



INNOVATION. AUTOMATION. ANALYTICS

**PROJECT ON**



**AmbitionBox**

***Title: “Workforce Insights from AmbitionBox: A Data-Driven Analysis”***



# About Us

- We are Sowmya and Ragasudha B.Tech graduates passionate about Data Science and data-driven problem solving.
- Hands-on experience with Python for data analysis, visualization, and web data extraction.
- Strong understanding of data preprocessing and exploratory analysis.
- Actively upskilling in machine learning and advanced analytics tools.

## Sowmya

- GITHUB - <https://github.com/SowmyaKona>
- LINKEDIN - <https://www.linkedin.com/in/sowmya-kona-152a862a7/>

## Ragasudha

- GITHUB - <https://github.com/ragasudha1104>
- LINKEDIN - <http://linkedin.com/in/kanaboina-ragasudha>

# *Project Overview*

- INTRODUCTION
- PROBLEM STATEMENT
- OBJECTIVE
- WEB SCRAPPING
- TOOLS USED FOR SCRAPPING
- DATA CLEANING STEPS
- DATA VISUALIZATION
- CONCLUSION

# *Introduction*

## *What is AmbitionBox?*

*AmbitionBox is an employee intelligence platform which Provides insights on:*

- Company culture*
- Salaries & benefits*
- Interview experiences*
- Career growth*

*It is widely used by job seekers and HR teams .However, it is difficult to manually identify hiring trends, salary patterns, and industry performance across different locations.*



# ***Problem Statement:***

Online platforms like AmbitionBox generate large volumes of company-level data including salaries, reviews, ratings, interviews, and job listings.

Job seekers lack transparent, data-driven insights to identify the best work environments. There is a need for data-driven analysis to summarize company performance, highlight industry trends, and visualize satisfaction levels effectively.

## **Goal:**

To analyze AmbitionBox data across industries and locations to identify patterns in:

- Salary distribution
- Hiring activity (Interviews & Jobs)
- Employee reviews and ratings
- Industry-wise and location-wise trends

# *Need of Ambition Box*

## *Why to Analyze AmbitionBox Data?*

- *Employee opinions influence employer branding*
- *Job seekers depend on reviews before joining*
- *Companies want to:*
  - *Improve retention*
  - *Compare themselves with competitors*
- *Raw web data needs structuring for insights*



# *Objective:*

- Collecting company-level data from AmbitionBox to analyze and compare companies using
  - Employee reviews
  - Ratings
  - Salaries
- Clean and standardize inconsistent web data to provide data-driven insights that help job seekers make informed career decisions based on real employee experiences and industry trends. To explore career growth patterns and work culture trends across different sectors.
- Visualize trends for decision-making

# *Web Scraping*

- AmbitionBox served as the core data source for this analysis.
- Web page structures were analyzed using browser developer tools to locate required data elements.
- Automated HTTP requests enabled efficient data retrieval.
- Python-based web scraping was implemented using Requests and BeautifulSoup.
- EDA techniques were applied to transform raw data into actionable insights.

# Web Page Inspection for Data Extraction

The screenshot displays a web browser window with the URL `ambitionbox.com/list-of-companies`. The page shows a list of companies, with the Accenture company card selected. The developer tools are open, showing the HTML structure of the selected card. The HTML elements identified include:

- `<meta itemprop="name" content="Accenture">`
- `<meta itemprop="alternateName" content="Accenture overview">`
- `<meta itemprop="url" content="https://www.ambitionbox.com/overview/accentur...>`
- `<meta itemprop="image" content="https://static.ambitionbox.com/alpha/company/photos/logos/accenture.jpg">`
- `<div class="companyCardWrapper_primaryInformation" data-widget="top_company_impressions">`
- `<div class="companyCardWrapper_companyLogo">`
- `<div class="companyCardWrapper_metaInformation">`
- `<div class="companyCardWrapper_header">`
- `<div class="companyCardWrapper_companyPrimaryDetailsTopSection">`
- `<a class="companyCardWrapper_companyName companyCardWrapper_companyName--m8">`
- `<h2 title="Accenture" class="companyCardWrapper_companyName">`
- `Accenture </h2> == $0`
- `</a>`
- `<!-->`
- `</div>`
- `<button type="button" arialabel="Follow" title="Follow" class="companyCardWrapper_FollowCTA g-btn g-btn--text g-btn--md">`
- `<!-->`
- `</div>`
- `<div class="companyCardWrapper_ratingWrapper">`
- `<div class="viewSalaryInsights">`
- `<!-->`
- `</div>`
- `</div>`
- `<div class="companyCardWrapper_tertiaryInformation">`

- *HTML elements containing company information were identified using browser developer tools*

## *Libraries Used:*

BeautifulSoup

Requests

•[RegEx]\*

 NumPy

 pandas

 matplotlib

 seaborn

# Raw Dataset

	company	ratings	types	location	reviews	salaries	interviews	jobs	benefits	photos	high	low	tots_oper
0	['TCS']	['3.4']	IT Services & Consulting	Bangalore	1.1L	9.5L	11.6k	1.1k	10.6k	87	Job Security	Promotions / Appraisal, Salary & Benefits, Wor...	434
1	['Accenture']	['3.7']	IT Services & Consulting	Bangalore	68.4k	6.3L	9k	27k	6.7k	48	NaN	Promotions / Appraisal, Salary & Benefits, Wor...	249
2	['Wipro']	['3.7']	IT Services & Consulting	Hyderabad	61.3k	4.7L	6.4k	4.4k	4.6k	99	Job Security	Promotions / Appraisal, Salary & Benefits, Wor...	368
3	['Cognizant']	['3.7']	IT Services & Consulting	Hyderabad	58.2k	5.9L	6.2k	778	5.5k	82	NaN	Promotions / Appraisal, Salary & Benefits, Wor...	229
4	['Capgemini']	['3.7']	IT Services & Consulting	Bangalore	49.9k	4.7L	5.3k	1.9k	3.6k	41	Work Life Balance, Job Security	Promotions / Appraisal, Salary & Benefits, Wor...	182
...	...	...	...	...	...	...	...	...	...	...	...	...	...
495	['NCR', 'Corporation']	['3.6']	IT Services & Consulting	Hyderabad	1.8k	9k	142	48	173	6	NaN	Promotions / Appraisal, Job Security, Skill De...	89

