Job Market Analysis By Web Scraping

Data collection and processing method: The job posting website used for this assignment is Ambition box and the search string used is 'Data Scientist' for India location. I have used webscraping to scrape the site data with python's *Requests* and *Beautiful Soup* library. The content was retrieved from the various HTML tags. The Python code for collecting data from the job portal website has been attached. The data generated from webscraping was stored in a csv file which was then coverted to a dataframe for further analysis. The intial sample size for the analysis was job details of 100 jobs, collected from 10 pages on the job portal site. The details collected for each job is its title, company name, salary, location, key skills and experience in years. Some jobs were posted for multiple locations. To handle this, separate job entries were made in the data frame for each location which effectively made the **sample size** as **150**.**Title, location, company name and skills are categorical variables while salary and experience are continous variables.** Out of 150 rows only 40 rows had posted the available salaries . Hence for analysis of salary a sample size of 40 was used but for the analysis of rest of the variables a sample size of 150 was used. With regards to handling the key skills data, a number of key skills had to be grouped together to correctly analyse their frequency. For example:

Python = "python|python programming|pandas|numpy|coding"
Machine_learning = "hadoop|opencv|cnn|machine learning algorithms|machine learning|azure machine learning|deep learning|computer vision|artificial intelligence|mlops|aiml|ml algorithms|ml|generative ai|ai platform|genai|ai|tensorflow|pytorch|neural networks|pyspark|big data|keras|rnn"
Data_science= "jupyter notebook|data cleansing|scikit-learn|advanced statistical|statistical models|algorithms|data science|data scientist|predictive analytics|text analytics|natural language processing|nlp|statistical modeling|data modeling|modeling|predictive modeling|regression|linear regression|logistic regression|regression analysis|classification"
SQL="sql|data management|database management|database design|data engineering"
Analytics="excel|powerpoint|excel powerpoint|spss|dashboards|data visualization|statistics|statistical analysis|statistical analyses|analytics|analytical|analytical skills|data analysis|data analytics|marketing analytics|customer analytics|business analytics|advanced analytics|distribution analytics|sales analytics|conversation analytics"
Other_prog_lang= "C++|Java|Scala|HTML|go|nodejs|PHP|perl"

Anything out of the above mentioned skills was grouped into "Other" skills.The Salary data extracted from webscraping was given in a range format which was later processed and changed into one single number by taking out the median salary. A similar procedure was followed for the range of Experience (in years)mentioned on the job portal website.Data visualization was done with Excel and Seaborn library functions such as barplot, countplot, boxplot, displot etc. were used for effective visualization of data.

Univariate Analysis

- Among the four category of job titles (Fig. 1), Data Scientist is the most in demand job with around 83% of Job postings.The rest of the categories individually hold less than 10% of the share in job postings.
- The number of different unique companies("All other companies" category) having just one job posting on the job portal is quite high as shown in the Fig. 2 below. IBM,Ford and Shell have the most number of multiple job postings.
- The highest % of job postings (Fig. 3) are in Bangalore(38%) followed by Chennai(16%).National Capital Region(NCR),Mumbai, Hyderabad and Pune cumulatively capture around 39% of the market.The rest of the locations are significantly low in %.
- For a Data Scientist role (Fig. 4), 65% of the time companies look for Data science,Machine learning,Analytics,Python and SQL skills.Other than Python, requirement for other programming languages like C++,Java,Html is significantly less.Knowledge of Cloud services like AWS,GCP and various other specific domain knowledge etc. constitute the "Others" skill category(32%).

