Maharshi Dayanand University Rohtak

Industrial Internship Report

on

"Airbnb New York City Analysis"

Prepared by
Ishita Bahamnia
M.Tech CSE-AIML
[2023-2025]

Executive Summary

This report provides details of the Industrial Internship provided by Upskill Campus and The IoT Academy in collaboration with Industrial Partner UniConverge Technologies Pvt Ltd (UCT). The internship focused on practical exposure through a data analytics project on Airbnb listings in New York City.

The objective was to analyze Airbnb listings using data-driven insights and machine learning techniques. The project included data preprocessing, exploratory data analysis (EDA), visualization, and predictive modeling. This internship allowed me to gain industrial experience, work on real-world data problems, and enhance my analytical skills.

Index

1. **Preface – Page 3**
2. **Introduction – Page 4**
    - **2.1 About UniConverge Technologies Pvt Ltd – Page 4**
    - **2.2 About Upskill Campus – Page 4**
    - **2.3 Objective – Page 4**
3. **Problem Statement – Page 4**
4. **Existing and Proposed Solutions – Page 5**
5. **Data Analysis and Model Design – Page 6**
    - **5.1 Data Collection and Preprocessing – Page 7**
    - **5.2 Exploratory Data Analysis (EDA) – Page 8**
    - **5.3 Feature Engineering and Modeling – Page 9**
6. **Performance Evaluation – Page 10**
7. **Data Insights and Visualization – Page 11**
8. **Learnings and Experience – Page 12**
9. **Future Work Scope – Page 12**
10. **References – Page 12**

Preface

This report summarizes the six-week internship experience, highlighting the importance of practical exposure in career development. The internship provided an opportunity to work on a real-world dataset, analyze Airbnb listing patterns, and develop predictive models to understand price trends and neighborhood dynamics.

The internship was well-structured with planned deliverables, including data collection, preprocessing, exploratory analysis, and model building. Special thanks to Upskill Campus, The IoT Academy, and UniConverge Technologies for providing this valuable experience.

Introduction

About UniConverge Technologies Pvt Ltd

UniConverge Technologies is a digital transformation company specializing in IoT, machine learning, cybersecurity, and cloud computing solutions. Their expertise spans across industrial automation, predictive maintenance, and data-driven insights for smart industries.

About Upskill Campus

Upskill Campus, in collaboration with The IoT Academy, facilitates industry-aligned internships and skill development programs. Their mission is to enhance employability by providing practical exposure and project-based learning opportunities.

Objective

- Gain hands-on experience in data analysis and machine learning.
- Work with real-world Airbnb data to derive meaningful insights.
- Develop predictive models for price estimation and market trends.
- Understand industry standards in data-driven decision-making.

Problem Statement

The project focused on analyzing Airbnb listings in New York City to:

1. Identify key factors influencing rental prices.
2. Understand neighborhood-wise pricing patterns.
3. Predict future pricing trends using machine learning models.

Existing and Proposed Solutions

Existing Solutions:

- Airbnb's dynamic pricing model considers seasonality and demand.
- Market analysis reports provide general trends but lack granular insights.

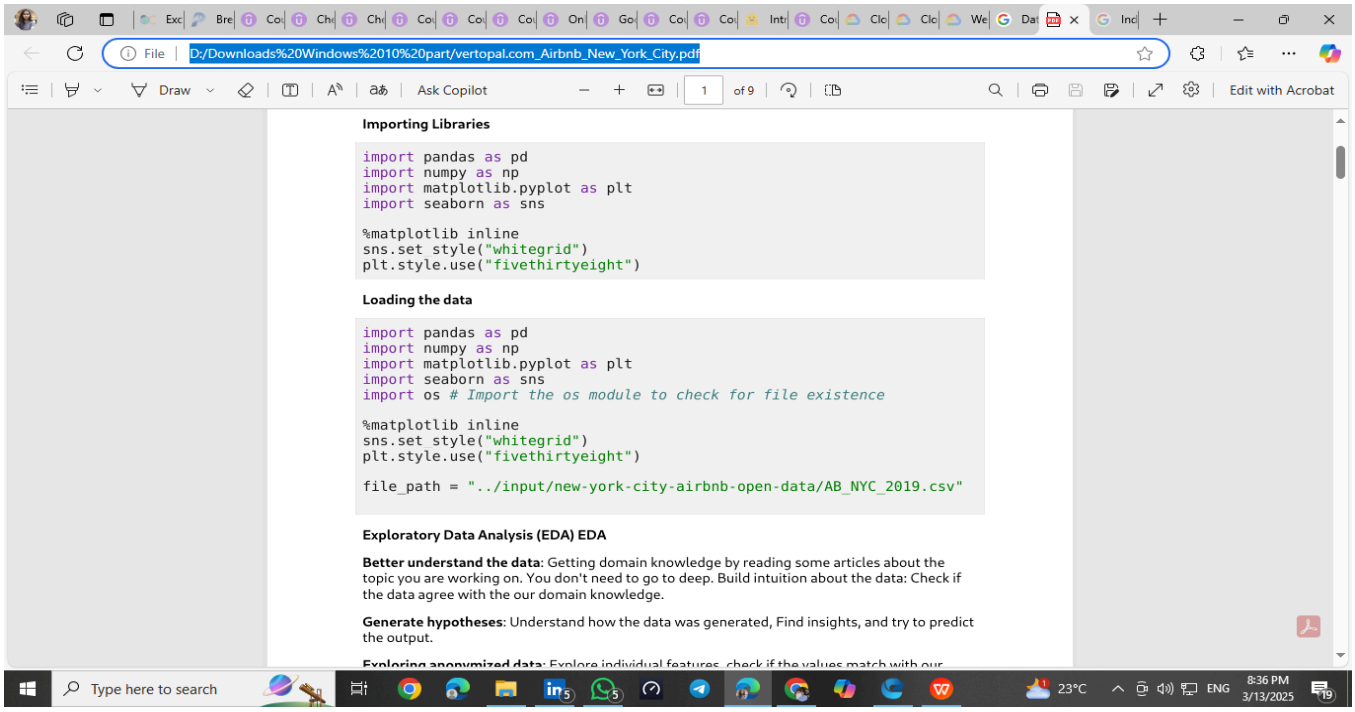
Proposed Solution:

