**ETL Project**

Group Members: Albe Eteme, Hussein Issa, Jane Kim, Shailja Mathur

**Proposal**

Amidst the pandemic many people unfortunately lost their jobs, however, demand for Data Scientists remains. We are creating a data set (ETL Project) to help fellow job seekers find qualifying positions.

**Extraction**

We obtained two datasets from the public platform Kaggle. The job listings in the data set are transformed to help find available Data Scientist positions.

The data set contains various job listings for data scientist positions, with fields such as:

* Salary Estimate
* Geographic Location
* Company Rating
* Job Description

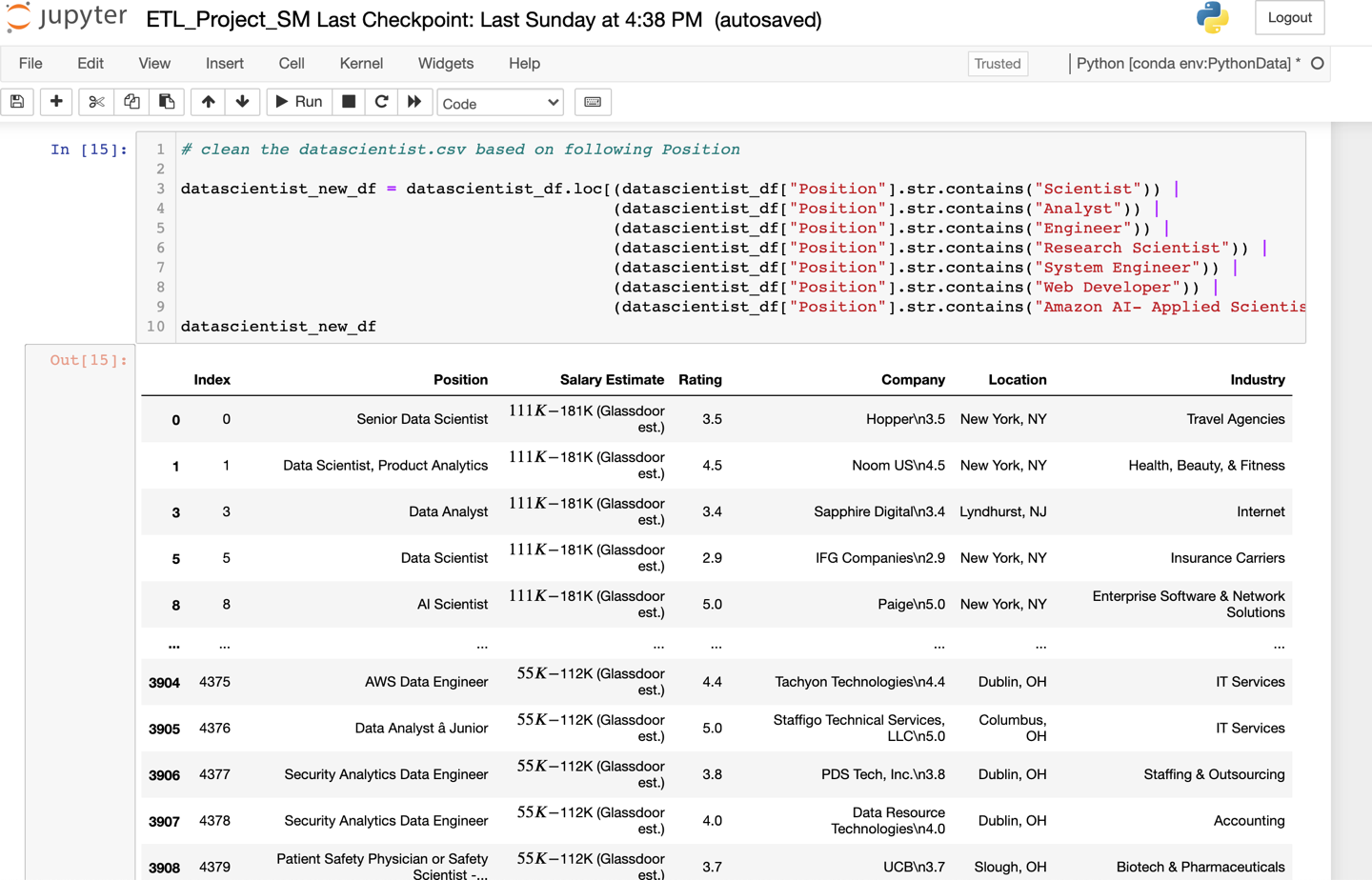
The Sources of our Dataset is as follows

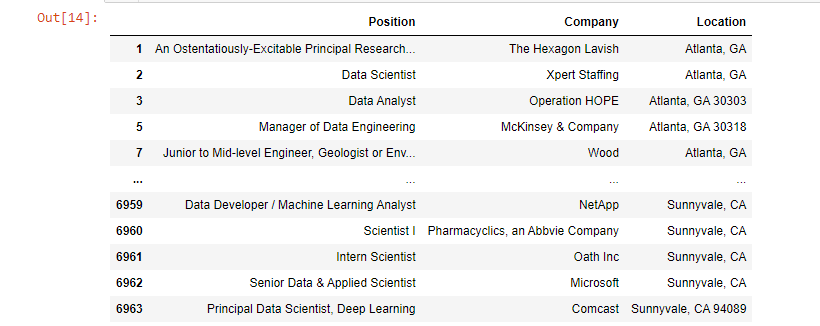
DataScientist.csv- <https://www.kaggle.com/andrewmvd/data-scientist-jobs>

