Airbnb's Impact on New York City vs Toronto

Isaac Wood

Abstract—In our research, we looked into the impacts that Airbnb has on urban areas, in particular, we focused on New York City and Toronto and seeked to discover how these cities are affected by Airbnb. Utilizing comprehensive datasets from Inside Airbnb, we employ advanced analytical techniques including GIS tools for mapping listing densities and machine learning models like Random Forest and XGBoost to predict revenue potentials. Ultimately, from this research, we were able to come to conclusions such as how the seasonality affects availability with decreased availability during the summer and holidays. We also make conclusions regarding what the most important features of an Airbnb listing are to determine the price of the listing, in both New York City and Toronto, that is the room type. This research on Airbnb's impact on New York City vs Toronto will help guide legislative bodies, city planners, Airbnb hosts, as well as consumers.

Index Terms—groupid, domain, Task keywords, type of project.

1 Introduction

The rise of Airbnb has significantly altered urban landscapes, influencing local economies, housing markets, and community dynamics. This disruptive influence has been particularly pronounced in major North American cities like New York City (NYC) and Toronto, where the platform has grown exponentially. The contrasting regulatory responses and economic impacts in these cities provide a compelling comparative study of Airbnb's effects on urban environments. This research is motivated by the need to understand how Airbnb's integration into these cities has reshaped housing availability, pricing strategies, and neighborhood characteristics, which are critical issues for urban planners and policymakers. We used Inside AirBnb's datasets for Toronto and NYC to carry on with our analysis.

1.1 Motivation and Research Problem

Our study investigates the specific impacts of Airbnb in NYC and Toronto, focusing on how the platform's presence correlates with changes in housing dynamics and local economies. The central problem is to delineate the role of Airbnb in urban housing markets and to assess its implications for residential communities in terms of rental prices, housing availability, and neighborhood transformation.

1.2 Main Contributions

To address this problem, we conducted a comparative analysis using a robust dataset from Airbnb listings, combined with demographic and economic data from both cities. We employed statistical and geospatial analysis techniques to extract patterns and derive insights into the platform's impact. The main contributions of our study are:

1) **Comparative Analysis of Airbnb Listing Densities:** How do listing densities differ between NYC and Toronto? This offers insights into urban demand for short-term rentals and their spatial distribution across city neighborhoods (RQ1).
2) **Seasonality Analysis:** How do specific times of the year effect Airbnb pricing strategies and occupancy?
3) **Economic Analysis of Airbnb Pricing Strategies:** Is it feasible to predict revenue potential of an Airbnb property using Random Forest and XGBoost?

• *Member 1, 2, and 3 are with School of Computing at Queen's University E-mail: put your emails here*

2 Related Work

"Inside Airbnb: Dallas" [1] provides a review of the in-depth analysis of Airbnb's impact on the Dallas housing market. The paper discusses the problem of how Airbnb's business model, particularly the dominance of entire home/apartment listings, affects the availability and affordability of housing in Dallas. It answers questions related to the growth of Airbnb listings, their distribution across different types of accommodations (entire homes vs. private or shared rooms), and the implications for local housing markets. The report's findings on the growth of entire home listings and their impact on housing availability provide a context for understanding the supply side of the Airbnb market. Areas with a high concentration of entire home listings might exhibit different pricing behaviors due to the scarcity of long-term rental options, potentially driving up short-term rental prices. This insight is particularly relevant for our project as it suggests that the availability of housing and the proportion of commercial listings could be key factors affecting price variations across different areas.

3 Methodology

3.1 Density Analysis Methods

3.1.1 Density Calculation

Listing densities were calculated to understand how Airbnb's presence is distributed spatially within each city. The process involved the following steps:

- **Spatial Grid Overlay:** Both cities were divided into a grid of equal-sized cells, each measuring 1 square kilometer. This grid facilitated the uniform aggregation of listings.