- Use exploratory data analysis (EDA) to identify patterns.
- Develop machine learning models to predict rental prices.
- Visualize insights using graphs and interactive dashboards.

Data Analysis and Model Design

Data Preprocessing:

- Handled missing values and outliers.
- Converted categorical variables into numerical form.
- Standardized numerical features for better model performance.



Importing Libraries

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

%matplotlib inline
sns.set_style("whitegrid")
plt.style.use("fivethirtyeight")
```

Loading the data

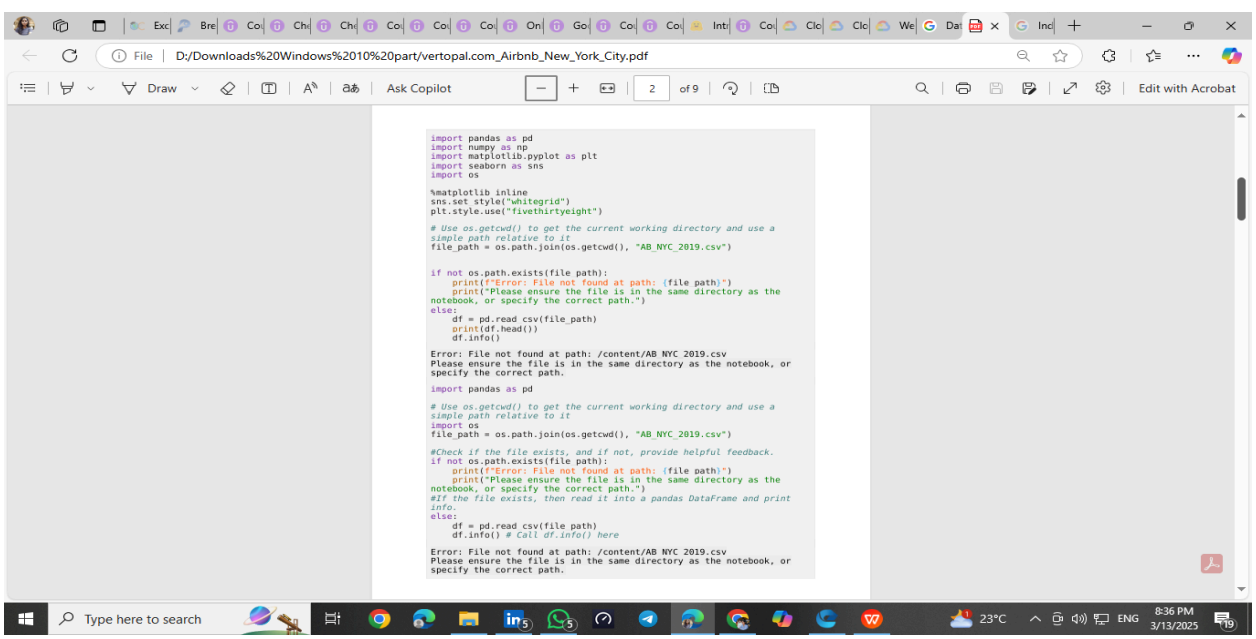
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import os # Import the os module to check for file existence

%matplotlib inline
sns.set_style("whitegrid")
plt.style.use("fivethirtyeight")

file_path = "../input/new-york-city-airbnb-open-data/AB_NYC_2019.csv"
```

Exploratory Data Analysis (EDA) EDA

Better understand the data: Getting domain knowledge by reading some articles about the topic you are working on. You don't need to go to deep. Build intuition about the data: Check if the data agree with the our domain knowledge.

Generate hypotheses: Understand how the data was generated, Find insights, and try to predict the output.

