ML JOB MARKET SEGMENTATION

A dissertation submitted in partial fulfilment of the requirements for the degree of

Master of Computer Applications
In
Computer Science and Engineering

Submitted by

AKASH JYOTI BORAH (CSM20022)

Under the supervision of

MR. SANJAY BASUMATARY
Managing Director
Feynn Labs

School of Engineering
Department of Computer Science and Engineering
Tezpur University
Napaam - 784028, Tezpur
Assam, India

June 2022

Department of Computer Science and Engineering

Tezpur University

DECLARATION

I hereby declare that the project presented in this report entitled **"ML JOB MARKET SEGMENTATION"** done as an Internship at Feynn Labs**,** submitted in partial fulfilment for the award of the degree of Master of Computer Applications in *Department of Computer Science and Engineering* during the period from January 2022 – June 2022, has been carried out by me and that it has not been submitted in part or whole to any institution for the award of any other degree or diploma.

Date: 15th June, 2022
Place: Tezpur

AKASH JYOTI BORAH
Roll no.: CSM20022
Department of CSE, Tezpur University

Department of Computer Science and Engineering

Tezpur University

CERTIFICATE

This is to certify that the project report entitled **"ML JOB MARKET SEGMENTATION"** submitted to the Department of Computer Science and Engineering, Tezpur University, in partial fulfilment for the award of the degree of Master of Computer Application(MCA), is a record of bona fide work carried out by **Akash Jyoti Borah** bearing roll number **CSM20022**, under my supervision and guidance.

All help received by him from various sources have been duly acknowledged.
No part of this report has been submitted elsewhere for the award of any other degree.

Date: 15thJune, 2022
Place: Tezpur

MR. SANJAY BASUMATARY
Supervisor

Managing Director, Feynn Labs

Department of Computer Science and Engineering

Tezpur University

CERTIFICATE

This is to certify that the project report entitled **"ML JOB MARKET SEGMENTATION"**, submitted to the Department of Computer Science and Engineering, Tezpur University, in partial fulfilment for the award of the degree of Master of Computer Application (MCA), is a record of work carried out by **Akash Jyoti Borah** bearing roll number **CSM20022**, at Feynn Labs, Guwahati.

No part of this report has been submitted elsewhere for the award of any other degree.

Date: 15th June, 2022
Place: Tezpur

Dr. Dhruba K Bhattacharyya
Internal Guide
Professor and Pro Vice-Chancellor
Department of CSE, Tezpur University

Department of Computer Science and Engineering

Tezpur University

CERTIFICATE BY THE EXAMINER

This project report entitled **"ML JOB MARKET SEGMENTATION"** submitted by **Akash Jyoti Borah** bearing roll number **CSM20022** in partial fulfilment of requirements for the degree of Master of Computer Application (MCA) of Tezpur University has been examined.

Internal examiner
Date:
Place:

External examiner
Date:
Place:

ACKNOWLEDGEMENT

Firstly, I would like to express our deep and sincere gratitude to our Project guide **Mr. Sanjay Basumatary**, Managing Director at Feynn Labs for his encouragement and valuable guidance in bringing shape to this project work. It was a great privilege and honour to work and study under her guidance.
Secondly, I am very much thankful to **Dr. Bhogeswar Borah**, Head of Department of Computer Science & Engineering for giving us the opportunity to present our work and ideas. I am thankful to my team member **Pranil Savale** for constantly supporting me throughout the internship period.
I am thankful to all the Professors and Faculty Members in the department for their teachings and academic support and also to our friends without whom this project couldn't have been successful.

Place:Tezpur

Date: 15th June2022

Akash Jyoti Borah

(CSM20022)

ABSTRACT

When it comes to work prospects, the scope of Machine Learning in India and other areas of the world is great in comparison to other professional disciplines. According to a Monster research, big data analytics and AI/ML would be the most in-demand talents in India in 2022. According to a Monster analysis, with fast tech adoption across industries and completely tech-enabled sectors like as IT and BFSI, the role of AI/ML will only rise in 2022. The average yearly income of an entry-level AI engineer in India is about 8 lakhs, which is much more than the average salary of any other engineering graduate. The compensation of an AI engineer at a high level might reach 50 lakhs.

A fresher can acquire a machine learning job if he or she possesses the necessary abilities. To have a successful career in the machine learning environment, newcomers must prepare how they will perform effectively and collaborate closely with others who have extensive expertise in the same sector.

To Start Career in ML/ AI Field, following skills are needed (may be acquired):

- Statistical Skill
- Mathematical skills and Probability Programming skills
- Advanced Signal Processing Techniques
- Distributed Computing
- Work on projects

CONTENTS

CHAPTER 1: INTRODUCTION

Segmentation is the process of isolating possible target groups in order to identify which ones would provide the best return on investment for your marketing efforts. Segmentation is determined by certain factors such as an individual's age, income, interests, and habits. When you segment different markets, you discover more about their underlying beliefs and what will eventually draw them to your brand. The basic unit of analysis for grouping consumers is the distinction between traditional market segmentation methods and the jobs-based segmentation methodology. The basic unit of analysis for traditional segmentation is the qualities of the clients themselves. A task that clients are attempting to complete is the fundamental unit of analysis for jobs-based segmentation. A market is traditionally characterised by the product and service categories specified by solution providers. According to Jobs Theory, a market is an aggregate of all available options, both provider and non-provider, that customers consider as being capable of meeting their demands in terms of getting a job done. Job segmentation provides a firm with a substantial competitive edge because it allows them to predict the value that consumers want—even before customers are aware of certain demands. A corporation may swiftly and efficiently improve existing offers and develop new ones that meet client demands better than competitor alternatives at the lowest feasible cost.

CHAPTER 2: PROBLEM STATEMENT

Finding Companies most probable to hire an **ML Engineer/Data Analyst Applicant** in respect to his/her skillset.

Data Collection/Scraping based on :

1. Geography
2. Company's field of work
3. Company size
4. Upcoming vacancies in respect to company's growth (IPO/Funding etc.)
5. Machine Learning/Data Analysis Skills currently most demanded in the market in respect to :
    i) Experience required
    ii) Time required to acquire the skill
    iii) Vacancies open
    iv) Salary etc.

We have to analyse Machine Learning Job Market in India with respect to the given problem statement using Segmentation analysis and outline the segments most optimal to apply or prepare for Machine Learning Jobs.

CHAPTER 3: BREAK DOWN OF THE PROBLEM STATEMENT USING FERMI ESTIMATION

The job market is the market in which companies look for workers and workers look for work. The job market is not so much a geographical location as it is a concept that depicts the competitiveness and interplay of many labour forces. It is sometimes referred to as the labour market. The job market has seen significant turmoil in the previous two years, owing mostly to the pandemic, and the situation is unlikely to improve in 2022.

Because the employment market is so dynamic, it's more crucial than ever to keep up with significant developments and trends. That is true not only if you are a recruiter or an employer; if you are a jobseeker, staying on top of things is your best chance of securing the best position.

A fermi estimate can help us understand this dynamic market by asking the following questions:

- What are the top skills companies are looking for?
- What is the most desired experience level in the industry?
- What are the companies that are actively offering jobs in this field?
- What are the locations that have more openings?
- What is the avg salary offered by a particular type of job in industry?

These breakups are helpful for making any decision by an individual who is willing to join a particular industry or an organisation to understand how much competition they may face if they enter a particular industry.

CHAPTER 4: DATA COLLECTION

The method of gathering, measuring, and evaluating correct insights for study using established approved techniques is characterised as data collection. On the basis of the facts gathered, a researcher might assess their hypothesis. In most situations, regardless of the subject of study, data gathering is the first and most significant stage. Depending on the information requested, the approach to data gathering differs for different topics of research. To begin, while collecting data, we should emphasise that empirical data serves as the foundation for both common-sense and data-driven job segmentation since it gives a solid foundation to work from. This information allows us to construct task segments as well as provide a comprehensive image and explanation of these segments. Although data is a valuable asset for every organisation, it does not serve any purpose until analysed or processed to get the desired results.

- Data Collection is one of the very crucial and difficult steps of data analysis.
- This part can be performed in various ways. For instance, from Kaggle, website or industry.
  perform market segmentation.
- Here, the data is a dataset from Naukri.com. It will be later used to perform market segmentation.
- This report is focused on analysing Machine Learning jobs in Indian market and draw valuable insights. The data is from the leading website Naukri.com for the year 2022 till March, IT CONSISTS OF 3878 RECORDS.
- The data is stored into 5 dictionaries which are “**roles**”, “**companies**”, “**locations**”, “**experience** “, and “**skills**”.

CHAPTER 5: DATA EXPLORATION AND DATA PRE-PROCESSING

The process of converting raw data into a comprehensible format is known as data pre-processing. Before using machine learning algorithms, the data quality should be evaluated which is done by data pre-processing.

Steps Involved in Data Preprocessing:

1. Data Cleaning:

The data may contain many redundant and missing value. Data cleaning is performed to handle this section. It entails dealing with missing data, noisy data, and so on.

