Market Segmentation Report

Problem statement - Job Market Segmentation

Submitted by,

Shashidhar Babu P V D

Pranjal Shrivastava

Ragesh K R

(Group B)

Contents

1. **Fermi estimation……**
2. **Data sources ……**
3. **Data pre-processing ……**
4. **Segment extraction……..**
5. **Profiling and describing potential segments……..**
6. **Selection of target segment and custom mixing……..**
7. **GitHub link……**

1. Fermi estimation

Our problem statement was to find companies most probable to hire an ML Engineer/Data Analyst applicant in respect to his/her skillset using Market Segmentation techniques. The problem involves the detailed analysis of various features like Experience, Skills, Location, Sectors, Company etc using machine learning techniques such as clustering. Our case study focuses on exploring different datas available and finding out the most favourable conditions for a Data Analyst/ ML engineer to get a job. On the way, we segmented the data as Geographic, Demographic, Psychographic, Behavioural segments and found the target segments.

2. Data Sources

Data collection was not an easy part of the project since the datas is not readily available. We explored many different data sources and a combined result of the same is used for the project. We organised the data in such a way that only necessary features were included and which can provide enough information that we needed.

Dataset 1
https://www.kaggle.com/c/kaggle-survey-2021/data

Dataset 2
https://www.kaggle.com/hrithikpawar/ai-jobs-in-india

Dataset 3
https://www.kaggle.com/usamakhan8199/current-data-science-jobs-in-india-by-glassdoor/version/1

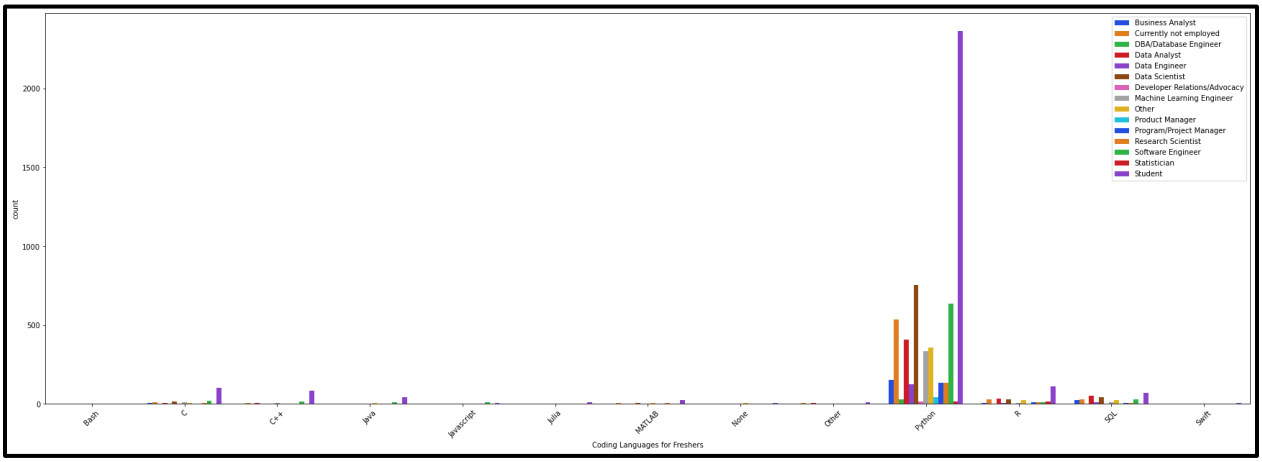
3. Data preprocessing

We used the '`kaggle_survey_2021_responses.csv`' dataset for the EDA part which contains information such as educational background, proficiency level in different coding languages and platforms, experience etc for the people all over the world. Among that we extracted the peoples

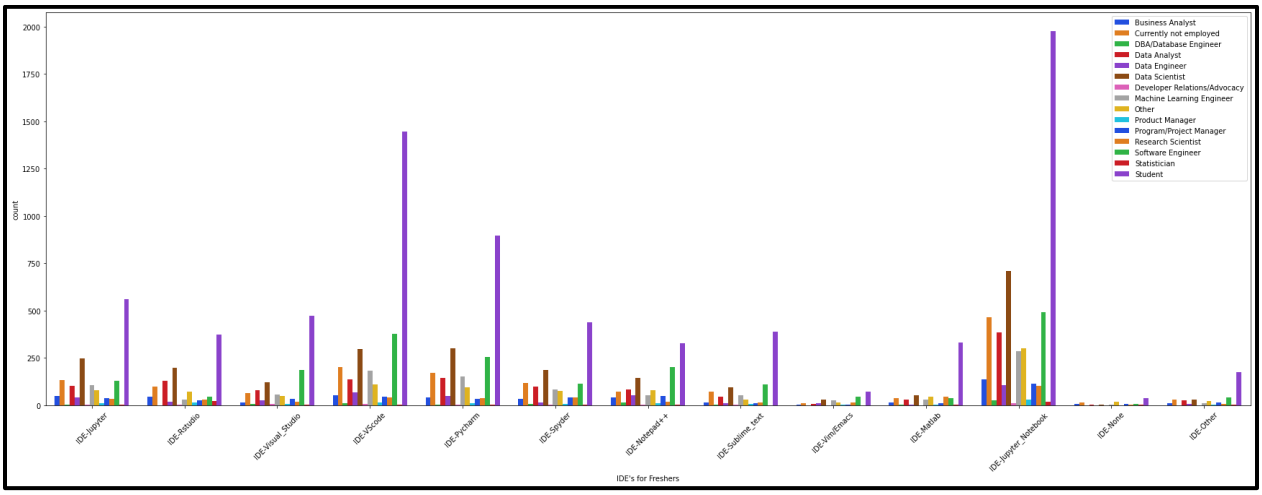
who are from India since we are dealing with the job market in India. We explored this data to analyse the trend of the employers in the data science and machine learning field.

Following are the visualisations of the data.

