Job Dataset Summary

The dataset focuses on job opportunities within the field of data science, encompassing various job roles and their corresponding job descriptions. It offers insights into the diverse range of positions available, providing valuable information for those interested in pursuing a career in this field.

In the Excel sheet, I have documented the progress and steps I have taken. The dataset includes various columns such as Job Title, Salary Estimate, Rating, Headquarters, Size, Founded, Type of Ownership, Industry, Sector, Revenue, Competitors, Job Description, Company Name, and Location.

Before proceeding with data cleaning, I took an overview of the dataset to familiarize myself with its contents.

The cleaning procedure involved the following steps:

1. Initially, I checked for any missing values in the Job Title and Salary Estimate columns, and fortunately, there were no missing values.
2. Upon further inspection, I discovered an unexpected value of "-1" appearing frequently in several columns, rendering them irrelevant. Unfortunately, these values differ across columns, making complete elimination impossible.
3. To begin cleaning, I started by removing unnecessary words that served no purpose in their respective columns. For instance, I removed words like 'Company,' 'Glassdoor,' 'employees,' and '(USD)' from the Type of Ownership, Salary Estimate, Size, and Revenue columns respectively.
4. Next, I replaced instances of "Not Applicable" and "-1" in the Revenue column with 'na' to maintain consistency.
5. Additionally, I created a new column called 'Company' and extracted the company names from the 'Company Name' column using the formula =LEFT(O2, SEARCH(CHAR(10), O2) - 1).
6. I added a new column named 'Job State' and extracted the state information from the 'Location' column using the Right function. In cases where the state was not

mentioned, I kept those rows unchanged.

7. I have created the "Lower Salary" column and used the following formula to extract the lower value: "=LEFT(RIGHT(B2,LEN(B2)-1),FIND("K",RIGHT(B2,LEN(B2)-1))-1)*1000".
    - RIGHT(B2,LEN(B2)-1): This function extracts a substring from cell B2, starting from the rightmost character. The length of the substring is determined by subtracting 1 from the total length of B2. This is done to remove the leftmost character from the original value in B2.
    - FIND("K", RIGHT(B2,LEN(B2)-1)): This function finds the position of the letter "K" within the substring obtained in the previous step. It searches from left to right within the substring and returns the position of the first occurrence of "K." This position is used in the next step.
    - LEFT(RIGHT(B2,LEN(B2)-1), FIND("K", RIGHT(B2,LEN(B2)-1))-1): This function extracts a substring from the right side of B2. It starts from the leftmost character of the substring and continues until the position of the "K" found in the previous step, minus 1. Essentially, it captures the numeric part of the value without the "K" character.
    - The extracted substring is then multiplied by 1000 to convert it from thousands to the actual value. This multiplication by 1000 is denoted by "*1000" at the end of the formula.
8. I then created a new column named "HS" where I extracted the higher value. For example, from the cell value "$121K-$131K," I extracted "$131K" using the formula '= "$" & TRIM(RIGHT(B2, LEN(B2) - FIND("-", B2) - 1))'. Let me explain how this formula works:
    - FIND("-", B2): This function searches for the position of the hyphen "-" within the text in cell B2. It returns the position of the first occurrence of the hyphen within the text.
    - LEN(B2): This function calculates the total length of the text in cell B2.
    - LEN(B2) - FIND("-", B2) - 1: This expression calculates the number of characters from the hyphen position to the end of the text in cell B2. It subtracts the position of the hyphen from the total length of the text and then subtracts 1.

- RIGHT(B2, LEN(B2) - FIND("-", B2) - 1): This function extracts a substring from the text in cell B2. It starts from the rightmost character of the text and continues for the number of characters calculated in the previous step.
- TRIM(RIGHT(B2, LEN(B2) - FIND("-", B2) - 1)): This function removes any leading or trailing spaces from the extracted substring. It ensures that there are no extra spaces before or after the value.
- "$" & TRIM(RIGHT(B2, LEN(B2) - FIND("-", B2) - 1)): This concatenates the dollar sign "$" with the trimmed substring obtained in the previous step. It adds the dollar sign to the extracted value.

