JOB MARKET SEGMENT ANALYSIS (ML Engineer/Data Analyst)

Team-Sonu

Sonu Kumar (Team Lead)
Akash Pratap Singh
Habiba Anjum
Tanmay Sandeep Pawar
Chandru J
C S Ayush Kumar

16th May 2022

1 Introduction

The scope of data analyst and machine learning in India is massive. Since most businesses, irrespective of their size, are wanting to be a part of this revolution, there are a lot of new vacancies opening for the right talent. listed sectors where maximum hiring of machine learning professionals is happening:

- BFSI – 38%
- E-commerce – 11%
- Energy – 13%
- Media – 7%
- Healthcare and pharmacy- 12%
- Retail – 6%

As per the Gartner report, only 75% of the registered business organizations are planning to or have already invested in machine learning and data analyst.

To keep up with technological innovations, most of the big business organizations are hiring employees that are skilled in these technical aspects. The result is a lucrative pay package and an exponential rise in career.

As the industry is still growing, you will soon get access to more opportunities. Also, the fact that machine learning and data analyst are so intertwined makes it possible for professionals to switch job positions if they desire.

2 Data Sources

Dataset : [https://www.kaggle.com/datasets/usamakhan8199/current-data-science-jobsin-india-by-glassdoor]

This Dataset consists list of AI Jobs in india .

This dataset is scraped from glassdoor .

Job titles which are interchangeably used ('Data Scientist', 'Data Engineer' ,'Machine Learning Engineer', 'Data Analyst').

Number of rows in data - 885

Number of columns /attribute in data - 19

2.1 Attribute Information:

- 'Job.Title' : Job title
- 'Job.Description': Job description
- 'Rating': Job rating
- 'Company.Name': Company name
- 'Location': Location of the company
- 'Headquarters': Company headquarters
- 'Size': the Size of the company
- 'Founded': The year it was founded
- 'Type.of.ownership': Public or private ownership
- 'Industry': Type of industry company belongs
- 'Sector': Type of sector company belongs
- 'Revenue': Yearly revenue by company
- 'Competitors': Company competitors
- 'Python': Does description has python keyword in it 1 if yes 0 if no
- 'R.Prog': Does description has R prog keyword in it 1 if yes 0 if no
- 'Excel': Does description has Excel keyword in it 1 if yes 0 if no
- 'Hadoop': Does description has Hadoop keyword in it 1 if yes 0 if no
- 'SQL': Does description has SQL keyword in it 1 if yes 0 if no
- 'SAS: Does description has SAS keyword in it 1 if yes 0 if no

3 Problem Statement:

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

4 Data Pre-Processing (Steps and Libraries Used)

Importing Libraries: firstly, we will import the libraries for our model, which is part of data pre-processing. The code is given below:

```
import pandas as pd
import numpy as np
from collections import Counter
import matplotlib.pyplot as plt
import warnings
import seaborn as sns
```

- Pandas we have imported to manage the dataset.
- Numpy imported to perform maths operation on top of our dataset.
- Counter Imported to compute frequency of required data.
- Matplotlib is for plotting the graphs.
- Seaborn is for data visualization internally is uses matplotlib.

5 Segment Extraction (Techniques used)

5.1 We have Used Basic Maths to solve our problem:

Example code is given below:

```
INPUT:

print(data['Location'].describe())

print('Unique locations are')

print(data['Location'].unique())


OUTPUT:


count 885

unique 32

top Bengaluru
```

```
freq 393
Name: Location, dtype: object
```

Describe() function gives the basic stats for given attributes:

- No of counts of the data that it is present in the rows.
- Unique times that the data is occurring.
- Top frequent occurring data among the selected data.

5.1.1 find the 10 most frequent Company.Name.

INPUT:

```
Company_Name_count = Counter(list(data['Company.Name']))
Company_Name_count.most_common(10)
```

OUTPUT:

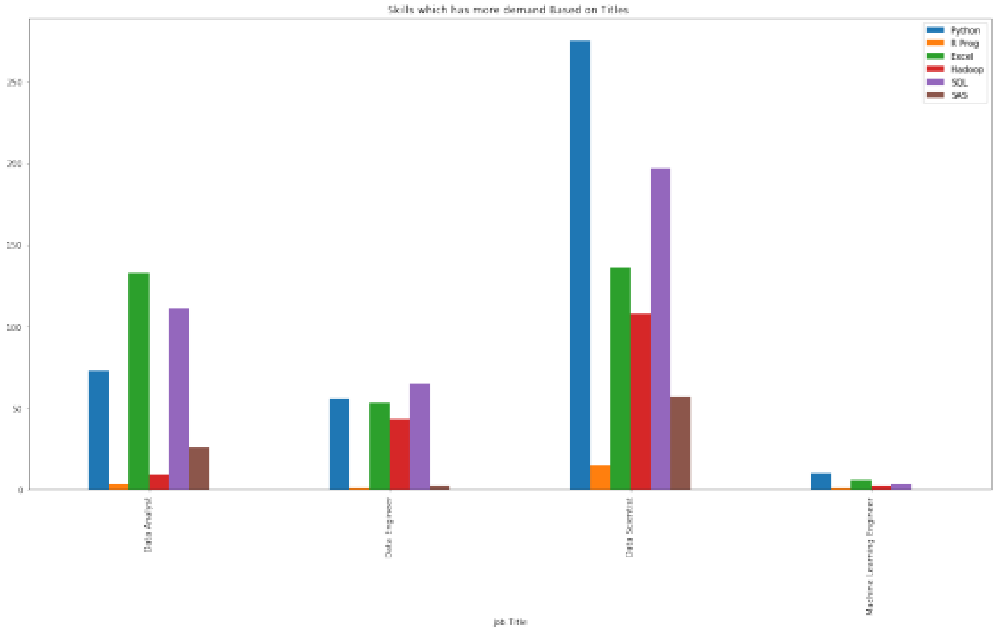
```
[('ZoomRx', 22),
('Amazon', 12),
('EY', 12),
('Sanofi', 12),
('Walmart', 12),
('Quantzig', 12),
('Citi', 12),
('String Bio', 12),
('Matelabs Innovations Pvt. Ltd.', 12),
('LogisticsNow', 11)]
```

COUNTER Function returns the frequency of data.

5.2 Graph is Used to get the results:

EXAMPLE:

5.2.1 Most Demanded Skills



Conclusions

1. Python is top most demanded skill is required for data scientist job.

2. for data Analyst Excel ,python is most demanded skills required

3. For Machine Learning Engineer most demanded skill is Python

4. For Data Engineer most demanded skills are SQL, python, excel.

6 Exploratory Data Analysis:

We start with Exploratory Data Analysis using stats on each attribute to get the intuitive understanding of attributes:

Exploratory data analysis (EDA) involves using graphics and visualizations to explore and analyze a data set.

Exploratory data analysis is a powerful way to explore a data set. Even when your goal is to perform planned analyses, EDA can be used for data cleaning, for subgroup analyses or simply for understanding your data better. An important initial step in any data analysis is to plot the data.

Exploratory data analysis is an investigative process in which you use summary statistics and graphical tools to get to know your data and understand what you can learn from them.