# *Data Cleaning Steps:*

- Removed unwanted characters such as [ ], ', and extra spaces from scraped text fields.
- Separated combined columns (e.g., “IT Services & Consulting | Bengaluru”) into Type and Location columns.
- Converted string-based ratings (like '[4.2]') into numeric float values for analysis.
- Converted textual numeric values such as '1.1L' and '9.3K' into actual numbers (110000, 9300).
- Handled missing values by replacing NaN entries with Unknown in columns like types, location, high, and low and with Zeros in columns like job and photos.
- Standardized salary, review counts ,etc. by converting text with suffixes like K and L into numerical values.
- Exported the cleaned dataset into a structured CSV/Excel format for further
- Exploration and visualization using Python

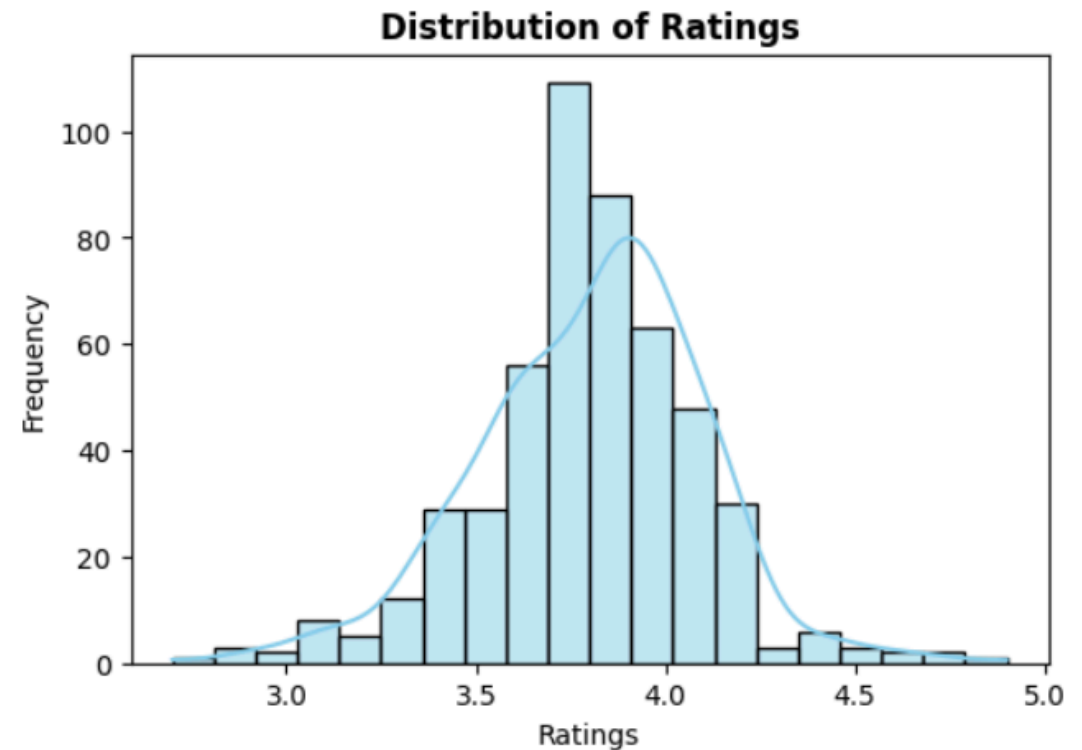
# Cleaned Data

	Company	Rating	Industry	Location	Reviews	Salaries	Interviews	Jobs	Benefits	Photos	Highly_Rated_For	Critically_Rated_For	Total_Operations_Spread
0	TCS	3.3	IT Services & Consulting	Bengaluru	110000.0	980000.0	11800.0	2500.0	10600.0	93.0	Job Security	Promotions / Appraisal, Salary & Benefits, Wor...	440
1	Accenture	3.7	IT Services & Consulting	Bengaluru	71700.0	650000.0	9200.0	37700.0	6700.0	49.0	Not Mentioned	Promotions / Appraisal, Salary & Benefits, Wor...	257
2	Wipro	3.6	IT Services & Consulting	Hyderabad	63700.0	480000.0	6700.0	7200.0	4600.0	107.0	Not Mentioned	Promotions / Appraisal, Salary & Benefits, Wor...	371
3	Cognizant	3.6	IT Services & Consulting	Hyderabad	60100.0	600000.0	6300.0	638.0	5500.0	86.0	Not Mentioned	Promotions / Appraisal, Salary & Benefits, Wor...	231
4	Capgemini	3.7	IT Services & Consulting	Bengaluru	51700.0	480000.0	5500.0	2100.0	3700.0	42.0	Work Life Balance, Job Security	Promotions / Appraisal, Salary & Benefits, Wor...	183
5	HDFC, Bank	3.8	Banking	Mumbai	50900.0	150000.0	3000.0	326.0	3200.0	86.0	Job Security, Skill Development / Learning	Promotions / Appraisal, Salary & Benefits	1835
6	Infosys	3.5	IT Services & Consulting	Bengaluru	47500.0	520000.0	8300.0	2500.0	4800.0	118.0	Job Security	Promotions / Appraisal, Salary & Benefits. Wor...	249

# *Histogram for “Ratings” Column*

This plot helps visualize how company ratings are spread across different score ranges.

