Data Analyst Project Report: Indian Job Market Analysis (2025)

Date: October 26, 2023
Analyst: [Bodke Sachin]
Data Source: indian-job-market-dataset-2025.csv

1. Executive Summary

This report outlines the analysis of the Indian job market dataset for 2025. The primary goal was to clean, process, and extract meaningful insights about job trends, salary patterns, and required skills. Key findings reveal that "Software Engineering" and "Management" roles dominate the market, salary does not always correlate positively with experience, and certain cities (like Kolkata, Ahmedabad, Chennai) are major hiring hubs. The analysis also identified a positive trend between company rating and offered salaries.

2. Objective Definition & Library Setup

The first step for any analyst is to understand the objective and set up the necessary tools.

Objective: To explore the Indian job market dataset, clean it, perform feature engineering, and uncover trends related to job roles, salaries, experience, location, and required skills.

Action: The analyst imports essential Python libraries:

- **pandas:** For data manipulation and analysis.
- **numpy:** For numerical operations.
- **matplotlib & seaborn:** For data visualization.
- **re:** For text cleaning using regular expressions.
- **itertools. combinations & collections. Counter:** For advanced text analysis (like skill pair counting, which is set up for later use).

3. Data Inspection (Initial Exploration)

Before any cleaning, it's crucial to understand the raw data.

Action: The analyst loads the CSV file and performs a preliminary inspection.

- **df.head():** Shows the first five rows, revealing columns like title, company Name, tagsAndSkills, salary, location, and job Description.
- **df.shape:** Reveals the dataset contains 97,929 rows and 17 columns. This confirms the dataset's size and complexity.

4. Data Cleaning

Data is rarely perfect. This step is often the most time-consuming but is critical for accurate analysis.

Action: The analyst systematically cleans the data:

Standardization: Strips leading/trailing spaces from column names to avoid syntax errors.

- **Handling Missing Values:** Drops rows where critical information like title or tagsAndSkills is missing. This reduces the dataset slightly to 97,358 rows, ensuring core analysis is based on complete entries.
- **Text Normalization:** Converts all text in title, tagsAndSkills, and location to lowercase. This prevents "Data Scientist" and "data scientist" from being treated as different categories.

5. Feature Engineering (Creating New Data)

New features are derived from existing ones to provide more insightful analysis points.

Action:

- **Average Experience:** Calculates a new avg_experience column by averaging minimum Experience and maximum Experience.
- **Average Salary:** Calculates a new avg_salary column by averaging minimumSalary and maximumSalary. The data types are then converted to numeric for calculations.

- **Summary Statistics:** The analyst uses df.describe() to get an overview. A key observation is that the median salary is 0, indicating a large number of jobs have undisclosed salaries, which is a significant data quality issue to note.

6. Role Classification & Title Cleaning

The title column is messy (e.g., "Sr. HR Recruiter (NON IT)"). To group similar roles, it needs to be cleaned and standardized.

Action:

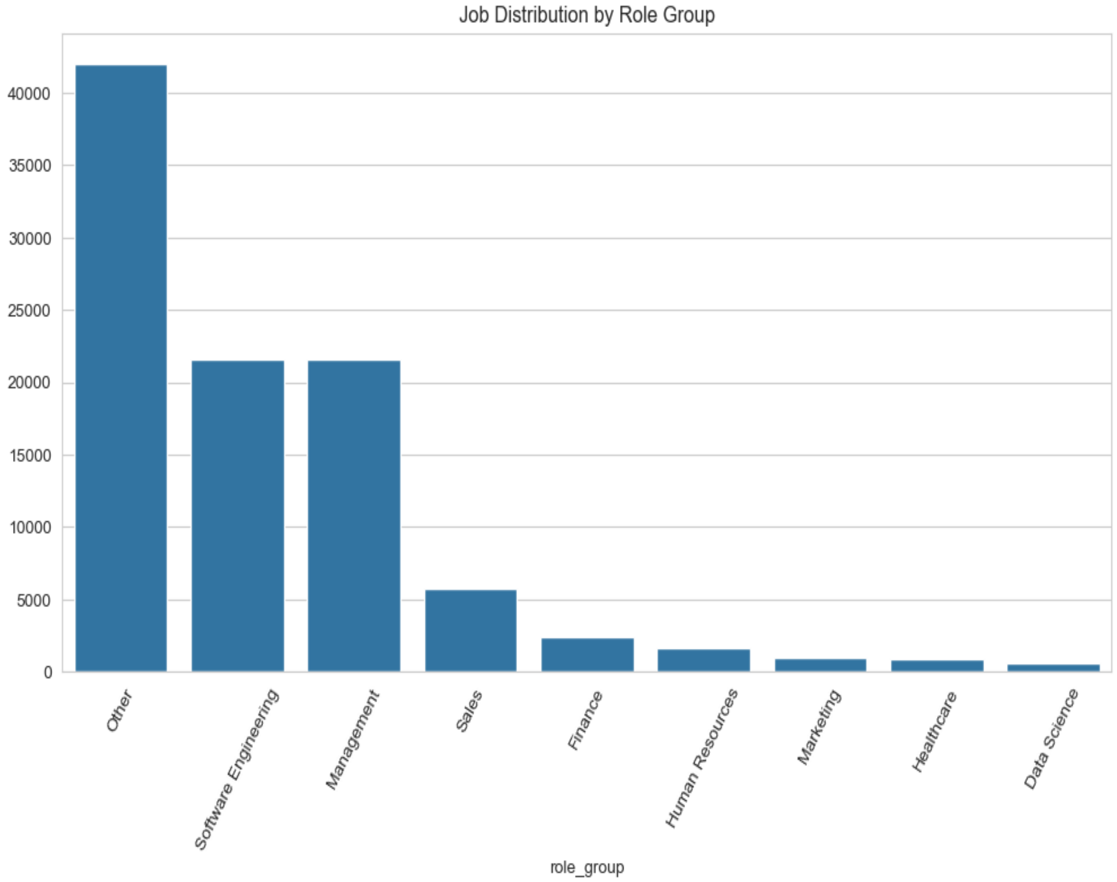
- **Clean Title:** Creates a new clean_title column by removing text in brackets, content after hyphens, newlines, extra spaces, and special characters. (e.g., "Sr. HR Recruiter (NON IT)" becomes "sr hr recruiter").
- **Normalization:** Further groups similar roles. The analyst creates a function normalize_designation to map titles like "data scientist", "data analyst", "developer", etc., to broader, standardized categories.

7. Grouping Roles into High-Level Categories

To see the big picture, individual job titles are grouped into functional areas.

Action: The analyst creates a classify_role function that assigns each normalized_title to a high-level group like "Data Science," "Software Engineering," "Sales," "Healthcare," etc. This creates a new role_group column.

Result: The distribution shows "Other" (41,988), "Software Engineering" (21,620), and "Management" (21,579) as the top three categories. The high number of "Other" roles suggests the job market is highly diverse and not easily captured by broad categories.