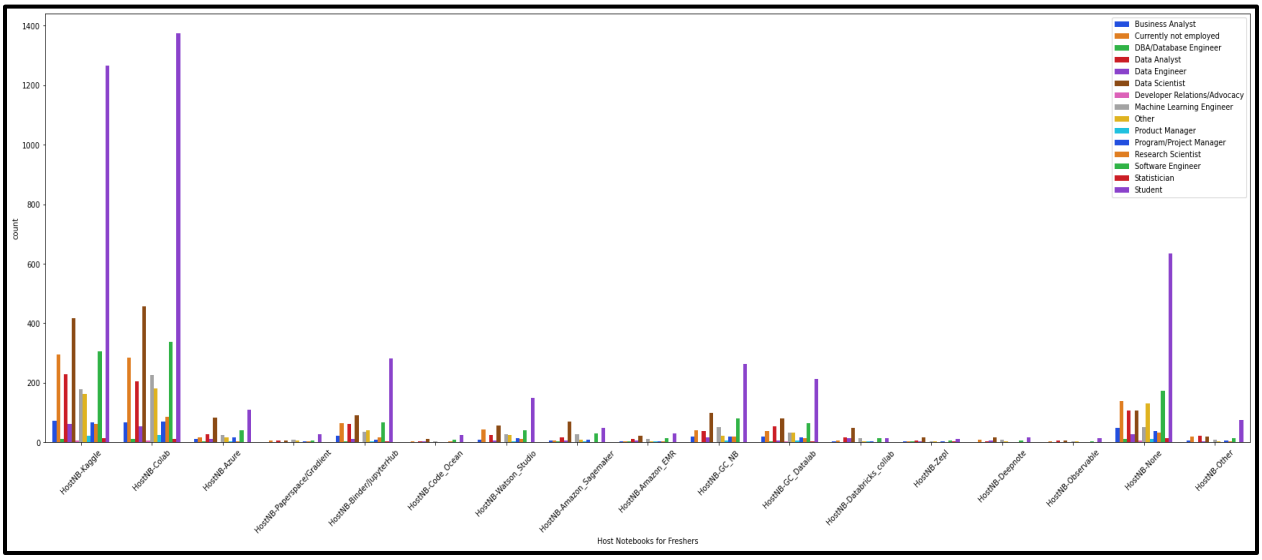
First we have plotted the different coding languages suggested for the freshers. Python is the universal favourite here & wins the vote without any ambiguity. SQL wins a distant 2nd while R is relegated to the 3rd place. Old time favourites such as C/C++ & Java are further in the back seats.



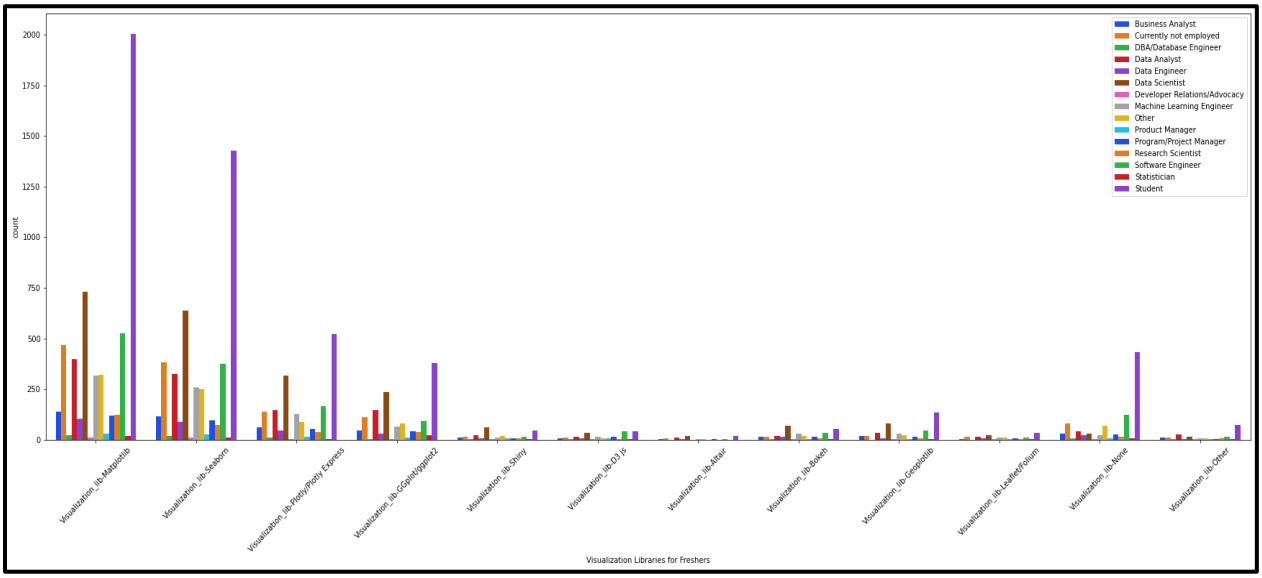
Next chart provides the details of the IDE used by various titles. The chart indicates that the Jupyter Notebook is not only popular with students but also with Data Scientists. VS Code is the next popular tool followed by Pycharm.



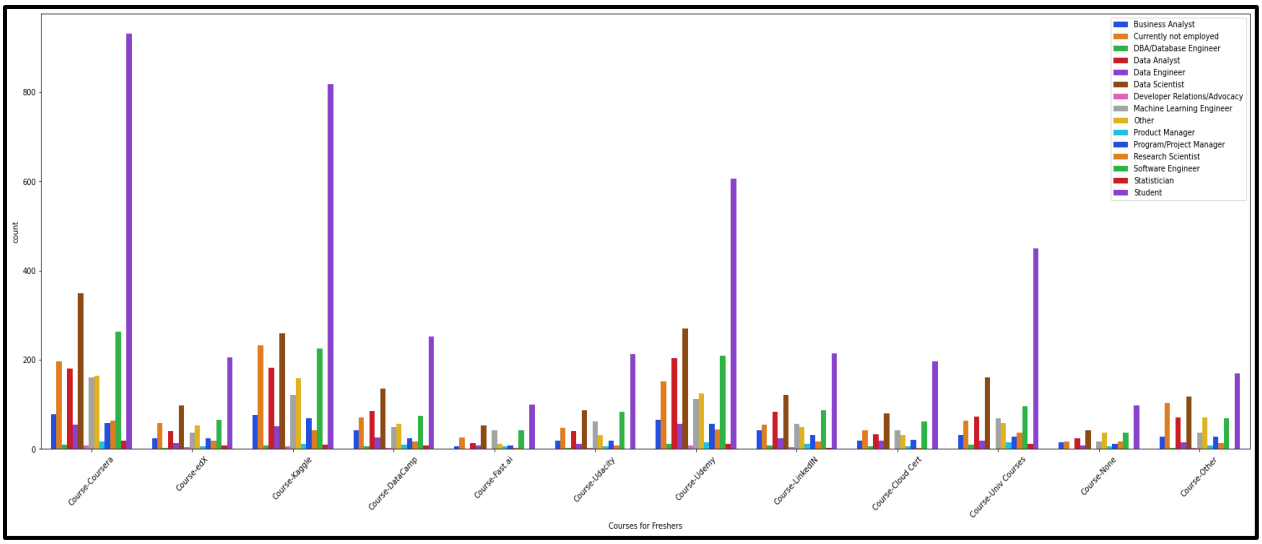
Next chart is for Hosted Notebooks by various titles. We can see the maximum number of Data Scientists are using Google Colab & Kaggle is at the 2nd position.



