**Capstone Project Final Report**

Trieu Vo

Yash Tadiyal

Nilay Anand

Ashutosh Kumar Singh

Rohit Meena

College of Professional Studies, Northeastern University

ALY 6140: Analytics Systems Technology

Prof. "Eric" Yang Liu

October 28th, 2021

**Introduction**

**Airbnb:** Airbnb is a platform that allows people to list, find, and rent lodging. This company is one of the greatest startup success stories of our time. Even with an attractive business plan and fantastic traction initially, the travel startup still had to pitch their company to potential investors. The founders of Airbnb are Brian Chesky and Joe Gebbia.

**Goal/Purpose of the Project:** Airbnb has requested to use predictive modelling so that they can create market specific forecasts with multiple variables. We are working as an Airbnb devoted team that forecasts and reports to optimize the existing predictive models so that they can use data to not only improve their service and search, but their hiring practices and customer groups as well.

**Problems:**

1. Price is an important concern for customers booking travel online.

2. Hotels leave you disconnected from the city and its culture.

3. No effortless way exists to book a room with a local or become a host.

**Solution:**

1. Make Money while hosting.

2. Share culture with a local connection to the city.

**Competitors:**

1. Rentobi.com

2. Craigslist

3. Couchsurfing.com

4. Hotels.com

**Question:**

1. How to increase the chances that the tenants will book the rentals after proposing properties with such features to them?

2. How to increase the chances that the owners' house will be rented more frequently by suggesting a set of features to them?

3. How can engagement be increased for houses that are less often rented or have not been booked in a long time?

4. Which is the most profitable zip code?

5. Which area does hiring require?

6. Which city has the high customer base and more property engagement needed?

**Methods Used:**

1. Linear Regression Model.

2. Random Forest Model.

3. K-NN Model.

4. Decision Tree.

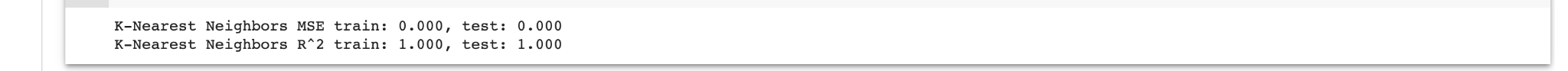
5. XG Boost.

**Exploratory Data Analysis**

**Data Extraction**

We gathered the information from Inside Airbnb because it is readily available and free. We have approximately 207,000 data samples from 28 cities in our dataset. We made a list in which we saved the URL and used the list to create a single Data frame. So, our dataset has 207,006 rows, 74 columns, and 74 attributes, such as "bedrooms," "bathroom," "room type," and so on. After that, we add the zip code URL, which will show us all of the zip codes in the United States, along with their coordinates (latitude and longitude).

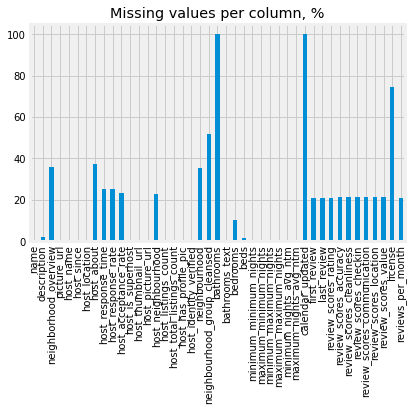
We created a model using the K-Nearest Neighbors technique, then tested it to see if it could predict the zip code based on the coordinates for our dataset. The average of the sum of squared differences between the actual value and the projected or estimated value is the mean squared error (MSE). MSE is always positive or larger than zero in value.



A number around 0 indicates that the estimator / predictor is of higher quality (regression model). The fact that the predictor has an MSE of zero (0) indicates that it is a perfect predictor. The ratio of Sum of Squares Regression (SSR) to Sum of Squares Total (SST) is known as R-Squared (SST). The amount of variance explained by the regression line is known as the Sum of Squares Regression. The goodness of fit is measured using the R-squared value. The better the regression model, the higher the R-Squared score. The R-Squared is bounded between 0 and 1 for the training dataset. If R-Squared is 1, the model fits the data exactly, with an MSE of 0 as a result.

​​Now we calculate the data type, the number of missing value and the percentage of missing values of each feature to selecting feature efficiently.

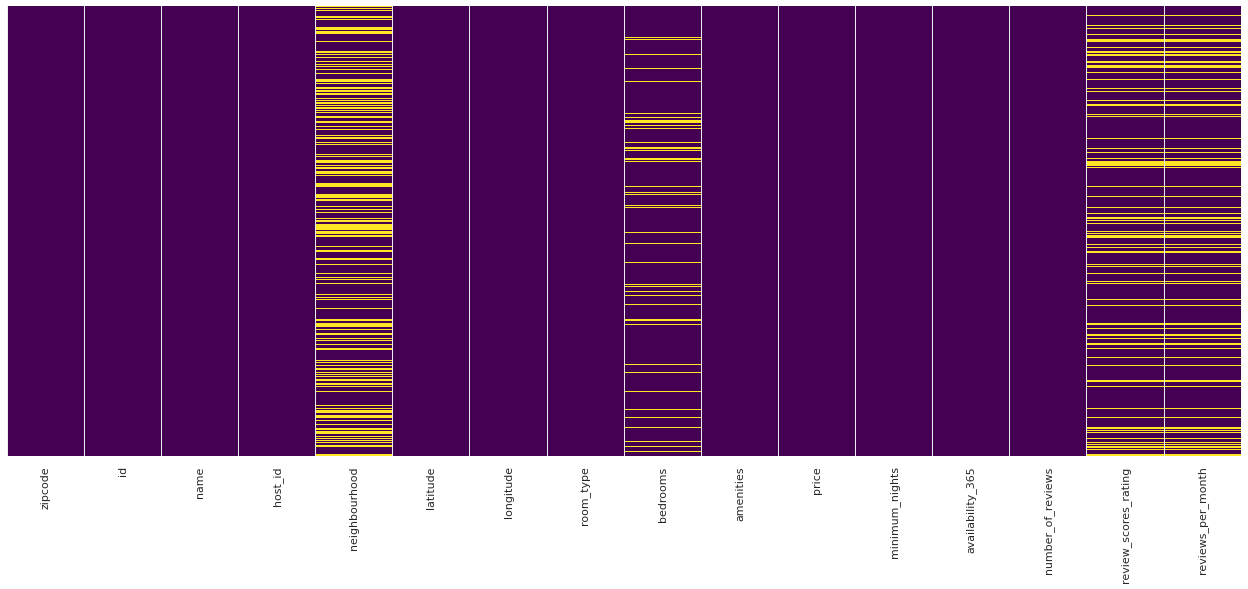
Here we visualize features that have missing value from our dataset.

