**CloudCredits**

**Internship**

**Project**

**Report**

Submitted By:

Vishalakshi Kumari

Email : [vishalakshi927@gmail.com](mailto:vishalakshi927@gmail.com)

Domain : Data Analytics

Project-1 : Airbnb Data Analysis Report

Project-2 : Iris Flower Dataset Analysis and Predictions

Github Repo link : https://github.com/Vishalakshi9/CloudCredits

Airbnb Data Analysis Report

1. **Problem Definition**

Objective:

The goal of this analysis is to understand the key trends in Airbnb listings, focusing on pricing, ratings, and other influencing factors. Specifically, we aim to:

Identify the distribution of ratings and prices.

Analyze relationships between price and features like number of reviews, bedrooms, and guest capacity.

Determine whether highly rated listings are priced significantly higher.

Data Source:

The dataset used for this analysis comes from an Airbnb CSV file containing 12,805 rows and 23 columns of listing details.

1. **Data Acquisition**

Dataset Overview:

The dataset contains various attributes related to Airbnb listings, such as:

Basic Details: Listing ID, Name, Host Name, Host ID

Location: Address, Country

Property Features: Bedrooms, Beds, Guests, Bathrooms, Studios

Pricing & Popularity: Price, Rating, Number of Reviews

Other Details: Check-in/out time, Amenities, House Rules

Initial Observations:

The dataset includes 23 columns, but not all are relevant for numerical analysis.

Some data columns require cleaning, such as rating, reviews, host\_id, and checkin/checkout times.

Missing values are present in columns like host\_name, checkin, and checkout.

1. **Data Cleaning and Preparation**

Steps Taken:

Removed an unnecessary column (Unnamed: 0).

Converted rating and reviews from object type to numeric.

Converted host\_id to integer.

Extracted checkin and checkout times.

Filled missing values in host\_name, checkin, and checkout with "Unknown".

**Cleaned Dataset Summary:**

12,805 listings analyzed.

All columns formatted correctly for analysis.

Missing values handled appropriately.

1. **Exploratory Data Analysis (EDA)**

Descriptive Statistics:

Feature Mean Median Min Max Std Dev

Rating 4.68 4.75 0.0 5.0 0.32

Reviews 23.6 5.0 0 1000 76.5

Price 8,000 6,500 500 250,000 10,200

Bedrooms 1.5 1.0 0 10 0.85

Guests 3.2 3.0 1 16 1.9

Key Findings:

Most listings have high ratings (average 4.68).

Price distribution is highly skewed, with some extreme values.

Some listings have zero bedrooms, indicating possible misclassification.

**Visualizations**

1. Distribution of Ratings

Most ratings are clustered around 4.5 to 5.0.

Very few listings have ratings below 3.

2. Distribution of Prices

Prices range from 500 to 250,000, but most listings are below 10,000.

Some outliers may need further investigation.

3. Price vs. Rating

No strong correlation between price and rating.

High-priced listings are found across all rating levels.

4. Correlation Matrix

Price is weakly correlated with ratings and reviews.

Number of guests and bedrooms have a higher impact on price.

5. Basic Statistical Analysis

Correlation Analysis

Feature Pair Correlation

Price vs. Rating -0.12 (Weak Negative)

Price vs. Reviews 0.05 (No Correlation)

Price vs. Bedrooms 0.68 (Moderate Positive)

Price vs. Guests 0.62 (Moderate Positive)

Number of bedrooms and guest capacity strongly impact price.

Ratings and reviews have little to no impact on price.

Hypothesis Testing: Does Rating Affect Price?

We conducted a t-test to check if highly-rated listings (rating ≥ 4.5) have significantly different prices compared to lower-rated listings (rating < 4.5).

T-Test Statistic: 1.32

P-Value: 0.18

Conclusion:

Since the p-value > 0.05, there is no significant difference in price between highly rated and low-rated listings.

6. Insights and Interpretation

Key Takeaways:

🔹 High ratings are common, but they do not significantly impact price.

🔹 Price is mostly influenced by the number of bedrooms and guest capacity rather than ratings or reviews.

🔹 Most listings are affordable (below 10,000), but some outliers exist.

🔹 There is no strong correlation between rating and price, meaning customers do not necessarily pay more for highly rated listings.

**Recommendations:**

✅ Hosts should focus on increasing guest capacity and bedrooms to justify higher prices.

✅ Potential price outliers should be examined for listing errors.

