SEPTEMBER 2021

AIRBNB CRIME DATA IN NYC

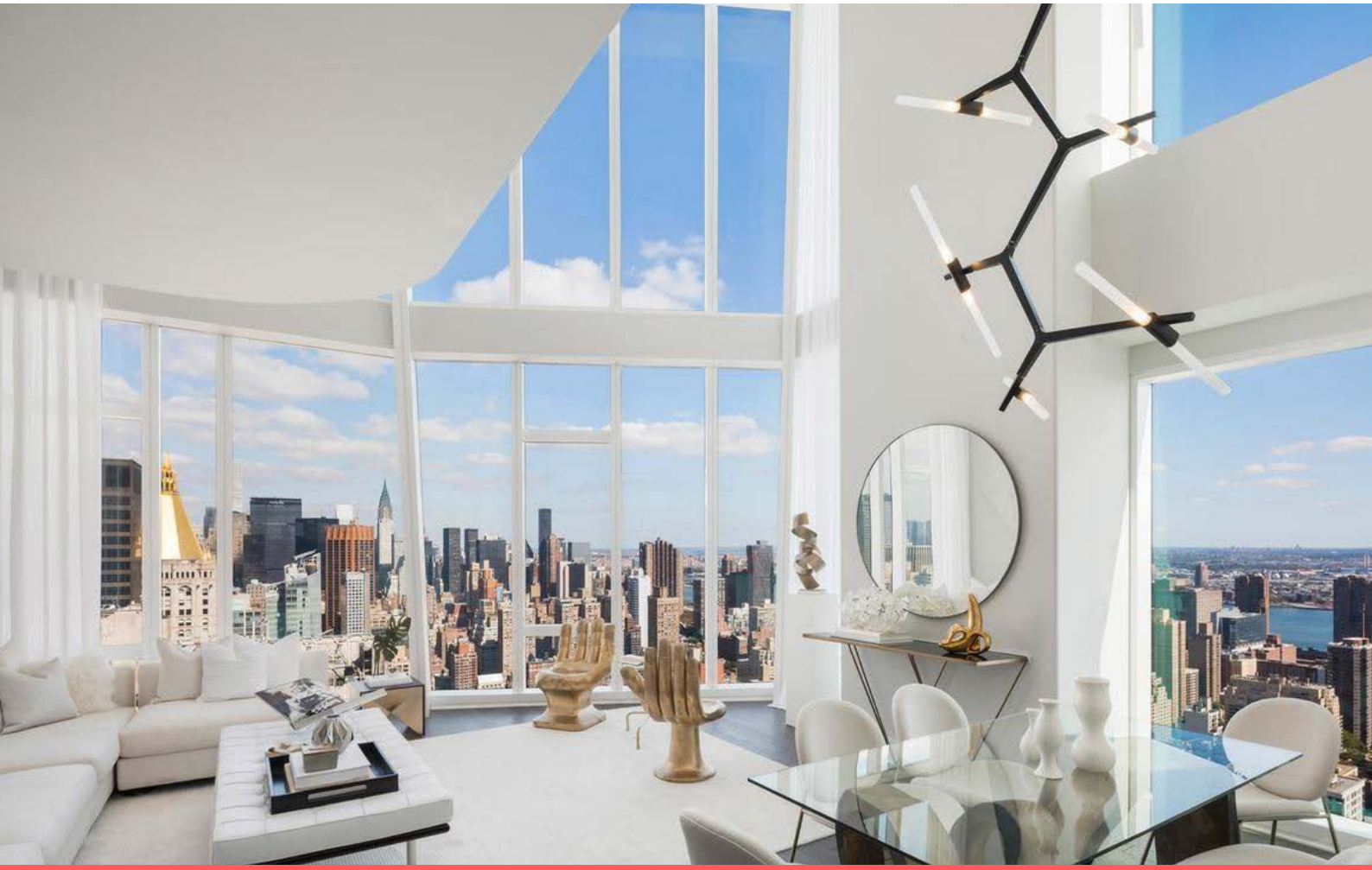
Elite - ETL

PREPARED BY:

Margarita Atehortua
Connor Byington
Phil Henson
Mehdi Rahman
Aaron Sentell

PROJECT SUMMARY

The project aims to prepare a dataset consolidating **Airbnb listings** and reviews and **crime data** in New York City so that an analyst can use the dataset for further examination. The combined data set may be attractive because we suspect there may be more Airbnbs in lower crime areas, and fewer mentions of crime in reviews in those areas.



