How Airbnb guests rates values of a listing in Seattle?
What matters to Airbnb guests?

Table of Contents:

Problem statement:

“Price is what you pay,
Value is what you got.” - Warren Buffett

Pricing strategies without factoring in customers’ values are called penetrating which means setting a low price and leaving the value in the hand of the customers, and shutting off margins from your competitors. This is what I observe with Airbnb recent dynamic pricing or smart pricing. Repeating notifications sent to ask me to lower my price to beat my neighbors from Airbnb are a signal of a change in the so-called sharing economy.

At a certain point, Airbnb is no longer a sharing economy platform but a marketing tools for big rental management companies. This set many legal feuds in different cities across the globe. How does Airbnb handle that? How does all of this affect Airbnb hosts and Airbnb guests?

Legal restrictions are now a factor in how Airbnb hosts operate. A winning Airbnb pricing strategy fascinates people because that could help hosts maximize the incomes. Many startups focus on that using machine learning, and algorithms, and they certainly form a hotel dynamic pricing system for Airbnb hosts. Airbnb themselves has their dynamic pricing team. However, in my opinion, all these pricing strategies undermine the importance of customers’ value.

In this project, I want to dig into what determines value of a listing to Airbnb guests. What are the most important features of a listing that they are looking for? Whether it is price, superhost, locations, description?

Understanding customers will be valuable to both Airbnb hosts and Airbnb themselves to strategize their marketing campaigns and their pricing methods to not only maximize profits, but to maintain the legacy of the sharing-economy.

Due to limited time, I cut out the word processing analysis for a more expanding project in the future. This project focuses on the number and categorical features of Airbnb data obtained from Kaggle.

We will try to answer: how to predict review_scores_value of a listing, and what are the most important features in determining that.

Exploratory data analysis Part 1: *Descriptive*

Airbnb Seattle dataset is from Kaggle.com. There are three csv files: listings, calendar and reviews. Listings have informations of each Airbnb listing in Seattle. Calendar have the available and unavailable information through the year of the listings. Reviews have the guests comments for the listings that stayed and reviewed.
Inspect all 3 files using:

a. .head() to see the first rows for the dataset to familiarize the with data structure.
b. .info() to have an overview of all the columns and inspect for missing values, how many rows are in the dataset.

listings.csv: has 3818 entries with 92 columns or features (which we will trim down in data wrangling)
calendars.csv: has 17466710 entries with 4 columns (we will use this in a short time series analysis)
reviews.csv: has 84849 entries with 6 columns with comments being the most important column that we will not analyze in this project

Data Wrangling:
calendars.csv: convert price into float by first removing $ sign then .astype to float64
listings.csv:

A. Trim down to 31 features using educated guess that these are the most important: 'name', 'summary', 'space', 'description', 'host_response_time', 'host_response_rate', 'host_acceptance_rate', 'host_is_superhost', 'neighbourhood_group_cleansed', 'property_type', 'room_type', 'accommodates', 'bathrooms', 'bedrooms', 'beds', 'bed_type', 'amenities', 'price', 'cleaning_fee', 'extra_people','minimum_nights', 'maximum_nights', 'instant_bookable', 'cancellation_policy', 'review_scores_rating', 'review_scores_accuracy', 'review_scores_cleanliness', 'review_scores_checkin', 'review_scores_communication', 'review_scores_location', 'review_scores_value'.
B. Get lengths of description features, and drop them: 'name', 'summary', 'space', 'description', 'amenities'
C. Convert host_response_time to scale, host_response_rate to float, host_is_superhost t/f to binary, bed_type to binary, property_type to 4 most common type and other categories. Then fill missing values with forward fill.
D. Convert price, extra_people, cleaning_fee from $ to float
E. Delete all entries without review_score_value which is our interest that we are predicting.