- **Listing Aggregation:** Listings were aggregated by their geographic coordinates into the corresponding grid cells.
- **Density Computation:** The density of listings per grid cell was computed by dividing the total number of listings in each cell by the area of the cell. This resulted in a measure of listings per square kilometer, which could then be compared across different regions within and between the two cities.

3.1.2 Hypothesis Testing of Densities

Statistical tests were used to assess whether differences in densities between comparable neighborhoods of the two cities were statistically significant. This was done using a 2-sample t-test. A 2-sample t-test is a statistical method used to determine whether the means of two independent samples are significantly different from each other. This test assumes that the samples are normally distributed and have similar variances. In the context of our study, the test was used to compare the average densities of Airbnb listings in each square kilometer grid cell for neighborhoods in New York City with those in Toronto. The null hypothesis states that there is no difference between the average densities of the two cities, while the alternative hypothesis suggests a significant difference. By setting a conventional significance level (usually 0.05), we can reject the null hypothesis if the p-value obtained from the t-test is less than this threshold, indicating that the observed differences in listing densities are statistically significant and not due to random chance.

3.2 Methods for Predicting Revenue Potential

3.3 Data Preprocessing

The initial phase of our analysis involved preprocessing the data obtained from Airbnb listings along with demographic and economic statistics from both cities. The preprocessing steps included:

- **Data Sampling:** Due to the large volume of data, we randomly sampled 10,000 listings from both New York City and Toronto to ensure manageability and computational efficiency.
- **Outlier Removal:** Outliers in the pricing data were identified and removed using the interquartile range (IQR). Prices that fell below the first quartile minus 1.5 times the IQR or above the third quartile plus 1.5 times the IQR were considered outliers and excluded from the analysis. This step was performed separately for each city to tailor the approach to local pricing distributions.
- **Data Cleaning:** Removal of incomplete entries, correction of inconsistencies, removing missing values, and dropping unimportant columns.

3.4 Feature Extraction

For our analysis, we selected features from the dataset based on their potential influence on the pricing and demand dynamics of Airbnb listings in New York City. The selected features include:

- **Neighbourhood Group:** The broader area in which the property is located, which impacts the desirability and pricing.
- **Neighbourhood:** The specific neighborhood, providing more granular control over location-based price variations.
- **Room Type:** Whether the listing is an entire home/apartment, private room, or shared room, which significantly affects pricing.
- **Minimum Nights:** The minimum number of nights required for booking, influencing the listing's availability and potential revenue.
- **Availability 365:** The number of days in the year the listing is available, indicating its potential to generate income.

These features were preprocessed for training as follows:

- **Handling Missing Values:** We ensured all selected features had complete data entries, particularly focusing on the 'price' feature to be used as the target variable.
- **Categorical Encoding:** The categorical variables ('neighbourhood group', 'neighbourhood', and 'room type') were encoded using one-hot encoding to convert them into a format suitable for model training.
- **Data Splitting:** The dataset was split into training and testing sets, with 80% used for training the model and 20% reserved for model evaluation.

3.5 Model Building

Our study employs two advanced machine learning models, Random Forest and XGBoost, which are well-suited for regression analysis due to their robust handling of non-linear data and their ability to manage overfitting. Here we delve deeper into each model's workings and their relevance to our study.

3.5.1 Random Forest Regressor

Random Forest is an ensemble learning method that operates by constructing a multitude of decision trees during training time and outputting the average of the predictions from individual trees for regression tasks [3]. This methodology is particularly effective due to several key features:

- **Bootstrap Aggregating (Bagging):** Each tree in a Random Forest is built from a sample drawn from the training set. This technique leads to better model performance because it decreases the model's variance without increasing the bias [3].
- **Feature Randomness:** When splitting a node during the construction of the tree, the choice of the split is no longer the best among all features. Instead, the split that is picked is the best among a random subset of the features. As a result, the forest model is less likely to overfit to the training data [3].