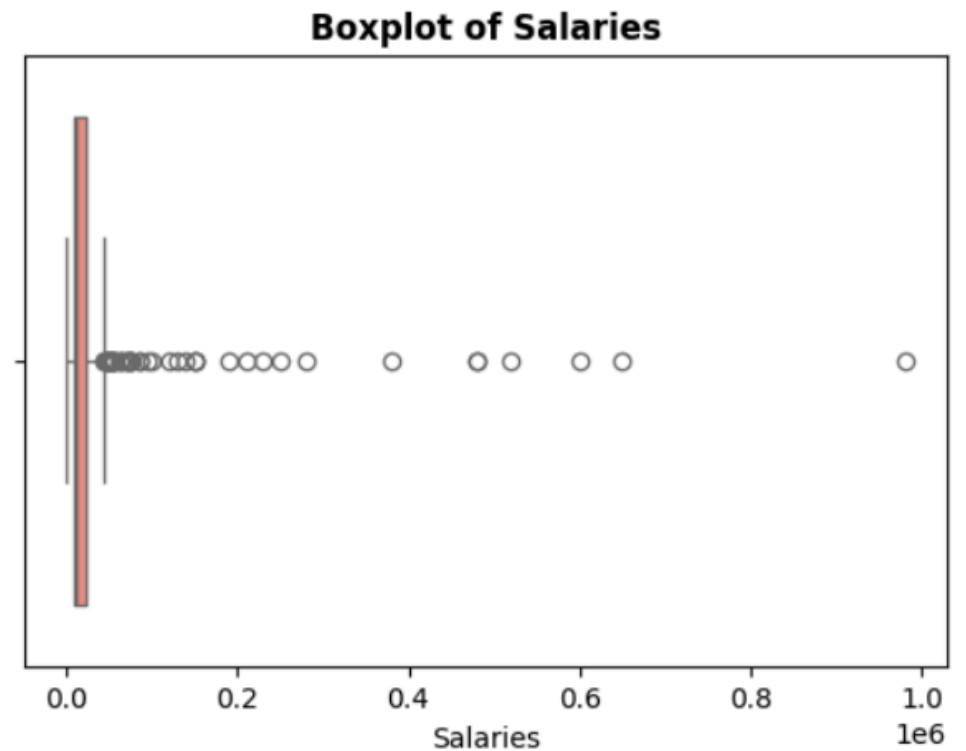
- It groups rating values into intervals and displays how frequently each range occurs.
- It highlights whether ratings are concentrated in a specific range or evenly distributed.
- It helps identify skewness whether ratings lean more toward higher or lower values.



## *Boxplot for “Salaries” Column*

This plot helps visualize how salary values are distributed and identifies extreme variations.

- It displays the median, quartiles (Q1 and Q3), and overall spread of salary data.
- It highlights the interquartile range (IQR), showing where most salary values are concentrated.
- It clearly identifies outliers unusually high or low salary values.

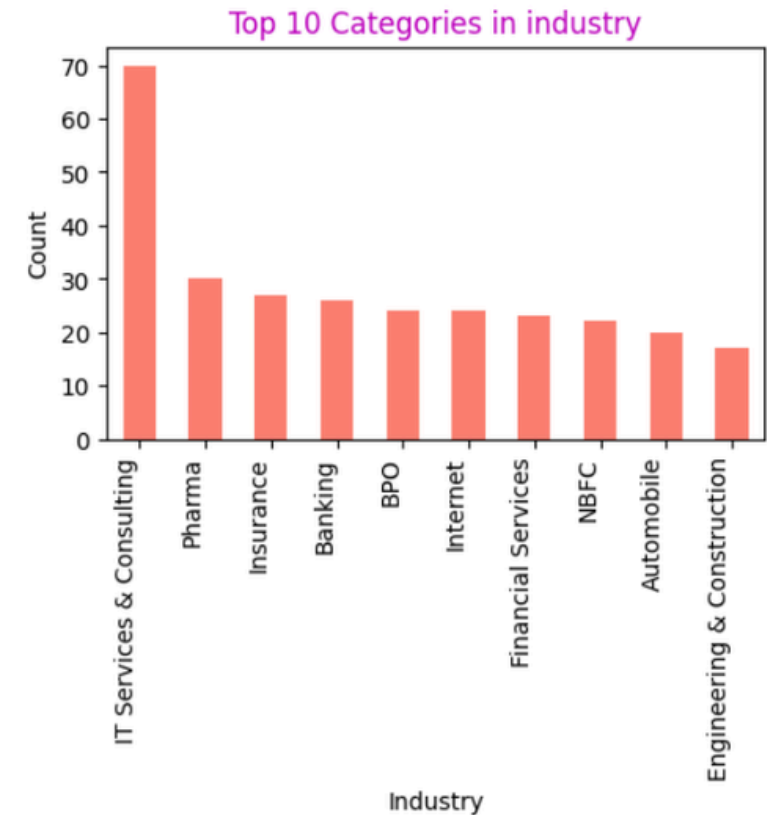


## *Bar Chart for “Industry” Column*

- Bar Chart of Top 10 Categories in Industry

This plot shows the frequency distribution of the top industries present in the dataset.

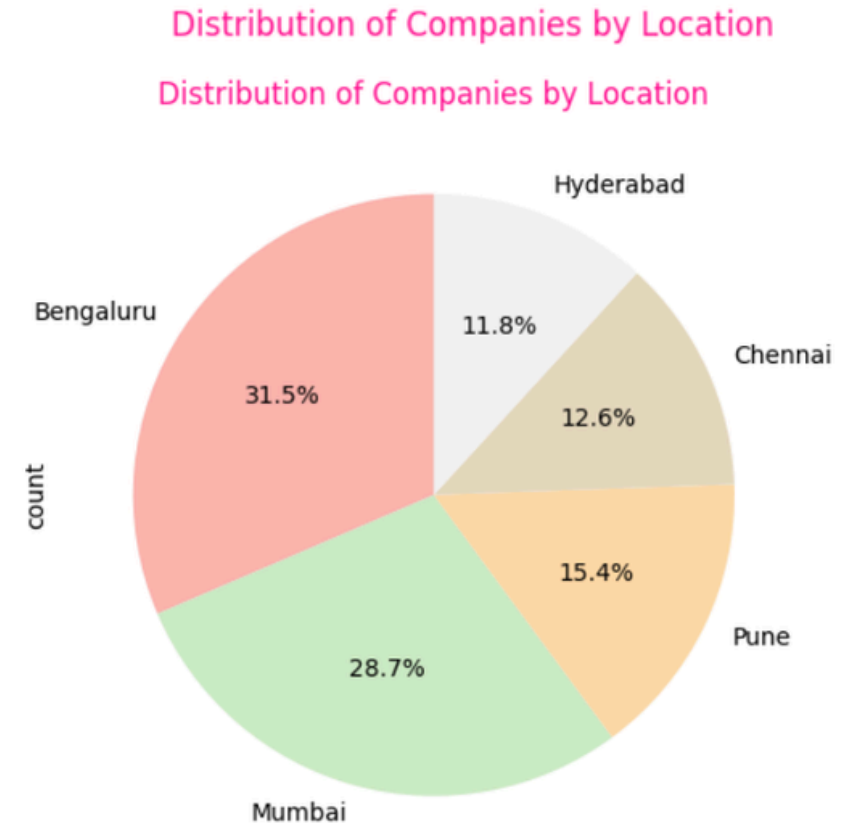
- It displays the count of companies belonging to each industry.
- It helps identify which industry dominates the dataset.
- It highlights variations in representation across different sectors.



