

Names: Aya Hajjeh and Bilan Aden
Kaggle usernames: ayahajjeh and BilanAdenn
Kaggle team name: Aya & Bilan
Airbnb Price Regression Project

One of the deliverables for the Airbnb price regression project is a project write-up. The goal of this write-up is for you to a) explain what you learned about the data, b) describe what processing and modeling choices you made, and c) hypothesize about how well your model worked given your choices. To give you a clearer understanding of what we expect in this write-up, we've provided this outline. Your project write-up should include the following sections and information.

Data

- Describe the dataset you worked with, including explanations of feature variables and the target variable

The dataset consists of Airbnb listings in New York City from 2008 to 2019 and our target variable was the price which was depicted on the training data set. The goal here was to build a regression model that estimated the price given specific airbnb features variable. Such features, included in the training and testing sets, are: ID, name, the host ID, the host name, the neighborhood group the airbnb is in, the neighborhood the airbnb is in, the latitude of the airbnb, the longitude of the airbnb, the room type, the minimum nights for booking, the number of reviews for this airbnb, the last review date, the reviews per month, the calculated host listings count, the availability per 365 days, the number of reviews last twelve months, the license, and lastly, the price column is only in the training dataset.


- Pick 3+ example data points, display them and describe them

The data points displayed here show the ID, name, host_ID, host_name, neighborhood_group, neighborhood, latitude, longitude, room_type, minimum_nights, number_of_reviews, last_review, reviews_per_month, calculated_host_listings_count, availability_365, number_of_reviews_ltm, license, price. For this specific apartment, it is located in Brooklyn and more specifically Fort Greene neighborhood, the latitude/longitude describe its specific 3 dimensionality location and it is an entire home apartment, last review was made in 2022 and the reviews it gets per month is super low, which makes sense why the number of reviews in the last 12 months is also a low number because the reviews

per month are less than 1.

0s	train.iloc[5]
id	a9918963
name	Brooklyn Apt with Patio in Heart of Fort Greene
host_id	50994782
host_name	Clay
neighbourhood_group	Brooklyn
neighbourhood	Fort Greene
latitude	40.68851
longitude	-73.97124
room_type	Entire home/apt
minimum_nights	2
number_of_reviews	31
last_review	2022-09-25
reviews_per_month	0.36
calculated_host_listings_count	1
availability_365	0
number_of_reviews_ltm	8
license	NaN
price	175
Name: 5, dtype: object	

Similarly, the same information is depicted here as well and some notable data is the reviews per month this specific apartment gets. It seems popular compared to our first data point, and the reviews are more recent in terms of date. And in terms of price it is cheaper which might lead to more people writing reviews on this specific place.

 train.iloc[8]

id	a726667728496451227
name	Charming 2 bedroom in a quiet tree lined street
host_id	188737645
host_name	John
neighbourhood_group	Bronx
neighbourhood	Pelham Gardens
latitude	40.86457
longitude	-73.84112
room_type	Entire home/apt
minimum_nights	3
number_of_reviews	10
last_review	2023-02-26
reviews_per_month	6.52
calculated_host_listings_count	1
availability_365	161
number_of_reviews_ltm	10
license	NaN
price	121
Name: 8, dtype: object	

And finally this specific apartment is in Brooklyn but underneath the neighborhood of L'Equipe. It also contains similar information described above.

```

▶ train.iloc[7]
📄 id a54364623
name Romantic and sunny 2-bedrooms rental unit Broo...
host_id 13187650
host_name L'Equipe
neighbourhood_group Brooklyn
neighbourhood Bedford-Stuyvesant
latitude 40.6893
longitude -73.95454
room_type Entire home/apt
minimum_nights 1
number_of_reviews 40
last_review 2023-02-28
reviews_per_month 3.05
calculated_host_listings_count 1
availability_365 44
number_of_reviews_ltm 38
license NaN
price 288
Name: 7, dtype: object

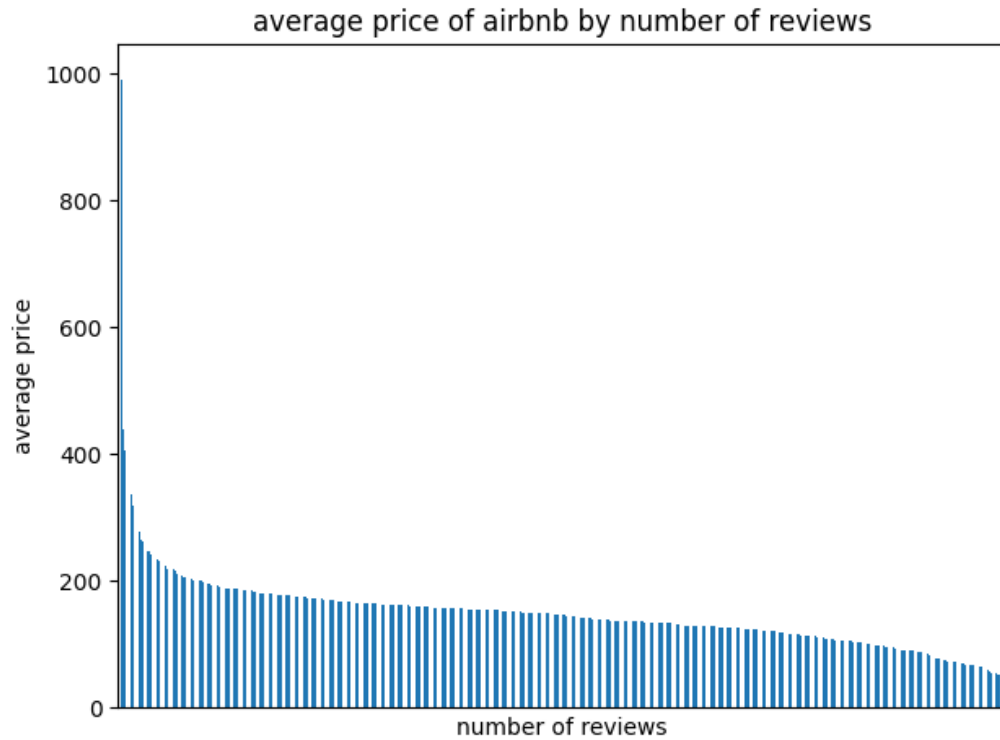
```

