DEVELOPMENT AND ANALYSIS OF US DATA SCIENCE JOBS DATABASE

GROUP 4

Harini Suram
Eshwari Vaishnavi Pampana
Venkata Mani Babu Karri
Revathi Duggineni

Guided by
Prof. Amir Manzour

Abstract:

In today's world data is everywhere. Over the past few years, the field of Data Science advanced so much, and more and more people are getting interested in starting their career in this awesome field. But for someone who is new to this field it might be difficult to get an overview about job roles, skills required, companies hiring and pay scale. The main motivation of this project is to help people who are interested to start their career in Data Science by providing all the required information in our database to get an overview about the job market based upon which they can start preparing themselves. The primary goal of our project is to develop a relational database about data science job market in US. We trying to include all the necessary information like companies offering jobs, pay scale, job roles and its location. Based on our goal, we collected required dataset and did necessary modifications to make our data clean. In order to reduce the complexity and easy extraction of information we categorized the data into multiple tables and a relation database that can relate multiple entries. We populated our database based on the finalized columns using Apify glassdoor web scraping tool. After populating our dataset to the required number of records. We finalized our tables and established relations between tables using ERD diagram. Then we imported the data into phpMyAdmin and wrote SQL queries for some of our research question and some interesting observations which are discussed below.

Background:

Unemployment is one of the most critical challenges in many countries. A lack of awareness of what employer’s demand from their employees is one of the fundamental causes of the high unemployment rate a skill gap results from a lack of understanding of what companies expect from their employees. People with a thorough knowledge of the job market might boost their career prospects significantly by properly preparing themselves in advance. This is also an excellent approach to reduce unemployment. The job market is as vast as an ocean. It isn't easy to cover every field in a single database. We chose data science as our field of interest because it is booming and gaining popularity these days and attracting an increasing number of people. The database we are constructing can inspire and guide people in other areas to create databases in their fields. Our objective is to encourage individuals to establish job market databases in other professions, like what we're doing with data science. This can reduce the skill gap, thereby reducing the unemployment rate. Some may claim that we can obtain all our information through job portals or other sources. However, the difficulty with job portals is that sifting through hundreds of positions is tough. Creating a database in this manner will make the process much easier. There is a great deal of information available online about job openings in data science. However, many of these sites leave out crucial information, such as the standards for a given company. Creating a database in which we can store all the data will aid us in fixing this challenge.

Related work:

Unemployment has been a big issue for many countries. Research has been done to find out the reason for the increase in the unemployment rate. Research by Pushova, L., Randjelovic, M., Jankovic, M., & Blazekovikj, M. (n.d.) found the skill gap as one of the reasons for the increase in unemployment. Harvard Business Review published a paper about how data science can change the future and Brynjolfsson, A. M. A. and E. (2012) termed 'Data Scientist' as 'The Sexiest Job of the 21st Century'. From this we can say that Data Science is going to be the next big thing in the Tech industry. There are so many datasets that talk about data science jobs. On Kaggle, there is a dataset labeled "Data Scientist Salary" that comprises information such as the minimum salary, maximum salary, average salary, job description, age of company in years, skill set necessary, and so on by Bhathi, N. (2021). The issue with this dataset, however, is that it does not cover a wide range of firms and job positions in data science and its related fields. This dataset is not very useful for the ones who are targeting the well-established firms like Meta, Google. Several articles and journals published information regarding the opportunities in data science like one by University of San Diego. (2021). All these articles and journals might be useful for many people. But most of them don't have key information. We can take it a step forward and place all the necessary information in one place. After a lot of research by Ganapati, S., & Ritchie, T. S. (2021) concluded that by reducing the skill gap we can reduce the unemployment rate to some extent, and we are trying to do so by building the database. We decided to design an interactive web interface for our database by which people extract data from the database even if they don't have proper knowledge about SQL which in turn will make this initiative to a larger group of people Duckett, J. (2011). There is an article about data science jobs and their salaries published by Towards Data Science(tds). (2019).

Methods:

We used tools like Kaggle, Apify Glassdoor Web Scraper, Excel, phpMyAdmin & Tableau. On Kaggle, we discovered a dataset that was relevant to our objective and contained most of the attributes we required. We generated our dataset using Apify Glassdoor Web Scraping tool by pulling data from the Glassdoor employment portal based on these attributes. Using Excel we performed data cleaning and finalized our dataset. Further, our we divided our dataset into five different tables and established the relations between these tables using Entity Relation Diagram(ERD). Created our database and gave all necessary constraints using phpMyAdmin and also retrieved required data by writing queries. Finally, made some interesting visualizations using Tableau to get some useful insights.

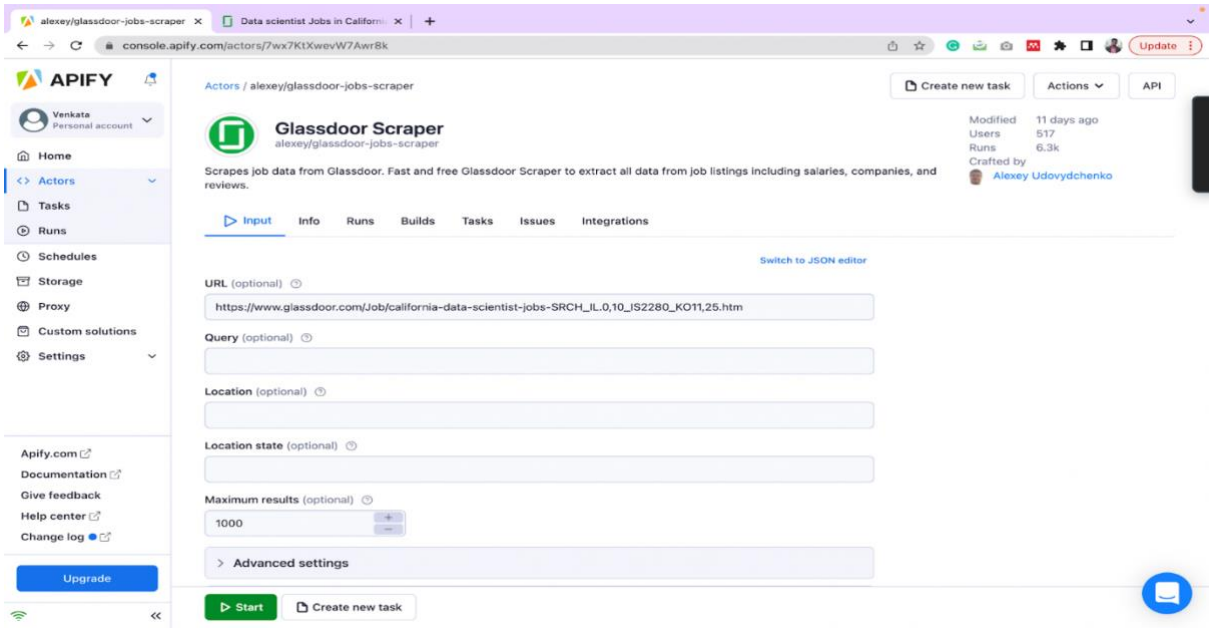


Fig:1 – Apify Glassdoor Scraper used for data collection.