✅ New hosts should aim for reviews rather than high ratings, as review count does not significantly impact price.

**7. Reporting and Visualization**

All charts and insights documented.

Correlation heatmaps and scatter plots help visualize relationships.

A cleaned dataset was saved for further analysis.

Final Data Export:

The cleaned dataset has been saved as:

📂/content/cleaned\_airbnb.csv

**Final Thoughts**

This analysis provides valuable insights into Airbnb listings, helping hosts optimize their pricing strategy and travelers make informed booking decisions.

Iris Flower Dataset Analysis and Predictions

1. **Problem Definition**

Objective:

This analysis aims to:

Understand trends in sepal & petal measurements

Identify correlations between features

Determine if species significantly differ in certain features

Build a classification model to predict Iris species

Data Source:

The dataset used is the well-known Iris dataset, containing 150 samples of iris flowers categorized into three species:

Iris-setosa

Iris-versicolor

Iris-virginica

Each sample includes sepal length, sepal width, petal length, and petal width as numerical features.

**2. Data Acquisition & Cleaning**

Loaded dataset with 150 rows and 5 columns

No missing values

Converted species names into numerical labels for ML modeling

1. **Exploratory Data Analysis (EDA)**

3.1 Summary Statistics

Feature Mean Min Max Std Dev

Sepal Length (cm) 5.84 4.3 7.9 0.83

Sepal Width (cm) 3.05 2.0 4.4 0.43

Petal Length (cm) 3.76 1.0 6.9 1.76

Petal Width (cm) 1.20 0.1 2.5 0.76

3.2 Key Observations

🔹 Petal length and petal width have the highest variation, indicating they may be key differentiators.

🔹 Sepal width has the least variation, meaning it may contribute less to species distinction.

🔹 Sepal length ranges between 4.3 cm - 7.9 cm, while petal length varies more (1 cm - 6.9 cm).

1. **Data Visualization & Insights**

4.1 Histograms of Features

Sepal length & width follow a near-normal distribution.

Petal length & width are more skewed, showing significant species separation.

4.2 Boxplots to Identify Outliers

Petal length & petal width show clear differences among species.

No extreme outliers observed, meaning data is well-behaved.

4.3 Pairplot Analysis

Petal length & petal width strongly correlate, separating species effectively.

Sepal features show overlap, making them less useful for classification.

4.4 Correlation Heatmap

Feature Pair Correlation

Petal Length & Petal Width +0.96 (Strong Positive)

Sepal Length & Petal Length +0.87 (Strong Positive)

Sepal Width & Petal Length -0.42 (Weak Negative)

Petal features are the most important for species classification.

Sepal width has almost no correlation with petal measurements.

**5. Statistical Analysis: Hypothesis Testing**

Question: Do different species have significantly different petal lengths?

📌 Test Used: ANOVA (Analysis of Variance)

📌 Hypothesis:

Null Hypothesis (H₀): There is no significant difference in petal length among species.

Alternative Hypothesis (H₁): Petal length differs significantly among species.

📊 Results:

ANOVA Test Statistic: 1179.03

P-Value: 3.05 × 10⁻⁹¹ (Extremely Small)

✅ Conclusion: Since p < 0.05, we reject the null hypothesis.

✅ Petal length is significantly different across species.

1. **Machine Learning: Iris Species Classification**

6.1 Model: Decision Tree Classifier

We trained a Decision Tree Classifier to predict the species based on sepal & petal measurements.

6.2 Data Preparation

Target Variable: species (encoded numerically: 0, 1, 2)

Features Used: sepal\_length, sepal\_width, petal\_length, petal\_width

Data Split:

80% training set (120 samples)

20% testing set (30 samples)

6.3 Model Performance

✅ Accuracy Score: 100% 🎯

✅ Classification Report:

Species Precision Recall F1-Score

Iris-setosa 1.00 1.00 1.00

Iris-versicolor 1.00 1.00 1.00

Iris-virginica 1.00 1.00 1.00

**7. Key Takeaways & Recommendations**

7.1 Insights from EDA

🔹 Petal length & petal width are the most important features for classification.

🔹 Sepal measurements are less useful for species distinction.

🔹 Species classification can be done with high accuracy using simple models.

7.2 Recommendations

✅ Use petal length & width as key identifiers for species.

✅ Decision Trees perform exceptionally well for this dataset (100% accuracy).

✅ Further analysis could explore Random Forest or Logistic Regression for comparison.