- Pick 3+ columns and describe why you hypothesize they might be useful

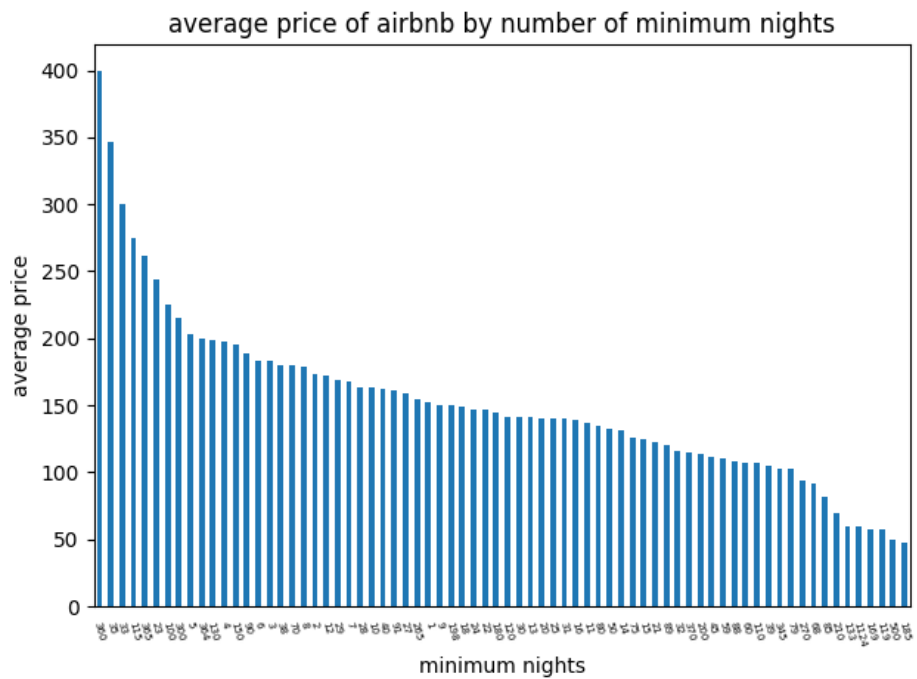
We selected three key graphs to provide valuable insights into the relationships within our data. These visualizations depict the connections between price and various factors, specifically neighborhood/neighborhood groups, room types, and the number of reviews. Our analysis reveals that room types play a significant role, as shared accommodations tend to be more budget-friendly compared to private hotel rooms, which are typically more expensive. Additionally, we observed that increased privacy often corresponds to higher prices. Another vital aspect we examined is the impact of neighborhood groups and individual neighborhoods on pricing. These graphs illustrate a clear correlation, with certain neighborhoods and neighborhood groups having higher or lower price ranges. For instance, the Bronx emerges as the most budget-friendly option, while Manhattan stands out as the most expensive location. In our analysis of reviews, we found that a reasonable price tends to attract more feedback, indicating a higher level of interest from guests. Conversely, when prices are excessively high, the number of reviews tends to be lower. This suggests that pricing can influence the level of customer engagement and feedback.

