# Mini Project on Visual Analytics

Visualisation for Data Analytics (CSC-40048)

## 1 Introduction

In the modern world data is treated as electricity. This is the era of Big Data and Artificial Intelligence age. There has been a significant data explosion, resulting in the emergence of modern technologies and smarter products. Every day, 2.5 exabytes of data are created. In the last decade, the demand for data has skyrocketed. Many businesses have built their operations around data. Data has given rise to new industries in the IT industry [1].

All the methods and tools required to analyze and evaluate a set of data are referred to as data analytics. Analytics is a wide phrase that encompasses all the tools and processes that are used to analyze and visualize data. We can extract valuable data from raw data and analyze it to provide patterns and numerical data that can assist in making intelligent judgments [2]. In Data analytics, we follow a set of preprocessing techniques and visualization methods to make the data easier for analysis. This data preprocessing step removes data discrepancies or duplicates that may otherwise degrade the accuracy of a model. Data preprocessing also ensures that no incorrect or missing values are present because of human error or bugs. In short, using data preprocessing techniques can improve the dataset's accuracy and completeness [3]. Data visualization is the process of displaying raw data in graphical representations that allow users, such as business analysts and executives, to examine the data and get deeper insights. People understand information easier through pictures than through raw reports, therefore this visual approach allows you to make rapid and effective judgments.

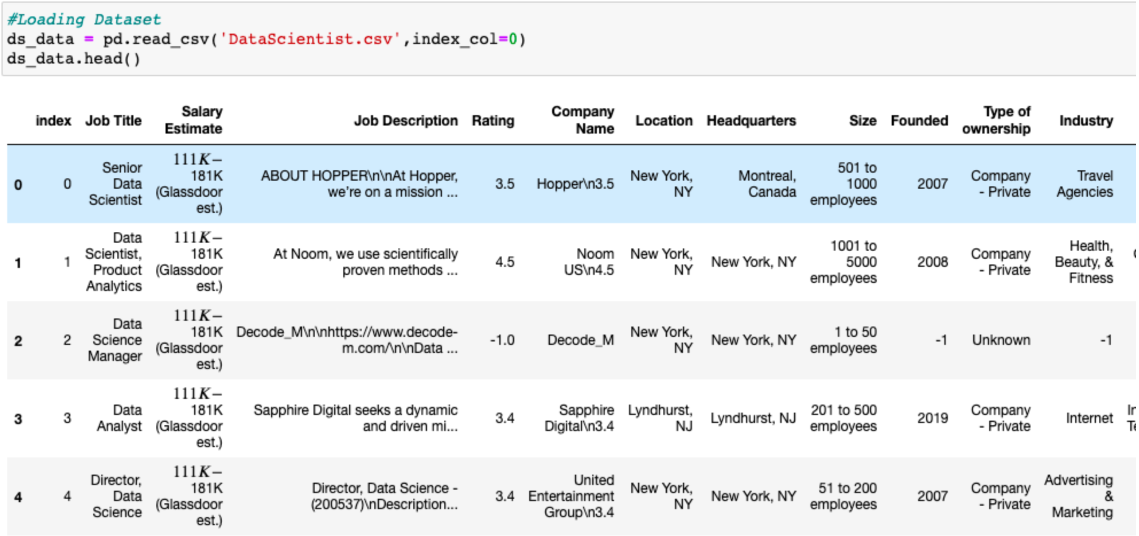
## 2 About Dataset

The dataset to work on this mini project is the 'dataScientist.csv'. This dataset is taken from Kaggle: https://www.kaggle.com/andrewmvd/data-scientist-jobs. Due to the pandamic crisis, many people lost their jobs; using this dataset, it is feasible to fine-tune the job search so that more people in need can find work.Picklesueat generated this dataset, which comprises over 3900 job listings for data scientist roles with the following attributes : index, Job Title, Salary Estimate, Job Description, Rating, Company Name, Location, Headquarters Size, Founded, Type of ownership, Industry, Sector, Revenue, Competitors, Easy Apply.To sum-up we could say our dataset consists of 3900 rows (records) and 16 columns (attributes).Using the dataset, we can find the top jobs based on salary and company rating, examine job descriptions for required skills and predict salary based on industry, geography, and corporate revenue.

## 3 Data Preprocessing

We need to import the required libraries for data analysis. We are importing the following libraries: pandas, matplotlib.pyplot ,NumPy and seaborn. As the first step of Data preprocessing, we need to load the dataset to our environment. We are using panda's inbuilt function read_csv() to load the dataset. Figure 1 depicts the dataset: 'DataScientist.csv'.

Pandas will create a dataframe for storing the dataset. A dataframe, like a spreadsheet, is a data structure that organizes data into a two-dimensional table of rows and columns. Because they provide a flexible and straightforward way of storing and working with data, DataFrames are one of the most frequent data structures used in modern data analytics. We need to eliminate unnecessary attributes in the dataframe. Because in figure 1, we could see that there is duplication of index values. We will remove the index attribute by del() function. To display the

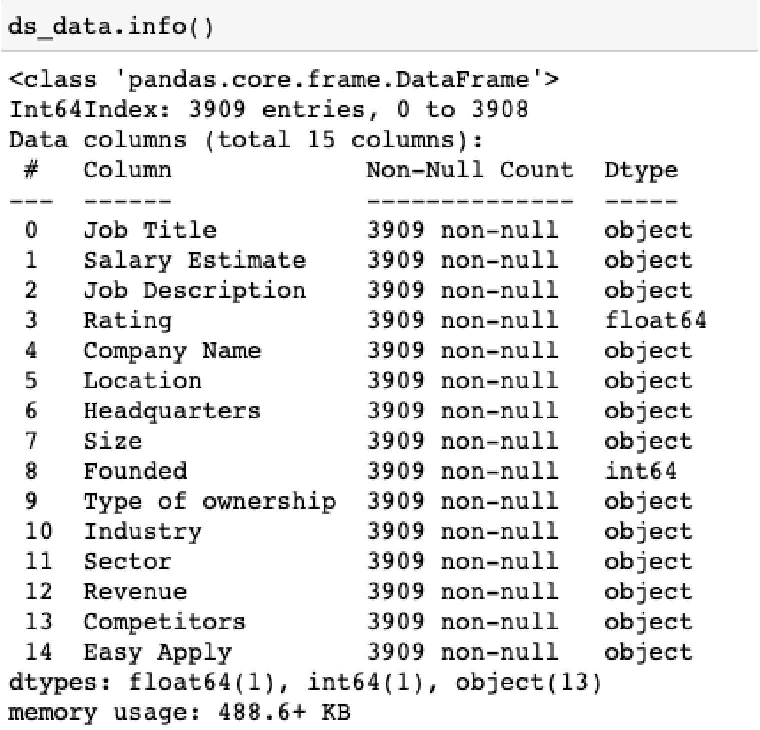


```
#Loading Dataset
ds_data = pd.read_csv('DataScientist.csv',index_col=0)
ds_data.head()
```