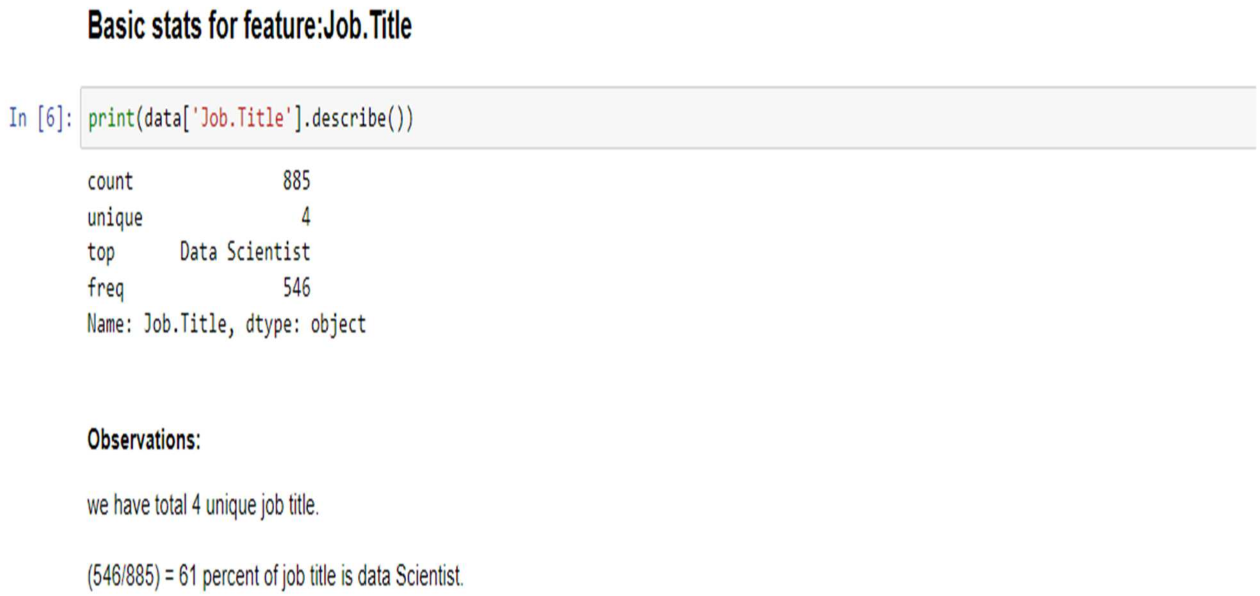
With EDA, you can find anomalies in your data, such as outliers or unusual observations, uncover patterns, understand potential relationships among variables, and generate interesting questions or hypotheses that you can test later using more formal statistical methods.

6.1 Typical goals are understanding:

- The distribution of variables in your data set. That is, what is the shape of your data? Is the distribution skewed?
- The relationships between variables.
- Whether or not your data have outliers or unusual points that may indicate data quality issues or lead to interesting insights.
- Whether or not your data have patterns over time.

6.2 UNDERSTANDING THE VARIOUS FEATURES USED IN DATASET:

6.2.1 Basic stats for feature:Job.Title



Basic stats for feature:Job.Title

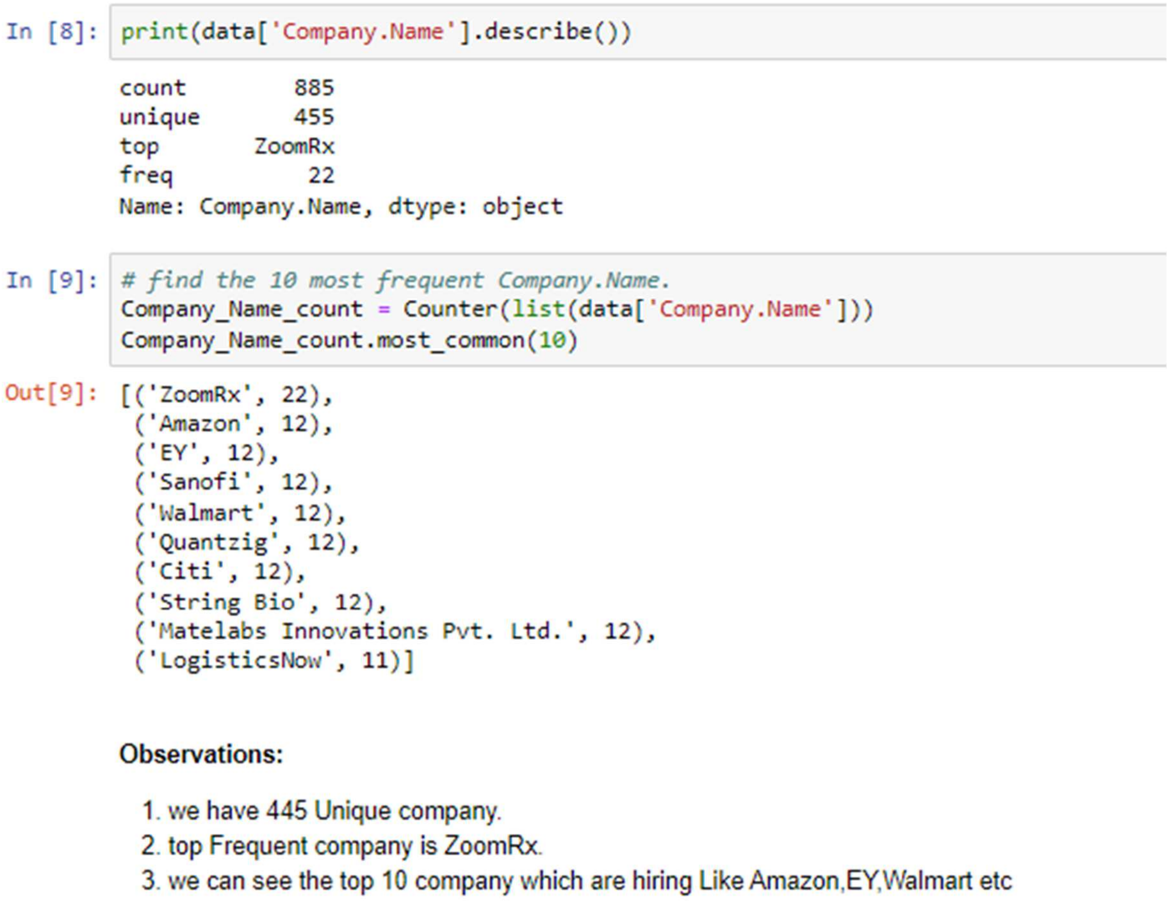
```
In [6]: print(data['Job.Title'].describe())

count                885
unique                 4
top       Data Scientist
freq                 546
Name: Job.Title, dtype: object
```

Observations:

we have total 4 unique job title.

(546/885) = 61 percent of job title is data Scientist.

6.2.2 Basic stats for feature:Company.Name



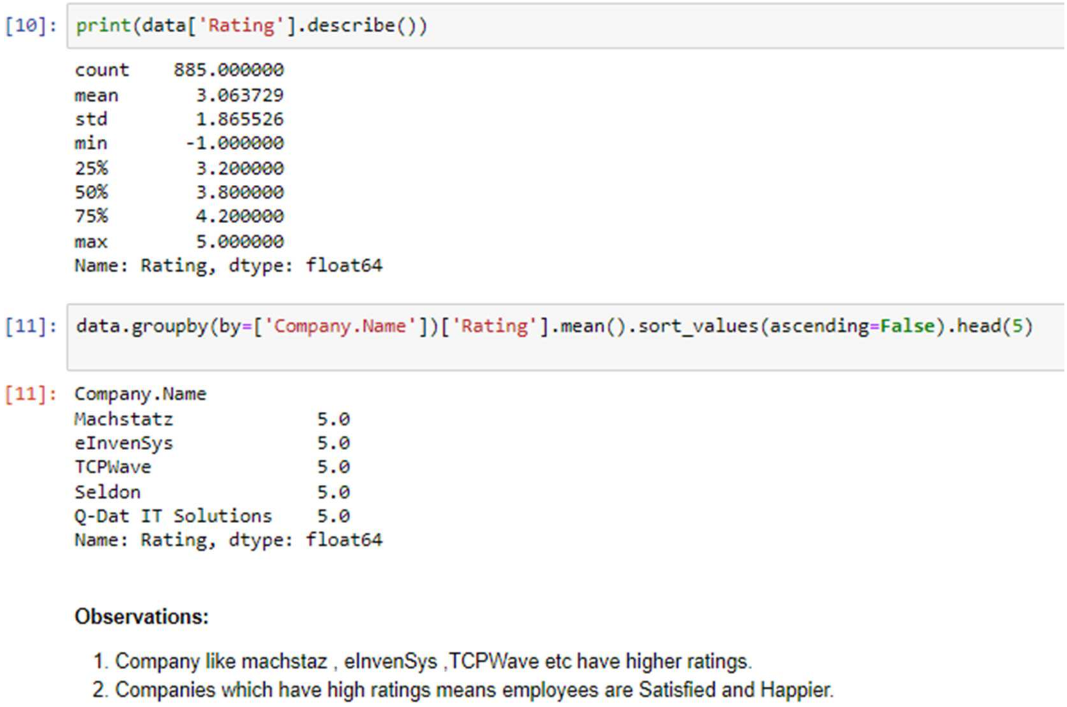
```
In [8]: print(data['Company.Name'].describe())

        count        885
        unique       455
        top       ZoomRx
        freq          22
        Name: Company.Name, dtype: object

In [9]: # find the 10 most frequent Company.Name.
        Company_Name_count = Counter(list(data['Company.Name']))
        Company_Name_count.most_common(10)

Out[9]: [('ZoomRx', 22),
         ('Amazon', 12),
         ('EY', 12),
         ('Sanofi', 12),
         ('Walmart', 12),
         ('Quantzig', 12),
         ('Citi', 12),
         ('String Bio', 12),
         ('Matelabs Innovations Pvt. Ltd.', 12),
         ('LogisticsNow', 11)]
```