IMPORTANCE OF THE PROJECT

RISE

of violent offenses in neighborhoods where more homes were converted to short-term rentals.

POSITIVE CORRELATION

between higher penetration of Airbnb properties in an area and a rise in violence.

EXPENSIVE

agreement and damage control from Airbnb to customers involved in serious crimes.

RESPONSIBILITES

Project breakout

- Build GitHub repository
- Read in data from CSV files
- Data transformation

- Clean data in Jupyter notebook
- Trimming unnecessary headers
- Filter by location
- Filter by crime
- Join relevant datasets

- Create ERD
- Create individual data tables
- Build relationships (primary / foreign keys)
- Create user-friendly views

EXTRACTION

The first step was to identify the best data to use for our project. The data was downloaded from Kaggle.com. The group identified the best data source that includes Airbnb listings and New York City crime records. The Airbnb listings CSV had more than 250,000 listings in 10 major cities, including information about hosts, pricing, location, and room type, along with over 5 million historical reviews. The New York City crime dataset was collected through thousands of calls to the NYC OpenData platform from 12/31/2005 to 12/31/2019.

After identifying the datasets, they were open in Jupyter Notebook, and all three datasets were read and transform into a Pandas DataFrame. Later, essential information was determined to reduce the size of the DataFrame and only focus on crucial criteria such as listing's location, reviews, and crimes that are considered violent. Before the transformation, the Airbnb listings CSV had 279,712 rows, and the NYC crime CSV had a total of 3,881,989 rows of information.

TRANSFORMATION

Several data transformation steps were performed after extracting the Airbnb listing, Airbnb review, and crime data CSVs into Pandas data frames. Since the target area for our SQL database was New York City (due to the availability of crime data), the Airbnb listing data frame was filtered to only include listings in New York City. After filtering out non-New York City Airbnbs, 37,012 rows remained in the Airbnb listing data frame.

Next, the crime data frame was trimmed to remove unnecessary columns using the drop function. Five columns were removed because they were internal area calculations used by the NYPD. These fields would not match our Airbnb dataset, which contains districts (boroughs) and latitude/longitude. An additional column (district) was inserted into the crime dataset to align its values with the Airbnb listing dataset.

In this step, the existing arrest_boro column, which contained the first letter of each borough, was used to determine the value of the new district column (which included the full name of the borough) using the map function.

