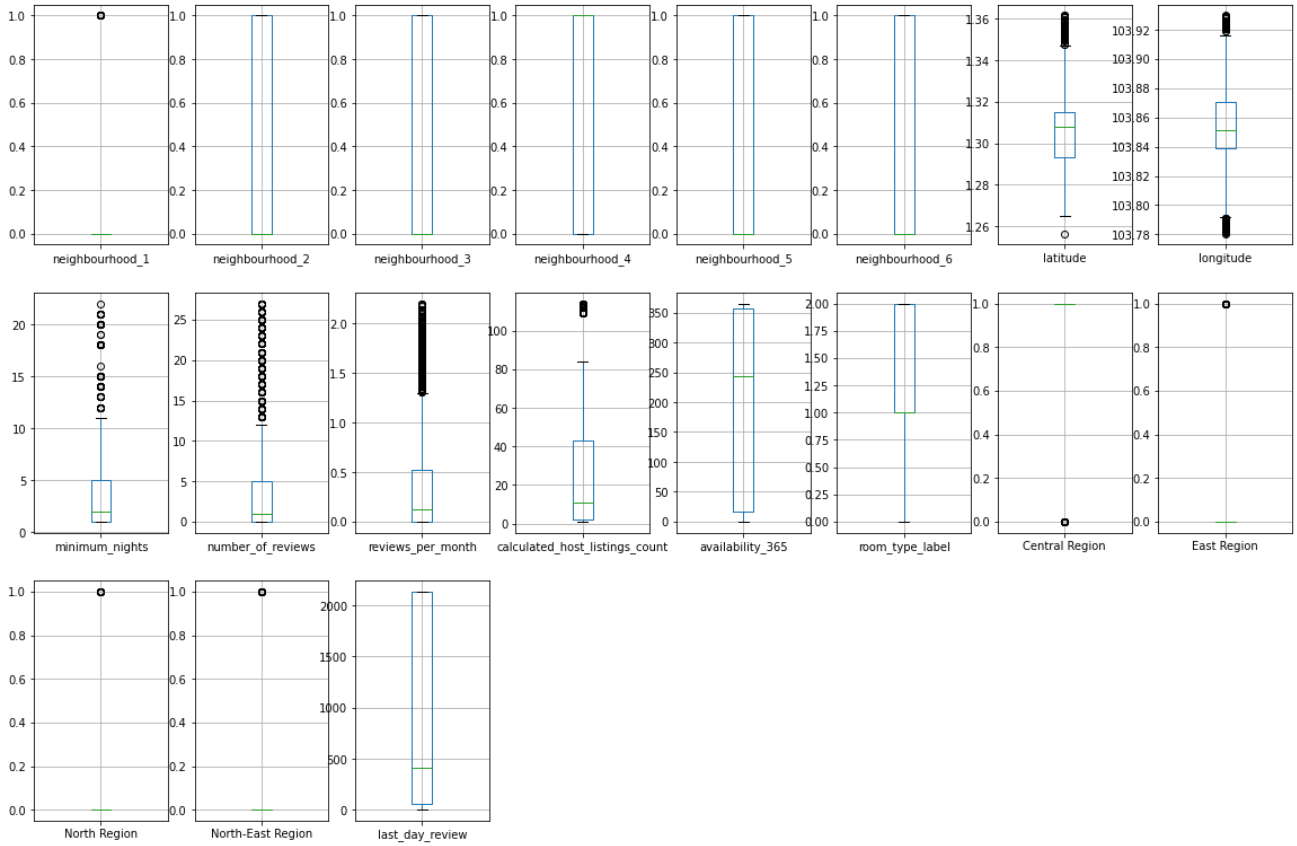


Figure 18: Distribution of attributes plotted using boxplot after removal of outliers