## *Pie Chart for “Location” Column*

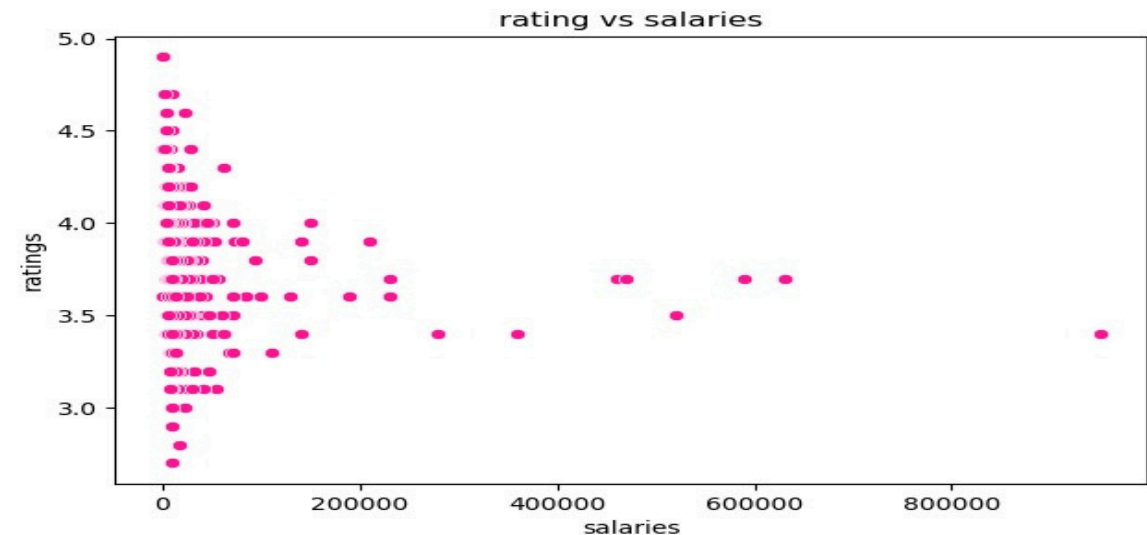
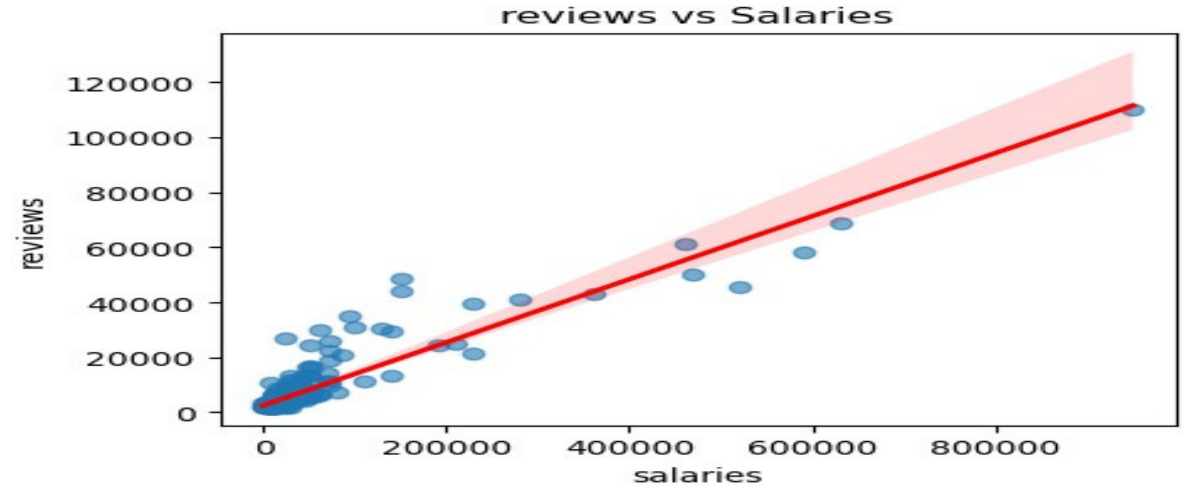
This chart shows the percentage distribution of companies across different cities.

- It represents how companies are geographically distributed in the dataset.
- Each slice indicates the proportion of companies in a specific location.
- It helps identify which city has the highest company concentration.