2. Data Transformation:

This step is done to change the data into forms suited for the mining process.

3. Data Reduction:

Data mining is a strategy for dealing with massive amounts of data. When dealing with massive amounts of data, analysis becomes more difficult. We employ data reduction techniques to eliminate this. It seeks to improve storage efficiency while lowering data storage and analysis costs.

Dimensionality Reduction:

Dimensional reduction is the process of reducing the number of variables from the data to ensure that the reduced data conveys maximum information.

Principal Component Analysis, or PCA, is a dimensionality-reduction approach that is frequently used to decrease the dimensionality of big data sets by reducing a large collection of variables into a smaller set that retains the majority of the information in the large set.

DATA PREPROCESSING

```
In [2]: data = pd.read_csv('pca_skills.csv')
        data2 = pd.read_csv('encoded_skills_data.csv')
        data3 = pd.read_csv('final_till_12_03.csv')
```

```
In [3]: data3.head()
```

Out[3]:

	Unnamed: 0	Unnamed: 0.1	roles	companies	locations	experience	skills	skill index
0	0	0	Data Scientist - Lead / Architect Wipro	Wipro	'Kochi', ' Kolkata', ' Pune', ' Gurgaon', ' Ch...	5-10	'data science', 'python', 'it skills', 'artifi...	-1.0
1	1	1	Urgent Requirement || Data Scientist || Noida	HCL	'Noida', 'Delhi'	3-8	'it skills', 'python', 'machine learning'	-1.0
2	2	2	Global Tax Automation & Operations - Data Scie...	Dell	'Bangalore'	3-5	'artificial intelligence', 'data science', 'it...	-1.0
3	4	4	Technical Architect/ Data Scientist	DMI Innovations Pvt. Ltd	'Noida', ' Pune', ' Chennai', 'Bangalore'	8-13	'it skills', 'python', 'data science', 'machin...	-1.0
4	5	5	Technical Architect/ Data Scientist	DMI Innovations Pvt. Ltd	'Noida', ' Pune', ' Chennai', 'Bangalore'	8-13	'it skills', 'python', 'data science', 'machin...	-1.0

```
In [4]: data2.head()
```

Out[4]:

	Unnamed: 0	python	machine learning	it skills	data science	computer science	artificial intelligence	r	java	sql	big data
0	0	1	0	1	1	0	1	1	0	0	1
1	1	1	1	1	0	0	0	0	0	0	0
2	2	1	1	1	1	0	1	0	0	0	0
3	3	1	1	1	1	0	1	0	0	0	0
4	4	1	1	1	1	0	1	0	0	0	0

```
In [5]: data.head()
```

Data Exploration

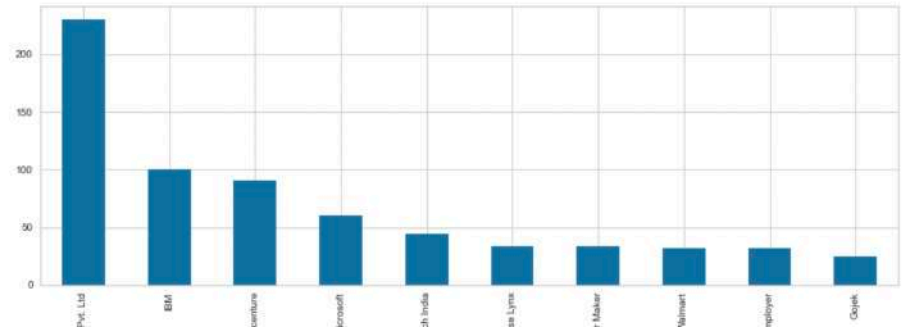
Out[5]:

	Unnamed: 0	pc1	pc2	pc3	pc4	pc5	pc6	pc7	pc8	pc9	pc10
0	0	0.998025	-0.102651	0.761568	0.396486	0.911977	0.713331	0.757557	-0.049060	-0.164950	-0.262696
1	1	0.652164	0.451644	-0.567728	-0.038726	-0.035953	-0.181696	-0.193841	-0.148095	-0.208024	-0.247326
2	2	0.906013	-0.460907	-0.151860	-0.131042	0.560132	0.479100	-0.480456	-0.067485	-0.101168	-0.076128
3	3	0.906013	-0.460907	-0.151860	-0.131042	0.560132	0.479100	-0.480456	-0.067485	-0.101168	-0.076128
4	4	0.906013	-0.460907	-0.151860	-0.131042	0.560132	0.479100	-0.480456	-0.067485	-0.101168	-0.076128

```
In [6]: x = data.iloc[:,1:]
        encoded_data = data2.iloc[:,1:]
        Y=encoded_data.to_numpy()
```

```
In [7]: f,ax=plt.subplots(figsize=(15,5))
        data3['companies'].value_counts().head(10).plot(kind = 'bar')
```

```
Out[7]: <AxesSubplot:>
```



Data Exploration

```
In [8]: encoded_data.head()
```

Out[8]:

	python	machine learning	it skills	data science	computer science	artificial intelligence	r	java	sql	big data
0	1	0	1	1	0	1	1	0	0	1
1	1	1	1	0	0	0	0	0	0	0
2	1	1	1	1	0	1	0	0	0	0
3	1	1	1	1	0	1	0	0	0	0
4	1	1	1	1	0	1	0	0	0	0

```
In [9]: pca = PCA(n_components=10)
x_pca = pca.fit_transform(encoded_data)
new_col = ['pc1','pc2','pc3','pc4','pc5','pc6','pc7','pc8','pc9','pc10']
x_pca_pd = pd.DataFrame(data = x_pca, columns = new_col)
```

```
In [10]: loadings = pca.components_
num_pc = pca.n_features_
pc_list = ["PC"+str(i) for i in list(range(1, num_pc+1))]
loadings_df = pd.DataFrame.from_dict(dict(zip(pc_list, loadings)))
loadings_df['variable'] = encoded_data.columns.values
loadings_df = loadings_df.set_index('variable')
loadings_df
```

Out[10]:

variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
python	0.674066	0.367058	0.108167	-0.211277	-0.237633	0.276762	-0.155087	0.298563	0.248480	-0.215787
machine learning	0.114402	-0.335711	-0.752470	-0.498959	-0.132286	0.011878	0.102683	0.111507	-0.002915	0.135693
it skills	0.609601	-0.036687	-0.264260	0.434944	0.123729	-0.083888	-0.054749	-0.414942	-0.411785	0.030736
data science	0.255520	-0.820957	0.460875	-0.019049	-0.164559	0.074248	-0.066456	0.046481	0.039312	0.085179
computer science	-0.216674	-0.064220	-0.247900	0.506279	-0.464039	0.628383	0.024323	-0.029752	0.139028	0.012371
artificial intelligence	-0.001671	-0.091594	-0.045006	-0.073267	0.760644	0.586547	-0.220159	0.034129	0.067545	0.086019

Loading PCA components

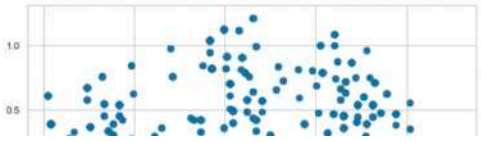
```
In [10]: loadings = pca.components_
num_pc = pca.n_features_
pc_list = ["PC"+str(i) for i in list(range(1, num_pc+1))]
loadings_df = pd.DataFrame.from_dict(dict(zip(pc_list, loadings)))
loadings_df['variable'] = encoded_data.columns.values
loadings_df = loadings_df.set_index('variable')
loadings_df
```

Out[10]:

variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
python	0.674066	0.367058	0.108167	-0.211277	-0.237633	0.276762	-0.155087	0.298563	0.248480	-0.215787
machine learning	0.114402	-0.335711	-0.752470	-0.498959	-0.132286	0.011878	0.102683	0.111507	-0.002915	0.135693
it skills	0.609601	-0.036687	-0.264260	0.434944	0.123729	-0.083888	-0.054749	-0.414942	-0.411785	0.030736
data science	0.255520	-0.820957	0.460875	-0.019049	-0.164559	0.074248	-0.066456	0.046481	0.039312	0.085179
computer science	-0.216674	-0.064220	-0.247900	0.506279	-0.464039	0.628383	0.024323	-0.029752	0.139028	0.012371
artificial intelligence	-0.001671	-0.091594	-0.045006	-0.073267	0.760644	0.586547	-0.220159	0.034129	0.067545	0.086019
r	0.079577	0.087531	0.201528	-0.304292	-0.009680	0.299896	0.710482	-0.507723	0.011843	-0.007750
java	0.161430	0.083359	-0.034246	0.177917	0.072815	-0.203345	0.067239	-0.103045	0.649735	0.670984
sql	-0.014597	0.214249	0.181486	-0.141886	-0.178282	0.193983	-0.049527	0.216874	-0.561054	0.683470
big data	0.126837	-0.064986	-0.040571	0.332862	0.229239	-0.053786	0.630215	0.637655	-0.078540	-0.043125

