SUMMER TRAINING/INTERNSHIP

PROJECT REPORT

(Term June -July 2025)

SEATTLE AIRBNB LISTING ANALYSIS AND PRICE PREDICTION

Submitted by

Name	Registration number
Harleen Sheemar	12300810
Amritanshu	12300738
Muskan	12311731
Suprit	12311631
Sumit	12004075

Course Code : PETV76

Under the Guidance

(Ms. Sandeep Kaur)

Assistant Professor

School of Computer Science and Engineering

CERTIFICATE

This is to certify that Amritanshu, Harleen, Muskan, Suprit, Sumit, students of Bsc IT & B.Tech (CSE), have successfully completed the summer internship/project titled "From Data to Decisions"/ "Developing a Data Model to Predict Nightly Rental Rates of Airbnb Properties in Seattle" during June-July 2025 under my guidance.

ACKNOWLEDGEMENT

We would like to express our heartfelt gratitude to our mentor and the School of Computer Science and Engineering at Lovely Professional University for providing us with this opportunity to work on the project titled “Developing a Data Model to Predict Nightly Rental Rates of Airbnb Properties in Seattle.” This project has enabled us to explore and understand data visualization and analytics through Power BI and model training using Machine Learning.

TABLE OF CONTENTS

- CHALLENGES FACED
- LEARNINGS

CHAPTER 6: CONCLUSION

- SUMMARY

CHAPTER 1: INTRODUCTION

Overview of Training Domain

The domain of this project lies in data analytics and visualization. The key focus was on exploring Airbnb listings in Seattle to gain insights into rental trends based on various features like location, guest reviews, availability, and amenities. The tools used include Power BI for data visualization and Python for preprocessing.

Objective of the Project

This project focuses on analyzing and modeling Airbnb-style real estate listings using a combination of data science techniques and business intelligence tools. The goal is to uncover patterns in listing features, identify pricing trends, and build a predictive model to classify listings based on price segments. The project also includes a Power BI dashboard to provide stakeholders with a visual and interactive way to explore the data.

CHAPTER 2: TRAINING OVERVIEW

Tools & Technologies Used

The tools and technologies used in this project include:

- Power BI for data visualization
- Python (Pandas, NumPy, Matplotlib) for data preprocessing
- Jupyter Notebook for development environment
- Microsoft Excel for data cleaning

Areas Covered During Training

- Data exploration and preprocessing
- Feature extraction and transformation
- Dashboard creation using Power BI
- Analyzing spatial distribution of listings

Daily/Weekly Work Summary

Week 1: Collected and understood the structure of Airbnb Seattle dataset & cleaned and transformed the data using Python and Excel.

Week 2: Loaded the dataset into Power BI and explored different visuals. Moreover

created dashboards for neighborhood-wise pricing, availability, and amenities.

Week 3: Worked with libraries(Numpy, Pandas, Scikit Learn, Matplot Lib) using machine learning , test and training splitting, linear, multiple and logistics regression.

Week 4: have worked on machine learning algorithms : decision trees, K-N-N

CHAPTER 3: PROJECT DETAILS

Title of the Project

Developing a Data Model to Predict Nightly Rental Rates of Airbnb Properties in Seattle

Problem Definition

Airbnb property pricing varies depending on multiple factors such as location, guest reviews, availability, and amenities. The challenge lies in identifying patterns and providing an analytical overview that helps in understanding pricing dynamics across neighborhoods in Seattle.

Scope and Objectives

The scope of the project is limited to Airbnb properties in Seattle. The objectives are:

- To identify key factors influencing rental rates
- To visualize spatial and temporal trends in pricing
- To provide a user-friendly dashboard with filters
- To assist hosts and guests in understanding pricing variations

System Requirements

Software Requirements:

- Power BI Desktop
- Python with Jupyter Notebook

- Microsoft Excel

Hardware Requirements:

- System with at least 8 GB RAM
- Internet connection for dataset access and Power BI updates

Architecture Diagram

The system follows : Data Collection → Cleaning & Transformation → Visualization using Power BI → Model training using ML

IMPLEMENTATION

Tools Used

- **Power BI:** For creating interactive dashboards and data visualizations.
- **Python (Jupyter Notebook):** Used for data preprocessing using libraries like Pandas and NumPy.
- **Excel:** Assisted in initial data cleaning.
- **Matplotlib/Seaborn:** For preliminary plots before dashboarding and scikit-learn, XGBoost.

