Airbnb Toronto

Final Report

Dongyan, Fahad, Huy, Jeni, Jordan, Mariana, Roma, Ryan
16-December-2021
17 Pages + Appendix

Table of Contents

1: Introduction & Problem Statement

Airbnb, is one of the most popular online housing platforms offering cheaper vacation rental alternatives for travellers compared to costlier hotel accommodations. Established in 2008, Airbnb has a proven and sustainable business model, evident by its resilience throughout the COVID-19 pandemic. Whereas many businesses hit financial hardship, Airbnb saw continued success despite pandemic travel restrictions and even became a public company in December of 2020.

As businesses reopen and economies in various sectors such as hospitality and tourism regain a foothold due to post-recovery from COVID-19, it is expected that the demand for travel will flourish, which will serve as a great opportunity to generate revenue in rental hosting.

Hosts pay commissions to Airbnb when customers book accommodations through Airbnb's online platform. Airbnb's top priority involves "perfecting the existing product by improving the entire end-to-end experience of their core service for both hosts and guests"[1]. We used predictive analytics to investigate how various features impact the price and review score of properties in Toronto, Ontario. Hosts will have the opportunity to explore potential areas for improvement, shape their reputation and increase the satisfaction of customers. This can lead to higher commissions for Airbnb with the goal of increasing revenue by 2% within the year 2022.

In addition to benefiting the company holistically and its shareholders; this strategy will also impact external stakeholders like property hosts and renters, along with internal stakeholders like strategic leadership, marketing teams, data engineers and programmers.

2: Research Methodology and Ethics

Airbnb actively uses data scientists to develop further business insights which have supported its overall success throughout the last several years.[5] Moreover, a number of analytical professionals have chosen to study Airbnb, its platform and structure, and have provided valuable opinions and insights. Two notable examples include Tian (2021) and Meijerink & Schoenmakers (2021).

Tian (2021) explored a number of theories and suggestions to increase host performance. These included utilizing the Natural Language Toolkit in Python to better understand the most popular words being used in reviews, room popularity and time, the popularity of each room type and their prices along with rating differences between super hosts and non-super hosts. Through exploring popular words being used in reviews, Tian was able to provide ideas into what features are of most value to customers. Tian suggested tangible actions that were tied to the top identified words: location, cleanliness, nice-host and everything. The interpersonal relationship between the host and customer is quite dynamic but certainly impacts overall perception and satisfaction. Additionally, Tian discovered that cleanliness and additional services and amenities, like having dishes, bedding, TV, and gaming services, increased satisfaction. Tian's research also recommends, promoting hosts to be fully engaged from July-December, improving response times to customers and a strong focus on overall cleanliness.[3]

Meijerink & Schoenmakers(2021) delves into understanding the dynamics of how and when reviews are left on Airbnb's platform. They discovered a positive linear relationship between customer stays and reviews, whereas in other industries reviews form more of a U-shape. That is to say that in most other industries customers who are either very satisfied or very dissatisfied are likely to leave reviews. Alternatively, in the sharing economy, the positive linear relationship suggests that customers are more likely to leave reviews only when they are very satisfied.

Through analysis and conducting their own data collection research, Meijerink & Schoenmakers argue the two main factors that skew towards positive ratings are non-anonymous reviews and that the reviewing process is two-way. A two-way reviewing process is one where a customer leaves a review about their host and the host leaves a review about their customer. In order to truly further understand customer satisfaction, the authors suggest Airbnb look at returning customers to the platform as well as hosts that have repeat customers.[4]

Airbnb datasets are widely available on public sites like Kaggle[7], inside Airbnb[6], and data.world[8] to name a few. Given its open accessibility, managing the data in an ethical way should be taken seriously. Prior to release, it's imperative that steps are taken to ensure the data is anonymized and that disclosure rules regarding the collection and use of data are followed to help minimize the potential exploitation of consumer and host information. Airbnb collects multiple forms of consumer data like geolocation information, usage information, log data and device information and payment and transaction information. Additionally, if a consumer links, connects or logs in through a social platform, Airbnb is authorized to collect personal information such as friends lists, profile information and background information. Collecting this information requires consumer consent, which customers should have the option of withdrawing consent at any time.[9]

3: Model Creation and Evaluation

For an appropriate Machine Learning model to be implemented, data cleaning of the Airbnb dataset was required. Exploratory analysis (EDA) of the original dataset (uncleaned) was first conducted through the use of Tableau. The data dictionary of the original dataset is shown in Table 1 of the Appendix. Graphs and plots to identify relevant features and the relationships between the features were generated from the EDA (e.g., Figure 1 and Figure 2 in the Appendix). Secondly, as is common with real world data, several features contained missing values and as result, imputation methods (Simple Imputer and Iterative Imputer from Sklearn Python library) were implemented in Python to replace the missing values. Thirdly, outliers were identified in the dataset using both Tableau and Python. We detected both negative and positive skewness in several features of the dataset. The interquartile range (IQR) method was selected to remove the outliers. Finally, using feature engineering, we created four features that we thought were relevant to our business objective. One feature created in the dataset was 'former city' based on Toronto districts. There were more than 40 neighborhoods and as a result, we decided to consolidate each neighborhood into a city. In Python, we webscraped the Wikipedia page https://en.wikipedia.org/wiki/Toronto, under the section entitled "Neighborhoods" to collect all the districts and associated neighbourhoods in Toronto. Using this method, we were able to group the neighbourhoods by district. Refer to Figure A and Figure B for visual representations in Tableau. The other three features were features were 'host_since_length' (length of time the host has been a host of an Airbnb), 'labels_host_acceptance_rate' (categorical variable containing levels of host acceptance rate) and 'labels_host_response_rate' (categorical variable containing levels of host response rate). Lastly, the cleaned dataset was generated for Modelling. A data profile of the cleaned dataset is shown in Figure 3.

As we had chosen two targets, we initially started with standard linear regression as a form of EDA to determine the contribution of predictors to each target variable. We did find there to be significant predictors for each target. Refer to Figure 7 and Figure 8. For modelling, we began with Linear regression with a training set assigned with 70% of the cleaned data. However, due to the complexity of the dataset, the results were subpar as shown with a low R squared (0.43) and high RMSE (48) for price as a target. A random forest regression method was then implemented, as we felt the complexity of our dataset would necessitate a more sophisticated machine learning method where decision trees were present. Our results showed a better R squared (0.6) and RMSE value (44.2), however we needed to identify the best model for our data. As such, we implemented the pycaret package to identify which machine learning regression model would be best suited for prediction of our target variables. The Linear gradient boost regression (LGBM) and Gradient boost regression (GBM) were identified as the best model for price and review score ratings respectively, based on metrics of R-squared score (0.61, 0.66) and Root Mean Squared Error (RMSE) (43.9, 2.36). Thus, we implemented LGBM on price and GBM on review score ratings as targets. The results showed that there was a significant difference in model prediction for price compared to review score ratings. The R-squared and RMSE values are shown in Figure 9 and Figure 10 of the Appendix for each model. As can be seen, LGBM and GBM were most suitable for predicting Price and Review ratings score respectively.

We used RMSE for price as we cared less about accuracy and more about the range of values as there are other unknowns which contribute to pricing. The tradeoff being the model may be less accurate, but is less of an issue. For score accuracy is key with such a small range, so we chose r-squared. Adjusted R-squared could also have been used. However, there is an obvious tradeoff as there is an overall reduction in accuracy due to the number of features, evident in test/train r-squared comparison.