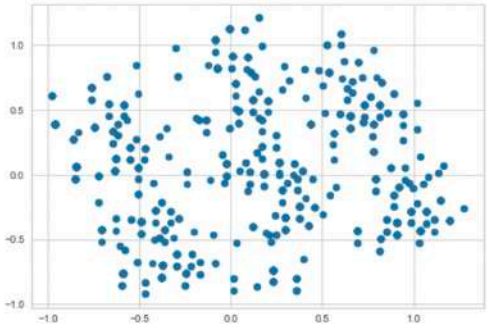
```
In [11]: plt.scatter(x_pca[:,0],x_pca[:,1])
```

```
Out[11]: <matplotlib.collections.PathCollection at 0x7f7e00114700>
```



Loading PCA Components and variable to one data frame

```
In [11]: plt.scatter(x_pca[:,0],x_pca[:,1])
Out[11]: <matplotlib.collections.PathCollection at 0x7f7e00114700>
```




```
In [12]: experience_split = data3['experience'].str[0:-1].str.split('-', expand=True)
         experience_split.head()
Out[12]:
```

	0	1
0	5	10
1	3	8
2	3	5
3	8	13
4	8	13

```
In [13]: experience_split[1] =  experience_split[1].str.strip()
         experience_split[1] = experience_split[1].str.replace('yr', '')
```

Plotting the x_pca components

```
In [13]: experience_split[1] =  experience_split[1].str.strip()
         experience_split[1] = experience_split[1].str.replace('yr', '')
         experience_split[1] = experience_split[1].str.replace(r'yr', '')
         experience_split[1].head()
Out[13]: 0    10
         1     8
         2     5
         3    13
         4    13
         Name: 1, dtype: object

In [14]: experience_split[0] = pd.to_numeric(experience_split[0], errors='coerce')
         experience_split[1] = pd.to_numeric(experience_split[1], errors='coerce')

In [15]: experience=pd.concat([experience_split[0], experience_split[1]], axis=1, sort=False)
         experience.rename(columns={0:'min_experience', 1:'max_experience'}, inplace=True)
         experience.head()
Out[15]:
```

	min_experience	max_experience
0	5	10
1	3	8
2	3	5
3	8	13
4	8	13

```
In [16]: data4=pd.concat([data3, experience], axis=1, sort=False)
         data4.head()
Out[16]:
```

	Unnamed: 0	Unnamed: 0.1	roles	companies	locations	experience	skills	skill index	min_experience	max_experi
0	0	0	Data Scientist - Lead / Architect Wipro	Wipro	'Kochi', ' Kolkata', ' Pune', ' Gurgaon', ' Ch...	5-10	'data science', 'python', 'it skills', 'artifi...	-1.0	5	

Data Pre-processing

```
In [16]: data4=pd.concat([data3, experience], axis=1, sort=False)
         data4.head()
```

Out[16]:

	Unnamed: 0	Unnamed: 0.1	roles	companies	locations	experience	skills	skill index	min_experience	max_experience
0	0	0	Data Scientist - Lead / Architect Wipro	Wipro	'Kochi', ' Kolkata', ' Pune', ' Gurgaon', ' Ch...	5-10	'data science', 'python', 'it skills', 'artifi...	-1.0	5	10
1	1	1	Urgent Requirement || Data Scientist || Noida	HCL	'Noida', 'Delhi'	3-8	'it skills', 'python', 'machine learning'	-1.0	3	8
2	2	2	Global Tax Automation & Operations - Data Scie...	Dell	'Bangalore'	3-5	'artificial intelligence', 'data science', 'it...	-1.0	3	5
3	4	4	Technical Architect/ Data Scientist	DMI Innovations Pvt. Ltd	'Noida', ' Pune', ' Chennai', 'Bangalore'	8-13	'it skills', 'python', 'data science', 'machin...	-1.0	8	13
4	5	5	Technical Architect/ Data Scientist	DMI Innovations Pvt. Ltd	'Noida', ' Pune', ' Chennai', 'Bangalore'	8-13	'it skills', 'python', 'data science', 'machin...	-1.0	8	13

```
In [17]: data4['avg_experience']=(data4['min_experience'].values + data4['max_experience'].values)/2
```

```
In [18]: data4.head()
```

Out[18]:

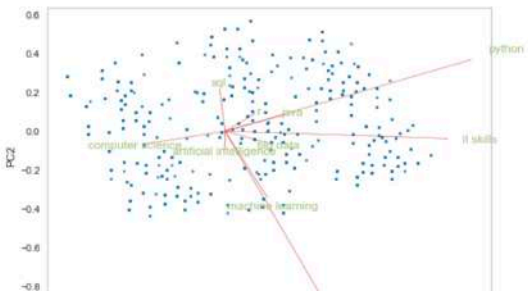
	Unnamed: 0	Unnamed: 0.1	roles	companies	locations	experience	skills	skill index	min_experience	max_experience	avg_experience
0	0	0	Data Scientist - Lead / Architect Wipro	Wipro	'Kochi', ' Kolkata', ' Pune', ' Gurgaon', ' Ch...	5-10	'data science', 'python', 'it skills', 'artifi...	-1.0	5	10	7.5
1	1	1	Urgent Requirement || Data Scientist || Noida	HCL	'Noida', 'Delhi'	3-8	'it skills', 'python', 'machine learning'	-1.0	3	8	5.5
2	2	2	Global Tax Automation & Operations - Data	Dell	'Bangalore'	3-5	'artificial intelligence', 'data science',	-1.0	3	5	4.0

Data Pre-processing

Biplot

```
In [19]: def myplot(score,coeff,labels=None):
             xs = score[:,0]
             ys = score[:,1]
             n = coeff.shape[0]
             scalex = 1.0/(xs.max() - xs.min())
             scaley = 1.0/(ys.max() - ys.min())
             plt.scatter(xs * scalex,ys * scaley,s=5)
             for i in range(n):
                 plt.arrow(0, 0, coeff[i,0], coeff[i,1],color = 'r',alpha = 0.5)
                 if labels is None:
                     plt.text(coeff[i,0]* 1.15, coeff[i,1] * 1.15, "Var"+str(i+1), color = 'green', ha = 'center', va = 'ce
                 else:
                     plt.text(coeff[i,0]* 1.15, coeff[i,1] * 1.15, labels[i], color = 'g', ha = 'center', va = 'center')

             plt.xlabel("PC{}".format(1))
             plt.ylabel("PC{}".format(2))
             plt.grid()
```

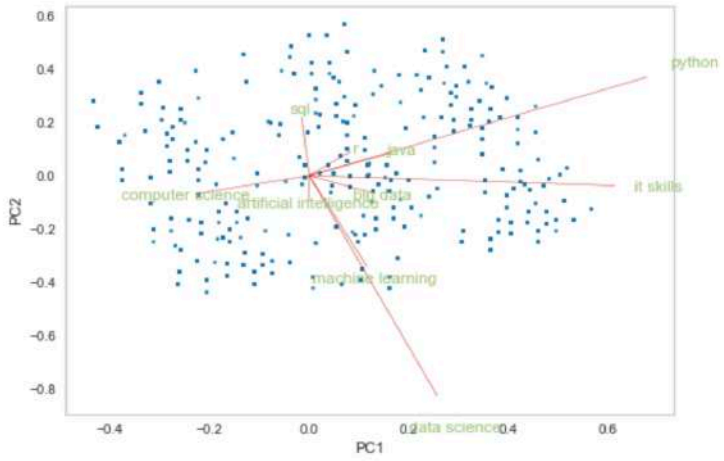
```
In [20]: myplot(x_pca[:,0:2],np.transpose(loadings[0:2, :]),list(encoded_data.columns))
         plt.show()
```



BIPLOT

```
    plt.xlabel("PC{}".format(1))
    plt.ylabel("PC{}".format(2))
    plt.grid()
```

```
In [20]: myplot(x_pca[:,0:2],np.transpose(loadings[0:2, :]),list(encoded_data.columns))
         plt.show()
```