3.5.2 XGBoost Regressor

XGBoost (eXtreme Gradient Boosting) is a highly sophisticated machine learning algorithm that has gained popularity due to its speed and performance. Unlike Random Forest, which builds each tree independently, XGBoost builds

one tree at a time, where each new tree corrects errors made by previously trained tree. This sequential tree building process results in performance improvements, though it can be more susceptible to overfitting if not properly tuned [2]. Key aspects of XGBoost include:

- **Gradient Boosting Framework:** XGBoost uses this framework to produce a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It optimizes both the loss function and the model complexity using gradient descent [2].
- **Regularization:** Includes both L1 and L2 regularization terms in the loss function which helps in reducing overfitting and improves model robustness [2].
- **Handling Sparse Data:** XGBoost can handle sparse data automatically with a sparsity-aware algorithm for missing data and automatic handling of missing data values [2].
- **System Optimization:** XGBoost is designed to be highly efficient and scalable, and implements advanced computing techniques such as cache awareness and block structure to optimize hardware performance [2].
- **Parameter Tuning:** Offers extensive options for tuning, including tree characteristics (depth, minimum child weight), boosting parameters (learning rate), and regularization (gamma) [2].

Both the Random Forest and XGBoost models are integral to our approach, providing robust tools to analyze complex datasets such as those involved in urban economic studies on Airbnb's impacts. These models are utilized not only for their predictive accuracy but also for their ability to handle large datasets with many features, making them ideal for this research.

Model Evaluation: Both the models' effectiveness was evaluated using the following metrics:

- **Root Mean Square Error (RMSE):** Measures the average magnitude of the error. The square root of the mean square error.
- **R-Squared (R2):** Represents the coefficient that represents what percentage of the predicted variable can expressed through our regression model; it is how much variance it can detect with our variables,

3.6 Seasonality Detection

Detecting seasonality in price and availability data involves analyzing the patterns and trends that occur at regular intervals over time. It is not part of the traditional data analysis where we involve training and test data to predict variables. The seasonality was detected with non-traditional techniques such as:

Visual Inspection:

We plotted the time series data of price and availability over time. Use line plots or scatter plots with time on the x-axis and price/availability on the y-axis. Visual inspection can often reveal obvious patterns such as regular peaks and troughs that occur at specific times of the year.

Seasonal Subsetting:

We aggregated the data over different time intervals (e.g., months, quarters) and compared the aggregated values across different periods. If there are significant differences in price and availability between certain periods (e.g., higher prices during summer months, lower availability during holidays), it suggests seasonality.

Comparing the price and availability data across different seasons or time periods to identify consistent patterns or trends is the way to detect seasonality. Looking for recurring events or factors that influence prices and availability at specific times of the year (e.g., holidays, tourist seasons, local events) was key.

4 Dataset

Here's a brief overview of each dataset:

Airbnb NYC Dataset: The Airbnb NYC dataset typically includes information such as:

Listing information: Details about individual listings, including property type (e.g., apartment, house, condo), number of bedrooms, bathrooms, and maximum occupancy. Pricing: Nightly rental prices for each listing. Availability: Availability of the listing for different time periods (e.g., availability for the next 30, 60, 90 days, or 365 days). Location: Neighborhood, latitude, and longitude of the listing. Reviews: Number of reviews, average rating, and other review-related metrics. Analyzing the NYC dataset can provide insights into the demand for short-term rentals in different neighborhoods, seasonal pricing variations, and the impact of various factors on rental prices.

Airbnb Toronto Dataset: Similarly, the Airbnb Toronto dataset contains comparable information specific to Toronto, including:

Listing details: Similar to the NYC dataset, including property type, bedrooms, bathrooms, and maximum occupancy. Pricing: Nightly rental prices for each listing. Availability: Availability of the listing for different time periods. Location: Neighborhood, latitude, and longitude. Reviews: Review-related metrics. Analyzing the Toronto dataset can reveal trends in the city's short-term rental market, the popularity of different neighborhoods among travelers, and the impact of events or local attractions on rental demand and pricing.