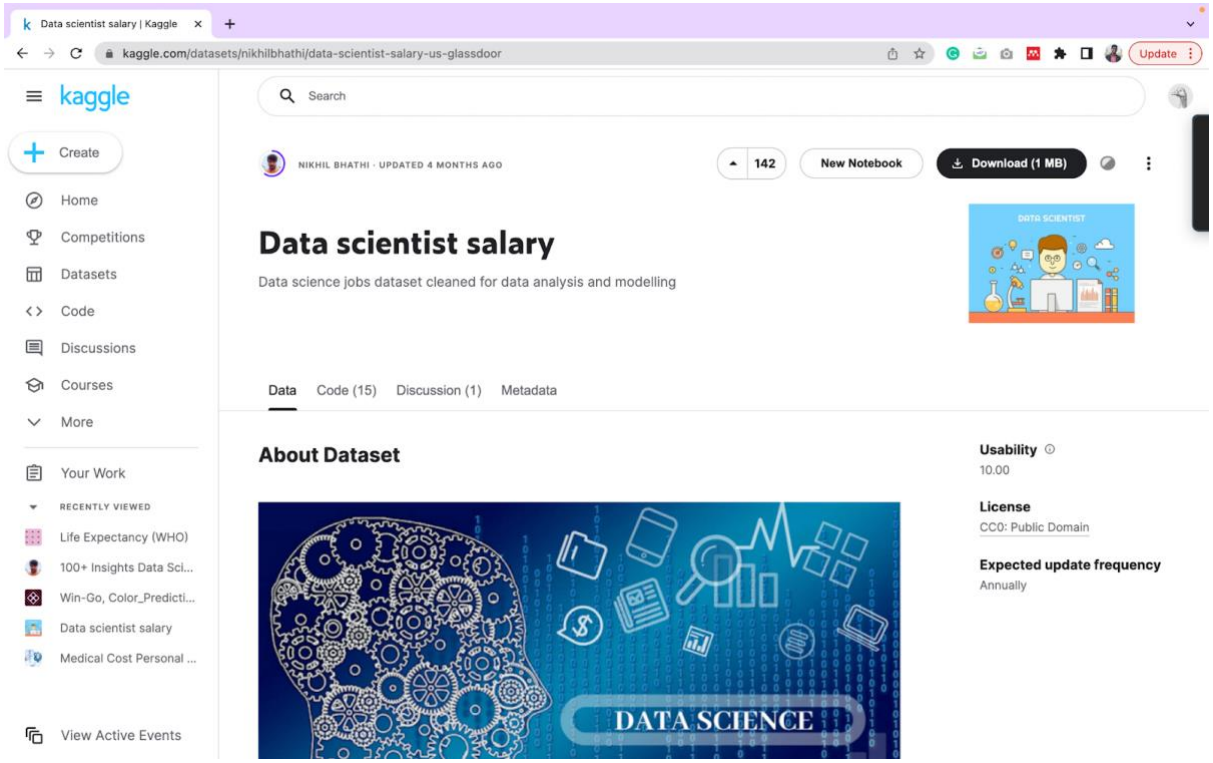


Fig:2 - Kaggle Dataset

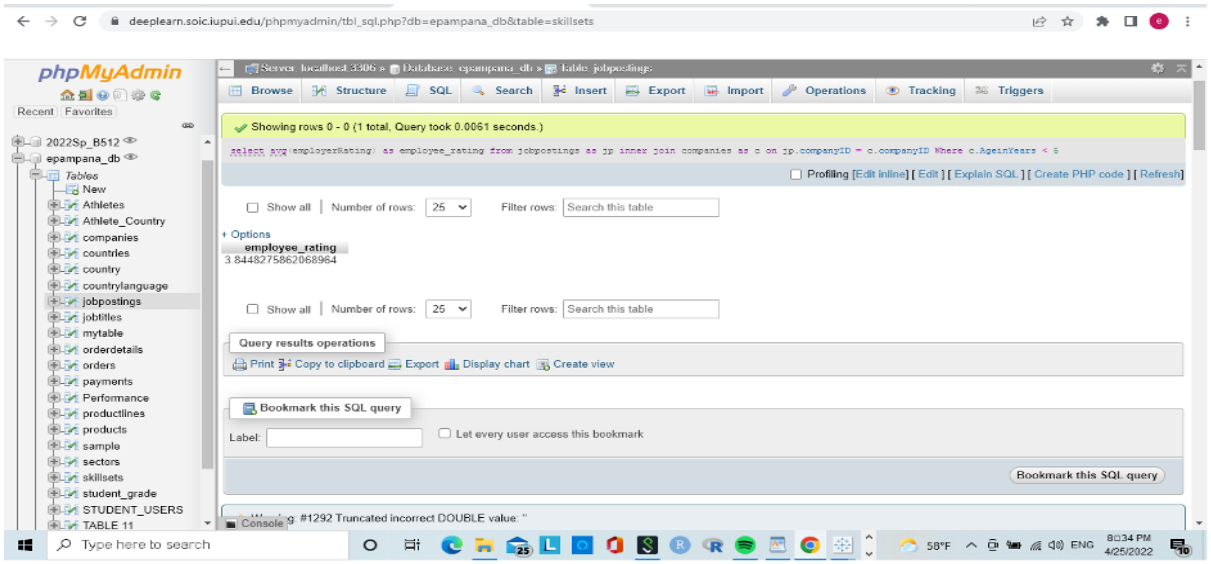


Fig:3-phpMyAdmin

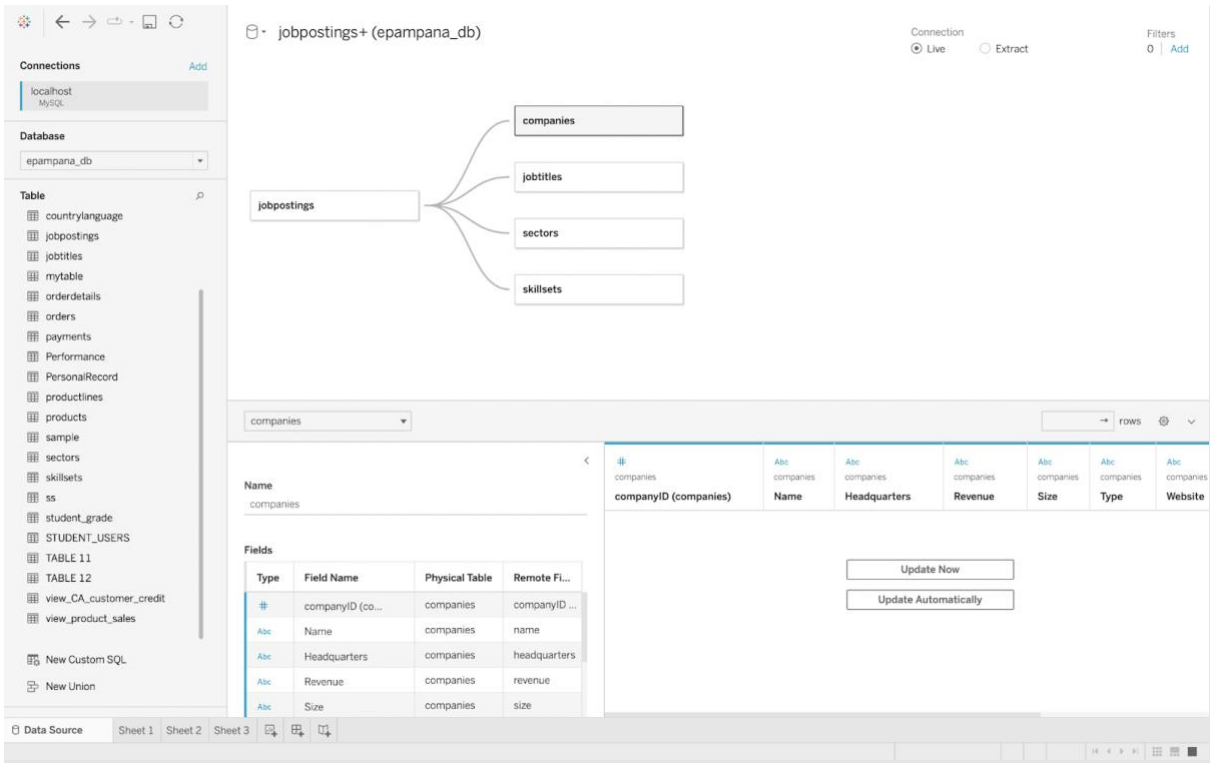


Fig:4 - Connecting phpMyAdmin database with Tableau

ENTITY RELATIONSHIP DIAGRAM

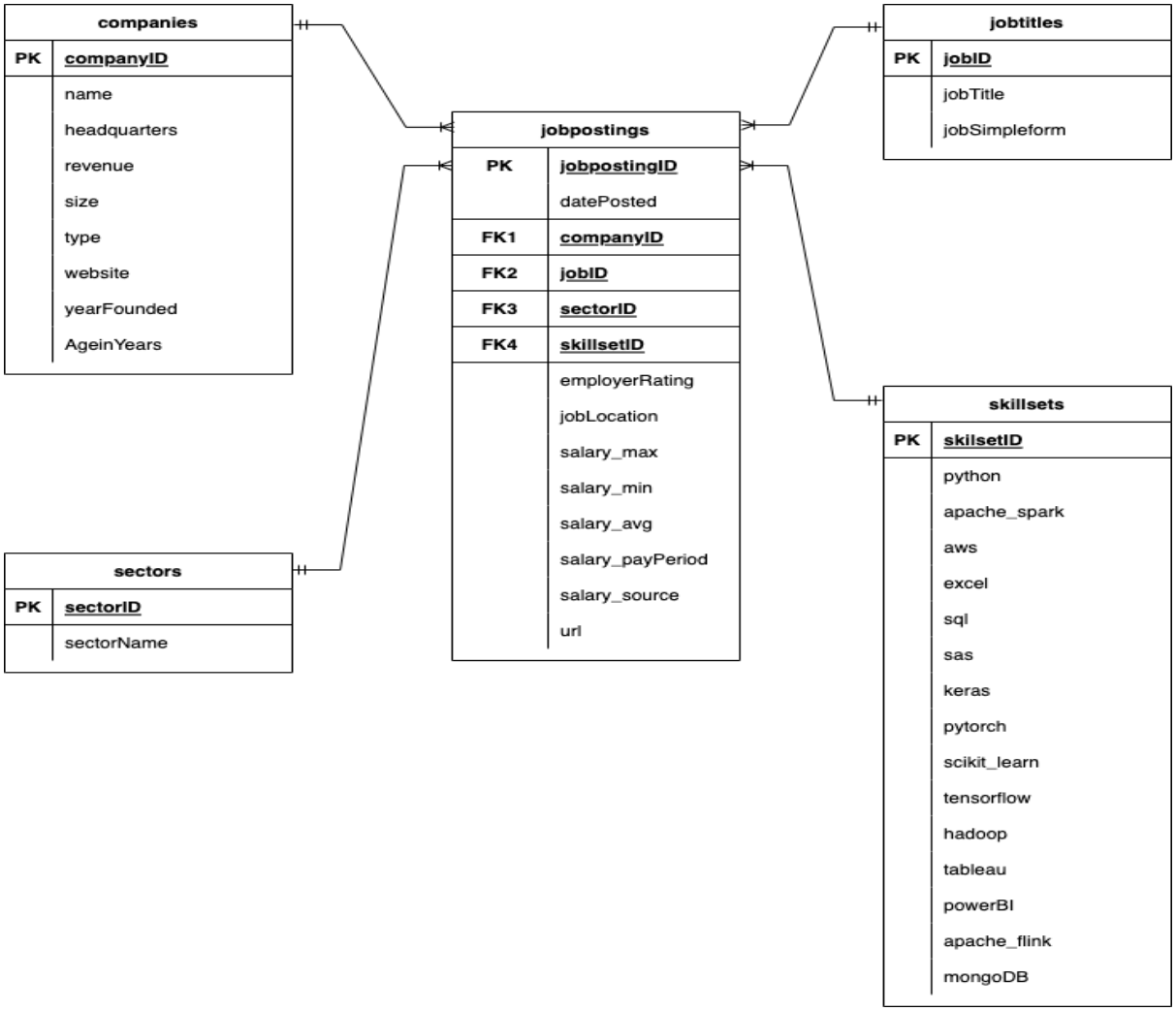


Fig:5 - ERD

Breakdown of ERD

In this Entity Relationship Diagram we have a total of 5 tables where Job postings serves as the parent table.

The relationship between the Companies table and the jobpostings is one or many, whereas the relationship between jobpostings and companies is one and only.

The relationship between the Sectors table and the jobpostings is one or many, whereas the relationship between jobpostings and sectors is one and only.

The relationship between the job titles table and the jobpostings is one or many, whereas the relationship between jobpostings and jobtitles is one and only.

The relationship between the skillsets table and the jobpostings is one or many, whereas the relationship between jobpostings and skillsets is one and only.