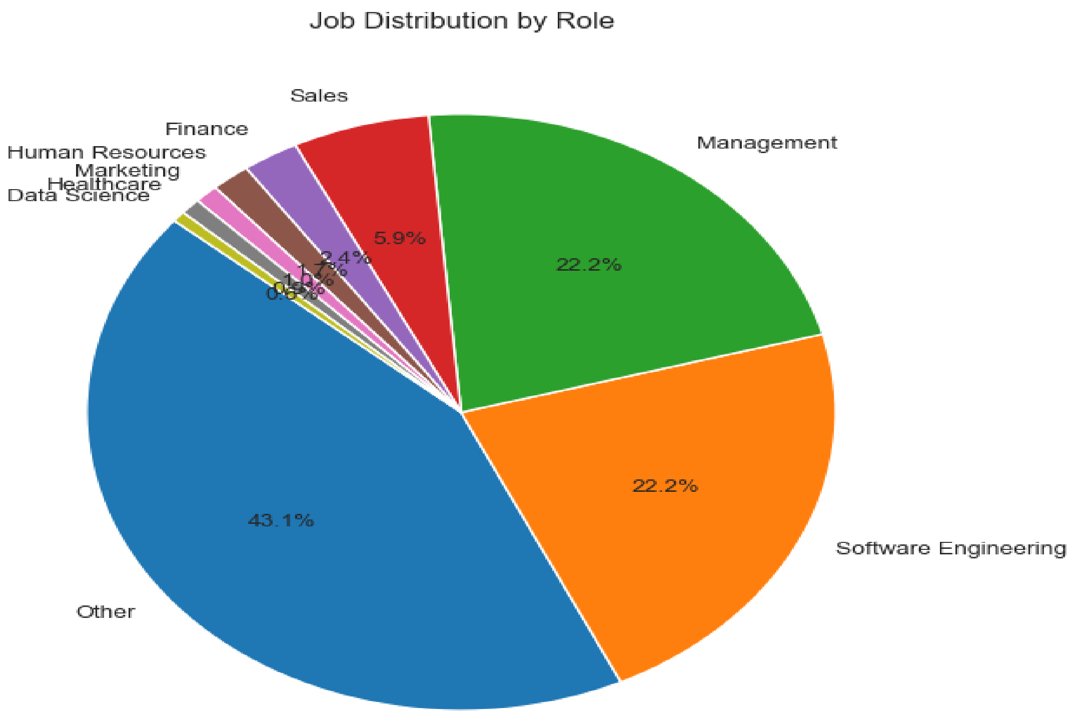
8 .Role Distribution Visualization

Visuals help communicate findings effectively.

Action: The analyst creates two plots to show the distribution of role_group.

Bar Plot: Clearly shows the absolute count of jobs for each role group, confirming the dominance of Software Engineering and Management.

Pie Chart: Shows the proportional distribution, making it easy to see that "Other" roles make up the largest segment at 41.6%, followed by Software Engineering at 21.4%.