Both datasets are valuable resources for researchers, analysts, and policymakers interested in understanding the dynamics of the short-term rental market, the sharing economy, and the broader impact on urban areas and local communities. They can also be used for predictive modeling, market forecasting, and studying the relationship between short-term rentals and the traditional hotel industry. Additionally, these datasets can be utilized for exploring themes such as affordability, gentrification, and regulatory implications associated with the growth of short-term rental platforms like Airbnb.

Preprocessing the Airbnb NYC and Airbnb Toronto datasets typically involves several steps to clean, transform, and prepare the data for analysis or modeling. Here's a general outline of the preprocessing steps we took:

Handle Missing Values:

Identify columns with missing values and decide on an appropriate strategy to handle them. Data Cleaning:

We checked for and clean any anomalies or inconsistencies in the data, such as outliers, incorrect data types, or invalid entries.

Feature Engineering:

Create new features that might be useful for analysis or modeling. For example, we changed the availability to represent the occupancy.

Encoding Categorical Variables:

We converted categorical variables into a numerical format suitable for analysis or modeling.

Data Splitting:

Split the data into training, validation, and testing sets when building predictive models. This ensures that we can evaluate the model's performance on unseen data. Handling Outliers:

Deciding on approaches to handle outliers in the data, such as removing them, or letting the outlines be expressed in the data. Feature Selection:

Select the most relevant features for the analysis or modeling task.

Basic Statistics of the Dataset:

Average density in NYC: 28.461909808574386

Average density in Toronto: 20.59460899034417

Median density in NYC: 8.927846765684603

Median density in Toronto: 8.116224332440547

Maximum count in NYC: 766

Maximum count in Toronto: 873

Average listing prices in NYC: 206.50294638809908

Average listing prices in Toronto: 177.03916842422416

5 Experiments and Results

5.1 Density Analysis (RQ1)

5.1.1 Visualizing Densities

To effectively convey the distribution and concentration of Airbnb listings in New York City and Toronto, heatmaps were created using geographic information system (GIS) tools. These heatmaps highlight the areas with the highest concentrations of listings in both cities, allowing for a visual comparison of urban hotspot locations. The color intensity on each heatmap corresponds to the listing density, with warmer colors indicating higher densities

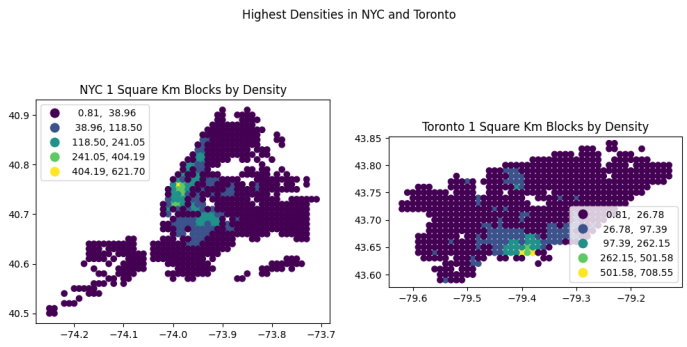

Fig. 1. Heatmap comparing Airbnb listing densities in Toronto and New York City.

5.1.2 Densest Neighbourhoods

The analysis identified neighbourhoods with the highest density of Airbnb listings in each city. For New York City, neighbourhoods such as Murray Hill, Hell's and the surrounding areas showed the highest densities, reflecting their popularity as tourist destinations. In Toronto, areas like Kensington-Chinatown and the Church-Yonge Corridor were identified as having the highest concentrations of listings. These findings help pinpoint where Airbnb's market penetration is most prevalent in each city.

