

New York AirBnb Data Mining

[Google Colab Environment Link](https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data)

<https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data>

Introduction

Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present a more unique, personalized way of experiencing the world. Our dataset describes the listing activity and metrics in NYC, New York for 2019. Throughout this project, our primary objective is to **uncover meaningful patterns within the New York City AirBnb dataset**. We are aiming to uncover patterns related to geographical data, pricing, and reviews in order to identify trends, popular regions, and potential correlations. Correlations are a key aspect of our project, since we will be using exploratory data analysis to understand how various factors within the dataset influence each other. Our main question then becomes how do we apply these findings to the New York City AirBnb market in a meaningful way? We want to focus on 3 to 4 major areas, including what we learn about different hosts and areas. We also want to explore what we can learn from predictions (locations, prices, reviews). We can also analyze which hosts are the busiest and why, which will allow us to provide insights that might reveal secrets about the New York Airbnb market. We could analyze if there are any noticeable differences in traffic among different areas and what the reason could be for these differences, and how that impacts Airbnb bookings in New York. Our findings will be valuable to hosts and guests and contribute to the broader understanding of the dynamics of short-term rental platforms such as Airbnb.

Objective

From this project, we anticipate to find correlations between the most booked AirBnbs and factors that affect Airbnb bookings, such as the host, room type, and location of the Airbnb. From the data, we will be making derivations and predictions about Airbnb bookings based on the data that we observe. Not only will we be using this data to explore the bookings process of AirBnbs, but we will be creating visualizations of the data to uncover patterns and reveal potential correlations between the data. We are hoping to empower Airbnb hosts with powerful insights that will help them drive booking metrics and learn more about the hidden rental

housing market!

Methodology

Before we start to analyze and create visualizations for the data, we preprocessed the provided data in order to explore attributes that introduce noise into the data, missing data, and outliers to ensure that the data is ready for analysis. For the visualization part of our project, we utilized several *Python* libraries (numpy to work with numerical data, pandas for data manipulation/cleaning, matplotlib and seaborn to create informative charts that visualize our data in several different ways) in order to support our preliminary analysis of the data. For the correlation analysis part, we used libraries like Scipy, Sklearn, and Statsmodels.api to perform more advanced statistical tests and prove our hypotheses. Overall, the main programming language used will be Python. Additionally, if we find any interesting or significant correlations between variables we can apply machine learning techniques to construct a model that tests how well we can predict one variable given 2 other ones, which can further deepen our insight into the Airbnb market in New York!

After the midterm report, the final will be focusing on identifying patterns related to the distribution for reviews, ratings, and trends over time. Our main question we want to answer is which metrics or features drive Airbnb listing price? This will be helpful to Airbnb hosts to know what to focus on in order to maximize profits and bookings for their listings, since they will take into account what truly drives pricing in this hidden housing market. First we will look into the correlation between price and availability out of 365 days a year, then correlation between location (neighborhood) and price, then correlation between price & number of reviews.

Preliminary Results

In the Preliminary Analysis we loaded and examined our dataset in order to get familiar with the different columns and attributes. This included getting familiar with packages as well. We learned about Numpy for numerical operations, used Pandas in order to turn our dataset into a Dataframe, and utilized Matplotlib in order to make the initial visualizations. Next, our goal for the midterm report was to have all the data clean, pre processed, and have a sense of what we want to achieve with this project going forward. We cleaned and preprocessed the dataset for analysis. This included handling missing data:

Here we can see that most of the columns/attributes do not have many missing values, however name, host_name, last_review, and reviews_per_month do have missing values. Since last_review, and reviews_per_month have more than 10,000

```
[62] missing_count = df.isna().sum()
      print(missing_count)

id                0
name              16
host_id           0
host_name         21
neighbourhood_group  0
neighbourhood     0
latitude          0
longitude         0
room_type         0
price            0
minimum_nights    0
number_of_reviews  0
last_review      10052
reviews_per_month 10052
calculated_host_listings_count  0
availability_365  0
dtype: int64
```

missing values, we will decide to drop the missing values.

```
[63] total_observations = df.shape[0]
      print(total_observations)

48895
```

There is a total of 48,895 rows, 10,052 *last_review* are missing/empty, 10,052 *reviews_per_month* are missing/empty. This is quite a large (20%) percentage of our dataset, but we don't want to drop the entire attribute, but instead we will drop only the missing values since we believe that reviews are going to provide meaningful insights. 16 names are missing/empty and 21 host_names are missing/empty We will be dropping these instances from the attribute since they contribute to such a small part of the dataset (<1%)

```
[64] # Dropping rows with missing "name" values
      df = df.dropna(subset=['name'])

      # Reset the index after dropping rows, so we don't have a pointer to an empty row
      df = df.reset_index(drop=True)
```