Methods

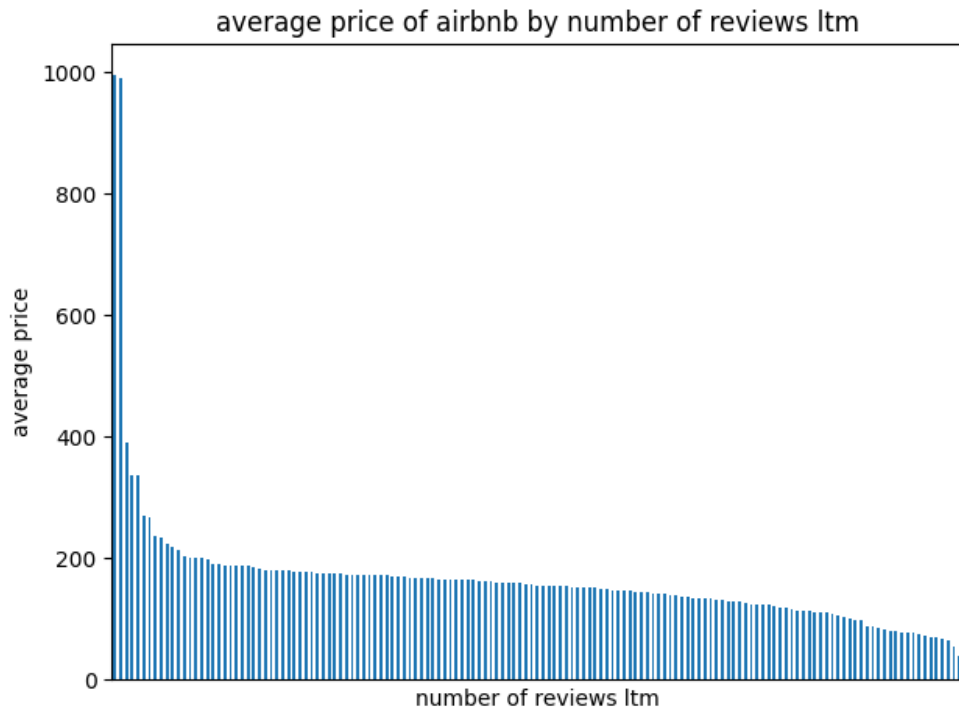
- Detail the steps you took to process your data
 - First we started with examining the shape of the training and test data sets and what type of features the data has
 - Then we captured the relation between the average price and the number of reviews



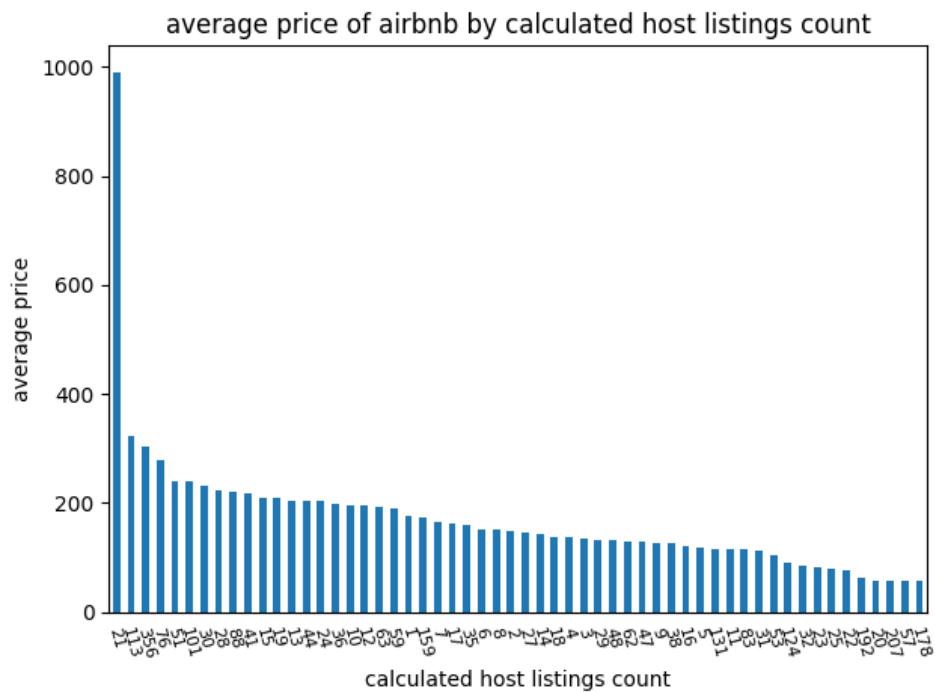
- Then the relation between the average price and the number of minimum nights



