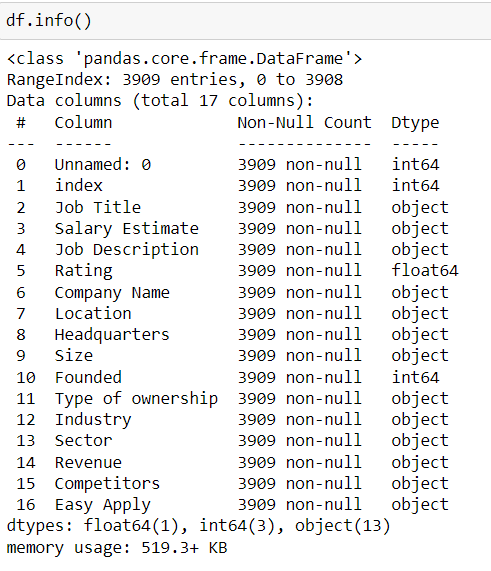
“Data Scientist Jobs” Dataset Analysis & Visualization in Python

**1. Description of the Scenario:**

In this case study, the ‘Data Scientist’ dataset is utilized to find the most hidden facts from this dataset to uncover the interesting findings regarding the job ratio in various regions [1]. To handle this objective python language is considered here for data preprocessing, data analysis and visualization by considering the different libraries like 'matplotlib' 'seaborn’ , ‘numpy’ and ‘pandas’ [2]. These are built-in libraries that are very powerful in finding helpful information from the vast dataset very efficiently. Various types of the graphs like histogram, bar charts, pie charts and correlation metrics are computed for better finings [3]. This dataset is consisted of the ‘3909’ jobs data along with the ‘16’ columns. The columns names are shown in fig 1. The types of the of the dataset variables are in form integer, objects, and floats.

Fig 1: Dataset Description



**2. Dataset Preprocessing:**

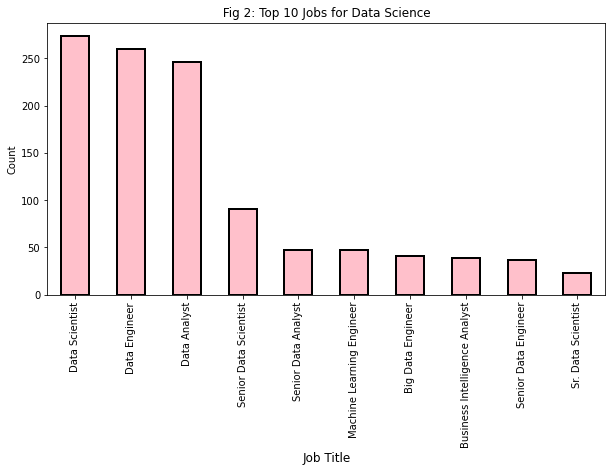
It is a significant step in data analysis to make ready the data for further processing. In the raw dataset, some anomalies and inconsistencies are handled by the data preprocessing step [4]. Moreover, the data cleaning has also undercome this phase. However, it is vital to apply the data preprocessing step to the dataset before making its visualisation. Python language is considered for data preprocessing because it can handle this step more efficiently. In the beginning, the missing values from the dataset are finds but there no missing values exit. Then, data cleaning of the various attributes is performed and remove the inconsistencies. There are some unwanted strings, characters in 'Revenue', 'Estimate salary', 'size'; these are removed to makes the features clean. Some unwanted columns are also removed, including the 'Index', ’Easy Apply’, ‘Competitors’ and 'unamed:0'.

Moreover, the four new features are computed from the two variables. The 'Size' feature is split into 'Maximum Size' and 'Minimum Size’ of employees for the whole dataset and creates new features. On the other hand, the ‘Salary Estimate’ variable is used to compute the upper and low bound of the salary by adding two columns in the dataset with 'salary\_estimate\_lower\_bound’ and ‘salary\_estimate\_upper\_bound’. Then, we removed the ‘Salary Estimate’ and ‘Size’ variables from the dataset. The ‘Company’ variable consists of some extra rating that is also cleaned by splitting it and clean the data for further step. The ‘Revenue’ variable is also cleaned because it has some ‘Unknown’ values. After all these steps, the data has become cleaned and ready for visualisation steps to find the datasets' useful hidden facts.