Observations:

1. we have 445 Unique company.
2. top Frequent company is ZoomRx.
3. we can see the top 10 company which are hiring Like Amazon,EY,Walmart etc

6.2.3

6.2.4 Basic stats for feature: Rating



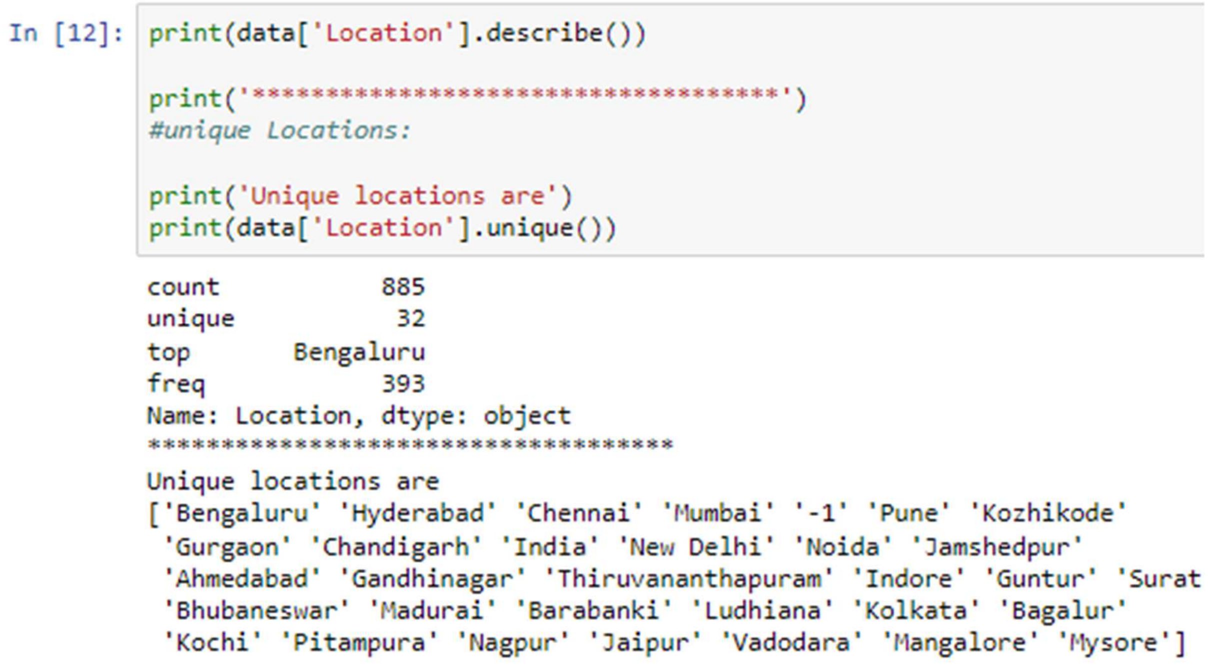
```
[10]: print(data['Rating'].describe())

      count    885.000000
      mean       3.063729
      std        1.865526
      min       -1.000000
      25%        3.200000
      50%        3.800000
      75%        4.200000
      max        5.000000
      Name: Rating, dtype: float64

[11]: data.groupby(by=['Company.Name'])['Rating'].mean().sort_values(ascending=False).head(5)

[11]: Company.Name
      Machstatz             5.0
      eInvenSys             5.0
      TCPWave               5.0
      Seldon                5.0
      Q-Dat IT Solutions    5.0
      Name: Rating, dtype: float64
```

Observations:

1. Company like machstaz , eInvenSys ,TCPWave etc have higher ratings.
2. Companies which have high ratings means employees are Satisfied and Happier.

6.2.5

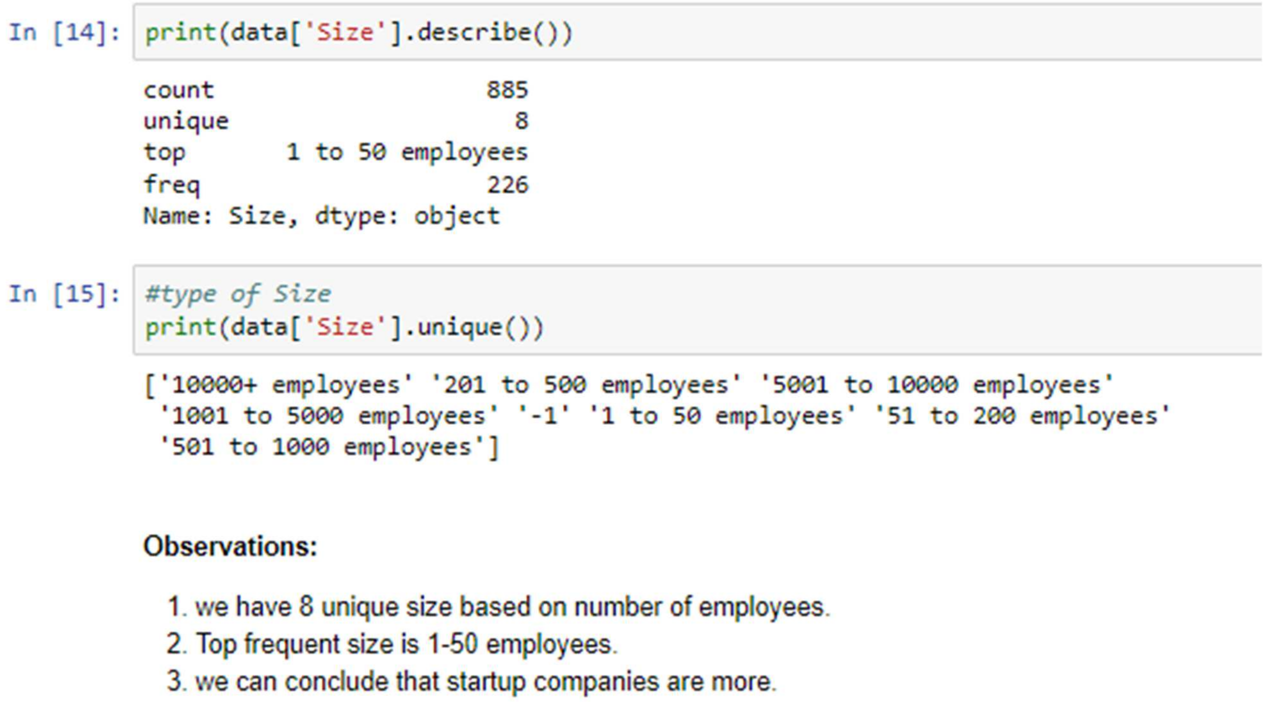
6.2.6 Basic stats for feature: Location



```
In [12]: print(data['Location'].describe())

         print('************************************')
         #unique Locations:

         print('Unique locations are')
         print(data['Location'].unique())
```

```
count            885
unique            32
top        Bengaluru
freq             393
Name: Location, dtype: object
************************************
Unique locations are
['Bengaluru' 'Hyderabad' 'Chennai' 'Mumbai' '-1' 'Pune' 'Kozhikode'
 'Gurgaon' 'Chandigarh' 'India' 'New Delhi' 'Noida' 'Jamshedpur'
 'Ahmedabad' 'Gandhinagar' 'Thiruvananthapuram' 'Indore' 'Guntur' 'Surat
 'Bhubaneswar' 'Madurai' 'Barabanki' 'Ludhiana' 'Kolkata' 'Bagalur'
 'Kochi' 'Pitampura' 'Nagpur' 'Jaipur' 'Vadodara' 'Mangalore' 'Mysore']
```

6.2.7 Basic stats for feature: Size



```
In [14]: print(data['Size'].describe())
```

```
count                    885
unique                     8
top       1 to 50 employees
freq                     226
Name: Size, dtype: object
```