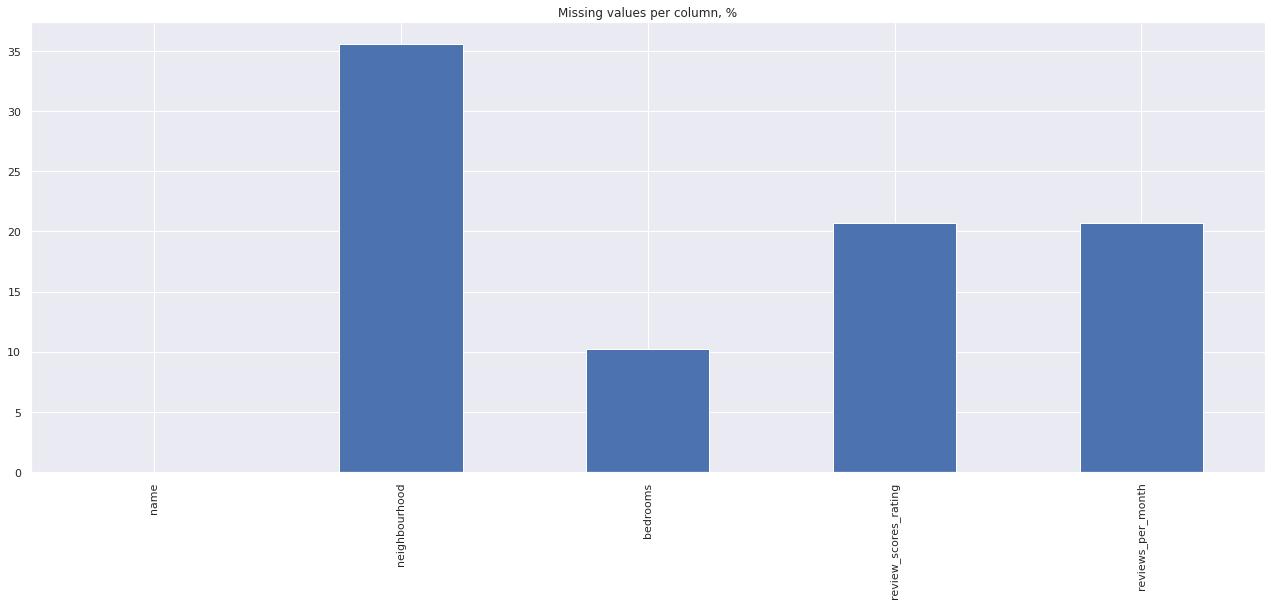


We avoid choosing bathroom, calendar\_updated and license because they have high rate of missing values.

**Data cleanup**

Then we start the feature selection and create a new data frame that only contains information of the selected features. The new dataset has 16 features and get the shape of our new feature dataset. We start the evaluating missing value by removing / filling. We Calculate the data type, the number of missing values and the percentage of missing values of each feature. We can see the frequency of missing values in each feature and also the percentage of missing values of features that have missing values.



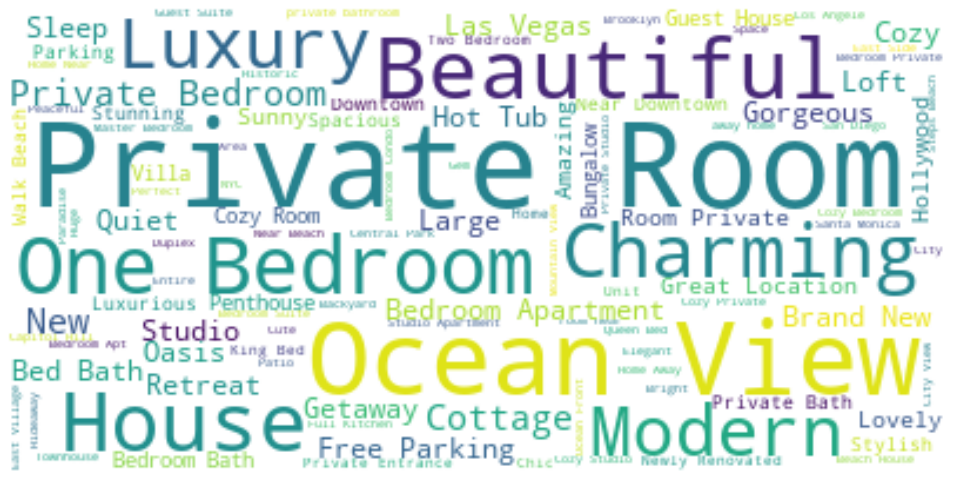


To begin, we fill missing values in the neighbourhood, review\_scores\_rating by propagating the most recent valid observation forward to the most recent valid observation, and then fill missing values in the bedrooms and review\_per\_month using mean values.

**Data Visualization**

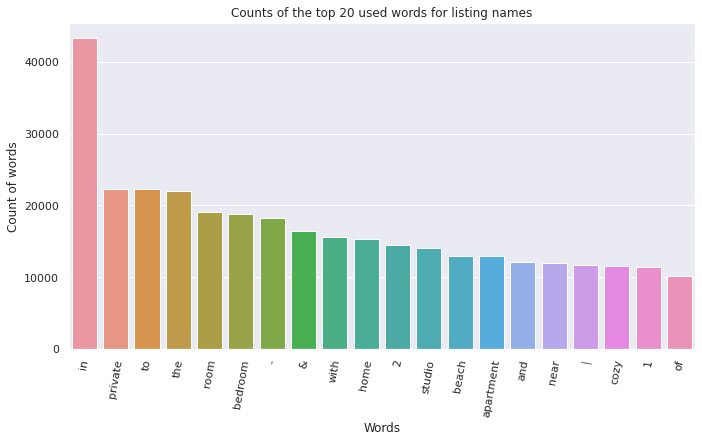
Data visualizations allow us to view the data in graphical or visual form making it easier to find any trend or patterns in the data.