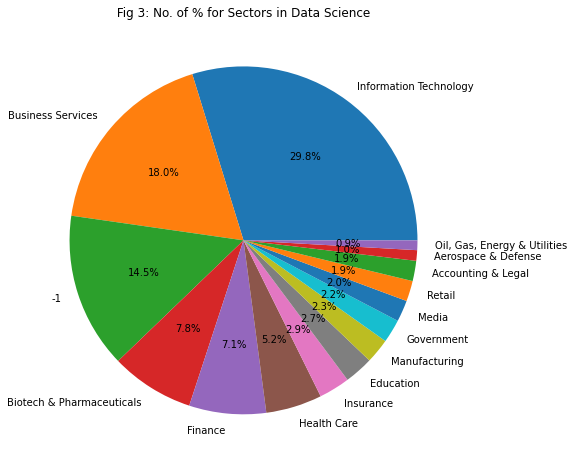
**3. Dataset Analysis & Visualization:**

Data analysis and visualization is indeed an essential element of data exploration. It is used to investigate the dataset in graphical shape in an attractive way with interesting hidden facts from dataset [5].

In fig 2, the top '10' jobs in data science are presented clearly with the bar plot's help. The job title on the x-axis and their count is on the y-axis. It is listed that the top '3' jobs in data science are a data scientist, data engineer and data analyst, respectively. These all three jobs have a significant number of counts compared to others jobs like lead data scientist and senior data engineer. So the most important job is a data scientist in the data science domain.

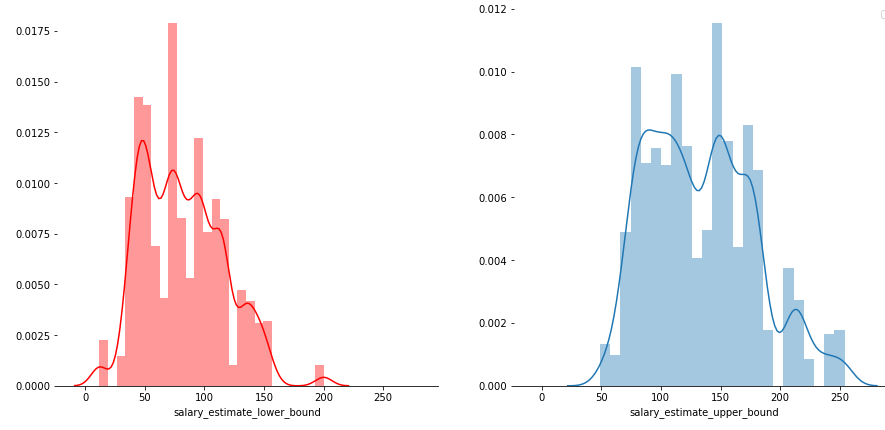


Different sectors utilised data science for better results by analysing the dataset. So, in this dataset, the % of each sector's usage data science is shown in the pie chart. The information technology sectors are using data science 29.8%, which is a high number compared to others. Although, the business service is in the second position in the usage of data science. The other domains (finance, biotech, health, education, media) percentage is also clearly illustrated for getting their usage idea regarding data science. The results are very clearly visualised for better interpretations.