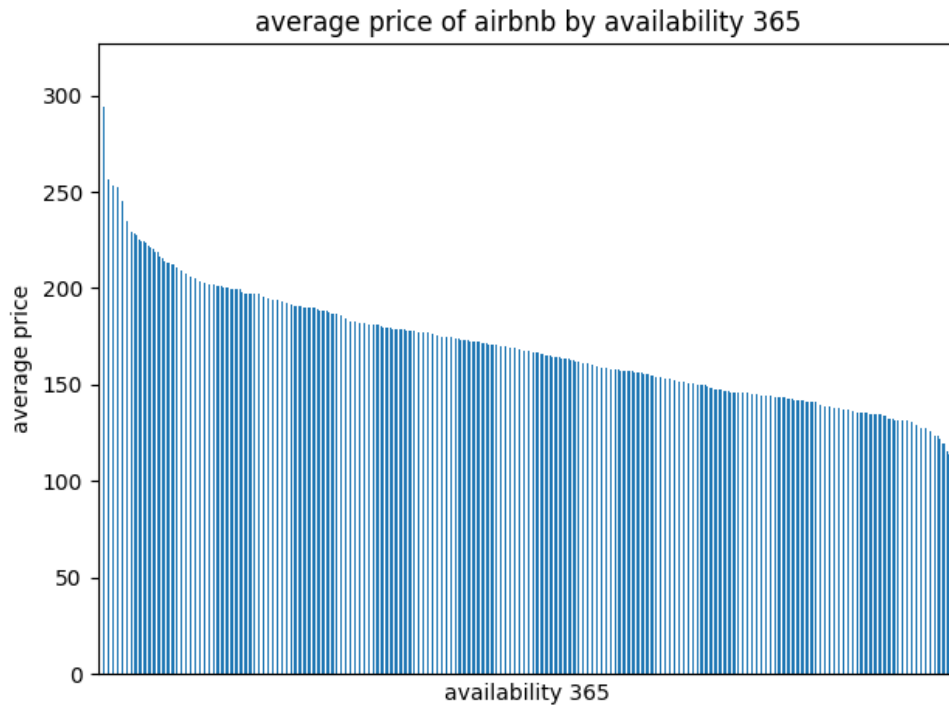
- Then the relation between the average price and the number of reviews ltm



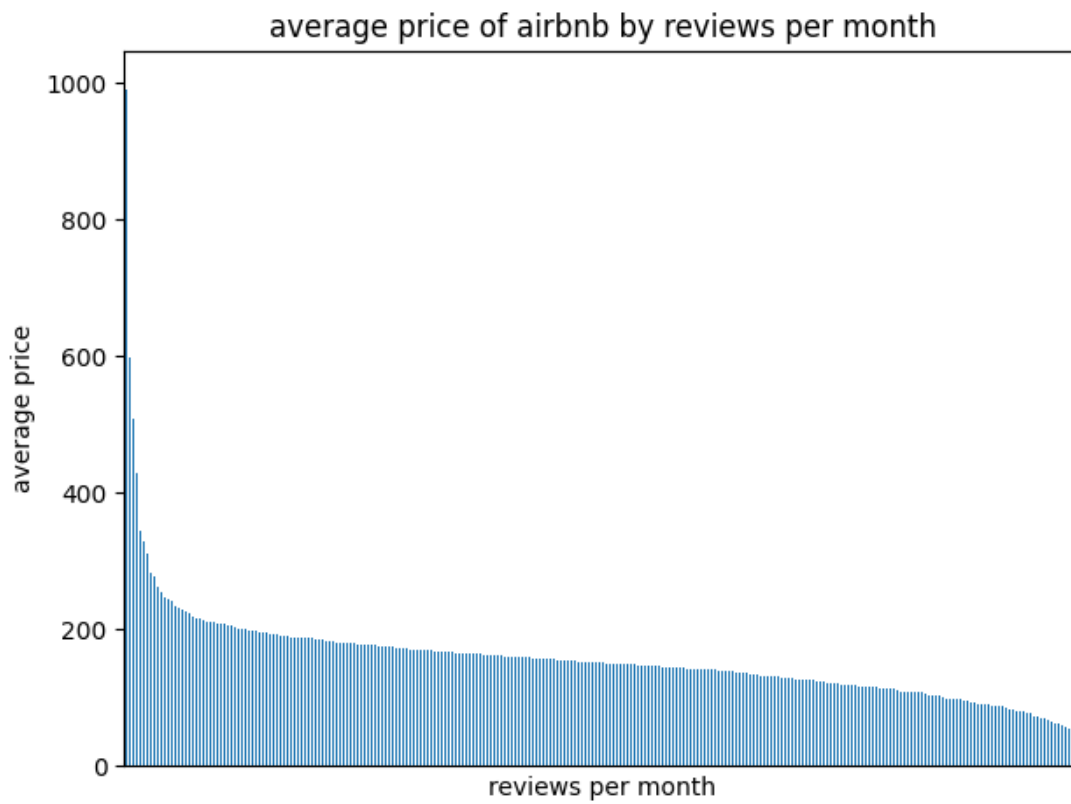
- Then the relation between the average price and the calculated host listings count