| | index | Job Title | Salary Estimate | Job Description | Rating | Company Name | Location | Headquarters | Size | Founded | Type of ownership | Industry |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | Senior Data Scientist | $111K-$181K (Glassdoor est.) | ABOUT HOPPER\n\nAt Hopper, we're on a mission ... | 3.5 | Hopper\n3.5 | New York, NY | Montreal, Canada | 501 to 1000 employees | 2007 | Company - Private | Travel Agencies |
| 1 | 1 | Data Scientist, Product Analytics | $111K-$181K (Glassdoor est.) | At Noom, we use scientifically proven methods ... | 4.5 | Noom US\n4.5 | New York, NY | New York, NY | 1001 to 5000 employees | 2008 | Company - Private | Health, Beauty, & Fitness |
| 2 | 2 | Data Science Manager | $111K-$181K (Glassdoor est.) | Decode_M\n\nhttps://www.decode-m.com/\n\nData ... | -1.0 | Decode_M | New York, NY | New York, NY | 1 to 50 employees | -1 | Unknown | -1 |
| 3 | 3 | Data Analyst | $111K-$181K (Glassdoor est.) | Sapphire Digital seeks a dynamic and driven mi... | 3.4 | Sapphire Digital\n3.4 | Lyndhurst, NJ | Lyndhurst, NJ | 201 to 500 employees | 2019 | Company - Private | Internet |
| 4 | 4 | Director, Data Science | $111K-$181K (Glassdoor est.) | Director, Data Science - (200537)\nDescription... | 3.4 | United Entertainment Group\n3.4 | New York, NY | New York, NY | 51 to 200 employees | 2007 | Company - Private | Advertising & Marketing |

Figure 1: Dataset: DataScientist.csv

basic information about the DataFrame, we can use the info() method. This method allows us to see details such as number of columns, column labels, column data types, memory usage, range index, and the number of cells in each column (non-null values). Figure 2 depicts the output of info().

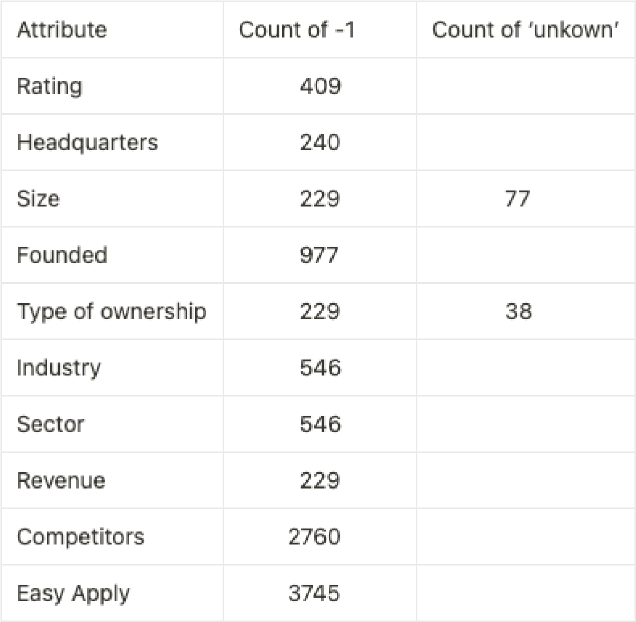


```
ds_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 3909 entries, 0 to 3908
Data columns (total 15 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Job Title          3909 non-null   object 
 1   Salary Estimate    3909 non-null   object 
 2   Job Description    3909 non-null   object 
 3   Rating             3909 non-null   float64
 4   Company Name       3909 non-null   object 
 5   Location           3909 non-null   object 
 6   Headquarters       3909 non-null   object 
 7   Size               3909 non-null   object 
 8   Founded            3909 non-null   int64  
 9   Type of ownership  3909 non-null   object 
 10  Industry           3909 non-null   object 
 11  Sector             3909 non-null   object 
 12  Revenue            3909 non-null   object 
 13  Competitors        3909 non-null   object 
 14  Easy Apply         3909 non-null   object 
dtypes: float64(1), int64(1), object(13)
memory usage: 488.6+ KB
```

Figure 2: Informtion of our dataset

When we observe the dataframe, we could see there is a lot of missing data, and it is addressed as –1 or 'unknown' in the records. We need to handle this missing data. So, I evaluated every column, Figure 3 represents the metric of missing data. We are deleting the attributes: Founded, Competitors and Easy Apply since it has more missing values. Also, we are deleting these attributes too: Job Description, Revenue and Sector, since these data does not provide insights for further analysis. Now we will replace the missing values with NaN. NaN stands for Not a Number and is one of the common ways to represent the missing value in computations. We can use NumPy.nan to represent the NaN.



| Attribute | Count of -1 | Count of 'unkown' |
|---|---|---|
| Rating | 409 | |
| Headquarters | 240 | |
| Size | 229 | 77 |
| Founded | 977 | |
| Type of ownership | 229 | 38 |
| Industry | 546 | |
| Sector | 546 | |
| Revenue | 229 | |
| Competitors | 2760 | |
| Easy Apply | 3745 | |

Figure 3: Table: Missing values information

Pandas provide functions such as isnull(), isna() to identify the missing values. We need to handle the missing values more seriously because it can impact the entire value of the data. There are two methods for handling missing values.

1. **Deleting the records having NaN values :** We can use DropNa() to drop rows or columns whose values are empty.
2. **Replacing the NaN value with statistically significant value :** We can utilize Panda's fillna() function to deal with missing values effectively. Using fillna(), Missing values can be replaced with a particular value or aggregate value, such as mean, median, or average.

We are following the first approach, because the 2nd approach is applicable only to numerical data and in our case most attributes have string datatype. Now we will drop the records or rows having NaN values using DropNa(). We can tune the function using 'how' attributes. We provided how='any' suggesting it will drop the records if any one of the columns contains NaN value. After dropping the NaN records, we need to reset the index values, for that we can use reset_index() function.

Now we can format the column data which can help in better analysis and visualization. For that, we displayed individual columns and verified whether it required any sort of formatting or not.