Exploratory data analysis Part 2: *Inferential Stats*
listings.csv

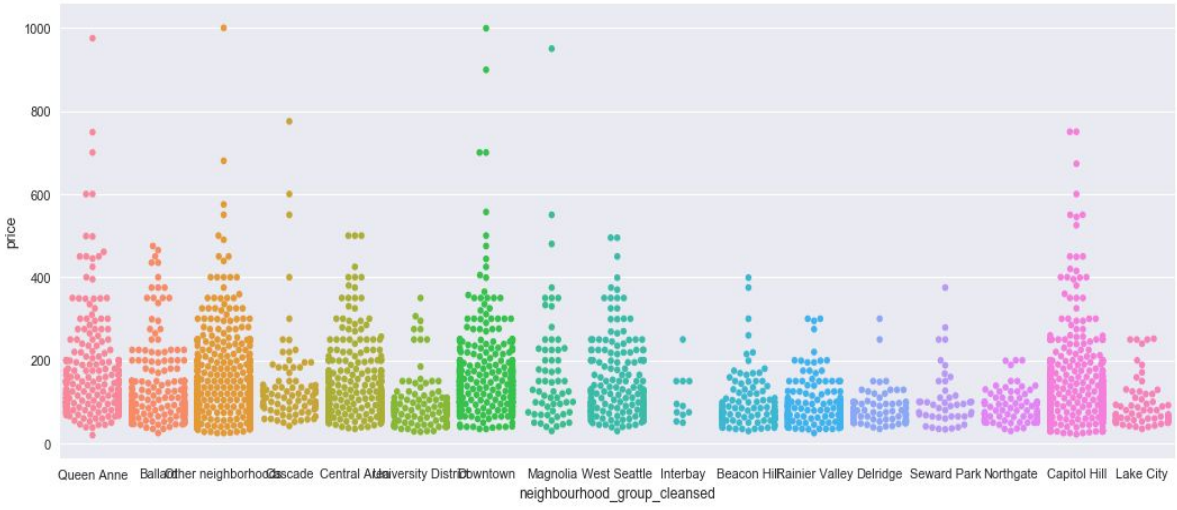
1. The highest count of a host listing is **502**. For a host to have 502 listings meaning they must be running a business in Seattle.

2. The licence requires column is all null. Seattle doesn't require a license for airbnb. Businesses and individuals alike can list their properties or spare rooms for Seattle with no restrictions.

3. The most counted listing property type is house (1733 listings) and apartment (1708 listings). They are other very interesting property type such as boat, RVs, tent and yurt. This shows a new trend for outdoor rentals. For data wrangling purpose: we use the top 4 categories: house, apartment, townhouse, and condominium, and the 5th as others.

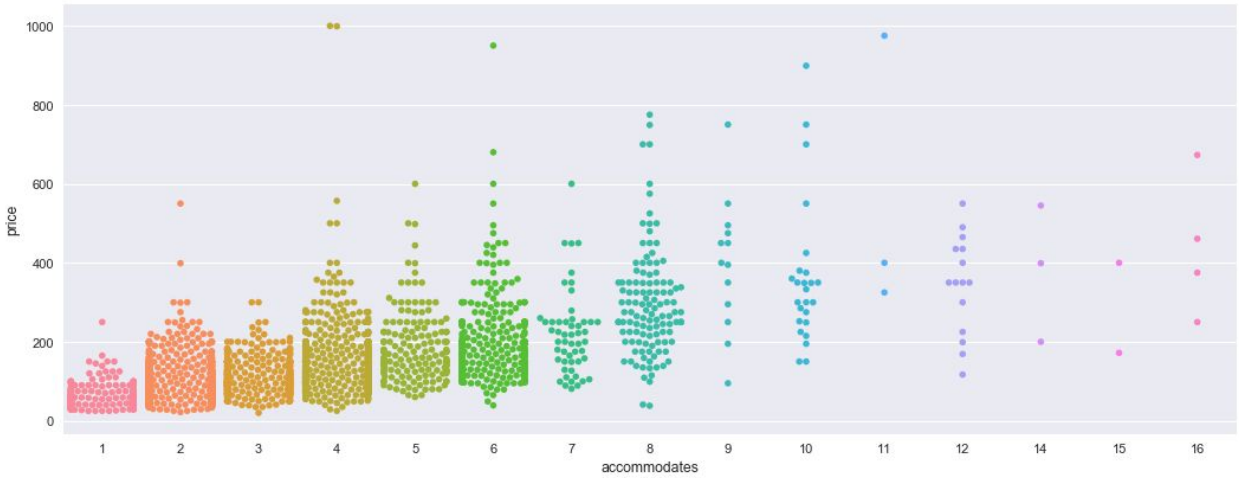
4. The highest counted bedroom size is 1 (2417 listings), and the highest counted for number that accommodates 2 people (1627 listings). We can interpret this as Airbnb is mainly used for individual or couple vacations or that most people rent out an extra room in their apartments.

5. There are 81 neighborhoods, then get grouped into 17 neighborhood groups. The last swarmplot is of neighborhood groups. We could notice that the higher priced listings are also in the more popular neighborhoods. The most popular and higher priced neighborhoods are: Capitol Hill, Downtown, Central and Queen Anne. Later we will notice this in the features of our model.