9. Then, I converted the value obtained from the "HS" column into currency using the formula '=VALUE(SUBSTITUTE(SUBSTITUTE(D2, "$", ""), "K", "000"))'.
    - SUBSTITUTE(D2, "$", ""): This function replaces the dollar sign "$" with an empty string in the value of cell D2. It effectively removes the dollar sign from the text.
    - SUBSTITUTE(result of step 1, "K", "000"): This function replaces the letter "K" with the string "000" in the result obtained from step 1. It is used to handle values that are represented in thousands. For example, if the original value was "1.5K", it would be transformed to "1.5000".
    - VALUE(result of step 2): This function converts the result obtained from step 2 into a numeric value. It treats the transformed text as a number and returns the corresponding numeric value.

Pivot tables and their Visualization

1. Industries with the Most Job Opportunities: In this analysis, we arranged the data by industry and calculated the count of jobs in each industry. This helps us identify which industry offers the most jobs. According to the analysis, the Biotech & Pharmaceuticals industry provides 66 jobs.
2. Most In-Demand Job Role: By analyzing the data, we determined the job role with the highest demand. We organized the data by the "job title" column and counted the occurrences of each job title. The most in-demand job role was "Data Scientist."

3. Industry-wise Company Organization: In this sheet, we categorized the companies according to their respective industries.
4. Average Rating of companies by Sector, Industry, and Ownership: This representation provides information about the average rating of companies based on sectors, industries, and Type of ownership
5. Number of Company by Size: The size of companies is determined by their number of employees. The most companies in the dataset are of 51 to 200 employees.
6. Most Founded Companies by Year: This analysis identifies the year with the highest number of company foundations. In this case, the most number of companies were founded in 2012, with a total of 34.
7. Job Count by State: By utilizing the "job_state" column, we determined the state with the highest number of job opportunities. According to the data, the state with the most jobs available is California (CA), with 165 jobs.

DashBoard_1

- The dashboard provides insights into the salary ranges of different job roles and their corresponding competitors. It allows users to filter the data based on specific job roles, enabling them to view the highest and lowest salaries offered within those roles by competitor companies. Additionally, the last graph illustrates the number of competitors for each company, with a value of -1 indicating no competitors exist.

DashBoard_2

- Pie chart 1
    - The pie chart represents the distribution of companies based on their size, which can be further filtered based on revenue. Additionally, the data can be filtered according to job title, allowing for a more specific analysis of company size within different job roles.
- Pie chart 2
    - The pie chart displays the number of available jobs in a specific state, which can be further filtered based on job roles. Additionally, the data can be filtered

according to the revenue generated by the companies offering those jobs, providing insights into the job market based on both job roles and company revenue.

Conclusion

- The dataset reveals that the job role of a data scientist is in high demand among companies.
- Industries such as biotech, IT services, and computer hardware offer a greater number of job opportunities compared to other industries. Job seekers in the field of data science can consider exploring opportunities in these industries.
- California is the state with the highest number of job opportunities, with a total of 165 available positions.
- Companies in the size range of 51-200 employees offer a significant number of job opportunities.
- A majority of companies were founded in 2012, indicating that these companies may provide more job opportunities. Job seekers can consider exploring companies established around this time period.
- When considering company Average ratings, Many company received the 5 rating among all companies.
- The dataset also provides information on the lowest and highest salary ranges for each job role. Based on the analysis, the average salary range is determined to be $181,000 in total.

Link to dataset -

Data Science Job Posting on Glassdoor

Web scrapped job posts from glassdoor for data science jobs

https://www.kaggle.com/datasets/rashikrahmanpritom/data-science-job-posting-on-glassdoor