**Keywords in the Name of Listings**



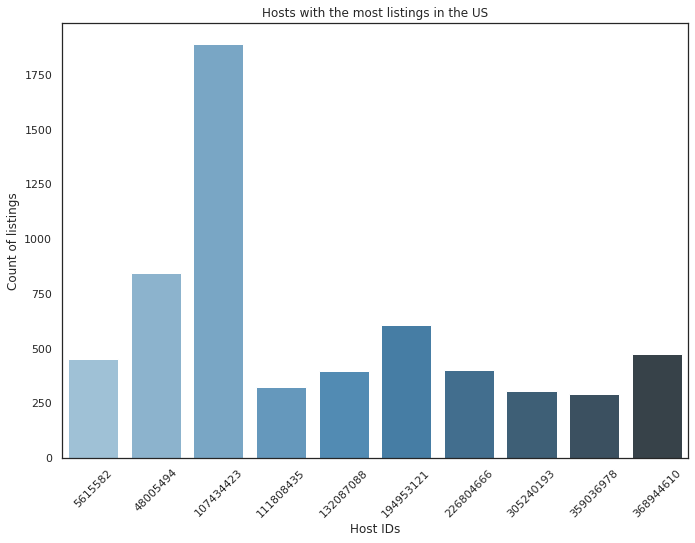
We start by looking into words used in the listing to attract customers.

Using a python library called word cloud to visualize most used words we find works like *Private Room, One Bedroom, Ocean View, Beautiful, Modern, Charming and Luxury.*



Further plotting the count of words, we see similar results in terms of frequency.

**Host ID**



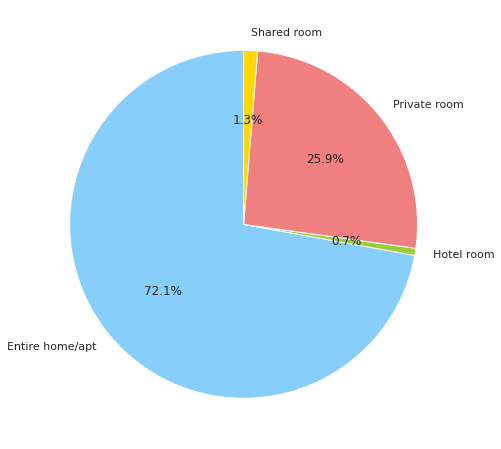
Next, we see the number of listings for each host. The host with the host id "107434423" has quite a lead compared to other hosts with 1891 listings on Airbnb.

**Mapping locations**



California, Chicago, Hawaii, and New York have the most dense collection of points. These locations have the highest number of listings in the data.

**Room Type**



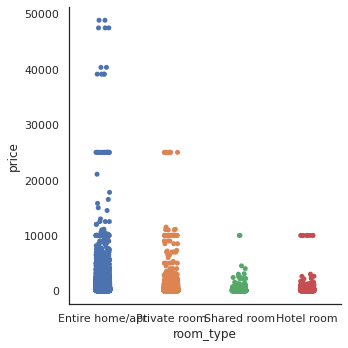
Rooms are divided into four types:

1. Entire place
2. Private room
3. Shared room
4. Hotel room

Entire place or apartment has the highest number of listings. Clearly most users of Airbnb want a place with privacy for them.

Hotel rooms on the other hand are the least, Airbnb listings are most likely privately owned and do not operate as a traditional business.

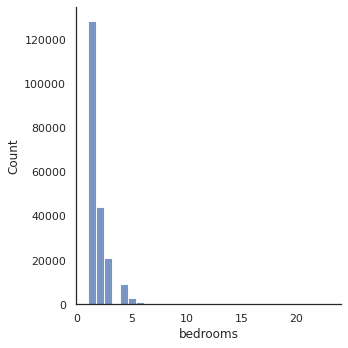
Pricing of rooms varies according to the type of rooms.



Entire homes have the highest price among all types of room with prices going as high as $50,000. Shared rooms or hotel rooms are lower average price in comparison.

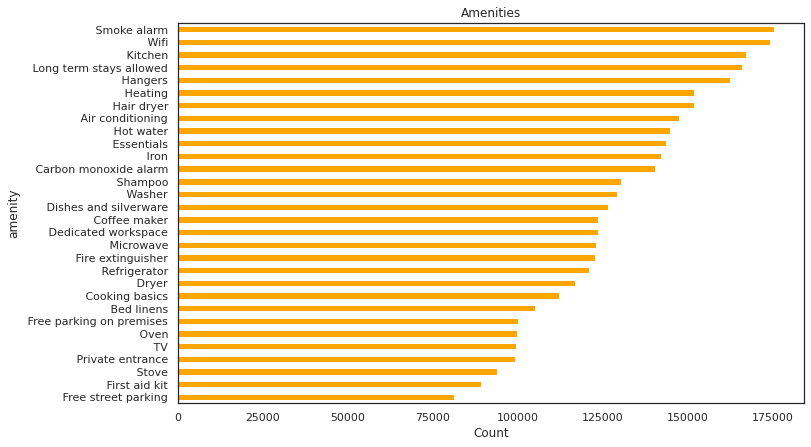
**Bedrooms**

Airbnb listings also offer a variation according to the number of bedrooms.



Most listings offer a single bedroom. Listing ranges from single to bedroom to more than 10 but bedrooms more than 6 are so few that they are not represented in the graph.