After the generation of both models, a plot of the most impactful features for each was also generated (Figure 3.1 & Figure 3.2). Hosts will be able to use this information to focus on

factors which will improve their revenue generation, and thereby airbnb as well. For price the most important factor was a unit's geographical location, likely due to distance from popular/business centers. To maximize profit, Airbnb company and Host Airbnb should consider increasing the price of listings in Old City Toronto and listings in districts close to Old City Toronto, conversely, they should decrease the price in districts far away from Old City Toronto.

Location is outside a current host's control, so we decided to focus on factors on which the host could control. Controllable factors such as rental availability as well as the quality and number of reviews that dictate the pricing for a rental listing. As shown from the feature importance plot (Figure 3.1), reviews per month, bedrooms and availability over 365 days were also strong predictors of price. In terms of number of bedrooms, the host could ensure that their listing or listings fit with the minimum required number of bedrooms. Certainly, a one bedroom rental unit would less likely be rented by a customer compared to a three bedroom rental unit. Airbnb should also increase the number of listings which have more than two bedrooms and minimize the number of listings with less than two bedrooms. This simple change would generate more revenue for Airbnb. The host should also consider making their listing available to multiple customers, this could be achieved by reducing the time that the listing is rented by one host which would allow the listing to be available to more customers. For Airbnb, a listing that is rented by multiple rather than a few customers would generate more revenue as the price of the listing could be increased.

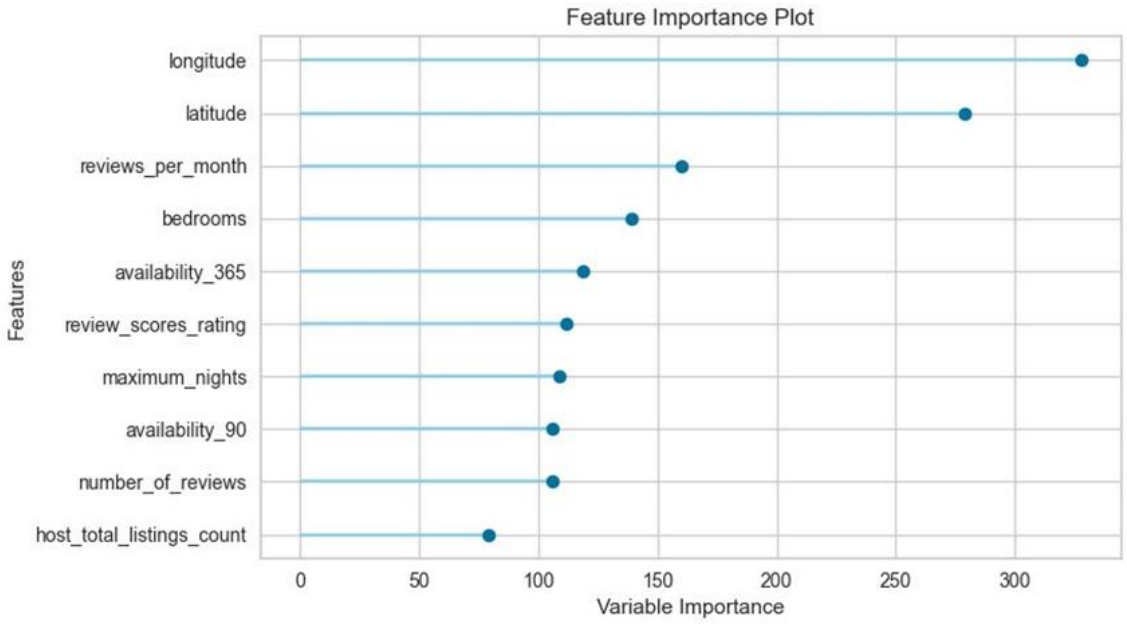

Figure 3.1. Price as Target variable

While it is simple to increase the amount of rental days, review scores are subjective. For ratings, all of the most important features were all review related (Figure 3.2). While all of the aspects are important, accuracy and cleanliness are much more so. It would thereby be recommended that hosts ensure that they are forthright with their posting and ensure that cleaning is done between visits. This should maximize their review ratings, which will thereby allow them to price higher. Hosts should also remind renters to rate them (If you enjoyed your stay, please consider reviewing us on airbnb) in order to maximize their reviews per month.

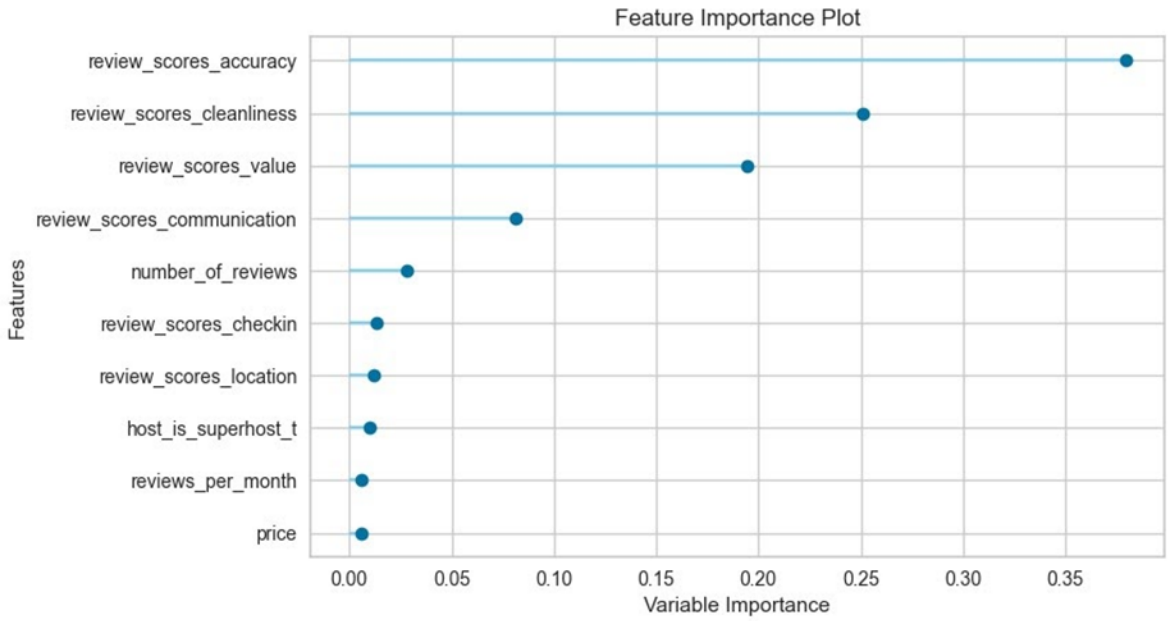

Figure 3.2. Review ratings score as Target variable

By maximizing review ratings and keeping the price at a reasonable level that matches with review rating score, higher commissions would be generated for Airbnb along with increased revenue. To maximize profit related to review score ratings, Airbnb should consider enforcing a cleanliness of listing bench score to ensure all listings score high on a score of cleanliness. Accuracy of review scores should also be ensured. A recommendation for producing a more accurate and reliable Machine learning model that incorporates prediction of both price and review score ratings would be to find Airbnb Toronto datasets that include additional relevant features such as number of parking lots, postal code,distance from subway and bus, aesthetic value, relative location to popular attractions and hotels, venues and crime levels that could be reliable and valid predictors.

4: Visualization with Dashboard

The following is the link to the Tableau Dashboard created for Toronto Airbnb:

https://public.tableau.com/app/profile/mariana6235/viz/Airbnb_Tableau/AirbnbToronro

The objective of this dashboard was to enable the Airbnb senior leadership team, and relevant stakeholders such as Airbnb hosts, to develop business and marketing strategies focused on emphasizing the features that visually demonstrate the highest positive impact on price per night and review scores.