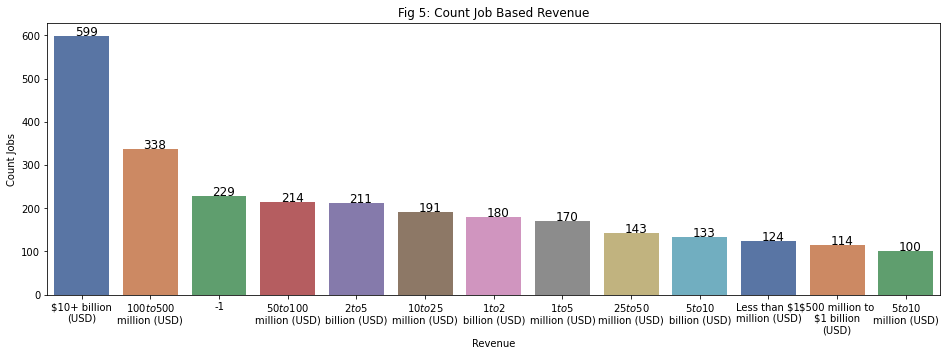


The two new features are computed from 'Salary Estimate', which are 'salary\_estimate\_lower\_bound’ and salary\_estimate\_upper\_boung’. In fig 4, a graphical depiction of these two features is plotted to obtain the thought of its distribution. The lower bound ranges are between 50-120 and upper bound 75-175 approximately. Their distributions are presented, and their number of counts can be seen on the y-axis. This presentation helps us understand these two features' ranges values with the reference line's help.

Fig 4: Salary estimates Upper and Lower bound

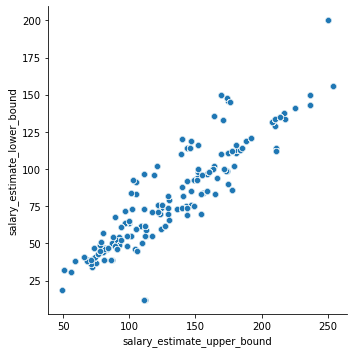


The 'Revenue' feature is also analysed and visualised by counting the jobs against each revenue. Fig 5 shows the count of jobs with respect to each revenue, which helps to understand the jobs count regarding revenue ranges in the dataset. The '$ 10 billion have a huge number of jobs count as compared to the others ranges. Moreover, the findings clearly show the different ranges of the revenue from $ 10 billion to $ 5 million.



The correlation is positive between upper and lower bound salaries and its is shown in fig 6 using the scatter plot. It shows that these are strongly and positive relation to each other.

Fig 6: Correlation score



Ultimately, after discovering some ideas, we are used visual representations to communicate the conclusions with readers. The findings of this case study about the dataset's visualization analysis are presented from fig 1-6. The hidden insights of the dataset for each feature are illustrated.

**4. Summary:**

In the end, the summarisation of the dataset's preprocessing, analysis and visualizations are going to be described. In the beginning, the description of the dataset and its scenarios is listed by importing the dataset using the python language. The data preprocessing is performed and remover the unnecessary columns, rows, strings to make the dataset clean. After that, some of the features include unnecessary information removed to make the dataset ready for visualisation. Moreover, '4' new features are computed using the 'Salary Estimate' and 'Size' features. Later on, the visualisation is performed on the cleaned data and finds useful insights from its visualisation. Our findings show that the data scientist is the most important job among all others.

Moreover, the information technology sector has 28.9 % data science as compared to the other sectors. Besides, the highest revenue is $10 billion with 599 job counts from the dataset. Other revenues are also listed to get the idea based on revenue. As a result, the dataset's visualisation helps them make the decision very efficiently and can be used for their solution as a baseline.

## **References**

[1] Glassdoor, "Data Scientist Jobs 3909 job listing for data scientist roles scraped from glassdoor," Kaggle, ed., 2020.

[2] W. McKinney, *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*: " O'Reilly Media, Inc.", 2012.

[3] V. S. N. Prasad, B. Siddiquie, J. Golbeck, and L. S. Davis, "Classifying computer generated charts." pp. 85-92.

[4] S. García, J. Luengo, and F. Herrera, *Data preprocessing in data mining*: Springer, 2015.

[5] Z. Wang, L. Sundin, D. Murray-Rust, and B. Bach, "Cheat sheets for data visualization techniques." pp. 1-13.