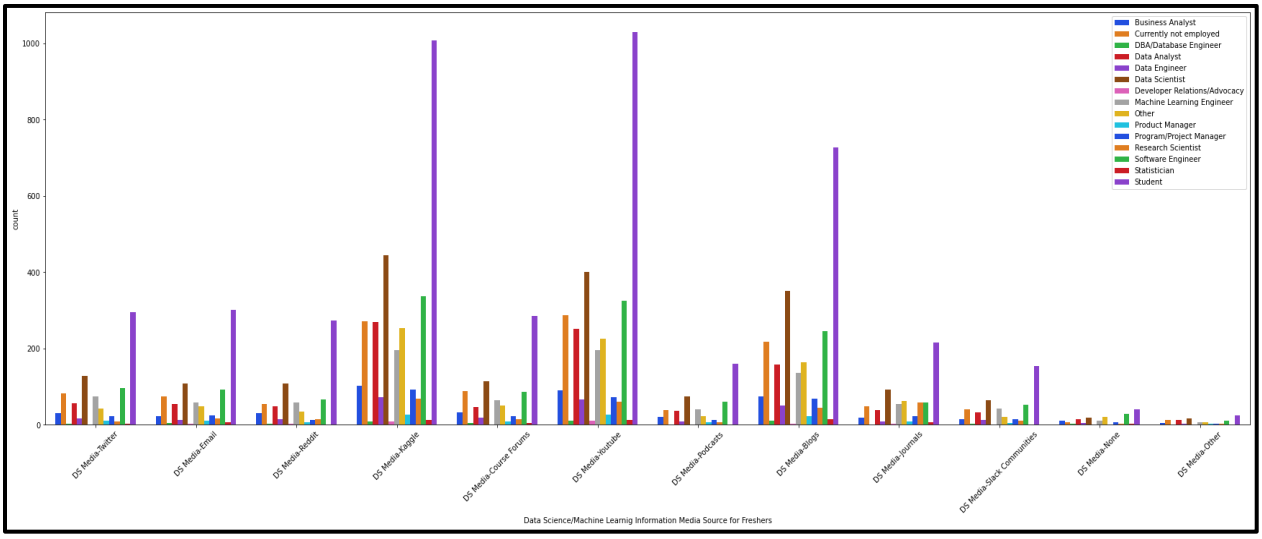
When we come to the visualisation libraries, Matplotlib, Seaborn & Plotly occupy the top 3 positions respectively.



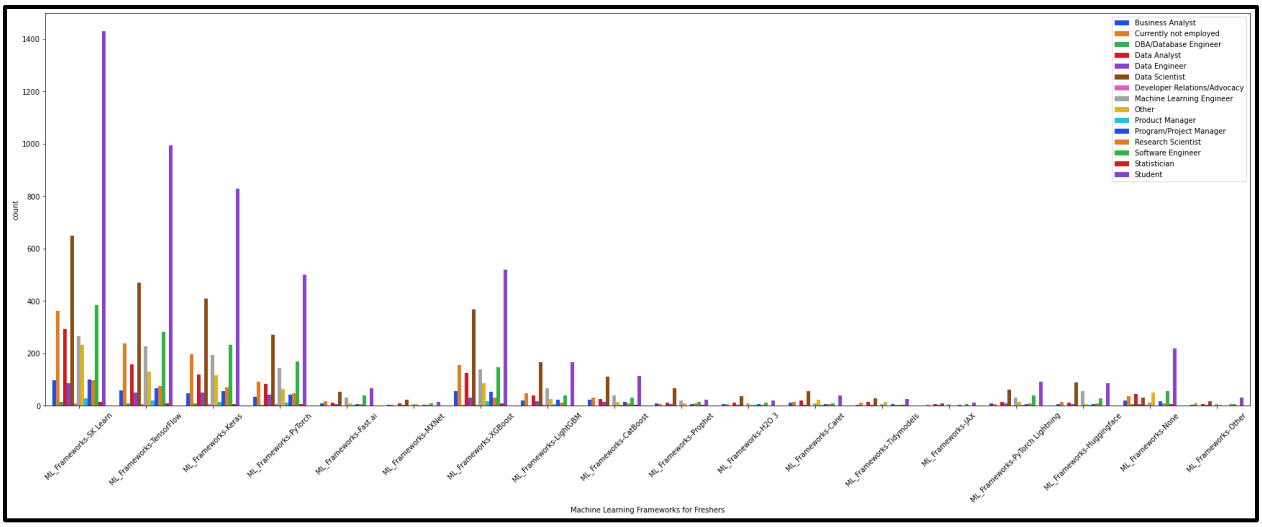
Next chart gives an overview about which online platform is preferred for educational courses by various titles. As seen in the chart below, Coursera is everyone's favourite with Kaggle in 2nd place & Udemy in 3rd position.