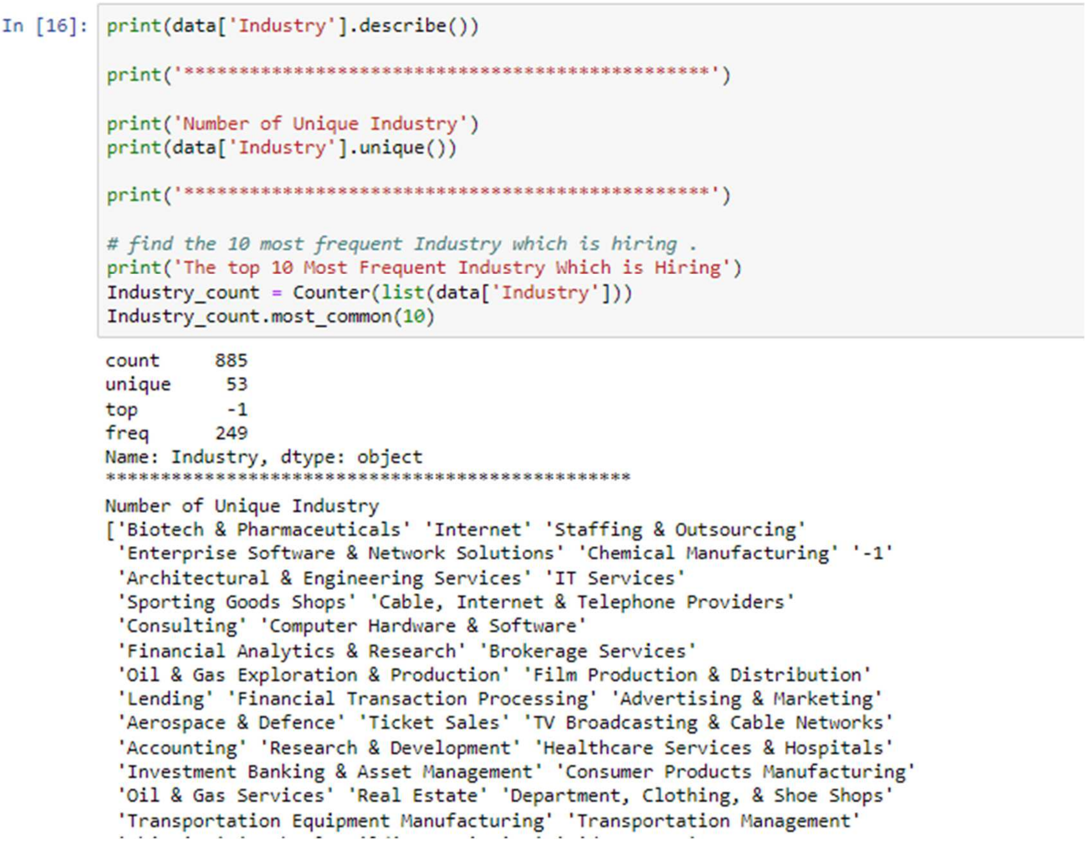
```
In [15]: #type of Size
         print(data['Size'].unique())
```

```
['10000+ employees' '201 to 500 employees' '5001 to 10000 employees'
 '1001 to 5000 employees' '-1' '1 to 50 employees' '51 to 200 employees'
 '501 to 1000 employees']
```

Observations:

1. we have 8 unique size based on number of employees.
2. Top frequent size is 1-50 employees.
3. we can conclude that startup companies are more.

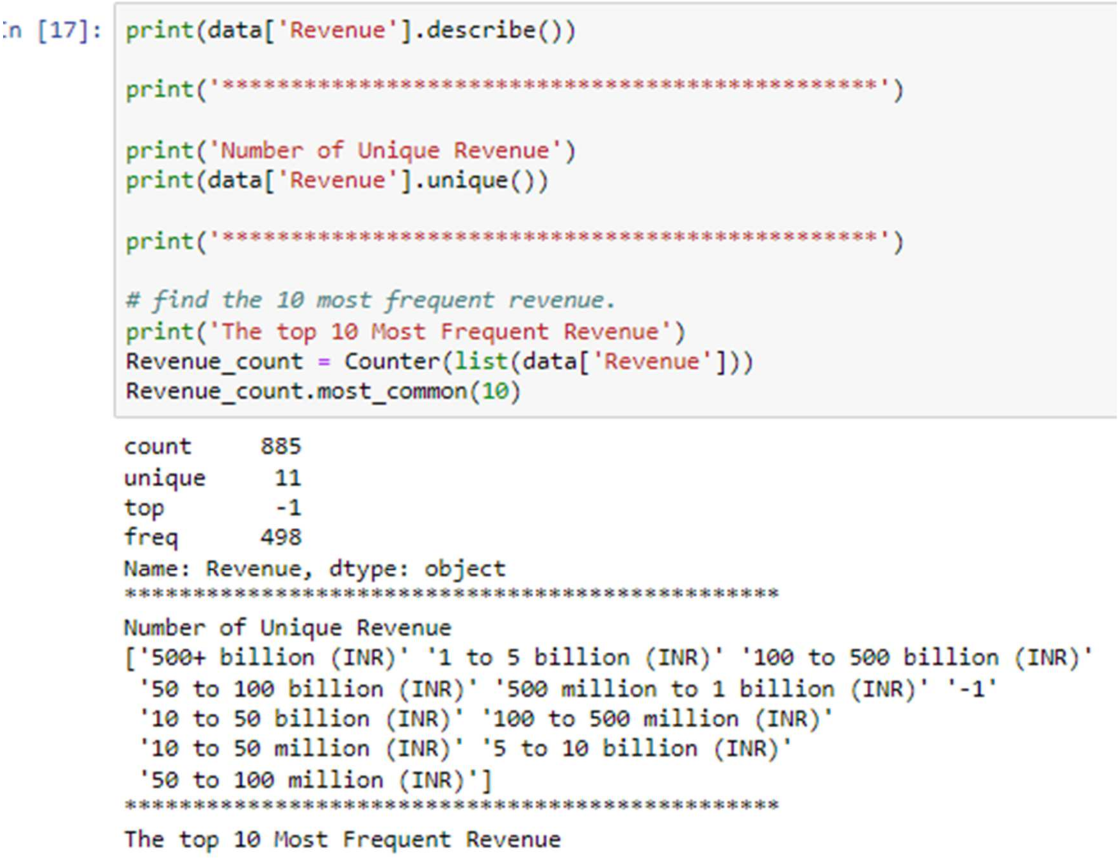
6.2.8 Basic stats for feature: Industry



```
In [16]: print(data['Industry'].describe())

         print('************************************************')

         print('Number of Unique Industry')
         print(data['Industry'].unique())

         print('************************************************')

         # find the 10 most frequent Industry which is hiring .
         print('The top 10 Most Frequent Industry Which is Hiring')
         Industry_count = Counter(list(data['Industry']))
         Industry_count.most_common(10)
```

```
count      885
unique      53
top         -1
freq       249
Name: Industry, dtype: object
************************************************
Number of Unique Industry
['Biotech & Pharmaceuticals' 'Internet' 'Staffing & Outsourcing'
 'Enterprise Software & Network Solutions' 'Chemical Manufacturing' '-1'
 'Architectural & Engineering Services' 'IT Services'
 'Sporting Goods Shops' 'Cable, Internet & Telephone Providers'
 'Consulting' 'Computer Hardware & Software'
 'Financial Analytics & Research' 'Brokerage Services'
 'Oil & Gas Exploration & Production' 'Film Production & Distribution'
 'Lending' 'Financial Transaction Processing' 'Advertising & Marketing'
 'Aerospace & Defence' 'Ticket Sales' 'TV Broadcasting & Cable Networks'
 'Accounting' 'Research & Development' 'Healthcare Services & Hospitals'
 'Investment Banking & Asset Management' 'Consumer Products Manufacturing'
 'Oil & Gas Services' 'Real Estate' 'Department, Clothing, & Shoe Shops'
 'Transportation Equipment Manufacturing' 'Transportation Management'
```