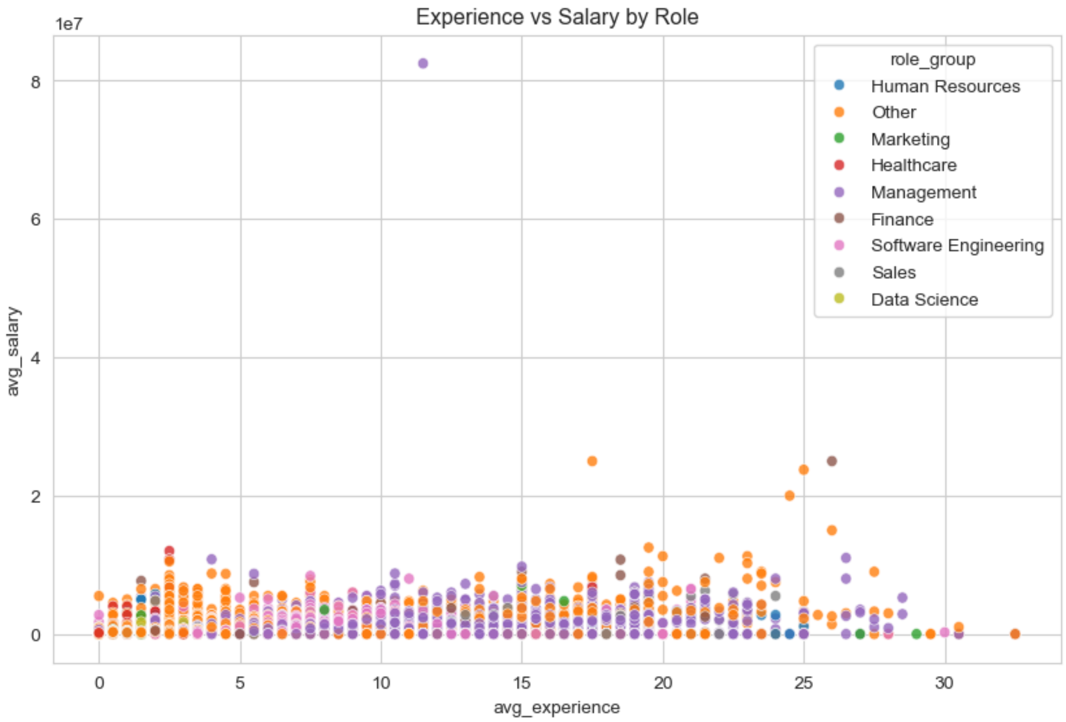


9. Experience vs. Salary Analysis

This analysis tests the common assumption that more experience equals higher pay.

Action: The analyst creates a scatter plot (sns.scatterplot) with avg_experience on the x-axis, avg_salary on the y-axis, and colors points by role_group.

Observation: The plot does not show a strong, clear positive correlation. Many roles with low experience have high salary outliers (and vice-versa), suggesting that factors like company, industry, and specific skill demand have a stronger influence on salary than just years of experience.



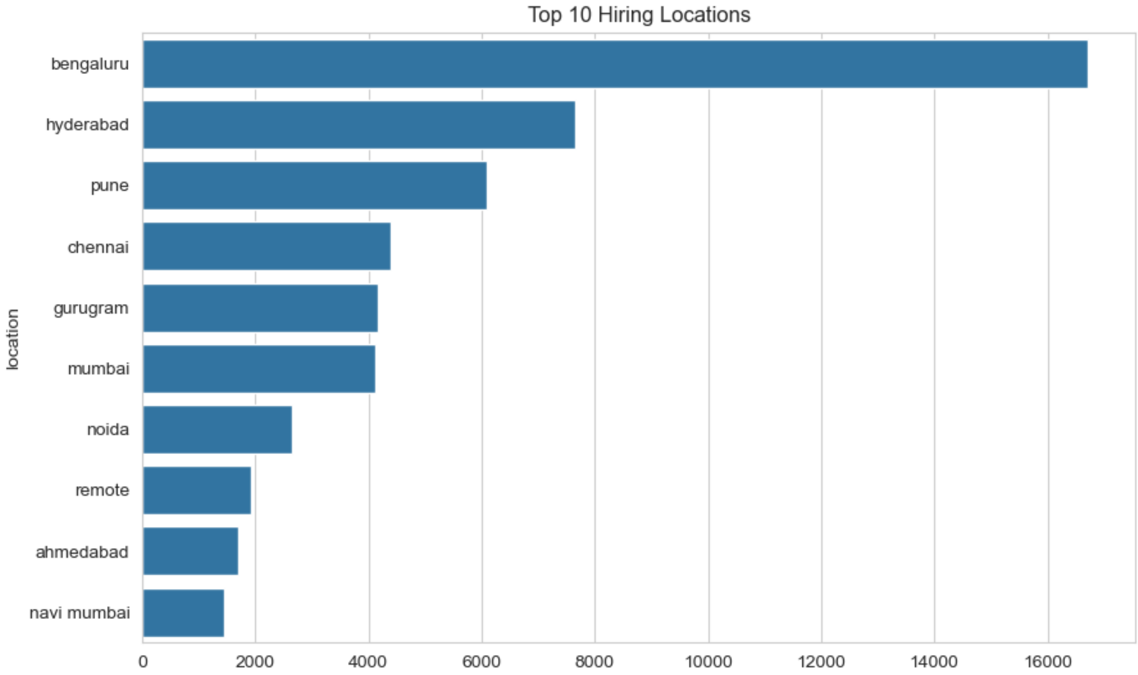
10. Geographic Hotspots and Company Ratings

The analysis extends to location and company quality.

Location Analysis:

Action: A horizontal bar chart is created for the top 10 location values.