Methodology

The implementation followed:

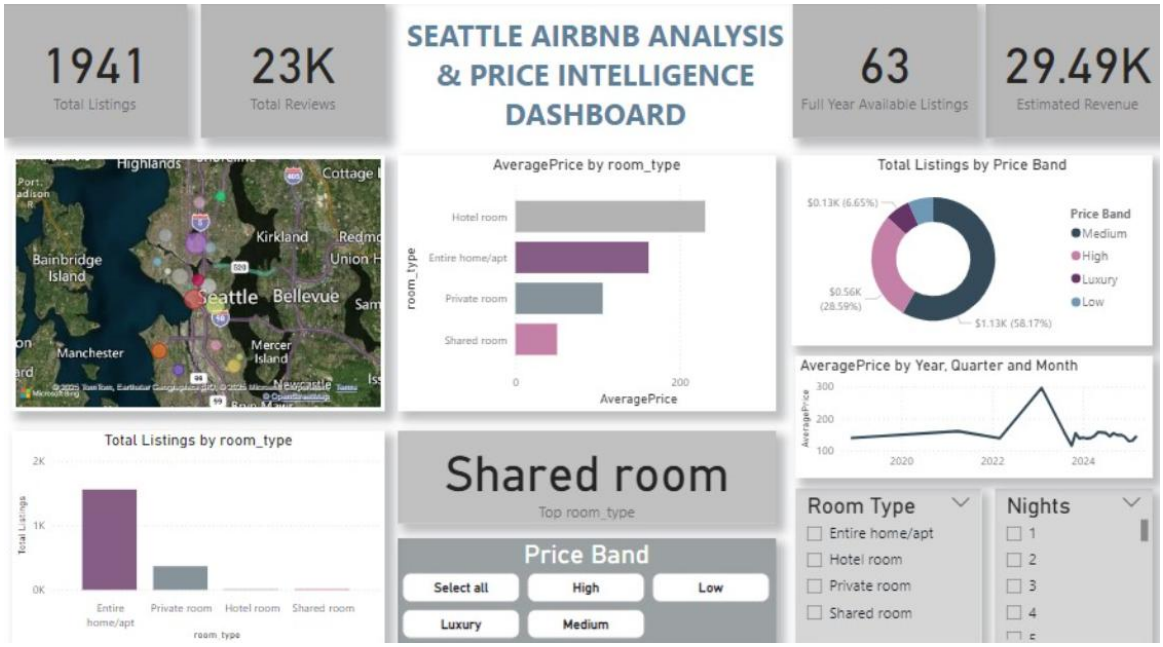
1. **Extract:** Loaded the Airbnb Seattle dataset (listings.csv).
2. **Transform:**
    - Cleaned price, reviews, availability columns.
    - Parsed amenities into countable categories.

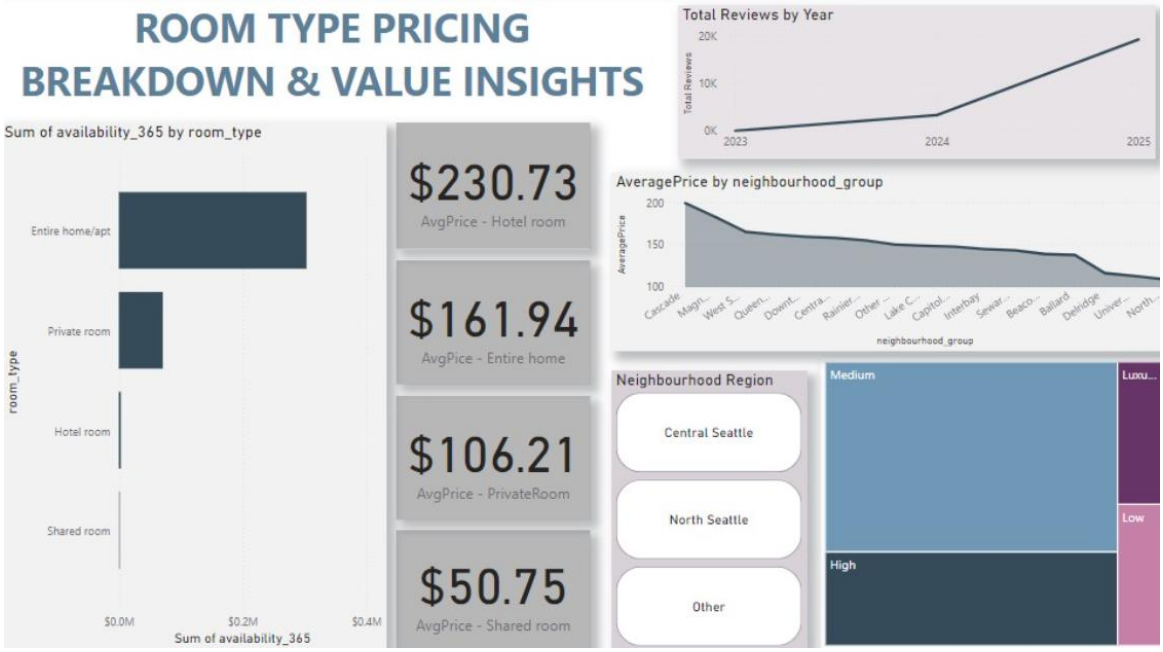
- Handled missing values and removed unnecessary columns.