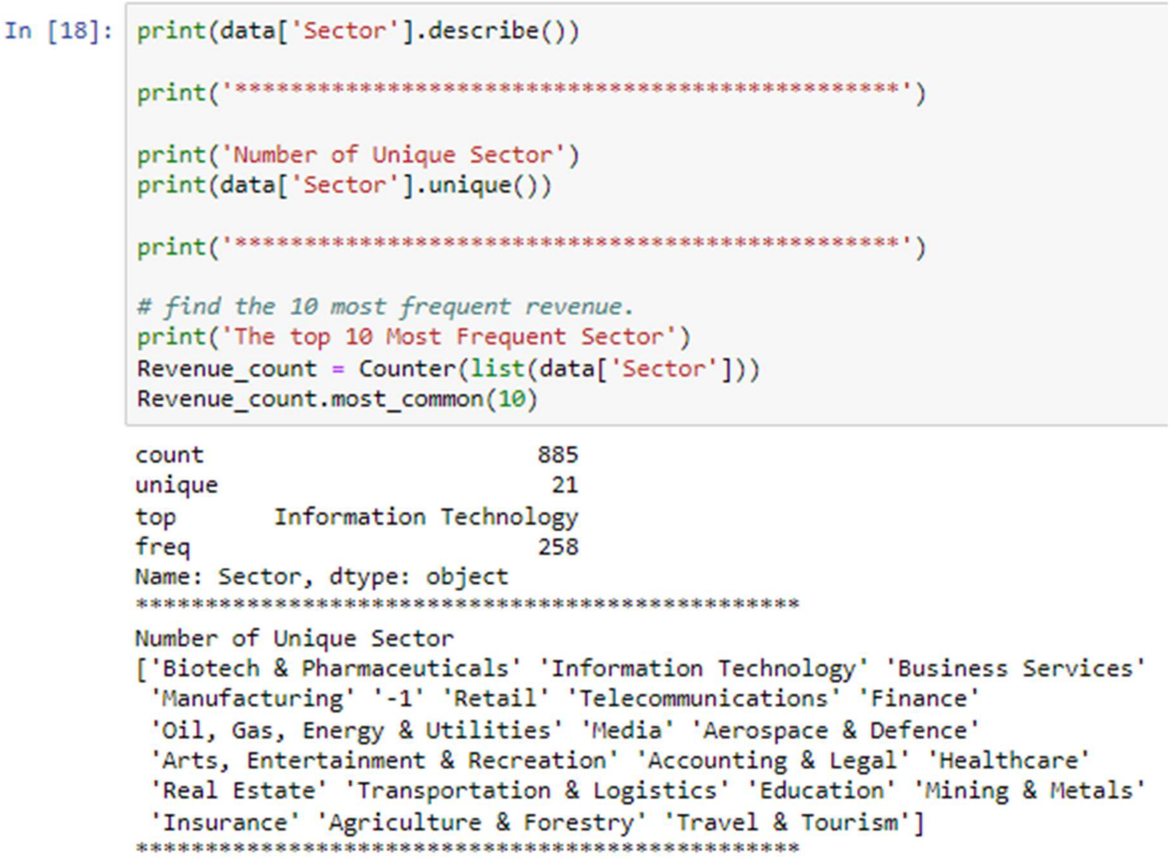
6.3 Basic stats for feature: Revenue



```
In [17]: print(data['Revenue'].describe())

         print('************************************************')

         print('Number of Unique Revenue')
         print(data['Revenue'].unique())

         print('************************************************')

         # find the 10 most frequent revenue.
         print('The top 10 Most Frequent Revenue')
         Revenue_count = Counter(list(data['Revenue']))
         Revenue_count.most_common(10)
```

```
count      885
unique      11
top         -1
freq       498
Name: Revenue, dtype: object
************************************************
Number of Unique Revenue
['500+ billion (INR)' '1 to 5 billion (INR)' '100 to 500 billion (INR)'
 '50 to 100 billion (INR)' '500 million to 1 billion (INR)' '-1'
 '10 to 50 billion (INR)' '100 to 500 million (INR)'
 '10 to 50 million (INR)' '5 to 10 billion (INR)'
 '50 to 100 million (INR)']
************************************************
The top 10 Most Frequent Revenue
```

6.3.1 Basic stats for feature: Sector

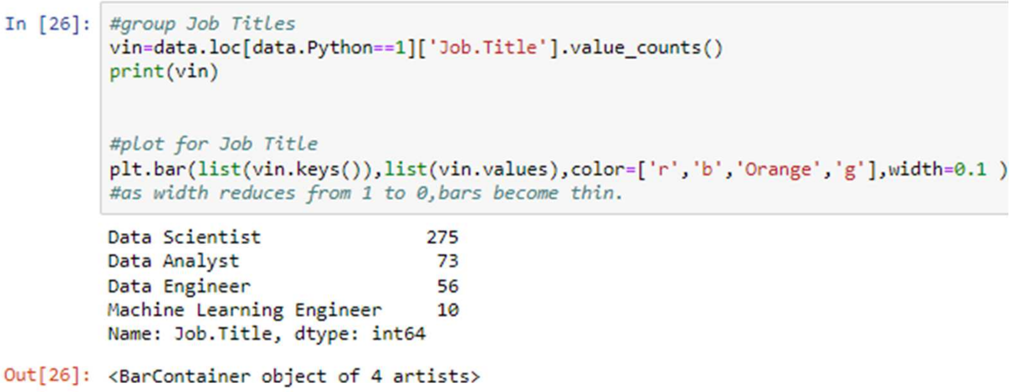


```
In [18]: print(data['Sector'].describe())

         print('************************************************')

         print('Number of Unique Sector')
         print(data['Sector'].unique())

         print('************************************************')

         # find the 10 most frequent revenue.
         print('The top 10 Most Frequent Sector')
         Revenue_count = Counter(list(data['Sector']))
         Revenue_count.most_common(10)
```

```
count                        885
unique                        21
top       Information Technology
freq                         258
Name: Sector, dtype: object
************************************************
Number of Unique Sector
['Biotech & Pharmaceuticals' 'Information Technology' 'Business Services'
 'Manufacturing' '-1' 'Retail' 'Telecommunications' 'Finance'
 'Oil, Gas, Energy & Utilities' 'Media' 'Aerospace & Defence'
 'Arts, Entertainment & Recreation' 'Accounting & Legal' 'Healthcare'
 'Real Estate' 'Transportation & Logistics' 'Education' 'Mining & Metals'
 'Insurance' 'Agriculture & Forestry' 'Travel & Tourism']
************************************************
```

7 Data Visualization:

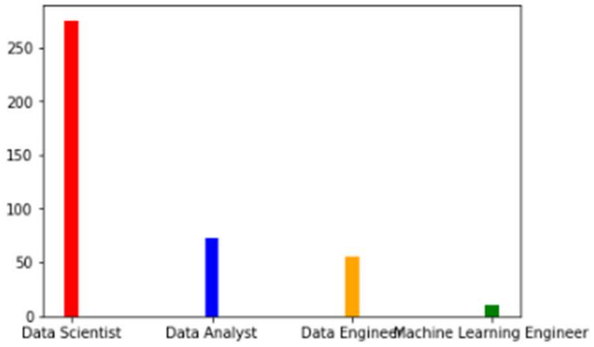
7.1 The most common search Titles



```
In [26]: #group Job Titles
         vin=data.loc[data.Python==1]['Job.Title'].value_counts()
         print(vin)


         #plot for Job Title
         plt.bar(list(vin.keys()),list(vin.values),color=['r','b','Orange','g'],width=0.1 )
         #as width reduces from 1 to 0,bars become thin.
```

```
Data Scientist               275
Data Analyst                  73
Data Engineer                 56
Machine Learning Engineer     10
Name: Job.Title, dtype: int64
```

```
Out[26]: <BarContainer object of 4 artists>
```

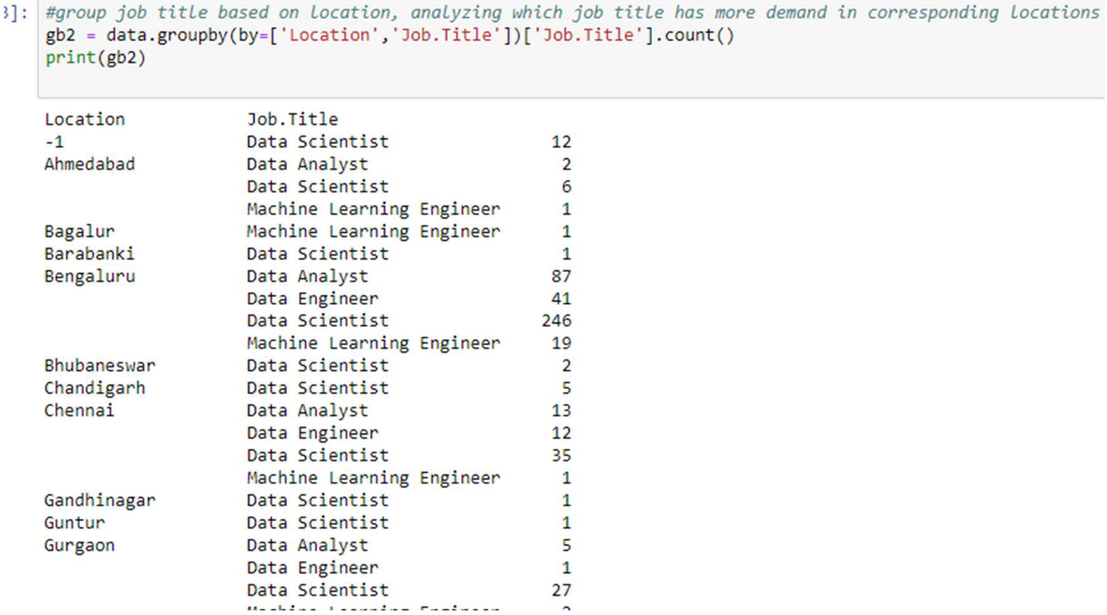


7.1.1 Conclusions:

1. Data Scientist,Data Analyst,Data Enginner , Machine Learning Engineer job titles are interchangeably used for search AI jobs.

