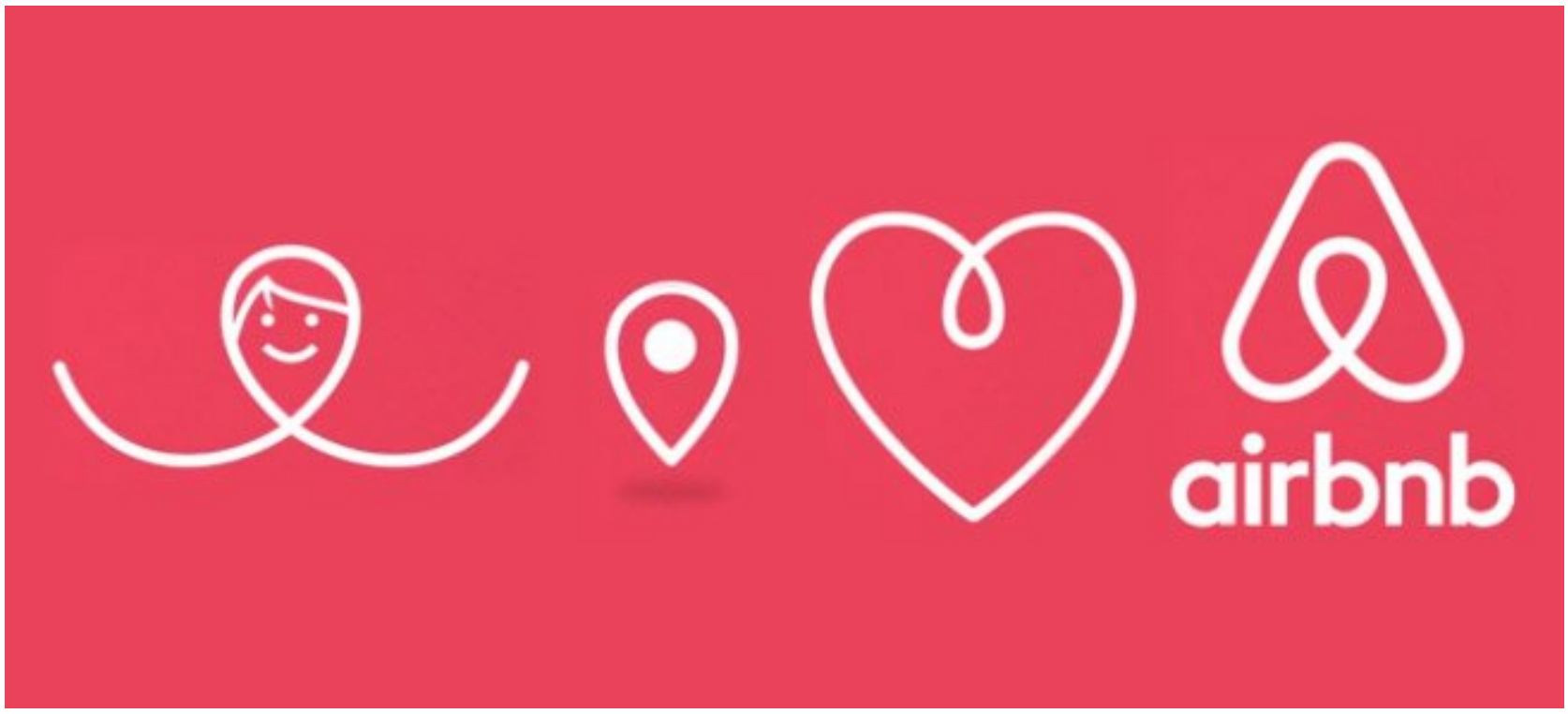


Airbnb Good Pricing vs Bad Pricing Classification

To answer how much you should price your listings

Table of Contents:

1. Problem Statement
2. Data Wrangling
3. Exploratory Analysis
4. Feature Selection
5. Modeling: not in milestone report
6. Evaluation: not in milestone report
7. Findings: not in milestone report
8. Conclusions: not in milestone report

Problem statement:

1. How to strategize pricing when you first set up an airbnb listing or how to adapt to it?
2. How to set up a successful airbnb listing from price to title? What features are not given enough attention and could be improved?
3. What make a listing profitable or successful.
4. What are the key factors in deciding that?
5. Is there a way to predict potential incomes?

