CFG +Masters A.I/ ML - Assignment 1 - Avril Childs

Selecting Serviced Accommodation in Essex - Occupancy and Revenue analysis and prediction

Real-World Scenario

A holiday lettings company is planning to expand its portfolio of serviced accommodations across Essex. The company seeks assistance in identifying the most suitable investment properties by analysing historical data on occupancy rates and annual revenue of similar listings. It also wishes to look at other possible influencing factors in a listing's earning potential.

Aims and Objectives

The project is structured into two main phases:

Phase 1 : Identify property and listing features that impact occupancy rates and annual revenue

Phase 2: Build a machine learning model that predicts which properties are likely to achieve high occupancy rates and revenue.

This assignment is focused solely on Phase 1, which involves

- Data cleaning
- Data standardisation
- Data preprocessing
- Data transformation
- Exploratory data analysis

Study Features (Independent Variables)

Category	Feature	Description	Data Type
Property	Bedrooms	Total bedrooms in the property	Discrete Numerical
	Bathrooms	Total bathrooms in the property	Discrete Numerical
	Guests	Maximum number of guests allowed	Discrete Numerical
	Location (Town)	Town in Essex where the property is situated	Categorical
Host	Host Ratings	Average rating given to the host	Continuous Numerical
	Number of Reviews	Total reviews received	Discrete Numerical

Amenity	Seafront	Whether the property is located near the sea	Binary
	Hot Tub	Availability of a hot tub	Binary
	Pets Allowed	Whether pets are allowed	Binary
	Smoking Allowed	Whether smoking is permitted	Binary
Listing	Length of Stay	Min/max nights allowed per stay	Discrete Numerical
	Active Days	Days the listing is bookable	Discrete Numerical
	Blocked Days	Days the listing is blocked	Discrete Numerical
	Price	Nightly price of the listing	Continuous Numerical

Target Variables

Annual Revenue Total earnings per listing over a 365-day period. This depends on price and occupancy.

Occupancy Rate Percentage of days a listing is booked versus days it is active

The target variables will be categorised into high and low for correlational study and modelling. The exploratory analysis within phase 1 will assist in determining the thresholds.

The Data

The data was retrieved from PriceLabs (https://www.pricelabs.co), a subscription-based platform used by property investors, landlords and agents. PriceLabs aggregates serviced accommodation data from platforms like Airbnb, Booking.com and Vrbo.

- Two .csv datasets were downloaded for analysis
- The file paths within the Juptyer Notebook to match user local file system in order to use the dataset correctly.

Data Cleaning and Standardisation

The following data cleaning and standardisation procedures were conducted:

Remove duplicate listings: Duplicates can occur within the dataset as one property can feature within 2 overlapping locations (towns). The data set was checked for any duplicate listings based on listing ID and discovered there were 186 duplicates. These duplicates were dropped.

Remove irrelevant columns: The columns that were irrelevant to further analysis were removed.

Missing Values: A heatmap was used to visualise the patterns of null/missing values within the dataset.

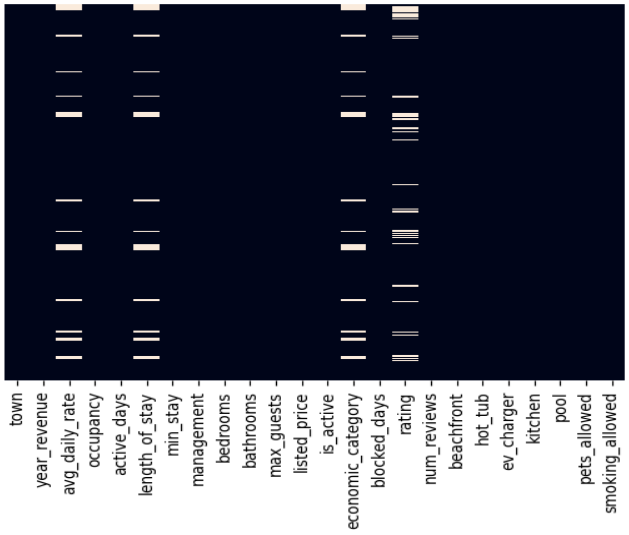

Fig.1 Heatmap of Missing values

Outlier management: Outliers within the numerical columns were visualised using boxplots. Further understanding and exploratory analysis was conducted to determine outlier causes and thresholds.