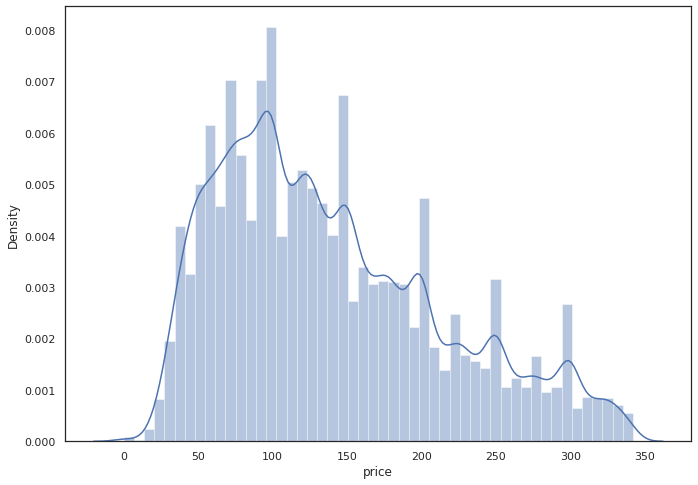
**Amenities**



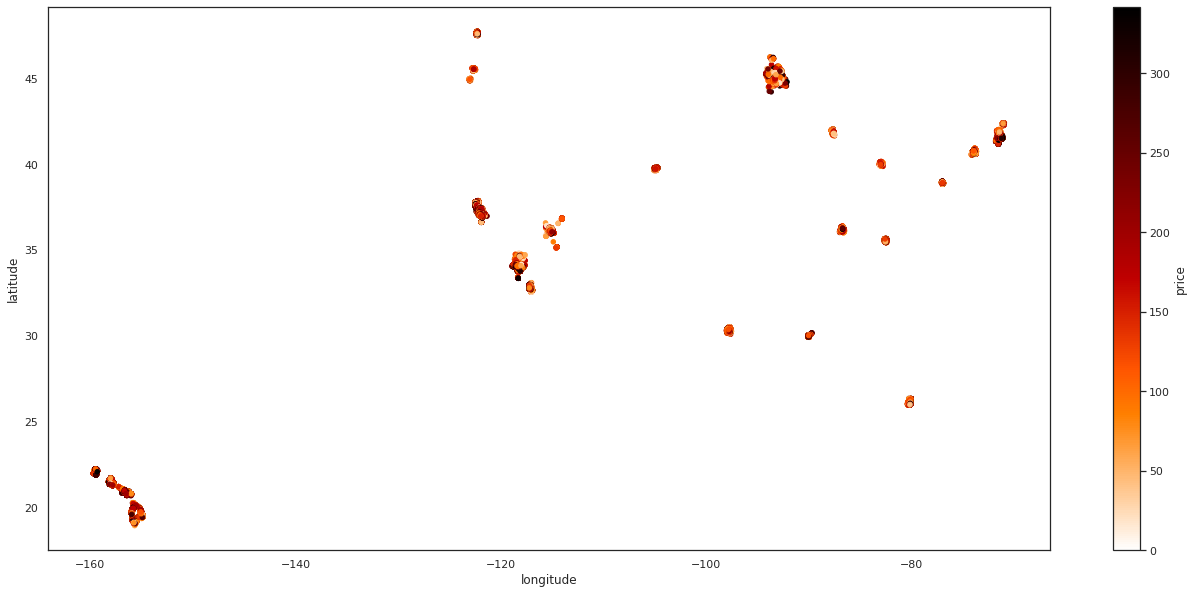
Deciding what amenities to offer customers is important for a host. We visualise the top 20 amenities that a hosts list on their property. We conclude that the most important one is a smoke alarm with safety being of utmost concern and required by law in each house. Second one is WiFi, the internet being a necessity in our age. Rest of the amenities by frequency are Kitchen, Long term stays, Hangers, Heating, Hair Dryer, Air Conditioning, Hot Water, Essentials, Iron, Carbon monoxide alarm, Shampoo, Washer, Dishes and silverware, Coffee Maker as so on. Refer to the above graph for more information.

**Price**

After removing the outliers from the list of prices we notice most of the listings have prices in the $50 to $150 range. $100 is the most frequent price within the listings.

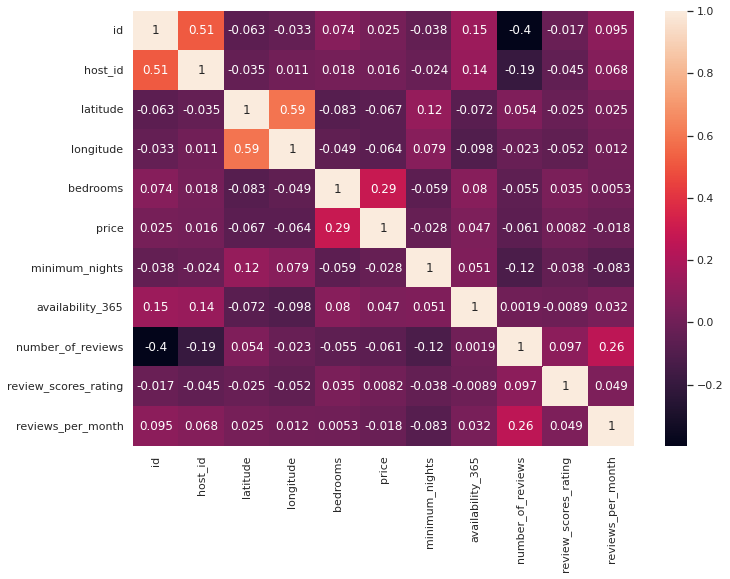


Visualizing price according to latitude and longitude:



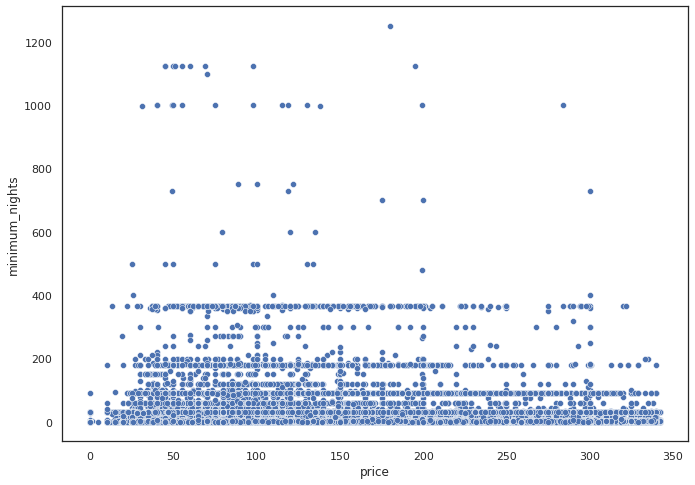
Although most places offer affordable pricing, the costliest regions seem to be within Hawaii, California, and New York with much higher prices compared to other regions.