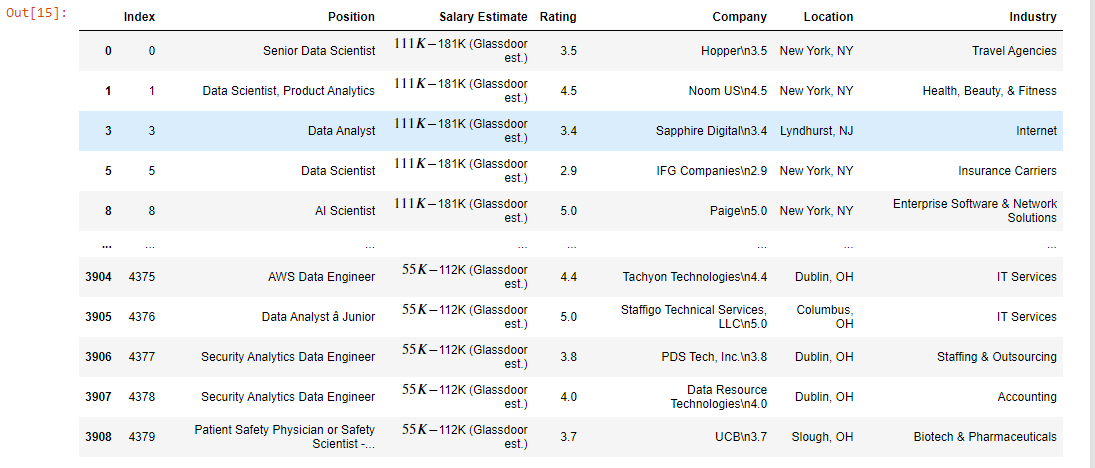
alldata.csv - <https://www.kaggle.com/sl6149/data-scientist-job-market-in-the-us>

**Transformation**

* Our first step is to clean the datasets involving figuring out which variables are not relevant.
* We then cleaned some of our CSVs to make merging easier, renaming/capitalizing column names and getting rid of the index column.
* Using Pandas on Jupyter Notebook we loaded our two CSV’s and transformed them into data frames.
* Removed multiple unnecessary columns:
  + Company HQ location
  + Size
  + Founded
  + Type of Ownership
  + Industry & Sector
  + Revenue
* Sorted job positions to specifically output certain job titles that will help with our hypothesis and research question. Cleaned the database to show Positions, Salary Estimate, Rating, Company, Location, and Industry.





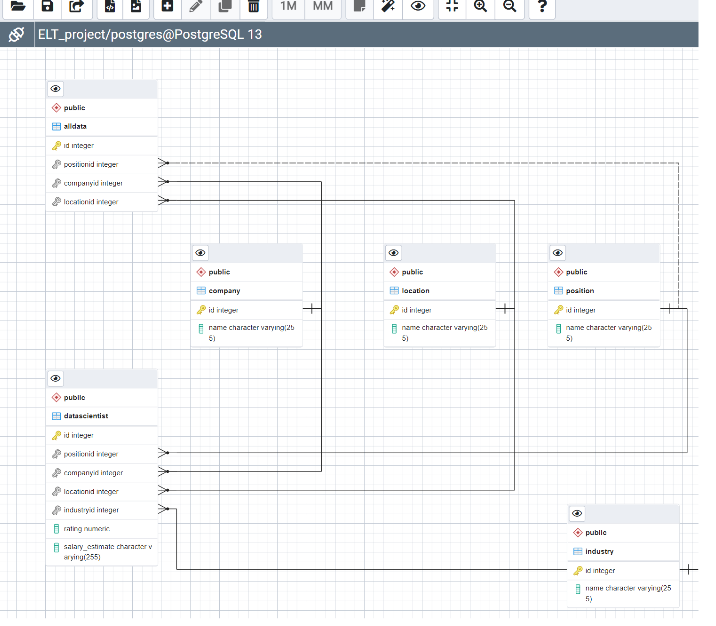


**Load**

After we pulled in the CSV files and loaded them into the data frames, we did an initial connection to the Postgres database using PG admin to store our original clean data sets.

Once loaded in we created the first set of tables (Company, Industry, Position, Location). Running queries and created tables with relevant information reconnecting the database.

Lastly generating our ERD chart for a cleaner look into everything done.



**Summary**

Overall, opening positions for Data Scientists seems to be in high demand even with the pandemic in 2020.

We had some step-backs and limitations to the amount of information we wanted to research, but still found solid evidence for our research questions.

**1. What is the salary range of a Data Scientist in a particular location?**  

Salary estimates through Glassdoor records shows higher demand for Data Scientists in major cities, with positions in New York ranging from 111k-181k and Chicago ranging from $69k-144k.

**2. Relationship between Salary and Company rating?**

The relationship between Salary and Company rating is little, but the companies with higher rating seem to be in major cities (higher salary)

**3. The objective of this part is to define the problem statement of this data to figure out what are the factors the researchers go for (employers, recruiters or employees)?**

These datasets only show opening positions for Data Scientists, the salary estimates, company name, and location during 2020 and not any requirements for each position which would have been good information.