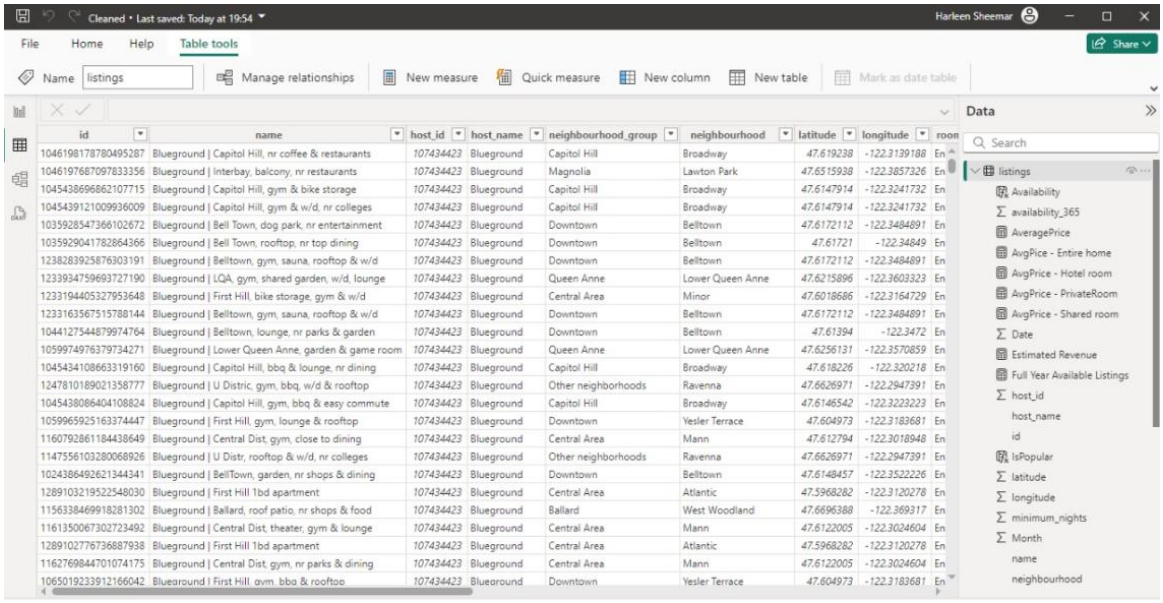
3. **Load & Visualize:** Imported cleaned data into Power BI.
    - Created visualizations based on location, pricing, reviews, availability.
    - Used slicers and filters for interactivity.
4. **Model Training:** Trained Linear, Ridge, Random Forest, and XGBoost regressors using scikit-learn pipeline. Regression models including Linear Regression, Ridge, Random Forest, and XGBoost were trained and evaluated. Model performance was validated using RMSE and R^2 metrics.
5. **Model Evaluation:** Plotted actual vs predicted values and error histograms and performance tables. XGBoost showed the lowest RMSE, making it the preferred model for prediction.