Fig 1. Percentage of Job Titles

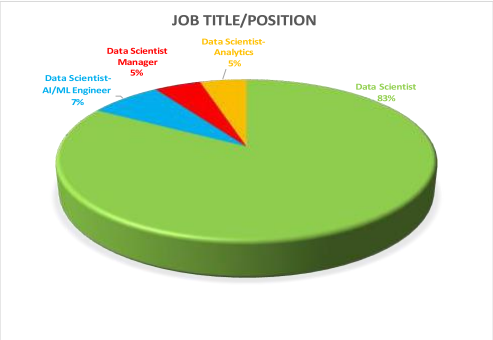


Fig 2. Number of Job postings in different companies

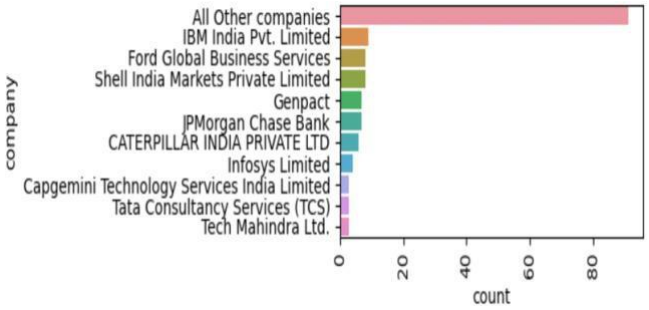


Fig 3. Percentage of Job postings in different locations

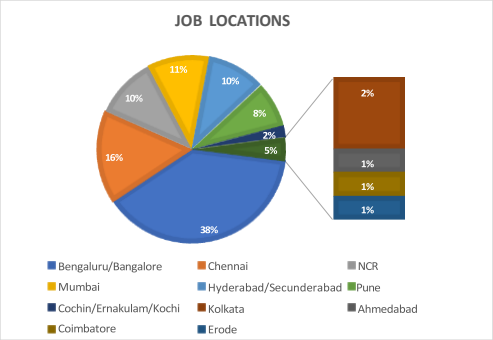
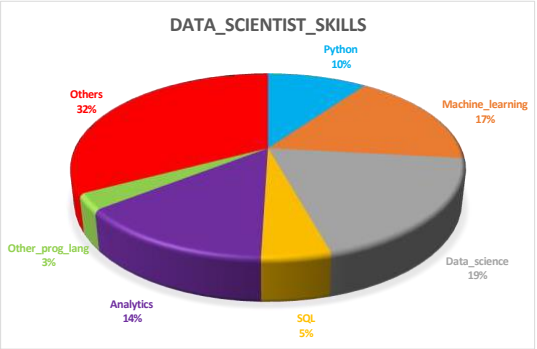


Fig 4. Percentage of Skills



- The salaries (Fig 5 & 6) are mostly centered around Rs.10-15 Lakhs/year with 80% earning less than Rs.18 Lakhs/year.
- The histogram for experience in years (Fig 7 & 8) is slightly right skewed with 60% of job postings requiring more than 5-14 years in experience.

Fig 5.Histogram of Salaries

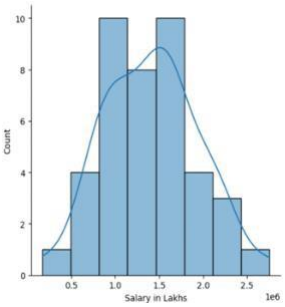


Fig 6. Cumulative distribution plot of Salaries

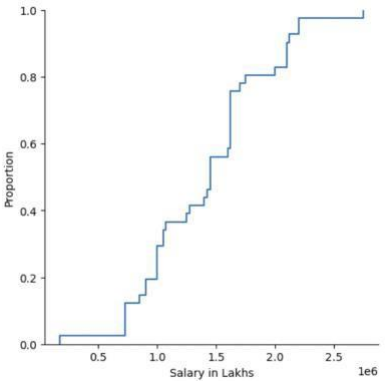


Fig 7.Distribution of Experience in Years

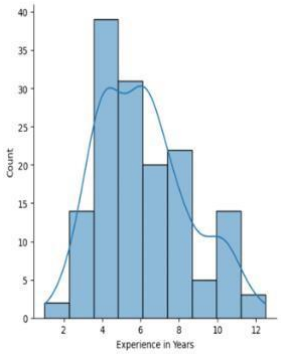
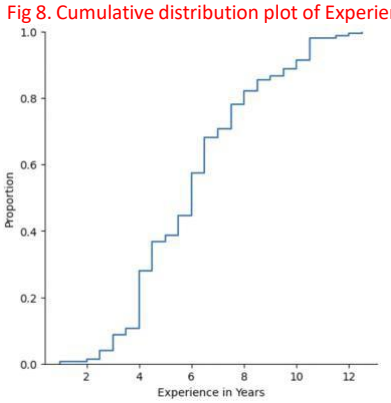


Fig 8. Cumulative distribution plot of Experience in Year



Bivariate Analysis

Note- *Salaries were available only for a smaller dataset of around 40 Jobs out of 150 jobs.The following 4 bullets and 4 figures (Fig. 9,10,11 & 12) have been analysed for 40 jobs.*

- Among all the job titles(Fig 9),the highest average salary is earned by Data Scientist Manager(Rs 18 Lakhs/Year) compared to a Data Scientist who earns an average of Rs 13 Lakhs/Year,although the uncertainity is more in case of Data Scientist Manager than Data Scientist as shown by error bars.
- The average salaries(Fig 10) offered in Pune(Rs 18 Lakhs/Year) is surprisingly higher than Bangalore(Rs 17 Lakhs/Year).Among other major cities Hyderabad offers an average salary of Rs 15 Lakhs/Year followed by Chennai and NCR in descending order.Mumbai has the lowest average salary.
- Data Scientist positions are available at all locations(Fig 11).NCR has the most number of different job titles.Currently, Data Scientist-AI/ML Engineer positions are only available in Hyderabad.Bangalore has the highest no. of available Data Scientist Manager positions with the highest average salary.
- The relationship between experience in years and salary in lakhs (Fig 12) is strongly linear and they have a Pearsons correlation coefficient as 0.75.

