**Data Analyst Job Search**

**Collaborators**

* Megan Moroney
* Nakouma Sako
* Wen Shon
* Matt Moormeier

**Summary**

Our ETL project is geared toward our fellow students who will be seeking employement as data analyst upon completion of our Data Visualization Bootcamp. We located a csv file on Kaggle of data analyst jobs on Glassdoor.com. We determined the top 50 cites with number of data analyst job postings to limit our job search. Then we web scrapped Zillow.com to calculate average apartment pricing for each of our 50 cities. The last dataset we brought in was current weather data for each city. Our goal is to provide quick information about the city that the most data analyst jobs are being offered.

**System Requirements**

* Chrome Web Browser
* Chromedriver
* Python environment running Python 3.7 with the following installations:
  + beautifulsoup4
  + numpy
  + pandas
  + requests
  + splinter
  + SQL

**Data Sources**

We gathered data analyst job listing from Kaggle.com [HERE](https://www.kaggle.com/andrewmvd/data-analyst-jobs) . To determine information about the cost of living we scraped [Zillow.com](https://www.zillow.com/) for apartments in each of our cities and calculated summary information. Finally we utilized the [OpenWeatherMap.org](https://openweathermap.org/) api to get a snapshot of weather in each city.

**Transformation Steps**

We created five separate jupyter notebooks to develop our python code; GlassDoorFinal, OpenWeatherFinal, Zillow\_Scrape\_Final and ZillowSummaryFinal..

**In the the GlassDoorFinal notebook we:**

* Import the DataAnalyst.csv file and drop unneccsary columns.
* Sort the dataframe by cities with the most jobs available and trim it to 50.
* Filter our analyst job dataframe and export to csv
* Create city dataframe and csv to export
* Confirmed there are 50 unique cities. Can use as primary key in SQL

Results

* Glassdoorfinal.csv for import into SQL
* Top50city.csv for import into SQL

**In the OpenWeatherFinal notebook we:**

* List of 50 cities were broken into 3 lists
* City Codes were looked up
* Python request.get function
* The JSON object for the 50 cities were loaded
* JSON strings were converted to dataframe
* Dataframes were concatenated and converted to csv files for export

Results

* Weatherfinal.csv for import into SQL

**In the Zillow\_Scrape\_Final notebook we:**

* Created dictionary of url strings to send to Zillow
* Looped through list and scraped Zillow for housing prices in our 50 cities
* We had to each run code so as not to get blocked

Results

* Created housing\_case.csv file for import into SQL

**In the ZillowSummaryFinal notebook we:**

* Imported the housing\_case.csv file
* Calculated Average Home Price, Max Home Price and Min Home Price for each city
* Exported zillowsummary.csv

Results

* Created zillowsummary.csv for import into SQL

**In Postgres we:**

* Create new database called DAJobSearch
* Create Tables – city, housing, jobs, weather
* Imported csv files into their tables

ERD included called: ERD