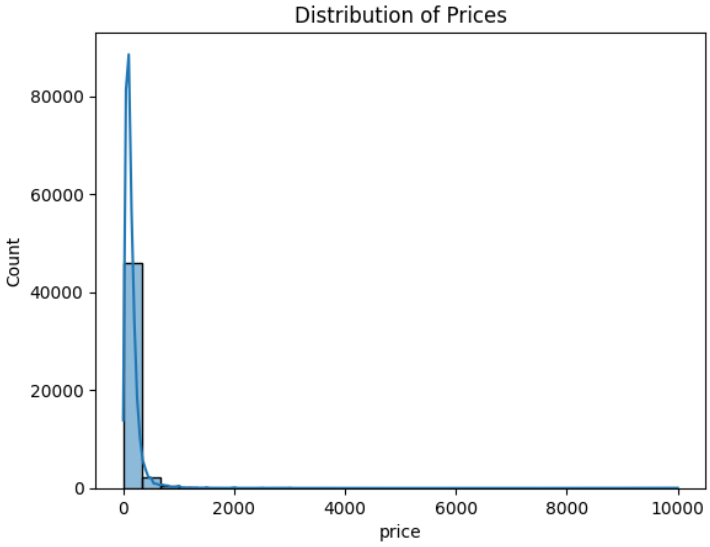
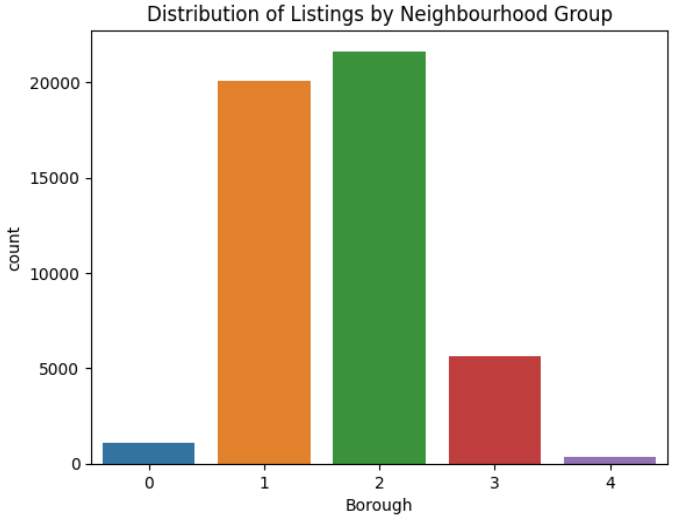
Introduction

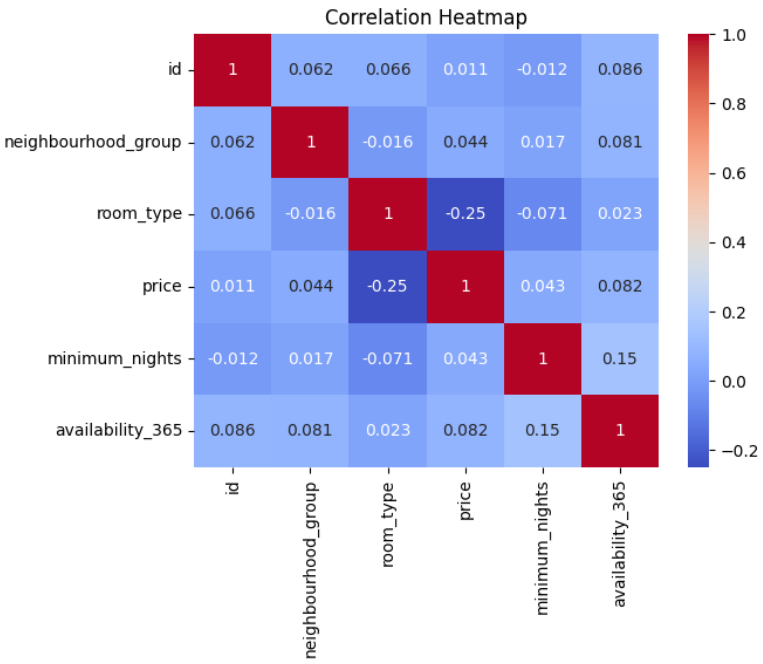
In today's ever evolving travel industry, predicting Airbnb prices has become a pressing challenge for those looking to travel as well as hosts. In this project, I aimed to analyze the New York City Airbnb dataset with the end goal of predicting listing prices. The main questions I planned to answer was understanding the factors influencing prices and creating a predictive model to assist hosts and potential guests in setting or understanding rental prices.

About the Data

I decided to use a dataset from Kaggle, specifically the "New York City Airbnb Open Data" dataset. The dataset consists of various features, including 'neighbourhood_group,' 'room_type,' 'price,' 'minimum_nights,' 'availability_365,' and others. The dataset had a manageable size and clear names. To better understand the data,I looked into the correlation between different features in relation to prices to see if there were any obvious existing patterns. Visualizations, such as histograms of the price distribution, a countplot of listings by borough, and a heatmap displaying feature correlation.







Methods

I started by pre-processing the data by dropping rows with missing values and encoding categorical features using label encoding. The model of choice was a Random Forest Regressor with 100 estimators.

I initially focused on a subset of features, including 'neighbourhood_group,' 'room_type,' and 'minimum_nights,' in order to build a baseline predictive model. In the next Iterations, I experimented with different feature sets and model parameters. For instance, in the second iteration, I incorporated 'calculated_host_listings_count' and adjusted the number of estimators to 150. In the third iteration, I further refined the features to use only 'neighbourhood_group', 'neighborhood', 'room_type', 'price'. I also adjusted the number of estimators to 200 in this 3rd iteration. Each iteration aimed to enhance model performance in order to capture additional insights.

Evaluation

I evaluated model performance using the Root Mean Squared Error (RMSE). The initial model achieved an RMSE of 225.71. The second iteration showed a mistake causing an increase with an RMSE of 225.61, and the third iteration achieved the lowest RMSE of 211.44. These results indicate that refining the feature set and adjusting parameters contributed heavily to a more accurate pricing prediction model.

Storytelling

Through this project, we were able to gain insights into the significant influence of certain features that have an effect on prices of Airbnb listings in New York City. While successfully answering the initial questions, there are opportunities for improvement, such as by exploring additional features and further refining the modeling process. This project has demonstrated the

importance of continuous improvement in model development and how that can help gain better, more accurate results..

Impact Section

This project is a perfect example of how using data can help us make smarter decisions in many areas of life such as fair prices for an Airbnb given the neighborhood. The impact is twofold, benefiting hosts by empowering them to set competitive prices and aiding guests in making informed choices. However, there is a need for caution which is essential in order to avoid unintentionally contributing to an abuse of the information or skewed data. Responsible and ethical data analysis practices, including careful feature selection and ongoing evaluation for biases, are crucial. The project aims to strike a balance between its positive impact on hosts and guests while considering potential implications for the Airbnb market dynamics.

Github Repository/Code & Data Used

The complete code is available in the GitHub repository: https://github.com/andrew273/ITCS-3162-Final-Project.

The dataset used can be accessed on Kaggle: https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data/.