Results:

1. I like to work in Texas, what are the companies that are offering data science jobs in Texas?

select distinct c.name from companies as c inner join jobpostings as jp on c.companyID = jp.companyID where jp.jobLocation like "%TX"

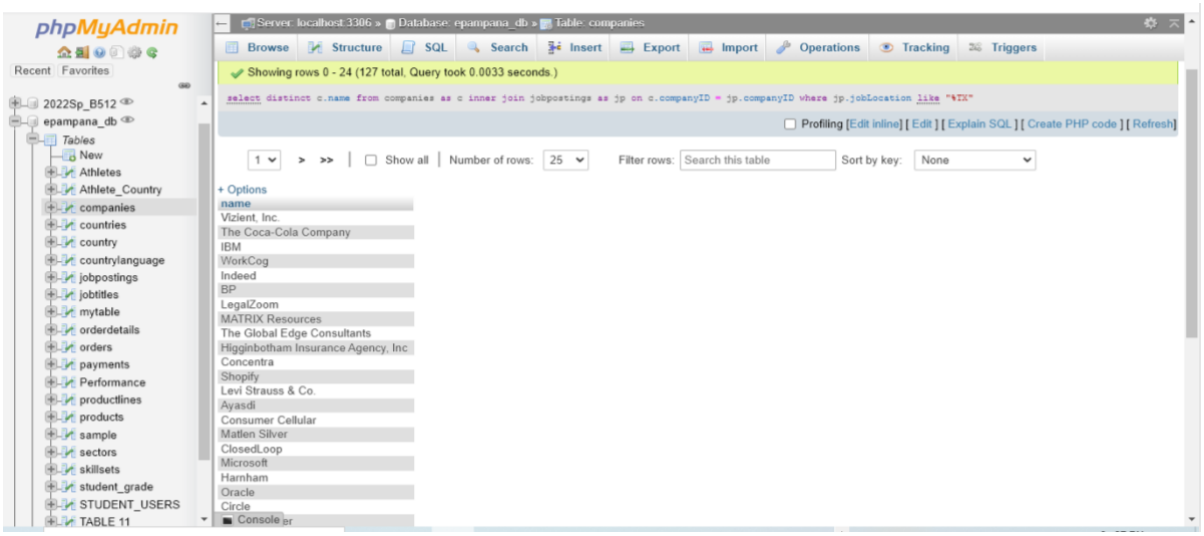


Fig:6 - Query 1

2. What are skills required to work as a Data Scientist at Adobe?

select s.* from skillsets as s inner join jobpostings as jp on s.skillsetID = jp.skillsetID inner join jobtitles as jt on jp.jobID = jt.jobID inner join companies as c on jp.companyID = c.companyID where c.name = "Adobe" and jt.jobSimpleform = "Data Scientist" limit 1

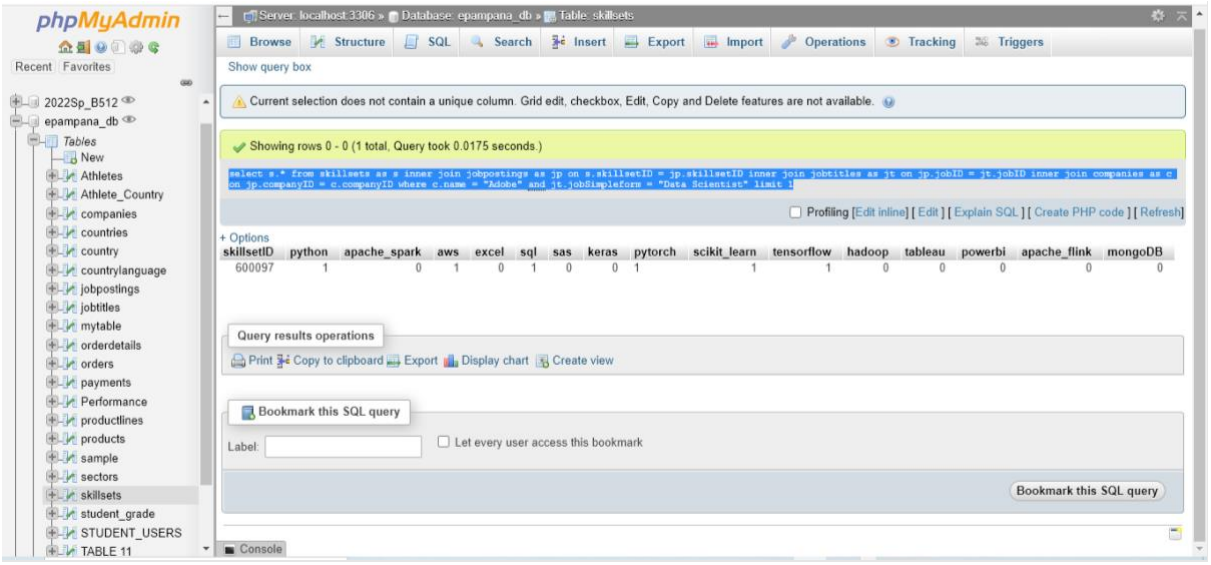


Fig:7- Query 2

3. Is it better to work for start-ups or well-established companies?

3a select avg(employerRating) as employee_rating from jobpostings as jp inner join companies as c on jp.companyID = c.companyID Where c.AgeinYears < 5

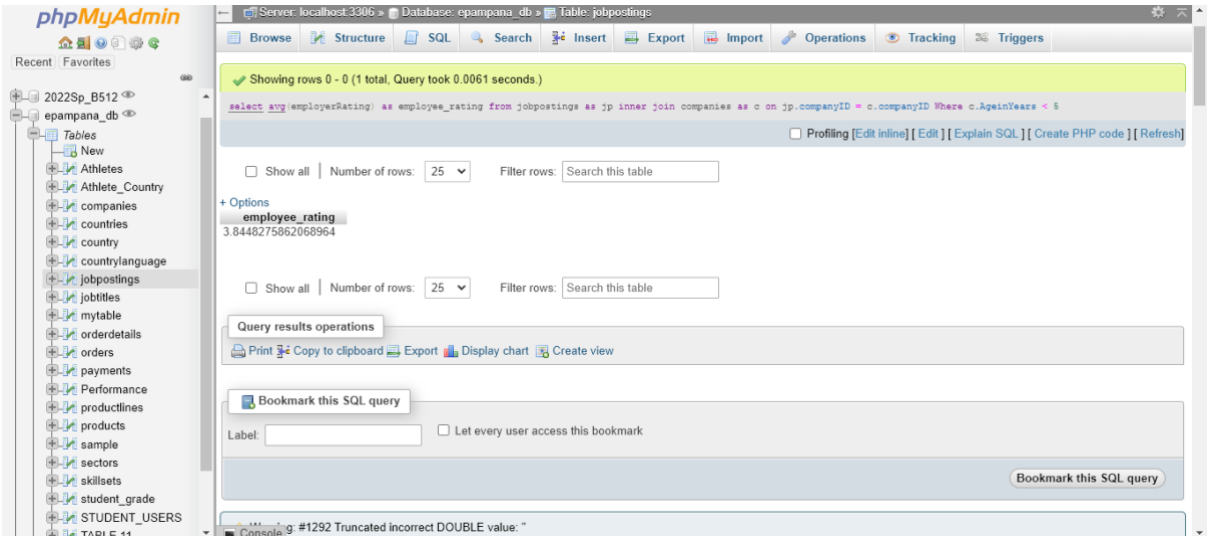


Fig: 8 - Query 3

3b select avg(employerRating) as employee_rating from jobpostings as jp inner join companies as c on jp.companyID = c.companyID Where c.AgeinYears >= 5