- **Job Title:** We could see some records have multiple job titles associated and it is separated by special symbols: comma, slash, and hyphens. We considered the first job title for our analysis. So, we split the job title based on these special symbols. In Pandas, we have the flexibility to add new functions as needed, for example: lambda functions. A lambda function can be applied to the Pandas data frame's columns and rows. An anonymous function that we can give in without having to define a name or anything else and it acts like a full-fledged function. We used lambda function to split the terms and save it on the same dataframe column.After the formatting, we varied the Job Title column and found that there is a space creating after the data scientist job titles. So, if we perform a value count operation, we will get separate results for 'data scientist' and 'data scientist '. So, we used lambda replace operation to treat this value as single entity.
- **Salary Estimate:** Here we can see the salary range of different companies and the source of the data collected. (Example: $111K-$181K (Glassdoor est.)). The source is mainly from

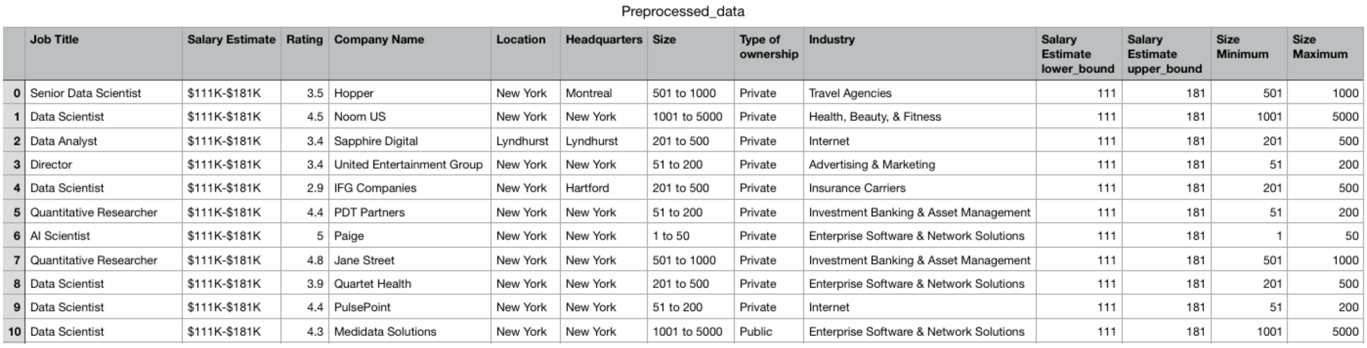
Glassdoor and Employers. In order to visualize and analyze the data based on salary perspective, we want to format this data. So, we eliminated the source string from the salary estimate. Used Lambda and split function for achieving this task (Example: $111K-$181K). We want to break the Salary Estimate column to lower and upper bound for further analysis, so we created two more columns: Salary Estimate_lower_bound and Salary Estimate_upper_bound by splitting the special symbol hyphen in the Salary Estimate. Finaly we filtered the lower (Example: 111) and upper (Example: 181) bound values with numerical data by replacing the $ and K value in the salary.

- **Company Name:** In this column, we could see that the rating value is appended to the end of the company name. We already have a rating attribute in our dataframe. So, we eliminated the rating values from the company name using the lambda spilt functions

```
0                                   Hopper\n3.5
1                                  Noom US\n4.5
2                        Sapphire Digital\n3.4
3              United Entertainment Group\n3.4
4                            IFG Companies\n2.9
```

- **Location:** As part of formatting, we removed the abbreviation terms of the location using the same using the lambda spilt logic.
- **Headquarters:** We done the same formatting as the location attribute for the headquarters column too.
- **Size:** It represents the employee size range of different companies. To make analysis based on employee size, we did the formatting like the salary attribute. We eliminated the string: 'employees' from the size column. Then we created 2 columns: Size_Minimum and Size_Maximum to store the minimum and maximum range of the size attribute, respectively.
- **Type of ownership:** Here we can see several types of ownerships. Here we analyzed the attribute value and applied a few formatting to form logical values for the term. Some records have more than one value and are separated by slashes (Example: College/University). Decided to show the first part of such terms. Few other records are separated with hyphen (Example: Company - Private). Decided to show the last part of such terms. Some other records are separated by 'or' (Subsidiary or Business Segment). Decided to show the last part of such terms. Finally, I kept the first word of the remaining values to denote the type of ownership.

Figure 4 indicates the preprocessed Dataframe.



Preprocessed_data

| | Job Title | Salary Estimate | Rating | Company Name | Location | Headquarters | Size | Type of ownership | Industry | Salary Estimate lower_bound | Salary Estimate upper_bound | Size Minimum | Size Maximum |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Senior Data Scientist | $111K-$181K | 3.5 | Hopper | New York | Montreal | 501 to 1000 | Private | Travel Agencies | 111 | 181 | 501 | 1000 |
| 1 | Data Scientist | $111K-$181K | 4.5 | Noom US | New York | New York | 1001 to 5000 | Private | Health, Beauty, & Fitness | 111 | 181 | 1001 | 5000 |
| 2 | Data Analyst | $111K-$181K | 3.4 | Sapphire Digital | Lyndhurst | Lyndhurst | 201 to 500 | Private | Internet | 111 | 181 | 201 | 500 |
| 3 | Director | $111K-$181K | 3.4 | United Entertainment Group | New York | New York | 51 to 200 | Private | Advertising & Marketing | 111 | 181 | 51 | 200 |
| 4 | Data Scientist | $111K-$181K | 2.9 | IFG Companies | New York | Hartford | 201 to 500 | Private | Insurance Carriers | 111 | 181 | 201 | 500 |
| 5 | Quantitative Researcher | $111K-$181K | 4.4 | PDT Partners | New York | New York | 51 to 200 | Private | Investment Banking & Asset Management | 111 | 181 | 51 | 200 |
| 6 | AI Scientist | $111K-$181K | 5 | Paige | New York | New York | 1 to 50 | Private | Enterprise Software & Network Solutions | 111 | 181 | 1 | 50 |
| 7 | Quantitative Researcher | $111K-$181K | 4.8 | Jane Street | New York | New York | 501 to 1000 | Private | Investment Banking & Asset Management | 111 | 181 | 501 | 1000 |
| 8 | Data Scientist | $111K-$181K | 3.9 | Quartet Health | New York | New York | 201 to 500 | Private | Enterprise Software & Network Solutions | 111 | 181 | 201 | 500 |
| 9 | Data Scientist | $111K-$181K | 4.4 | PulsePoint | New York | New York | 51 to 200 | Private | Internet | 111 | 181 | 51 | 200 |
| 10 | Data Scientist | $111K-$181K | 4.3 | Medidata Solutions | New York | New York | 1001 to 5000 | Public | Enterprise Software & Network Solutions | 111 | 181 | 1001 | 5000 |

