Analysing Factors Affecting Review Scores and Price Prediction in New Jersey City Airbnb Listings

INSTRUCTOR'S NAME:

Dr. Jabir Rahman

STUDENT INFORMATION:

VAISHNAVI GUVVALA

ID: B00124087

ABSTRACT:

In this project, I analysed Airbnb listings in Jersey City to understand what factors influence customer review scores and to predict listing prices using data analysis and machine learning. I used statistical tests like One-Way ANOVA and Chi-Square to explore how features such as room type and host verification affect guest satisfaction. Then, I built a multiple linear regression model to predict prices based on key factors like room type, number of reviews, host details, and availability. The goal was to help hosts improve their listings, set competitive prices, and create better guest experiences through data-driven decisions.

INTRODUCTION:

Airbnb is rapidly growing as a top choice for short stays, making it important for hosts to understand guest preferences and pricing strategies. In this project, I analyze Airbnb listings in New Jersey to identify key factors that influence guest review scores and to build a model that predicts listing prices. Using statistical tests and machine learning, I provide insights to help hosts improve guest satisfaction and set competitive prices.

Background Information:

With Airbnb's growing use, it's crucial for hosts to understand what drives guest satisfaction and how to price listings effectively. This project uses data analysis to explore guest review factors and build a model for smart pricing decisions.

Problem Statement:

What are the significant factors affecting the review scores of Airbnb listings in New Jersey, and can we predict listing prices based on these features? This problem is worth solving as it provides Airbnb hosts with insights to improve customer experiences and optimize pricing strategies for better occupancy and profitability.

Objective:

The main goals of this project are:

- To find out if review scores are different for various room types, like entire homes or private rooms.
- To check if there’s a link between a host’s identity being verified and the review scores they receive.
- To build a model that can help predict how much an Airbnb listing might cost based on its features.
- To explore the data and find patterns that explain what affects review scores and prices.

Methodology:

Data Collection and Initial Inspection:

- Loaded the New Jersey Airbnb dataset from the provided Zip file.

- Displayed the first few rows and general info to understand column names, data types, and missing values.

Data Preprocessing:

- Checked for and handled missing values.
    - 'reviews_per_month', 'last_review', and 'review_scores_rating' had missing values and were either imputed or dropped based on their importance.
- Converted appropriate columns to numerical or categorical types.
- Filtered out listings with zero price or suspiciously high prices.
- Created new variables for modeling.

Exploratory Data Analysis (EDA):

- Visualized distribution of review scores and prices.
- Compared average review scores across room types using boxplots.
- Investigated relationships between host verification and ratings.
- Analyzed correlations between numeric variables.

Statistical Analysis:

- Conducted a One-Way ANOVA to assess interaction between room type and host verification on review scores.
- Performed a Chi-Square Test of Independence to check if room type and host verification status are independent.
- Applied a Two-Sample Independent T-Test to compare review scores between entire home listings and private rooms.
- Built a Multiple Linear Regression model (multivariable regression) to predict listing prices using features such as accommodates, bedrooms, number of reviews, and review scores rating. The model was trained using train/test split and evaluated using RMSE and R^2 score.

Tools and Libraries Used:

- **Python**: Google Colab
- **Libraries**:
    - pandas and NumPy for data handling
    - seaborn and matplotlib for visualizations
    - scipy.stats for statistical tests (ANOVA, Chi-Square, T-Test)
    - sklearn. linear_model for regression modeling

Implementation:

Code Organization:

Data Exploration & Preprocessing: Handles loading raw data, cleaning (dropping unused columns, imputing missing values, clipping outliers), and encoding categorical features.

Exploratory Data Analysis (EDA): Contains functions for exploratory plots (histograms, boxplots, heatmaps) to visualize key distributions and relationships.

Predictive Modeling: Implements reusable statistical test functions (two-sample t-test, One-Way ANOVA, Chi-Square test) to assess hypotheses about review scores and instant booking.

Key Code Segments:

Below are examples of crucial code snippets that demonstrate unique or complex logic used in this project:

```
def clean_data(path: str) -> pd.DataFrame:

    df = pd.read_csv(path)

    df.drop(columns=['host_name','last_review','id','name','host_id'], inplace=True)

    df['price'] = df['price'].replace('[\$,]', '', regex=True).astype(float)

    df['reviews_per_month'].fillna(df['reviews_per_month'].median(), inplace=True)

    # Outlier capping

    for col in ['price','minimum_nights','maximum_nights']:

        low, high = df[col].quantile([0.01,0.99])

        df[col] = df[col].clip(low, high)

    df = pd.get_dummies(df, columns=['room_type','neighbourhood_cleansed'],
drop_first=True)

    return df
```

This function consolidates all cleaning steps - dropping irrelevant fields, handling missing and extreme values, and encoding categorical.