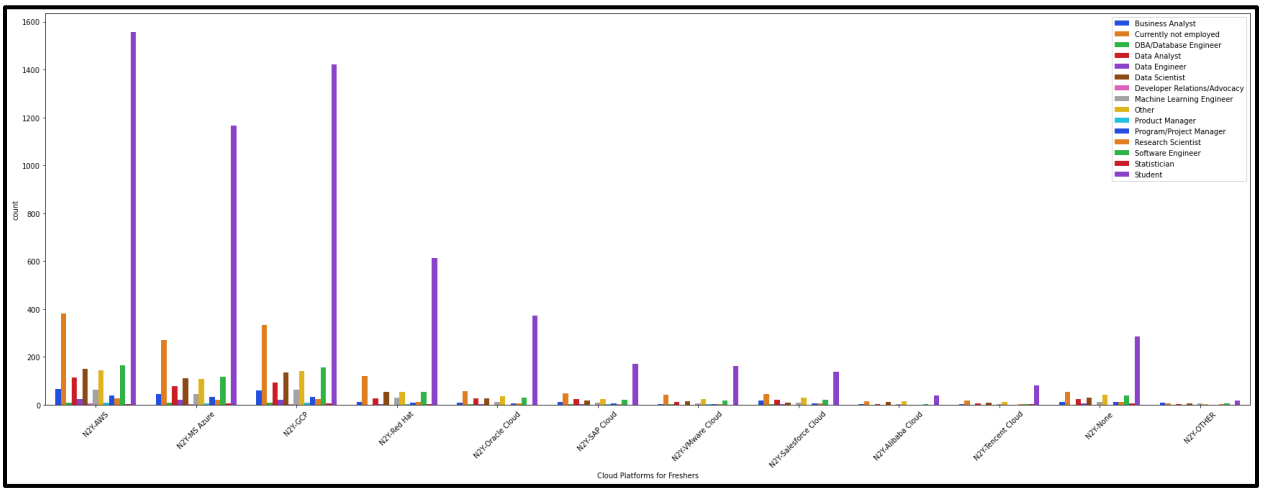
Next Chart is to understand which is the media platform used by various titles to gain information about Data Science. We can see Kaggle tops the charts here! Followed by YouTube & blogs.



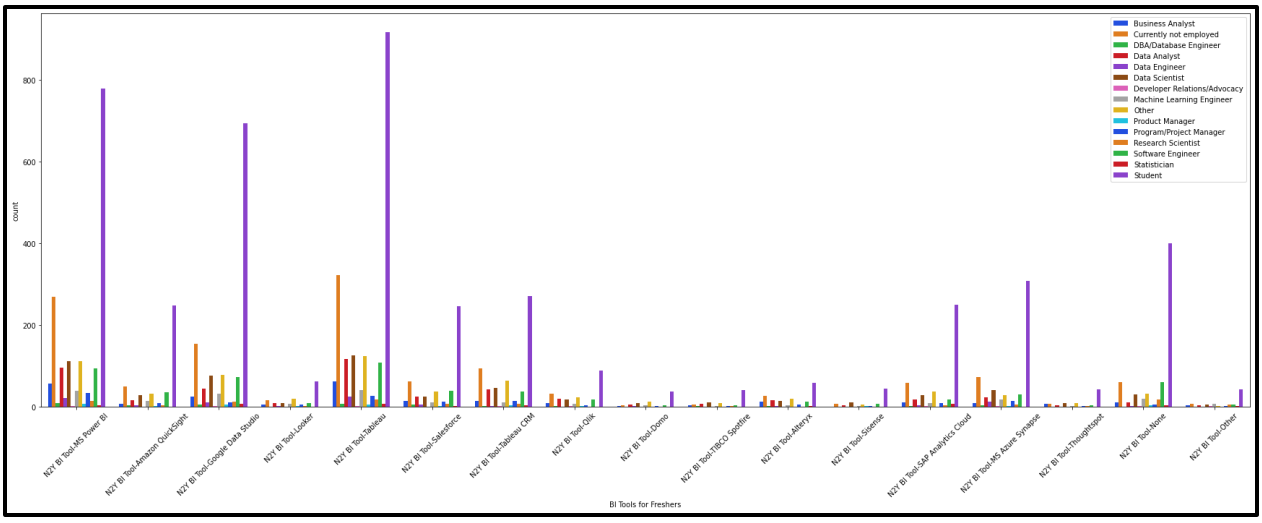
Now we will see which are the ML frameworks used by various titles in the next chart.
Scikit Learn is the most favourite here followed by TensorFlow & Keras.



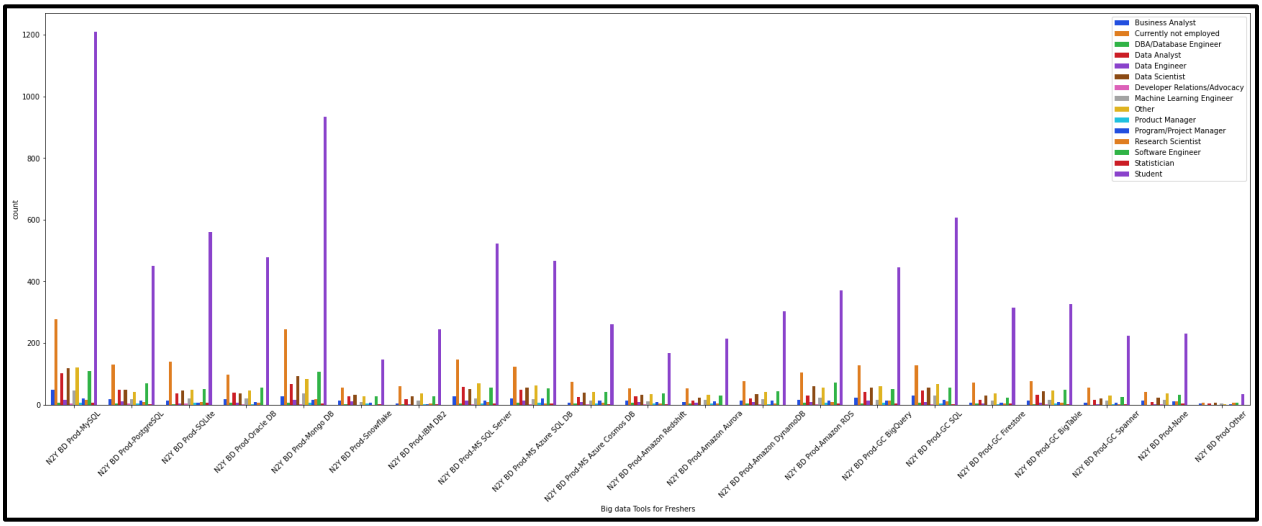
Next chart is for Best Cloud for developer experience with Coding experience of the participant. Amazon wins the vote (irrespective of coding experience) here & Google Cloud is at 2nd place with Azure coming 3th.