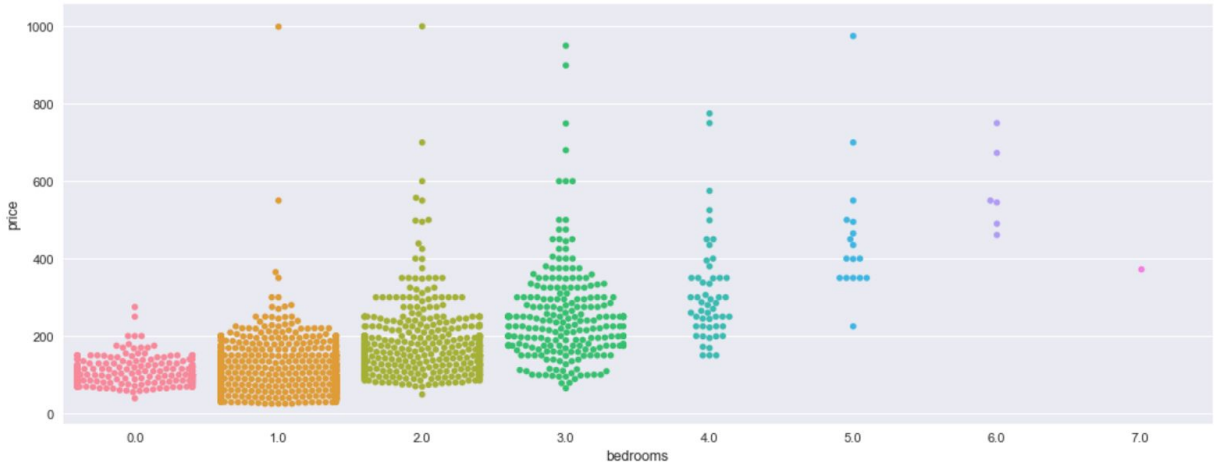


6. This swarmplot shows the prices of listings that accommodates from 1 to 16 people. Interesting pattern to notice:

1. The even number are priced higher than the odd number.
2. Listing for more than 9 people are very little. There is not market for that?
3. The optimal number for accommodates is 2 to 6.



7. There are some similarities between the prices of listings that have 0-3 bedrooms and accommodate 1-6 people. The pattern after accommodates 7 and has more than 3 bedrooms has little data points to analyze. This could be due to lack of properties (supply) or lack of renters (demand)