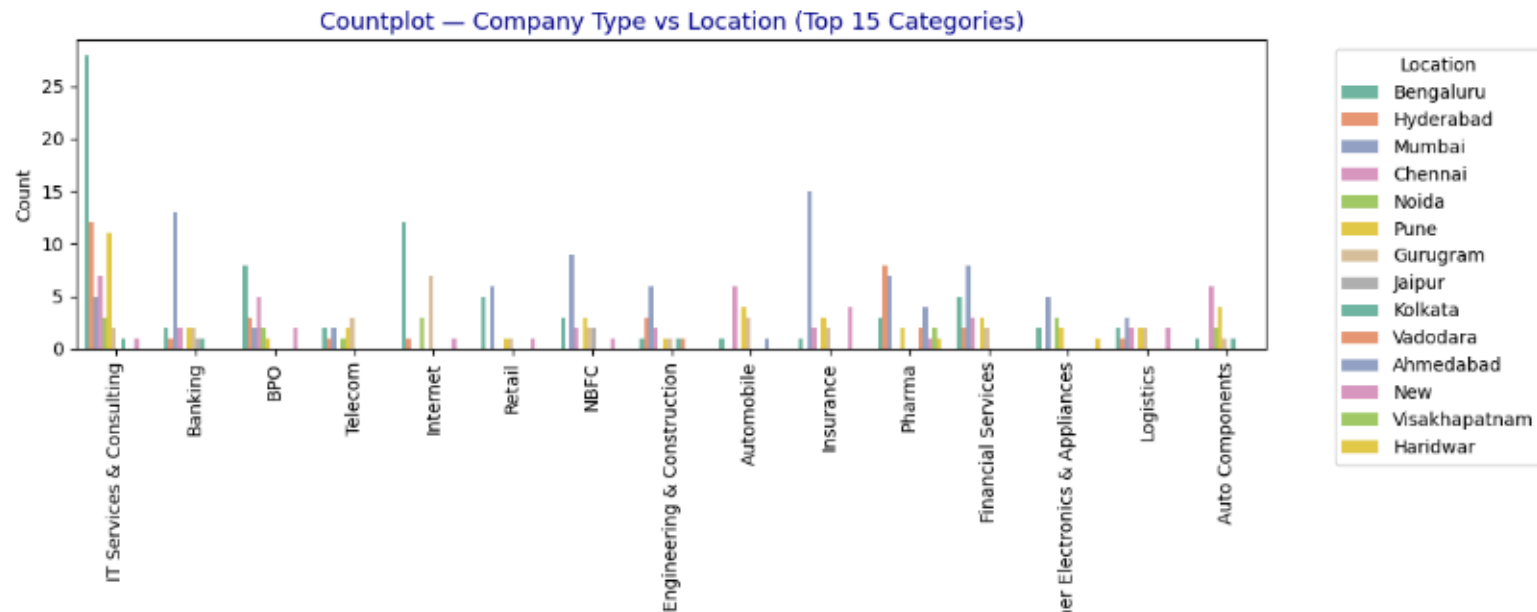
# *Scatter Plot and Regression plot*

- The regression plot of reviews vs salaries shows a clear positive correlation — companies offering higher salaries tend to receive more reviews, indicating strong employee engagement.
- The scatter plot of rating vs salaries reveals that most ratings remain between 3.0 and 4.5 regardless of salary range, suggesting that pay alone does not significantly affect overall ratings



# *Count Plot for “Industry” and “Location”*

- Graph Shows the Distribution of company types across different locations..
- IT Services dominates across most major cities.
- Bengaluru and Hyderabad show strong industry concentration.
- Metro cities have diversified industry presence.
- Smaller cities have limited representation in certain sectors.



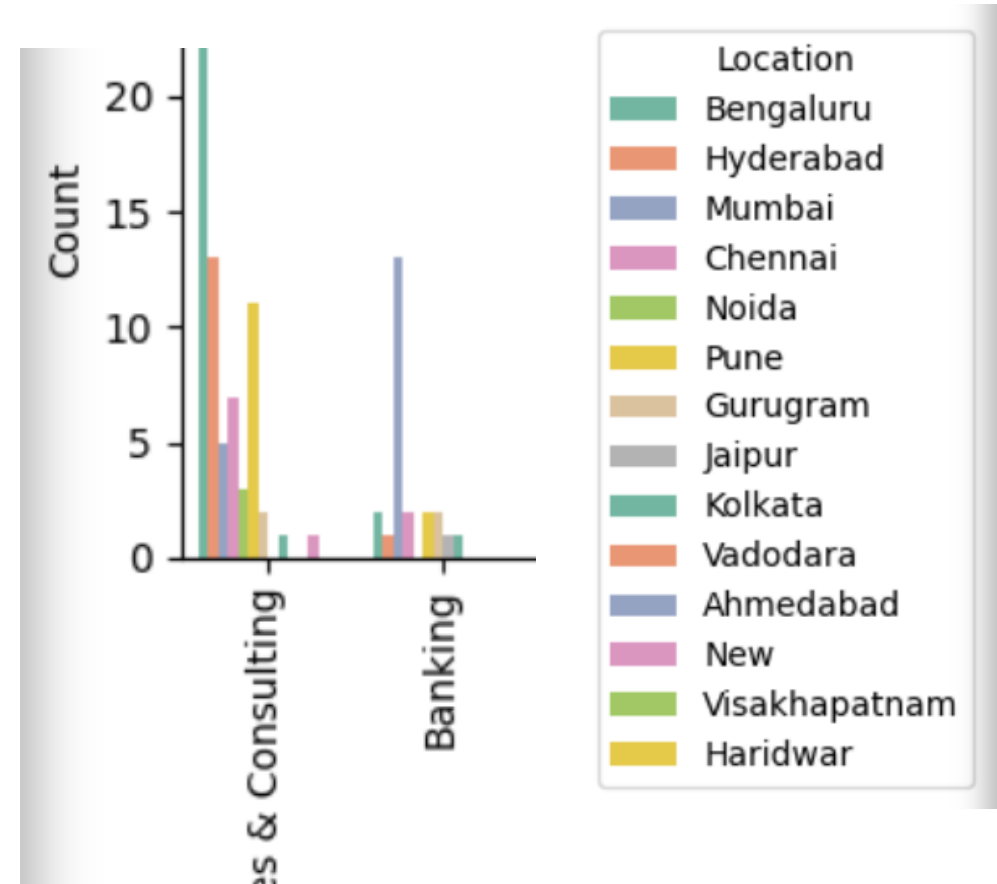
# *Detailed Analysis of Count Plot*

## **IT Services & Consulting vs Location**

- Bangalore dominates IT Services & Consulting.
- Hyderabad shows the second strongest presence.
- Pune and Chennai have moderate representation.
- Other cities like Mumbai, Noida, have smaller counts