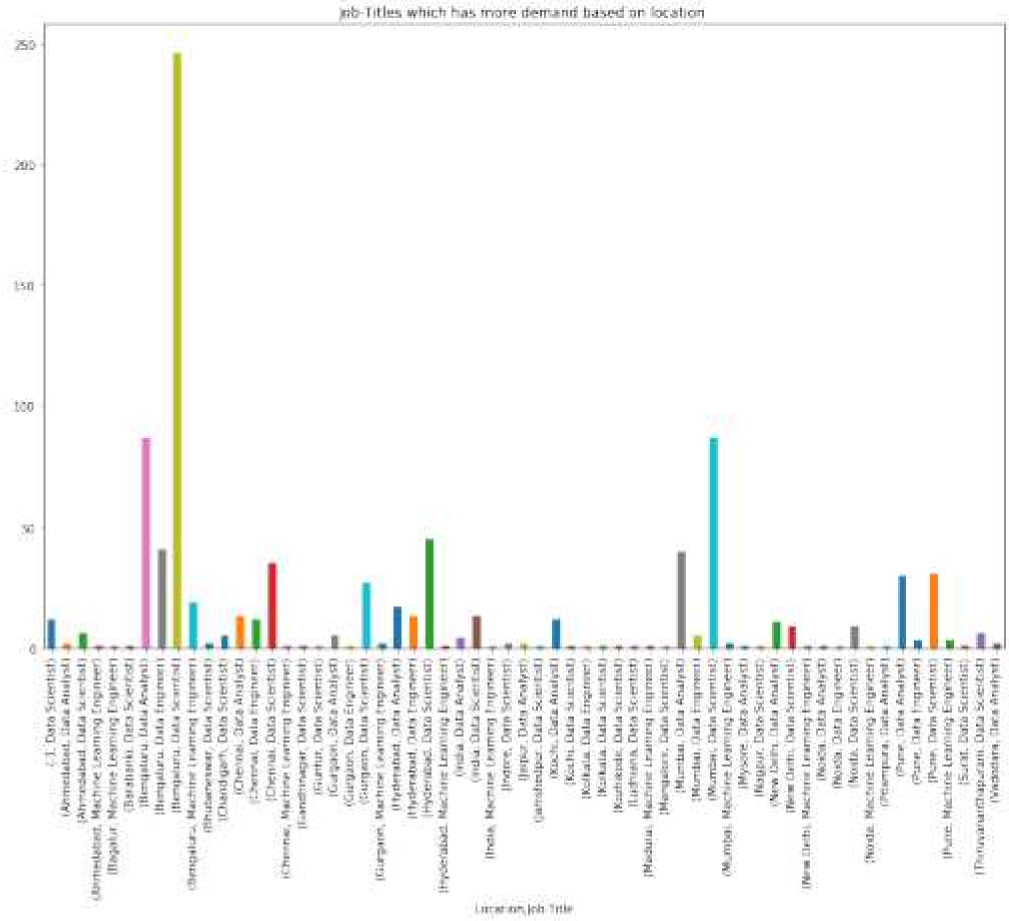
2. Data Scientist , Data Analyst are most frequent job titles used.

7.2 Job Vacancies based on location corresponding to job titles

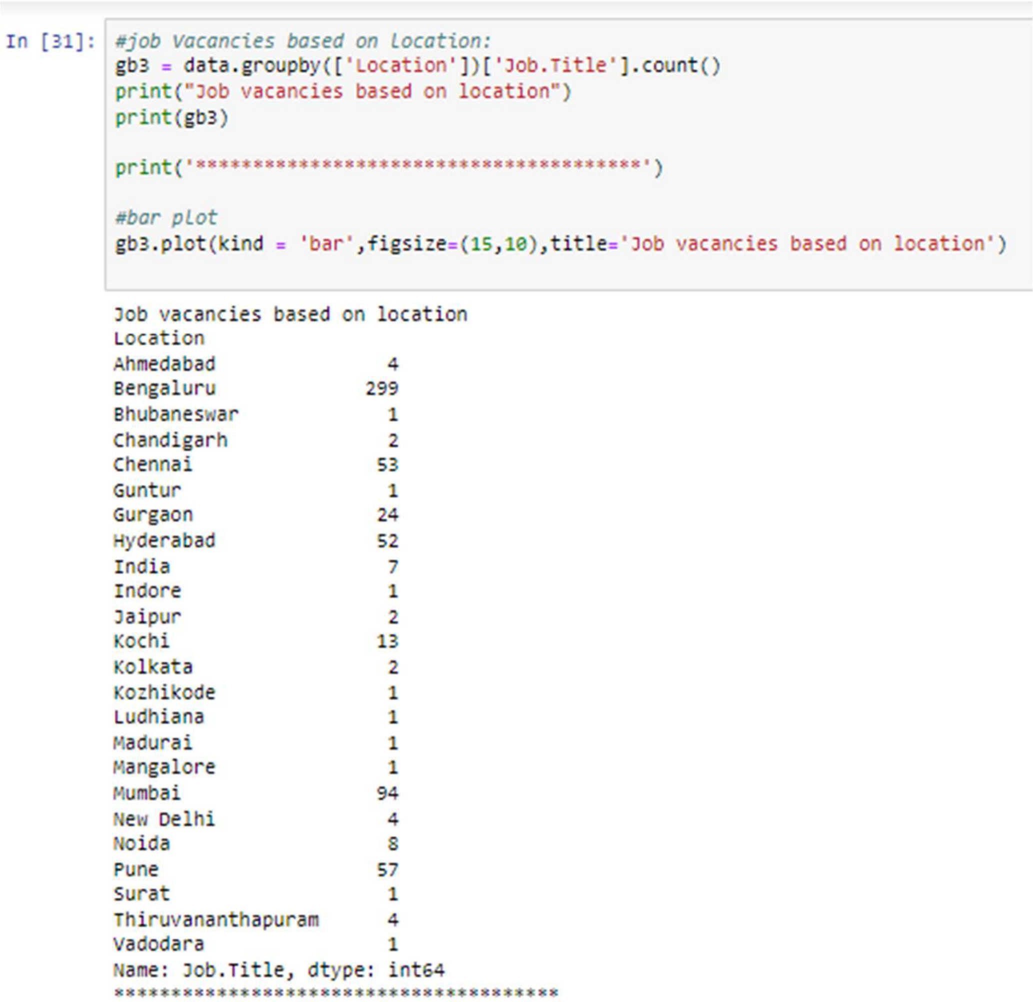


```
3]: #group job title based on location, analyzing which job title has more demand in corresponding locations
    gb2 = data.groupby(by=['Location','Job.Title'])['Job.Title'].count()
    print(gb2)
```

```
Location            Job.Title
-1                  Data Scientist               12
Ahmedabad           Data Analyst                  2
                    Data Scientist                6
                    Machine Learning Engineer     1
Bagalur             Machine Learning Engineer     1
Barabanki           Data Scientist                1
Bengaluru           Data Analyst                 87
                    Data Engineer                41
                    Data Scientist              246
                    Machine Learning Engineer    19
Bhubaneswar         Data Scientist                2
Chandigarh          Data Scientist                5
Chennai             Data Analyst                 13
                    Data Engineer                12
                    Data Scientist               35
                    Machine Learning Engineer     1
Gandhinagar         Data Scientist                1
Guntur              Data Scientist                1
Gurgaon             Data Analyst                  5
                    Data Engineer                 1
                    Data Scientist               27
                    Machine Learning Engineer     2
```