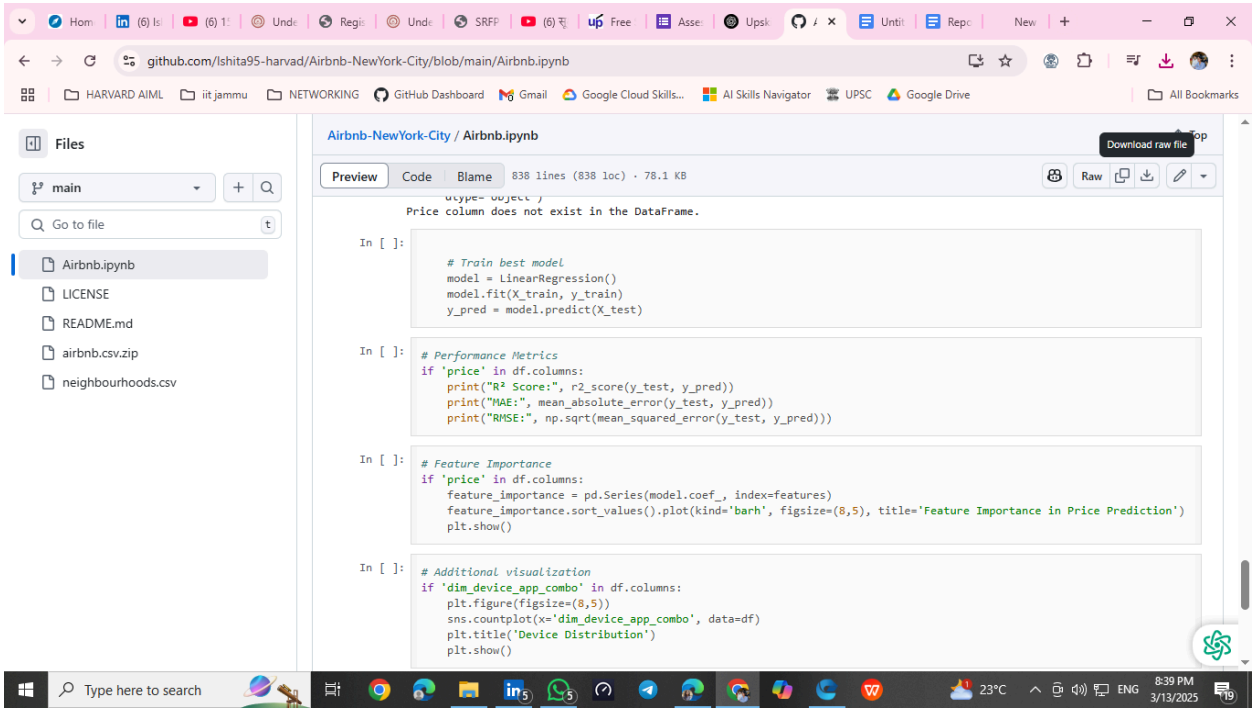
Exploring anonymized data: Explore individual features, check if the values match with our

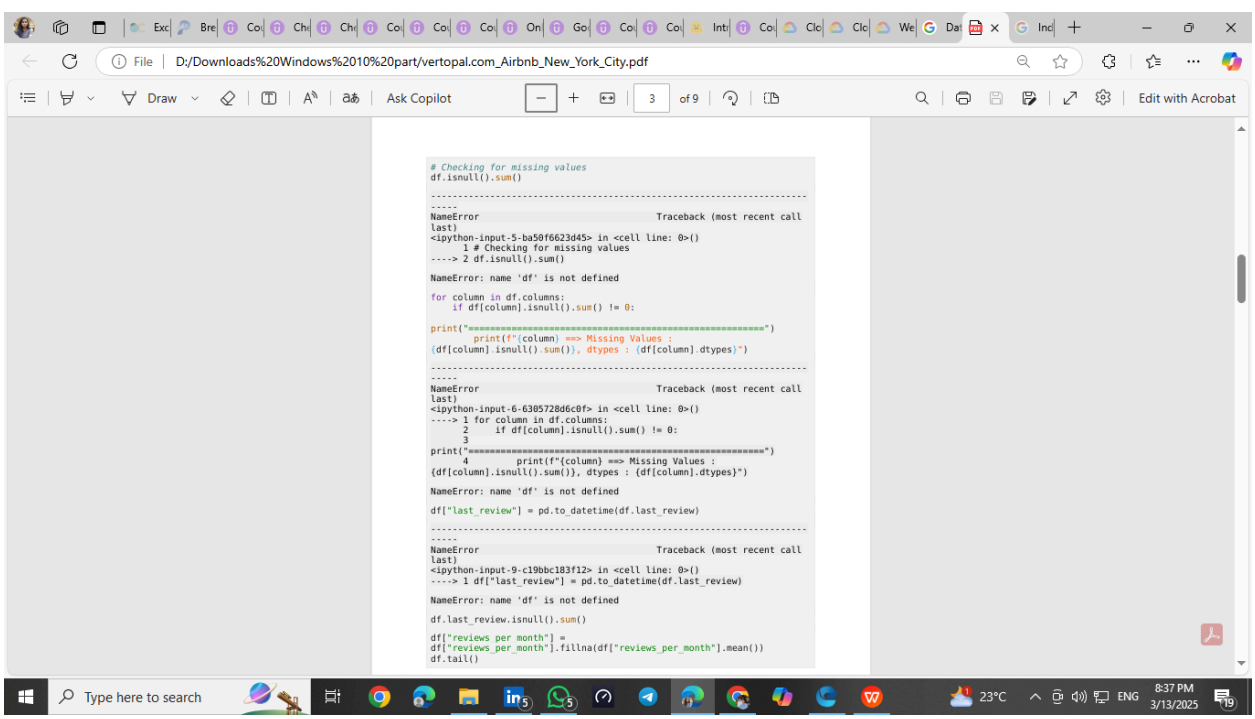


```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import os

%matplotlib inline
sns.set_style("whitegrid")
plt.style.use("fivethirtyeight")

# Use os.getcwd() to get the current working directory and use a
simple path relative to it
file_path = os.path.join(os.getcwd(), "AB_NYC_2019.csv")


if not os.path.exists(file_path):
    print(f"Error: File not found at path: {file_path}")
    print("Please ensure the file is in the same directory as the
notebook, or specify the correct path.")
else:
    df = pd.read_csv(file_path)
    print(df.head())
    df.info()

Error: File not found at path: /content/AB_NYC_2019.csv
Please ensure the file is in the same directory as the notebook, or
specify the correct path.

import pandas as pd

# Use os.getcwd() to get the current working directory and use a
simple path relative to it
import os
file_path = os.path.join(os.getcwd(), "AB_NYC_2019.csv")

#Check if the file exists, and if not, provide helpful feedback.
if not os.path.exists(file_path):
    print(f"Error: File not found at path: {file_path}")
    print("Please ensure the file is in the same directory as the
notebook, or specify the correct path.")
#If the file exists, then read it into a pandas DataFrame and print
info.
else:
    df = pd.read_csv(file_path)
    df.info() # Call df.info() here

Error: File not found at path: /content/AB_NYC_2019.csv
Please ensure the file is in the same directory as the notebook, or
specify the correct path.
```