BIPLOT

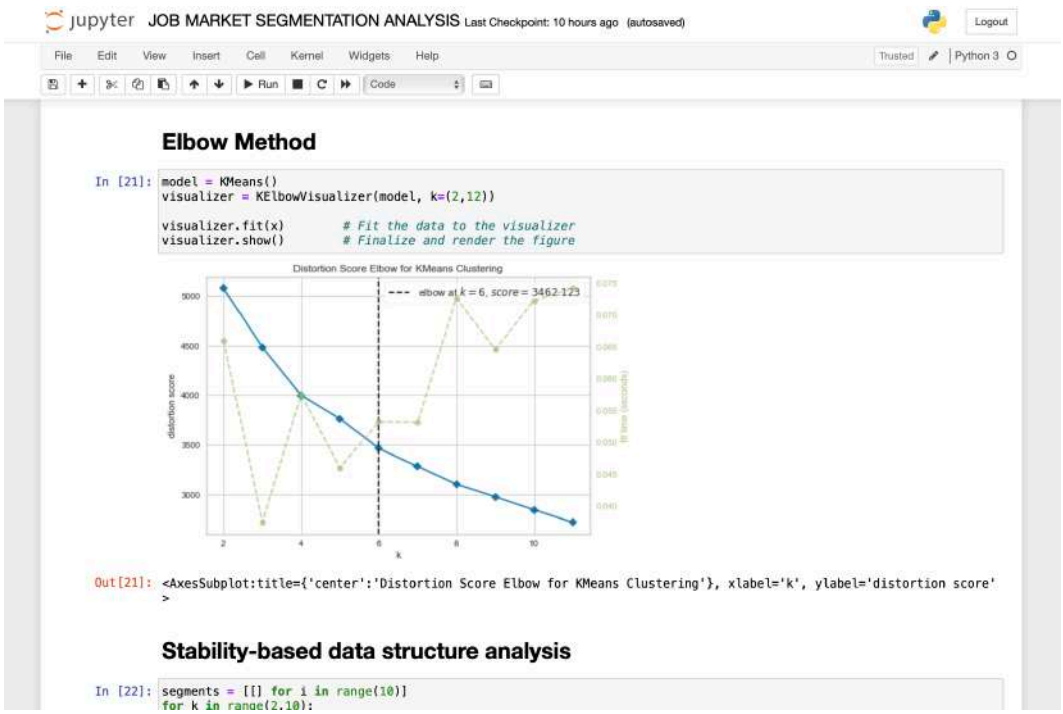
The PCA biplot represents variables with calibrated axes and data as points, allowing one to project the observations onto the axes to approximate the variables' original values.

CHAPTER 6: EXTRACTING AND DESCRIBING SEGMENTS

Elbow Method:

The Elbow method is a popular technique wherein we run k-means clustering for a range of clusters k (let's say from 1 to 10) and calculate the sum of squared distances from each point to its assigned centre for each value (distortions).

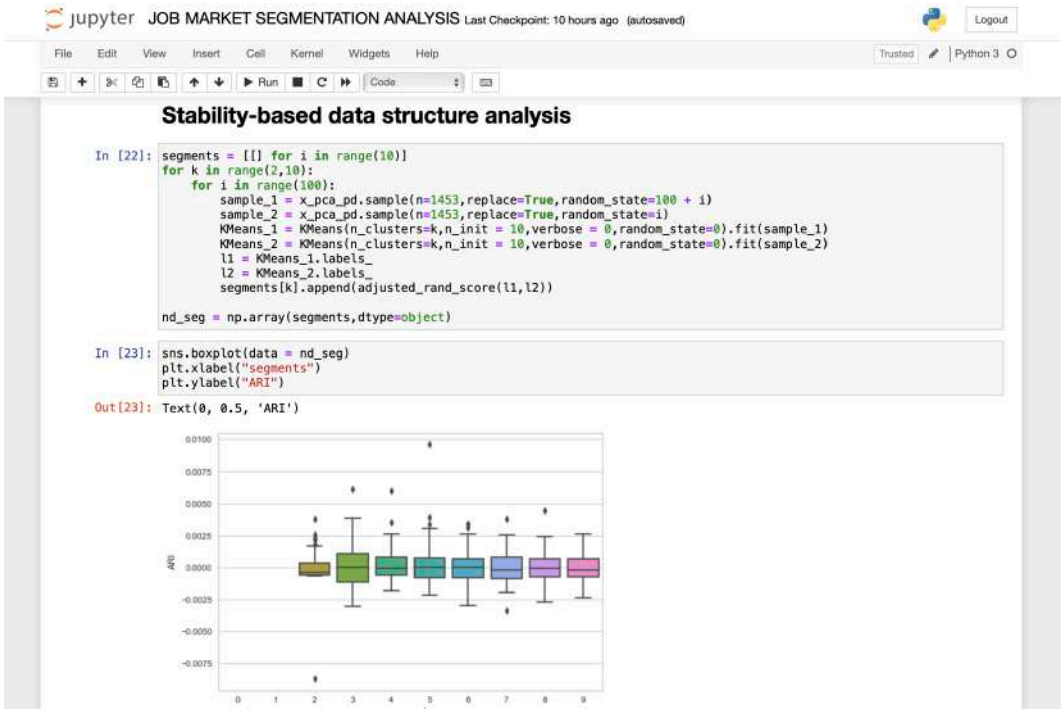
When the distortions are plotted and the plot resembles an arm, the optimal value of k is the "elbow" (the point of inflection on the curve).



Elbow Method

Data Structure Analysis:

- The term validation in the context of market segmentation is typically used in the sense of assessing reliability or stability of solutions across repeated calculations after slightly modifying the data, or the algorithm.
- Data structure analysis provides valuable insights into the properties of the data. These insights guide subsequent methodological decisions. Stability- based data structure analysis provides an indication of whether natural, distinct, and well-separated market segments exist in the data or not.
- If there is structure in the data, be it cluster structure or structure of a different kind, data structure analysis can also help to choose a suitable number of segments to extract.
- Segment Level Stability Analysis: Segment Level Stability Analysis is performed to protect against discarding solutions containing interesting individual segments from being prematurely discarded.



Stability Based Data Structure Analysis

CLUSTERING:

Clustering algorithms try to find natural clusters in data, the various aspects of how the algorithms to cluster data can be tuned and modified. Clustering is based on the principle that items within the same cluster must be

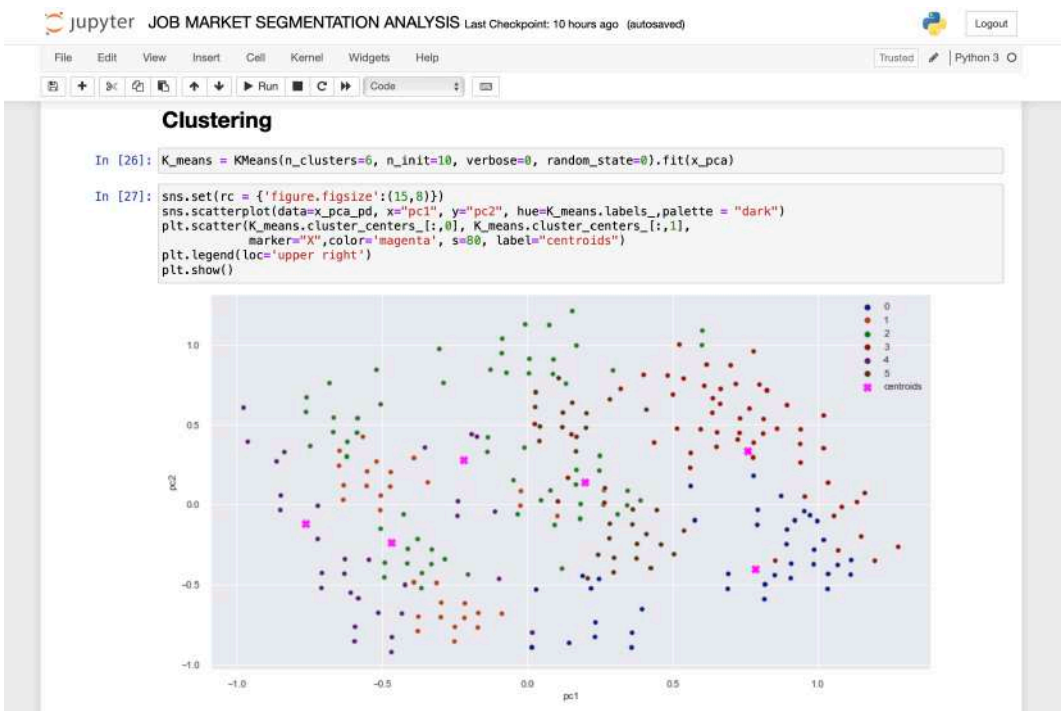
similar to each other. The data is grouped in such a way that related elements are close to each other.