**Feature Correlation**



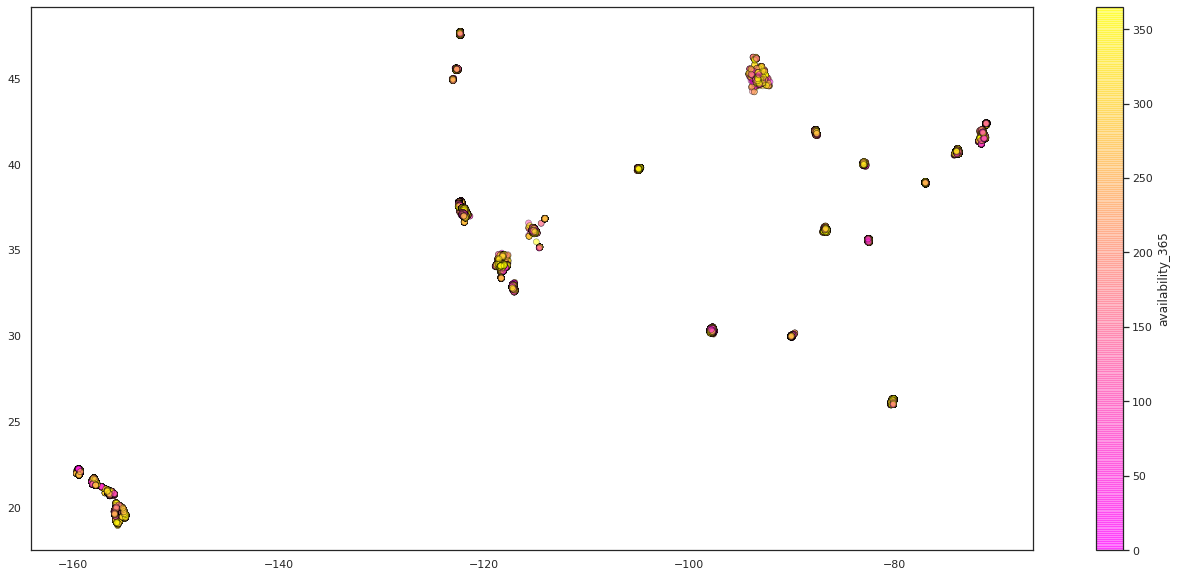
Mapping the correlation between features, we summarise that there is no real correlation between the features. Price and number of bedrooms have the highest correlation with 0.29 showing some relationship between the number of rooms and the price of listing, although it is not significant enough. So, there are other features that affect the price of the listing. Also, the number of reviews and reviews per month have some correlation which is to be expected.

**Minimum Nights**



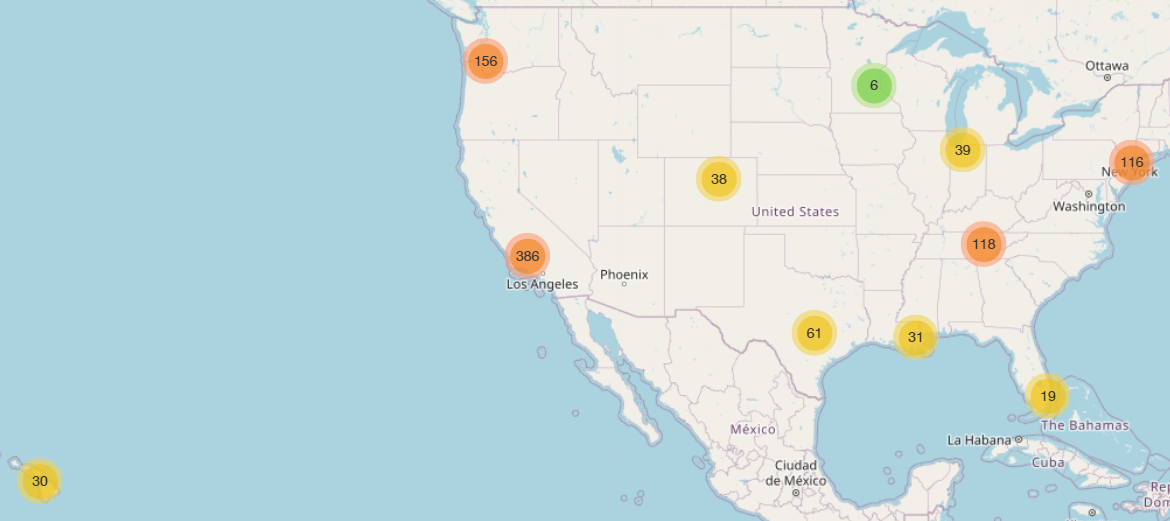
Plotting the price and minimum number of nights in a scatter plot, the price varies a lot within the same number of nights. Although we can conclude that most listings allow under 200 nights of minimum stay.

**Room Availability**

****

Plotting the coordinates with availability within 365 days, we can conclude that most listings were available to over 200 days in the last year.

**Number of reviews**



Visualizing 1000 locations which have the highest number of reviews, we can see states which are hottest, most visited are California, Washington, New York and Tennessee.

We can concentrate on these locations to find out the most profitable zip code.