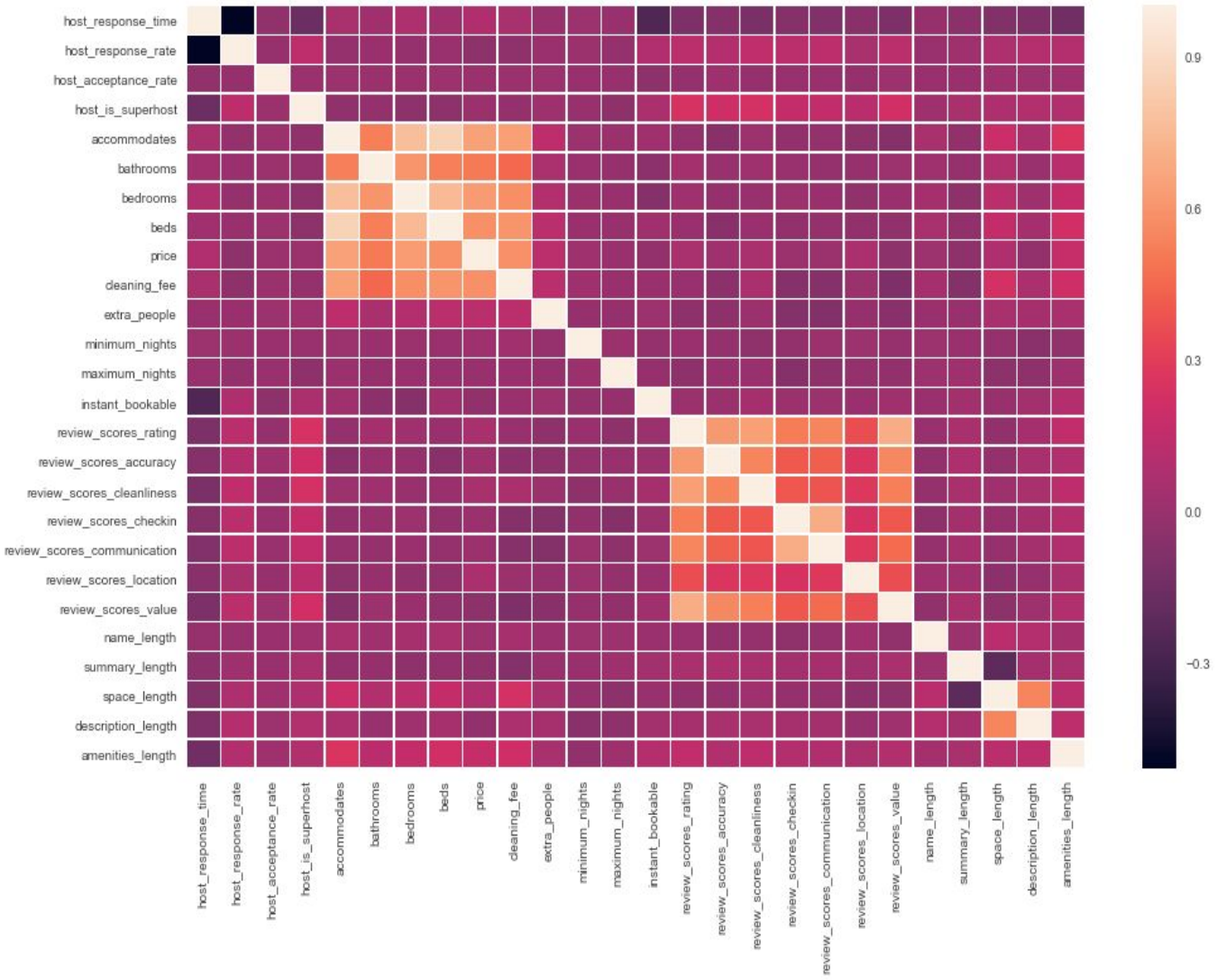


Even though these accommodates more than 7 and listings with more than 4 bedrooms can be dropped from our data as outliers. We decide to keep them.

After some exploratory data analysis, we recognize some patterns in regarding to favorite neighborhoods, numbers of bedrooms and accommodate capacity. Let's see we can see the correlation in a matrix.

8. We plot a matrix of correlations shows 4 groups of correlations:

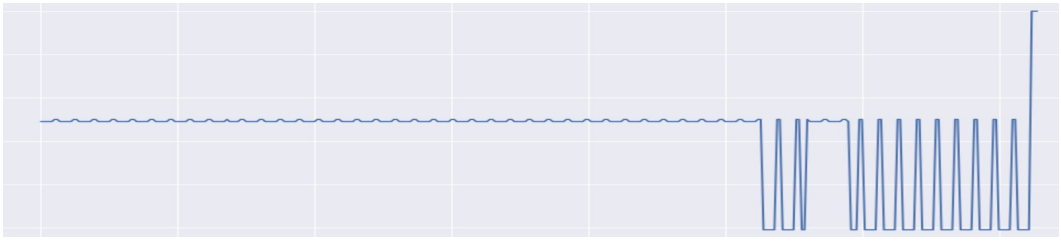
- Description length and summary length: hosts spend about the same amount of time to write descriptions and summaries. However space length and summary length are the opposite.
- Rating correlation: overall ratings and individual scores are highly correlated.
- Price, beds, accommodates, bedrooms, bathrooms, cleaning fee are highly correlated: people use these features to price their listing. But is it the right way?
- Host response time and host response rate are not correlated. And so is host response time and instant bookable.



calendars.csv

We conduct some time series analysis using data from calendars to explore whether there is any pricing strategies patterns, and found 4 interesting patterns as belows:

A. Weekend spike:



The weekend has a higher rate than weekdays, hence we see the spikes.

B. Summer rate:

This listing has a higher rate set for the summer.

C. One rate:

This listing sets at a fixed rate all year round.

D. Holiday rate: Peaks on Christmas and high rate during summer

Preprocessing

We did a thorough job in data wrangling of listings.csv including getting lengths of description features, convert object to categorical, binary and spectrum.

We remove all other ratings score accept our target: review_scores_value (since they are highly correlated meaning they are the same)

We then get dummies variables for our categorical features: 'neighbourhood_group_cleansed', 'property_type', 'room_type', 'bed_type', 'cancellation_policy'

We add our target, and drop the review_scores_value before split train and test data.

Now, we are ready for modeling.

Modeling:
We use 4 different models to predict the review_scores_value of Airbnb listings with 31 features.

A. *Linear Regression*
B. *Decision Tree*
C. *Random Forest*
D. *Gradient Boosting*

Import each model from sklearn, fit data into train, test.
Plot the top 10 coefficients of linear regression.
Plot the top 10 important features of decision tree, random forest and gradient boosting.