K-Means Clustering :

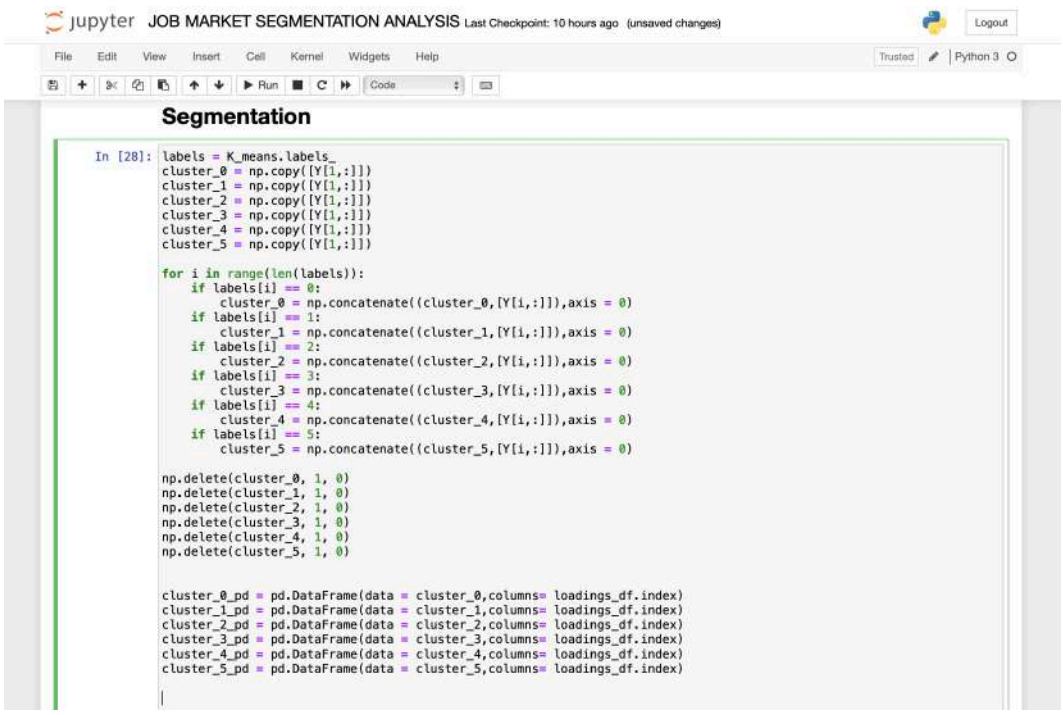
K-Means clustering is an unsupervised machine learning algorithm that divides the given data into the given number of clusters. Here, the “K” is the given number of predefined clusters, that need to be created.
It is a centroid based algorithm in which each cluster is associated with a centroid. The main idea is to reduce the distance between the data points and their respective cluster centroid.

The algorithm takes raw unlabelled data as an input and divides the dataset into clusters and the process is repeated until the best clusters are found.

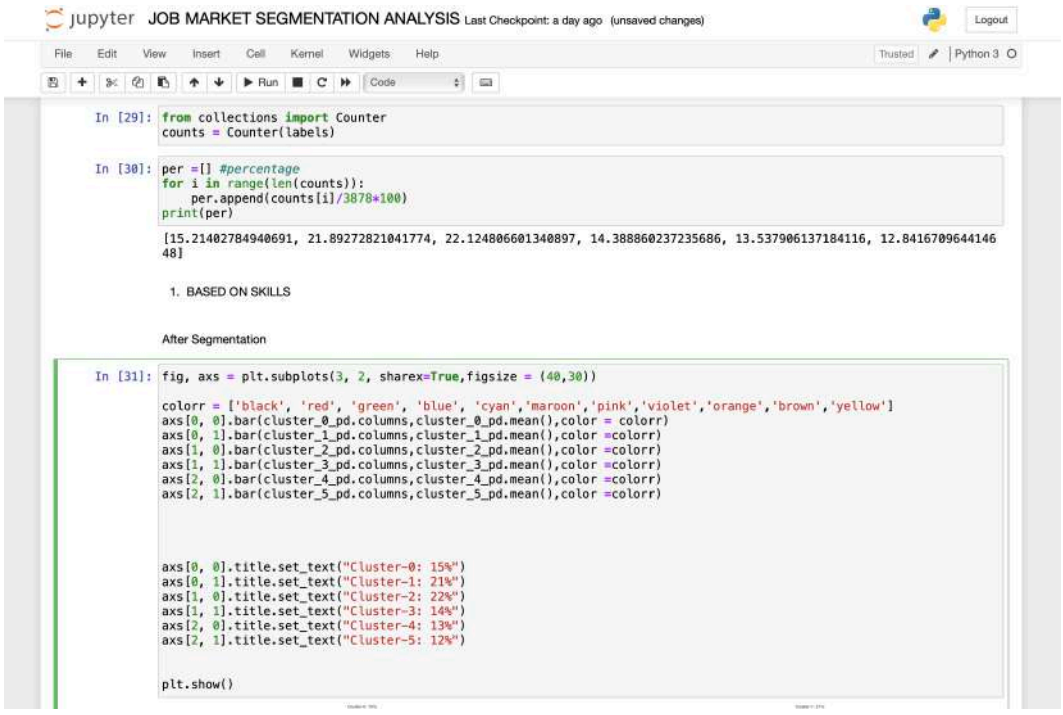
K-Means is very easy and simple to implement. It is highly scalable, can be applied to both small and large datasets. There is, however, a problem with choosing the number of clusters or K. Also, with the increase in dimensions, stability decreases. But, overall K Means is a simple and robust algorithm that makes clustering very easy.



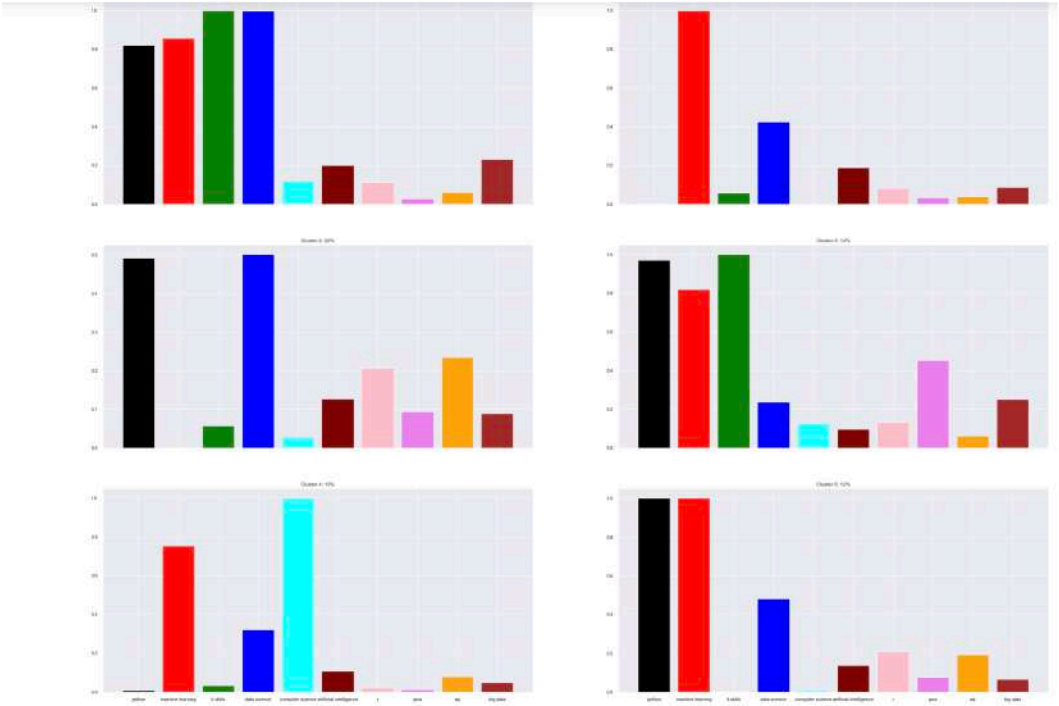
Caption



Segmentation



Based on skills



Profiling Segments

2. Based on Companies

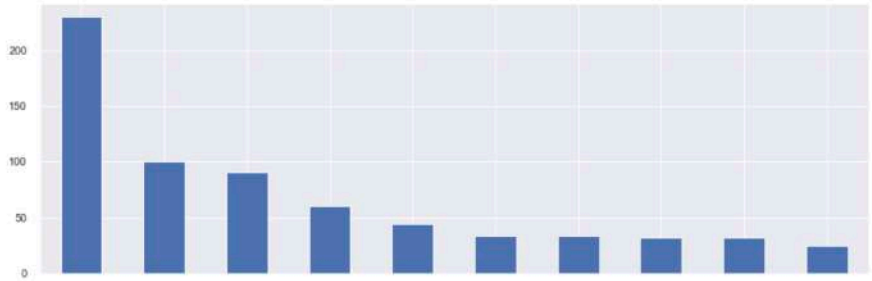
```
In [32]: data3['companies'].value_counts().head(10)
```

```
Out[32]: Huquo Consulting Pvt. Ltd    230
         IBM                         100
         Accenture                    91
         Microsoft                    60
         CarbyneTech India            44
         Diverse Lynx                 34
         Career Maker                 34
         Walmart                      32
         First Employer               32
         Gojek                        25
         Name: companies, dtype: int64
```

Before Segmentation

```
In [33]: f,ax=plt.subplots(figsize=(15,5))
         data3['companies'].value_counts().head(10).plot(kind = 'bar')
```

```
Out[33]: <AxesSubplot:>
```

