Project Plan: Data Amalgamation for Airbnb Success Prediction

Student: Ananya Praveen Shetty, Apoorva Shastry

Date: September 16, 2025

Objective: The objective is to determine the plan of strategy to incorporate three separate datasets into a single enriched dataset. The enriched dataset will be used to train efficient and robust classification models with the purpose to predict the success of Airbnb listings in Los Angeles.

Dataset Inventory

Our analysis will include three data sets, each providing a separate level of information.

Dataset 1: Airbnb Listings (listings.csv)

Description: This is our key dataset. It contains comprehensive details on individual listings on Airbnb, their geolocation (longitude and latitude), their price, their room types and their guest rating score.

Role: It forms our analysis foundation. A row is a single listing on Airbnb that we aim to analyze and categorize at the end.

Dataset 2: U.S. Census Data (ACSDT5Y2023.B19013-Data.csv)

Description: This is a dataset of socioeconomic information from the U.S. Census Bureau. The key variable in our undertaking is the median household income and is presented by ZIP Code Tabulation Area (ZCTA).

Role: This dataset provides valuable socioeconomic context to each listing's neighborhood.

Dataset 3: Walkability Index (Walkability_Index.csv & 2020_Census_Tracts.geojson)

Description: This dataset contains a "Walkability" score of how pedestrian-friendly an area is based on density and how readily amenities are accessible. The score is at the Census Tract level. This requires two files: one with the score and a second with the geographical map shapes (.geojson).

Role: This data contains locational convenience context, a significant factor to consider for travelers and tourists.

---

2. Amalgamation Plan & Join Strategy

The amalgamation will be a sequential, two-phase process. We will progressively enrich our primary dataset.

Phase 1: Creating `Dataset 1+2` (Listings + Income)

- **Challenge:** The Airbnb data has geographic coordinates, while the Census data has ZIP codes. There is no common column to join on directly.
- **Planned Solution: Reverse Geocoding**
    - We will process each Airbnb listing's `latitude` and `longitude` to determine its corresponding `zip_code`. This creates the necessary common key.
- **Join Type: Attribute Left Join**
    - We will perform a **left join**, starting with the Airbnb dataset (`Dataset 1`) and adding columns from the Census dataset (`Dataset 2`). This ensures that every original Airbnb listing is preserved in our new table.

Phase 2: Creating `Dataset 1+2+3` (Listings + Income + Walkability)

- **Challenge:** The amalgamated `Dataset 1+2` has point coordinates, while the Walkability data is organized by Census Tract polygons (geographic areas). Again, there is no simple common column.
- **Planned Solution: Spatial Join**
    - This is a more advanced join that works on geographic location. The process will check each Airbnb listing's coordinate point to see which Census Tract polygon it falls **within**.
- **Join Type: Spatial Left Join**
    - We will perform a **spatial left join**, starting with our `Dataset 1+2` and adding the `Walkability` score from `Dataset 3`. This, again, guarantees that no listings are dropped during the merge.

---

3. Data Relationships & Data Loss

- **Relationship Between Datasets:** The final combined dataset will be a richer version of the first. The first listings dataset is the "skeleton," while the Census and Walkability datasets add the "flesh and muscle" by providing necessary context on the surroundings in which each listing is located.
- **Data Loss:** There is no loss of Airbnb listings (rows) in this process. Using left joins alone, we will always retain our initial dataset. Nonetheless, listings will or will not contain an income or walk score (e.g., if it is a new ZIP code and is not in the Census dataset). For these rows, the respective new columns will contain a NaN (Not a Number) value. This is a signal of missing data and is not a loss of data and will be corrected in the cleaning steps ahead of modeling.

4. Questions for Discussion

As we finalize our plan to merge these datasets, we have a few strategic questions we'd like to raise:

1. **Add Strategy and Missing Data**
   Do we preserve all the Airbnb listings and impute the null values (with left joins), or do we choose to have inner joins so our dataset is smaller but neater?
2. **Geographic Scale**
   Since income data are at the ZIP code level while walkability is at Census Tract level, could combining these two levels bring in bias? Is one level usually a better predictor?
3. **Model Choice**
   Will Logistic Regression be enough to capture relationships like "walkability is more important in high-income areas" or will we need higher-complexity models like Random Forest or Gradient Boosting?
4. **Other Data Sources**
   Aside from Census income and walkability, what other public data (e.g., crime levels, transit availability, density of businesses) could improve our predictions?