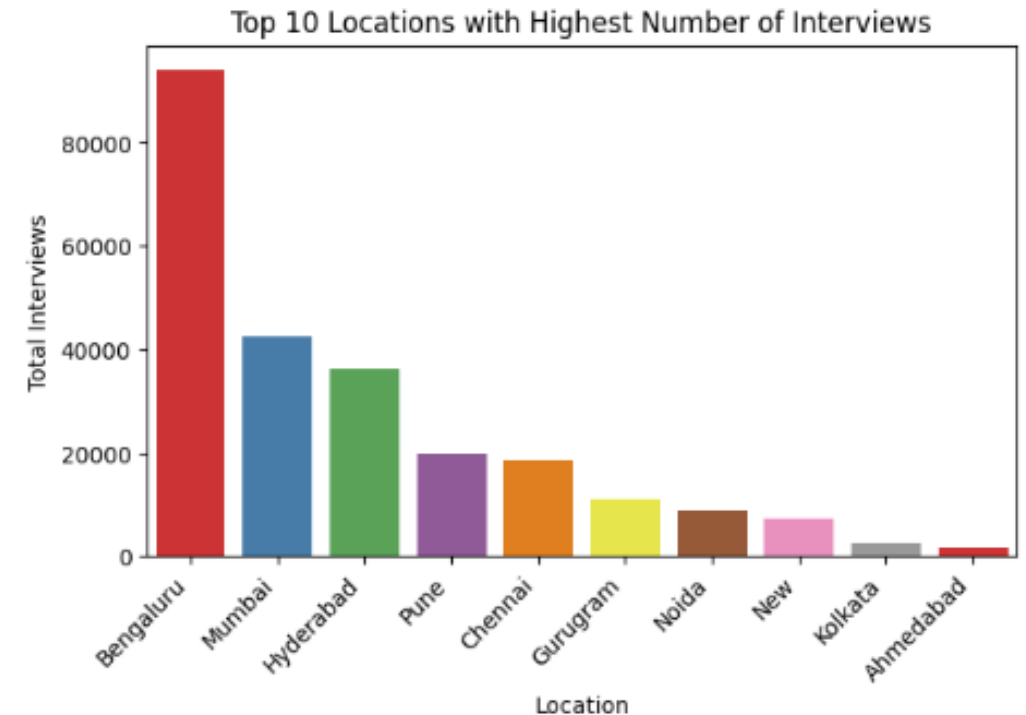
## **Banking vs Location**

- Mumbai has the tallest bar (highest count).
- Hyderabad and Chennai show moderate presence.
- Bangalore has relatively fewer banking companies compared to IT.
- Most other cities have smaller numbers.



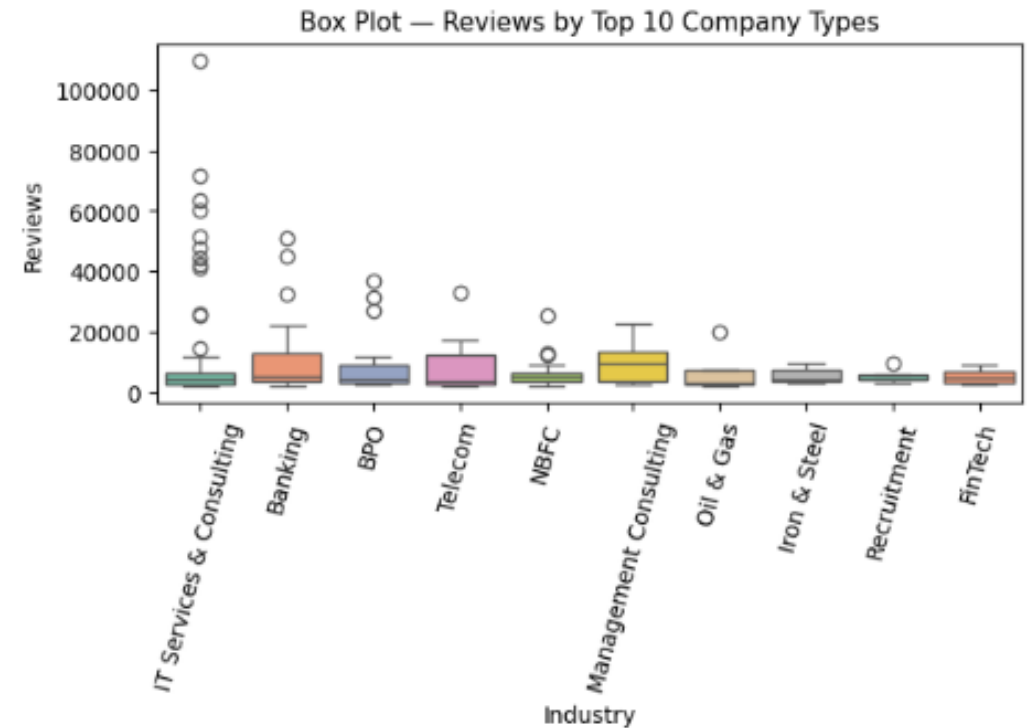
## *Bar Plot for Location and Total Interviews*

- Bengaluru leads with the highest number of interview opportunities, indicating strong hiring activity.
- Mumbai and Hyderabad emerge as the next major recruitment hubs.
- Interview activity is largely concentrated in metro cities.
- A significant gap exists between top-tier cities and lower-ranked locations.
- This suggests that employment opportunities are geographically concentrated in major urban centers.