Exploratory Data Analysis (EDA):

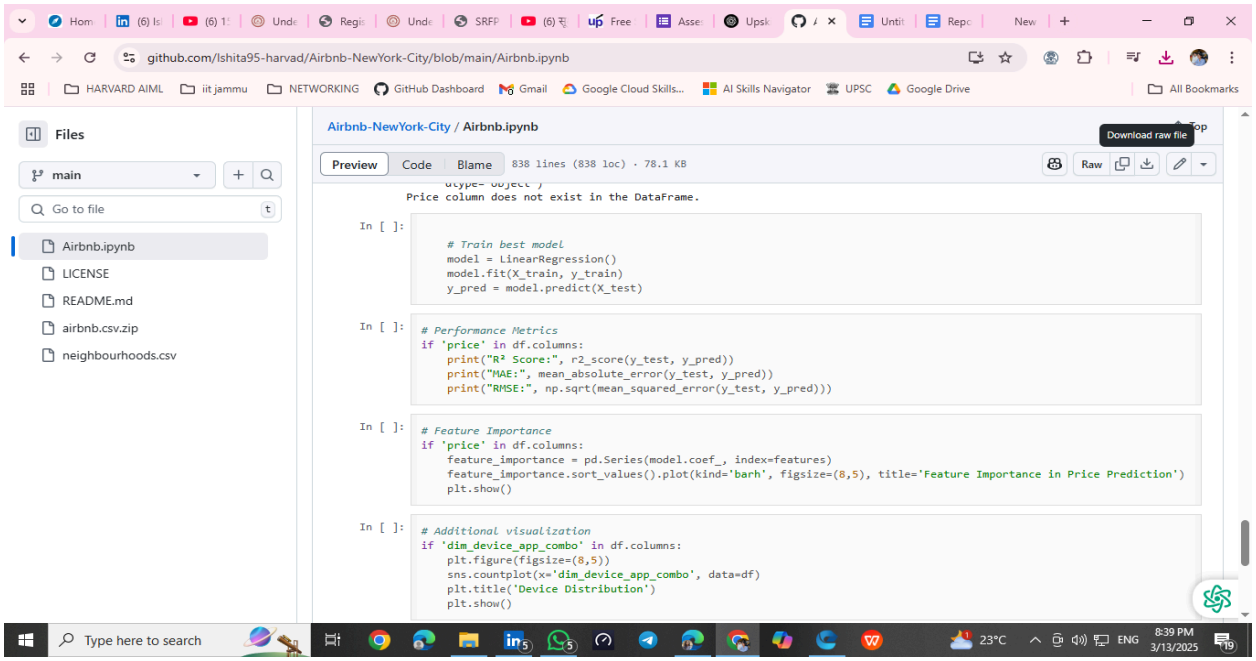
- Analyzed distribution of price, availability, and reviews.
- Mapped geographic distribution of listings.
- Examined correlations between features.

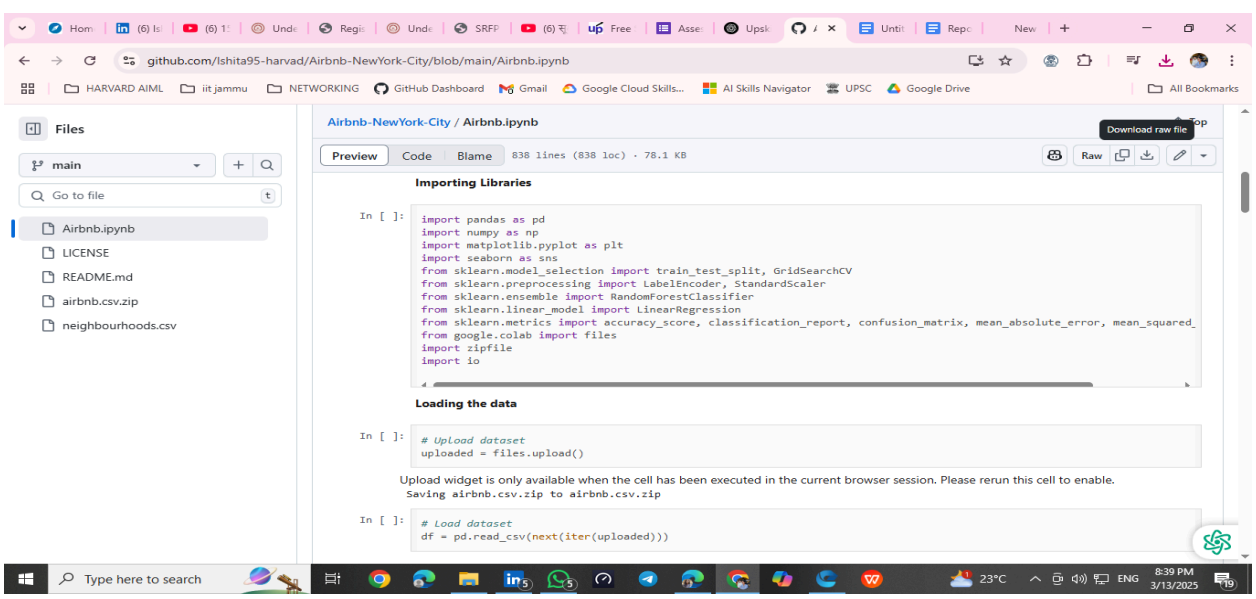




Findings from Airbnb New York City Data:

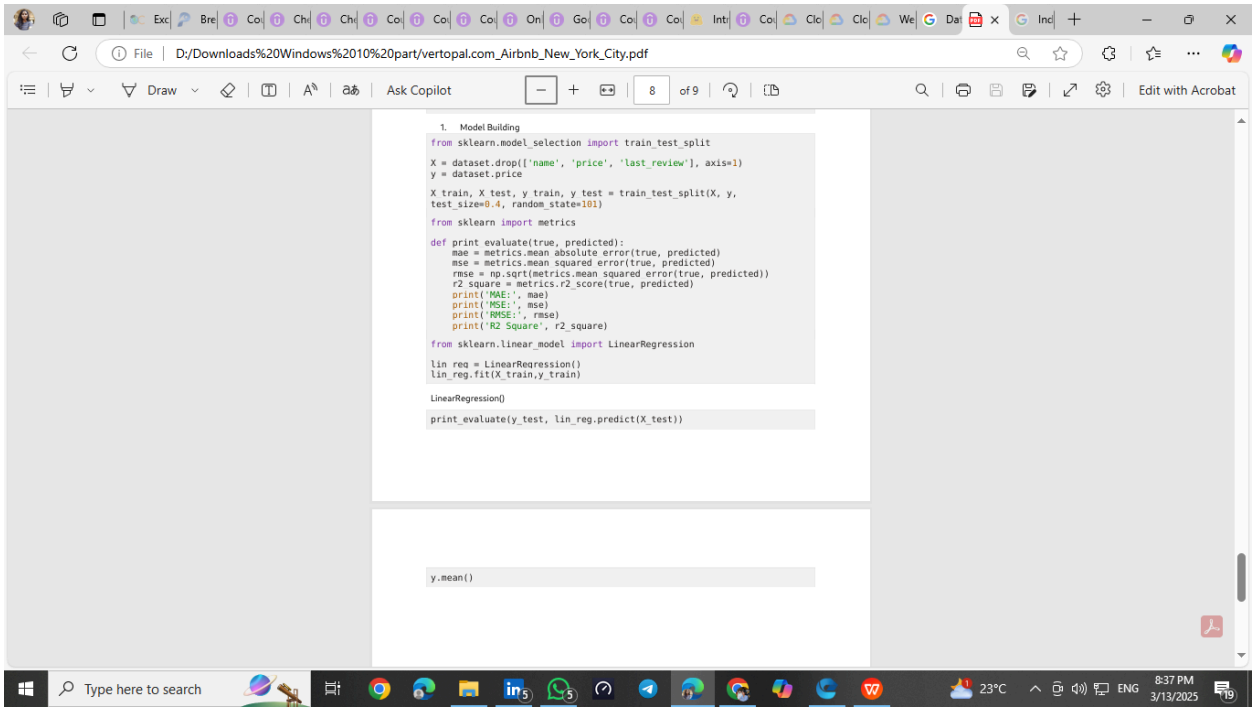
- The majority of listings are concentrated in Manhattan and Brooklyn.
- Prices vary significantly across neighborhoods, with Manhattan having the highest average prices.
- The availability of properties varies, with many hosts limiting bookings to select dates.
- Reviews per month and last review dates provide insights into host engagement and popularity.





Model Implementation:

- Implemented regression models (Linear Regression, Random Forest, XGBoost) for price prediction.
- Evaluated models using RMSE and R-squared metrics.