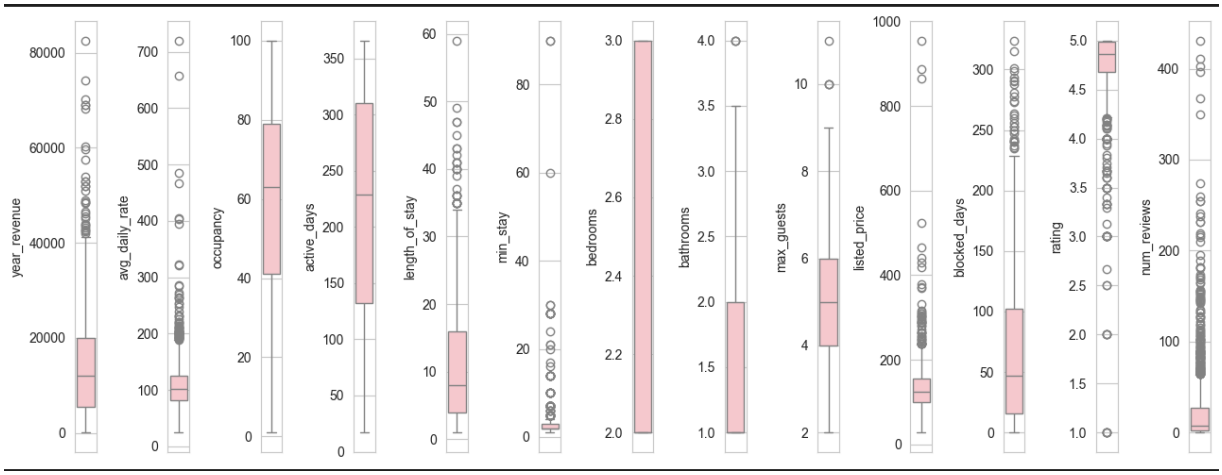

Fig. 2. Boxplots for outlier detection

Winsorization method was used to reduce impact of extreme values on the dataset and was applied to the following:

- Annual revenue
- Average Daily Rate
- Listed Price

Capping (Thresholds) was deemed more appropriate in addressing outliers for the following:

- Minimum length of stay
- Number of blocked days

Exploratory Data Analysis

Descriptive analysis of features were analysed using comparative tables and data visualisations such as pie charts, distribution graphs.