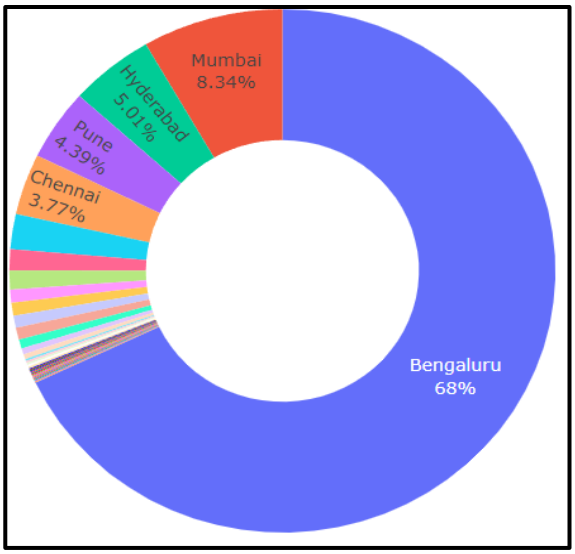
Following chart gives an overview of BI tools used by the freshers. We can note Tableau & Power BI continue being the favourites with Google Data Studio overtaking Qlik for the 3rd position.



The next chart is for the selected choice of Big Data tool by Title of the participant. We can note MySQL is the favourite of all titles & takes the top position by a long margin. Mongo DB at 2nd position is a favourite with Data Scientists.

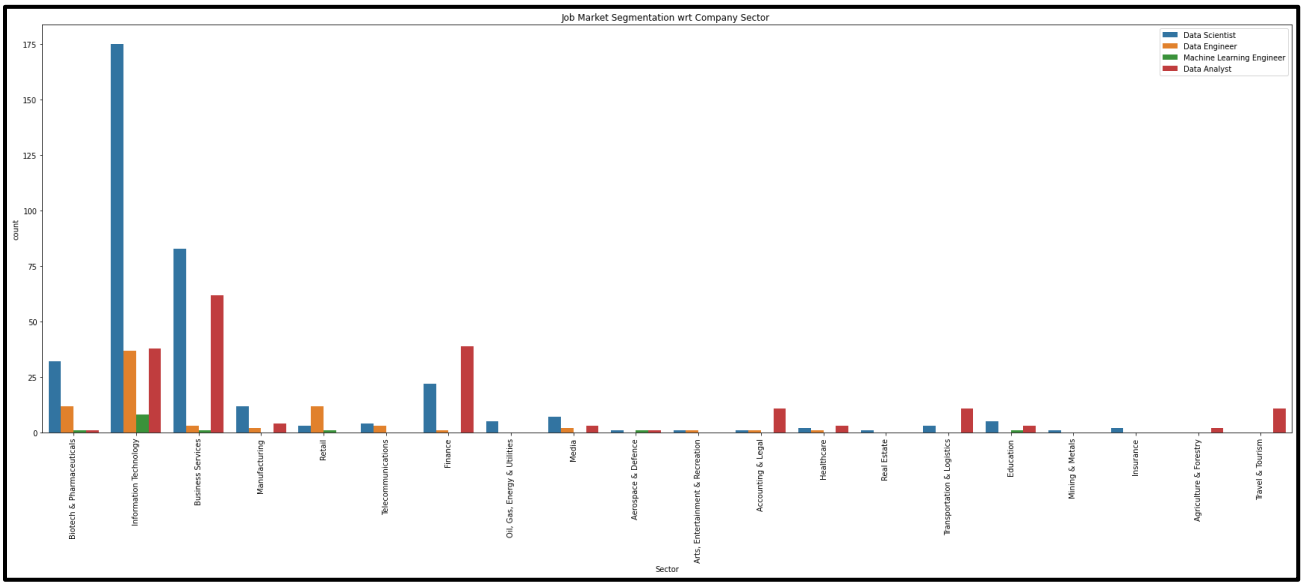


We used ‘`AIjobs.csv`’ and ‘`CJD_new.csv`’ for the purpose of clustering analysis and segmentation. From both these datasets we extracted the location details and a combined data frame was formed. A visualisation of the job market with respect to the location has been done using a pie chart which shows the location wise job openings.



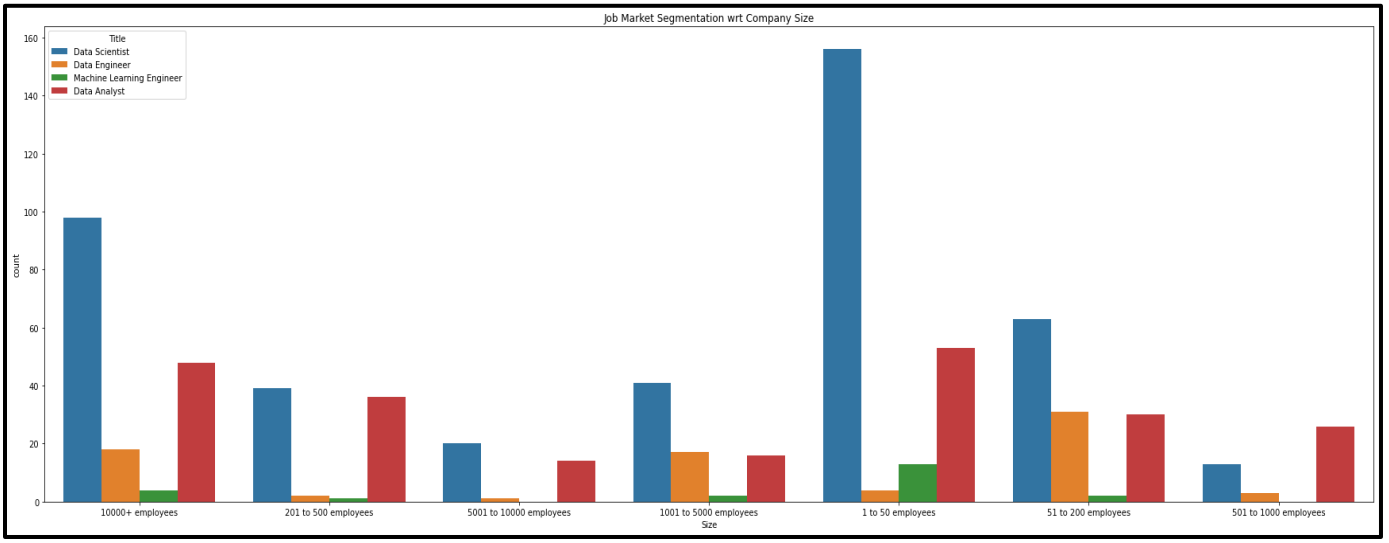
It shows that Bengaluru has the maximum number of openings.

Further we analysed different sectors and different roles like Data scientist, Data engineer, ML engineer, Data analyst in them and the result is visualised using a count plot.



Count for Data Scientist has the highest value in the information technology sector. The overall result shows that sectors such as Information technology, business services, Finance, biotech & pharmaceuticals have the highest opening for Data Scientist, Data Engineer, ML engineer, Data analyst roles.

