**ETL Project**

By Danica Rios & Anusha Ramachandran

**Extraction**

We used data sets from Redfin and Kaggle

* Redfin data for San Francisco real estate market – current listings only
* Flight Search Results from Kaggle
* Airbnb Data from [Inside Airbnb](http://insideairbnb.com/get-the-data.html?source=post_page---------------------------)

**Transformation**

We cleansed each dataset by

* Checking for duplicates since we used the MLS # as the primary key for the Redfin data
* Deleted new construction listings since they didn’t have an MLS #
* Redfin didn’t allow us to download more than 350 per csv. Therefore, we needed to download the data by zipcode within San Francisco
* We also cleansed the search engine results since Postgres didn’t allow special characters

We will be using three Python packages for our work:

* [advertools](https://github.com/eliasdabbas/advertools): To connect to the Google CSE API and receive SERPs in a table format.
* [pandas](https://pandas.pydata.org/): For data manipulation, reshaping, merging, sorting, etc.
* [matplotlib](https://matplotlib.org/): For data visualization.

**Load**

* Once data was cleaned, we loaded it into Postgres