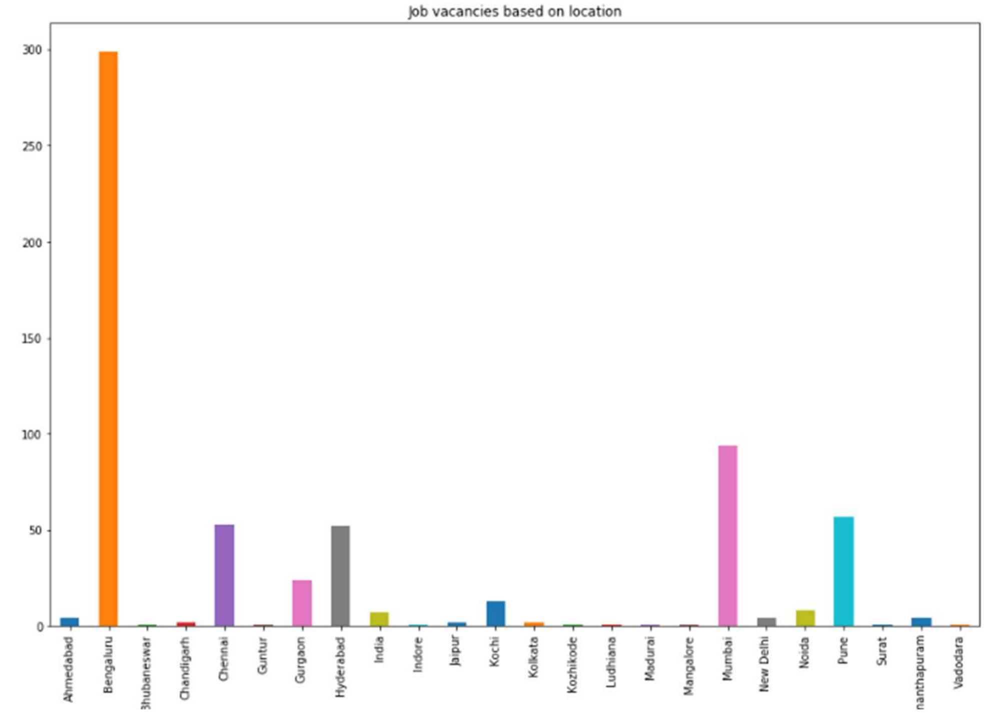


7.3 Job Vacancies Based on Location:

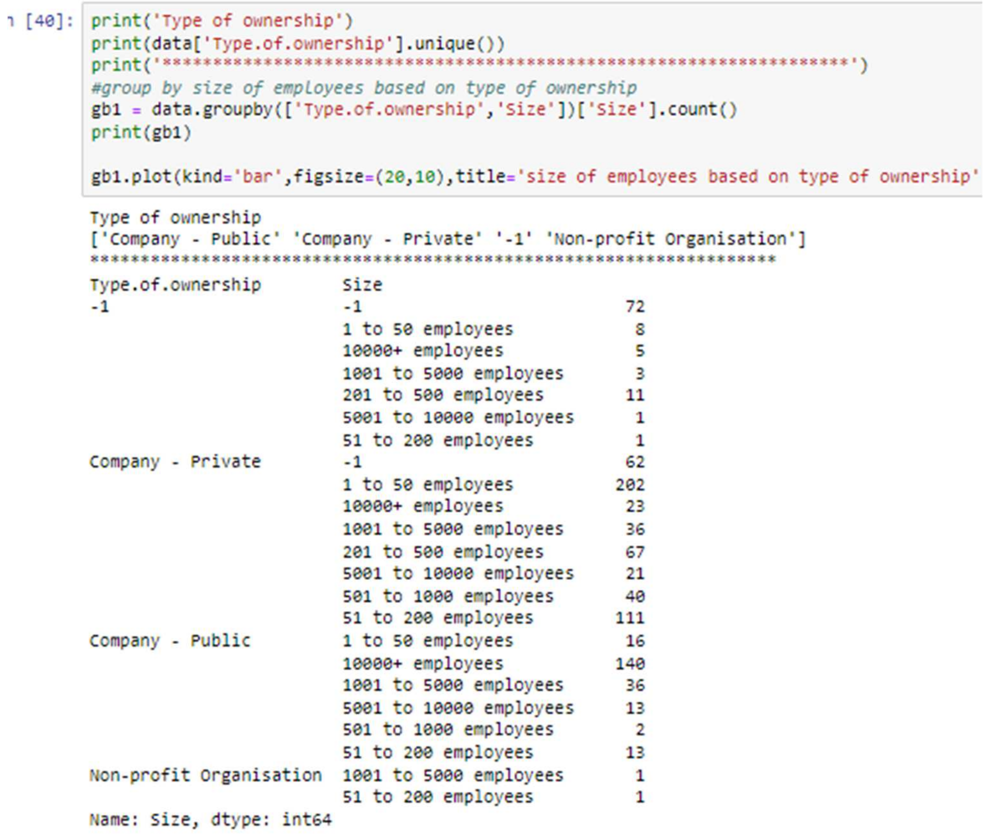


```
In [31]: #job Vacancies based on Location:
         gb3 = data.groupby(['Location'])['Job.Title'].count()
         print("Job vacancies based on location")
         print(gb3)

         print('***************************************')

         #bar plot
         gb3.plot(kind = 'bar',figsize=(15,10),title='Job vacancies based on location')
```

```
Job vacancies based on location
Location
Ahmedabad               4
Bengaluru             299
Bhubaneswar             1
Chandigarh              2
Chennai                53
Guntur                  1
Gurgaon                24
Hyderabad              52
India                   7
Indore                  1
Jaipur                  2
Kochi                  13
Kolkata                 2
Kozhikode               1
Ludhiana                1
Madurai                 1
Mangalore               1
Mumbai                 94
New Delhi               4
Noida                   8
Pune                   57
Surat                   1
Thiruvananthapuram      4
Vadodara                1
Name: Job.Title, dtype: int64
***************************************
```