Figure 4: Preprocessed Dataframe

## 4 Data Visualisation

After completing the data preprocessing task, now it is the time for visualization.We are going to visualize the following goals:

- **Top 10 companies based on the job vacancies:** We decided to show a bar graph of the top ten companies which have the highest number of jobs vacancies for data scientist roles. For that we created a dataframe to count the unique Company name entries and sorted it in descending order to get the highest number on the top. We used value_counts() function to get a series containing the counts of unique values. The resulting object will be in descending order by default, with the first element being the most frequent occurrence. Then we used head() for extracting only the top 10 entries. The code snippet for the dataframe creation is indicated below:

  ```
  df1 = ds_data["Company Name"].value_counts().head(10)
  ```

  Once we captured the required data in a dataframe, we can proceed to visualizing it. We used the Seaborn library's bar plot to generate the graph. The relationship between a numeric and a categorical variable is depicted using a bar plot. For x-axis, we provided the company names that we generated in the dataframe and for y-axis, we provided the value counts associated with the corresponding companies. We labeled x-axis and y-axis respectively with appropriate name and provided suitable title for our bar graph. Since some of the company names are lengthier, we used xticks(rotation=90) function to represent the x-labels in vertical fashion so that it will not override with the neighboring labels. The resulting overall diagram is depicted in Figure 5.

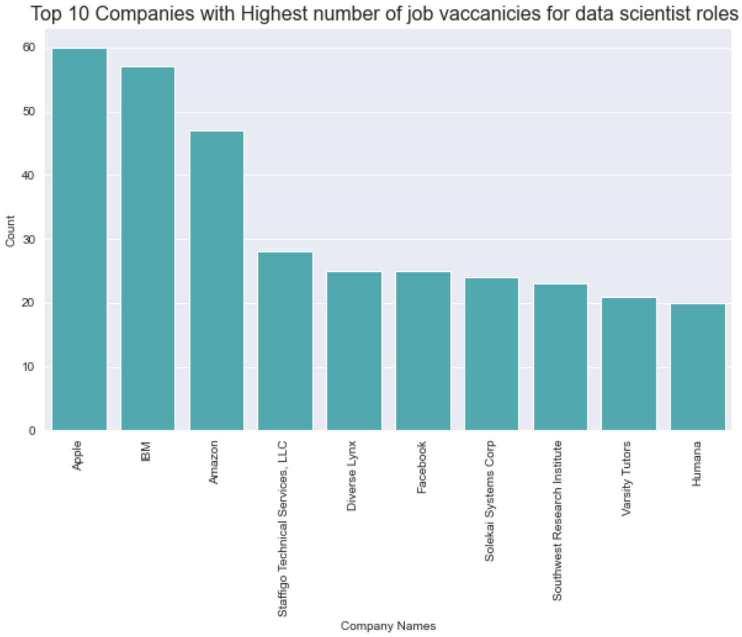
  

  Figure 5: Bar plot illustrating the top 10 companies based on the job vacancies.

  From Figure 5, we could observe that Apple has the highest number of job vacancies which is around 60.IBM has the second highest number and Amazon has the third highest number. The remaining companies have an average of job vacancies between the range 20 to 30. By looking at this metric, job applicants can apply to the companies having higher job vacancies.

- **Most Popular Data Science Job Titles:** Now we can visualize on the most popular data science job titles. We can use a bar graph to represent the most popular job titles in the dataset. For that we created a dataframe to count the unique job titles records and sorted it in descending order to get the highest number on the top. We used the same value_counts() function to get a series containing the counts of unique values. Then we used head() for highlighting only the top 10 entries which represent the most popular job tiles. The code snippet for the dataframe creation is indicated below:

```
df2 = ds_data["Job Title"].value_counts().head(10)
```

After the dataframe creation, now it is the time for visualization part. Since it is a bar graph, we can use the Seaborn library's bar plot feature. For x-axis, we provided the job titles that we created in the dataframe and for y-axis, we provided the value counts associated with the corresponding job titles. We labeled x-axis and y-axis respectively with appropriate name and provided suitable title for our bar graph. Since some of the job titles are lengthier, we used xticks(rotation=90) function to represent the x-labels in vertical fashion so that it will not override with the neighboring labels. The resulting overall diagram is depicted in Figure 6.

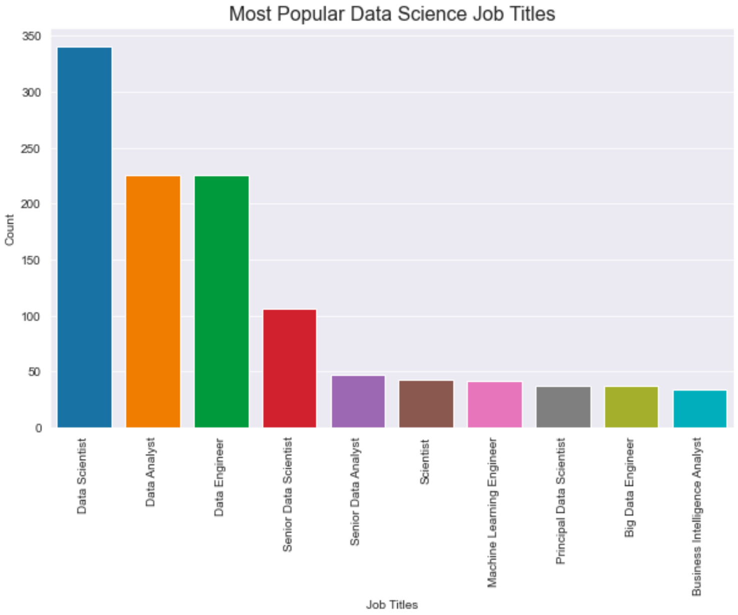


Figure 6: Most Popular Data Science Job Titles

From the graph (Figure 6), we could find that Data scientist job title has the highest occurrence, which is around 350. Following that Data Analyst and Data Engineer job titles have the second highest number, which is around 225. These are the three job titles that are popular in the data science industry. If a candidate is searching for a career in the data science industry, he can look into this graph to get an idea of job titles that are popular in this domain. If we further investigate the graph, we can find that Senior data scientist and Senior data Analyst job titles are enlisted in top 10 list. But their numbers are far less than the top 3 job titles. We presume this is because it is a senior role and needs more experience and requirements for satisfying these roles. Some of the other promising job titles in the data science industry are the following: Machine learning engineer, Principal Data Scientist, Big Data Engineer, and Business Intelligence Analyst.