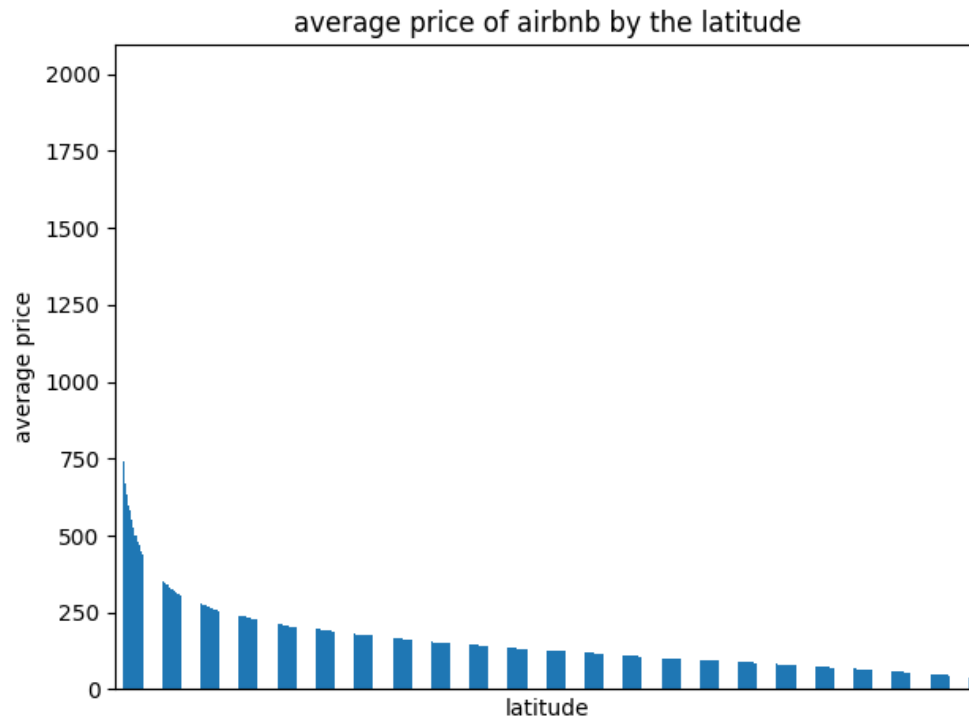
- Then the relation between the average price and the availability 365



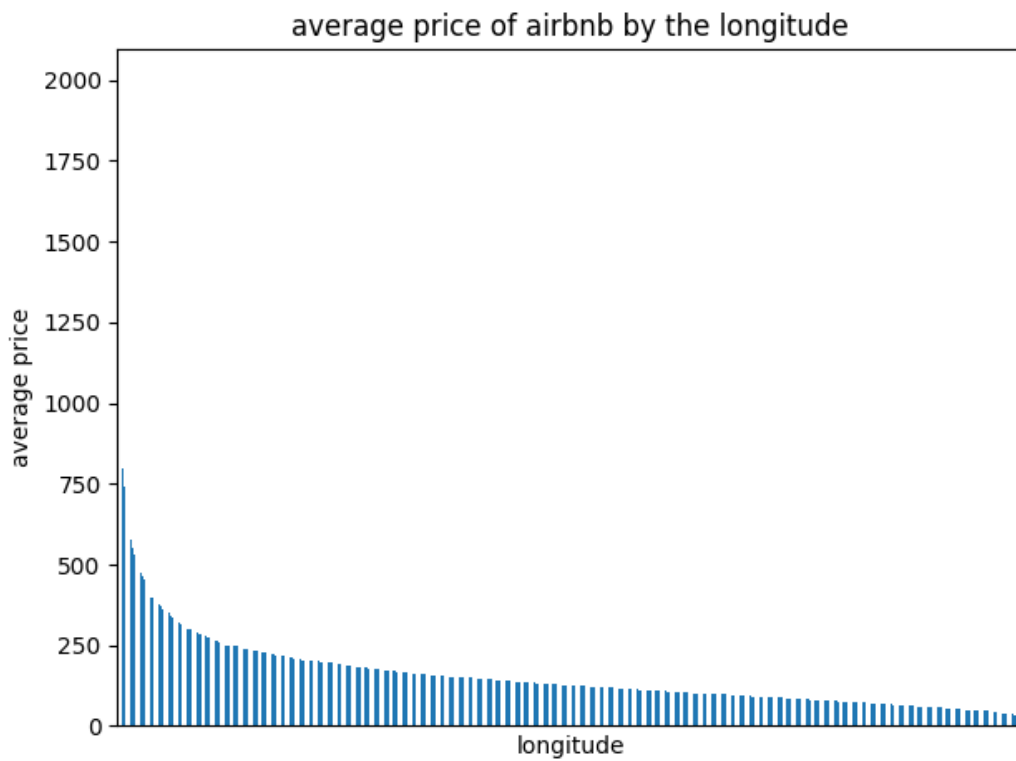
- Then the relation between the average price and the number of reviews per month