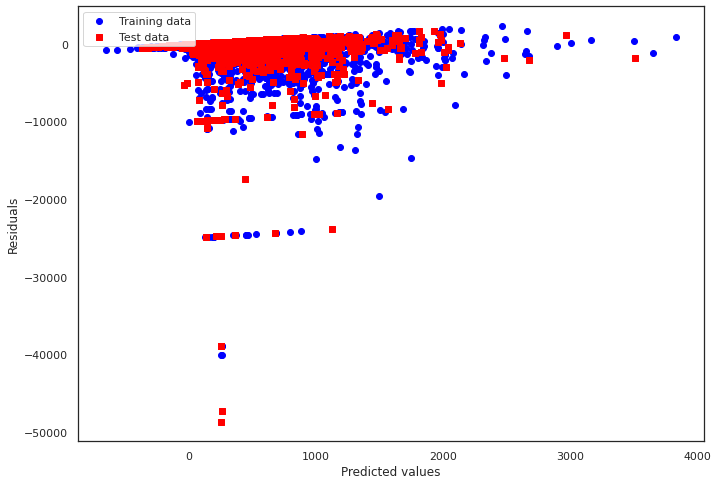
**Predictive Models**

**Linear Regression**

Linear regression model is one of the models that we’ve used to predict the price of a property on the basis of multiple factors such as location, amenities, etc. Linear regression is an approach used for modeling the relationship between one or more independent variables and a scalar dependent variable. Simple linear regression is used when there is only one independent variable, whereas multiple linear regression is used when there are numerous input variables.

Linear Regression MSE train: 349827.902, test: 418670.400

Linear Regression R^2 train: 0.092, test: 0.075

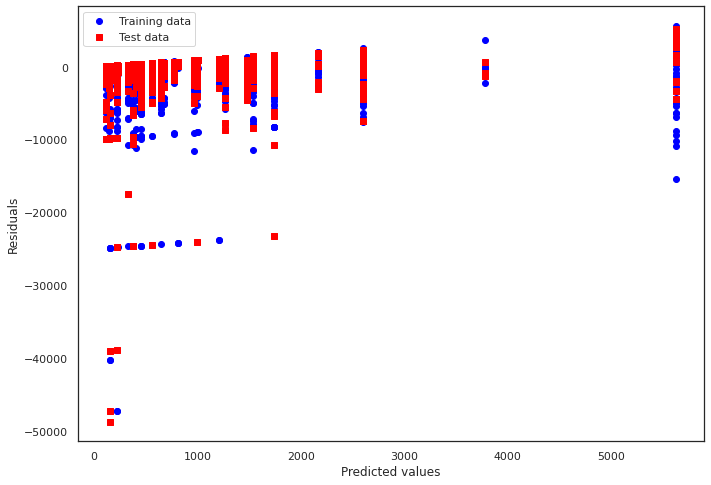


**Decision Tree**

Decision Tree is the second model used to predict the price of a property on the basis of its attributes. It is a non-parametric supervised learning approach to regression. It's used to build a model that learns basic decision rules from data attributes and predicts the value of a target variable. The decision criteria get more complicated as the tree grows deeper, and the model becomes more accurate.

Decision Tree MSE train: 328616.156, test: 403920.103

Decision Tree R^2 train: 0.147, test: 0.108

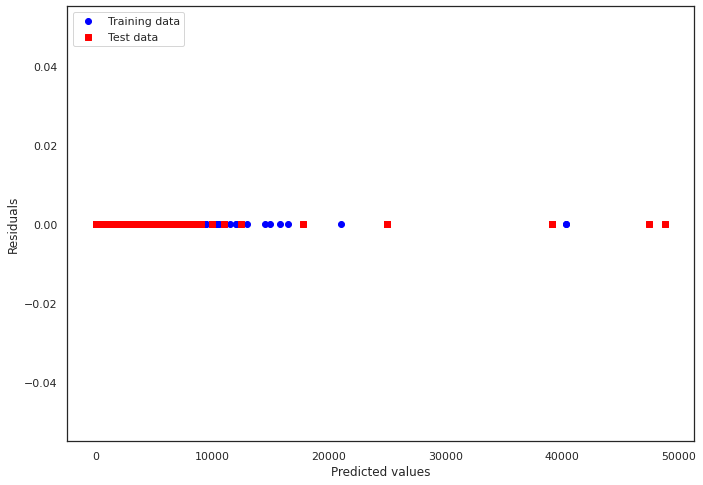


**K-Nearest Neighbor (KNN)**

The supervised learning technique K-nearest neighbor is the final model that is used to predict the price of a property on the basis of external factors through regression. By computing the distance between the test data and all of the training points, KNN predicts the price for the zipcode data. Then choose the K number of points that are the most similar to the zipcode data. The KNN method analyzes the likelihood of test data belonging to each of the 'K' training data classes, and the class with the highest probability is chosen. In regression, the value is the average of the 'K' training points chosen.

K-Nearest Neighbors MSE train: 0.000, test: 0.000

K-Nearest Neighbors R^2 train: 1.000, test: 1.000

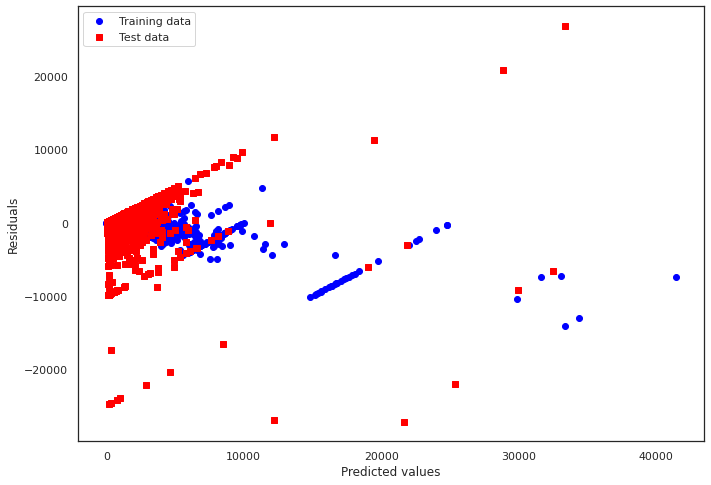


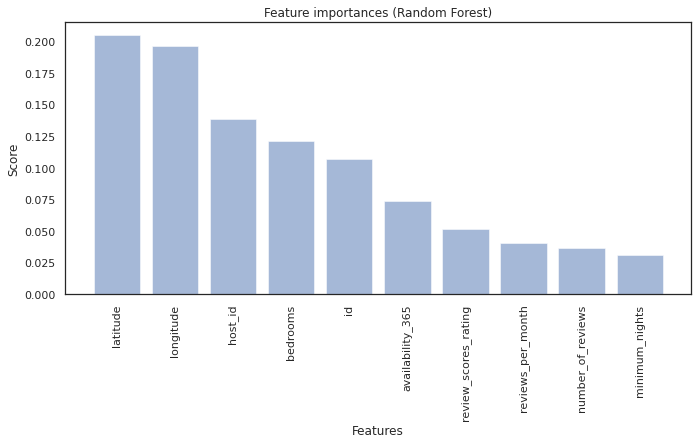
**Random Forrest**

Random forest is used to analyze and predict which of the features and attributes of the listed rental property are relevant from the customer’s perspective. It’s a decision tree-based approach for modeling predictions and behavior analysis which comprises multiple decision trees, each of which represents a unique categorization of data fed into the random forest. Random forest is based on the idea that a large number of substantially uncorrelated models (trees) working together as a group will outperform any of the individual component models.