Evaluations: Using mean square of errors to see how close our predictions are

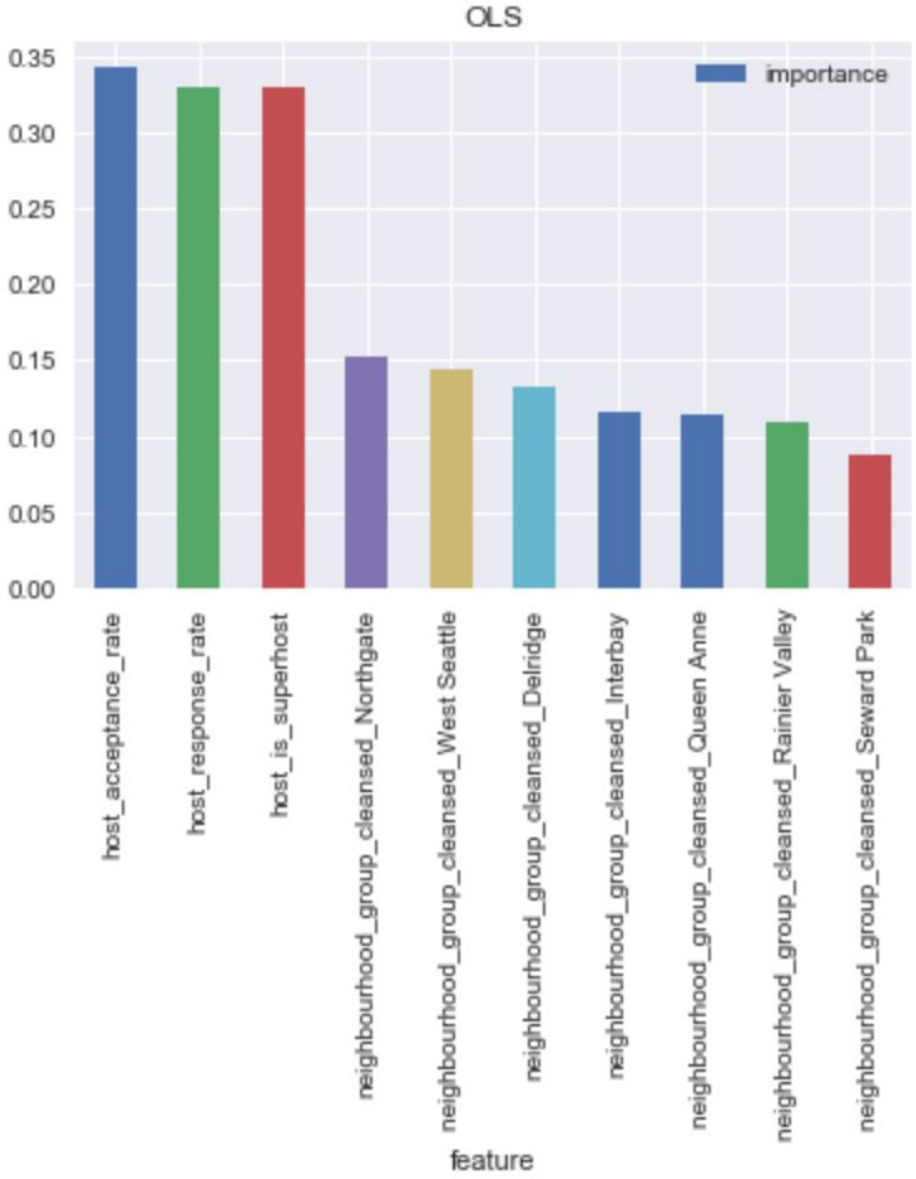
Model	Mean square of errors
Linear Regression	0.6977697963551963 lowest
Decision Tree	1.8637822325921867 highest
Random Forest	0.7058711712461408
Gradient Boosting	0.8669365301845464