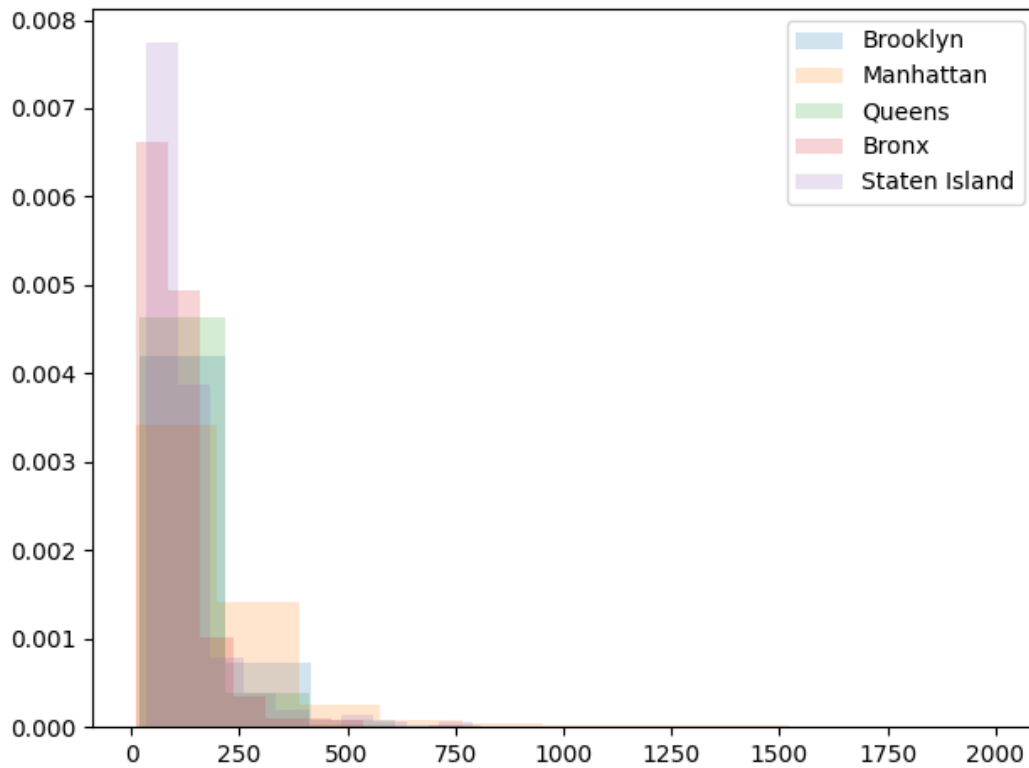
- Then the relation between the average price and the latitude



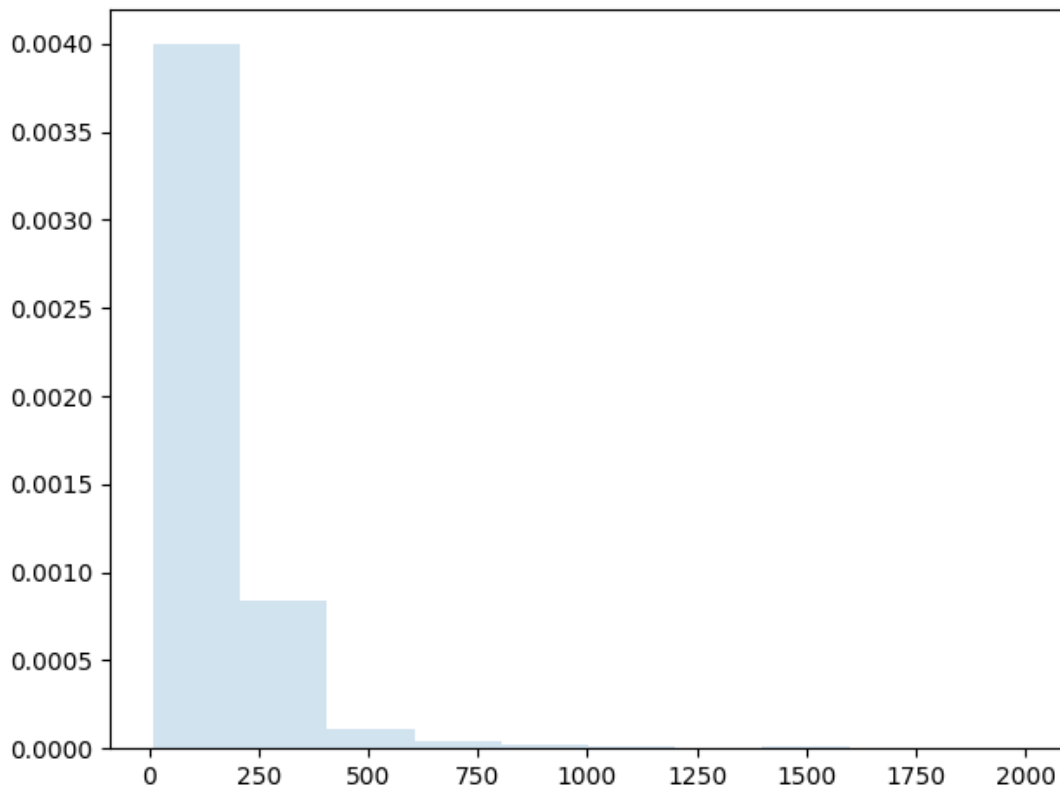
- Then the relation between the average price and the longitude