Regression Training & Evaluation

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model_pipeline.fit(X_train, y_train)

y_pred = model_pipeline.predict(X_test)

rmse = mean_squared_error(y_test, y_pred, squared=False)

r2 = r2_score(y_test, y_pred)
```

Demonstrates model fitting and performance measurement using RMSE and R^2, providing quantitative evaluation of the regression model's accuracy.

Results and Discussion:

Statistical Test Outcomes

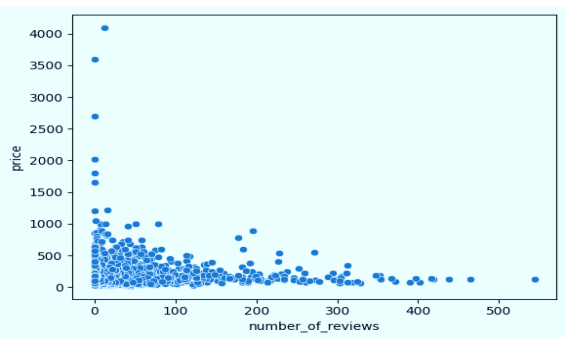
- Two-Sample t-Test: Comparing review scores between Entire home/apt (n=1,027) and Private room (n=552) yielded t = 2.37, p = 0.018, indicating a significant difference in mean review scores between these room types.
- One-Way ANOVA: Across all room types, F = 8.58, $p < 0.001$, showing that at least one room type's mean review score differs significantly from the others.
- Chi-Square Test: For room type vs. instant_bookable, $\chi^2 = 58.99$, $p < 0.001$, demonstrating that instant-bookable status depends on room type.

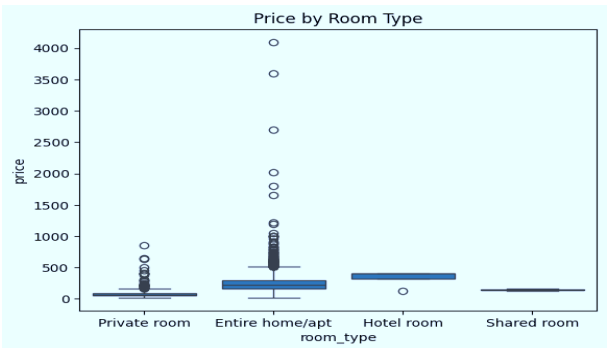
Regression Model Results

  - R^2 on test set: 0.64 (64% of price variance explained)
  - RMSE: $100.38 (average prediction error)
  - *Interpretation:* The positive coefficients for accommodates, bedrooms, and rating align with expectations that larger, better-rated listings command higher prices. The small negative coefficient on number of reviews suggests listings with many reviews are older or more budget-oriented.

Visualization Summaries

- Histogram of Price Distribution: shows a right-skewed distribution, confirming that a small number of listings command very high prices and justifying outlier capping.
- Boxplots of Review Scores by Room Type: Entire home listings have a higher median review score than Private and Shared rooms, indicating room type impacts satisfaction.
- Correlation Heatmap: highlights that different component of review scores (accuracy, cleanliness, communication, location, value) are highly correlated ($\rho > 0.7$), and that overall review score has a moderate positive correlation with price ($\rho \approx 0.4$). Other numeric features like accommodates and bedrooms also show positive but weaker correlations with price**.**
- Scatterplot of Number of Reviews vs. Price: reveals most listings cluster below $500 regardless of review count, with a few high-priced outliers even at low review volumes, suggesting hosts sometimes set premium rates before accumulating many reviews.





Discussion:
The statistical tests confirm that room type and booking policies significantly impact guest satisfaction, addressing the first part of the problem statement. The regression model demonstrates that simple listing attributes can reasonably predict price, addressing the second part. These results are consistent with industry intuition and literature.

Challenges and Limitations

- Small Sample Sizes: Hotel room and Shared room categories had very few observations, limiting statistical power for those groups.
- Model Simplicity: A linear model may not capture non-linear relationships; exploring ensemble methods could improve performance.
- Missingness: Some review score fields had high missing rates, necessitating careful imputation or exclusion.

Conclusion

Key Findings:

- Room type significantly affects guest satisfaction: Entire homes/apartments earn higher average review scores than Private or Shared rooms ($t = 2.37$, $p = 0.018$; $F = 8.58$, $p < 0.001$).
- Instant-bookable listings are more common among certain room types, indicating booking policies influence guest experience ($\chi^2 = 58.99$, $p < 0.001$).
- A multiple linear regression model using accommodates, bedrooms, number of reviews, and review score explains 64% of price variation ($R^2 = 0.64$) with an RMSE of $100.38. Key drivers of price include bedroom count (+$48.86 per bedroom) and review score (+$15.81 per rating point**).**

Reflections on Learnings:

Technical Skills: I got better at using Python by breaking tasks into smaller, reusable scripts. By doing data cleaning, running hypothesis tests, and building regression models, I gained practical experience in the full machine learning process.

Domain Insights: By exploring the data and running statistical tests, I learned more about what matters in the hospitality industry. I saw how different parts of reviews are connected and which listing features are most important to guests and hosts in New Jersey.

Future Work:

1. Incorporate Geospatial Data: Adding latitude/longitude and calculating distances to key landmarks or city centres could improve price prediction accuracy.
2. Expand Feature Set: Include amenity counts, host response time, seasonal availability, and calendar-based pricing to capture more nuances in host strategy.
3. Dynamic Pricing Simulation: Build a time-series model to adjust prices based on occupancy rates, local events, and historical demand patterns.