Findings

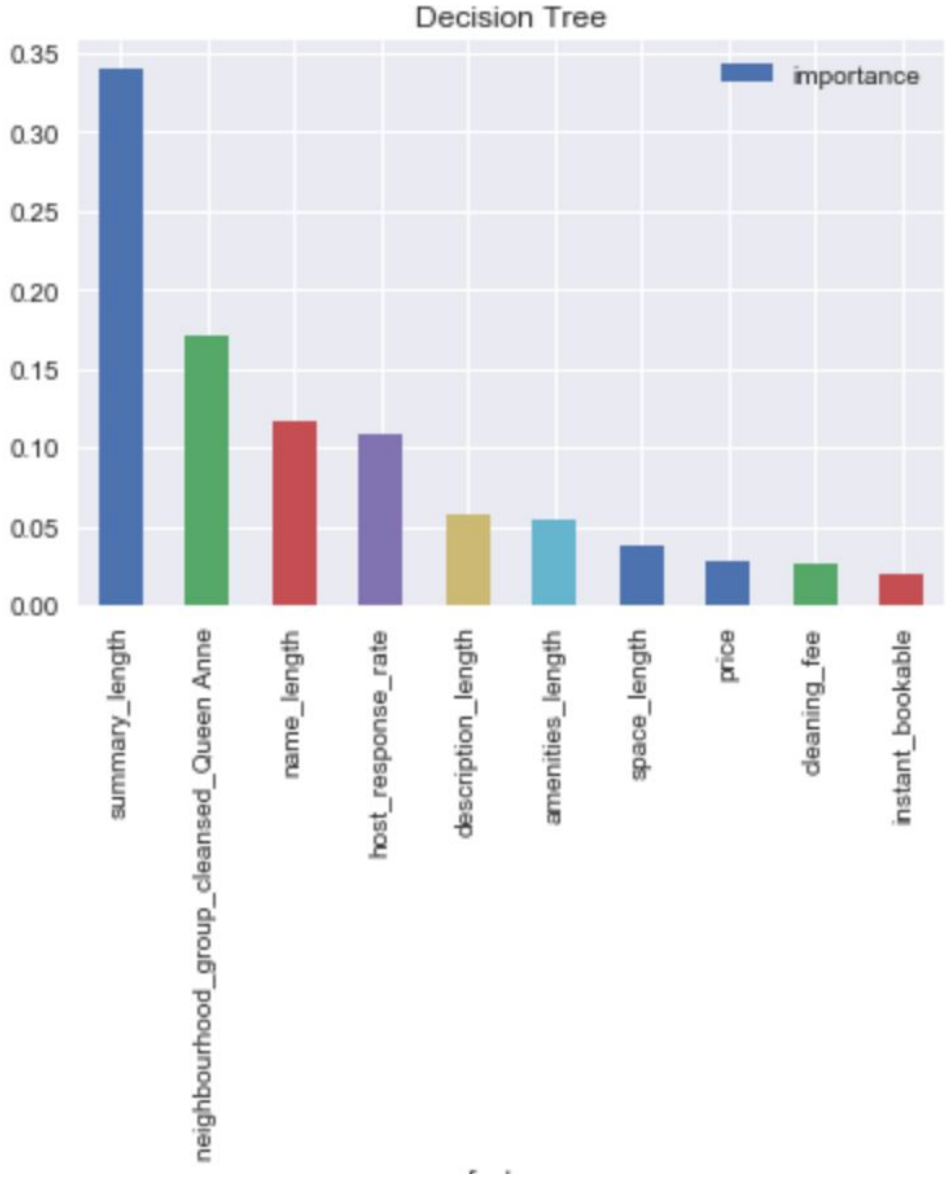
A. Linear Regression:

With a mean square of errors of 0.6977697963551963

10 features that determines the value score are host related and location related. But location score has the weakest correlation with value score. Is that an indicator to the difference between locations and neighborhoods? Or a preference in neighborhood or what guests prefer as individuals are difference than a location as convenience to move around? One thing we learn from this is that what can determine the value score, does not necessarily be the most important thing or preferred things. Later in other models we can see that the neighborhoods group changes quite interestingly.