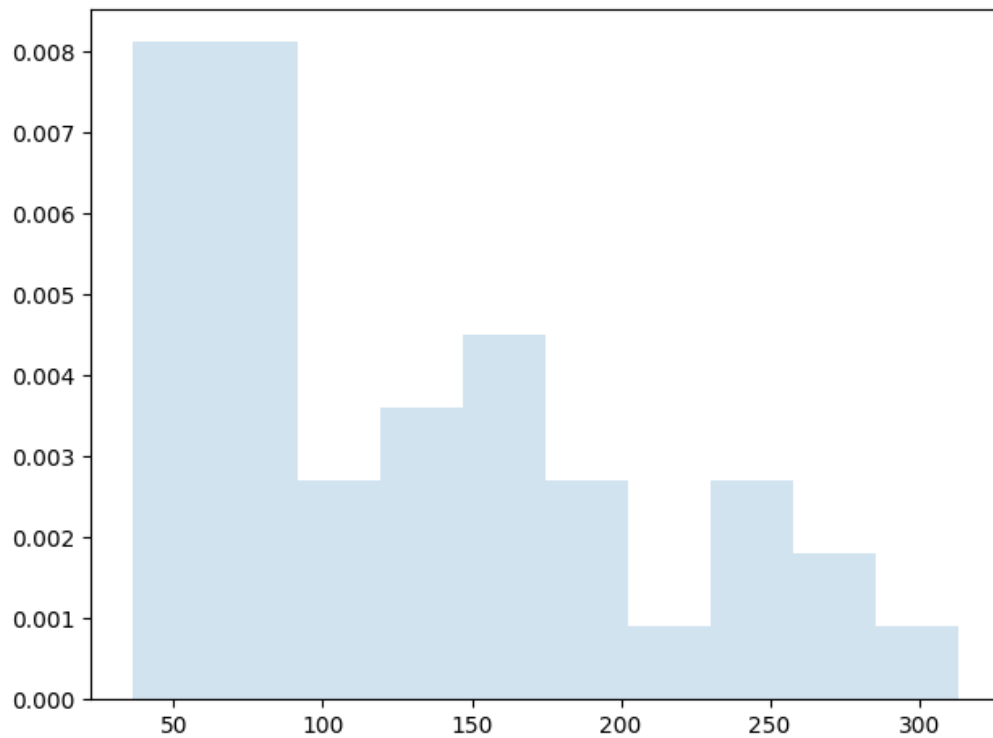
- Then the relation between the neighborhood group and the price



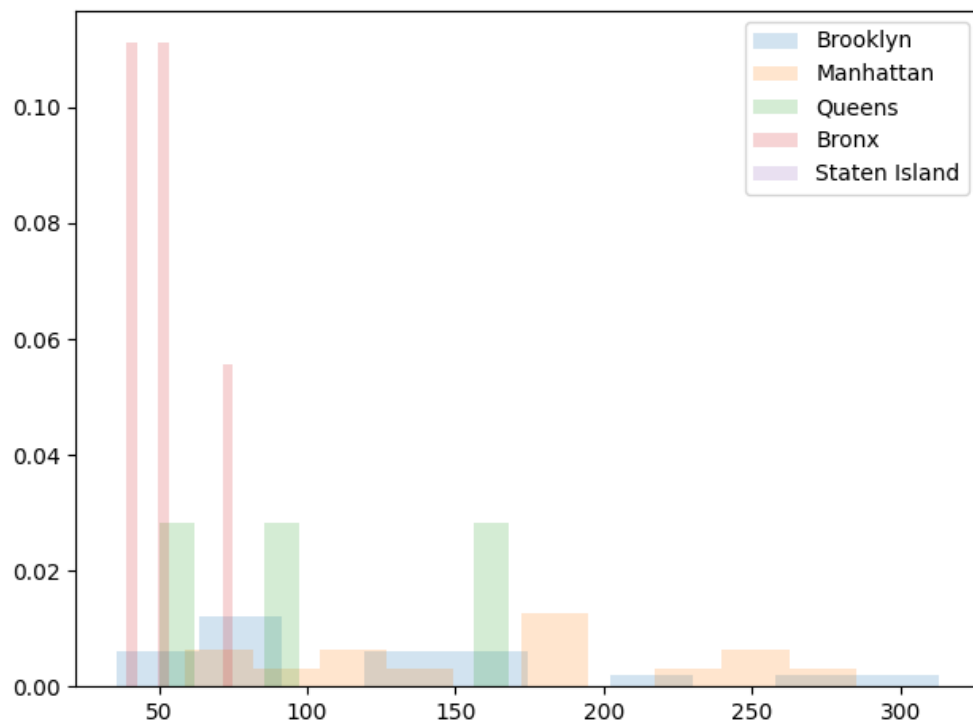
- Then we tried to understand the distribution of our data as a whole to examine if we need to reweight the training data