Random Forest MSE train: 32832.664, test: 244568.729

Random Forest R^2 train: 0.915, test: 0.429





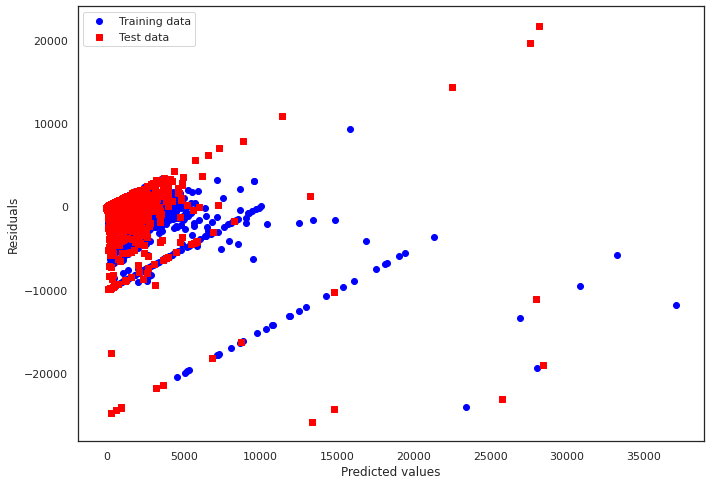
We can see that factors have the most influence on price: longitude, latitude, host\_id and bedrooms.

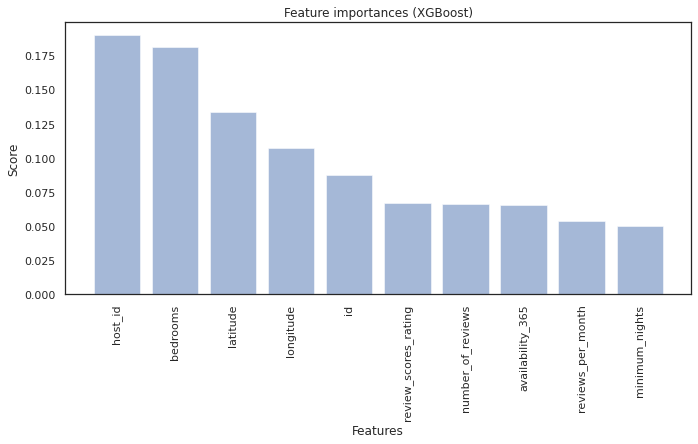
**Extreme Gradient Boosting (XGBoost)**

XGBoost is an algorithm used to predict the price of the properties and the influence of different factors on the rent of the listed rental property. The XGBoost technique produces decision trees sequentially to solve regression problems (rather than in parallel and independently, as Random Forest does), with each succeeding tree aiming to minimize the preceding tree's faults. Moreover, unlike Random Forest, where trees are developed to their full potential, boosting uses trees with fewer splits.