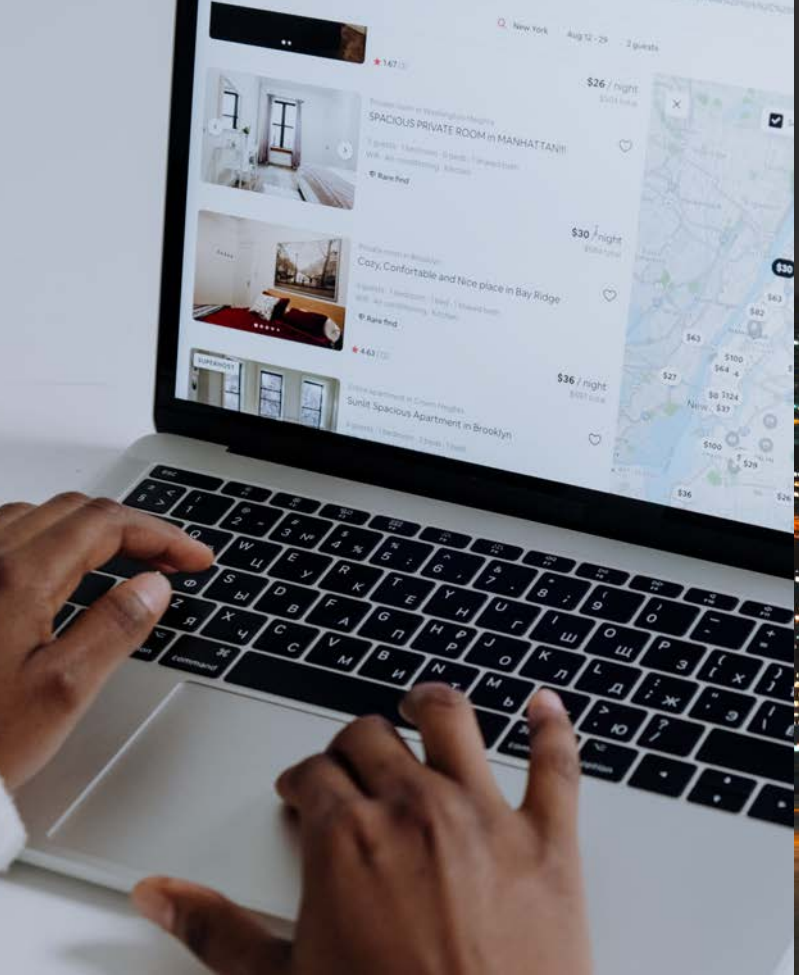
In addition to adding a new column to align the naming of boroughs in the crime and Airbnb listing data frames, the crime data frame was filtered to only include violent crimes. The reason for focusing on violent crimes (violence, sexual violence, and robbery) was based on group consensus that these crimes were more likely to be relevant to Airbnb listing abundance and reviews. After filtering out non-violent crimes, 476,599 rows remained in the NYC crime data frame.

There were no transformation steps performed on the Airbnb review data frame.

LOAD

After data transformation was complete, an entity-relationship diagram was created to design the architecture for the database. Three tables were created: nyc_abnb_listings, abnb_reviews, and nyc_crime. The nyc_abnb_listings table was linked to the abnb_reviews table via the listing_id. In addition, the nyc_abnb_listings table was linked to the nyc_crime table via the latitude and longitude fields.

Later, when the ERD was completed on quickdatabasediagrams.com, the PostgreSQL query was generated to create the tables in the database. After running the query, the CSV files were loaded into the database. While the abnb_reviews and nyc_crime tables accepted the data without issue, there was an issue that had to be troubleshot to get the nyc_abnb_listings table to take the CSV file. Listing ID 5298458 had an ‘ at the end of the address field that caused PostgreSQL to present the following error: “unterminated CSV quoted field in Postgress.” After deleting the apostrophe, the CSV was imported successfully, and the database was ready for further applications.



FUTURE USES

- App to monitor crimes in NYC near the Airbnb listings that the user is considering to stay.
- App for women traveling alone to know better which zones have fewer crime rates and choose a safe short-term rental.
- Data input for short-term rental websites to control listings that appear on the website.

REFERENCES

- Bhat, M. A. (2021). Airbnb Listings & Reviews. Retrieved from https://www.kaggle.com/mysarahmadbhat/airbnb-listings-reviews?select=Airbnb Data
- Carville, O. (2021). Airbnb Is Spending Millions of Dollars to Make Nightmares Go Away. Retrieved from https://www.bloomberg.com/news/features/2021-06-15/airbnb-spends-millions-making-nightmares-at-live-anywhere-rentals-go-away
- Fussell, S. (2021, July 14). Why Do Some Crimes Increase When Airbnbs Come to Town? Retrieved from https://www.wired.com/story/why-some-crimes-increase-when-airbnbs-come-town/
- Karella, A., Patel, J., & Bizzu, N. (2021). NYC Crime Stats. Retrieved from https://www.kaggle.com/ajkarella/nyc-crime-stats