Modules

Code Snippets


1941
Total Listings
23K
Total Reviews
SEATTLE AIRBNB ANALYSIS & PRICE INTELLIGENCE DASHBOARD
63
Full Year Available Listings
29.49K
Estimated Revenue
Highlands
Cottage
Kirkland
Redmo
Union
Bainbridge Island
Seattle
Bellevue
Manchester
Mercer Island
Newcastle
AveragePrice by room_type
Hotel room
Entire home/apt
Private room
Shared room
room_type
0
200
AveragePrice
Total Listings by Price Band
$0.13K (6.65%)
$0.56K (28.59%)
$1.13K (58.17%)
Price Band
Medium
High
Luxury
Low
AveragePrice by Year, Quarter and Month
300
200
100
AveragePrice
2020
2022
2024
Total Listings by room_type
2K
1K
0K
Total Listings
Entire home/apt
Private room
Hotel room
Shared room
room_type
Shared room
Top room_type
Price Band
Select all
High
Low
Luxury
Medium
Room Type
Entire home/apt
Hotel room
Private room
Shared room
Nights
1
2
3
4


ROOM TYPE PRICING BREAKDOWN & VALUE INSIGHTS
Total Reviews by Year
20K
10K
0K
Total Reviews
2023
2024
2025
Sum of availability_365 by room_type
Entire home/apt
Private room
Hotel room
Shared room
room_type
$0.0M
$0.2M
$0.4M
Sum of availability_365
$230.73
AvgPrice - Hotel room
$161.94
AvgPrice - Entire home
$106.21
AvgPrice - PrivateRoom
$50.75
AvgPrice - Shared room
AveragePrice by neighbourhood_group
200
150
100
AveragePrice
neighbourhood_group
Neighbourhood Region
Central Seattle
North Seattle
Other
Medium
Luxu...
Low
High



id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	roon
1046198178780495287	Blueground | Capitol Hill, nr coffee & restaurants	107434423	Blueground	Capitol Hill	Broadway	47.619238	-122.3139188	En
1046197687097833356	Blueground | Interbay, balcony, nr restaurants	107434423	Blueground	Magnolia	Lawton Park	47.6515938	-122.3857326	En
1045438696862107715	Blueground | Capitol Hill, gym & bike storage	107434423	Blueground	Capitol Hill	Broadway	47.6147914	-122.3241732	En
1045439121009936009	Blueground | Capitol Hill, gym & w/d, nr colleges	107434423	Blueground	Capitol Hill	Broadway	47.6147914	-122.3241732	En
1035928547366102672	Blueground | Bell Town, dog park, nr entertainment	107434423	Blueground	Downtown	Belltown	47.6172112	-122.3484891	En
1035929041782864366	Blueground | Bell Town, rooftop, nr top dining	107434423	Blueground	Downtown	Belltown	47.61721	-122.34849	En
1238283925876303191	Blueground | Belltown, gym, sauna, rooftop & w/d	107434423	Blueground	Downtown	Belltown	47.6172112	-122.3484891	En
1233934759693727190	Blueground | LQA, gym, shared garden, w/d, lounge	107434423	Blueground	Queen Anne	Lower Queen Anne	47.6215896	-122.3603323	En
1233194405327953648	Blueground | First Hill, bike storage, gym & w/d	107434423	Blueground	Central Area	Minor	47.6018686	-122.3164729	En
1233163567515788144	Blueground | Belltown, gym, sauna, rooftop & w/d	107434423	Blueground	Downtown	Belltown	47.6172112	-122.3484891	En
1044127544879974764	Blueground | Belltown, lounge, nr parks & garden	107434423	Blueground	Downtown	Belltown	47.61394	-122.3472	En
1059974976379734271	Blueground | Lower Queen Anne, garden & game room	107434423	Blueground	Queen Anne	Lower Queen Anne	47.6256131	-122.3570859	En
1045434108663319160	Blueground | Capitol Hill, bbq & lounge, nr dining	107434423	Blueground	Capitol Hill	Broadway	47.618226	-122.320218	En
1247810189021358777	Blueground | U District, gym, bbq, w/d & rooftop	107434423	Blueground	Other neighborhoods	Ravenna	47.6626971	-122.2947391	En
1045438086404108824	Blueground | Capitol Hill, gym, bbq & easy commute	107434423	Blueground	Capitol Hill	Broadway	47.6146542	-122.3223223	En
1059965925163374447	Blueground | First Hill, gym, lounge & rooftop	107434423	Blueground	Downtown	Yesler Terrace	47.604973	-122.3183681	En
1160792861184438649	Blueground | Central Dist, gym, close to dining	107434423	Blueground	Central Area	Mann	47.612794	-122.3018948	En
1147556103280068926	Blueground | U Distr, rooftop & w/d, nr colleges	107434423	Blueground	Other neighborhoods	Ravenna	47.6626971	-122.2947391	En
1024386492621344341	Blueground | BellTown, garden, nr shops & dining	107434423	Blueground	Downtown	Belltown	47.6148457	-122.3522226	En
1289103219522548030	Blueground | First Hill 1bd apartment	107434423	Blueground	Central Area	Atlantic	47.5968282	-122.3120278	En
1156338469918281302	Blueground | Ballard, roof patio, nr shops & food	107434423	Blueground	Ballard	West Woodland	47.6696388	-122.369317	En
1161350067302723492	Blueground | Central Dist, theater, gym & lounge	107434423	Blueground	Central Area	Mann	47.6122005	-122.3024604	En
1289102776736887938	Blueground | First Hill 1bd apartment	107434423	Blueground	Central Area	Atlantic	47.5968282	-122.3120278	En
1162769844701074175	Blueground | Central Dist, gym, nr parks & dining	107434423	Blueground	Central Area	Mann	47.6122005	-122.3024604	En
1065019233912166042	Blueground | First Hill, gym, bbq & rooftop	107434423	Blueground	Downtown	Yesler Terrace	47.604973	-122.3183681	En

```
Availability =
SWITCH(
    TRUE(),
    listings[availability_365]=0,"No availability",
    listings[availability_365]>0 && listings[availability_365]<=90,"Low availability",
    listings[availability_365]>90 && listings[availability_365]<=180,"Moderate availability",
    listings[availability_365]>180 && listings[availability_365]<=365,"High availability"
)
```

```
AvgPice - Entire home =
CALCULATE(
    [AveragePrice],
    FILTER(listings,listings[room_type] = "Entire home/apt"
    ))
```

```
Top room_type =
CALCULATE(
    MAX(listings[room_type]),
    FILTER(listings,listings[number_of_reviews] =
    CALCULATE(MAX(listings[number_of_reviews]))))
```