If we did this correctly, we should see 16 less rows $48,895 - 16 = 48,879$

```
[65] total_observations = df.shape[0]
      print(total_observations)

48879
```

Now, let's do the same for all 3 other attributes (drop the rows with missing instances)

```
[66] df = df.dropna(subset=['host_name'])
      df = df.reset_index(drop=True)

      df = df.dropna(subset=['last_review'])
      df = df.reset_index(drop=True)

      df = df.dropna(subset=['reviews_per_month'])
      df = df.reset_index(drop=True)
```

```
[67] total_observations = df.shape[0]
      print(total_observations)
```

38821

Now, let us confirm that there are no more missing values in our dataset, and all of the columns are still there:

```
[68] # let's confirm that there are no missing values:
      missing_count = df.isna().sum()
      print(missing_count)
```

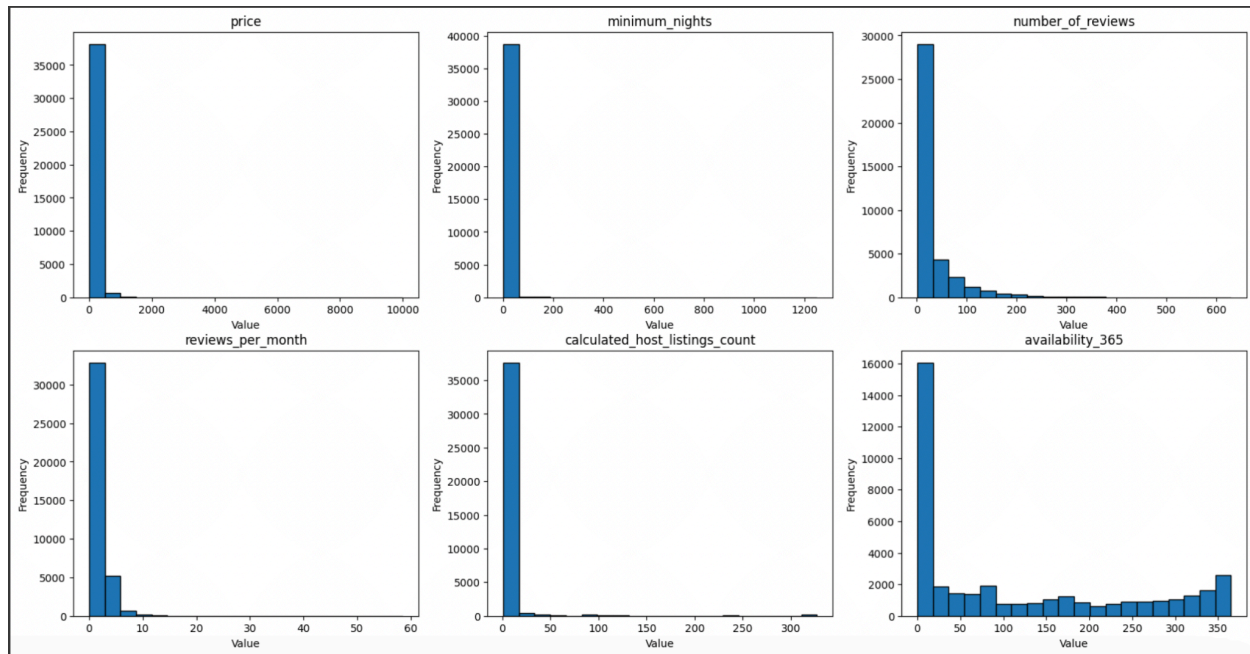
```
id          0
name         0
host_id      0
host_name    0
neighbourhood_group  0
neighbourhood  0
latitude     0
longitude    0
room_type    0
price        0
minimum_nights  0
number_of_reviews  0
last_review   0
reviews_per_month  0
calculated_host_listings_count  0
availability_365  0
dtype: int64
```

We cleaned the dataset and removed missing values (but we did not remove any attributes yet)! That could potentially be something we do during noise filtering. We will not be worried about removing outliers, since this is something we want to explore further in our analysis. Do we or do we not remove outliers? Visualization for outliers: focusing on numerical attributes

- > creating histograms to just get a feel of the data spread for each of the numerical variables
- > histograms are the top choice for numerical data

- price

- minimum_nights
- number_of_reviews
- reviews_per_month
- calculated_host_listings_count
- availability_365



These histograms give us a good initial understanding of the data, including some outliers. Since the primary purpose of our project is to understand trends and correlations (and potentially find interesting hidden patterns), we will not be removing outliers. Removing missing/empty data should be enough.