The visualisation of the job market with respect to the company size shows the general trend of the employers with respect to the company size.



It clearly shows that for Data scientist roles, employers prefer startup companies with 1-50 employees.

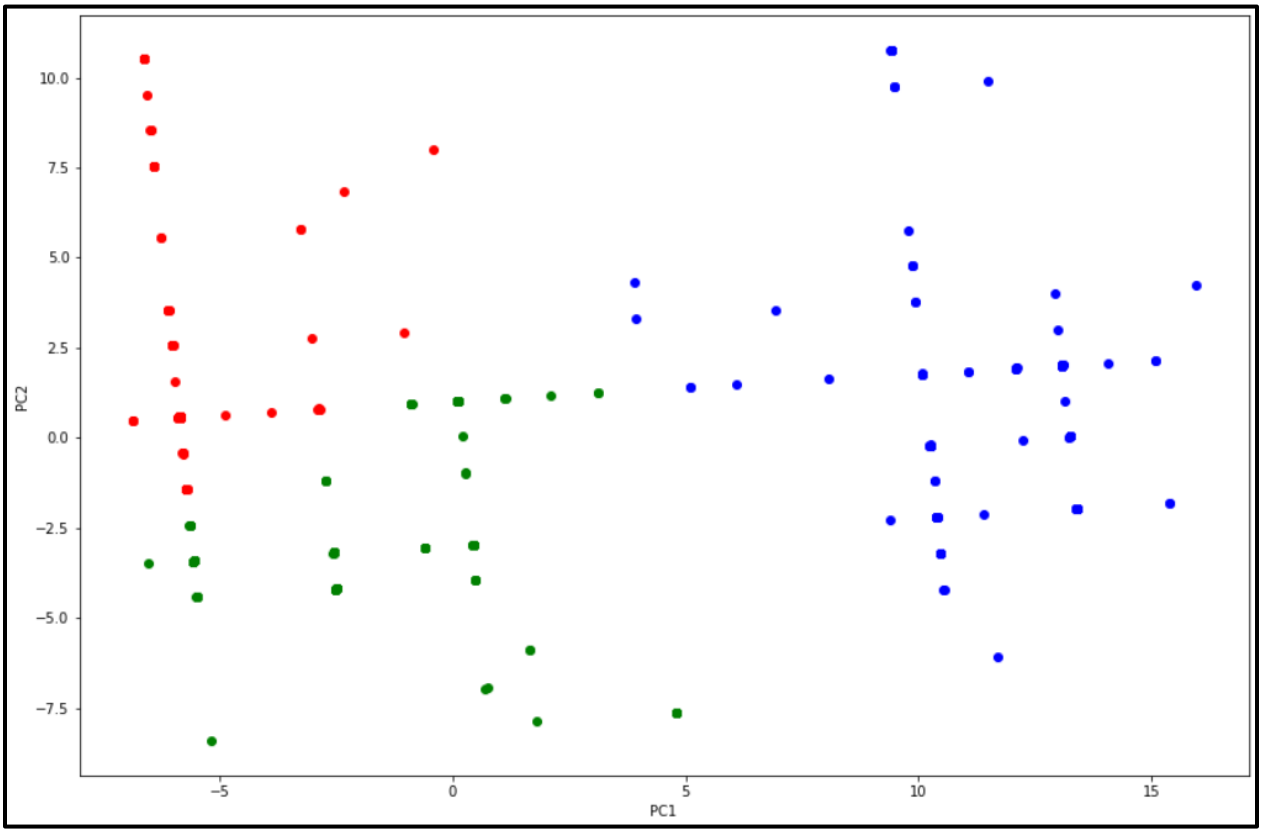
In order to deal with the categorical variables, we have used LabelEncoder from scikit-learn library. Features such as `'Title','Type.of.ownership','Sector','LocationCity'` from '`CJD_new.csv`' are converted into corresponding labels using LabelEncoder.

4. Segment Extraction

In order to extract the segments, we used K-means clustering technique. The optimal number of clusters(K) is found using the elbow method and validation was done using the silhouette_score method. And which is found to be 3. K-means clustering technique is applied on the data frame containing label encoded features - `'Title','Type.of.ownership','Sector','LocationCity'.`

5. Profiling and describing potential segment

We have obtained 3 distinct clusters, with a good amount of separation between them, the clustering experiments can be called successful. The 3 clusters obtained vary based on job title, location of the company, company sector and type of ownership. Hence, we can have 3 distinct segments here. The segments range in mainly by factors of job title, sector and location of the company. We have used principal component analysis for visualisation purposes.



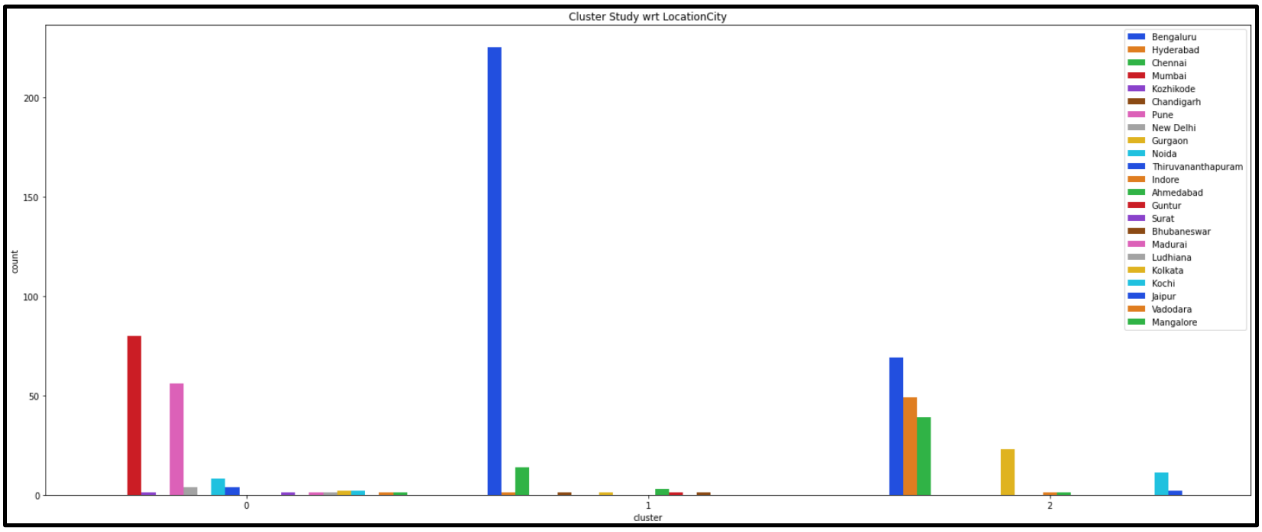
6. Selection of target segment and custom mixing

