ETL Project

By: Jillian Nemec and Celine Villareal

**Extraction**

This project used two datasets from Kaggle, a public API platform, which highlight Airbnb data from the city of Chicago and major U.S. cities. Both datasets were archived in a .csv format and were the most recent datasets available. Chicago data was collected between 2014 – 2020, while U.S. data was collected between 2018 – 2020. Both datasets contain the following information:

* Airbnb host ID
* Listings and property information
* Price
* Minimum number of nights allowed per booker

**Transformation**

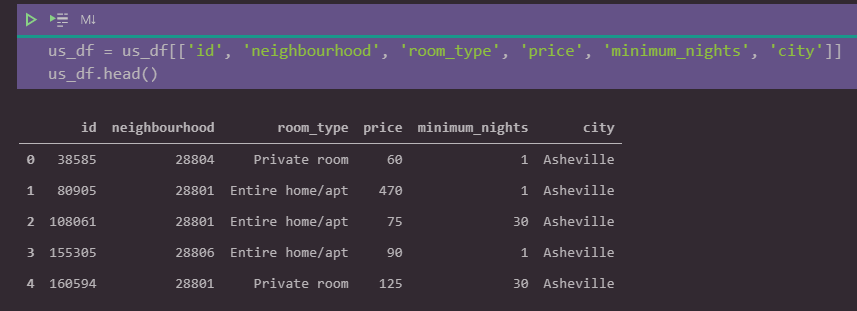
Both .csv files were loaded into separate pandas data frames in order to clean and transform the datasets before filtering. A copy of each of the data frames was created to save the original, untouched dataset. The first step was to identify the specific columns or variables that were relevant for analysis **(Figure 1)**. This project dropped the following columns from both datasets:

* Name
* Host\_id
* Host\_name
* Neighbourhood\_group
* Latitude
* Longitude
* Number\_of\_reviews
* Last\_review
* Reviews\_per\_month
* Calculated\_host\_listings\_count

The remaining relevant columns were then saved into the data frame:

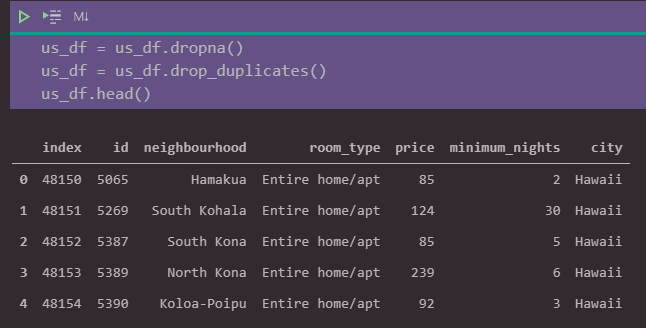
* Id
* Neighborhood
* Room\_type
* Price
* Minimum\_nights
* City

**Figure 1: Filtered Relevant Columns**

****

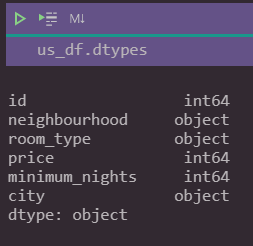
The next step was to remove any duplicate variables and to drop unfilled rows **(Figure 2)**. In particular, the dataset containing U.S. data had multiple undocumented rows.

**Figure 2: Removing Duplicates and NaN rows**

****

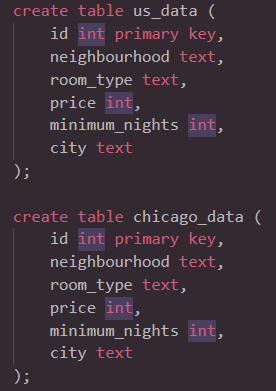
Lastly, we extracted the data types of each column. The data types provided the base information needed to create tables for filtering in a SQL database **(Figure 3)**.

**Figure 3: Extracting Data Types of Columns**



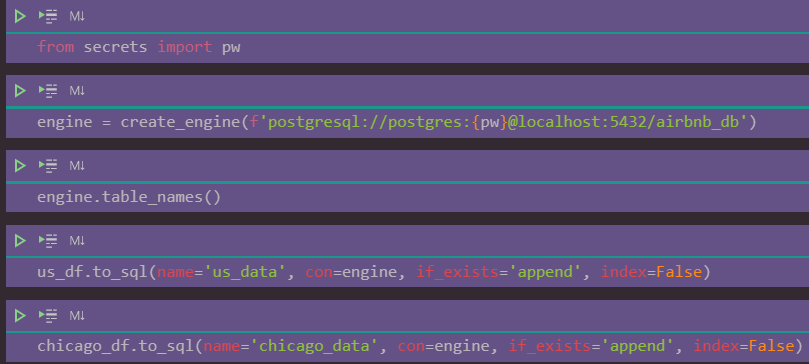
**Load**

Both datasets were transferred to the local PostgreSQL database via SQLAlchemy. Before transferring the data to PostgreSQL, respective tables were created in PostgreSQL to match the columns from the final pandas data frames **(Figure 4)**. The primary key for both the Chicago and U.S. data was the id, which is the unique identifier for each Airbnb location.