At the top of the dashboard, the motivation for the tool has been stated. The Tableau dashboard consists of five main graphs. On the top left, there is a map that displays the regions that correspond to a former city. The points on the map correspond to the geographic location of an Airbnb listing. If the viewer hovers over a point on the map, details are shown such as the unique Airbnb ID, location and the review score and price per night which are the variables of most interest. To easily describe the following 4 graphs included in the dashboard, assume that the “Average of” filter is set to “Review Score (%)”. On the top right, there is a graph that compares the number of host listings to review scores. Along the middle, there is a bar that shows the distribution of review scores across the different room types that were taken into consideration. On the bottom left, the graph shows the relationship of the former city to the review score which accounts for location being a factor in listings.

Lastly, the bottom right graph shows the length of time that a host has been around and how that affects the review scores of customers. As previously stated, these descriptions were based on the assumption that the “Average of” filter is set to “Review Score (%)”. By using sheet hiding however, the ability to show the same graphs with a different metric is made possible. In other words, if the viewer changes the “Average of” filter to “Price per Night ($)”, the graphs yield the same comparisons and the metric of review score (%) instead changes to price per night ($).

This filter was created to focus on one of the two main factors of the analysis. The dashboard enables viewers to make decisions on the impact of certain features on price and/or review scores independently.

In general, two graphs correlate to the location of Airbnb listings which is useful to distinguish the areas that are most lucrative. Two different graphs were created to demonstrate the relationship of certain characteristics (e.g. number of host listings) of hosts to the average price per night or review score. The last graph that doesn't fall into either of these groupings was created to compare the differences in room types for Toronto Airbnb listings.

In addition to the graph visuals, the dashboard contains nine filters. Starting at the top of the dashboard, the first two filters alter the view of the Airbnb listings shown in the map by selecting the desired range for price and/or review score. Using shapes as filters, the following filter makes changes to all graphs depending on the room types selected. The "Average of" filter alters the view of all graphs to use price per night or review scores as the primary metric. The following three filters also affect all graphs and regard the location and host information. The last two of these filters are binary filters that are reflected on all graphs and represent the impact of being able to book instantly and if the host is a superhost. These filters were created with the intent that viewers can manipulate various features and visually see the effect of the desired metric specified. Although not a filter, this interactive dashboard also contains a link when the bottom left corner logo is clicked on. Adding this link is convenient to direct viewers to current Airbnb Toronto listings.

5: Recommendations and Conclusion

Airbnb, an organization that generates billions in annual revenue, has the opportunity for significant growth as restrictions related to the COVID-19 pandemic are lifted and the industry experiences a travel rebound.[1]The ability to increase price and improve review scores will lead to additional commission revenue for Airbnb, with a target increase of 2% in 2022.

In relation to increasing price, our modelling shows that three main features have the largest impact. These features are: number of reviews a listing has, the ability to book a property instantly and the location of the property. As seen in figure 5.1, data shows that the more reviews a property has, the higher the average price.

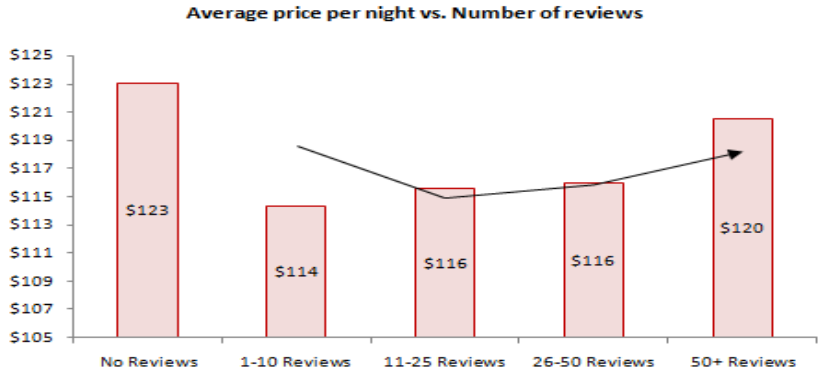

Figure 5.1

There are various strategies Airbnb can implement to increase the likelihood a renter will leave a review. A number of experts believe "...the best way to encourage customers to write reviews is to just ask."[2] Indeed, weaving requests for reviews into various existing customer communications, personalizing the communications and outlining to customers the value of leaving reviews are all strategies that can be considered. Offering the right incentives for customers to leave reviews can be another powerful tool to increase overall reviews. Incentives can include things like; a loyalty discount or providing advanced insights into new listings or deals.

The location and the ability to book a property instantly each have a high impact on the overall price. Figure 5.2 outlines the differentials in price per night based on the location of the property. More central locations yield a higher price per night due to the proximity of desirable amenities and attractions.

Figure 5.2

An intelligent algorithm that provides in-the-moment pricing recommendations to hosts as well as the option for flexible pricing, where the price is automatically adjusted based on the booking window, could increase revenue. Equipping hosts with the knowledge, through this system-generated algorithm, will allow hosts to be empowered to make smart business decisions at the moment which will improve and maximize their pricing strategy.

As seen below in figure 5.3, there is a direct relationship with the price per night and both types of property and the ability to book the property instantly. Booking a full house/apartment, for example, shows booking instantly is not an advantage, renters pay more for booking in advance whereas for every other property type the opposite is true. This is likely a result of an entire house/apartment rental for an endeavour requiring pre-planning like travelling with an entire

family or group of friends. On the other hand, renters requiring smaller accommodations are likely travelling in smaller groups and will pay more for last minute bookings.

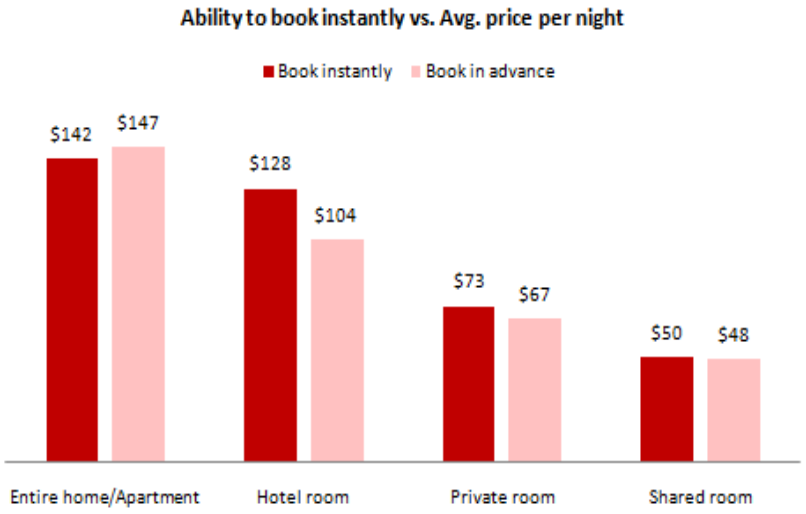

Figure 5.3

In relation to improving review scores, our modelling shows three main features which will have the largest impact. These features are: the accuracy of the listing, the location of the property, and the cleanliness. As depicted in figure 5.4, data shows the higher the accuracy rating, the higher the review score.

