Jobs Database for ‘Analyst’ Jobs:

The objective of this project is to create a database for job finders who are looking for any kind of analyst job roles, select from which job sites they will apply for that job. The database will consist of data from the most visited job sites like Indeed, Glassdoor, Monster etc., with columns Job Id, Job role, Job Description, Skills, Company, Location, Job site (link to the site), twitter info of the job post and company, twitter handles of the companies, most used hashtags, reviews, and ratings of the companies.

Using BeautifulSoup as well as Selenium in python, we scrape the necessary data required and the cleaning and munging of the data is done in Jupyter notebook. Later this cleaned data is converted into ‘.csv’ format in stored in SQL. By then storing this data in a SQL table we can perform the necessary operations such as Searching and Filtering the data.

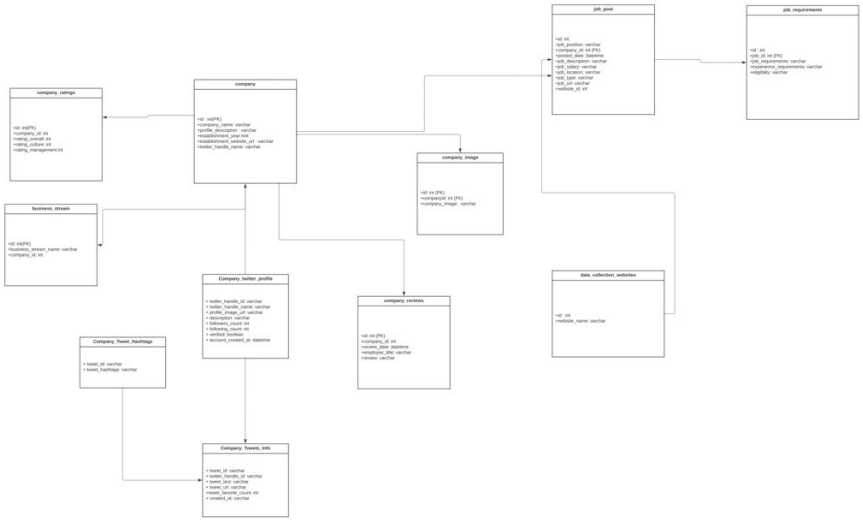


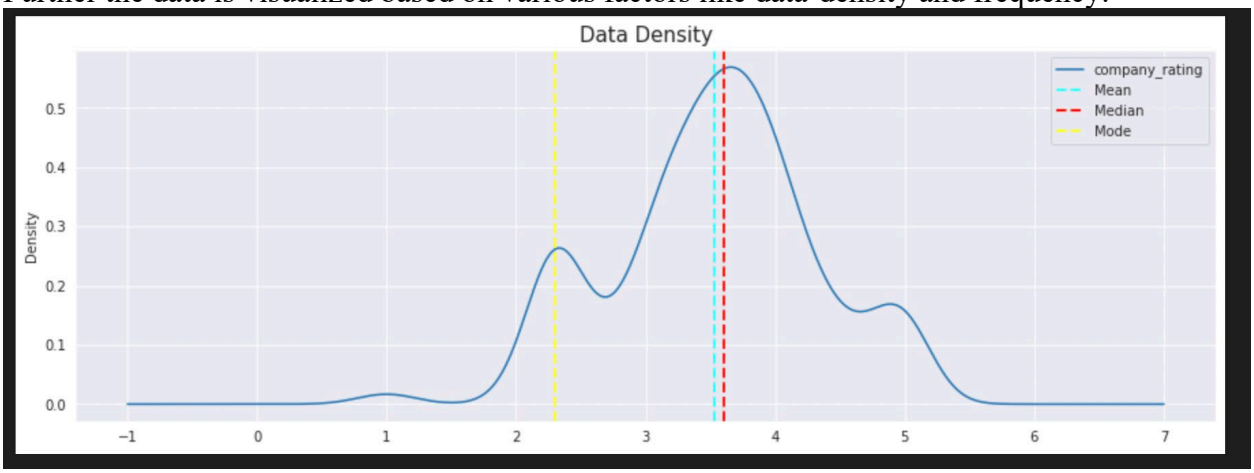
Figure 1: ER Diagram

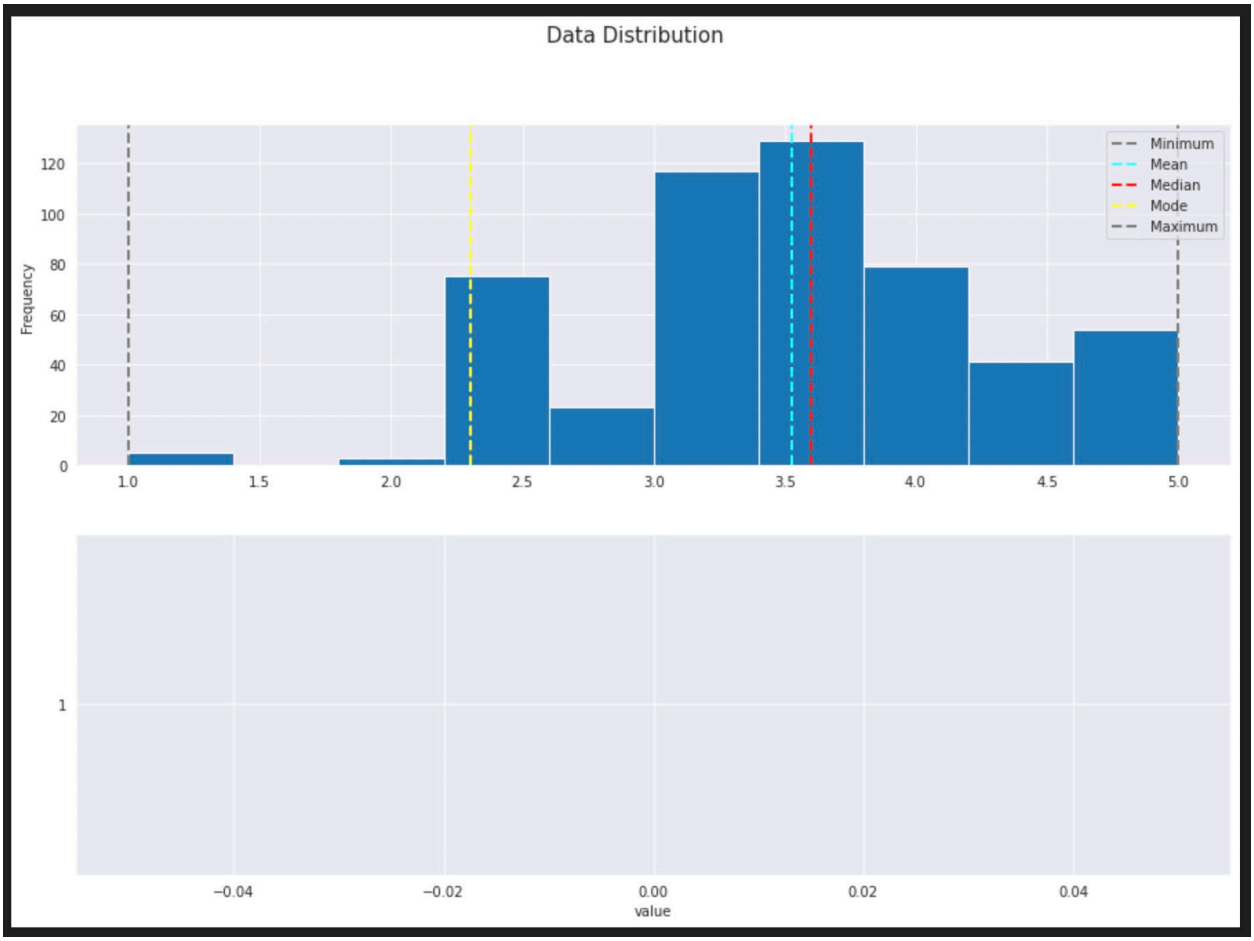
Sources of data:

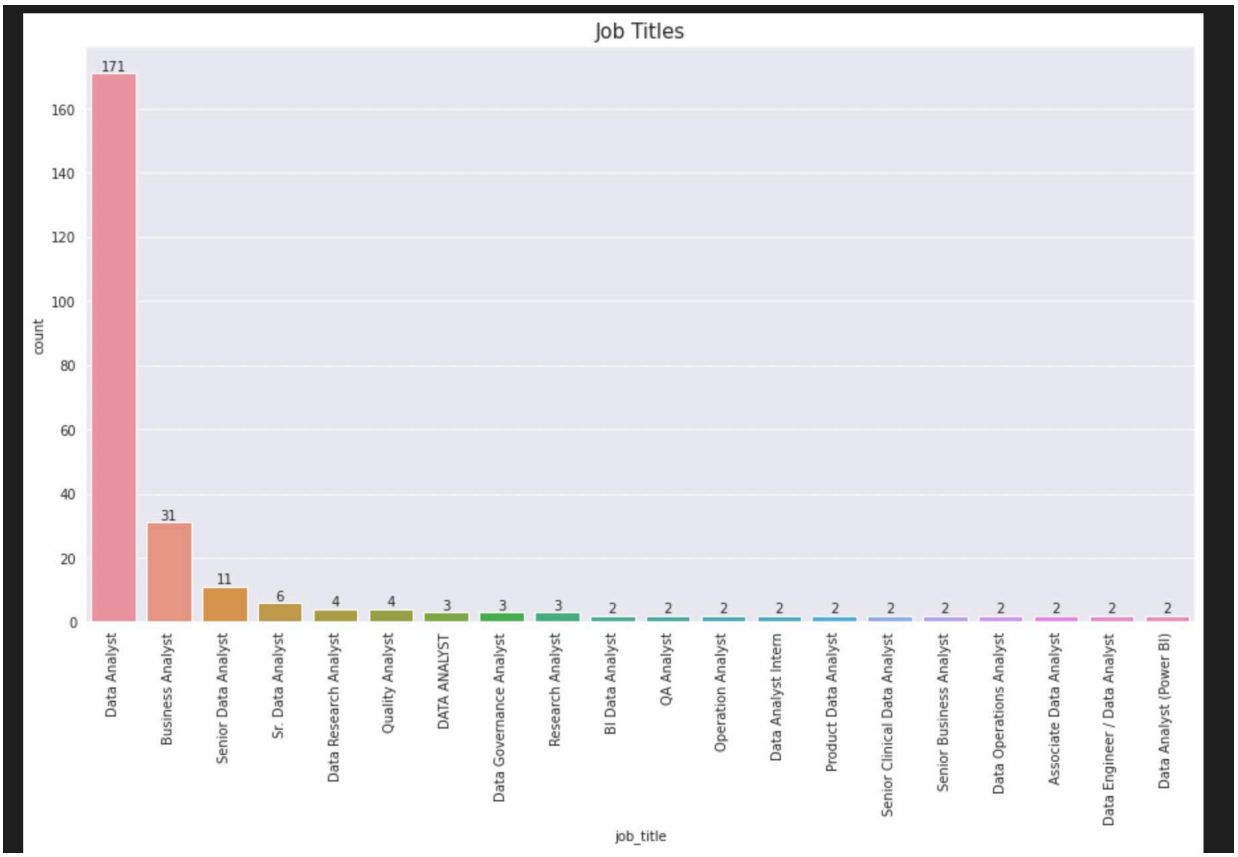
To obtain relevant and high-quality data we choose to scrape data from websites like LinkedIn, Twitter, Indeed, Monster, etc.
The data gathered from these websites is scraped using python libraries like BeautifulSoup and Selenium.
Also, data used in the database is collected and downloaded from data repositories and websites like ‘kaggle.com’.