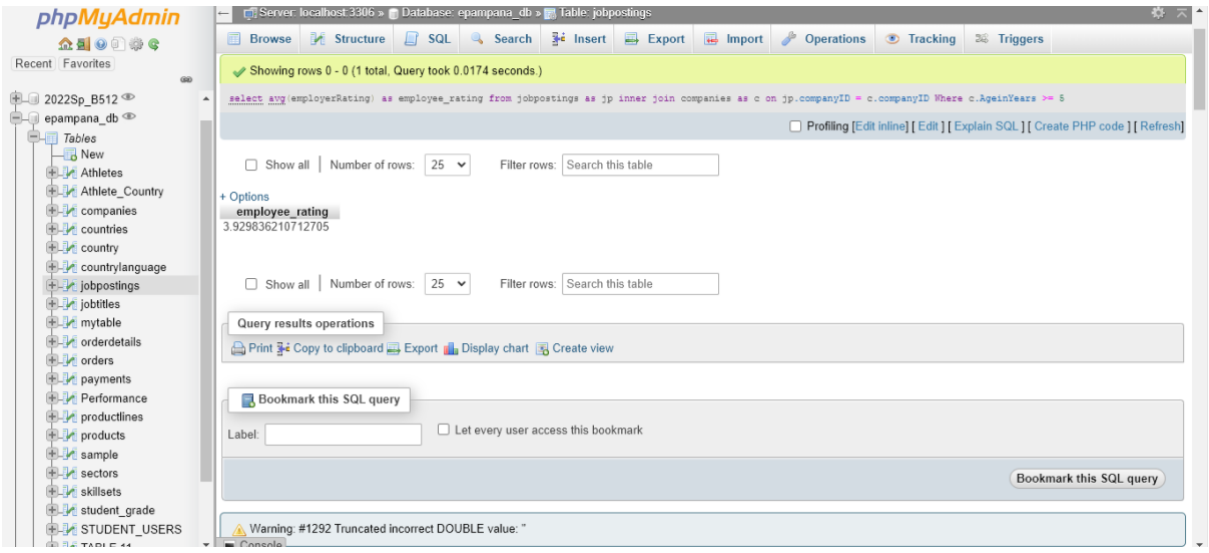


Fig: 9 – Query 4

4. Which companies are paying more for data science role like relatively newer companies or older companies?

4a select avg(jp.salary_avg) from jobpostings as jp inner join companies as c on jp.companyID = c.companyID where c.AgeinYears < 5

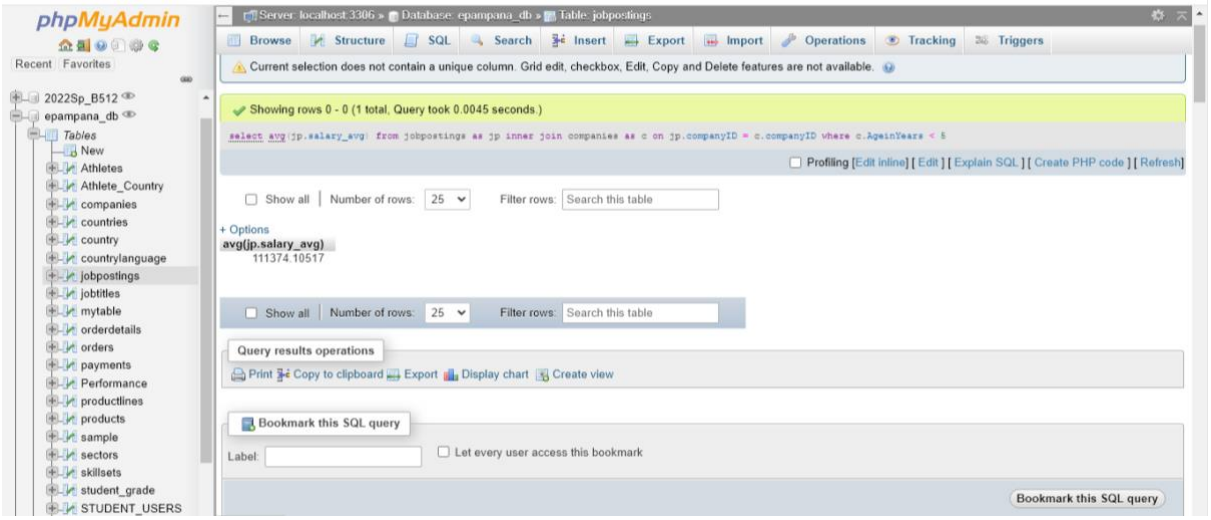


Fig:10 – Query 5

4b select avg(jp.salary_avg) from jobpostings as jp inner join companies as c on jp.companyID = c.companyID where c.AgeinYears >= 5