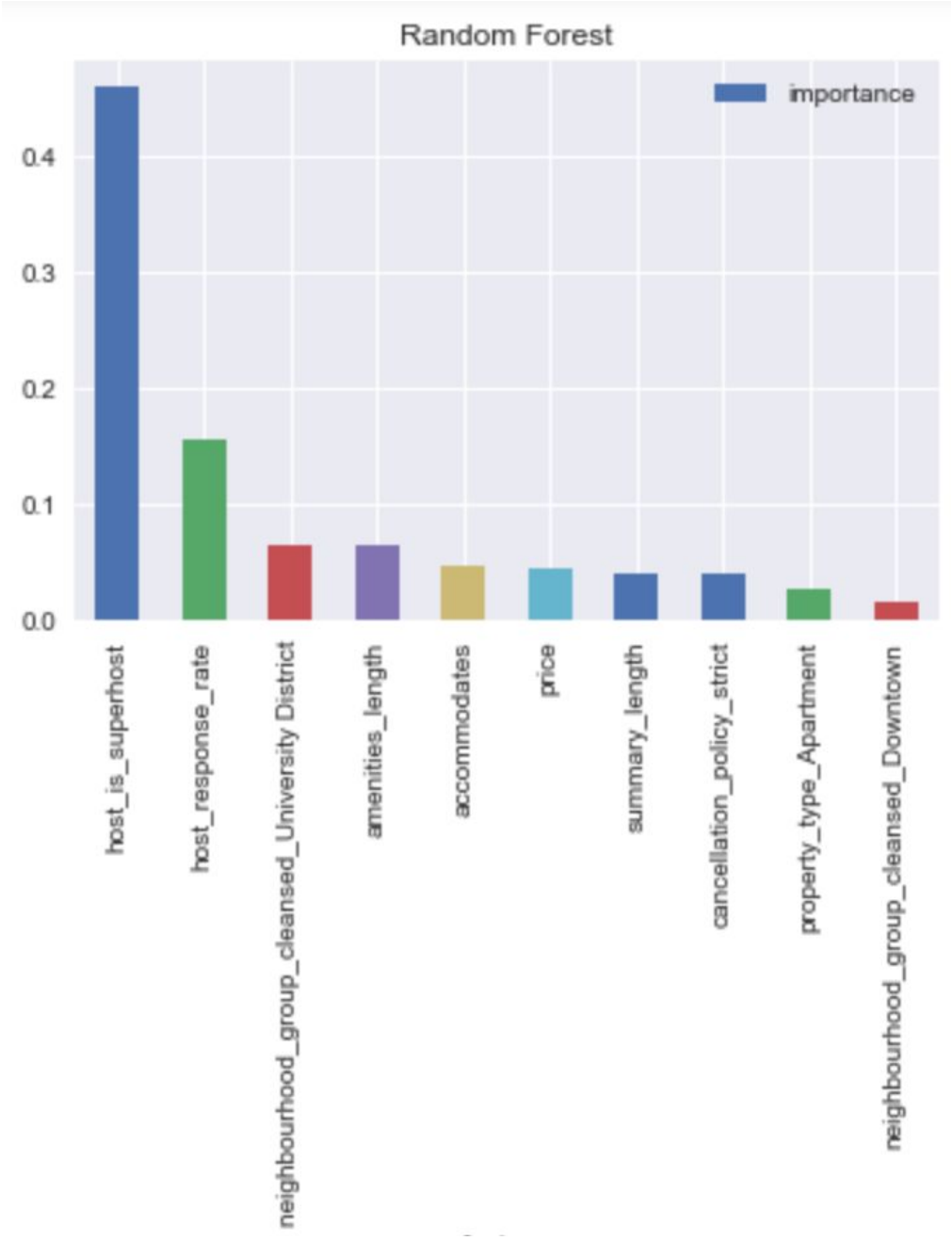
B. Decision Tree: with a mean square of errors of 1.8637822325921867
Decision Tree has the highest mean square of errors or the least accurate model (even though we don't use accuracy score here as our evaluation)
10 most important features are completely different from OLS, but at the same time makes more sense as explanation to what are the most important features in determine values of a listing.



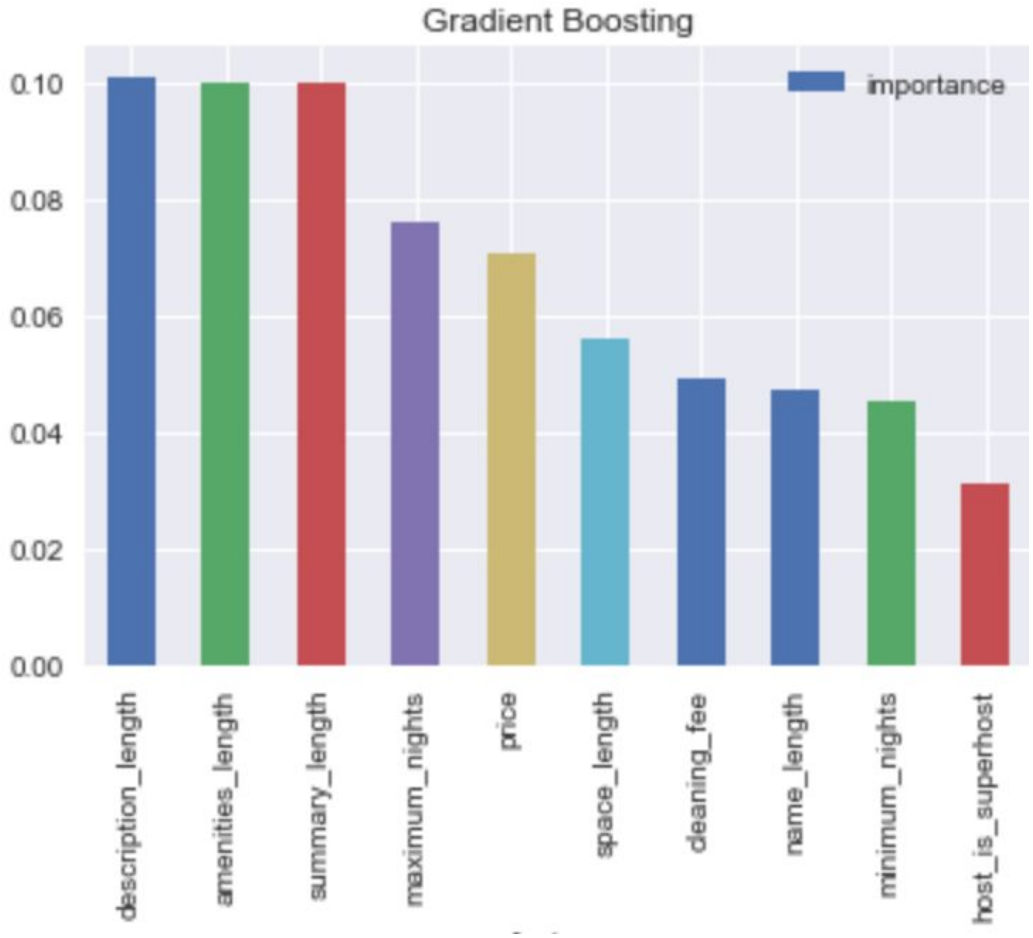
C. Random Forest:

With a mean square of errors of 0.7058711712461408, close enough to OLS, with a diverse group of 10 most important features in which 9 of the top 10 features weigh almost evenly, with the indicator that superhost is the determining factor in a high value score.

Notice how the neighborhoods changed from Queen Anne in Decision Tree model to University District and Downtown in Random Forest? While other top features are almost the same in both models: price, summary length, host response rate.



D. Gradient Boosting: with a mean square of errors of 0.8669365301845464. In this model, we do not see any neighborhood group in the top 10 important features, and all top 10 features weigh evenly. Gradient Boosting is the closest to the correlation matrix.



It is hard to decide which is the best model depending on the mean square of errors. OLS might have the lowest MSE, also the fastest model, but Random Forest makes more sense to me as what determines values.

If numbers speak the truth, there is an underlying relation of value score and what guests value subjectively. As different neighborhoods of Seattle fit different styles and personalities, location might not determine one's value score, but a fitting neighborhood could determine that.

Conclusion:

With 31 features, we predict the review_scores_value with 4 different models producing acceptable results in all 4 models. Linear regression carries the best or lowest mean square of errors. Presumably it is the best model among 4.

We temporarily conclude that review_scores_value depends heavily on host acceptance rate, host response rate, if host is superhost, and if the listings are in neighborhoods listed in the features with the top 10 coefficients. When taking a careful look into to list of neighborhoods, we reconsider that linear regression is our best predictor.

We match the 7 neighborhood groups with our swarmplot, and point out that some of these neighborhoods such as Interbay or Seward Park have significantly less listings than other neighborhoods. Factoring the number, their importance factors in deciding a target is bigger. This fits the analogy, 'Big fish in a small pond." Since there are not that many listings in these groups, it is easier to predict the value score of these groups.

We explore the second alternative best model: Random Forest. The top 10 features make sense. Host response rate and superhost, not host acceptance rate are the top 2 determiners. It is agreeable that a listing has higher value if the host is a superhost. Airbnb has strict requirements for superhosts having 4.7 or above rating (on the scale of 5), no cancellation, response time in 24 hours. These requirements are even higher than a 5 star hotel.

The two neighborhood groups make it in the top 10 of Random Forest model are University District and Downtown, which makes the most sense since they both have great public transportation and lively atmosphere. Amenities, price, accommodates all affect the value score. Interesting thing, cancellation policy strict has great value. And apartments are the favorite choice of property type in Seattle, my wild guess: apartments mean highrises, they have better view than houses that are not in Queen Anne.

We conclude that Random Forest is the best model with no bias like OLS that groups listings in to groups to predict.

Now that we know the most important factor in deciding a listing value, our nextfuture-steps would be analyzing reviews.csv, especially the comments columns. Our approach is natural language processing to predict if a review has a good or bad value score. Our second approach is unsupervised clustering to find any pattern in airbnb guests' comments.

No doubt, Airbnb knows what they are doing by setting up Superhost status with strict requirements. Competing to the hotel industry and other copycats, they stand out thanks to their invest in technology with data science being an important part. Even though, their pricing strategies are not the best for the hosts, maybe they are the best for the guest. Airbnb understands they excel with the guests support, they are guest oriented.