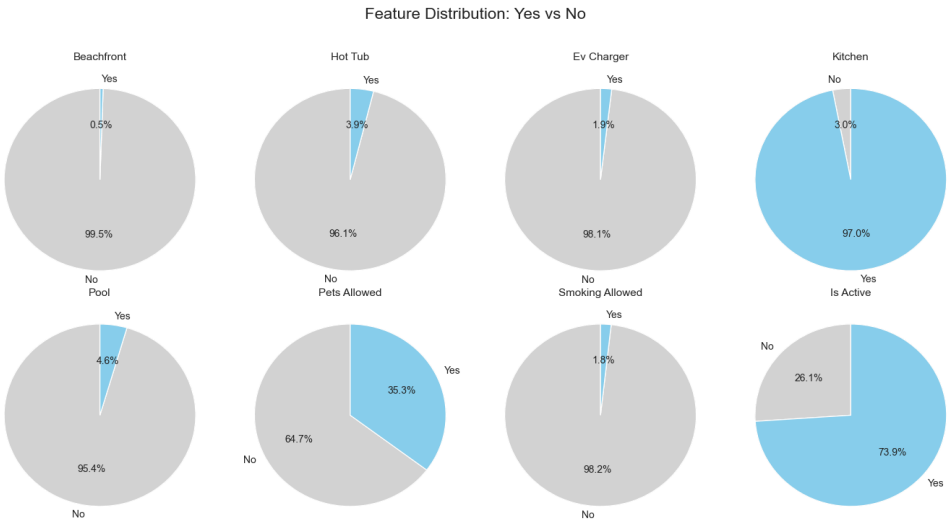

Fig 3. Pie chart demonstrating proportion of yes/no values in binary features

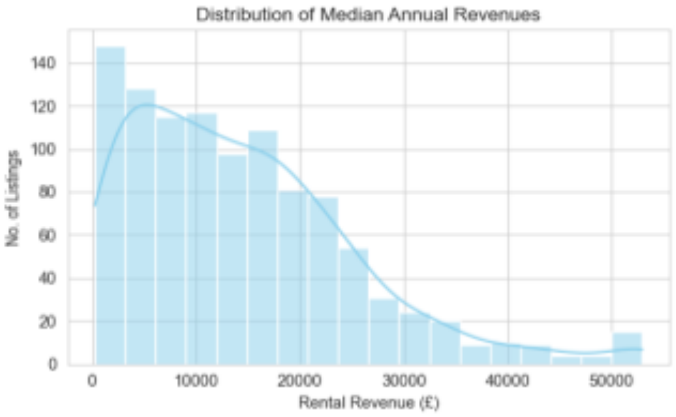

Fig. 4 Distribution plot for annual rental revenue

Both revenue and occupancy data and their percentile values were analysed to determine threshold values of “High” and “Low” which is used for correlational/comparative study and modelling:

High Revenue (£): ‘year_revenue’ > = 12500
Low Revenue (£): ‘year_revenue’ < 12500

High Occupancy (%): ‘occupancy’ > = 60
Low Occupancy (%): ‘occupancy’ < 60

Comparative Analysis of features and target variables were dealt with depending on whether the feature is a continuous/discrete numerical or categorical/ binary.

Continuous/ discrete numerical data were visualised using violinplots.

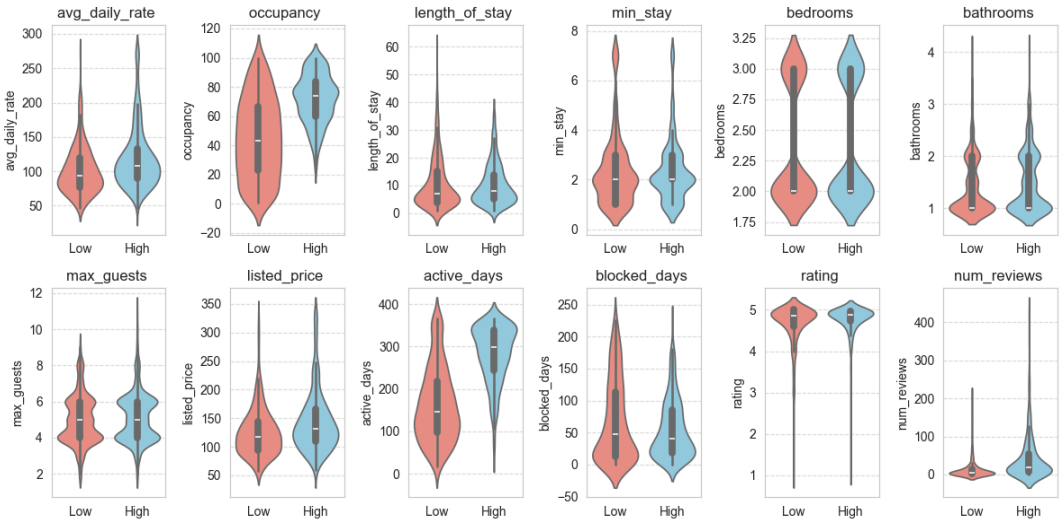

Fig.5 Numerical Features vs Revenue Violinplots using High vs Low Revenue Markers