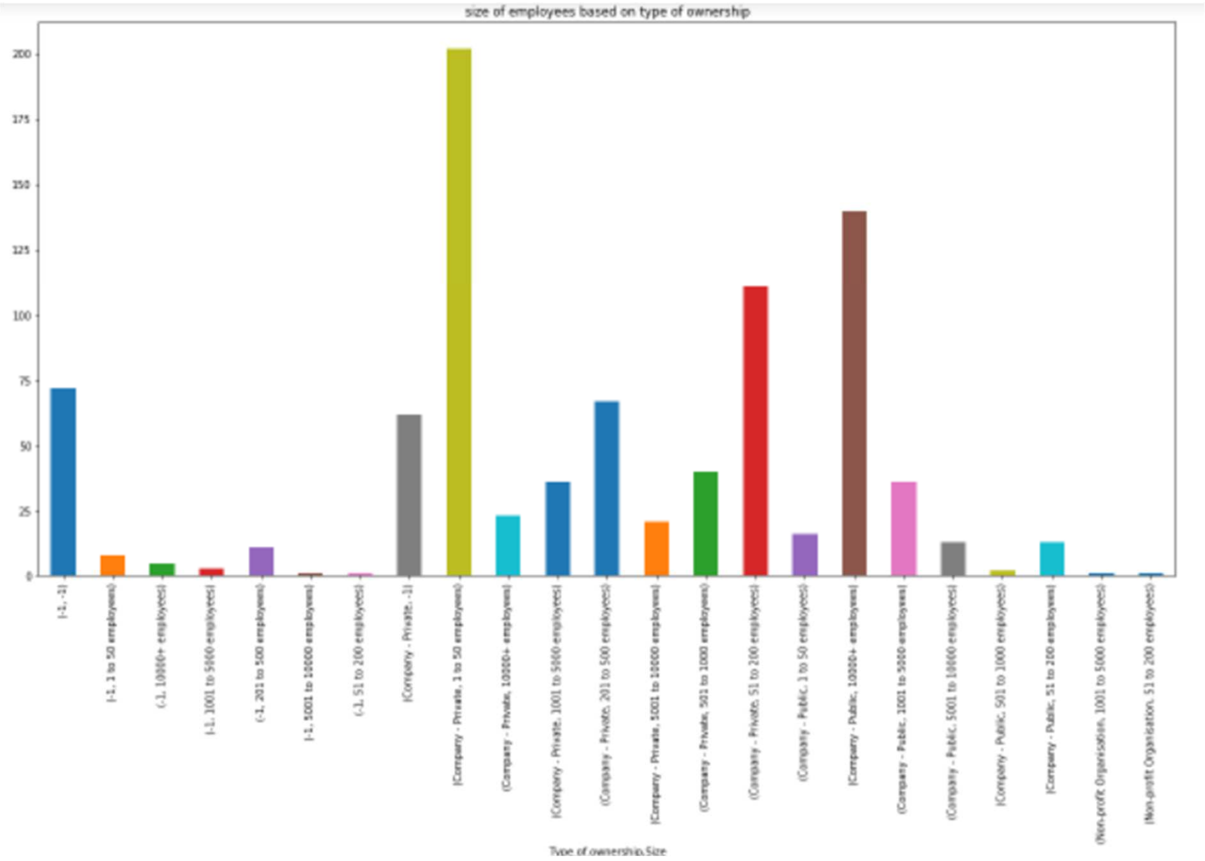


7.4 Size of employees based on type of ownership:

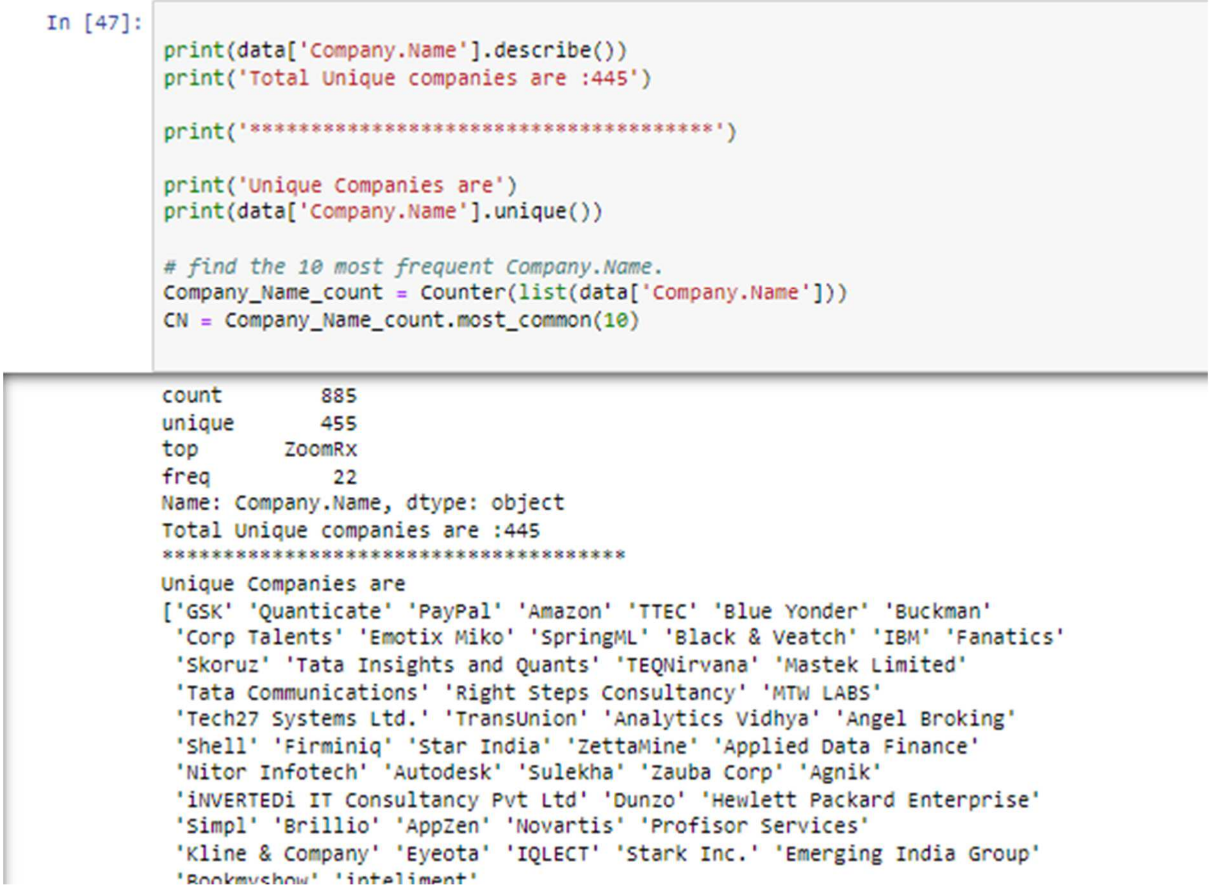


```
ı [40]: print('Type of ownership')
        print(data['Type.of.ownership'].unique())
        print('**********************************************************************')
        #group by size of employees based on type of ownership
        gb1 = data.groupby(['Type.of.ownership','Size'])['Size'].count()
        print(gb1)

        gb1.plot(kind='bar',figsize=(20,10),title='size of employees based on type of ownership'
```

```
Type of ownership
['Company - Public' 'Company - Private' '-1' 'Non-profit Organisation']
*********************************************************************
Type.of.ownership        Size
-1                       -1                          72
                         1 to 50 employees            8
                         10000+ employees             5
                         1001 to 5000 employees       3
                         201 to 500 employees        11
                         5001 to 10000 employees      1
                         51 to 200 employees          1
Company - Private        -1                          62
                         1 to 50 employees          202
                         10000+ employees            23
                         1001 to 5000 employees      36
                         201 to 500 employees        67
                         5001 to 10000 employees     21
                         501 to 1000 employees       40
                         51 to 200 employees        111
Company - Public         1 to 50 employees           16
                         10000+ employees           140
                         1001 to 5000 employees      36
                         5001 to 10000 employees     13
                         501 to 1000 employees        2
                         51 to 200 employees         13
Non-profit Organisation  1001 to 5000 employees       1
                         51 to 200 employees          1
Name: Size, dtype: int64
```



7.5 Companies Which are Hiring:



```
In [47]:

print(data['Company.Name'].describe())
print('Total Unique companies are :445')

print('*************************************')

print('Unique Companies are')
print(data['Company.Name'].unique())

# find the 10 most frequent Company.Name.
Company_Name_count = Counter(list(data['Company.Name']))
CN = Company_Name_count.most_common(10)
```

```
count        885
unique       455
top       ZoomRx
freq          22
Name: Company.Name, dtype: object
Total Unique companies are :445
*************************************
Unique Companies are
['GSK' 'Quanticate' 'PayPal' 'Amazon' 'TTEC' 'Blue Yonder' 'Buckman'
 'Corp Talents' 'Emotix Miko' 'SpringML' 'Black & Veatch' 'IBM' 'Fanatics'
 'Skoruz' 'Tata Insights and Quants' 'TEQNirvana' 'Mastek Limited'
 'Tata Communications' 'Right Steps Consultancy' 'MTW LABS'
 'Tech27 Systems Ltd.' 'TransUnion' 'Analytics Vidhya' 'Angel Broking'
 'Shell' 'Firminiq' 'Star India' 'ZettaMine' 'Applied Data Finance'
 'Nitor Infotech' 'Autodesk' 'Sulekha' 'Zauba Corp' 'Agnik'
 'iNVERTEDi IT Consultancy Pvt Ltd' 'Dunzo' 'Hewlett Packard Enterprise'
 'Simpl' 'Brillio' 'AppZen' 'Novartis' 'Profisor Services'
 'Kline & Company' 'Eyeota' 'IQLECT' 'Stark Inc.' 'Emerging India Group'
 'Bookmyshow' 'inteliment'
```

8 Conclusions:

[Q] Find Companies most probable to hire an ML Engineer/Data Analyst Applicant in respect to his/her skillset.

- [Ans] Top Most Companies which are hiring are 'ZoomRx', 'Amazon','EY','Walmart','Citi','Sanfoi', 'Matelabs Innovations Pvt. Ltd. etc.

[Q] To analyse Machine Learning/Data Analyst Job Market in India and outline the segments most optimal to apply or prepare for Data Analyst/ Machine Learning Jobs.

- [Ans] Top Most Locations Which Are Hiring Machine Learning/Data Analyst in India Are Beangaluru, Mumbai, Pune ,Chennai, Hyderabad.
- Top Most Demanded Skills for prepare Data Analyst/ Machine Learning Jobs are Python, Excel, SQL.

Github Link - https://github.com/Adrg01/Job-Market-Analysis