- **Job vacancies based on the geographical aspects:** Next, we will visualize the job vacancies based on the geographical aspects. We can use a bar graph to represent the most popular job locations in the dataset. We are using bar graphs because they are the most logical ones to infer our goals. So, we need to create dataframe for this task to count the distinct locations and sort it in descending order to get the highest numbers on the top. Here also we are using the value_counts() function to get a series containing the counts of distinct values. We are using head() command to get the top 10 entries which represent the most popular job locations. The code snippet for the dataframe creation is indicated below:

```
df3 = ds_data["Location"].value_counts().head(10)
```

Now we have the required data in our dataframe, we will plot the bar graph using Seaborn library's bar plot feature. For x-axis, we provided the count associated with corresponding job locations and for y-axis, we provided the job locations. We labeled x-axis and y-axis respectively with appropriate name and provided suitable title for our bar graph. We created a horizontal bar graph in this case for better visualization. The resulting overall bar graph is illustrated in Figure 6.

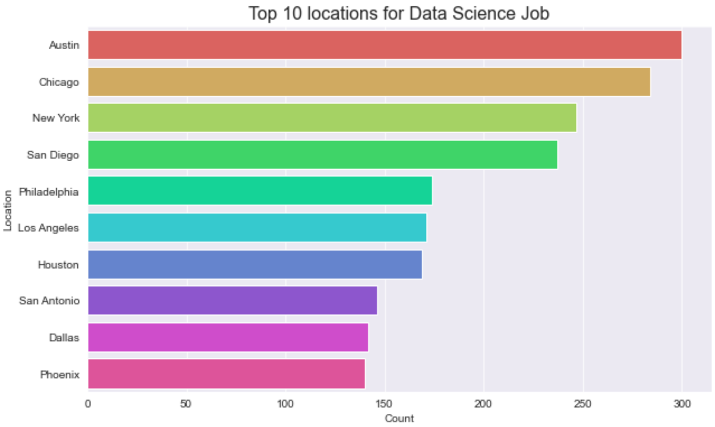
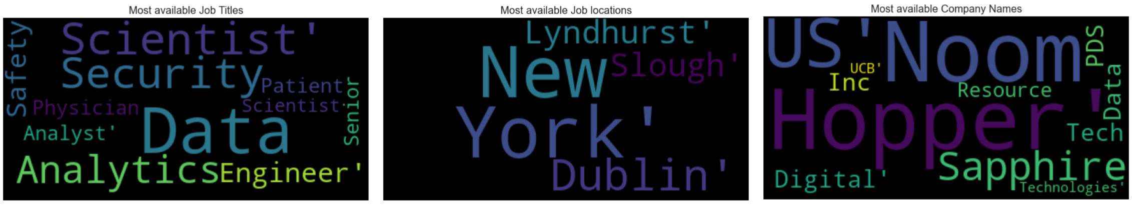


Figure 7: Job vaccanices basesd on the location attribute.

From the horizontal bar graph, we could observe that Austin is the best location to prosper the carrier in data science field. Because we could see a higher number of jobs opportunities are there, which is around 300. Following Austin, Chicago comes 2nd place in terms of job opportunities. In Chicago there are around 275 job vacancies. Newyork and San Deago come 3rd and 4th position. In these locations also a job aspirant can find jobs opportunities which are about more than 200.Following are the rest of the best job locations that the data science job candidate can explore: Philadelphia, Los Angeles, Houston, San Antonio, Dallas, Phoenix

- **Word cloud visualization on companies, job titles and locations:** Word clouds, also known as tag clouds, are graphical representations of word frequency that emphasize words that appear frequently in a source text. The larger the term in the image, the more frequently it appeared in the document. We visualized the popular companies, job titles and locations based on job opportunities by a bar chart. Here by using word cloud, we can illustrate the word frequency of the same in the dataset. It helps to recognize how your job candidate feels about the overall job opportunities. Employment seekers can summarize their thoughts on a certain topic or keyword, as well as how popular it is in the job market.



Job titles.

Job locations

Companies

Figure 8: Word cloud representation of attributes: Job title, Job locations & Companies.

We imported dedicated following libraries for creating word clouds: WordCloud and STOPWORDS. We passed our preprocessed attributes such as companies, job titles and locations to the WordCloud() function to generate corresponding word clouds. Figure 8 shows the word cloud that we created. From the word clouds, we could see most available Job Title are Data Scientist, Analyst and Engineer. The most available job locations are Newyork, Dublin and Lyndhurst and most available Companies are Hooper, Sapphire and Noom.

- **Top Head Quarters of Job Holder Company:** We can demonstrate the metric of top Headquarters of Data Science job holder companies. Here we will use a pie chart to represent the percentage of job openings from various headquarters of job holder companies. So as a first step, we need to create a dataframe with the required conditions. We got the distinct count using the value_count () function and used default head() function to bring the top 5 entries from the records for Headquarters attribute. The code snippet for the dataframe creation is indicated below:

```
df4 = ds_data["Headquarters"].value_counts().head()
```

  So now we will create the pie chart from the above-created date frame. We used Matplotlib.pyplot function to draw the pie chart and used Seaborn library's few functions to tweak the pie chart into a more appealing fashion. For example, we used Seaborn's color palette feature to improve the coloring scheme of our pie chart. The final pie chart we generated is represented in Figure 9.