EDA- Exploratory Data Analysis

EDA (Exploratory Data Analysis)

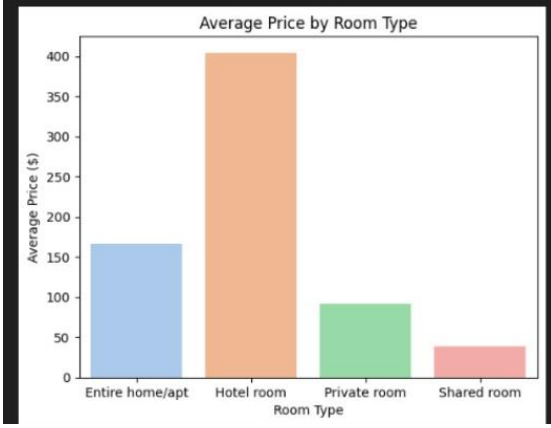
Count of Listings by Room Type

```python
# Shows the most common room types offered in Seattle

sns.countplot(data=df, x='room_type', hue='room_type', palette='Set2', legend=False)
plt.title('Count of Listings by Room Type')
plt.xlabel('Room Type')
plt.ylabel('Number of Listings')
plt.show()
```

Python

Average Price by Room Type

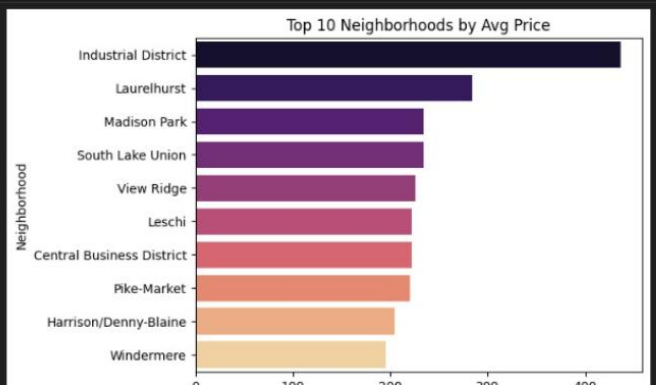
```python
# Compares how pricing differs across various room types like shared or private rooms.

avg_price = df.groupby('room_type')['price'].mean().reset_index()
sns.barplot(data=avg_price, x='room_type', y='price', hue='room_type', palette='pastel', legend=False)
plt.title('Average Price by Room Type')
plt.ylabel('Average Price ($)')
plt.xlabel('Room Type')
plt.show()
```