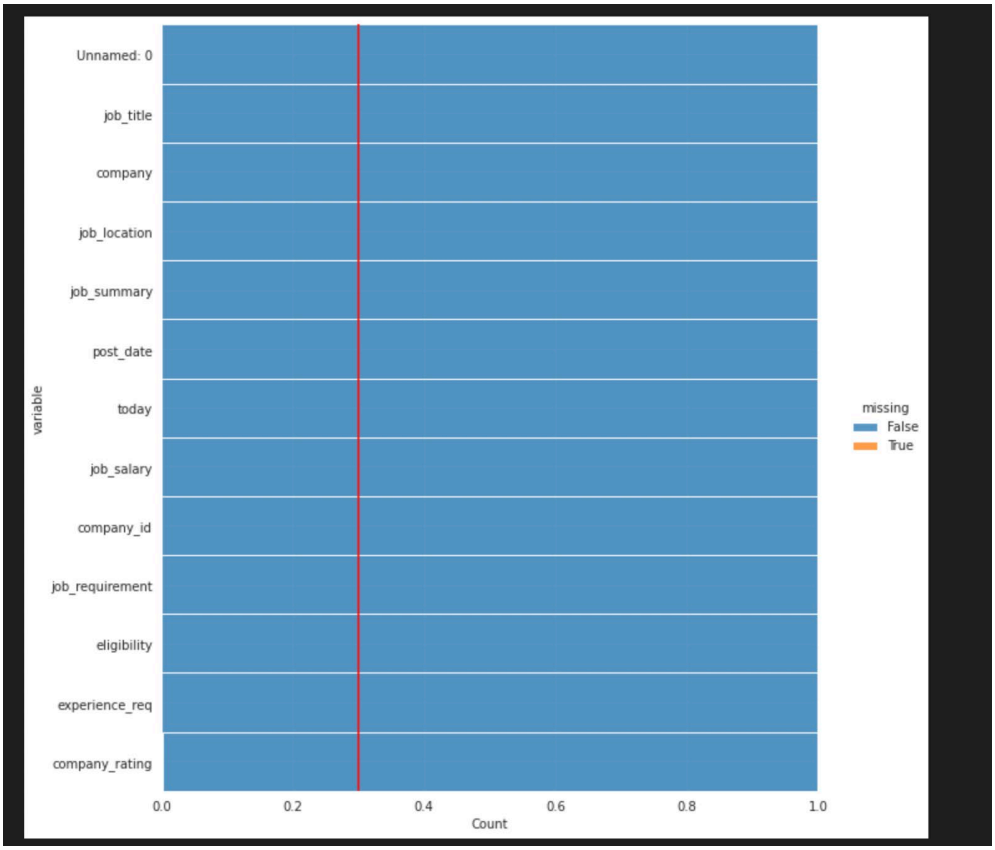
Data Visualization:

Further the data is visualized based on various factors like data-density and frequency:






Job Titles
count
job_title
171
31
11
6
4
4
3
3
3
2
2
2
2
2
2
2
2
2
2
2
Data Analyst
Business Analyst
Senior Data Analyst
Sr. Data Analyst
Data Research Analyst
Quality Analyst
DATA ANALYST
Data Governance Analyst
Research Analyst
BI Data Analyst
QA Analyst
Operation Analyst
Data Analyst Intern
Product Data Analyst
Senior Clinical Data Analyst
Senior Business Analyst
Data Operations Analyst
Associate Data Analyst
Data Engineer / Data Analyst
Data Analyst (Power BI)


variable
Unnamed: 0
job_title
company
job_location
job_summary
post_date
today
job_salary
company_id
job_requirement
eligibility
experience_req
company_rating
missing
False
True
0.0
0.2
0.4
0.6
0.8
1.0
Count

Data Validation:

The data gathered is checked for its validity which includes three factors:

Completeness: refers to the extent to which an entity contains the information needed to describe a real-world object. The presence of null values, which are typically regarded as missing values, in tables in relational database systems can be used to determine how full a table is.