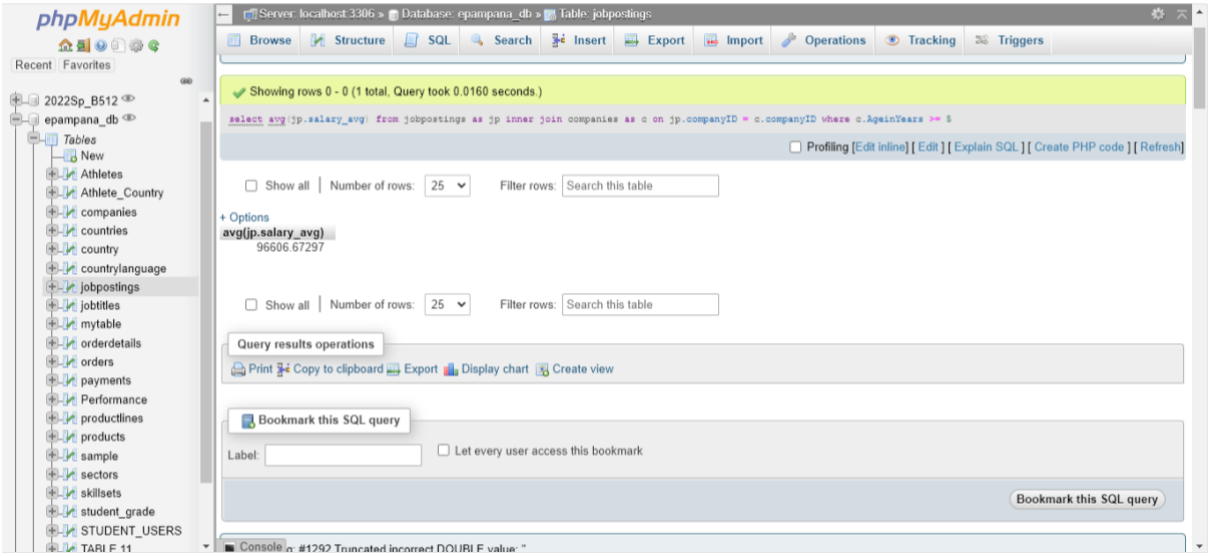


Fig:11 – Query 6

5. What top 3 required skills to work as data scientist in the US?

select sum(ss.python),sum(ss.apache_spark),sum(ss.aws),sum(ss.excel),sum(ss.sql),sum(ss.sas),sum(ss.keras),sum(ss.pytorch),sum(ss.scikit_learn),sum(ss.tensorflow),sum(ss.hadoop),sum(ss.tableau),sum(ss.powerBI),sum(ss.apache_flink),sum(ss.mongoDB) from skillsets as ss inner join jobpostings as jp on ss.skillsetID = jp.skillsetID inner join jobtitles as jt on jp.jobID = jt.jobID where jt.jobSimpleform = "Data Scientist"

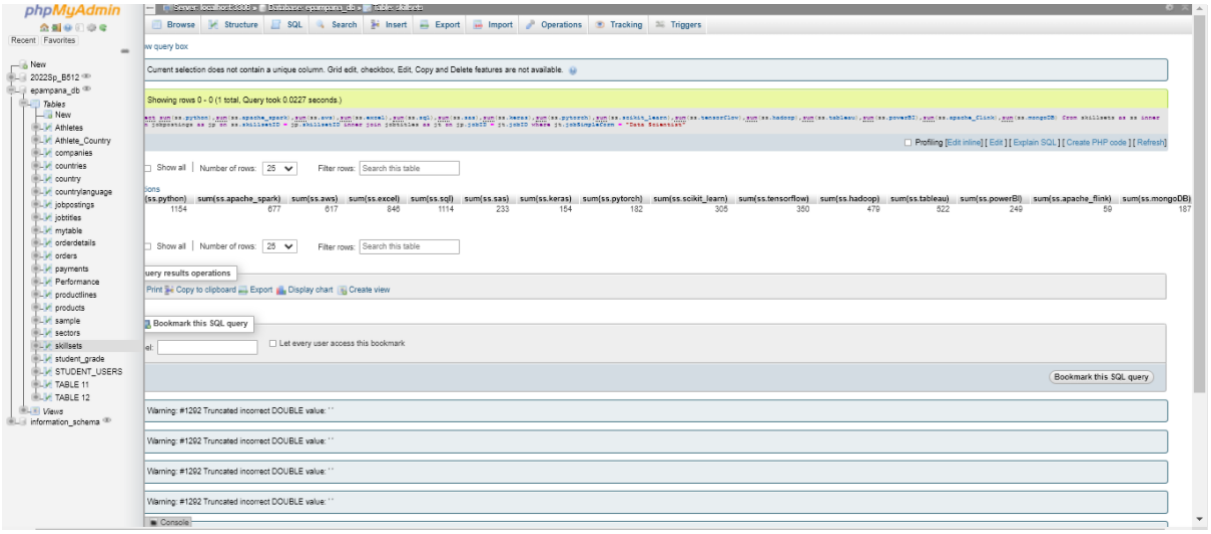


Fig:12 – Query 7

DATA VISUALIZATIONS

We used tableau to visualize our data. We connected the phpMyAdmin with the Tableau using MySQL Connector.

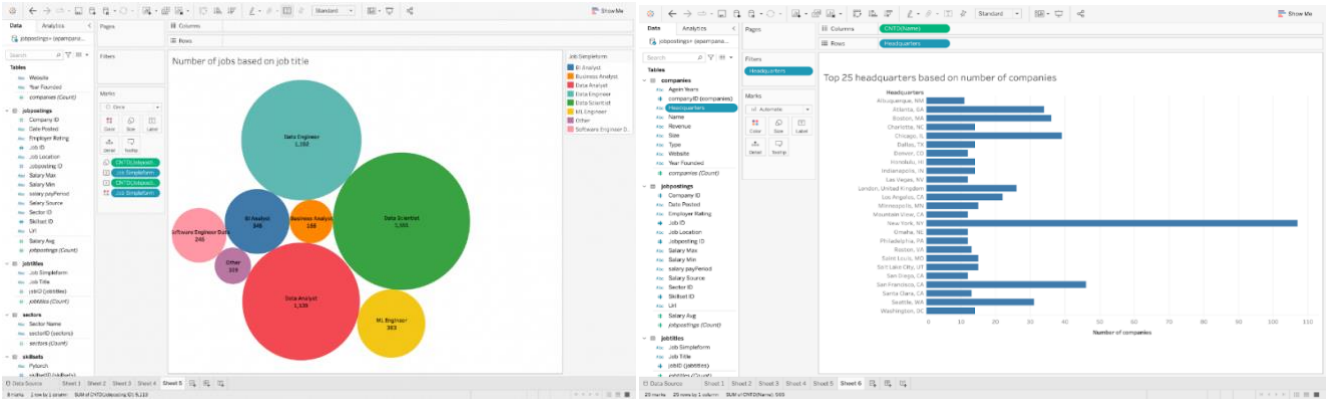


Fig: 13 & 14

We created a visual representation of our dataset using a bubble chart based on the number job postings for different job roles.

The top 25 headquarters locations were shown using a horizontal bar chart depending on the number of companies in each location. We can see that the majority of the companies have their headquarters in New York.

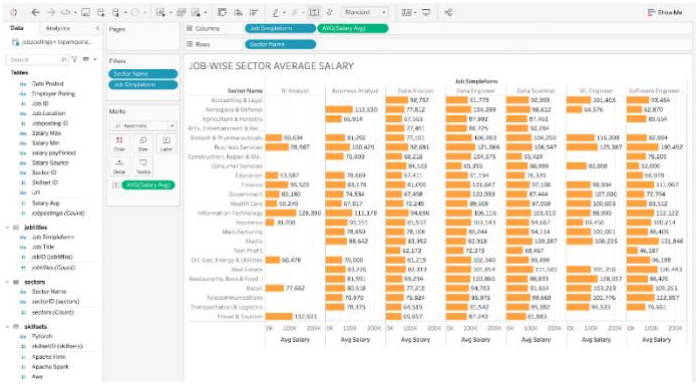


Fig: 15

The average salary in each sector for various job roles is represented in Fig:15. The graph reveals that the business services and information technology industries have the highest average incomes across all job roles.