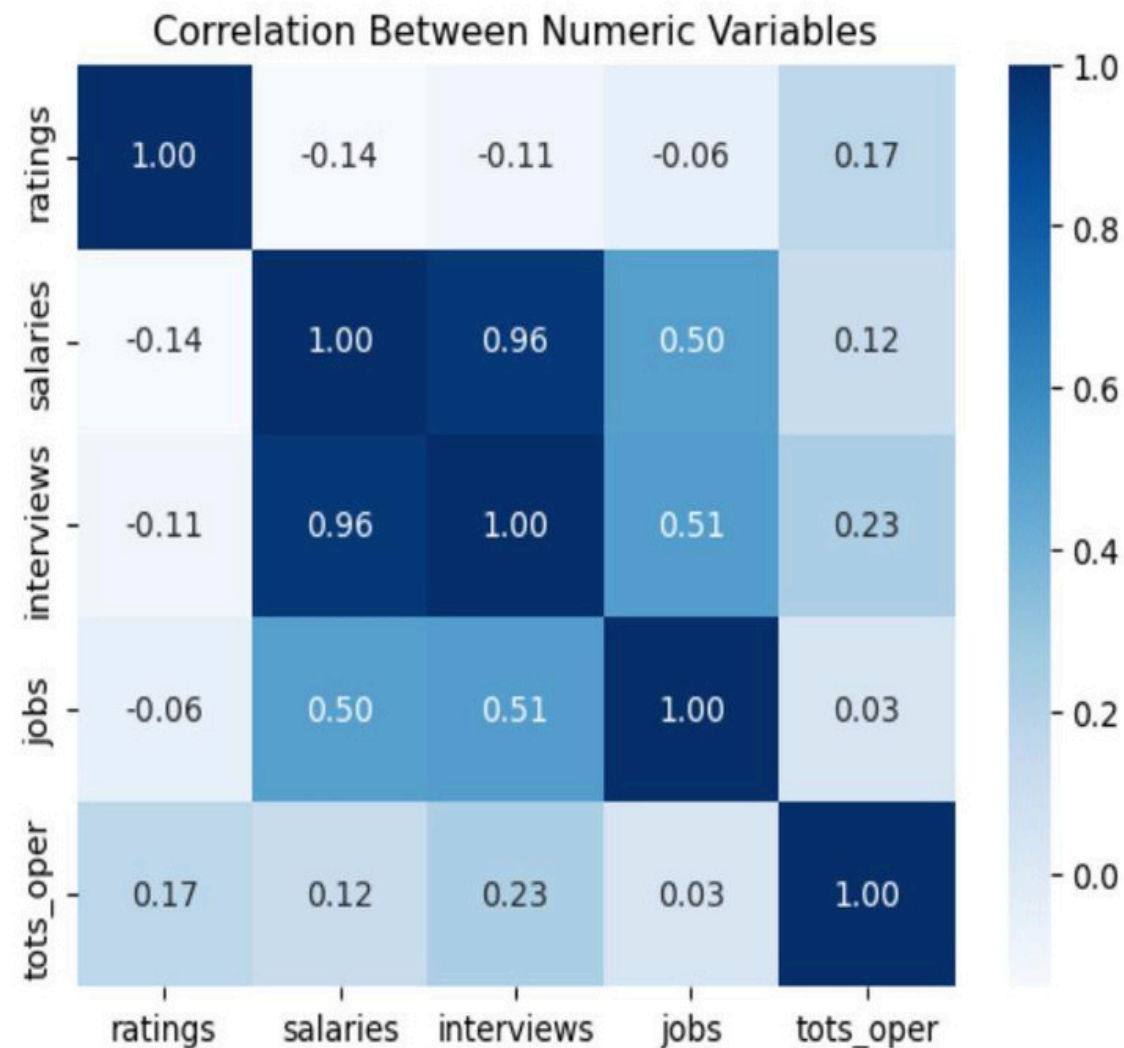
# *Box Plot for Reviews and Industry*

- This box plot shows the distribution of reviews across top 10 industries.
- IT Services & Consulting and Banking show higher review variability.
- Some industries have extreme outliers, indicating very high review counts.
- Median reviews differ across industries, showing variation in employee engagement.



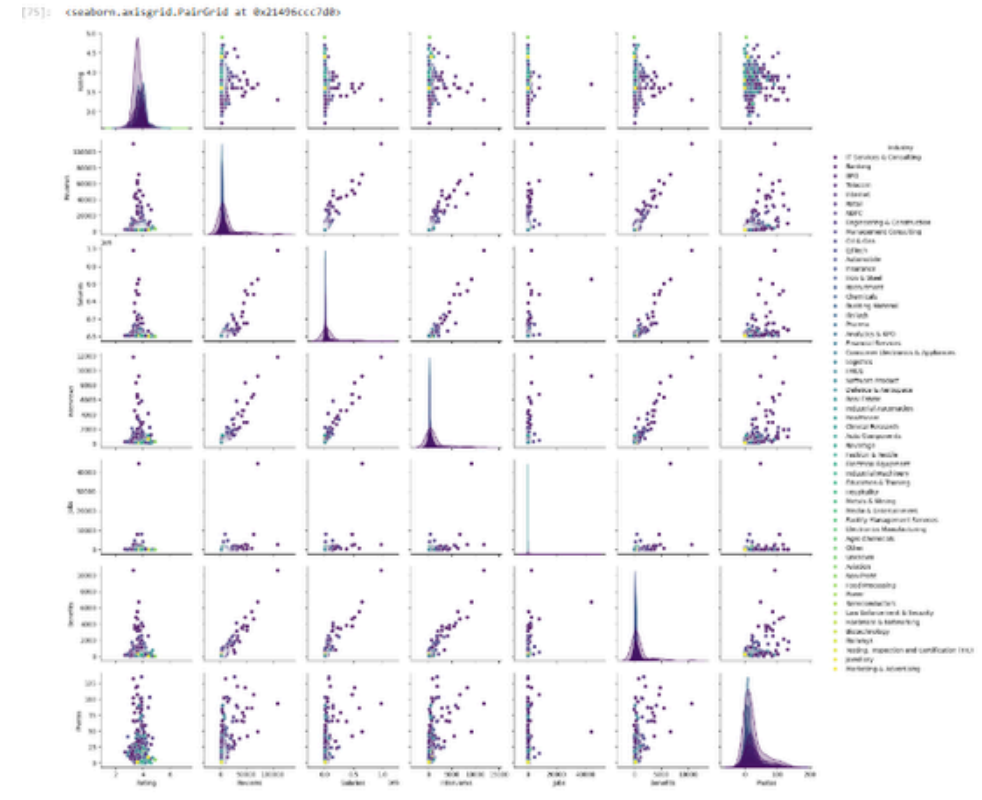
# Correlation Heatmap

- Salaries and Interviews (0.95) show a very strong positive correlation.
- Jobs and Interviews (0.50) have moderate positive correlation.
- Salaries and Jobs (0.49) also show moderate relationship.
- Rating has weak correlation with most variables.
- Total\_Operations\_Spread has low correlation with other features.



## Pair Plot

- Salaries, Interviews, and Reviews show strong positive relationships.
- Higher salary companies tend to have more interviews and reviews.
- Ratings show weak correlation with salary and hiring metrics.
- Most variables are right-skewed (many low values, few very high values).
- Presence of outliers in salary and review counts.