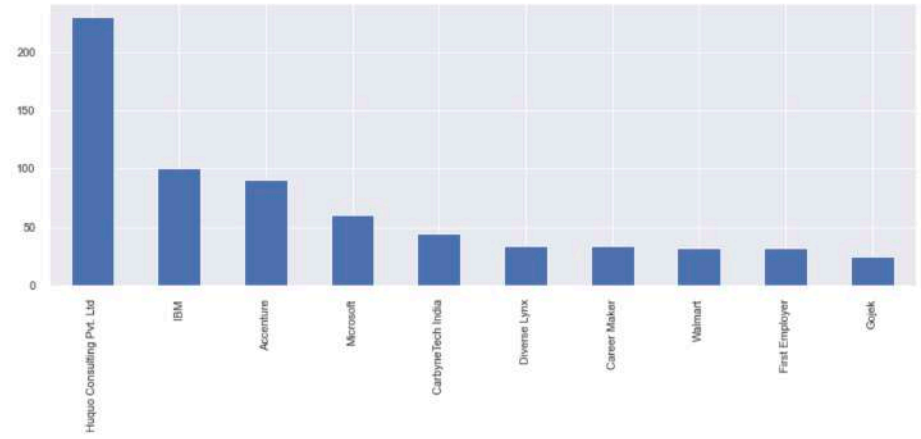


Based on Companies

Before Segmentation

```
In [33]: f,ax=plt.subplots(figsize=(15,5))
         data3['companies'].value_counts().head(10).plot(kind = 'bar')
```

```
Out[33]: <AxesSubplot:>
```



After Segmentation

Cluster 0

Before Segmentation

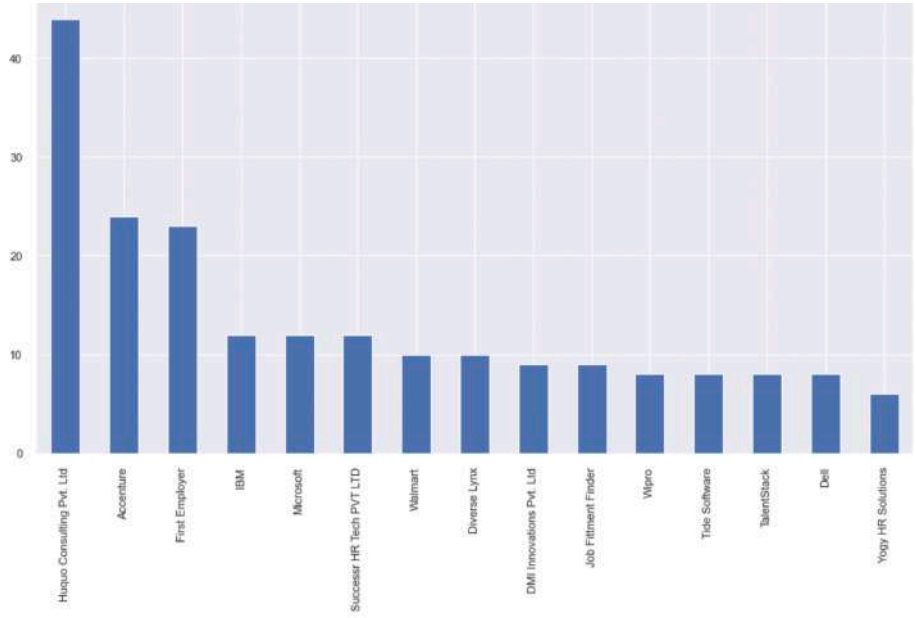
After Segmentation

Cluster 0

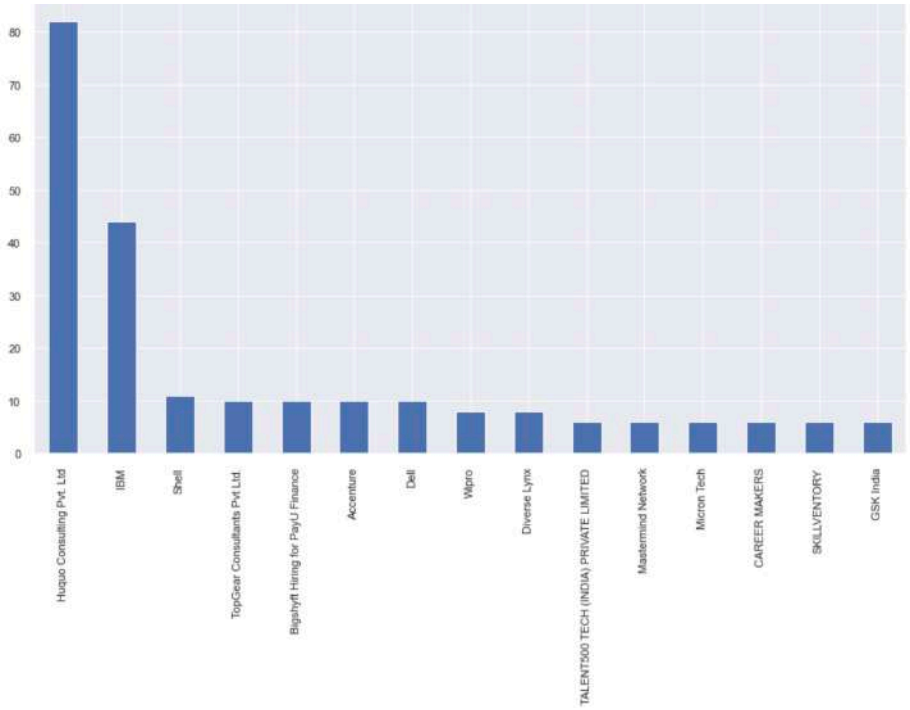
```
In [34]: data4['labels'] = labels
         desired_cluster  = data4[data4['labels']==0]
         o = desired_cluster['companies'].value_counts()
         o[:15].plot.bar()
```

```
Out[34]: <AxesSubplot:>
```

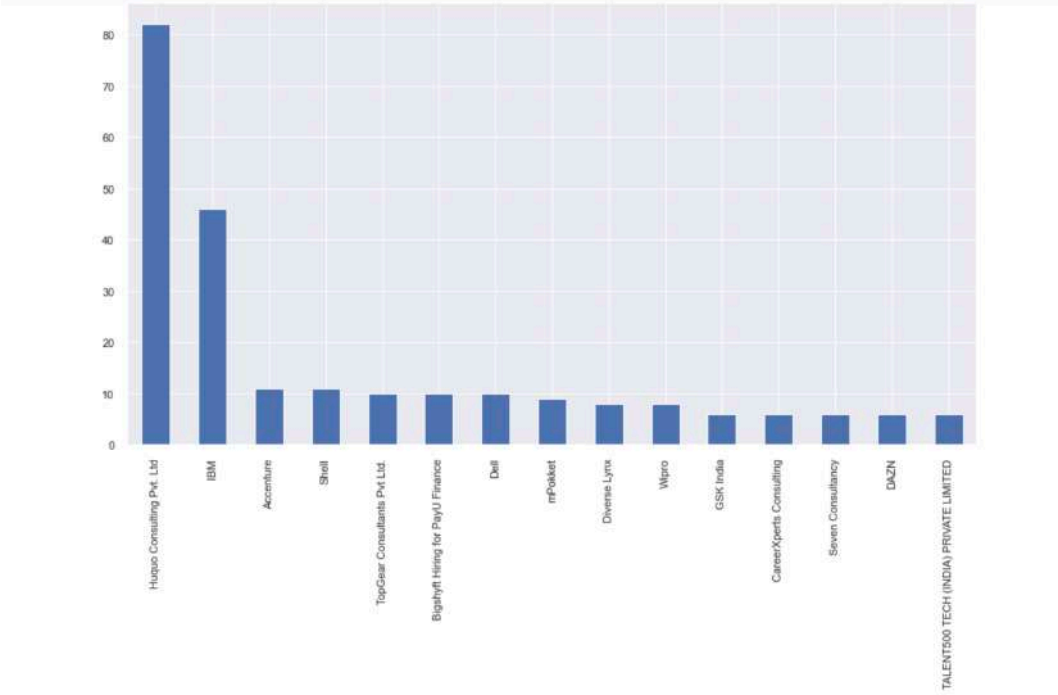

After Segmentation



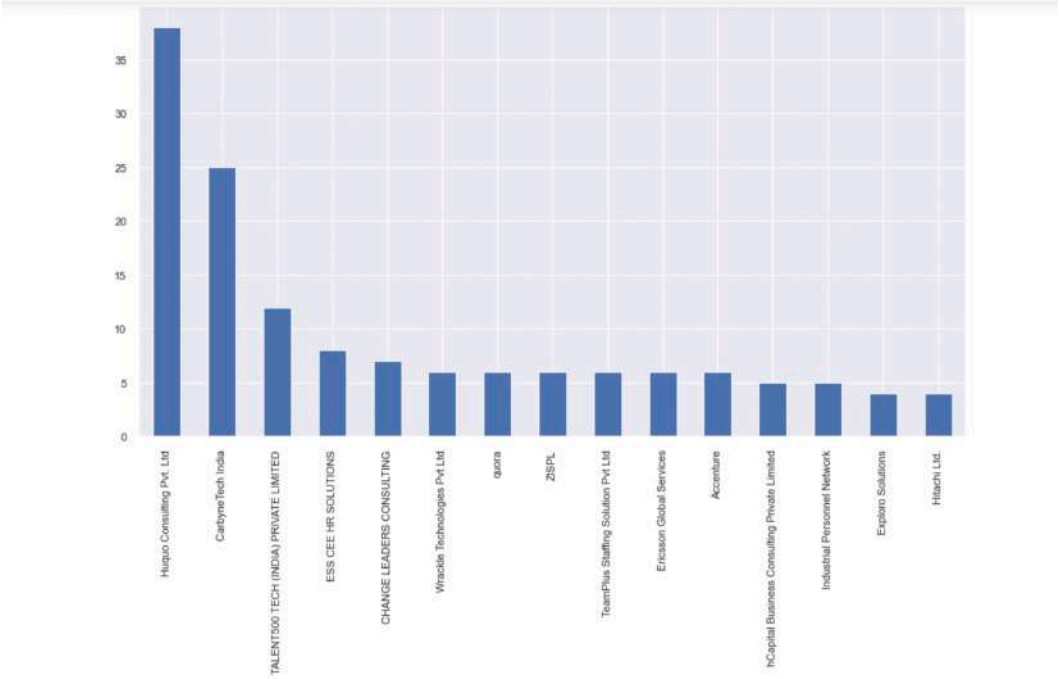
Cluster 0



Cluster 2



Cluster 3



Cluster 5

For the company's analysis based on experience demanded, it was observed that Wipro, GlobalLogic and Gojek didn't appear in top numbers before the segmentation and appeared after the segmentation was carried out for the minimum and average experience data.