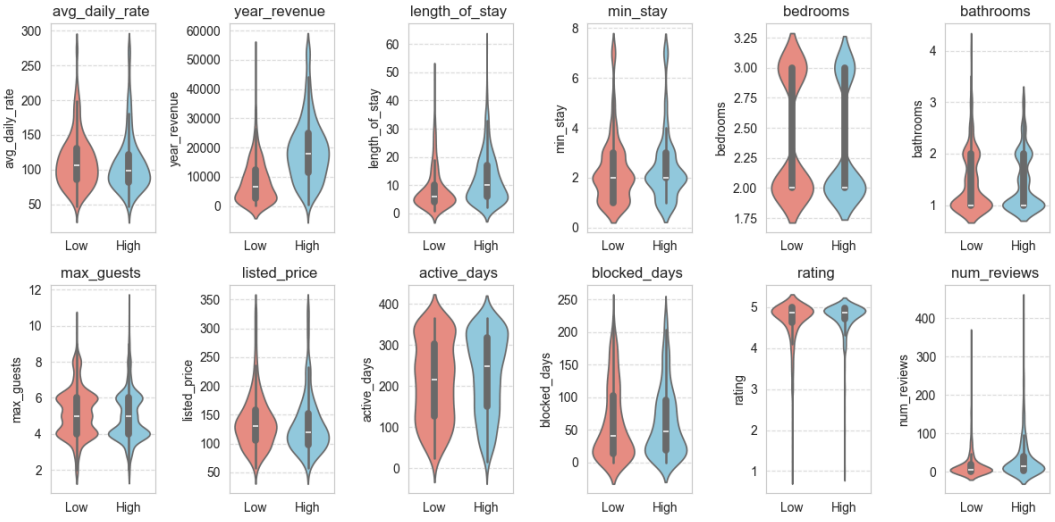

Fig. 6 Numerical Features vs Occupancy Violinplots using High vs Low Occupancy marker