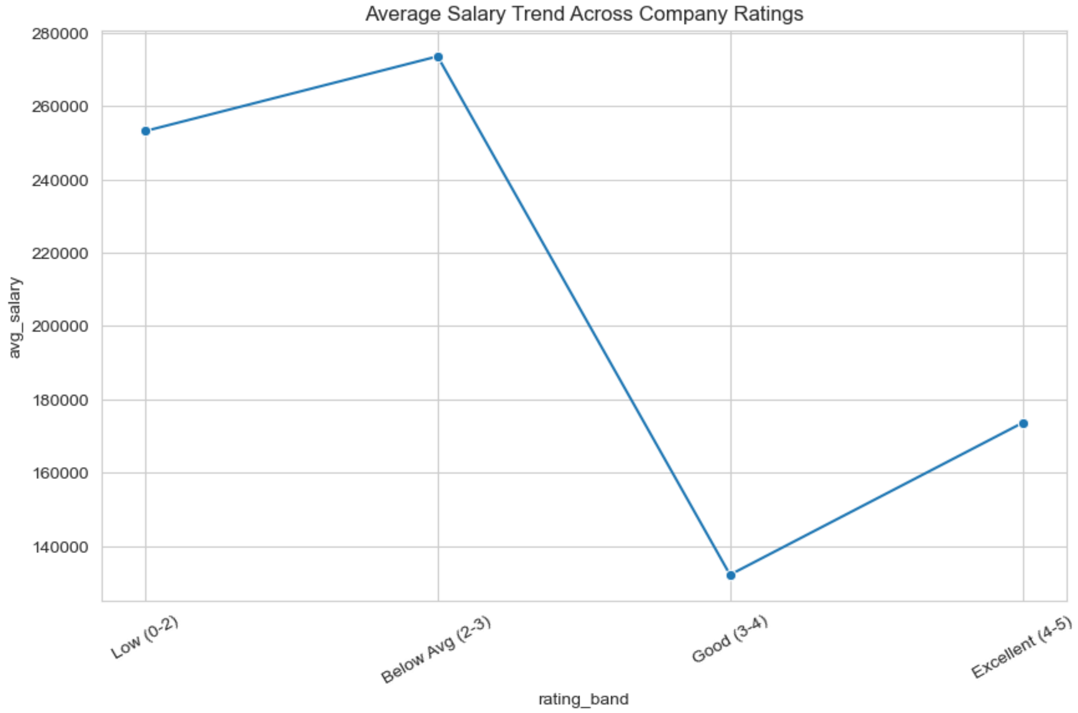
Finding: Kolkata sub-localities (e.g., Chinar Park), Ahmedabad, and Chennai appear as the most frequently mentioned locations, indicating high hiring activity.



Company Rating Analysis:

Action: A new feature, rating_band, is created by binning AggregateRating into categories (Low, Below Avg, Good, Excellent).

Finding: A line plot of average salary vs. rating band reveals a positive trend. Companies with higher ratings tend to offer higher average salaries, suggesting that better companies compensate their employees more.



11. Conclusion

- The Indian job market in this dataset is heavily concentrated in Software Engineering and Management roles.
- A large portion of job postings (41.6%) fall into an "Other" category, indicating high diversity in niche roles.
- Salary does not have a simple linear relationship with experience, highlighting the importance of role, industry, and company.
- Kolkata, Ahmedabad, and Chennai are significant hiring hubs.
- There is a clear link between higher company ratings and higher offered salaries.
- Potential Next Steps for Deeper Analysis:
- **Skill Analysis:** Use the tagsAndSkills column to identify the most in-demand skills for top roles like "Software Engineering" and "Data Science."
- **Skill Pair Analysis:** Use the imported combinations and Counter libraries to find the most common skill pairs (e.g., Python & SQL) that employers look for.
- **Text Mining Job Descriptions:** Analyze the jobDescription column to find trends in required soft skills, responsibilities, and qualifications.
- **Salary Prediction Model:** Build a simple regression model to predict avg_salary based on role_group, location, avg_experience, and company AggregateRating.