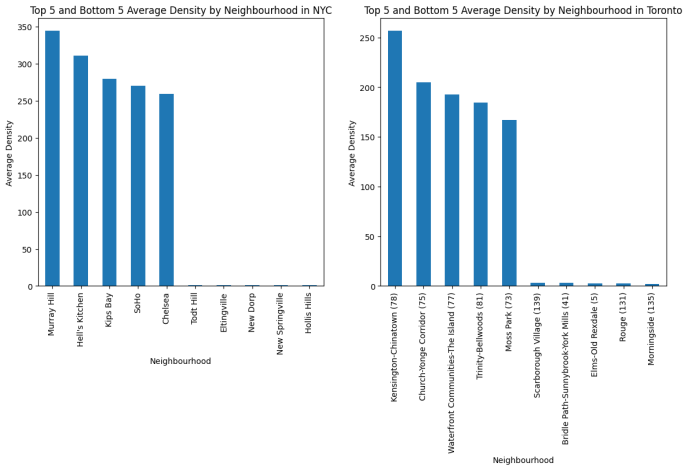

Fig. 2. Neighbourhood Density Across NYC and Toronto with the top 5 highest densities and the bottom 5 lowest densities.

5.1.3 Comparing NYC's and Toronto's Densities

The comparison of Airbnb listing densities between New York City and Toronto was quantitatively assessed using a 2-sample t-test. This statistical test was chosen to determine if there are significant differences in the average listing densities per square kilometer between the two cities.

5.1.3.1 Statistical Test Setup: The null hypothesis for our test assumed that there is no significant difference in the Airbnb listing densities between New York City and Toronto. The alternative hypothesis posited that a significant difference does exist.

5.1.3.2 Test Results: The results of the 2-sample t-test led to the rejection of the null hypothesis, indicating that the differences in listing densities between New York City and Toronto are statistically significant. This suggests that the two cities have markedly different Airbnb market characteristics and penetration levels.

5.1.3.3 Implications of Results: The rejection of the null hypothesis has several implications:

- **Market Dynamics:** The significant difference in densities may reflect differing urban policies, tourism appeal, or economic conditions between New York City and Toronto.
- **Strategic Planning:** For Airbnb and potential hosts, these insights can guide strategic decisions about where to focus marketing and development efforts based on the saturation and potential growth areas in each city.

5.1.4 Limitations

While the analysis provides valuable insights into the differences in Airbnb listing densities between New York City and Toronto, there are limitations that must be considered. The assumption that listing density correlates directly with market saturation may overlook other influential factors such as seasonal variations, local regulations, or economic events that could affect Airbnb activity.

5.2 Seasonality Analysis

5.2.1 Null and Alternative Hypothesis

Time and seasons are represented differently in the Airbnb datasets as time is represented with availability. The null hypothesis will be that the seasonality does not have any effect on the price and occupancy rate of Airbnbs in Toronto and NYC. An alternate hypothesis is that Christmas time and summer time will increase the price and the occupancy of Airbnbs in both Toronto and New York City.

5.2.2 Correlation Matrix Analysis

	availability_30	availability_60	availability_90	availability_365	price
availability_30	1.000000	0.934552	0.862161	0.393416	0.076501
availability_60	0.934552	1.000000	0.966671	0.434935	0.063721
availability_90	0.862161	0.966671	1.000000	0.484116	0.056923
availability_365	0.393416	0.434935	0.484116	1.000000	0.026108
price	0.076501	0.063721	0.056923	0.026108	1.000000

Fig. 3. Toronto correlation matrix comparing the seasonality of availability and price

	availability_30	availability_60	availability_90	availability_365	price
availability_30	1.000000	0.936554	0.870733	0.403313	0.020942
availability_60	0.936554	1.000000	0.967117	0.471680	0.019820
availability_90	0.870733	0.967117	1.000000	0.540942	0.019549
availability_365	0.403313	0.471680	0.540942	1.000000	0.023260
price	0.020942	0.019820	0.019549	0.023260	1.000000

Fig. 4. New York City correlation matrix comparing the seasonality of availability and price

Based on the results given by the correlation matrix of both Toronto and New York City; they display similar patterns. Both seem to accept the null hypothesis where seasonality does not seem to affect the prices or the availability. However, this is a preliminary statistical analysis in the experimentation process and is not to be taken as a conclusion.