# Analytical Questions & Insights

## Q1. Which Company Has the Highest Rating?

```
[60]: df.sort_values(by="Rating", ascending=False)[["Company", "Rating"]].head(1)
```

```
[60]:
```

	Company	Rating
267	Marpu Foundation	4.9

## Q2. Top 5 Highest Rated Companies

```
[61]: df.sort_values(by="Rating", ascending=False)[["Company", "Rating"]].head(5)
```

```
[61]:
```

	Company	Rating
267	Marpu Foundation	4.9
177	Kogta Financial India Limited	4.7
93	Indian Army	4.7
16	Ienergizer	4.6
261	Indian Air Force	4.6

## Q3. Which Company Has the Most Reviews?

```
[62]: df.sort_values(by="Reviews", ascending=False)[["Company", "Reviews"]].head(1)
```

```
[62]:
```

	Company	Reviews
0	Tcs	622.5

## Q4. Which Location Has the Highest Hiring Activity?

```
[63]: df.groupby("Location")["Interviews"].sum().sort_values(ascending=False).head(1)
```

```
[63]:
```

Location	Interviews
Bengaluru	41788.0

Name: Interviews, dtype: float64

## Q5. Which Industry Has Highest Average Rating?

```
[64]: df.groupby("Industry")["Rating"].mean().sort_values(ascending=False).head(5)
```

```
[64]:
```

Industry	Rating
Jewellery	4.400000
Defence & Aerospace	4.333333
Non-Profit	4.250000
Railways	4.200000
Electrical Equipment	4.200000

Name: Rating, dtype: float64

## Q6. Which Industry Has Highest Average Salary?

```
[65]: df.groupby("Industry")["Salaries"].mean().sort_values(ascending=False).head(5)
```

```
[65]:
```

Industry	Salaries
Agro Chemicals	622.5
NBFC	622.5
Industrial Automation	622.5
Industrial Machinery	622.5
Insurance	622.5

Name: Salaries, dtype: float64

## Q7. Is Salary Strongly Related to Rating?

```
[66]: df[["Salaries", "Rating"]].corr()
```

```
[66]:
```

	Salaries	Rating
Salaries	1.000000	-0.103849
Rating	-0.103849	1.000000

## Insight:

- Correlation value is low.
- Salary is not strong predictor of satisfaction.

## Q8. Which Company Has Maximum Operations Spread?

```
[67]: df.sort_values(by="Total_Operations_Spread", ascending=False)[["Company", "Total_Operations_Spread"]].head(1)
```

```
[67]:
```

Company	Total_Operations_Spread
88 Chola Mandalam Investment Finance	622.5

# Analytical Questions & Insights

## Q9: Which Industry Receives Most Criticism?

```
j: df["Critically_Rated_For"].value_counts().head(5)
```

```
j: Critically_Rated_For
Promotions / Appraisal      166
Not Mentioned                92
Promotions / Appraisal, Salary & Benefits, Work Satisfaction  37
Promotions / Appraisal, Salary & Benefits                    35
Promotions / Appraisal, Company Culture, Work Satisfaction  13
Name: count, dtype: int64
```

## Q10: Top 10 Companies with Most Job Openings

```
j: df.sort_values("Jobs", ascending=False)[["Company", "Jobs"]].head(10)
```

```
j:
      Company  Jobs
0          Tcs  622.5
124       Cgi Group  622.5
48    Maruti Suzuki  622.5
52         Paytm  622.5
30    Ernst Young  622.5
54    Bajaj Finance  622.5
342  Fortis Healthcare  622.5
58        Exl Service  622.5
380    Adani Group  622.5
407  Apollo Pharmacy  622.5
```

## Q11: Which Location Offers Highest Average Salary?

```
j: df.groupby("Location")["Salaries"].mean().sort_values(ascending=False).head()
```

```
j: Location
Ahmedabad      622.5
Jaipur         622.5
Jamnagar       622.5
Jamshepur      622.5
Kanpur         622.5
Name: Salaries, dtype: float64
```

## Q12: Most Frequently Highly Rated Aspects ?

```
j: df["Highly_Rated_For"].value_counts().head(5)
```

```
j: Highly_Rated_For
Not Mentioned      161
Job Security        36
Job Security, Company Culture, Work Life Balance  24
Salary & Benefits  22
Work Life Balance   18
Name: count, dtype: int64
```

## Q13: Most Common Industries ?

```
j: df["Industry"].value_counts().head(5)
```

```
j: Industry
IT Services & Consulting  71
Pharma                   30
Insurance                27
Banking                  26
Internet                 24
Name: count, dtype: int64
```

## Q14: Most Frequently Criticized Aspects ?

```
j: df["Critically_Rated_For"].value_counts().head(10)
```

```
j: Critically_Rated_For
Promotions / Appraisal      166
Not Mentioned                92
Promotions / Appraisal, Salary & Benefits, Work Satisfaction  37
Promotions / Appraisal, Salary & Benefits                    35
Promotions / Appraisal, Company Culture, Work Satisfaction  13
Promotions / Appraisal, Job Security                        11
Promotions / Appraisal, Job Security, Work Satisfaction     11
Promotions / Appraisal, Salary & Benefits, Company Culture  10
Job Security, Promotions / Appraisal, Company Culture        9
Promotions / Appraisal, Salary & Benefits, Skill Development / Learning  9
Name: count, dtype: int64
```

```
j: df["Company"].value_counts()
```

```
j: Company
Tcs          1
Elasticrun   1
Fortis Healthcare  1
Nobroker     1
Movate       1
Oppo         1
Dell         1
Havells      1
Cars24       1
Motherson Automotive Technologies And Engineering  1
```

# *Challenges*

## **1.403 Forbidden Error During Scraping**

- Website blocked automated requests.
- Required adding proper headers (User-Agent).

## **2.Inconsistent Data Formats**

- Salary values in “K” and “L” format.
- Mixed data types (object instead of numeric) and Special characters like [ ] '.

## **3.Hidden Missing Values**

- Some nulls were not visible before type conversion.
- Conversion exposed additional missing values.

## **4.Extreme Outliers**

- Salary and review columns were highly skewed.
- Required IQR-based outlier detection

# Conclusion

- **Strong Salary–Interview Relationship**

Salaries and interviews show the highest positive correlation, indicating that companies conducting more interviews generally offer higher pay.

- **Bangalore as the Key Employment Hub** Bangalore leads in both salary range and interview volume, followed by Hyderabad and Mumbai as major recruitment centers.

- **IT Services Dominate the Market** IT Services & Consulting holds over 60% industry share and ranks among the top salary-paying sectors, followed by Management Consulting and NBFCs.

- **Positive Company Reputation and Ratings** Most companies maintain ratings between 3.5 and 4.2; TCS, Accenture, and Wipro lead with the highest positive reviews and strong employee approval.

**Thank You**  
For Your Attention!

Any Questions