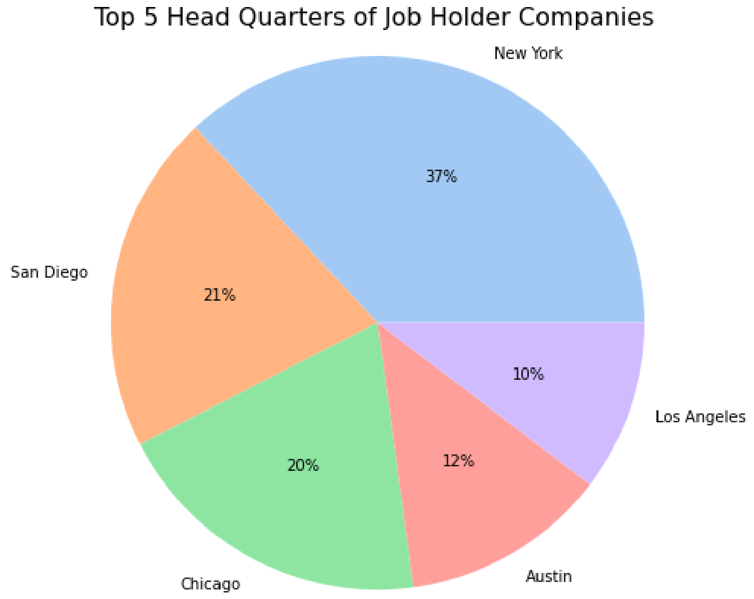


Figure 9: Top 5 Head Quarters of Job Holder Company

  So, by seeing this pie chart in Figure 9, we will get an idea about the percentile volume of top headquarters which are controlling the wheel of data science careers across the globe. Newyork has the highest percentage over the top 5 headquarters. 37% of the data science job market is controlled by the headquarters in Newyork. San Diego and Chicago comes with 21% and 20% respectively. Following these Los Angeles and Austin comes in 4th and 5th position in terms of percentage volume. By comparing the bar chart that we developed to find the job vacancies based on the geographical aspects, we could see that most of the trending locations are nearer or within the headquarters' location. We suggest it is better to have carrier exposure from the headquarters location because they might have more IT resources and technological support that boost our career prospects. So, we can assume it is better to opt for carriers from these headquarter regions, because these locations have the potential to grow more in the future.

- **Top 5 industries for datascience:** Here we are going to show the top industries in the data science world. We are using a pie chart to indicate the portion of industries that help in the growth of data science. So, using this pie chart, job candidates can choose the industries which they already have the domain knowledge. We created a dataframe required for the industry attribute. The code snippet for the dataframe creation is indicated below:

```
df5 = ds_data["Industry"].value_counts().head(5)
```

Using the dataframe created, like the earlier task, we used Matplotlib.pyplot function to draw the pie chart and used Seaborn library's few functions to tweak the pie chart into a more appealing manner. Since some of the names of the industries are very lengthy, we cannot place randomly over the pie chart portions so, we used legend() to label elements plotted on our pie chart. The pie chart we generated is represented in Figure 10.

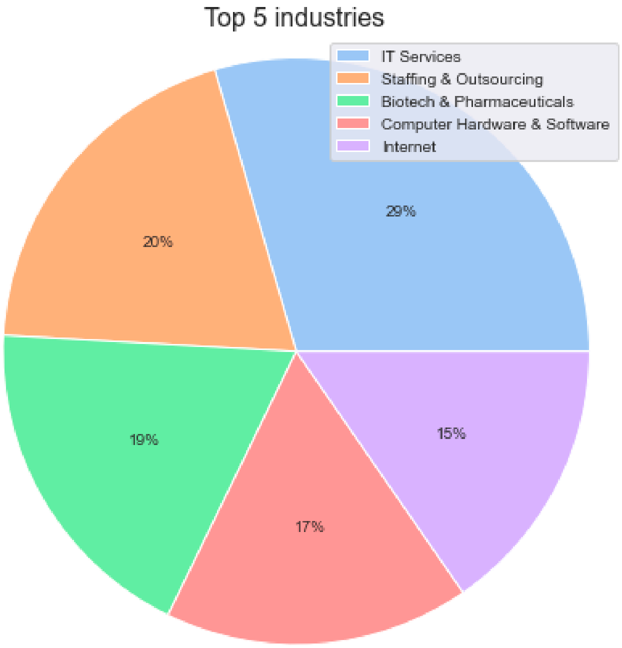


Figure 10: Top 5 industries for datascience

If the job aspirant wants to be a part of data science world, He needs to focus on these industries and learn the domain knowledge of each industry. If the job aspirant already has the domain knowledge of the industry, then he can apply job roles of that industry. Since it is a data scientist's role, knowing the domain knowledge is more crucial and gets the advantage of getting the role than the normal candidate who has no prior domain knowledge. From the pie chart in Figure 10, I could see the IT Service industry is the top industry of all and have 29% of job opportunities. Both staffing & outsourcing and Biotech & Pharmaceuticals industries have almost around 20% job vacancies available. By looking at this data, if the candidate has domain knowledge of pharmaceutical background and wishing to seek data science roles, they can choose these job opportunities.

- **Salary distribution for data scientist roles:** In this, we are going to analyze the salary distribution for the data scientist positions. To visualize the data from salary perspective, we already done some preprocessing techniques to extract the numerical information from the salary estimate attribute. We created a new column such as Salary Estimate_upper_bound and Salary Estimate_lower_bound for retaining the upper and lower bound values of salary estimates. We are using the upper bound salary column to represent the whole salary distribution of the entire data science jobs. By using the seaborne library, we created histogram for representing the salary distribution. Figure 11 shows the histogram we generated.

  By observing the histogram, we could see the maximum salary of data scientists job role offered is $250K and the minimum salary is around $25K. If we observe the salary distribution, maximum job role opportunities offer a salary range of $70K to $180K. If the job applicant wants the generic overview of salary distribution of the data scientist job opportunities, he can refer to this histogram (Figure 11).

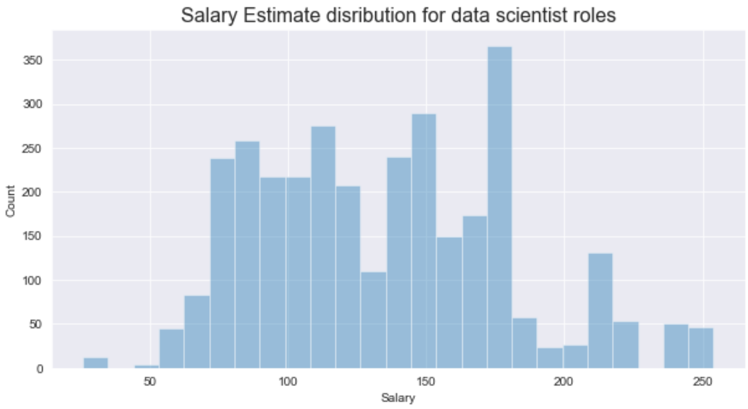


Figure 11: Histogram for Salary distribution

- **Top companies providing the highest salary:** Here we are going to display the information of top companies who are providing the highest salary for data scientist job roles. From the salary distribution we found that the maximum salary offered is $250K.So using that information we created dataframe of records having a salary greater than or equal to $250K. Then we created another dataframe using the filtered dataframe that we created before to count the distinct values of the job offering companies. As usual we used default head function to generate the top 5 records. The code snippet for the dataframes creation is indicated below:

  ```
  df6 = ds_data.loc[(ds_data['Salary Estimate_upper_bound'] >= 250)]

  df7 = df6["Company Name"].value_counts().head()
  ```

  From the filtered dataframe now it is possible to plot the bar graph using Seaborn library's bar plot feature. For x-axis, we provided the company names and for y-axis we provided the count associated with corresponding company names. The resulting overall bar graph is illustrated in Figure 12.