5.2.3 Detecting Seasonality

Using statistical analysis methodologies described in the methods section of the report, seasonality was detected within the data. A visual representation of the analysis can be shown between both cities:

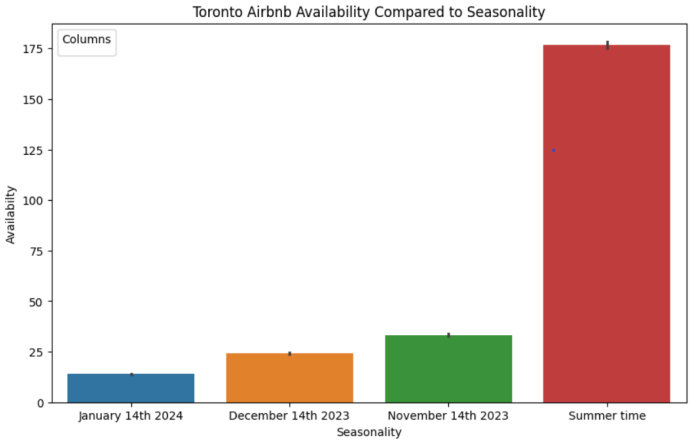

Fig. 5. Toronto Seasonality Bar Chart

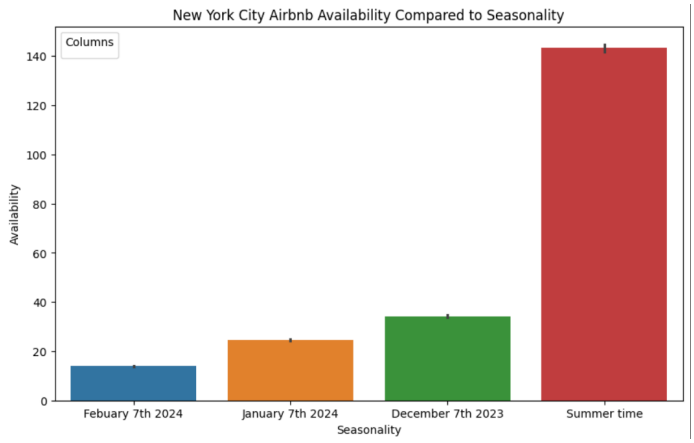

Fig. 6. New York City Seasonality Bar Chart

Based on the analysis done to detect the seasonality of the data, the data seems to point more to rejecting the null hypothesis and accepting the alternative hypothesis. In Toronto, wintertime has around a 60 percent availability rate and it goes down to about 50 percent around Christmas time which is also around to be the rate in the summer. This shows that occupancy rates increase due to seasonality in Toronto. We experience a similar trend with New York where the availability rate is at 60 percent during the winter, but then it drops to 40 percent during Christmas time and summer time. This describes a similar trend to the seasonality in Toronto. The data in the seasonality comparison to price is quite dense, however using data analysis one can extrapolate that prices seem to jump a little bit during Christmas time compared to those year-round. However, due to the sheer size of the data and its incompleteness, it is safe to conclude that for Toronto, the seasonality of summer and Christmas only seem to affect prices slightly on the higher side. The New York City data is more dense than the Toronto dataset, however, from the data analysis, we can extrapolate a conclusions. There does seem to be a trend in the price of the Airbnbs and the Christmas season. There does seem to be a reasonable uptick in the price of Airbnbs and their prices.

5.3 Airbnb Pricing Strategies

5.3.1 *Price Analysis*

The average of the listings was computed by performing a simple calculation of summing the total price of all the listings and then dividing it by the number of listings. This revealed that the average listing price in NYC was 30 dollars more than the average listing price in Toronto.

5.3.2 *Random Forest*

Before performing Random Forest Tests we needed to remove outliers and take 1000 samples from both the NYC and Toronto datasets. Then we used price as target and try to predict how It changes based on the 'neighbourhood group', 'neighborhood', 'room type', 'minimum nights', 'availability 365 features.