Final Results and Discussion

After the midterm report and preliminary results, we decided to focus on hosts and number of reviews, specifically identifying patterns related to the distribution of reviews, ratings, and trends over time. Our goal is to identify patterns related to the distribution of reviews, ratings, and trends over time. *Our main question we want to answer is which metric/feature drives price?* This will be helpful to airbnb hosts to know what to focus on in order to maximize

profits and bookings for their listings, since they will take into account what truly drives pricing in this hidden housing market.

(code below):

Correlation analysis between price and reviews:

```
[51] columnsToAnalyze = ['number_of_reviews', 'price']  
     dataClusters = df[columnsToAnalyze].dropna()  
  
     scaler = StandardScaler()  
     scaled_data = scaler.fit_transform(dataClusters)
```

```
[54] corrl = df[columnsToAnalyze].dropna()
```

```
[55] correlationData = corrl.corr()
```


```
[56] print(correlationData)
```

	number_of_reviews	price
number_of_reviews	1.000000	-0.035924
price	-0.035924	1.000000

Analysis of correlation:How price impacts number of reviews:

As the price increases, there is a very weak negative correlation with the number of reviews, indicating that, on average, higher-priced listings tend to have slightly fewer reviews. The correlation coefficient is approximately -0.036%, suggesting a minimal linear relationship between price and the number of reviews

Correlation analysis between availability out of 365 days and price

```
✓  columnsToAnalyze = ['availability_365', 'price']  
dataClusters = df[columnsToAnalyze].dropna()  
  
scaler = StandardScaler()  
scaled_data = scaler.fit_transform(dataClusters)
```

```
✓ [67] corrl = df[columnsToAnalyze].dropna()
```

```
✓ [68] correlationData = corrl.corr()
```

```
✓ [69] print(correlationData)
```

```
Os  
          availability_365    price  
availability_365    1.000000  0.078276  
price              0.078276  1.000000
```

Analysis of correlation: How yearly availability impacts number of reviews: As the price increases, there is a very weak positive correlation with the number of reviews, indicating that, on average, higher-priced listings tend to have more availability. The correlation coefficient is approximately 0.078%, suggesting a minimal linear relationship between price and the availability out of 365 days

Analysis of correlation (ANOVA): How location impacts price:

Here, our null hypothesis is that there is no correlation between the neighborhood and the price of the Airbnb:

Ho = neighborhood does not have an effect on price

Ha = neighborhood does have an effect on price

```
[ ] from statsmodels.formula.api import ols
    import statsmodels.api as sm

[ ] model = ols('price ~ neighbourhood', data=df).fit()
    anova_table = sm.stats.anova_lm(model, typ=2)
    print(anova_table)
```

	sum_sq	df	F	PR(>F)
neighbourhood	1.109660e+08	217.0	14.145412	0.0
Residual	1.395519e+09	38603.0	NaN	NaN

The ANOVA suggests that there is a statistically significant difference in mean prices among different neighbourhoods. The 'neighbourhood' variable appears to have an effect on prices. Since the P-value is 0, we reject the Null hypothesis H_0 that neighborhood does not have an effect on price. Therefore, one of the best features to predict price is location.

We hope that our results will provide helpful predictions that can be utilized by Airbnb hosts in the future so that they can maximize their profits and provide a more tailored experience to their customers. More specifically, we found that the positive correlation between the price and the availability_365 attributes entail that listing an Airbnb at a higher price may result in a trade off with the amount of bookings they receive in a given year. Another compelling finding from our research is that there exists a cause and effect relationship between price and neighborhood attributes, meaning that certain neighborhoods have Airbnbs that are on the higher end of pricing while others tend to be on the lower end. With this, Airbnb hosts could utilize this trend to ensure that they are pricing their properties accordingly to their respective locations so that they stay competitive on the market and are not missing out on any potential profit. Finally, we found a negative correlation between the number of reviews and the price. With this, hosts can be mindful during pricing that if they decide to price their Airbnb higher, then this could lead to fewer reviews and possibly their Airbnb being less popular and having a lower level of discourse among the public.

Next Steps:

In the next steps, one avenue that we could explore is looking at air bnbs in other major cities, like Washington DC, Boston, or Chicago for example. We could collect data for other cities and compare it with the data for airbnbs in New York. We could also see if the hosts that we explored in this assignment rented out air bnbs in the other cities and what patterns emerge from this data. From this we could compare and contrast the reviews, prices, and other features to see if they fit the models that we made based on the current data. Another next step would to look at the air bnb data for the current year and see how it has changed compared to the data that we explored this time, which form 2019. We could observe patters and see how the data has changed since 2019 to determine what air bnbs users rent the most. Another thing that we would look at in our next steps, would to word frequency analysis, and see which words correlates to hosts and prices to see how positive descriptions impact the price. Through this assignment, we have learned the skills that we would need to apply to pursue theses next steps in the future.