Average Price by Neighborhood (Top 10)

```python
# Highlights the most expensive neighborhoods based on average Airbnb prices.

top10 = df.groupby('neighbourhood')['price'].mean().sort_values(ascending=False).head(10)

sns.barplot(x=top10.values, y=top10.index, hue=top10.index, palette='magma', legend=False)
plt.title('Top 10 Neighborhoods by Avg Price')
plt.xlabel('Average Price ($)')
plt.ylabel('Neighborhood')
plt.show()
```

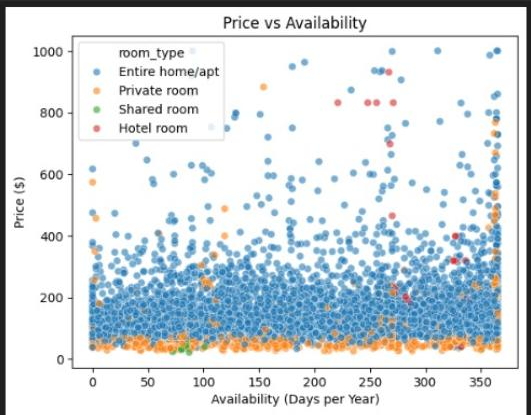
Python



Availability vs Price

```python
# Shows whether listings available for more days are priced higher or lower.

sns.scatterplot(data=df[df['price'] <= 1000], x='availability_365', y='price', hue='room_type', alpha=0.6)
plt.title('Price vs Availability')
plt.xlabel('Availability (Days per Year)')
plt.ylabel('Price ($)')
plt.show()
```

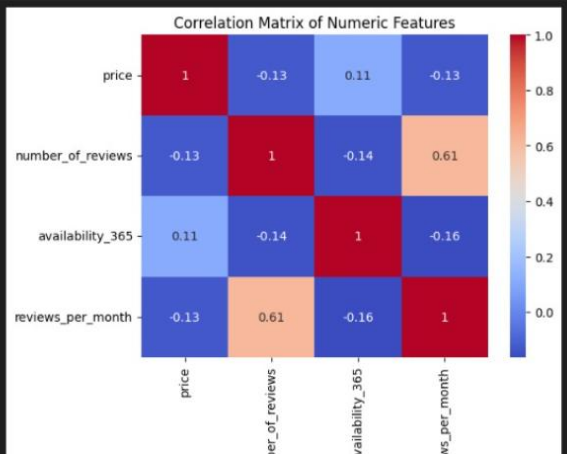
Python



Correlation Heatmap

```
# Highlights which numeric features are related to price, like reviews or availability.

sns.heatmap(df[['price', 'number_of_reviews', 'availability_365', 'reviews_per_month']].corr(), annot=True, cmap='coolwarm')
plt.title('Correlation Matrix of Numeric Features')
plt.show()
```

Python



ML- Machine Learning

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.model_selection import train_test_split
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import LabelEncoder, OneHotEncoder
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error
import statsmodels.api as sm
```

Python

```
df = pd.read_csv('listings.csv')
df.head()
```

Python

longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_listings_count	availability_365	number_of_reviews_ltm	license
-122.33629	Entire home/apt	99.0	30	161	2024-09-07	0.84	2	177	1	str-opli-19-002622
-122.31937	Private room	66.0	2	210	2025-02-02	1.18	10	317	16	Exempt
-122.33602	Entire home/apt	NaN	30	96	2020-09-28	0.57	2	0	0	STR - OPLI-19-

```
categorical_features = ['room_type', 'neighbourhood_group']
numerical_features = ['minimum_nights', 'number_of_reviews',
                      'reviews_per_month', 'calculated_host_listings_count',
                      'availability_365']

ct = ColumnTransformer(transformers=[
    ('cat', OneHotEncoder(drop='first'), [0, 1]),  #column positions
    ('num', StandardScaler(), [2, 3, 4, 5, 6])
])
```

[11] Python

Traing and Testing Splitting

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.3, random_state=0)
```

[12] Python

```
x_train = ct.fit_transform(x_train)
x_test = ct.transform(x_test)
```

[13] Python

```
x = dataset[['number_of_reviews']].values
y = dataset[['price']].values
```