The results yielded very similar R2 scores with the NYC being 0.327 and Toronto being 0.314. The mean Absolute Errors were 51.06 for NYC and 51.18 dollars for Toronto.

The slight variance in results can be explained by NYC having an extra feature that is the neighborhood group and Toronto not having this feature, resulting in a slight difference when using random forests to predict the price feature.

5.3.3 *XGBoost*

For XG Boost, the NYC dataset had a significantly higher R2 value of approximately 50 and the Toronto dataset had a R2 value of approximately 40. They both had relatively similar RMSE scores with NYC having an Error of approximately 61 dollars and Toronto having an error of 60 dollars. This tells us that the NYC dataset has a stronger correlation between it's features compared to the Toronto dataset. Hyper-parameter tuning: I made the max depth equal to the amount of features used and adjusted the parameters based on the results. The best learning rate found was a low value of 0.1 with the 100 trees used and increasing it to 0.3 and 0.2 led to lower R2 values. The Toronto Dataset worked better with a slightly higher subsample value of 0.8 compared to the 0.75 of NYC and a lower max depth of 5 compared to the 6 of NYC.

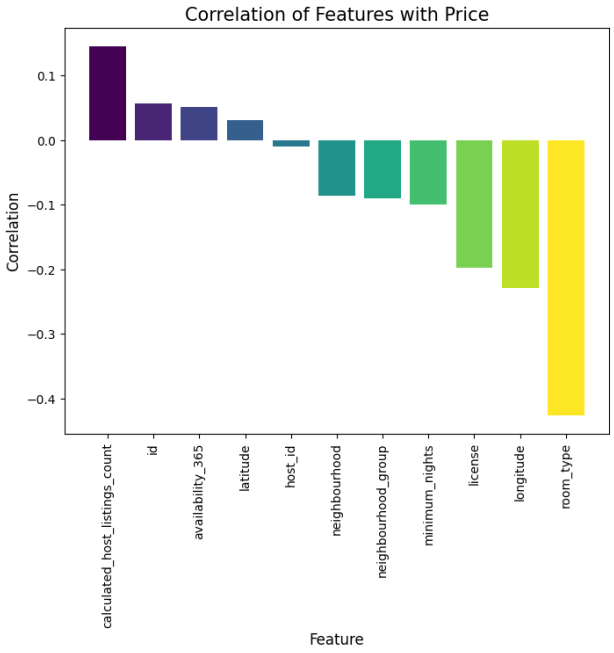

Fig. 7. Bar Chart of Correlation of Features with Price for NYC

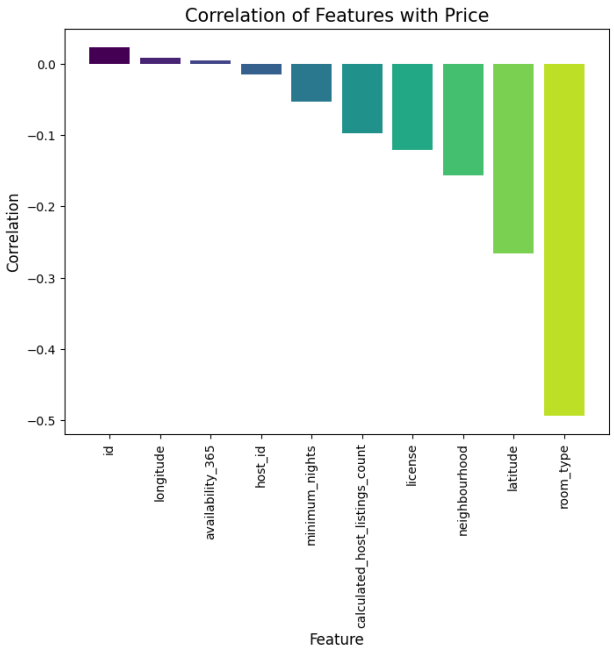

Fig. 8. Bar Chart of Correlation of Features with Price for Toronto

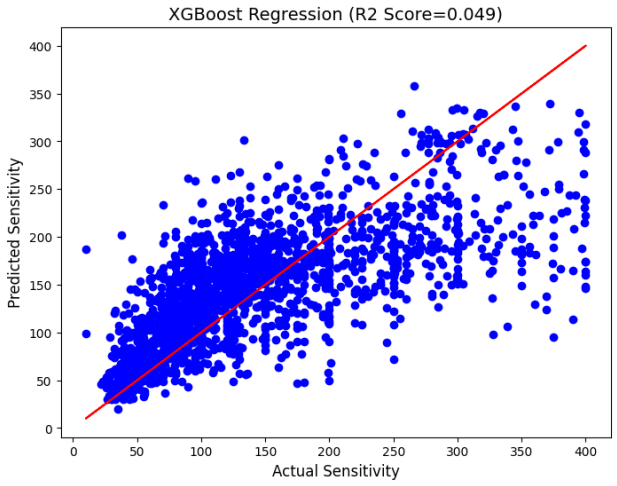

Fig. 9. XGBoost Regression Plot for NYC

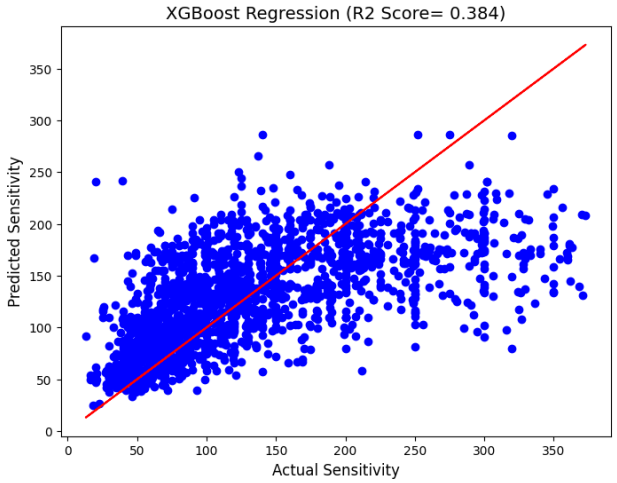

Fig. 10. XGBoost Regression Plot for Toronto

6 Replication Package

https://github.com/twentysixfiftythree/CISC-372-Project

Get the data from here:
https://insideairbnb.com/get-the-data

7 Conclusion and Future Work