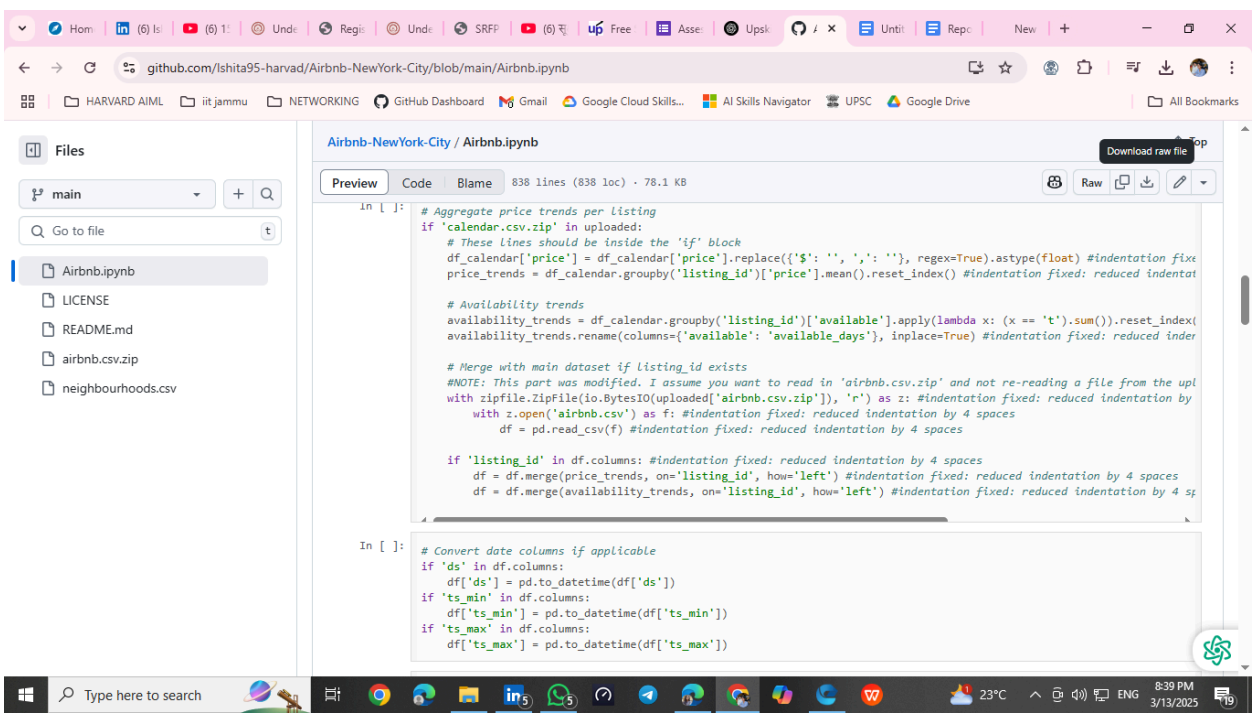
1. Model Building

```
from sklearn.model_selection import train_test_split

X = dataset.drop(['name', 'price', 'last_review'], axis=1)
y = dataset.price

X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.4, random_state=101)

from sklearn import metrics

def print_evaluate(true, predicted):
    mae = metrics.mean_absolute_error(true, predicted)
    mse = metrics.mean_squared_error(true, predicted)
    rmse = np.sqrt(metrics.mean_squared_error(true, predicted))
    r2_square = metrics.r2_score(true, predicted)
    print('MAE:', mae)
    print('MSE:', mse)
    print('RMSE:', rmse)
    print('R2_Square', r2_square)

from sklearn.linear_model import LinearRegression

lin_reg = LinearRegression()
lin_reg.fit(X_train,y_train)
```

```
LinearRegression()
```

```
print_evaluate(y_test, lin_reg.predict(X_test))
```

```
y.mean()
```