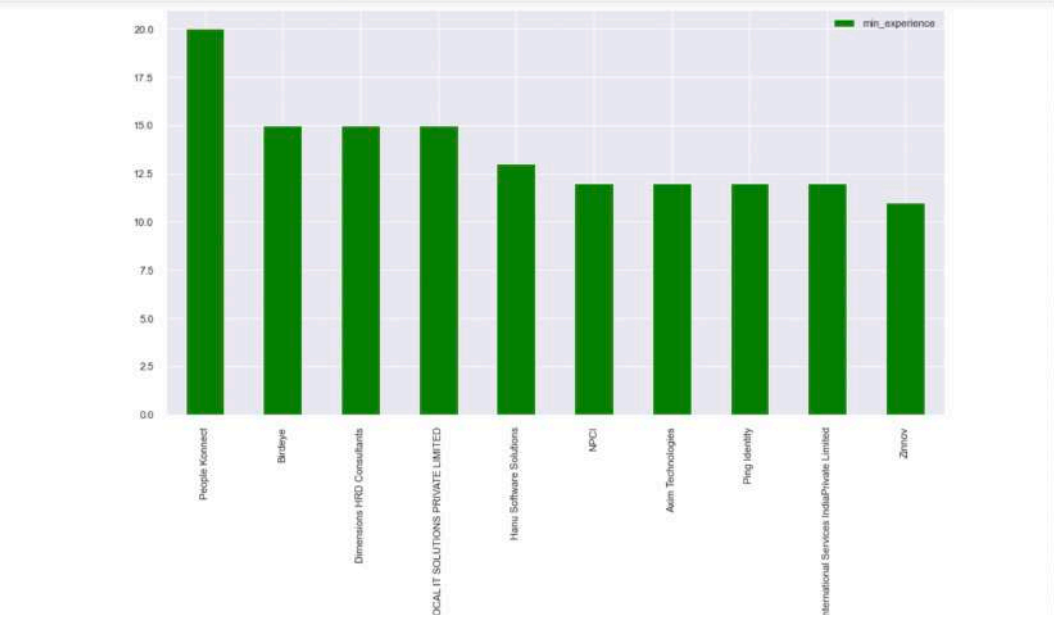
Target segment in job market based on the experience in totality would be the top 10 companies as shown in the above plot.

3. Based on Experience

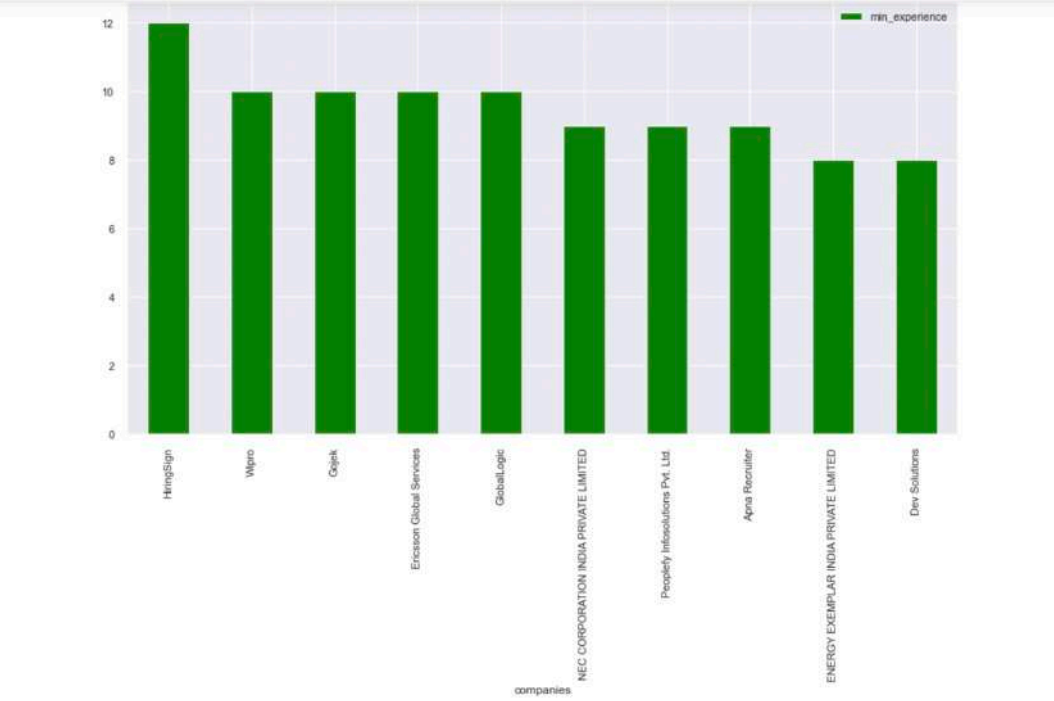
Before Segmentation

```
In [66]: data4[['min_experience','companies']].groupby(["companies"]).median().sort_values(by='min_experience',ascending=Fals
Out[66]: <AxesSubplot:xlabel='companies'>
```

Based on Experience



Top 10 companies with minimum experience before Segmentation



Top 10 companies with minimum experience after Segmentation

DESCRIBING POTENTIAL SEGMENTS :

Cluster 0: Cluster 0 contains companies which are inclined towards hiring people with Python skills on Data Science and Machine Learning.

Cluster 1: Cluster 1 contains companies which are likely to hire people with skills are not oriented towards Data Analysis.

Cluster 2: Cluster 2 contains companies which are inclined towards hiring people with Python and R skills on Data Science.

Cluster 3: Cluster 3 contains companies which are inclined towards hiring people with Python skills on Machine Learning.

Cluster 4: Cluster 4 contains companies which are likely to hire people with skills are not oriented towards Data Analysis.

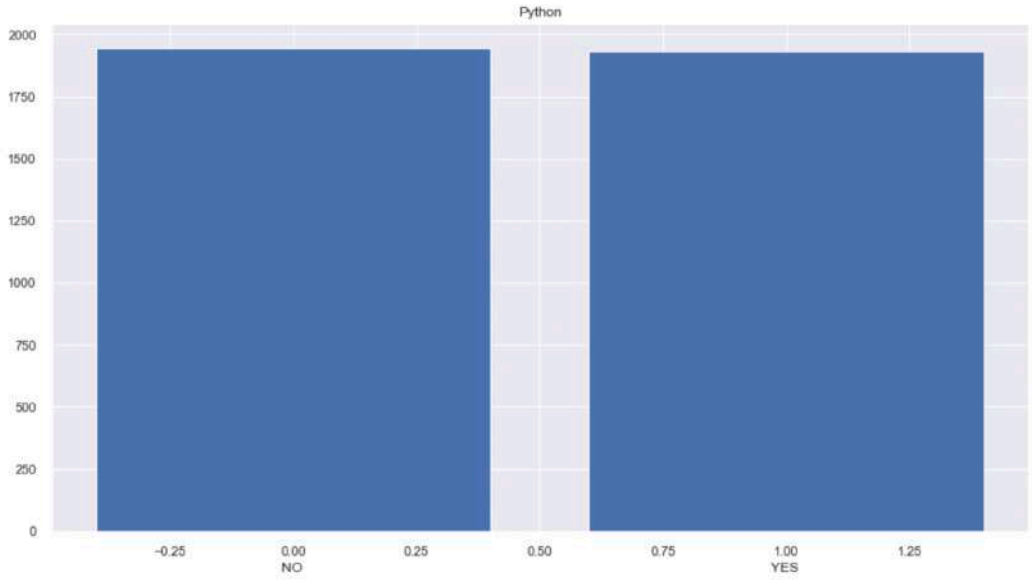
Cluster 5: Cluster 5 contains companies which are likely to hire people with skills of Python, Machine Learning and minimal Data Science.

The most demanded skills for the recruiters are Python, Data Science, Machine Learning and other IT skills.

