**Final Report: Airbnb Equity & Pricing Analyzer**

**1. Project Title**

**Airbnb Equity & Pricing Analyzer**

**2. Objective**

The project aims to design a data-driven application to:

* Predict Airbnb listing prices using machine learning.
* Help **hosts** optimize pricing and amenities.
* Help **travelers** discover high-value, affordable, and safe stays.
* Promote **equitable tourism** by identifying underserved neighborhoods with growth potential.

**3. Dataset Overview**

**Primary Dataset:** Airbnb NYC Open Data

* Fields: Price, Room Type, Bedrooms, Amenities, Reviews, etc.
* Source: Inside Airbnb (public dataset)

**Supplementary Dataset:** ZIP Code Income Data

* Source: <https://simplemaps.com/>
* Used to calculate income-based equity metrics

**4. Methodology**

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

We performed 10+ EDA visualizations to understand trends:

* Price distribution
* Room type vs price
* Availability vs income
* Income vs price
* Amenities vs price
* ZIP code median pricing
* Property type distribution
* Correlation heatmap

**B. Feature Engineering**

* Extracted amenities count
* Created price per person feature
* Converted room type and host verifications to numerical forms
* Calculated host listing ratio

**C. Modeling**

* **Algorithm:** Random Forest Regressor
* **Target:** price
* **Features:** 16 engineered variables (e.g., bedrooms, amenities, income, host metrics)
* **Evaluation Metrics:**
  + MAE
  + RMSE
  + R^2

**D. SHAP Analysis**

* Used SHAP values to identify impactful features
* Provided amenity and room-type suggestions to increase predicted price

**E. Equity Score Calculation**

ZIP Code Equity Score = Weighted Score of:

* Income (40%) — lower income = higher score
* Price (30%) — affordable pricing = higher score
* Rating (30%) — higher quality = higher score

**5. Application Features (Streamlit App)**

**A. Introduction**

* Project overview, purpose, and dataset description

**B. EDA**

* Interactive charts with explanations
* Sliders and filters for exploring price trends

**C. Model Section**

* Train model button with live metrics display
* MAE, RMSE, R^2 visualized with st.metric
* Actual vs predicted plot

**D. Host Mode**

* Enter listing features to get predicted price
* Get actionable upgrade suggestions using SHAP
* ZIP-level price comparison table

**E. Traveler Mode**

* Filters: max price, beds, ZIP code, income tier, room type
* Runtime prediction based on guest preferences
* Top recommended listings by value score

**F. Communities Mode**

* Equity score visualization (bar + map)
* Summary metrics (ZIPs evaluated, avg score)
* Listings in high-equity ZIPs with filters
* Explanation of scoring in st.expander

**6. Results and Insights**

* The model predicts price accurately with ~84% R^2
* Hosts can gain ~$20–$50 by optimizing amenities and listing types
* Several underserved ZIPs have high-value listings for travelers
* Equity score highlights areas where tourism can be used as a tool for local upliftment

**7. Tools and Technologies**

* Python (pandas, scikit-learn, matplotlib, seaborn, pydeck)
* Streamlit for app deployment
* SHAP for interpretability
* Open datasets (Airbnb, US Census)

**8. Conclusion**

This project bridges machine learning and social impact by providing:

* Transparent pricing insights
* Support for underserved communities
* Smart tools for hosts and travelers

It demonstrates a full-cycle data science pipeline, from EDA to ML to deployment — with a focus on fairness and usability.

**9. Future Work**

* Deploy publicly via Streamlit Cloud or as executable
* Integrate Google Maps API for listings map
* Add sentiment analysis of reviews for quality score
* Explore equity over time (trend analysis)

**10. Team Member**

**Name:** Muhammad Tayyab

**Course:** Introduction to Data Science  
**Instructor:** Muhammad Nadeem Majeed  
**University:** PUCIT