The most preferred location is Bengaluru. All the clusters suggest that the major share of vacancy is in the IT capital of the country, i.e. Bengaluru, followed by Mumbai and Hyderabad. Further it is seen that almost all the clusters are concentrated towards the Information Technology and Business Services sector and Private Ownership. Fresher

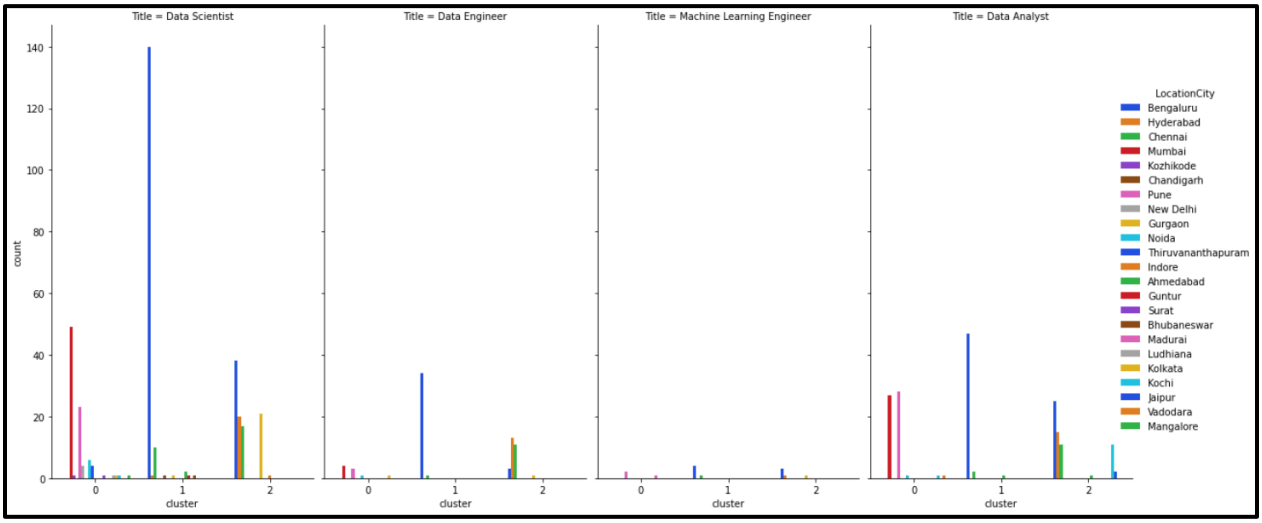
data analysts and machine learning engineers can try their hands on getting a job in these ‘A’ grade cities along with Pune, Chennai ,and Ahmedabad.

The segment has a considerable market for freshers. The most wanted skills are having knowledge about a Programming language, Machine learning, Big data tools, BI tools, and Cloud platforms. Hence, job seekers could concentrate on strengthening their knowledge in the above mentioned domains.