Optimization Strategy

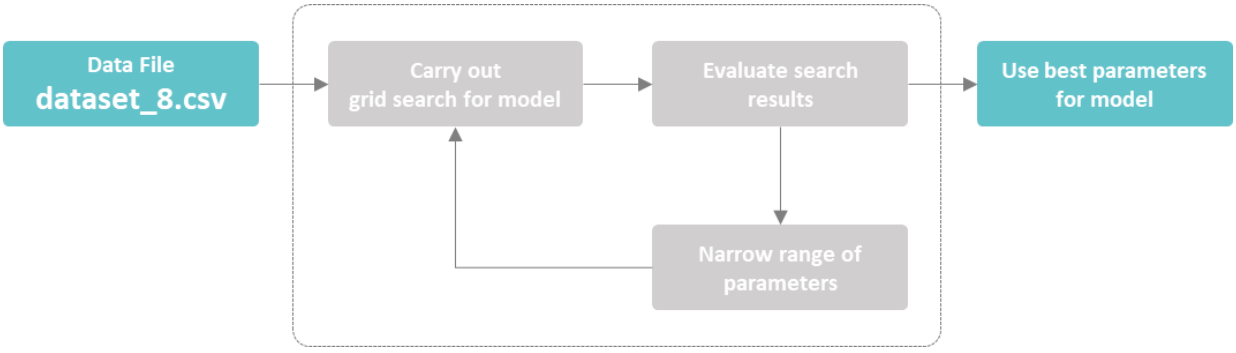


Figure 19: Diagram showing how hyperparameters are selected

The optimization strategy involves performing both grid search and coarse-to-fine search to narrow in on the best parameters.

GridSearchCV is a function in scikit-learn that determines the best hyperparameters for a model by passing in a range of parameter values. Choosing which hyperparameters to optimize is difficult as some are more sensitive than others and are dependent on the choice of the model.

Hence, a wide range of values (with log intervals) is first set for the parameters. The range of values that do not get a good score are then eliminated until the best hyperparameters are found. This strategy helps to

narrow in on only the high performing hyper-parameters. A 5-fold cross validation is also performed to avoid overfitting.

Files: 3. Prediction -> Hyperparameter Tuning -> tune_RF.ipynb, RF_cv_results_n.csv

Evaluation of Results

Name (10 visualized)	Notes	Runtime	MAE	MSE	RMSE
Baseline RF	dataset_1	1m 9s	37.671	3430.372	58.569
RF 2	dataset_2	1m 23s	37.716	3366.003	58.017
RF 3	dataset_3	33s	43.651	4262.584	65.288
RF 4	dataset_4	1m 16s	37.738	3364.382	58.003
RF 5	dataset_5	1m 15s	38.019	3406.588	58.366
RF 6	dataset_6	1m 44s	37.858	3397.09	58.285
RF 7	dataset_7	25s	36.759	2706.845	52.027
RF 8	dataset_8, robust scaler	28s	34.952	2460.915	49.608
RF 9	dataset_8	29s	34.638	2415.228	49.145
Tuned RF	dataset_8	2m 28s	34.565	2397.147	48.961

Figure 20: RMSE Test Scores

The above figure shows the RMSE scores from the initial dataset used (dataset_1) to the final dataset used (dataset_8) to using the tuned parameters for prediction on the test set. The RMSE score has decreased from 58.569 (Baseline model) to 48.961 (Tuned model).