From our dataset we selected these numbers of particular skills and visualised them as follows:

Python

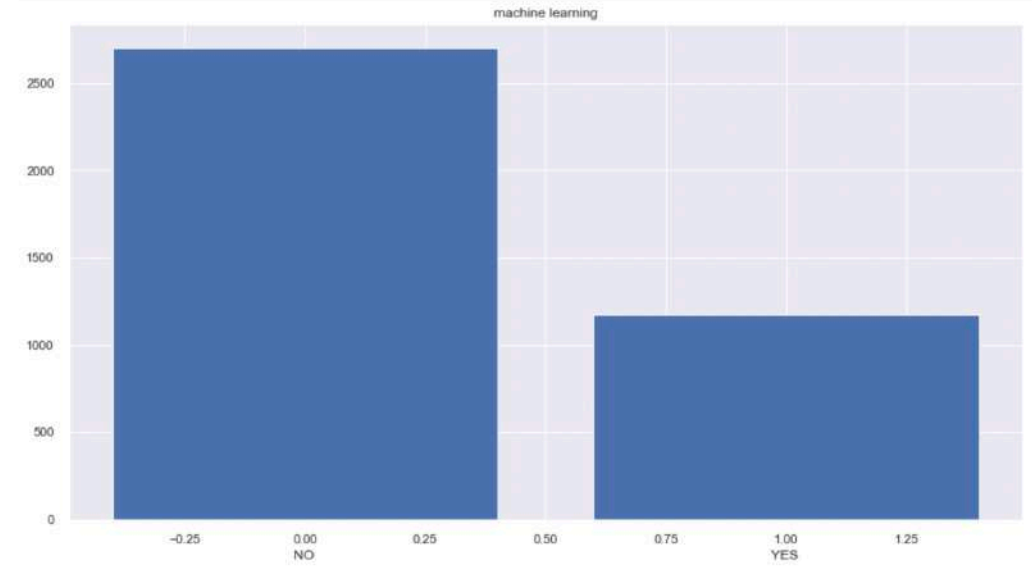
```
In [48]: skills =list(encoded_data.columns)
         skills_dict = dict.fromkeys(skills)
         skill_count =[]
         for i in range(10):
             skill_count.append(list(encoded_data.iloc[:,i].value_counts()))
         incr = 0
         for keys in skills_dict.keys():
             skills_dict[keys] = skill_count[incr]
             incr+=1
         print(skills_dict)

         {'python': [1947, 1931], 'machine learning': [2704, 1174], 'it skills': [2618, 1260], 'data science': [1955, 1923],
         'computer science': [3184, 694], 'artificial intelligence': [3313, 565], 'r': [3380, 498], 'java': [3455, 423], 'sq
         l': [3436, 442], 'big data': [3395, 483]}

In [49]: plt.bar(data='python',height=skills_dict['python'],x = [0,1])
         plt.title('Python')
         plt.xlabel('NO
         plt.show()
```



Python

Machine Learning

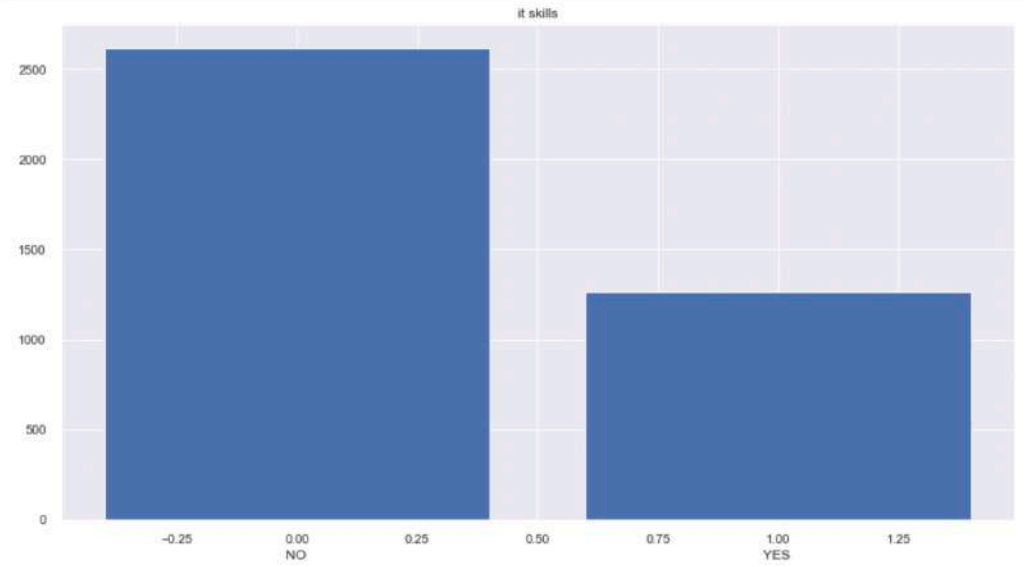
```
0]: plt.bar(data='machine learning',height=skills_dict['machine learning'],x = [0,1])
    plt.title('machine learning')
    plt.xlabel('NO
    plt.show()
```



Machine Learning

IT Skills

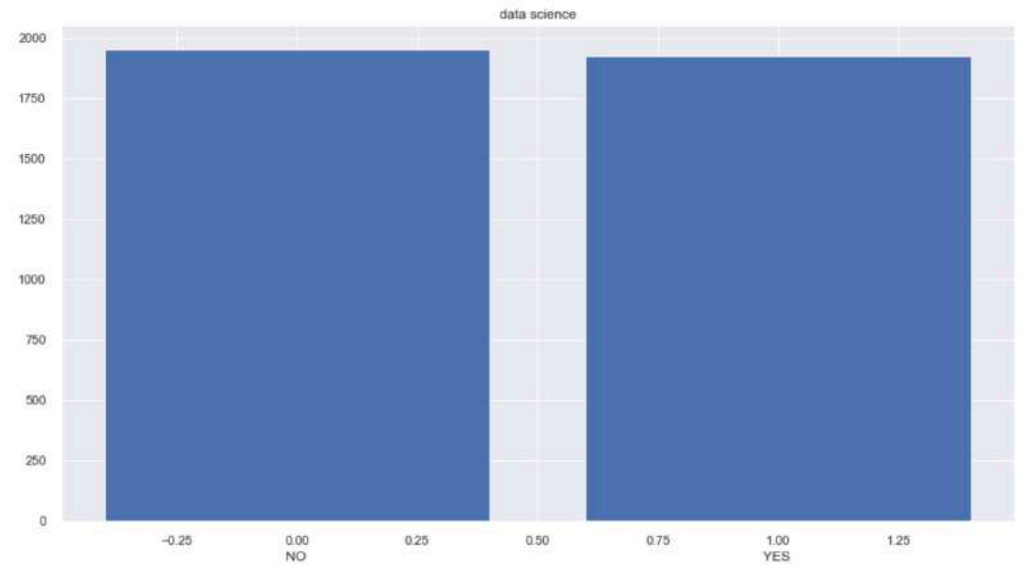
```
1]: plt.bar(data='it skills',height=skills_dict['it skills'],x = [0,1])
    plt.title('it skills')
    plt.xlabel('NO
    plt.show()
```



IT Skills

Data Science

```
2]: plt.bar(data='data science',height=skills_dict['data science'],x = [0,1])
    plt.title('data science')
    plt.xlabel('NO
    plt.show()
```



Data Science

CHAPTER 7: CONCLUSION

We conducted job market segmentation using data from Naukri website (record size: 3878) and discovered that By analyzing the trend, we have observed cluster 0 contains companies which are inclined towards hiring people with Python skills on Data Science and Machine Learning. Cluster 1 contains companies which are likely to hire people with skills are not oriented towards Data Analysis. Cluster 2 contains companies which are inclined towards hiring people with Python and R skills on Data Science. Cluster 3 contains companies which are inclined towards hiring people with Python skills on Machine Learning. Cluster 4 contains companies which are likely to hire people with skills are not oriented towards Data Analysis. Cluster 5 contains companies which are likely to hire people with skills of Python, Machine Learning and minimal Data Science. The most demanded skills for the recruiters are Python, Data Science, Machine Learning and other IT skills.

For the company's analysis based on experience demanded, it was observed that Wipro, HiringSign, Global Logic and Gojek etc. didn't appear in top numbers before the segmentation and appeared after the segmentation was carried out for the minimum, average and maximum experience data.

Overall, the cluster analysis we conducted would be a wonderful tool for both individuals and employers to locate the needed experience and forthcoming expertise in the Data Science/ML job market, and it would make it easier for any tech company to find a suitable applicant.

BIBLIOGRAPHY

GITHUB Link :

https://github.com/akashjborah97/Machine-Learning-Job-Market-Segmentation-Analysis

BOOKS:

1. Market Segmentation Analysis by Sara Dolnicar, Bettina Grun and Friedrich Reisch

REFERENCES:

1. https://marutitech.com/artificial-intelligence-and-machine-learning
2. https://builtin.com/data-science/step-step-explanation-principal-component-analysis
3. https://en.wikipedia.org/wiki/Fermi_problem