Consistency: The degree to which a set of semantic rules are violated such as a specific data type, an interval for a numerical column, or a set of values for a categorical column.

Accuracy: The correctness of the data and can be measured in two dimensions: syntactic and semantic. Semantic accuracy contrasts a value with its actual representation, while syntactic accuracy compares a value's representation with a domain of definition that corresponds.

Further

Use Cases:

Based on the data gathered from various sources and the goal of the database we created UseCases and wrote SQL queries and Relational Algebra for the same.

Below are some of the examples:

```
1. What is the salary for Data Analyst?
Query:
SELECT job_salary FROM job_post WHERE job_position='Data Analyst';

Relational_Algebra:
π job_salary
 σ job_position = "Data Analyst" job_post



2. list the companies with ratings greater than 4 ?
Query:
SELECT c.company_name,cr.ratings_overall FROM company_ratings cr inner join company c on cr.company_id = c.id WHERE cr.ratings_overall > 4;

Relational_Algebra:
π c . company_name, cr . ratings_overall
 σ cr . ratings_overall > 4
  (ρ cr company_ratings ⋈ cr . company_id = c . id
   ρ c company)



3. What are the salaries offered in Boston?
Query:
SELECT job_salary FROM job_post WHERE job_location ='Boston';

Relational_Algebra:
π job_salary
 σ job_location = "Boston" job_post



4. What is the salary for Data Analyst?
Query:
select job_salary from job_post where job_position='Data analyst';

Relational_Algebra:
π job_salary
 σ job_position = "Data analyst" job_post
```

Figure 2: Use Cases Examples

Data Normalization:

The data which is already cleaned, munged, and validated is further checked to see if it is in a normalized form and satisfies the conditions for 1NF, 2NF and 3NF.
If not, we made changes in those tables to convert them into a normalized format.

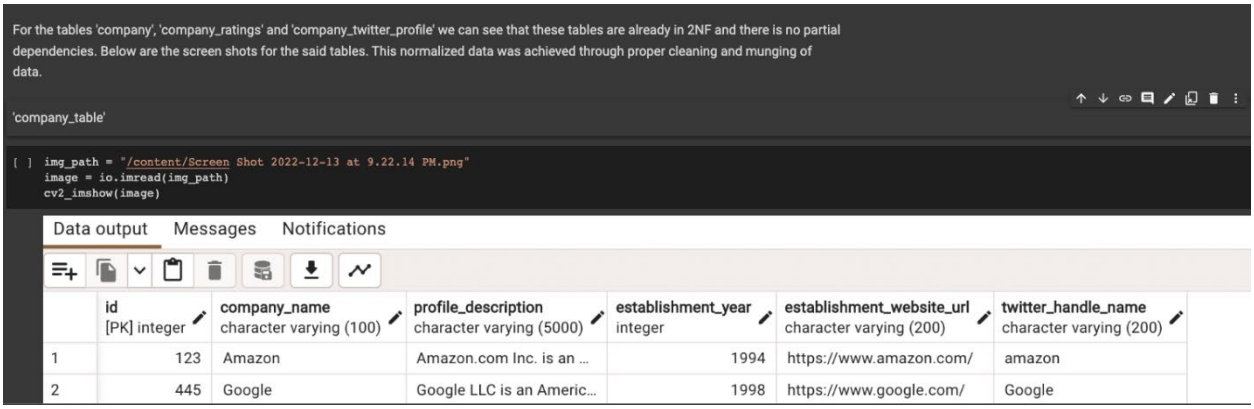


Figure 3: Normalization