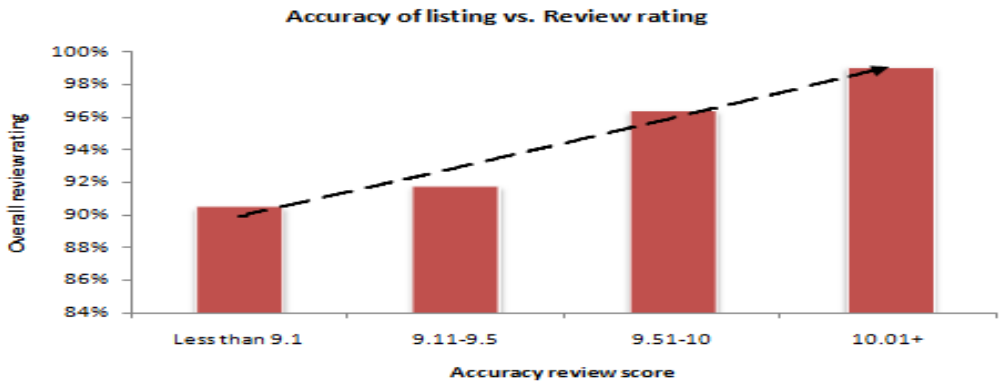

Figure 5.4

Renters can feel deceived if the listing details are not an accurate representation of the property. To address this, Airbnb can implement an internal audit process which identifies

individual listings, and host trending, where properties are receiving low accuracy scores. Through this process Airbnb can mandate an improvement plan, and even require corrective action, to improve the listing accuracy.

Data shows that the cleanliness of the property has a high impact on the overall review score. To market the importance of cleanliness to hosts Airbnb should implement a cleanliness bench score that requires host adherence.

Collecting additional information would be immensely helpful in strengthening this predictive model going forward. An indication of popular tourist sites in proximity to the property, understanding why people are visiting, the livability rating of the area and buildings, the distance from the building to public transportation and parking could all be features that would add value. Additional feature engineering to identify days of the week, and seasons throughout the year of travel could also be helpful in understanding booking trends. Broadening the geographic region and implementing the model in all areas will be the best way for Airbnb to move forward in implementing holistic solutions.

6: References

- 1. https://news.Airbnb.com/Airbnb-fourth-quarter-and-full-year-2020-financial-results/

Airbnb. (2021, February 25). *Airbnb fourth quarter and full year 2020 financial results*. Airbnb Newsroom. Retrieved from https://news.airbnb.com/airbnb-fourth-quarter-and-full-year-2020-financial-results/.

- 2. https://databox.com/how-to-encourage-customers-to-write-reviews

Greene, J. (2021, March 25). *14 proven ways to encourage customers to write reviews*. Databox. Retrieved from https://databox.com/how-to-encourage-customers-to-write-reviews.

- 3. https://www.proquest.com/docview/2582650420?parentSessionId=WjZCm4uRVhBSvV3WgukCh0%2BsK5nj02S%2FPTUiiXNzLXo%3D&pq-origsite=summon&accountid=3455

Tian, Z. (2021). Use python data analysis to gain insights from airbnb hosts. *Advances in Mathematical Physics, 2021* doi:http://dx.doi.org/10.1155/2021/1079850

- 4. https://www.proquest.com/docview/2526827719?parentSessionId=4%2BUqlfBUGYXp6Ce8RLqqwOwsb38BKmU8YYLW%2Fu7a33Y%3D&pq-origsite=summon&accountid=3455

Meijerink, J., & Schoenmakers, E. (2021). Why are online reviews in the sharing economy skewed toward positive ratings? linking customer perceptions of service quality to leaving a review of an airbnb stay. *Journal of Tourism Futures, 7*(1), 5-19. doi:http://dx.doi.org/10.1108/JTF-04-2019-0039

- 5. https://venturebeat.com/2015/06/30/how-we-scaled-data-science-to-all-sides-of-Airbnb-over-5-years-of-hypergrowth/

Riley Newman, A. (2015, June 30). *How we scaled data science to all sides of Airbnb over 5 years of hypergrowth*. VentureBeat. Retrieved from https://venturebeat.com/2015/06/30/how-we-scaled-data-science-to-all-sides-of-airbnb-over-5-years-of-hypergrowth/.

- 6. http://insideAirbnb.com/get-the-data.html
- 7. https://www.kaggle.com/search?q=Airbnb
- 8. https://data.world/datasets/Airbnb
- 9. https://www.Airbnb.com/help/article/2855/privacy-policy

Airbnb. (2020, October 30). *Privacy policy - airbnb help centre*. Airbnb. Retrieved from https://www.airbnb.ca/help/article/2855/privacy-policy?locale=en&_set_bev_on_new_domain=1639220868_NGZkZDA4MzIyN2Iy.

7: Appendix

Link to Tableau Dashboard:

https://public.tableau.com/app/profile/mariana6235/viz/Airbnb_Tableau/AirbnbToronro

Link to Github code:

https://github.com/Datasciencecap101/TorontoAirbnb_Machine-Learning-Code.git

Gmail Account:

sheridancapstone2021@gmail.com

Datascience101

Github Account:

sheridancapstone2021@gmail.com

pw: Datascience101

Username:Datasciencecap101

Figure A. Listings by Neighborhood in Toronto as displayed in Tableau from Original Dataset

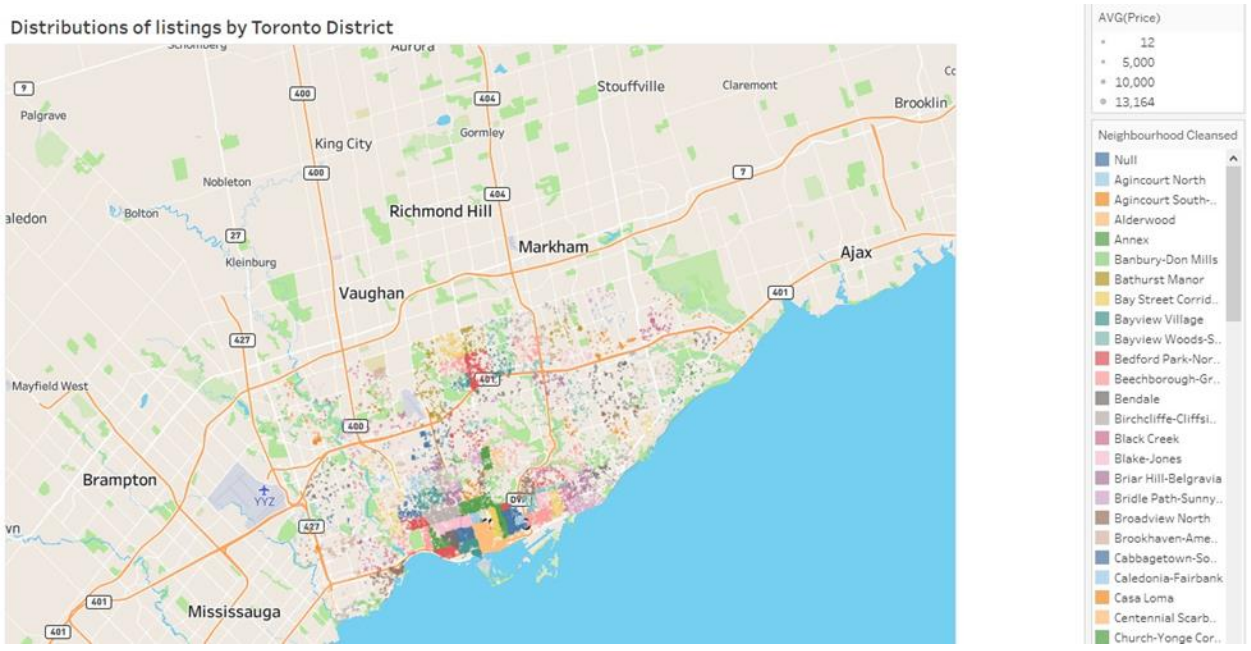

Figure B. Listings by District in Toronto as displayed in Tableau from Original Dataset

