ETL Project: due 10/19/2019

Team: Adelia Manuel, John Knight, Rashid Khokhar

Project Proposal:

Following project Proposal was submitted on Oct 12, 2019:

> Using Kaggle as our data resource, we plan to show the relationships between Airbnb listings from both east and west coasts.
>
> So, before your next vacations, reach to us for the best listings in in your area!

Proposal was approved by instructional team and we proceeded with the work

Finding Data:

One of the data sources suggested in the README document containing instructions and guidelines for the project mentioned https://www.kaggle.com as a potential data source.

Team visited the website and started focusing on the AirBnB data and thus the above mentioned proposal was submitted.

New York was selected as the representative city from the East Coast, while Seattle was selected as a West Coast city.

There were multiple data sources for both these cities. After some discussion and looking through various data, one each CSV file was selected for the two cities. These two CSV files appeared to have similar data but not exactly the same data and/or format.

The two files are:

AB_NYC_2019.csv
seattle_01.csv

These two (2) files were downloaded and stored at a subfolder named "Resources" under the folder where Jupyter Notebook work was done.

Data Cleanup & Analysis:

Data was then worked on in the Jupyter Notebook. The two datasets were extracted from the respective CSV files into 2 Pandas DataFrames. Then columns of interest were extracted and renamed the same way across both datasets.

Then the duplicate records were removed from each of the DataFrames.

Several additional checks were made to assure data consistency and worthiness.

Subsequently, a database was created in the PostgreSQL system. The database was named: ETL_Proj_DB

From within the Jupyter Notebook, the two DatFrames were pushed out to two tables within the ETL_Proj_DB database. The two tables were: NY & Seatle

(Please note that within the Jupyter Notebook when confirming the table names via the "engine.table_names()" more than these 2 table names will show up as some additional test tables were also created)

Rest of the data analysis and compiling was done via SQL in the PostgreSQL.

SQL query was created to find average prices and the average number of reviews for each room_type for both data sets. Results of these queries were saved as two new tables: NY_Stats & SE_Stats. A FULL JOIN was run on these two tables to compare average prices for each kind of accommodation.

Result of the FULL JOIN of two tables is shown below:

a Output

	room_type text	ny_avg_price numeric	se_avg_price numeric
1	Entire home/apt	211.79	129.85
2	Private room	89.78	66.40
3	Shared room	70.13	48.76

From above table, it is quite clear that for similar accommodation categories there is a substantial price difference between NY and Seattle. Similar comparisons can be made for other cities which should help one to plan a more economical vacation!