XGBoost MSE train: 124562.931, test: 242405.204

XGBoost R^2 train: 0.677, test: 0.434

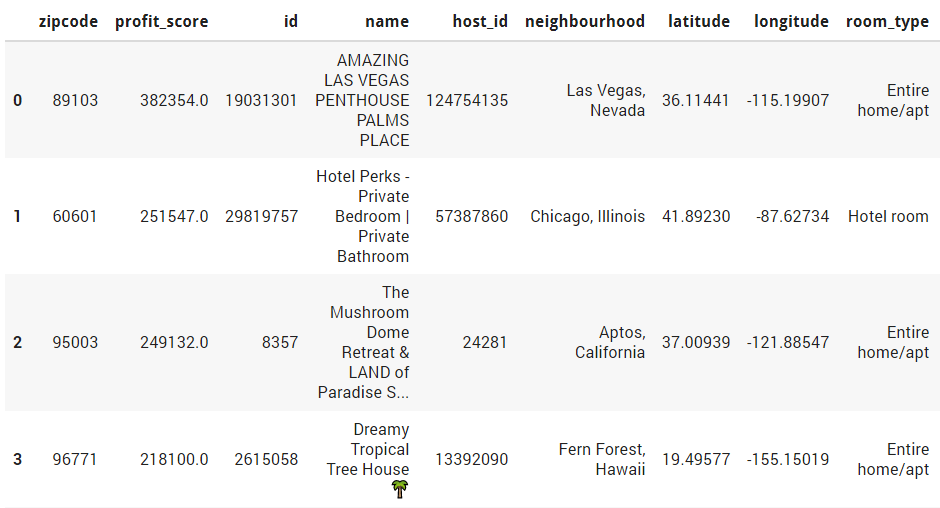


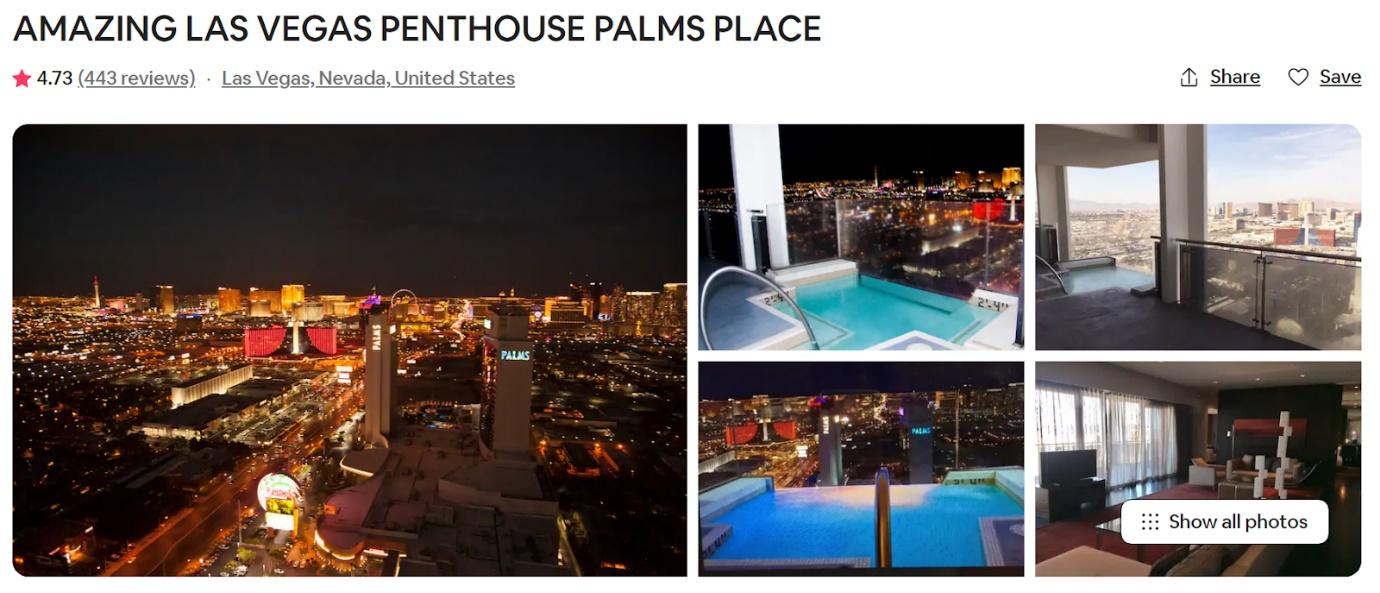


We can see that factors have the most influence on price: host\_id, longitude, latitude and bedrooms.

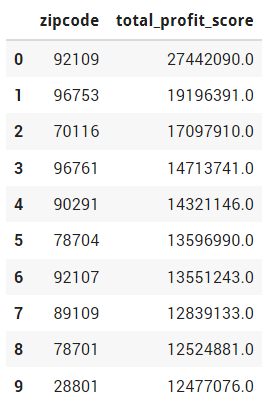
**Find the most profitable properties**

**Profit score of 1 listing = daily price \* number of reviews (number of visits)**





**Find the most profitable zip code**



1. Group by zip code
2. Calculate the total profit score of each zip code

The most profitable zip code is: **92109**. This zip code is in San Diego, California.

**Interpretation & Conclusions**

In this project, we have investigated the Airbnb dataset with over 200,000 samples from 28 cities, with 16 different features. We have gone through steps such as Data extraction, Data cleanup, Data visualization and Predictive Models Building. We have analyzed different aspects of the dataset and finally figure out the most profitable properties and zipcodes.

These are the questions and our answers:

***What is the most profitable property?***

Amazing Las Vegas Penthouse Palms Place. The zipcode of this place is **89103**.

***What is the most profitable zipcode?***

The most profitable zipcode is: **92109**. The zipcode is in San Diego, California.

***How to increase the chances that the owners' house will be rented more frequently?***

The hosts should offer entire home / apartment room type. They also need to equip these things in their listings: Smoke alarm, Wifi, Kitchen, Long term stays allowed, Hangers, Heating, Hair dryer, Air conditioning, Hot water… Besides, they should add these words in the name of their listing: private room, ocean view, beautiful, one bedroom, modern, charming, luxury. Because these are the words that the renters usually look for.

This will assist the hosts in obtaining more renters and increasing their revenue.

**References**

Airbnb, A. (n.d.). *Get the data - inside Airbnb. adding data to the debate.* Inside Airbnb. Retrieved October 6, 2021, from <http://insideairbnb.com/get-the-data.html>

Ray, S. (2015, August 14). *Regression techniques in machine learning - analytics Vidhya*. Analytics Vidhya. Retrieved October 6, 2021, from <https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/>

Shin, T. (2020, January 5). All machine learning models explained in 6 minutes. Medium. Retrieved October 6, 2021, from <https://towardsdatascience.com/all-machine-learning-models-explained-in-6-minutes-9fe30ff6776a>

Aleksandradeis. (2019, January 30). Airbnb Seattle Reservation Prices Analysis. Kaggle. Retrieved October 31, 2021, from <https://www.kaggle.com/aleksandradeis/airbnb-seattle-reservation-prices-analysis>.

bavalpreet26. (2019, October 21). Singapore\_Airbnb. Kaggle. Retrieved October 31, 2021, from <https://www.kaggle.com/bavalpreet26/singapore-airbnb/>.

Brittabettendorf. (2019, March 6). Predicting prices: XGBoost &amp; Feature Engineering. Kaggle. Retrieved October 31, 2021, from <https://www.kaggle.com/brittabettendorf/predicting-prices-xgboost-feature-engineering>.

Dgomonov. (2020, August 3). Data Exploration on NYC airbnb. Kaggle. Retrieved October 31, 2021, from <https://www.kaggle.com/dgomonov/data-exploration-on-nyc-airbnb>.

qwertyqwerty 11111 gold badge11 silver badge55 bronze badges, Marcelo Villa-PiñerosMarcelo Villa-Piñeros 5, &amp; TarasTaras 16.5k44 gold badges3030 silver badges8383 bronze badges. (1968, May 1). Convert lat/lon to zip/postal code using Python. Geographic Information Systems Stack Exchange. Retrieved October 31, 2021, from <https://gis.stackexchange.com/questions/352961/convert-lat-lon-to-zip-postal-code-using-python>.

Government, U. S. (2013). All US ZIP codes with their corresponding latitude and longitude coordinates. comma delimited for your database goodness. source: Http://www.census.gov/geo/maps-data/data/gazetteer.html. Gist. Retrieved October 31, 2021, from <https://gist.github.com/erichurst/7882666>.

1.6. nearest neighbors. scikit. (n.d.). Retrieved October 31, 2021, from <https://scikit-learn.org/stable/modules/neighbors.html#nearest-neighbors-regression>.