**ETL Project**

**Team Members:**

*Team BLP*

* Bosco Sitati
* Lorenza Culotta
* Pauline Li

**ETL Outline:**

Airbnb listings and Zillow home value datasets by Chicago neighborhood. Postgres and Pandas will be utilized for the ETL process.

**Extract**

**Dataset used:**

* “listings”: Airbnb Listings | Source: <http://insideairbnb.com/get-the-data.html> | Format: CSV.
* “neighborhood”: Zillow Home Values Index | Source: <https://www.zillow.com/research/data/> | Format: CSV

**Transform**

Most of the transformation needed to be done on the neighborhood dataset, in order to obtain the average price per neighborhood:

* Filter by State (IL) and County (Cook) to align to listings.csv
* Filter by year and select only 2019-2020 (available: 1996-2020) to align dates to listings.csv
* Create a AvgPrice column to store the average price per neighborhood

We finally merged the two datasets on “neighborhood”, in order to add the average price and the price/year columns to the listings.csv

**Load**

* The Airbnb database was created using pgAdmin
* To create the tables, a DB Diagram was created using the Quick Database Diagrams app (<https://app.quickdatabasediagrams.com/#/d/xOvKdU>). Three tables were created:
  + Listings: store data from listings.csv
  + Neighborhood: store data from neighborhood.csv
  + Listings\_neighborhood: store data from merged df
* DataFrames were loaded into the database using jyputer notebook (“df.to\_sql” function)

**Challenges**

* We had to be generous with the VARCHAR in the listings table because names of the listings often contain description of the listings themselves.
* Filtering of the data was complicated by the different date format (European vs American)

**If we had more time….**

* We would have spent a bit more time to find a way to retrieve neighborhood names from latitude and longitude coordinates, thus allowing us to use other datasets, like the “crime.csv” dataset found on the CPD website.
* We would have run some analyses, like:
  + Cost of listing vs house prices
  + trend of pricing over time
  + trend of Airbnb pricing over time