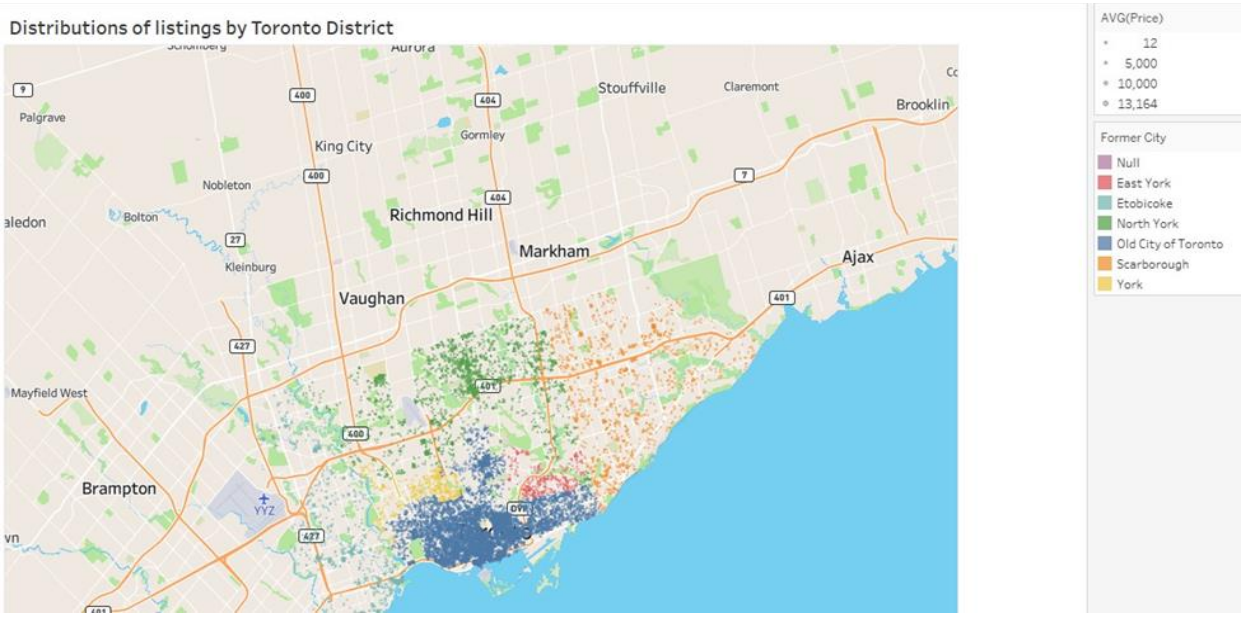

Table 1. Data dictionary showing the fields or features in the AirBnb Data Set with description and the type of feature is stated.

In total there were 73 features in the original dataset that were not cleaned. A colour scheme was implemented to show distribution of features in the unclean or original dataset that were irrelevant, had more than 40% missing values, were target features, contained redundant data or were relevant features.

Legend:

Orange-Irrelevant features

Blue-Features with more than 40% NaN values

Red-Target features

Dark Magenta-Redundant features

Purple-Relevant features

Number of Variables	Column Name	Description	Feature type(e.g., Numeric, String)
	id	Unique listing id (Primary Key)	Numeric
1	listing_url	Link to the rental property listing on Airbnb	String
2	scrape_id	Identifier for scraper	Numeric
3	last_scraped	Last date listing was scraped	Numeric
4	name	Title of Posting	String
5	description	Description of Posting	String
6	neighborhood_overview	Overview of neighborhood	String
7	picture_url	Link to the main vacation rental listing image on Airbnb	Image
8	host_id	Unique id for each host	Numeric
9	host_url	Link to the host profile on Airbnb	String
10	host_name	Host Name	String
11	host_since	Date an individual became a host	Numeric
12	host_location	Location of Listing	String
13	host_about	Description of host - relationship status, interests and hobbies	String
14	*host_response_time*	Time to respond to customer booking inquiry; ranges from 1hour to few days	Numeric

15	*host_response_rate*	Ranges from 0% to 100% for reply to booking inquiries	Numeric
16	*host_acceptance_rate*	Ranges from 0% to 100% response for acceptance of booking	Numeric
17	host_is_superhost	Is either t(true) or f(false)	Boolean
18	host_thumbnail_url	Host thumbnail URL	Image
19	host_picture_url	Host Picture URL	Image
20	host_neighbourhood	Description of neighborhood listing	String
21	host_listings_count	Current number of host listings	Numeric
22	host_total_listings_count	Total number of listings made by host	Numeric
23	host_verifications	Identifies how the host has completed the identity verification process	String
24	host_has_profile_pic	Host Profile Pic	Image
25	host_identity_verified	Identifies if the host has completed the verification process by indicating true or false	Boolean
26	neighbourhood	Specific location of Toronto area	String
27	neighbourhood_cleansed	Represents one of boroughs in Toronto in which a listing resides	String
28	neighbourhood_group_cleansed	No values (N/A)	N/A
29	latitude	The angular distance of a location or object north or south of the Earth's celestial equator	String
30	longitude	The angular distance of a location or object east or west of the meridian	String

31	property_type	Type of property (e.g., Entire house)	String
32	room_type	Specific type of room (e.g., Entire home/apt)	String
33	accommodates	How many people can stay (e.g., 6)	Numeric
34	bathrooms	No values, NA	N/A
35	bathrooms_text	Number of bathrooms in property	Numeric
36	bedrooms	Number of bedrooms in property	Numeric
37	beds	Number of beds in property	Numeric
38	amenities	List of amenities such as shampoo available in property	String
39	*price*	Price of stay per night	Numeric
40	minimum_nights	Minimum nights can be booked by same individual	Numeric
41	maximum_nights	Maximum nights can be booked by same individual	Numeric
42	minimum_minimum_nights	Same values as Minimum nights	Numeric
43	maximum_minimum_nights	Same values as Maximum_nights	Numeric
44	minimum_maximum_nights	Same values as Maximum_nights	Numeric
45	maximum_maximum_nights	Same values as Maximum_nights	Numeric
46	minimum_nights_avg_ntm	Average minimum nights can be booked by same individual	Numeric
47	maximum_nights_avg_ntm	Average maximum nights can be booked by same individual	Numeric
48	calendar_updated	No values (N/A)	N/A

49	has_availability	True or False if listing is available	Boolean
50	availability_30	Availability of Property in 30 days	Numeric
51	availability_60	Availability of Property in 60 days	Numeric
52	availability_90	Availability of Property in 90 days	Numeric
53	availability_365	Availability of Property in 365 days	Numeric
54	calendar_last_scraped	Date last scraped	Numeric
55	number_of_reviews	Total number of reviews that a listing has received from customers	Numeric
56	number_of_reviews_ltm	The number of reviews that a listing has received last twelve month	Numeric
57	number_of_reviews_l30d	The number of reviews that a listing has received per 130 days	Numeric
58	first_review	Date of first review by customer	Numeric
59	last_review	Date of last review by customer	Numeric
60	***review_scores_rating***	***Customer-provided score rating (0% to 100%); A customer-provided review score attributed to a listing*** ***based on overall experience and satisfaction***	Numeric
61	review_scores_accuracy	Accuracy of review scores (0 to 10)	Numeric
62	review_scores_cleanliness	Cleanliness score (0 to 10)	Numeric
63	review_scores_checkin	Over-all check in score (0 to 10)	Numeric
64	review_scores_communication	Score on communication with host (0 to 10)	Numeric
65	***review_scores_location***	***Score on location based on factors such as nearby transportation, noise level (0 to 10)***	Numeric

66	review_scores_value	Over- all value/quality/experience (0 to 10)	Numeric
67	license	No values (N/A)	N/A
68	instant_bookable	True or False if customer can instantly book	BOolean
69	calculated_host_listings_count	Calculated number of listings by host	Numeric
70	calculated_host_listings_count_entire_homes	Number of host listings which are entire homes	Numeric
71	calculated_host_listings_count_private_rooms	Number of host listings which are private rooms	Numeric
72	calculated_host_listings_count_shared_rooms	Number of host listings which are shared homes	Numeric
73	reviews_per_month	Number of Customer reviews of the host accommodation or accommodations per month	Numeric