Bar charts were used for categorical/ binary features

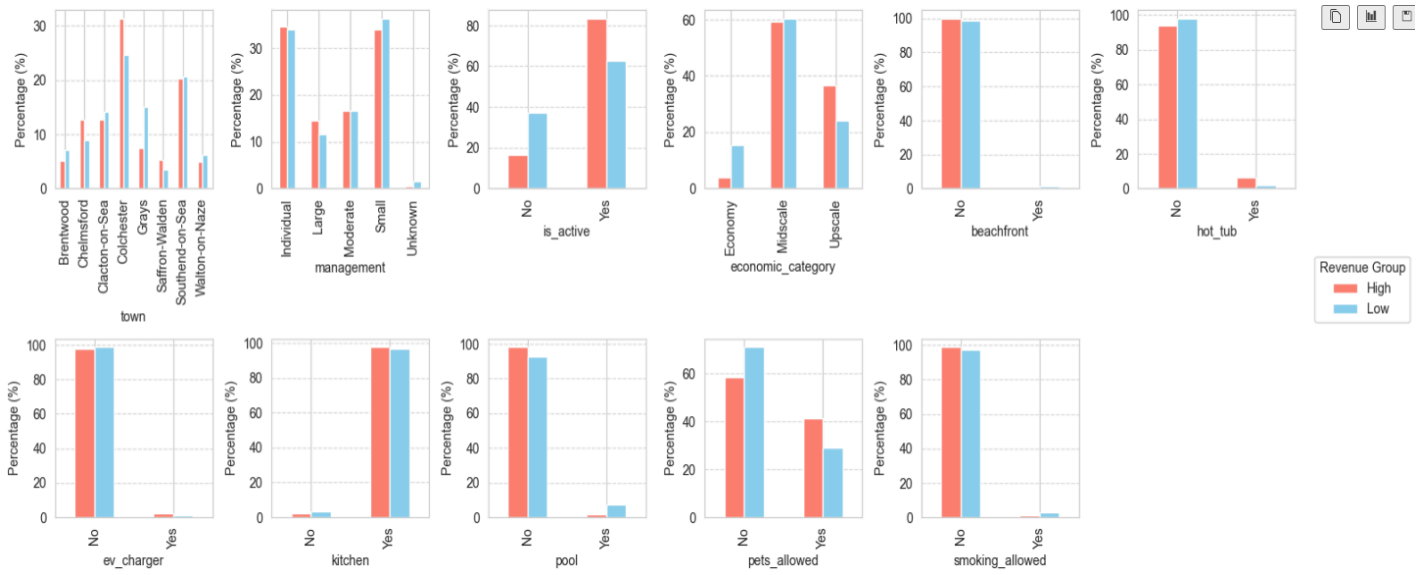

Fig 7. Bar chart feature analysis for annual revenue

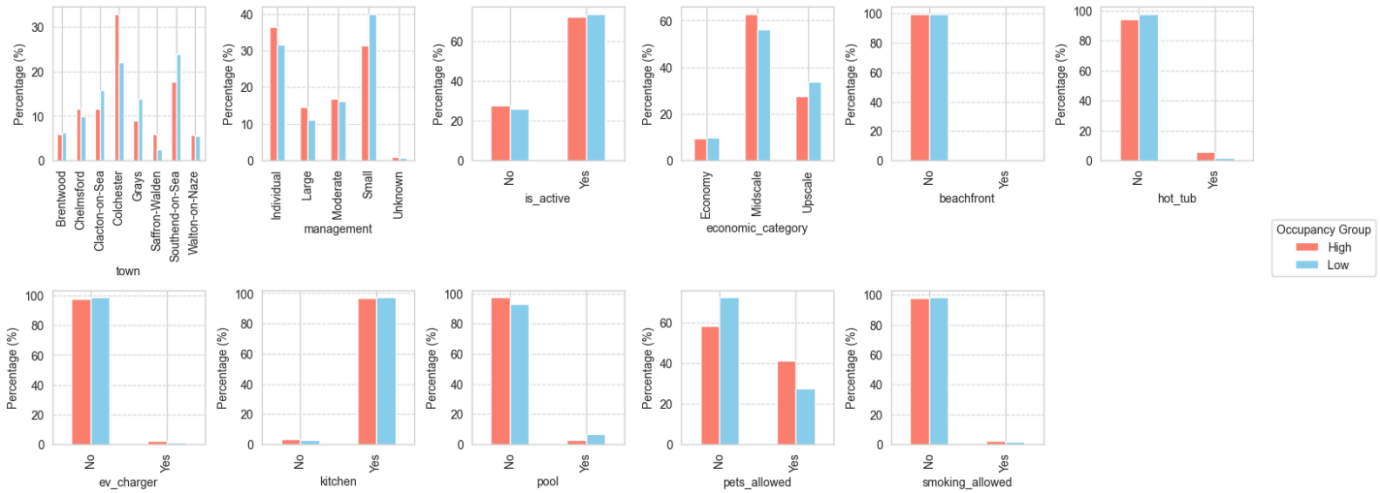

Fig 8. Bar chart feature analysis for Occupancy

Data Pre-Processing

Data pre-processing was also conducted to transform categorical/binary features into numerical values using Scikit Learn module ‘preprocessing, LabelEncoder.

Column values transformed were ‘town’, ‘management’, ‘is_active’, ‘economic category’, ‘beachfron’, ‘hot_tub’, ‘ev_charger’, ‘kitchen’, ‘pool’, ‘pets_allowed’ and ‘smoking_allowed.

As median values were used for ‘High’ and ‘Low’ markers for the target variables, there is near equal weight between these categories.

Correlational analysis

Correlational analysis was conducted using heatmaps to determine which features show the strongest relationship between the target variables. For increased readability, this was limited to the top 15 features (k = 15).