Fig 9.Salary of different job titles

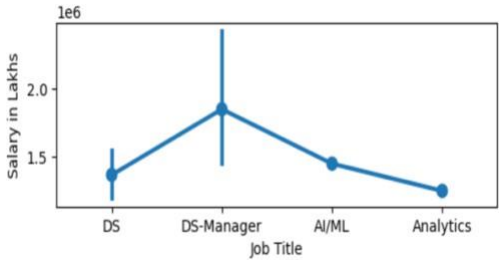


Fig 10.Salary at different locations

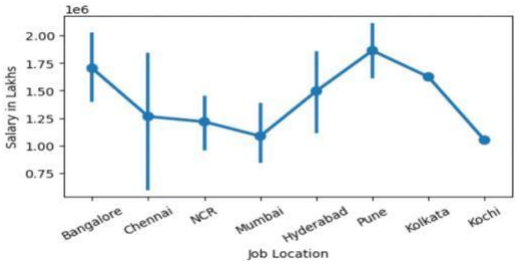


Fig 11.Salary of different job titles at different locations

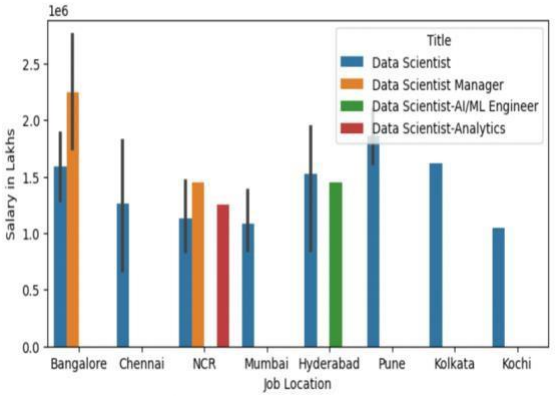
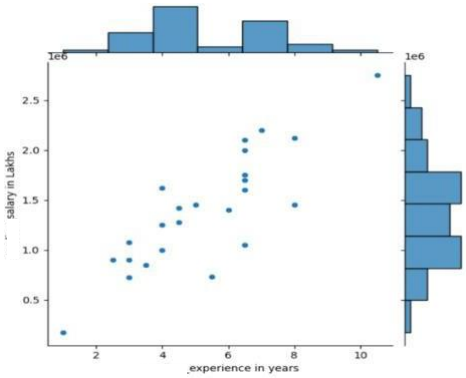


Fig 12. Correlation of salary and experience



Note-*.The following bullets and figures (Fig.13,14,15 & 16) have been analysed for 150 jobs.*

- Bangalore, Chennai and NCR have all the different categories of job title postings (Fig 13) with significantly higher number of Data Scientist jobs in Bangalore compared to other cities.Delhi follows next but the ratio of jobs posted for Bangalore to Delhi is around 5:1.
- The average experience required (Fig 14) for Data Scientist Manager is 8years whereas for all the rest of the titles it is 6years.
- Mumbai (Fig 15) requires the lowest average experience in years compared to other cities for its job postings.
- For Data Scientists-AI/ML engineer (Fig 16), NCR requires the lowest average experience in years; for Data Scientists-Manager,Chennai requires the highest; for Data Scientists Analytics,Bangalore requires the lowest average experience in years.

Fig 13.Count of different job titles at different locations

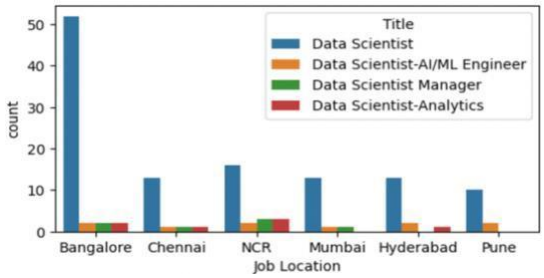


Fig 14.Boxplot of Experience of different job titles

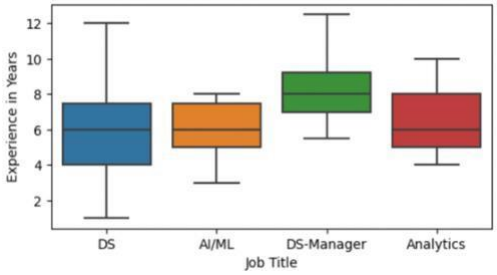


Fig 15.Boxplot of Experience at different locations

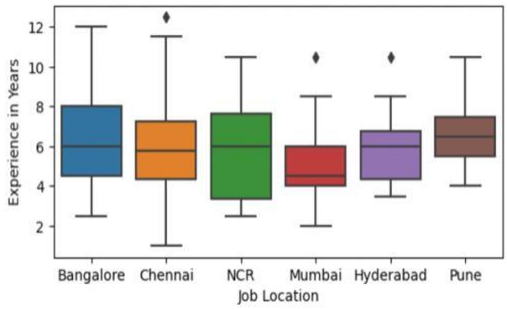


Fig 16. Experience of different job titles at different locations