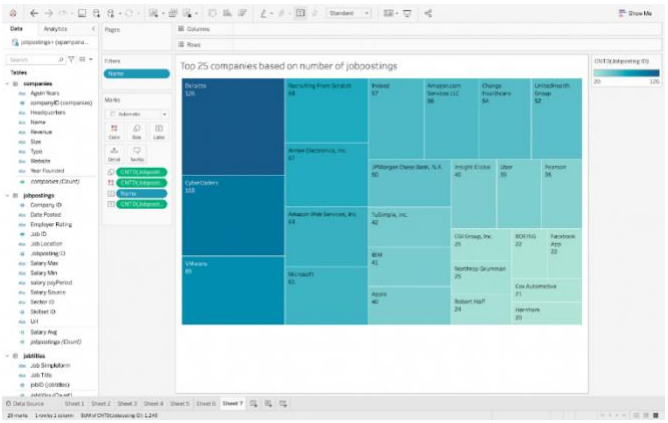


Fig: 16

We used a tree map in tableau to represent the top 25 companies based on the number of job postings in each of them. We can see from the graph above that Deloitte appears to have the most job posts, while Harnham has the least.

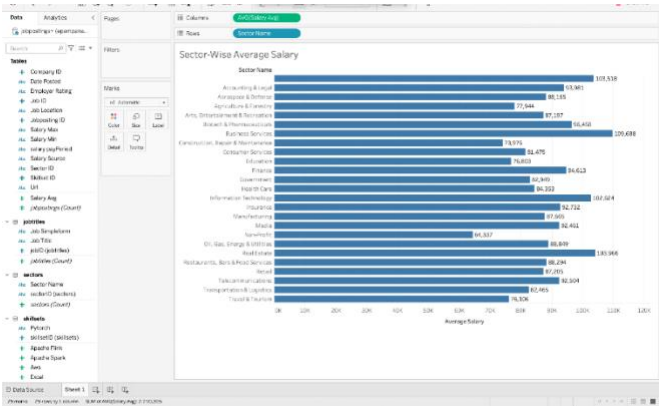


Fig:17

We created a horizontal bar chart based on the average salary in each sector in Fig17. According to the above graph, the business services sector has the highest average salary, while the non-profit organization sector has the lowest.

Discussion:

Special Observations: We observed that the sector, Business services, had the highest average salaries and Non-Profit organizations had the lowest average salaries. Data scientists have the

most number of jobs while the BI analysts have the lowest number of jobs available. Amazon Web services had the highest number of jobs. Software Data engineers had the highest salaries and data analysts had the lowest salaries.

Challenges:

Data collection: Dataset from Kaggle had limited data and was not sufficient to meet our objectives. Therefore, we decided on populating our own database. Data population: Using Apify, we could retrieve only 30 records at once, so we had to repeat this process several times to populate our huge database. Data cleaning: Since our data contained so many null values and duplicates, we had to delete those records and populate more data to meet our requirement. Initially, we intended to use the dataset we acquired from kaggle.We thoroughly reviewed and understood the data and decided on the tables and created the ERD. We then began cleaning the data, removing all null values and other irrelevant information, only to realize that the remaining information was extremely limited and insufficient to meet our objectives. We then decided to create our own database utilizing data from Glassdoor's job posts (https://www.glassdoor.com/member/home/index.htm). We did extensive research and discovered that web scraping is an efficient way of extracting data from the Glassdoor website and populating our database. We used Apify to scrape data from the Glassdoor website (https://console.apify.com/actors/7wx7KtXwevW7Awr8k/?addFromActorId=7wx7KtXwevW7Awr8k#/console). We could only retrieve 30 records at a time, therefore we had to repeat the process numerous times to populate our database. We had to conduct a lot of data cleansing this time because the data from Glassdoor included a lot of nulls and duplicates. The process of populating our database was a very time-consuming and difficult task. Also, importing the data was another challenge because it was a huge dataset that we created.

Conclusion:

In order to tackle the challenges we faced in our project, we created our own database for data science jobs in the US. Following the creation of our database, we generated approximately 30 questions that a data science aspirant would like to know , and we were able to answer all of these research questions by writing queries, indicating that our database is functioning properly and is capable of answering the majority of the questions we have, which is our primary project goal. We were able to successfully connect our database to Tableau and build some simple visualizations using the data we had. We conducted a survey where students were asked to determine the most efficient way to obtain the necessary information about data science jobs, and the majority of respondents stated that using the database made the process much simpler and easier than using other sources like job portals or articles. This implies that our database did a great job at fulfilling our objective to give the data science aspirants a good overview of the job market. Since we have done this process of gathering and processing data in one specific sector, that is data science, we strongly believe that we can do it in other fields as well. As a result, if this is properly implemented, more people will have access to job data in every field in a much better and simpler way, enabling them to gain a better understanding and knowledge about the jobs in their field. This will allow users to better understand the employer's requirements and expectations, allowing them to prepare well in advance, reducing the skill gap between applicants and companies, and ultimately reducing unemployment.

References:

Bhathi, N. (2021). Data scientist salary. Kaggle. https://www.kaggle.com/nikhilbhathi/data-scientist-salary-us-glassdoor

Data scientist salaries [by industry, region, Education & More]. University of San Diego. (2021). https://onlinedegrees.sandiego.edu/data-scientist-salary-guide/

Data scientist salary -the Ultimate Guide for 2021. ProjectPro. (n.d.). https://www.projectpro.io/article/data-scientist-salary-the-ultimate-guide-for-2021/218

Davenport,T.H., Patil,D.J.(2012). Data scientist: The sexiest job of the 21st Century. Harvard Business Review. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Duckett, J., 2011. HTML & CSS: design and build websites (Vol. 15). Indianapolis, IN: Wiley.https://ghnet.guelphhumber.ca/files/course_outlines/AHSS_3080_Thomas_Borzecki(05).pdf

Ganapati, S., & Ritchie, T. S. (2021). Professional development and career-preparedness experiences of STEM Ph.D. students: Gaps and avenues for improvement. PloS one, 16(12), e0260328. https://doi.org/10.1371/journal.pone.0260328

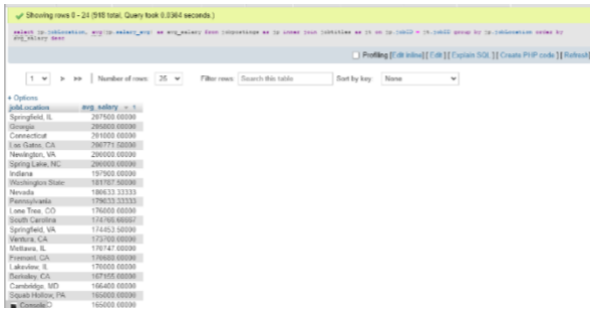
Levine, M. V. (2013). The skills gap and unemployment in Wisconsin: Separating fact from fiction. https://dc.uwm.edu/cgi/viewcontent.cgi?article=1017&context=ced_pubs