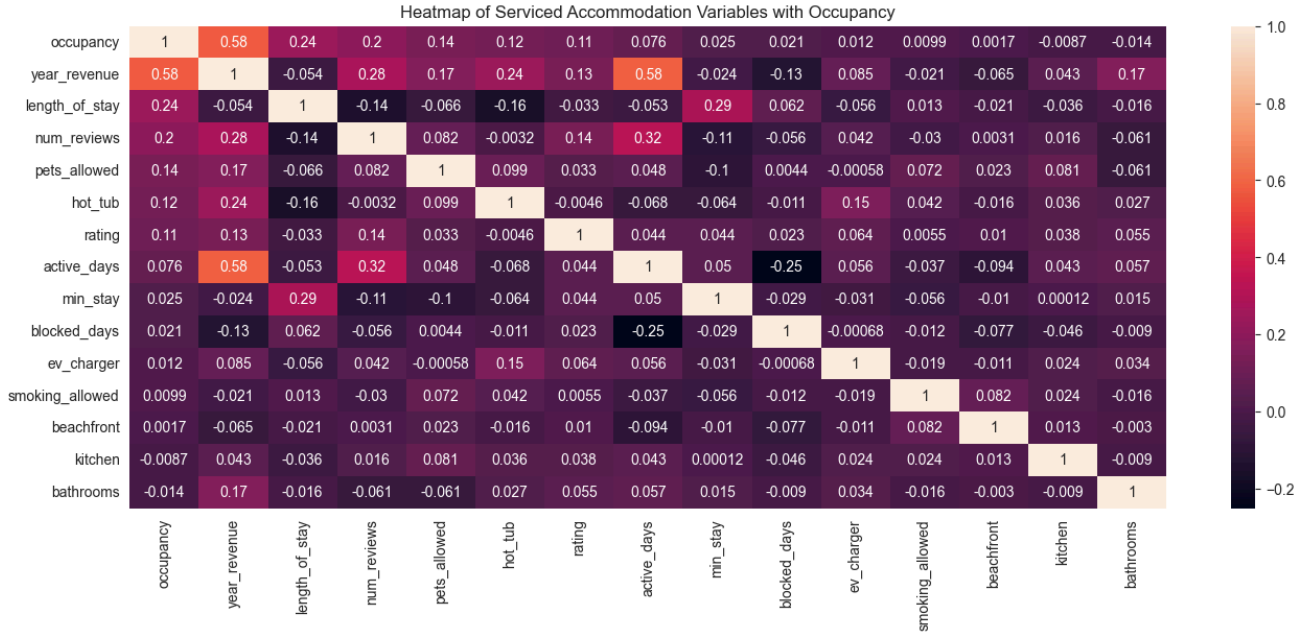

Fig 9. Heatmap showing correlationship between features and Occupancy

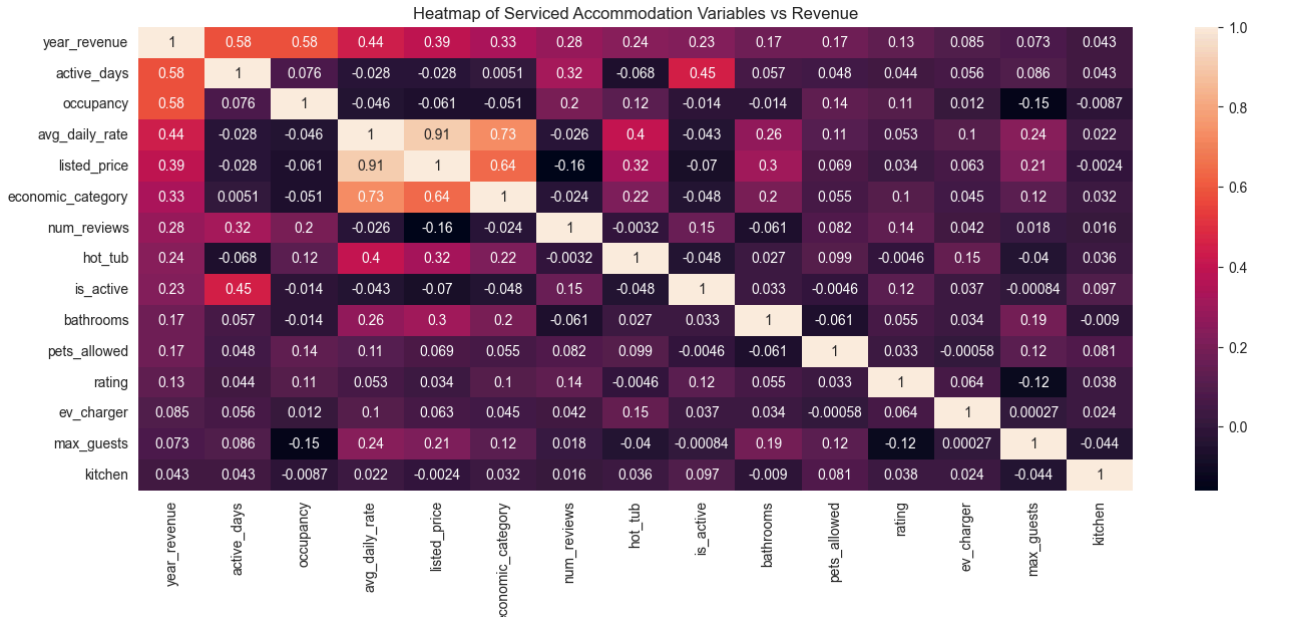

Fig 10. Heatmap showing correlationship between features and Annual Revenue

Summary of Findings and Conclusions

Whilst calculating correlation coefficients have not shown strong correlations between property features, occupancy rates and annual revenue, some patterns have emerged.

Property Features
Chelmsford and Colchester appear to be the most likely areas of higher occupancy rates and higher annual return.
There is no evidence to show that more bedrooms, beachfront properties or availability of a pool or EV charger on the properties will yield better earnings.
There is some weak evidence that the availability of a hot-tub, properties considered within a higher economic bracket, more bathrooms and are pet-friendly may be slightly more profitable.

Listing and Host Features
Listed price and average daily rate of a property has the strongest relationship with revenue but does not impact occupancy rates. For both target variables, the listing has had a positive relationship with occupancy and revenue, suggesting that customers are more likely to book a property if that property has had a record of reviews. However, rating has a weak correlation between both groups.