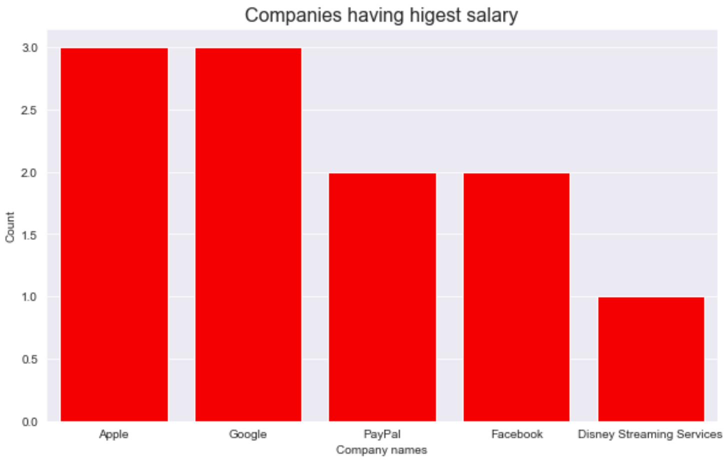
  

  Figure 12: Top companies based on the highest salary offering

  From the bar graph, we could see Companies like Apple and Google pay highest salary for data scientist job roles. PayPal, Facebook and Carbon3D are the other tech giants who offer the maximum pay. If the job aspirant is seeking to work in companies paying the highest salaries, then these are his best options available.

- **Employee size distribution for data scientist jobs:** In this, we are going to analyze the employee size distribution for the data scientist roles in different companies. To visualize the data from an employee size perspective, we already done some preprocessing techniques to extract the numerical information from the Size feature. We created new columns such as Size_Minimum and Size_Maximum to store the minimum and maximum range of the size attribute, respectively. Here we are using the maximum employee size column to portray the distribution of job vacancies based on a company's employee size. We used the Seaborn library's distplot function to draw the required histogram distribution. Figure 13 depicts the histogram we generated.

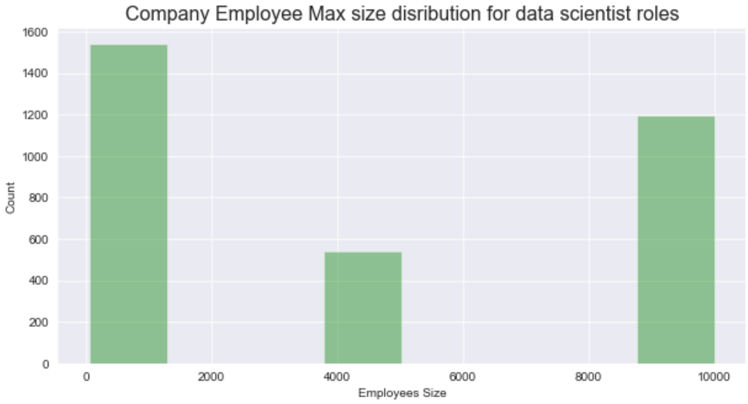


Figure 13: Employee size distribution for data scientist jobs

By observing the histogram, we could see the maximum employee size for data scientist job offering companies are more than or equal to 10000 and minimum is between the range of 50 to 1000.If the job applicant wants the general overview of employee size distribution of the data scientist job opportunities in different companies; he can refer to this histogram (Figure 13).

- **Companies having the highest number of employees:** Here we are going to highlight the details of top companies having more employee size. From the employee size distribution, we found that the maximum employee size is 10000 plus. In the dataset more than 10000 were represented by 10000+, we preprocessed it to 10000 for numerical classification.So 10000 denotes 10000 plus employee's size. So, using this knowledge, we created a dataframe of the records having employee size maximum i.e., 10000. Then we created another dataframe using the dataframe that we created now to count the unique values of companies in the dataset. The code snippet for the dataframes creation is indicated below:

```
df8 = ds_data.loc[(ds_data['Size_Maximum'] == ds_data["Size_Maximum"].max())]

df9 = df8["Company Name"].value_counts().head()
```

From the filtered dataframe we created the bar graph using Seaborn library's bar plot function. For x-axis, we provided the company names and for y-axis we provided the count associated with corresponding company names. The bar graph generated is demonstrated in Figure 14.

From the bar graph, we could note that Southwest Research Institute company has more job vacancies with maximum employee size in their teams. The rest of the companies like U.S Navy, Apex Systems, Zoom and TikTok have the highest employee size but only have half of the count for job vacancies compared to Southwest Research Institute company. If the job seeking person wants to work in companies having higher employee size, then these companies are the right choices. We assume that the higher the employee size, the higher the job openings and job applicant can purse the data science-oriented jobs from these companies.

- **Types of Ownership of Employment providers:** Here we are going to visualize the several types of ownership of employment providers. We are creating a bar graph to represent the various ownership types. We have preprocessed thoroughly the ownership attribute. By

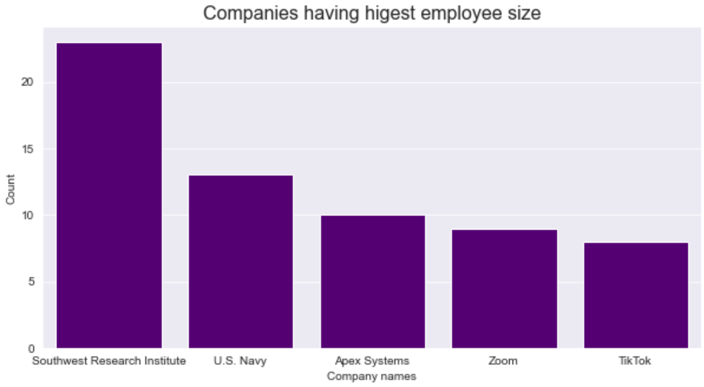


Figure 14: Bar chart- Companies having highest employee size

using that we will create a dataframe containing the distinct of ownership types. We sorted it based on ownership types. The code snippet for the dataframe creation is indicated below:

```
df10 = ds_data["Type of ownership"].value_counts().sort_index()
```

Since it is a bar graph, we are using the Seaborn library's bar plot function. For x-axis, we provided the type of ownerships that we created in the dataframe and for y-axis, we provided the value counts associated with the corresponding type of ownerships. Since some of the job titles are lengthier, we used xticks(rotation=90) function to represent the x-labels in vertical fashion so that it will not override with the neighboring labels. The resulting overall bar chart is depicted in Figure 15.