Figure 1. Correlations between numerical features and target variable being price

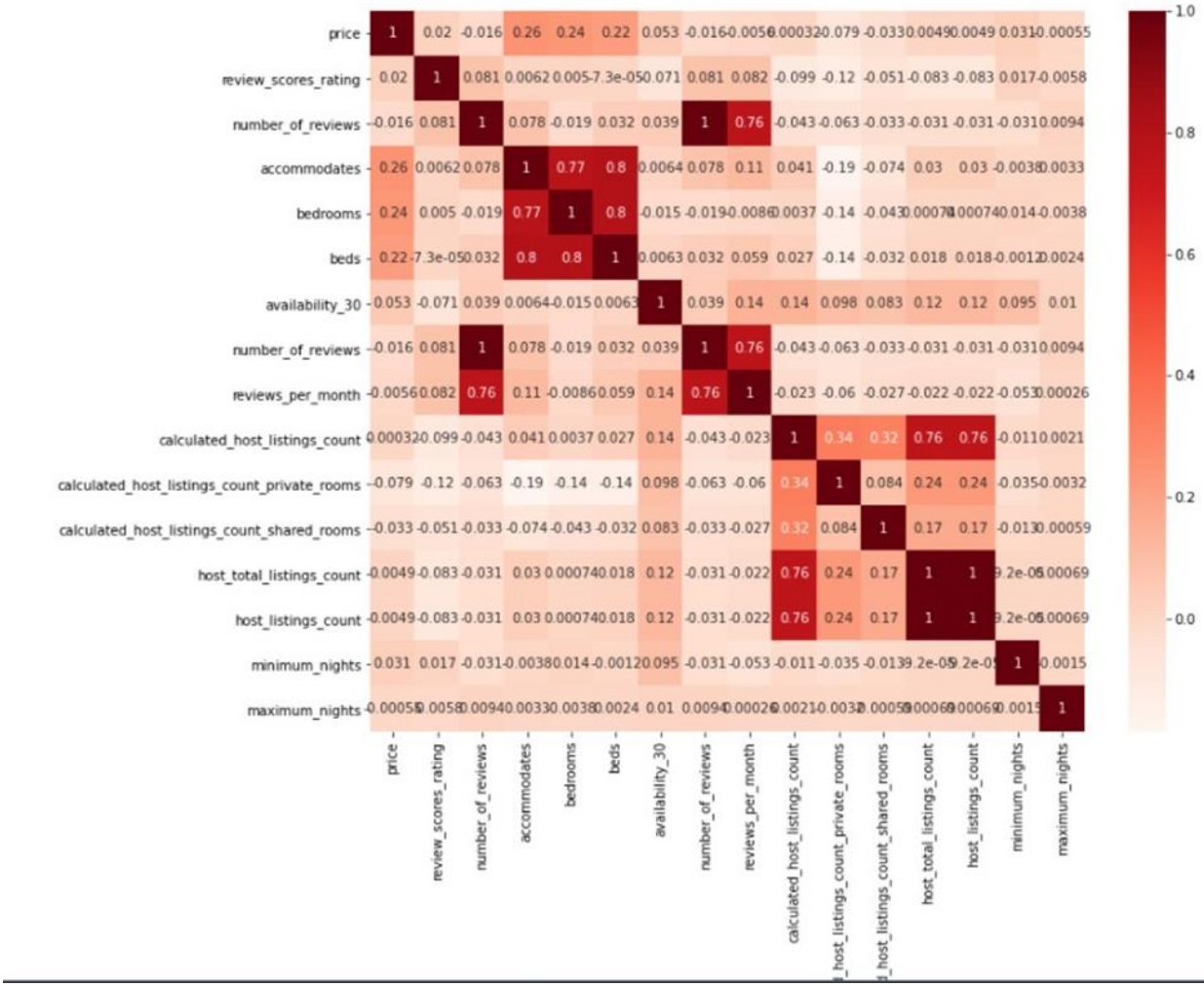

Figure 2. Sample EDA of Original Dataset

Figure 3. Feature dataset profile of the Cleaned dataset that was used for Modeling

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 19343 entries, 0 to 19342
Data columns (total 41 columns):
 #   Column                                        Non-Null Count  Dtype
---  ------                                        --------------  -----
 0   bedrooms                                      19343 non-null  float64
 1   beds                                          19343 non-null  float64
 2   host_total_listings_count                     19343 non-null  float64
 3   review_scores_rating                          19343 non-null  float64
 4   review_scores_accuracy                        19343 non-null  float64
 5   review_scores_cleanliness                     19343 non-null  float64
 6   review_scores_checkin                         19343 non-null  float64
 7   review_scores_communication                   19343 non-null  float64
 8   review_scores_location                        19343 non-null  float64
 9   review_scores_value                           19343 non-null  float64
 10  reviews_per_month                             19343 non-null  float64
 11  neighbourhood_cleansed                        19343 non-null  object
 12  latitude                                      19343 non-null  float64
 13  longitude                                     19343 non-null  float64
 14  accommodates                                  19343 non-null  int64
 15  price                                         19343 non-null  float64
 16  minimum_nights                                19343 non-null  int64
 17  maximum_nights                                19343 non-null  int64
 18  availability_30                               19343 non-null  int64
 19  availability_60                               19343 non-null  int64
 20  availability_90                               19343 non-null  int64
 21  availability_365                              19343 non-null  int64
 22  number_of_reviews                             19343 non-null  int64
 23  number_of_reviews_ltm                         19343 non-null  int64
 24  number_of_reviews_l30d                        19343 non-null  int64
 25  instant_bookable                              19343 non-null  object
 26  calculated_host_listings_count                19343 non-null  int64
 27  calculated_host_listings_count_entire_homes   19343 non-null  int64
 28  calculated_host_listings_count_private_rooms  19343 non-null  int64
 29  calculated_host_listings_count_shared_rooms   19343 non-null  int64
 30  former_city                                   19343 non-null  object
 31  room_type                                     19343 non-null  object
 32  host_since                                    19343 non-null  object
 33  host_response_time                            19343 non-null  object
 34  host_is_superhost                             19343 non-null  object
 35  host_acceptance_rate                          19343 non-null  int64
 36  host_response_rate                            19343 non-null  int64
 37  labels_host_acceptance_rate                   19343 non-null  category
 38  labels_host_response_rate                     19343 non-null  category
 39  host_since_year                               19343 non-null  int64
 40  host_length                                   19343 non-null  int64
dtypes: category(2), float64(14), int64(18), object(7)
memory usage: 5.8+ MB
```

Figure 4. Correlation for Price target feature

