Data Science Jobs

Team Name: Kicking Kangaroos

Project Description:

The purpose of this project is to extract data from Indeed to determine the total number of data scientist job openings and company information throughout the United States. Another dataset was also scraped from Zillow to determine housing prices in different cities where there is a high demand for data scientists.

Assigned Tasks:

- ***Andy Kwon***: responsible for scrapping Indeed to retrieve data science roles and salaries.
- ***Wayne Tseng:*** responsible for scrapping Indeed to determine which companies have the highest demands for data science roles.
- ***Fabiola Cartagena:*** responsible for scrapping Zillow to retrieve housing prices by location (cities).
- ***Hyun Soo Kim:*** responsible for creating ERD, and loading data to postgreSQL.

Assumptions:

- No typos were made in the postings
- Indeed postings without the location information was assumed to be remote position
- For locations marked as remote, we used Los Angeles for the project
- For missing data, such as company rating, it was considered unknown or no rating

Scraping Indeed.com for Salary, Location, and Company Data:

The purpose of the project was to build a relational database around indeed.com scraping for location, salary, company name, and job titles. Then, we would scrape Zillow for housing prices in major metropolitan areas, and further scrape for each companies' average salaries. However, we came across several challenges while scraping Iindeed.com, and we had to adjusted the script before finalizing the scraping:

1. Much of the scraped data did not specify a location of the workplace. We gave several thoughts about the reasons behind this occurrence and came to the conclusion that due to the COVID-19 pandemic, most of the jobs had become remote. Therefore, it was important for us to keep the data as it is, instead of considering it null data.
2. Similarly much of the salary information was missing. We also came to the conclusion that due to COVID-19, the salaries had become negotiable.
3. Much of the initial dataset before data cleansing were duplicates from companies constantly updating the same information over and over.
4. Indeed.com was scraped for company information, and ratings. A table was created with the rating for each company. Some rating values have missing information, which is not a major concern for this project since the information is subjective.

Scraping Zillow Data:

1275 properties were scraped from Zillow using Python Notebook. The data included full address and listing prices. Information was saved in multiple csv files, and were later merged into a single extract. The property location was a single string, therefore data was split (using the .split function) into different columns so that it could be loaded to PostgreSQL. Some of the challenges were as follows:

1. Zillow has an API and provided the key instantly, but it takes a couple of days to activate it.
2. Zillow is not fond of scraping. We were only able to scrape by city and only 40 listings were provided. A script was created to sort through multiple pages, but the website limits the number of results. This made the process of obtaining the data more time consuming and somewhat manual.
    a. Note: in the first attempt to scrape the website, the results were greater than 150, but on the second attempt, the results were limited.
3. We made an attempt to combine the various city results into a single DataFrame, but information was missing, so we decided to save the results into csv files and merged the extracts.

Relational Database Creation:

Some of the challenges we faced were as follows:

- Each company/position had varying amounts of information posted in the website. A limited number of companies contained location, and salary information. Separate additional research will be needed if trying to obtain more information on the role.
- Values in the salary column are given in range, instead of average salaries. If our own analysis on the salary amount is to be done, additional data cleaning will be needed.
- Posgresql was very strict on how the tables are formulated with how columns are related. When different scraped csv files did not completely match one another from different sites, it caused a lot of issues.

Analysis:

With collected data, basic analysis of finding which job to apply for can be done.

Indeed was scraped for data scientist and data analyst positions in a few cities. The cities for each position can be joined with the data from Zillow to determine rent or housing prices in the area where the company is located. The results from the analysis can assist job seekers determine if any potential salary increase will cover the cost of living.

Based on the selected role, Indeed provides the company's rating. Although Indeed is not known for accurate ratings or consistent information, this could work as a good deterrent for which company to be more wary.

As mentioned above, much of the data scraped from Indeed.com pertaining to job locations and salaries were missing. However, if they were to be considered null values and be removed, the entire dataset shrinks quite significantly. Therefore, it was necessary to come up with reasoning behind why they were missing in the first place. The problem arose because most jobs switched to remote instead of specific locations after COVID-19 pandemic occurred. With this in mind, it was in fact significant to keep the data, instead of dropping them from columns.