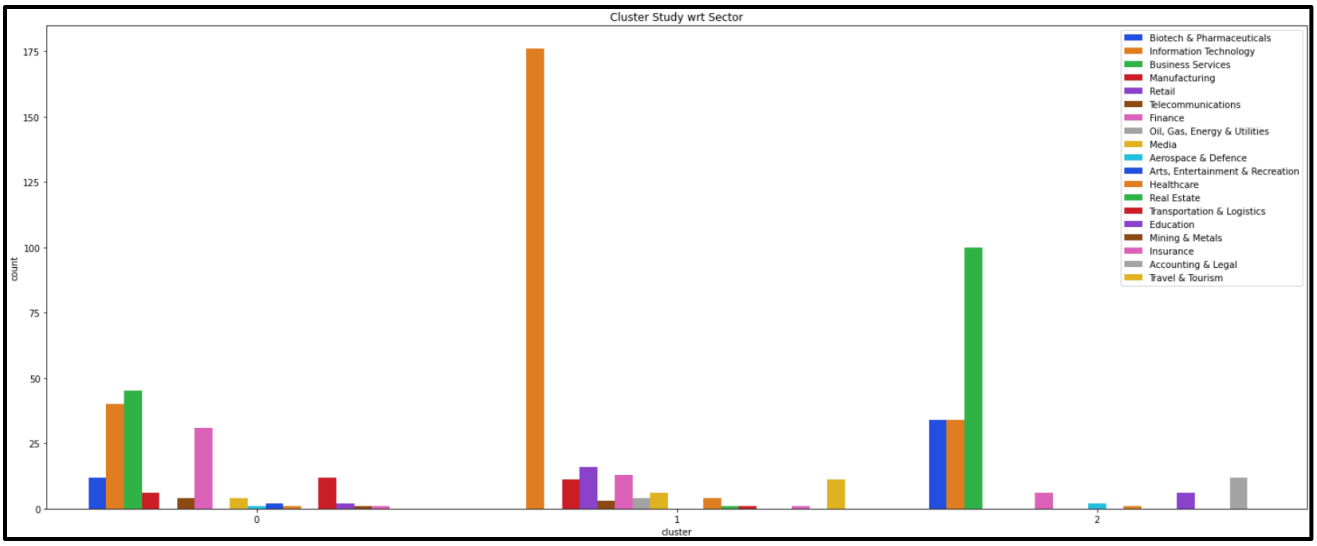
Cluster Study visualisation



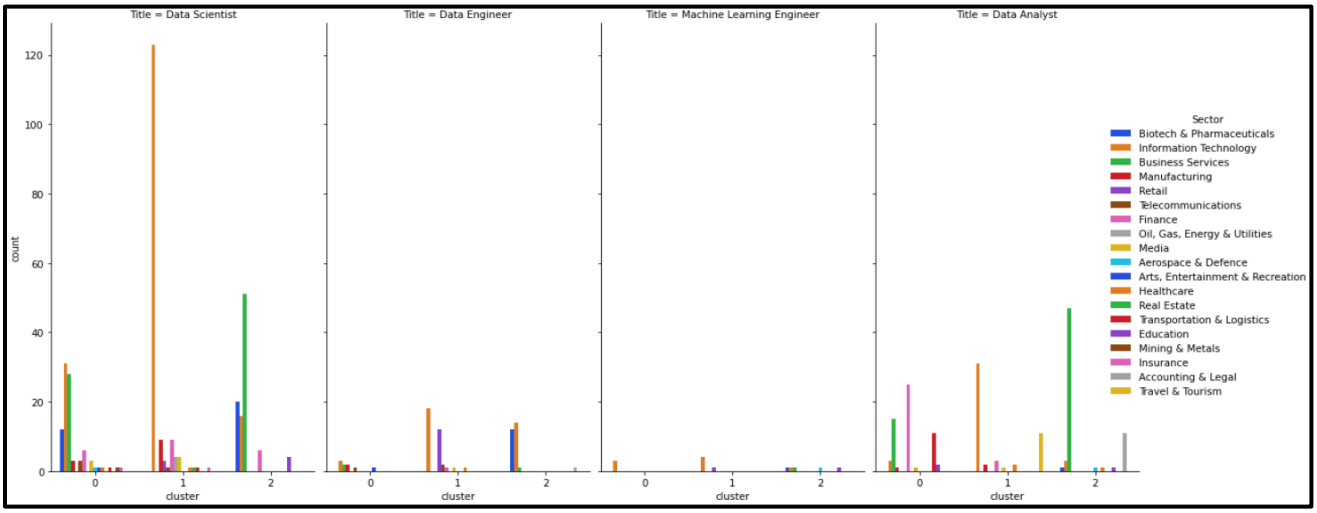
Above graph shows Cluster distribution based on Company Location.



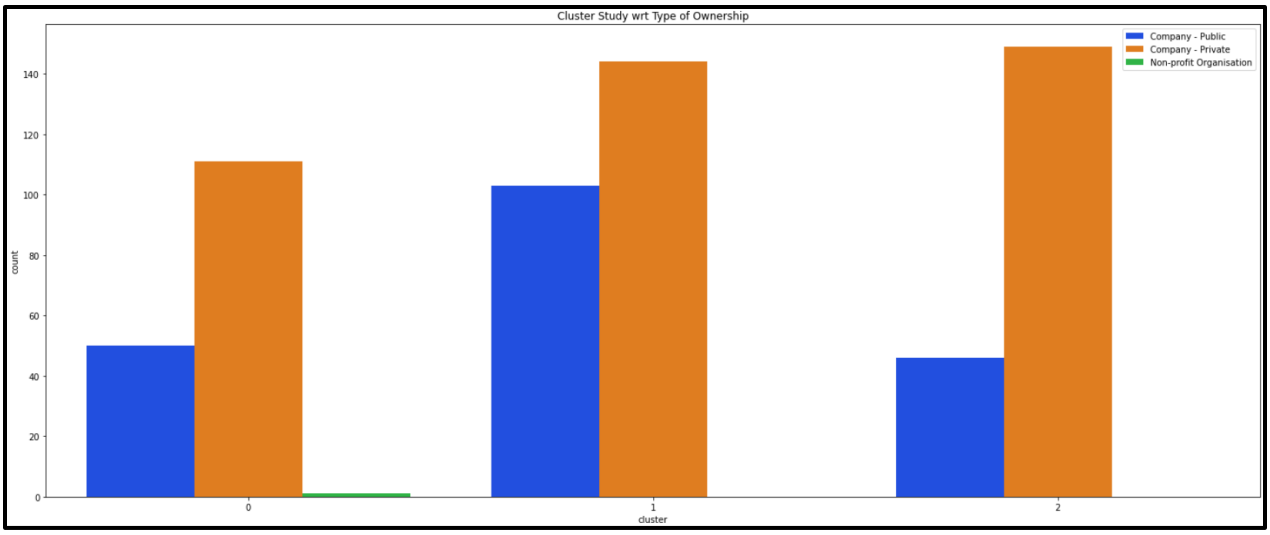
Above graph shows Job Title opportunities stats based on Company Location.



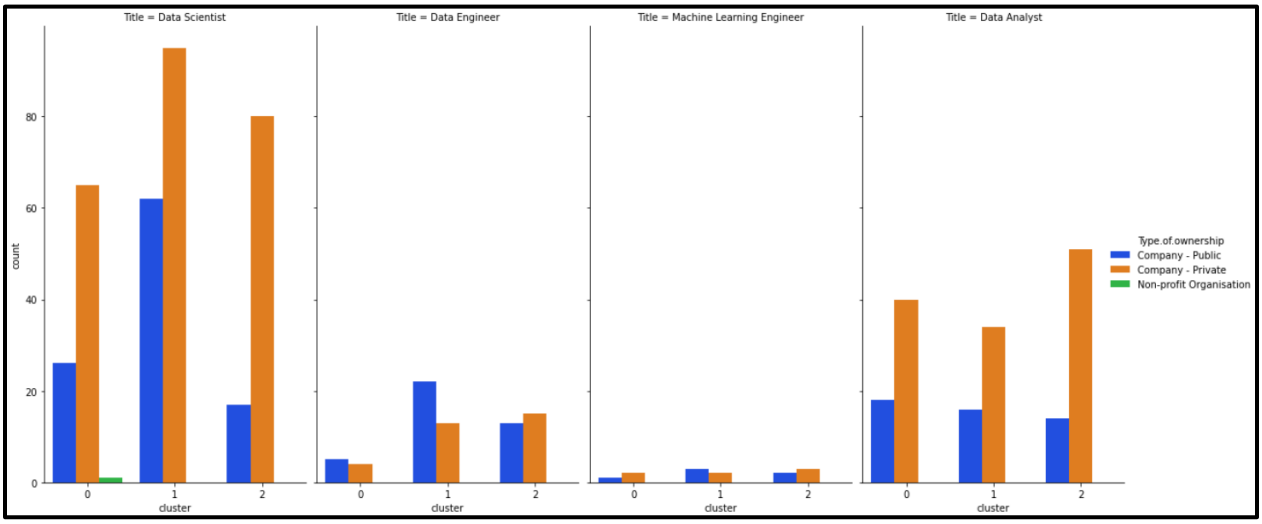
Above graph shows Cluster distribution based on Company Sector.



Above graph shows Job Title opportunities stats based on Company Sector.



Above graph shows Cluster distribution based on Company's type of ownership.



Above graph shows Job Title opportunities stats based on Company's type of ownership.

7. GitHub link

https://github.com/Pranjal755/Job_Market_Segmentation