Muhammad Faizan Kazi
23276745

Influence of Population and it's Density on Airbnb Listings and Pricing in US Cities

Question

"Do cities with larger populations have more Airbnb listings? And does higher population density affect higher average Airbnb prices & room type preferences (e.g., shared vs. private rooms)?"

Data Source 1

The name of the dataset is "US Airbnb Open Data" which is available on Kaggle. It contains a vast record of Airbnb listings in several cities of the United States of America.

The reason why I chose this dataset is because it covers most of the USA and has well defined columns. It gives every listing an id, name, the city it belongs to, the type of the room and most importantly, the price of the listing. These attributes will help me to merge this dataset with US cities Database and see how the demographic of the cities affect the listings and pricing of the rooms.

The structure of the dataset is such that it is in a tabular format, specifically CSV which contains a fixed number of columns. The dataset contains both numerical (Price), categorical features (Room type) and textual columns (name, neighbourhood, host name etc). The quality of the dataset can be claimed as stable and good, since it accurately reflects most of the cities with a good spread of the USA. The dataset contained "Nulls" and a few missing entries. Regarding the timeliness, it is a quite relevant dataset as it provides the facts and figures from 2023, so it is a relevant source of data, keeping in mind the scope of our project to focus on North America.

The licence it uses is CC0 1.0 Universal and since it is public domain, I can work with this dataset without asking permission. The link to the metadata is - Metadata & licence details - Licence details.

Data Source 2

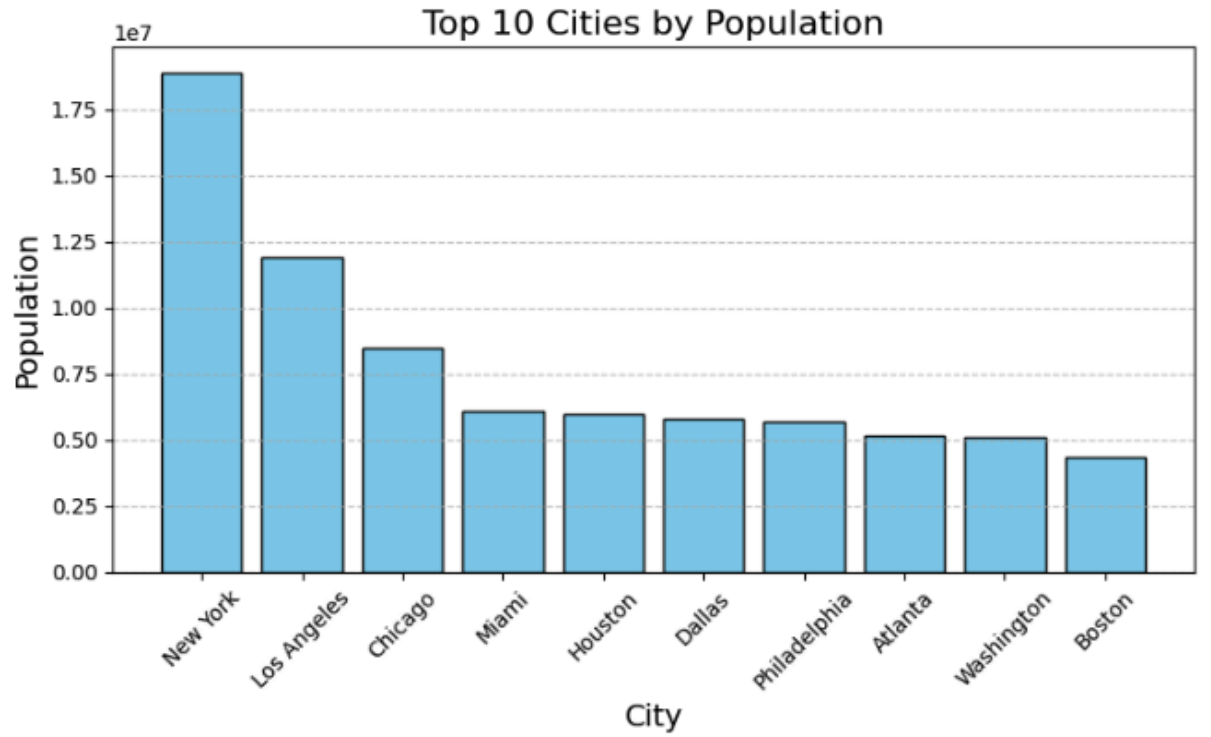
The name of the dataset is "United States Cities Database" which is available on Simplemaps. It contains demographic data of 30,844 cities of the United States of America.

The reason why I chose this dataset is because it covers most of the cities of the USA and has well defined columns. It gives every city with its name, population and the estimated population per square kilometer. These attributes will help me to merge this dataset with US cities Database and see how the demographic of the cities affect the listings and pricing of the rooms.

The structure of the dataset is such that it is in a tabular format, specifically CSV which contains a fixed number of columns. The dataset contains both numerical (Population, density), categorical features (incorporated) and textual columns (name, city, state name etc). The quality of the dataset can be claimed as stable and good, since it accurately reflects most of the cities with a good spread of the USA. The dataset contained only 2 “Nulls”. Regarding the timeliness, it is a quite relevant dataset as it provides the facts and figures from 2024, so it is a relevant source of data, keeping in mind the scope of our project to focus on North America.

The licence it uses is Creative Commons Attribution 4.0 and since it is public domain, I can work with this dataset without asking permission. The link to the metadata is - Metadata & licence details - Licence details. I will be giving the

attribution - give appropriate credit , provide a link to the license, and indicate if changes were made.



Data Pipeline

Technology used by the pipeline was python 3.13. Firstly, zip files are downloaded and both the datasets are extracted. Then, it processes the fetched data and cleans them by removing any NULL or missing values so that joins further don't cause duplicated values on NULL key value joins. One problem I encountered was that in Airbnb Data, the column name neighbourhood matches the city column of the cities dataset more, instead of the city column of Airbnb data, so then I analysed that there are more joins on neighbourhood & city level (inner join). For meta-quality measures, I would say the requests model checks if the extracted data says status 200 when fetching data, then processes further. This is done if ever the API is down, it would not return 200, and the pipeline won't process further. Apart from that, the click module is used for print statements with different colours for different types of prints, red for error, magenta for processing. Doing this would help me to deal with any errors easily.

Result and Limitations

The output data contains merged data of both datasets. So, in an overview, every airbnb listing would also have the demographics of the city it is listed in. I join both

the datasets on city level and save the dataset in CSV format. The reason for choosing CSV is the ease of use and short code needed to read it, which will make the pipeline optimized and will follow the modern standards of using CSV for data projects. One problem I can see is that not each and every city matches using the key(city) due to change in naming conventions which might affect the completeness of my final report. One solution I can think of is trying to join them using longitudes and latitudes which exist in both datasets.