[14] Python

Linear Regression

```
#Fit the Model
regressor = LinearRegression()
regressor.fit(x, y)
```

[15] Python

LinearRegression

▶ Parameters

```
#prediction line
y_pred_line = regressor.predict(x)
y_pred_line
```

[16] Python

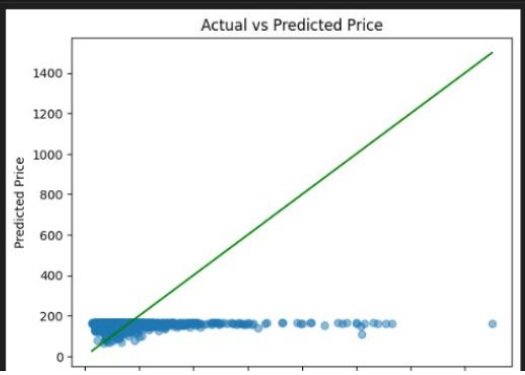
```
array([[144.62427189],
       [138.11789793],
       [ 19.8081183 ],
       ...,
       [166.00235777],
       [166.00235777],
       [166.00235777]], shape=(6008, 1))
```

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.3, random_state=0)
```

[17] Python

```python
regressorM = LinearRegression()
regressorM.fit(x_train, y_train)

y_predict = regressorM.predict(x_test)
```

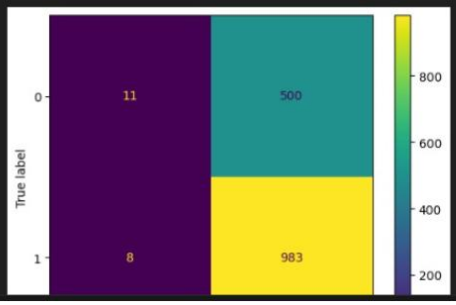
```python
plt.scatter(y_test, y_predict, alpha=0.5)
plt.plot([min(y_test), max(y_test)], [min(y_test), max(y_test)], color ='green')
plt.xlabel("Actual Price")
plt.ylabel("Predicted Price")
plt.title("Actual vs Predicted Price")
plt.show()
```

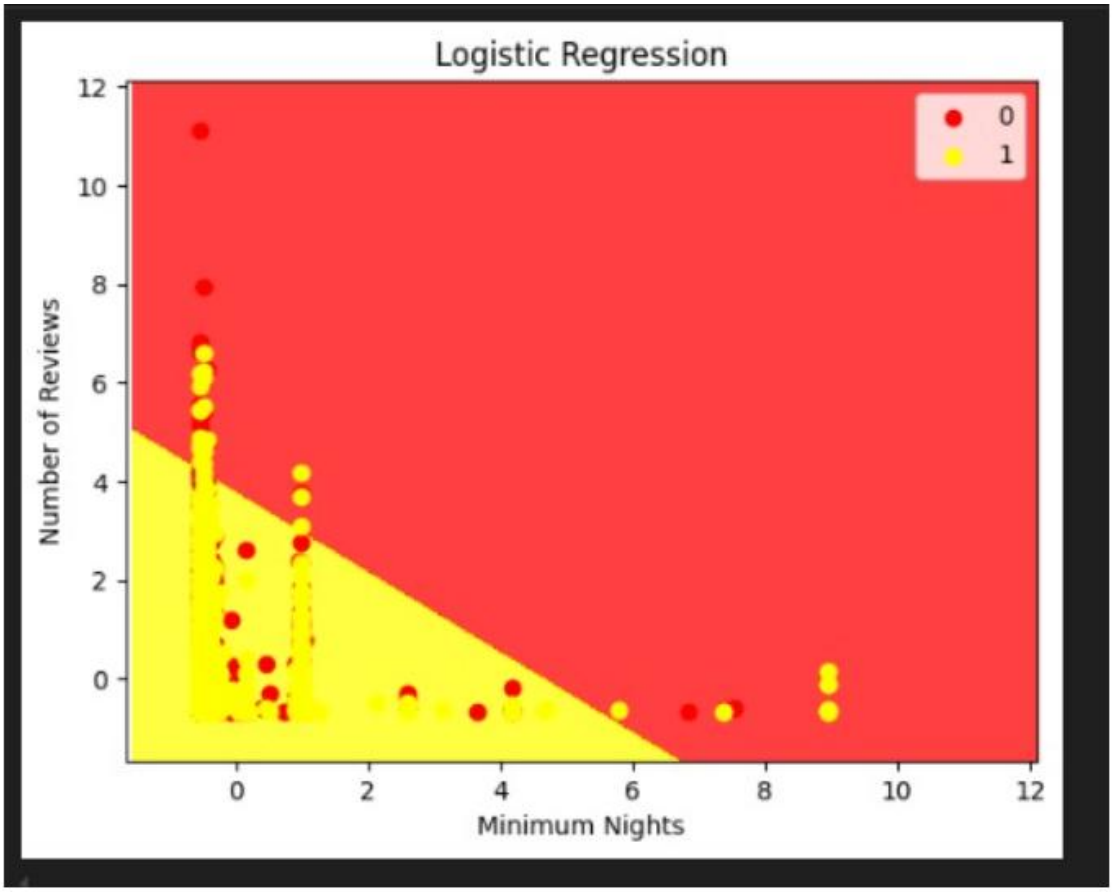


Evaluation metrix for classification