In conclusion, we have discovered that the average and median density of Airbnb listings is greater in New York City than in Toronto, this result makes sense given that New York City has a denser population than Toronto. The maximum amount of listings in Toronto exceeds the amount in New York City, this is a somewhat surprising result given that New York City is generally thought of as a more attractive destination to visit. The average listings of prices in New York City is a bit more expensive than Toronto, this makes sense given that renting in New York City costs more as well. It was determined for both New York City and Toronto that specific times of the year affect the availability of Airbnb's and thus allows hosts to change their listing price accordingly to maximize profit. In the wintertime, the availability rates increase in both New York City and Toronto, during Christmas time, these availability rates drop to a similar level to the lower availability rates or Airbnb listings in the summer. Based on correlation graphs, we are able to conclude that the type of room for an Airbnb listing is the most significant factor in predicting the price. The correlation graphs provide further insights such as how the latitude of a listing is important for predicting price in Toronto but not New York City and how longitude is important for predicting price in New York City but not Toronto.

In future work, we will seek to explore how Airbnb impacts other cities besides New York City and Toronto, and then we plan to compare those differences to New York City and Toronto. We also look forward to further investigations on the impacts that Airbnb has on New York City and Toronto by applying artificial intelligence using deep learning.

References

[1] M. Cox. Inside Airbnb data hid the facts in New York City. *Inside Airbnb*, 2023.
[2] guest_blog. Introduction to xgboost algorithm in machine learning. 2024.
[3] IBM. What is random forest. 2024.