These questions will be answered for individual airbnb hosts, corp rental hosts as well Airbnb themselves to improve their user experience.

These questions will lead to a dynamic pricing system that ideally can be integrated into Airbnb website and app to assist the hosts to maximize their incomes suited for their goals.

Data Wrangling:

The dataset of Seattle airbnb listings is obtained from Kaggle. In the dataset, there are three files: listings.csv, calendar.csv, and reviews.csv

Cleaning dataset steps include:

1. Remove $ of columns: price, weekly price, monthly price, cleaning fee, security deposit and convert them into floats.
2. Convert percentage into floats.
3. Convert date into datetime format.

Wrangling dataset steps include:

1. Divide calendar into subsets of each listing to plot a time series of price.
2. Remove outliers with price over $800

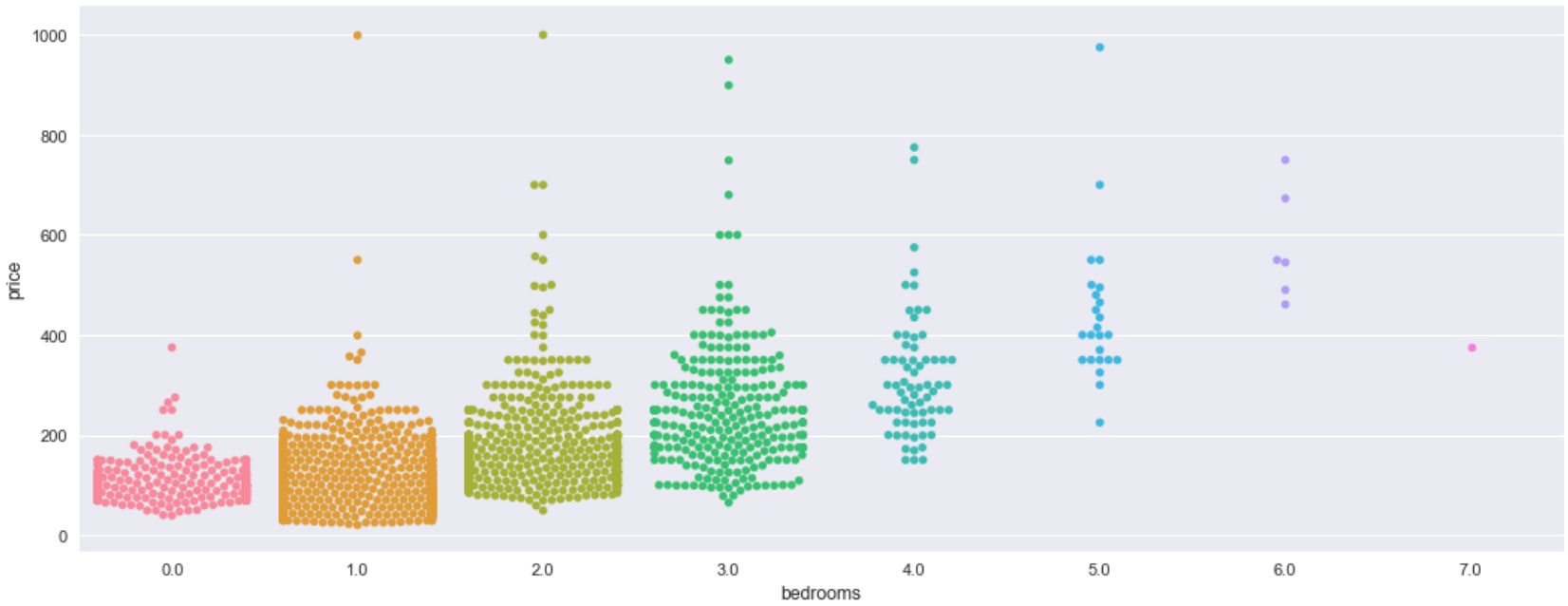
Exploratory analysis:

Exploratory analysis steps:

1. Inspect the data frame of each file 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.
    c. .describe() to summarize mean, average, median, standard deviation of the numerical columns.
2. Find the top 10 categories in categorical columns: neighborhood, property type, amenities, bedrooms: using .value_counts().head(10)
3. Histograms for distribution of interested features
4. Swarmplots for categorical features vs price.
5. Find max of interested features.

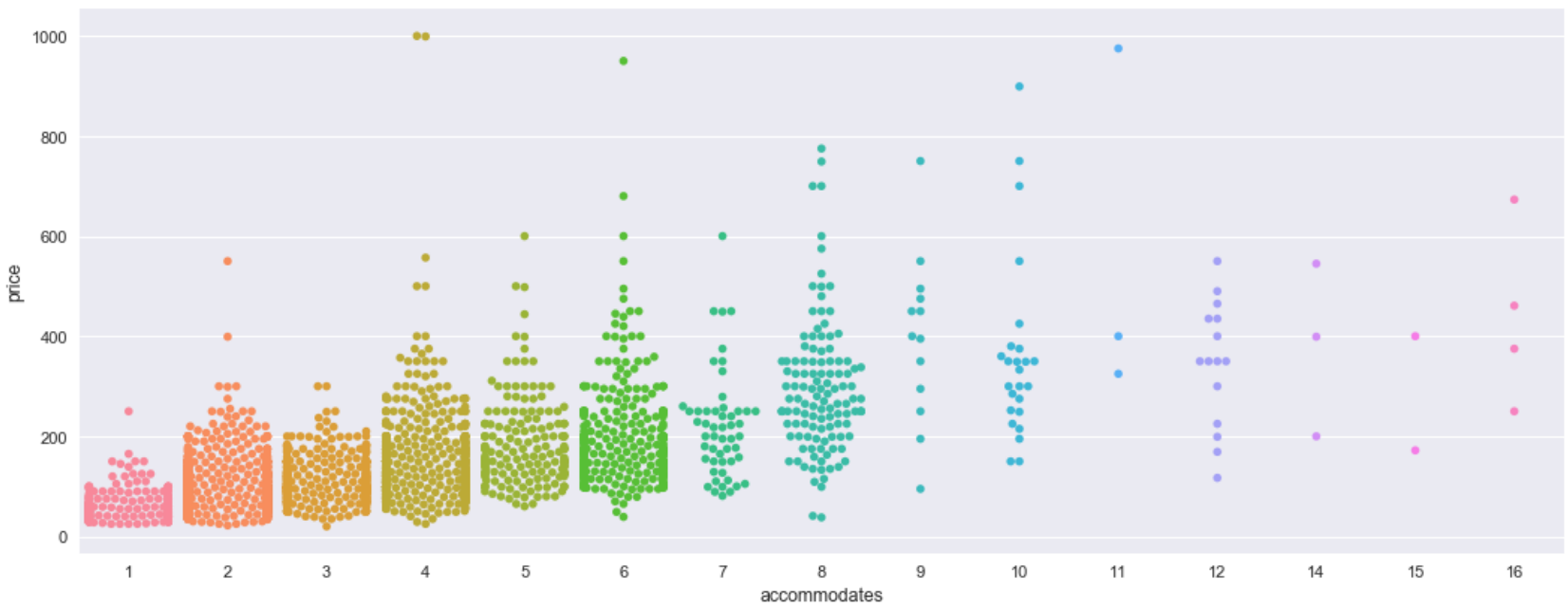
Exploratory analysis stand out findings that will be used for feature selection part:

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.
4. The highest counted bedroom size is 1 (2417 listings), and the highest counted for number that accommodates 2 people (1627 listings).
5. There are 81 neighborhoods, then get grouped into 17 neighborhood groups.