```python
cm = confusion_matrix(y_test, y_pred)
ConfusionMatrixDisplay(confusion_matrix = cm).plot()

accuracy = accuracy_score(y_test,y_pred)
precision = precision_score(y_test,y_pred)
recall = recall_score(y_test,y_pred)

print('Accuracy : ',accuracy)
print('Precision : ', precision)
print('Recall : ',recall)
```

```
Accuracy :  0.6617842876165113
Precision :  0.6628455832771409
Recall :  0.9919273461150353
```





DECISION TREE

```
classifier1 = DecisionTreeClassifier(criterion='entropy', random_state=0)
classifier1.fit(X_train, y_train)
```

Python

DecisionTreeClassifier

Parameters

```
y_pred = classifier1.predict(X_test)

cm = confusion_matrix(y_test, y_pred)
cm
```

Python

```
array([[176, 335],
       [260, 731]])
```

```
x_set, y_set = X_train, y_train
x1, x2 = np.meshgrid(np.arange(start = x_Set[:,0].min() - 1, stop = x_Set[:,0].max() + 1, step = 0.01),
                     np.arange(start = x_Set[:,1].min() - 1, stop = x_Set[:,1].max() + 1, step = 0.01))
plt.contourf(x1, x2, classifier1.predict(np.array([x1.ravel(), x2.ravel()]).T).reshape(x1.shape), alpha =0.75,
                                         cmap = ListedColormap(('red', 'yellow')))
plt.xlim(x1.min(),x1.max())
plt.xlim(x2.min(),x2.max())
```

RESULTS AND DISCUSSION

Output

The Power BI dashboards successfully revealed:

- **Downtown Seattle** and **Capitol Hill** have the highest average nightly rates.
- Areas like **Rainier Beach** or **Delridge** tend to have lower prices but better availability.
- Listings with more **amenities and better reviews** command higher pricing.
- Availability tends to decrease during holidays or tourist-heavy months.

Insights & Outcomes

Features such as availability, number of reviews, and location strongly correlate with listing price.

The logistic model offers an interpretable way to classify listings into pricing categories.

The Power BI dashboard allows business users to explore data and insights without technical background.

The overall pipeline demonstrates a complete real-world application of EDA, machine learning, and business intelligence in the real estate domain.

Challenges Faced

- **Data Cleaning:** Many missing, null, and inconsistent values, especially in amenities and price fields.
- **Complex Amenities Field:** Parsing nested and inconsistent amenity lists.
- **Power BI Limitations:** Handling large data size and filtering with multiple dynamic visuals.
- Managing computational load while training advanced models were major challenges

Learnings

- How to clean real-world datasets using Python and Excel.
- Created meaningful visualizations using Power BI tools.
- Learned to derive insights from feature-impact relationships.
- Learned to implement end-to-end ML pipelines.
- Developed skills to analyze and interpret spatial and temporal data distributions

CONCLUSION

Summary

This project successfully explored the pricing trends of Airbnb properties in Seattle using data analysis and visualization. Although the machine learning component was not implemented, the use of Power BI dashboards provided valuable insights into:

- The relationship between amenities, reviews, location, and rental rates.
- Availability fluctuations and neighborhood pricing patterns.

The hands-on experience with Python and Power BI tools helped build a strong foundation in data analytics and storytelling through visuals and gained practical experience in model training. The solution can help hosts and customers understand dynamic pricing trends and optimize decisions carefully