```
price                                          1.000000
accommodates                                   0.595983
bedrooms                                       0.540802
beds                                           0.493700
calculated_host_listings_count_private_rooms   0.412672
calculated_host_listings_count_entire_homes    0.339368
latitude                                       0.275433
review_scores_location                         0.147562
availability_30                                0.092657
host_since_year                                0.078497
host_length                                    0.078497
review_scores_rating                           0.076440
review_scores_cleanliness                      0.072589
availability_60                                0.068199
maximum_nights                                 0.059920
availability_90                                0.055047
calculated_host_listings_count                 0.035810
availability_365                               0.033946
review_scores_checkin                          0.033755
reviews_per_month                              0.032303
review_scores_accuracy                         0.028995
host_acceptance_rate                           0.024670
minimum_nights                                 0.020220
number_of_reviews                              0.013625
review_scores_communication                    0.009418
review_scores_value                            0.008806
longitude                                      0.005433
number_of_reviews_ltm                          0.001903
host_total_listings_count                      0.000012
number_of_reviews_l30d                              NaN
calculated_host_listings_count_shared_rooms         NaN
host_response_rate                                  NaN
Name: price, dtype: float64
```

Figure 5. EDA of Cleaned dataset. Beds, Price by Toronto District

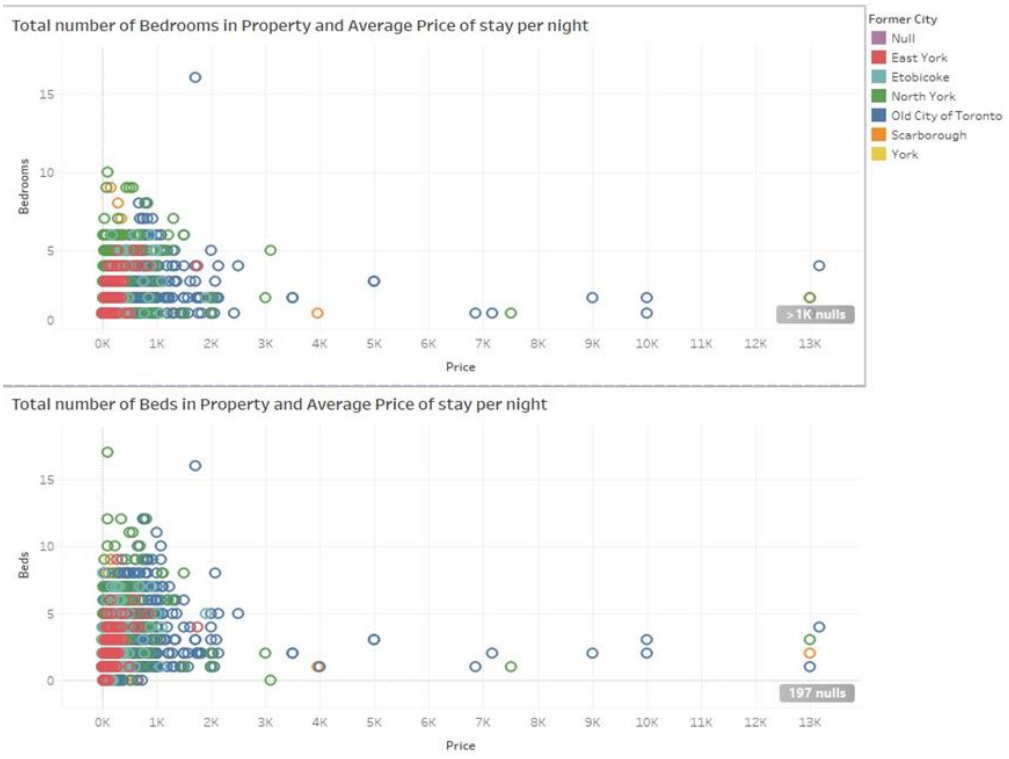

Figure 6. Room type and Average price per day of listing by District in Toronto

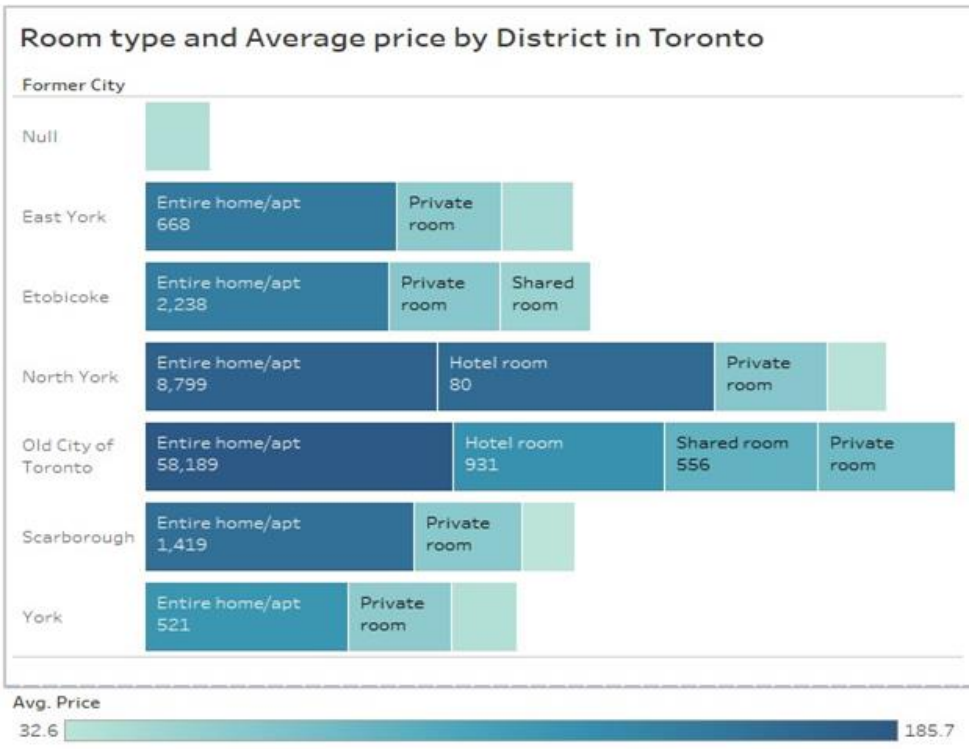

Figure 7. Output from Standard Linear regression with target being price

OLS Regression Results

Dep. Variable:	price	R-squared:	0.532
Model:	OLS	Adj. R-squared:	0.531
Method:	Least Squares	F-statistic:	509.3
Date:	Wed, 17 Nov 2021	Prob (F-statistic):	0.00
Time:	20:48:00	Log-Likelihood:	-1.0255e+05
No. Observations:	19343	AIC:	2.052e+05
Df Residuals:	19299	BIC:	2.055e+05
Df Model:	43		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.663e+04	1059.930	15.688	0.000	1.46e+04	1.87e+04
longitude	82.3151	9.308	8.844	0.000	64.071	100.559
latitude	-233.2391	15.835	-14.729	0.000	-264.277	-202.201
bedrooms	29.1728	0.931	31.343	0.000	27.348	30.997
beds	-3.4195	0.779	-4.387	0.000	-4.947	-1.892
host_total_listings_count	1.8417	0.278	6.631	0.000	1.297	2.386
review_scores_rating	1.2362	0.145	8.551	0.000	0.953	1.520
review_scores_accuracy	1.1270	1.394	0.808	0.419	-1.606	3.860
review_scores_cleanliness	4.7269	0.744	6.356	0.000	3.269	6.185
review_scores_checkin	-12.3456	1.990	-6.204	0.000	-16.246	-8.445
review_scores_communication	-2.9191	2.289	-1.275	0.202	-7.407	1.568
review_scores_location	13.6364	1.522	8.960	0.000	10.653	16.620
review_scores_value	-5.7392	0.808	-7.100	0.000	-7.324	-4.155
reviews_per_month	2.2493	0.647	3.476	0.001	0.981	3.518
accommodates	11.4470	0.423	27.082	0.000	10.618	12.275
host_acceptance_rate	0.3453	0.076	4.542	0.000	0.196	0.494
host_length	0.4061	0.167	2.437	0.015	0.079	0.733
instant_bookable_t	-2.0346	0.797	-2.552	0.011	-3.598	-0.472
former_city_Etobicoke	8.5782	3.570	2.403	0.016	1.581	15.576
former_city_North York	23.4454	3.097	7.572	0.000	17.376	29.515
former_city_Old City of Toronto	20.4164	2.803	7.283	0.000	14.921	25.911
former_city_Scarborough	3.1213	3.197	0.976	0.329	-3.144	9.387
former_city_York	6.1394	3.529	1.740	0.082	-0.778	13.057
room_type_Hotel room	5.4936	6.379	0.861	0.389	-7.009	17.997
room_type_Private room	-35.5520	1.495	-23.776	0.000	-38.483	-32.621
room_type_Shared room	-51.9939	2.971	-17.500	0.000	-57.817	-46.170

Figure 8. Standard linear regression with target as 'review score ratings'

OLS Regression Results

Dep. Variable:	review_scores_rating	R-squared:	0.644
Model:	OLS	Adj. R-squared:	0.643
Method:	Least Squares	F-statistic:	811.1
Date:	Sun, 21 Nov 2021	Prob (F-statistic):	0.00
Time:	21:39:57	Log-Likelihood:	-44489.
No. Observations:	19343	AIC:	8.907e+04
Df Residuals:	19299	BIC:	8.941e+04
Df Model:	43		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	106.4247	53.004	2.008	0.045	2.531	210.318