Files: 3. Prediction -> final_RF_test.py, preprocess_test.py

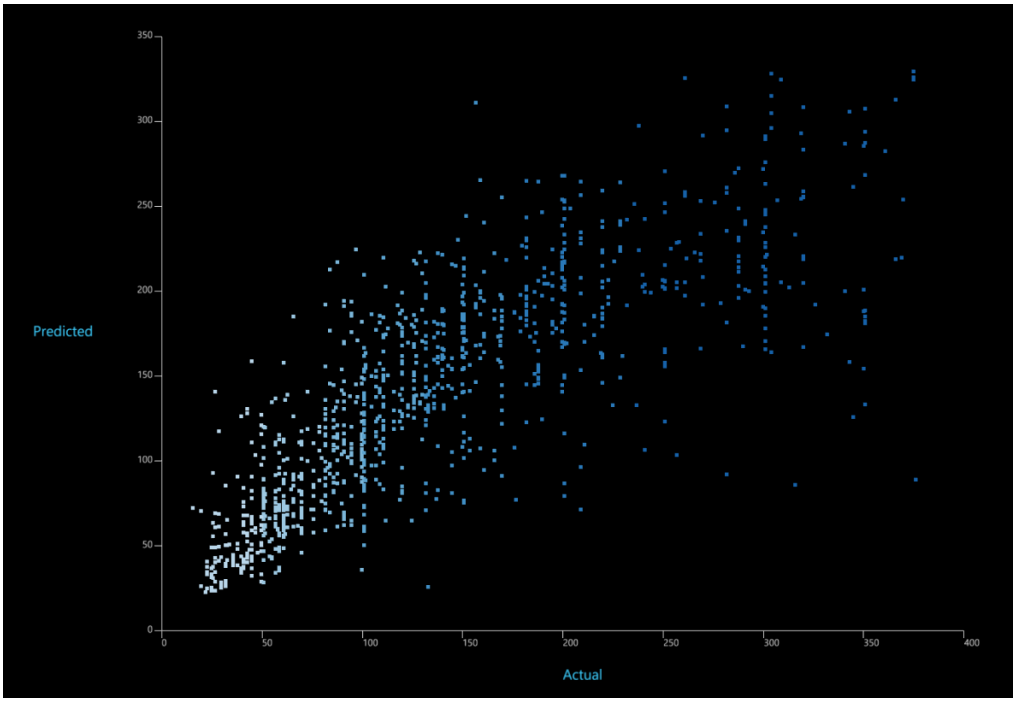


Figure 21: Plotting Predicted versus Actual values

Files: 3. Prediction -> Plots -> scatterplot.ipynb, plot_scatter.csv

Recommendations

Prediction

From the test results, what we can conclude is that the prediction of the prices may not be that accurate even after the tuning of hyperparameters. The error rate of 48.961 is still high. The error rate starts to increase as the price of the listing increases beyond 250 as seen in Figure 21. We can improve the prediction results by:

1. **Getting the host data features**
   By recognising if a host is a superhost, we are able to deduce which listings are recognised as reliable in the market and eliminate listings which are not.
2. **Amenities provided by the host**
   Providing a myriad of amenities may encourage a guest to register their booking for a host. If we were to get a list of available amenities provided by the host, this can help us to accurately decide the pricing strategy
3. **No. of rooms and no. of guests**
   A host may decide to increase his/her price based on the number of rooms and guests that can stay in an apartment. As this information was not known before, this could potentially skew results as we assume that listings have the same number of rooms and guests.
4. **Creating new features such as distance to the nearest MRT or bus stop**
   Analysing the sentiment of the reviews uncovered that guests often recommended a location based on the convenience to certain locations such as the MRT station and bus stops.
5. **Taking into consideration additional fees**
   The Airbnb website [2] recommends that hosts need to keep in check the additional fees (service fees, cleaning fees, etc) to determine a price that will work for the guests. This could be a possible reason why some listings fail to attract guests. By eliminate these listings with suboptimal prices, we are able to improve the performance of the model.

Using current data

Although the weak performance of the model may not be that helpful in determining the optimal pricing of an Airbnb room, we can still utilize what we have uncovered in the data exploration of the project. Airbnb optimizes the pricing of the listings available through Smart Pricing. [3] It does so by allowing hosts to set a maximum and minimum price. The price of the listing changes based on factors such as the lead-time, listing popularity, listing details, market popularity, etc. As we know the number of listings based on the neighbourhoods, this information can be used to inform the host of the range of the prices in their neighbourhood, letting them make an informed decision to price their maximum and minimum prices. Furthermore, a ranking of best priced listings can be collated from the dataset. This serves as a reference for a new host to decide how he can improve his listings, whether is it taking better high-quality photos or providing the right set of amenities for the guests.

Additionally, regions with high number of listings may be subjected to more competitive prices. While it may not be ideal for a host, it is still essential information that the host can use to decide their maximum and minimum prices.

Conclusion

In this project, what was done was to first explore the dataset. By understanding the nature of the dataset, what can be worked on, we can use the information to guide us in how we can pre-process the data and choose an appropriate model. The next step was to perform data pre-processing and cleaning. A combination of pre-processing methods was selected based on whether it led to a decrease in overall RMSE score. Following the training of the model, the evaluation of the models is carried out by testing the tuned model on the test set. Recommendations are then provided based on the insights and observations on the results

Appendix

Pandas Profiling Reports

https://github.com/pandas-profiling/pandas-profiling

SandDance Visualization Tool (any graphs with a black background)

https://marketplace.visualstudio.com/items?itemName=msrvida.vscode-sanddance&ssr=false#overview

Test Results of Regression models

https://wandb.ai/todayisagreatday/Airbnb%20Project?workspace=user-todayisagreatday

Preprocessing and Final Test Result

https://wandb.ai/todayisagreatday/Airbnb%20Tuning?workspace=user-todayisagreatday

VADER Library

https://pypi.org/project/vaderSentiment/

GridSearchCV

https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

Regression Models from sklearn

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.TweedieRegressor.html

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.PoissonRegressor.html

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesRegressor.html

Bibliography

[1]"What do the different home types mean?", *Airbnb*, 2020. [Online]. Available: https://www.airbnb.com.sg/help/article/317/what-do-the-different-home-types-mean. [Accessed: 02- Nov- 2020].

[2]"Set a price strategy", *Airbnb*, 2020. [Online]. Available: https://www.airbnb.com.sg/resources/hosting-homes/a/set-a-price-strategy-15. [Accessed: 02- Nov- 2020].

[3]"What's smart about Smart Pricing?", *The Airbnb Blog - Belong Anywhere*, 2020. [Online]. Available: https://blog.atairbnb.com/smart-pricing/. [Accessed: 02- Nov- 2020].