This swarmplot shows the prices of listings in each bedroom size category 0 to 7. As shown above, listings with 1 bedroom has the highest count of 2417. The pattern we notice from the plot is:

1. The listings with price above $800 could be outliers. Might need to remove them for the model.
2. A listing with 7 bedrooms is priced less than 6 bedrooms. That could be another outlier.
3. Starting price of 1 bedroom listings is lower than 0 bedroom (studio) listings. A guess is that studio listings tend to be entire property, while 1 bedroom get separated into entire 1 bedroom property or 1 private bedroom.

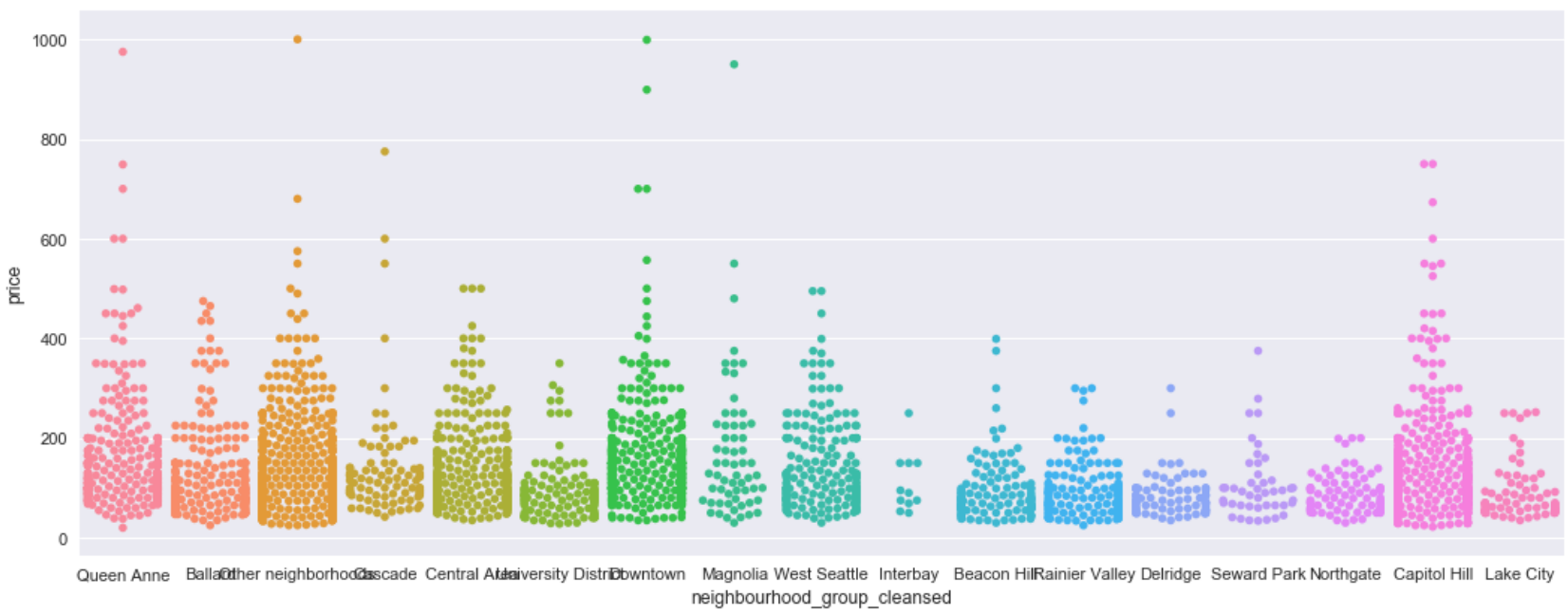


This swarmplot shows the prices of listings that accommodates from 1 to 16 people. Interesting pattern to notice:

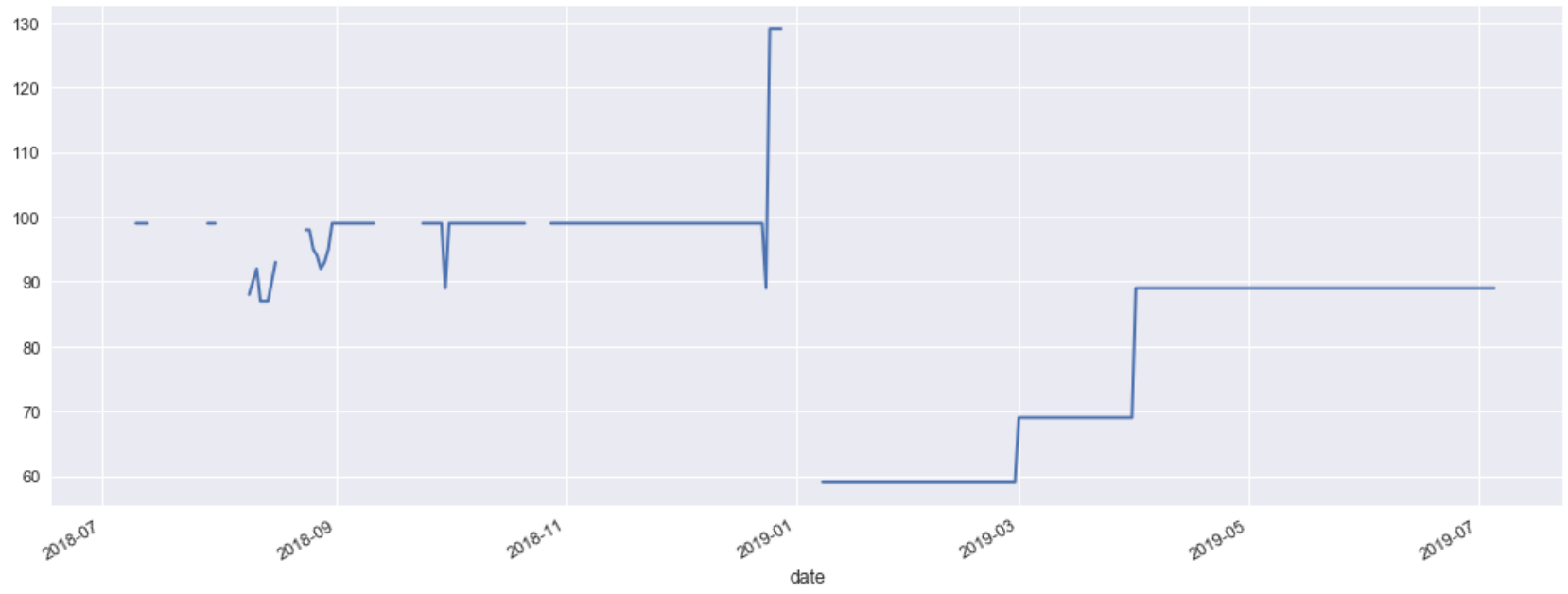
1. The even number are priced higher than the odd number.
2. Listing for more than 9 people are very little. There is not market for that?
3. The optimal number for accommodates is 2 to 6.

I choose swarmplots because it shows the difference between each categories as well as the volume of each categories.

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.



This is a time series plot of a randomly selected listing: price over a year. We notice it doesn't have dynamic hotel pricing, but a set price at just below $100. There is a surge for the holidays: Christmas and New Year, and then lower price for the winter months January to March.

Exploratory Data Analysis Inferences:

There are patterns in our interested features: bedroom, accommodates, neighborhood and date, which is good sign.
In order to determine the success of a listing we also need to looks into reviews: scores and numbers of reviews.
The dataset does not have total revenue or occupancy rate, hence we need to make some assumptions in order to calculate these numbers: every booking has a review, and the average length of each booking is the minimum nights

MINIMUM OCCUPANCY RATE: (numbers of reviews x minimum nights) / availability_365

ESTIMATED REVENUE: (numbers of reviews x *minimum nights) x* price

MAXIMUM REVENUE: availability_365 x price
The next step would be feature selection using Chi-squared or LASSO.