longitude	0.3017	0.463	0.651	0.515	-0.607	1.210
latitude	-2.1074	0.791	-2.664	0.008	-3.658	-0.557
bedrooms	-0.0135	0.047	-0.284	0.777	-0.106	0.079
beds	0.0036	0.039	0.094	0.925	-0.072	0.080
host_total_listings_count	-0.0386	0.014	-2.797	0.005	-0.066	-0.012
review_scores_accuracy	2.5273	0.067	37.797	0.000	2.396	2.658
review_scores_cleanliness	1.7785	0.035	51.237	0.000	1.710	1.847
review_scores_checkin	1.2208	0.099	12.381	0.000	1.028	1.414
review_scores_communication	2.6206	0.112	23.355	0.000	2.401	2.841
review_scores_location	1.0985	0.075	14.573	0.000	0.951	1.246
calculated_host_listings_count_entire_homes	0.0044	0.025	0.175	0.861	-0.045	0.054
calculated_host_listings_count_private_rooms	-0.0336	0.048	-0.701	0.483	-0.128	0.060
calculated_host_listings_count_shared_rooms	-2.407e-16	8.76e-17	-2.747	0.006	-4.12e-16	-6.89e-17
host_acceptance_rate	-0.0019	0.004	-0.495	0.620	-0.009	0.006
host_length	0.0086	0.008	1.041	0.298	-0.008	0.025
instant_bookable_t	0.0129	0.040	0.325	0.745	-0.065	0.091
former_city_Etobicoke	0.0568	0.177	0.320	0.749	-0.291	0.405
former_city_North York	0.2630	0.154	1.707	0.088	-0.039	0.565
former_city_Old City of Toronto	0.0384	0.140	0.275	0.783	-0.235	0.312
former_city_Scarborough	0.1517	0.159	0.955	0.340	-0.160	0.463
former_city_York	0.2510	0.175	1.431	0.152	-0.093	0.595

Table 2. Training set metrics for models

70% training set	R-squared	RMSE
Price_Linear Regression	0.53	48
Review score ratings_Linear Regression	0.64	2

Price_Random Forest Regression	0.6	44.2
Review score ratings_Random Forest Regression	0.65	2.43
Price_LGBM	**0.61**	**43.9**
Review score ratings_GBM	**0.66**	**2.36**

Figure 9.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
gbr	Gradient Boosting Regressor	1.4718	5.5544	2.3551	0.6627	0.0248	0.0156	0.4440
lightgbm	Light Gradient Boosting Machine	1.4543	5.6000	2.3648	0.6599	0.0249	0.0154	0.0530
br	Bayesian Ridge	1.6357	5.8661	2.4207	0.6437	0.0255	0.0174	0.1230
ridge	Ridge Regression	1.6354	5.8700	2.4215	0.6435	0.0256	0.0174	0.0170
lr	Linear Regression	1.6356	5.8706	2.4216	0.6434	0.0256	0.0174	0.0500
omp	Orthogonal Matching Pursuit	1.6158	5.8721	2.4220	0.6433	0.0256	0.0172	0.2360
rf	Random Forest Regressor	1.4686	5.8923	2.4258	0.6420	0.0256	0.0156	1.2090
et	Extra Trees Regressor	1.5153	6.6061	2.5680	0.5986	0.0271	0.0161	1.3480
ada	AdaBoost Regressor	2.0287	7.7024	2.7744	0.5319	0.0291	0.0214	0.2250
en	Elastic Net	2.5643	10.4956	3.2389	0.3625	0.0342	0.0272	0.0240
lasso	Lasso Regression	2.6037	10.7556	3.2789	0.3467	0.0346	0.0277	0.0220
dt	Decision Tree Regressor	1.8296	11.3538	3.3676	0.3100	0.0355	0.0194	0.0490
huber	Huber Regressor	2.6800	11.7639	3.4292	0.2853	0.0362	0.0285	0.7370
llar	Lasso Least Angle Regression	3.2357	16.4627	4.0569	-0.0002	0.0428	0.0345	0.2700
dummy	Dummy Regressor	3.2357	16.4627	4.0569	-0.0002	0.0428	0.0345	0.0180
knn	K Neighbors Regressor	3.2473	17.5998	4.1945	-0.0693	0.0442	0.0346	0.0730
par	Passive Aggressive Regressor	3.5355	20.9040	4.3938	-0.2633	0.0459	0.0372	0.0580

Figure 10.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
lightgbm	Light Gradient Boosting Machine	31.4060	1929.9485	43.9074	0.6199	0.3567	0.3129	0.0690
rf	Random Forest Regressor	32.2055	2027.4467	45.0038	0.6006	0.3645	0.3223	1.2620
gbr	Gradient Boosting Regressor	33.3540	2108.1634	45.8943	0.5848	0.3776	0.3377	0.4250
et	Extra Trees Regressor	32.5954	2165.4108	46.5085	0.5735	0.3744	0.3250	1.4540
lr	Linear Regression	36.0115	2366.6580	48.6258	0.5340	0.4371	0.3769	0.2890
ridge	Ridge Regression	35.9928	2367.1384	48.6306	0.5339	0.4337	0.3761	0.0210
br	Bayesian Ridge	36.0759	2382.4473	48.7876	0.5309	0.4300	0.3757	0.1350
omp	Orthogonal Matching Pursuit	36.8443	2497.4200	49.9512	0.5083	0.4290	0.3869	0.2380
lasso	Lasso Regression	37.5344	2563.1525	50.6071	0.4954	0.4301	0.3998	0.2840
en	Elastic Net	41.0248	2878.7650	53.6444	0.4332	0.4714	0.4715	0.0910
huber	Huber Regressor	39.2292	2924.5352	54.0645	0.4242	0.4641	0.3890	0.6670
ada	AdaBoost Regressor	51.5616	3523.7673	59.3484	0.3061	0.5769	0.6938	0.3750
dt	Decision Tree Regressor	42.7730	4046.5708	63.5813	0.2023	0.4991	0.4187	0.0520
knn	K Neighbors Regressor	51.7494	4584.2339	67.6944	0.0975	0.5855	0.5938	0.0800
llar	Lasso Least Angle Regression	57.0030	5082.7976	71.2876	-0.0004	0.6439	0.7275	0.2960
dummy	Dummy Regressor	57.0030	5082.7975	71.2876	-0.0004	0.6439	0.7275	0.0220
par	Passive Aggressive Regressor	61.4767	6126.6620	76.5601	-0.1999	0.7654	0.7576	0.0400

Figure 11. Listings by Neighborhood in Toronto as displayed in Tableau Dashboard using Cleaned Dataset