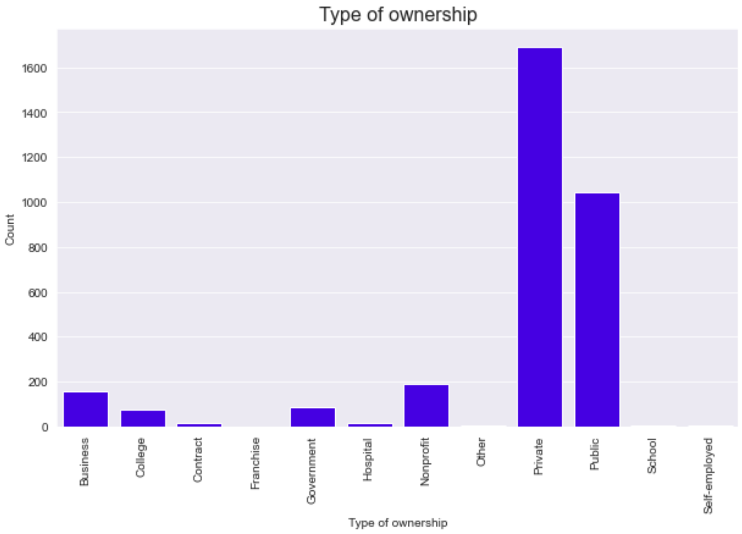


Figure 15: Types of Ownership of Employment providers

From the bar chart (Figure 15), we could identify those 12 types of ownerships of employment providers are there and out of these 12, Private ownership companies are providing more job opportunities. We could see more than 1600 job openings for private types. The public type contributes the second position which is about 1000 job opportunities.By analyzing this bar graph job seekers will get the idea of the ownership types of different employment providers.

- **Comparison of top companies having best and worst user rating:** Here we are going to compare the visualization of user rating on employment. We are creating 2 bar graphs to show maximum and worst user ratings. From the dataset, we could see maximum rating is 5, so we created a dataframe of records having rating equal to 5. Similarly, we another dataframe in which we considered the worst user ratings as less than 2. Then we created dataframes from the conditional dataframe we created now to count the distinct company names. The code snippets for the dataframes creation are indicated below:

```
df12 = ds_data.loc[(ds_data['Rating'] == ds_data["Rating"].max())]

df13 = df12["Company Name"].value_counts().head(5)

df14 = ds_data.loc[(ds_data['Rating'] <= 2.0)]

df15 = df14["Company Name"].value_counts()
```

To draw bar graphs, we are using the Seaborn library's bar plot function. For x-axis, we provided the company name and for y-axis, we provided the value counts associated with the corresponding companies. The generated bar graphs are illustrated in Figure 16 and 17.

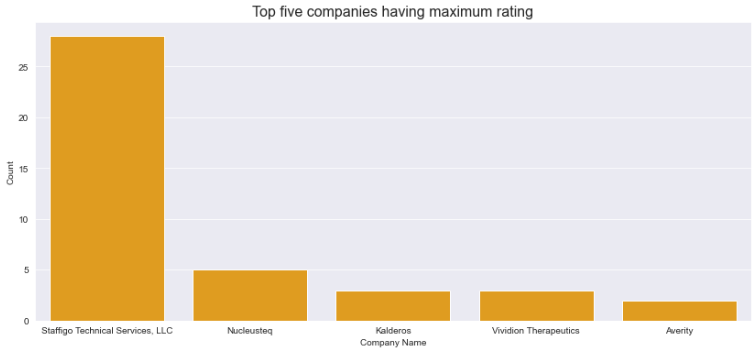


Figure 16: Top five companies having maximum rating

From the maximum rating bar chart (Figure 16), we could say Staffigo Technical Services company has job vacancies with top user rating. Following are the rest of the top 5 companies having best user rating: Nucleusteq,Vividion Therapeutics, Kalderos & Averity. From the worst rating bar chart (Figure 17), we could see there are 9 companies which have less than 2 user ratings. This user rating is collected from previous employees and Glassdoor.So if the job aspirant wants to work in a company which has the highest user rating, he can apply to the job roles in maximum rating bar chart (Figure 16). If the job seeker does not want to work in lesser user rating companies, please make sure to skip the companies in the worst rating bar chart (Figure 17).

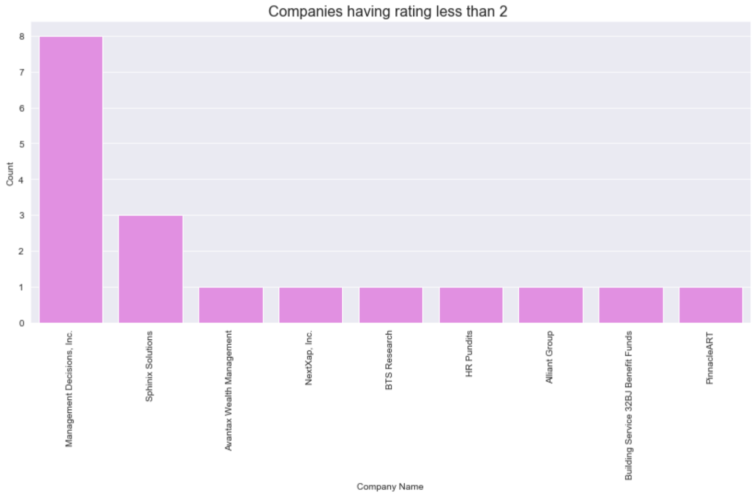


Figure 17: Companies having rating less than 2

## 5 Conclusion

In conclusion, through this mini project we learned more about data preprocessing, analysis, and visualization in Python. The dataset provided for this task was like real-world data. Because it was noisy and contained null values and was not in the right format. Using various data preprocessing techniques, we learned how to make raw data into an appropriate format. We also learned how to evaluate data for useful insights and then visualize the results. Data visualization lets users produce meaningful business insight quickly and effectively using bar graphs, pie charts, tables, and visual representations. The data visualization gives a clear image of what the data signifies and makes data easier to understand for those making business choices and critical decisions.

## References

[1] "What is the Purpose of Data Science? Know Its Importance," DataFlair, Mar. 31, 2019. https://data-flair.training/blogs/purpose-of-data-science/ (accessed April 28, 2022).

[2] TechnologyHQ, "Importance of Data Analytics in the Modern World - TechnologyHQ," technologyhq, Feb. 01, 2021. https://www.technologyhq.org/importance-data-analytics-modern-world/ (April 28, 2022).

[3] A. Joby, "What Is Data Preprocessing? 4 Crucial Steps to Do It Right." https://learn.g2.com/data-preprocessing (April 28, 2022).