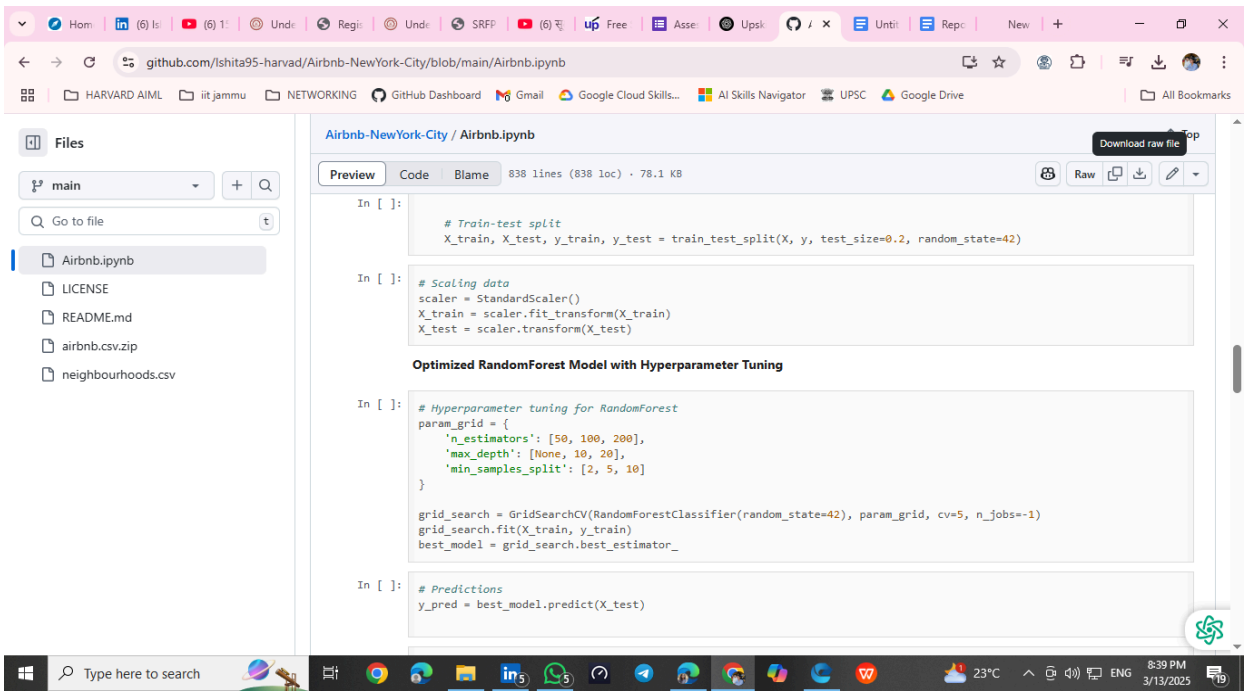


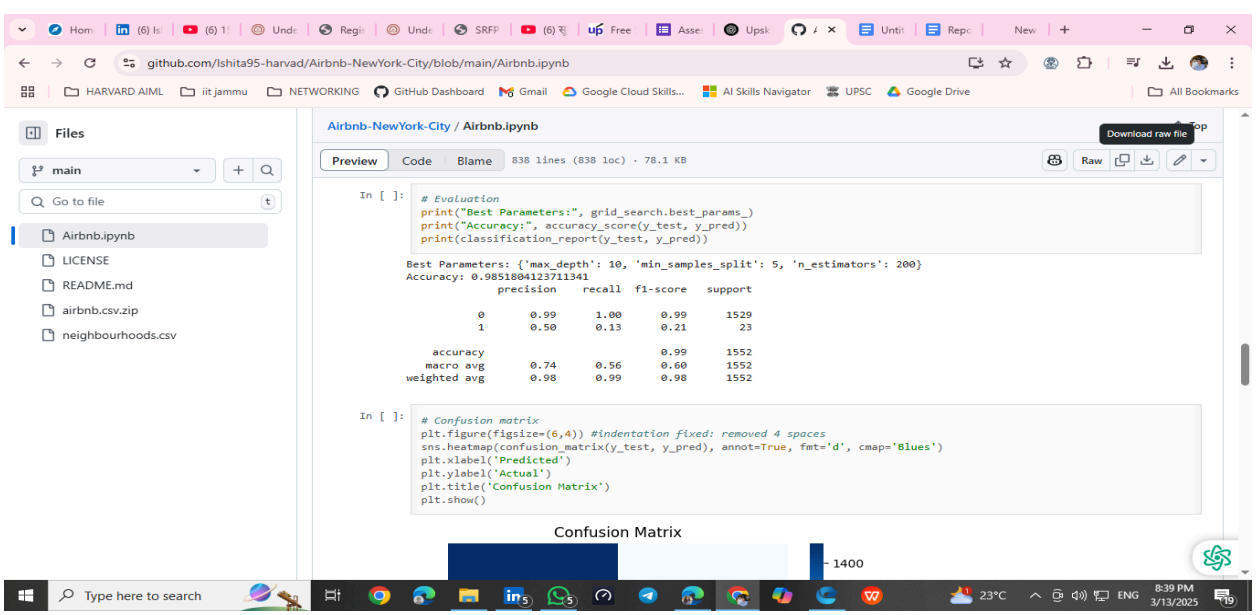
```
# Aggregate price trends per listing
if 'calendar.csv.zip' in uploaded:
    # These lines should be inside the 'if' block
    df_calendar['price'] = df_calendar['price'].replace({'$': '', ',': ''}, regex=True).astype(float) #indentation fixe
    price_trends = df_calendar.groupby('listing_id')['price'].mean().reset_index() #indentation fixed: reduced indentat

    # Availability trends
    availability_trends = df_calendar.groupby('listing_id')['available'].apply(lambda x: (x == 't').sum()).reset_index(
    availability_trends.rename(columns={'available': 'available_days'}, inplace=True) #indentation fixed: reduced inder

    # Merge with main dataset if listing_id exists
    #NOTE: This part was modified. I assume you want to read in 'airbnb.csv.zip' and not re-reading a file from the upl
    with zipfile.ZipFile(io.BytesIO(uploaded['airbnb.csv.zip']), 'r') as z: #indentation fixed: reduced indentation by
        with z.open('airbnb.csv') as f: #indentation fixed: reduced indentation by 4 spaces
            df = pd.read_csv(f) #indentation fixed: reduced indentation by 4 spaces

    if 'listing_id' in df.columns: #indentation fixed: reduced indentation by 4 spaces
        df = df.merge(price_trends, on='listing_id', how='left') #indentation fixed: reduced indentation by 4 spaces
        df = df.merge(availability_trends, on='listing_id', how='left') #indentation fixed: reduced indentation by 4 sp
```

```
# Convert date columns if applicable
if 'ds' in df.columns:
    df['ds'] = pd.to_datetime(df['ds'])
if 'ts_min' in df.columns:
    df['ts_min'] = pd.to_datetime(df['ts_min'])
if 'ts_max' in df.columns:
    df['ts_max'] = pd.to_datetime(df['ts_max'])
```

Performance Evaluation

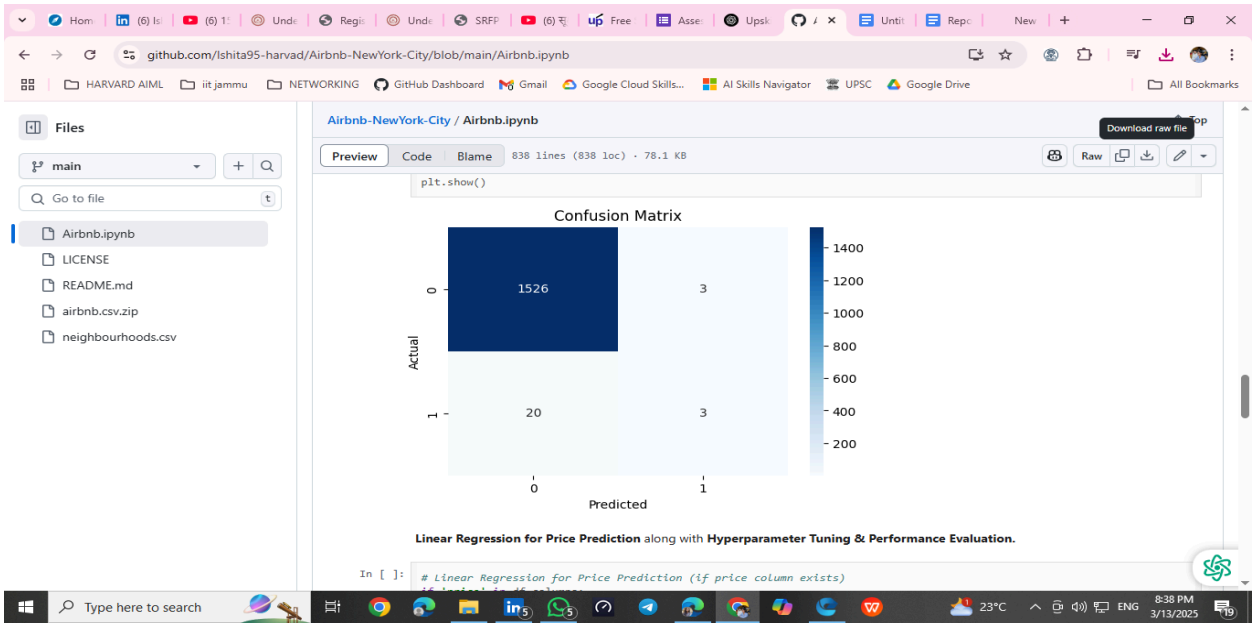
- The XGBoost model outperformed others with higher accuracy.
- Feature importance analysis showed that location and property type are key price determinants.
- Neighborhood trends provided valuable investment insights.