Pushova, L., Randjelovic, M., Jankovic, M., & Blazekovikj, M. (n.d.). Youth unemployment in Europe: Causes and solutions. KNOWLEDGE - International Journal. http://www.ikm.mk/ojs/index.php/kij/article/view/4371

Roy, A. (2016). Data Science - a career option for 21st century: Job prospect in data science. ResearchGate.https://www.researchgate.net/publication/299353360_Data_Science_-_A_Career_Option_for_21st_Century_Job_Prospect_in_Data_Science

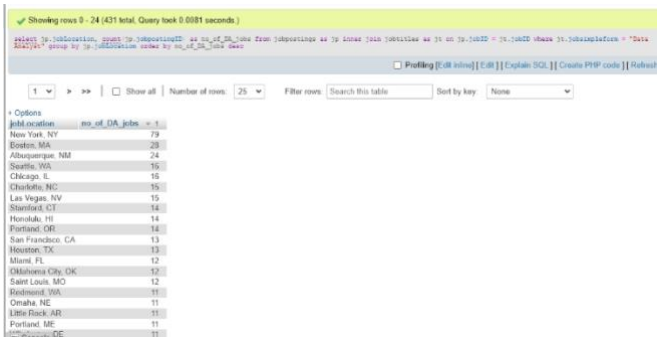
Appendix

1. Which state pays higher salaries for Data Engineer?



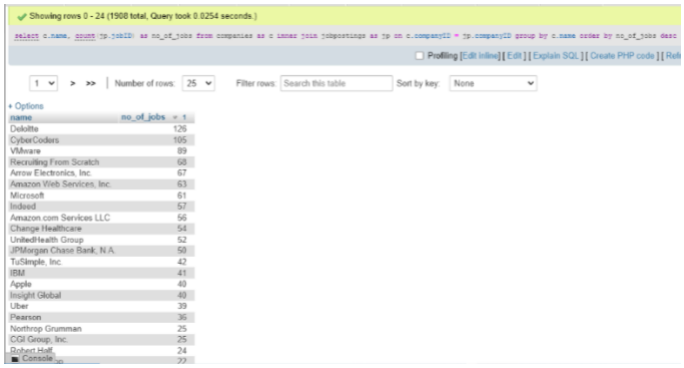
```
select jp.jobLocation, avg(jp.salary_avg) as avg_salary
from jobpostings as jp inner join jobtitles as jt on
jp.jobID = jt.jobID group by jp.jobLocation order by
avg_salary desc
```

2. Which place has more Data Analyst jobs with average salary greater than 75000?




```
select jp.jobLocation, count(jp.jobpostingID) as no_of_DA_jobs from jobpostings as jp inner join jobtitles as jt on jp.jobID = jt.jobID where jt.jobsimpleform = "Data Analyst" group by jp.jobLocation order by no_of_DA_jobs desc
```

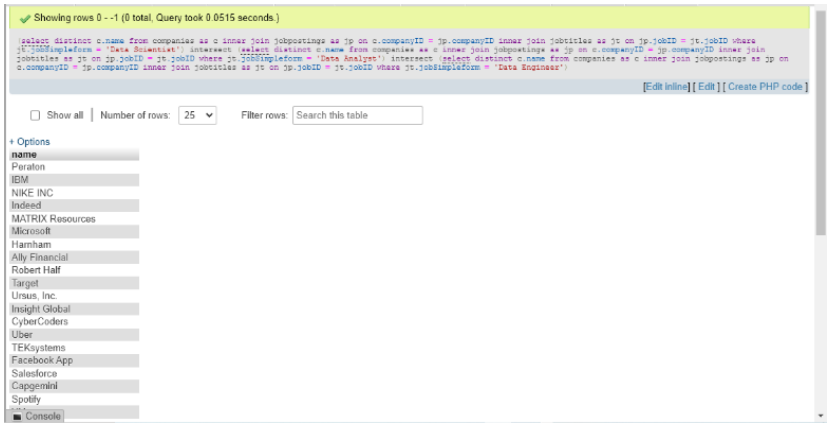
3. Name the company that is offering the greater number of data science jobs?



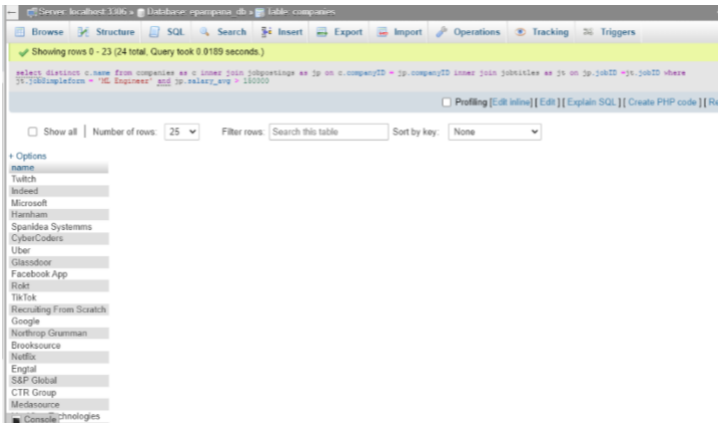

```
select c.name, count(jp.jobID) as no_of_jobs from companies as c inner join jobpostings as jp on c.companyID = jp.companyID group by c.name order by no_of_jobs desc;
```

4. List the companies that are offering Data Scientist, Data Analyst and Data Engineer jobs?

(select distinct c.name from companies as c inner join jobpostings as jp on c.companyID = jp.companyID inner join jobtitles as jt on jp.jobID = jt.jobID where jt.jobSimpleform = 'Data Scientist') intersect (select distinct c.name from companies as c inner join jobpostings as jp on c.companyID = jp.companyID inner join jobtitles as jt on jp.jobID = jt.jobID where jt.jobSimpleform = 'Data Analyst') intersect (select distinct c.name from companies as c inner join jobpostings as jp on c.companyID = jp.companyID inner join jobtitles as jt on jp.jobID = jt.jobID where jt.jobSimpleform = 'Data Engineer')



5. I want to become a ML Engineer and want the list of companies that are paying more than 150000 for ML Engineers?




```
select distinct c.name from companies as c
inner join jobpostings as jp on
c.companyID = jp.companyID inner join
jobtitles as jt on jp.jobID =jt.jobID where
jt.jobSimpleform = 'ML Engineer' and
jp.salary_avg > 150000
```

Other queries and visualization are in the below link:

https://docs.google.com/document/d/1pNE_RE-U_A97jOPBG6Ezc55Q0bs2on0fFeYmT_ONe8s/edit