- Then tried to reweight our data to reduce the number of data points we have for prices between 0 and 250



- Lastly, we tried to examine the relation between the reweighted train data and the neighborhood group



- Describe the steps you took to build your model

- If you tried other models along the way, describe that process
 - First, for model #1, we started by building a regression model based on the features that we thought were most relevant according to the steps we took to process our data earlier. The features we considered for model #1 were ["reviews_per_month", "calculated_host_listings_count", "number_of_reviews_ltm", "minimum_nights", "number_of_reviews"] and our model scored 148.79405. To make it better, we thought maybe some features were not as relevant as we thought so we tried model #2 based off model #1 after eliminating the minimum_nights feature. The second model scored 149.57810, worse than the first model. Third, we eliminated the reviews per month feature instead of the minimum nights feature. Our third model even scored worse with a score of 150.40412. For our model #4, we tried to reweight the training data set because the vast majority of prices fell in the range 0 - 250 dollars. The reweighted model scored 158.77534, even worse than all the previous models. For model #5, we decided to go back to the original data set and not reweight the data since it scored worse. Moreover, we decided to include the neighborhood group feature in this model. We didn't include it first because we didn't know how to handle string columns, but then we converted the column into integer values – or in other words, we labeled the categories in the neighborhood column. The fifth model had scored the best so far with a score of 145.03586. Then, we decided to do the same model as model #5 but include the room type feature as well. We followed the same algorithm as before by labeling different categories in the room type column. Now, with the features from model #1 and the room type and neighborhood group features, our sixth model scored the best with a score of 135.94095.
- Detail the model you built and give a description of the hyperparameters and training
 - The features I included in the final model, ie model #6, are ["neighbourhood_group", "room_type", "reviews_per_month", "calculated_host_listings_count", "number_of_reviews_ltm", "minimum_nights", "number_of_reviews"]. We used RandomForestClassifier for our regression training from the Scikit Learn, with hyperparameters max_depth = 5, random_state = 27. We played with the hyperparameters a bit so we can find the best fitting model.

Procedure

- Justify the metric you used to evaluate your model (i.e. the metric described on Kaggle)
The metric used here was the Root Mean Square Error and this metric has been used to evaluate the model because RMSE is computed by first squaring the difference between the predicted value and the actual target value for each data point and then taking the square root of the average of these squared differences. In essence, it involves finding the squared errors for each data point and then finding the mean of these squared errors, followed by taking the square root of that mean and this is helpful because it penalizes the outliers.

Results

- List the score you achieved using Kaggle

- If you tried other models along the way, detail the scores you achieved using those models
 - model #1 score: 148.79405
 - model #2 score: 149.57810
 - model #3 score: 150.40412
 - model #4 score: 158.77534
 - model #5 score: 145.03586
 - model #6 score: 135.94095 (this is the final model we went with!)
- Describe potential future work: what would you do differently next time?
 - There isn't an even distribution of data along the price ranges. For example, most of the data points aggregate in the price range 0 - 250 dollars. Less data points are between 250 - 750 dollars, and there are barely any data points with prices above 750 dollars compared to the two previous price ranges. We already tried reweighting the data to make the distribution more even, but we just got a worse score in Kaggle. However, that try didn't have many features and that could be the reason why the score got worse, therefore, we might reweight the best model we got (model #6, the model with the most comprehensive and relevant features we built) to have a more even distribution in the data point prices. We also would add the neighborhood feature when building the model in the future as we anticipate it might be super relevant. Thus far, we only included the neighborhood group feature but not the neighborhood feature. Moreover, we might also try other training methods other than the random forest classifier while spending more time fine tuning the hyperparameters.