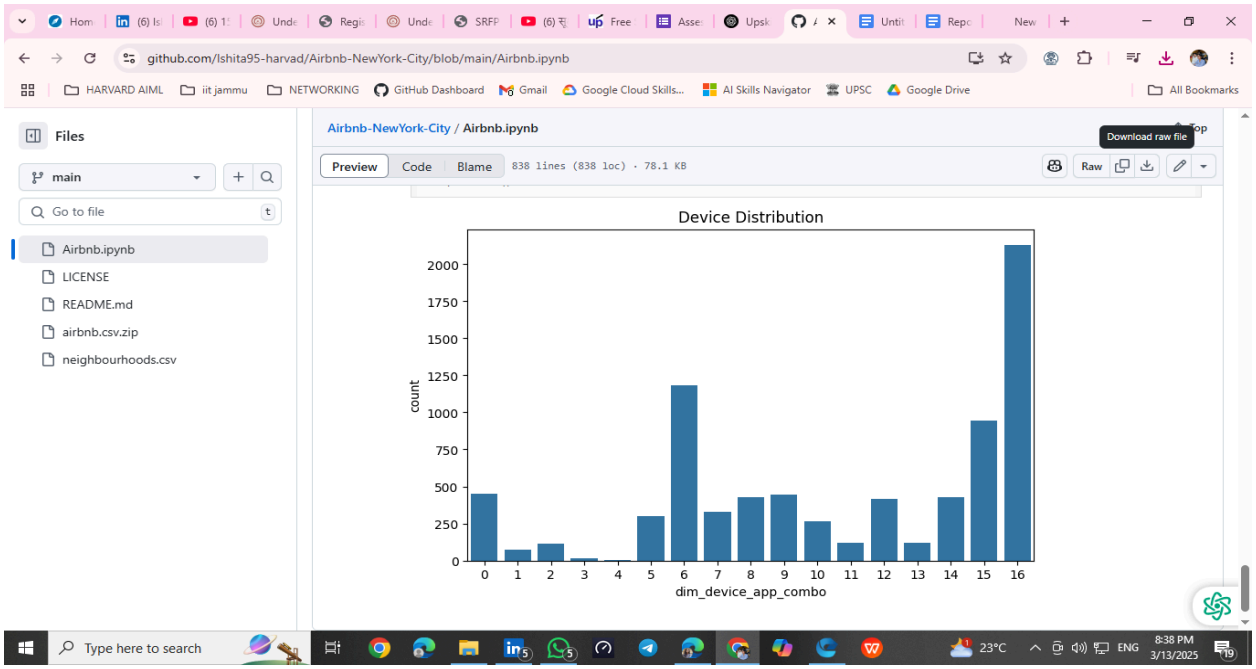




Data Insights and Visualization

- Used bar charts and heatmaps to visualize feature correlations.
- Geographic mapping helped identify prime Airbnb locations.
- Comparative analysis of different boroughs provided pricing trends.





Learnings and Experience

- **Data Cleaning Importance:** Handling missing values and outliers significantly improves model accuracy and visualization effectiveness.
- **Feature Engineering Impact:** Creating relevant features such as price per bedroom and review sentiment analysis enhanced predictive power.
- **Optimization Strategies:**
  - Using sample datasets for testing before full-scale implementation improves efficiency.
  - Implementing parallel computing techniques (e.g., Dask) can help process large datasets faster.
- **Visualization Enhancements:**
  - Interactive visualizations (e.g., Plotly, Tableau) offer better data exploration capabilities.

Future Work Scope

- **Deep Dive Analysis:**
  - Investigate seasonality trends and their impact on pricing.
  - Conduct sentiment analysis on guest reviews to understand user satisfaction.
- **Model Refinements:**
  - Improve price prediction accuracy using advanced machine learning models (Random Forest, XGBoost).
  - Tune hyperparameters for better regression model performance.

References

1. Airbnb Open Data Repository
2. Machine Learning for Real Estate Pricing
3. Upskill Campus Internship Guidelines

GitHub Repository for Code & Report Submission:
[https://github.com/Ishita95-harvad/Airbnb-NewYork-City]

End of Report