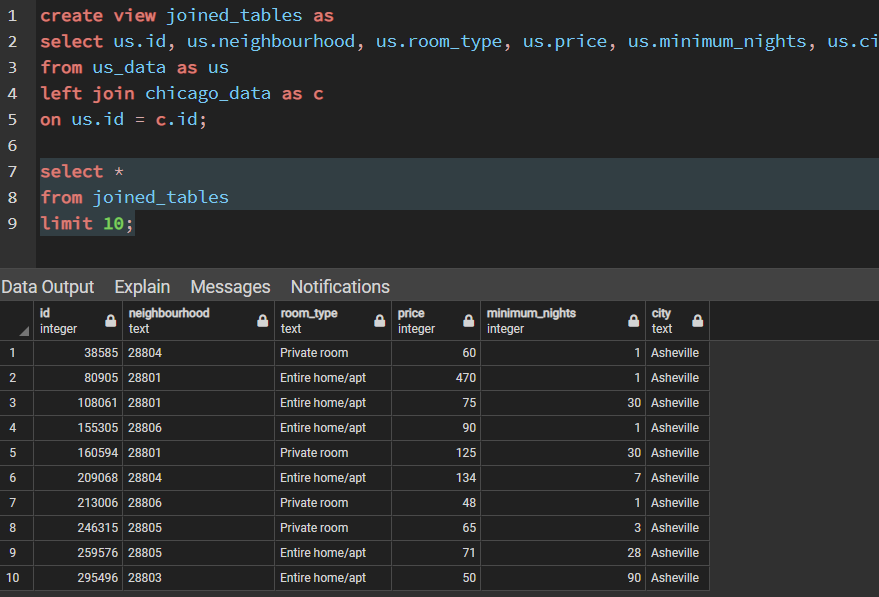
**Figure 4: Query to Create Tables on PostgreSQL**

After creating the tables, the data was transferred to PostgreSQL, a relational database, via a created connection **(Figure 5)**. The data frame information was transferred to each respective SQL table.

**Figure 5: Transferring Datasets to PostgreSQL**

****

Once the data was loaded onto PostgreSQL, we were able to perform multiple queries to filter desired criterion. Criterion were saved into table views to save queries of data for fast reference. For example, the Chicago and U.S. data were joined into a view table called ‘joined\_tables’ on primary key id **(Figure 6)**.

**Figure 6: Example Query**

**Summary: Analysis of Data**

**Figure 1: Top 10 U.S. Cities with the Most Airbnb Locations**

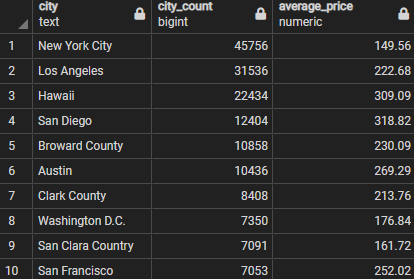
****

Figure 1 Analysis: New York City currently has the most Airbnb locations with an average rental price of $149.56

**Figure 2: Top 10 Cities with the Highest Average Prices Per Stay**

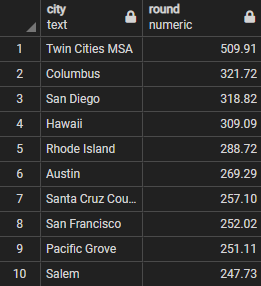


Figure 2 Analysis: Interestingly, the top 10 most expensive Airbnb locations are cities not situated in the most populous areas of the U.S. Nonetheless, Hawaii is a popular travel destination.

**Figure 3: Airbnb Room Types and their Perspective Counts and Average Prices**

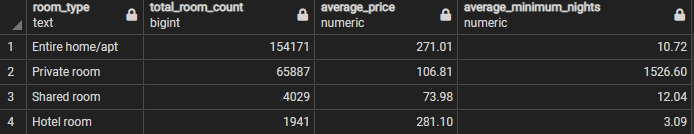


Figure 3 Analysis: The room type most prevalent for Airbnb is the entire home/apt with an average price of $271.01 and average night stay of 10.72 days. In contrast, Hotel Rooms come in last in terms of count and average night stay; but, has the most expensive average price of $281.10. Lastly, a private room has the longest average night stay of 1526.60 days.