INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

AmbitionBox

Title: Web Scraping and Exploratory Data Analysis of AmbitionBox Company Data

About Me

- I am Bhargavi B.Sc graduate with a strong passion for **Data Analytics and Visualization**.

- Skilled in **exploring real-world datasets** to uncover meaningful insights.

- Experienced in **web data extraction and cleaning** using Python tools.

- Continuously learning to enhance expertise in **data science and analytics tools**.

- **GITHUB –** https://github.com/bhargavi7122

- **LINKEDIN –** http://www.linkedin.com/in/kundrapu-bhargavi

Preface

- INTRODUCTION

- BUSINESS PROBLEM

- OBJECTIVE

- WEB SCRAPPING

- TOOLS USED

- DATA CLEANING STEPS

- DATA VISUALIZATION

- CONCLUSION

INNOMATICS
RESEARCH LABS

Introduction

- AmbitionBox is a leading Indian platform that provides company reviews, salary insights, interview experiences, and job listings.

- Founded in **2015** by IIT Madras alumni, acquired by Naukri.com (Info Edge India).

- Helps job seekers make informed career decisions through authentic employee feedback.

- Offers company ratings, benefits data, and workplace culture insights.

INNOMATICS
RESEARCH LABS

Business Problem

- Employee perception data is **unstructured and scattered** across online platforms like AmbitionBox.
- Companies struggle to **benchmark their reputation and employee satisfaction** against competitors.
- 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.
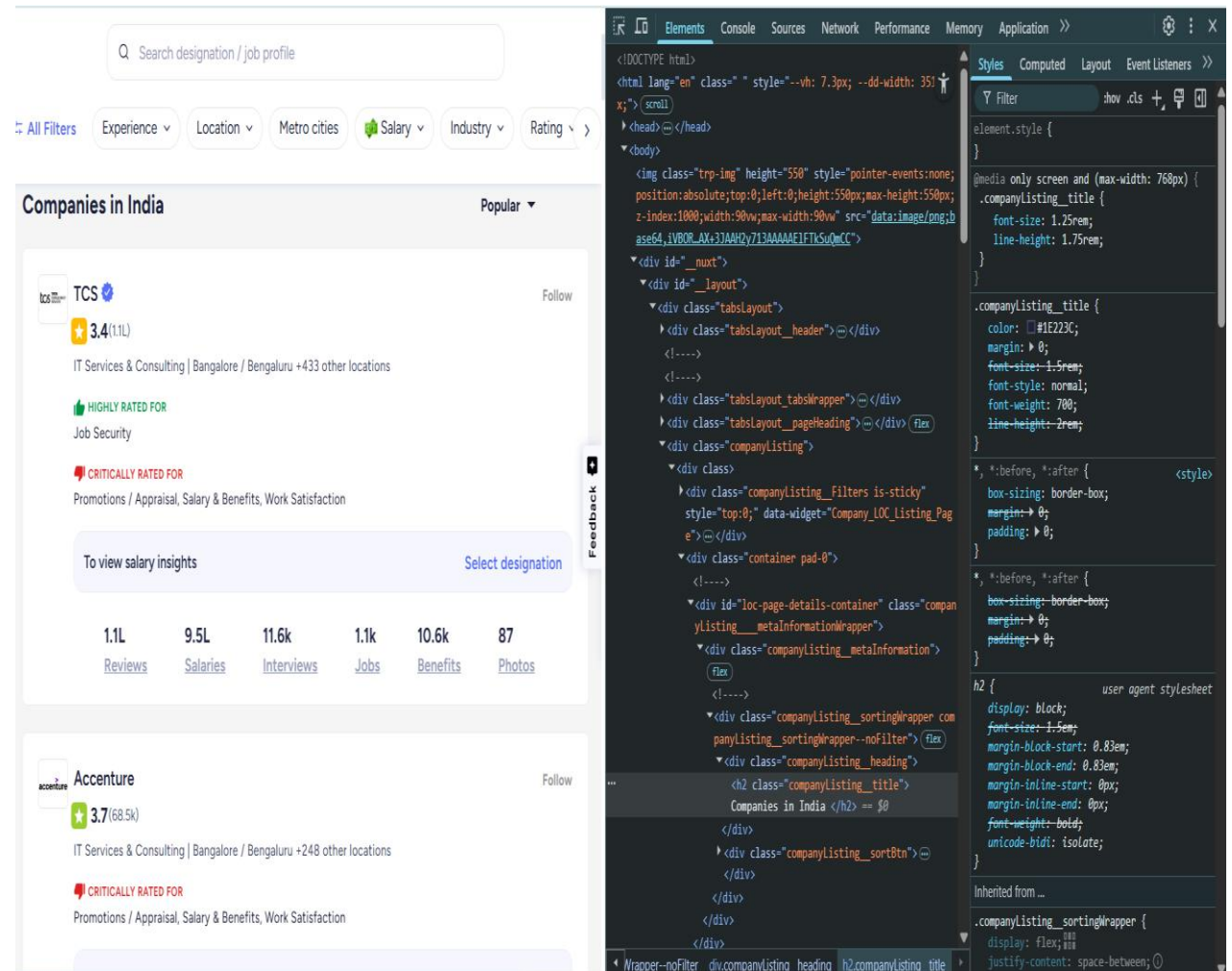
Objectives

- To **analyze and compare companies** using employee reviews, ratings, and salary insights from AmbitionBox..

- 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.

Web Scraping

- Ambition box was selected as the data source

- Used **browser developer tools** (Inspect Element) to locate relevant HTML tags Data.

- Used **BeautifulSoup and Requests** to extract data from Ambition box.

- Sent **HTTP requests** to fetch web pages.

- Performed **Exploratory Data Analysis (EDA)** to identify trends.

Tools Used:

Beautiful Soup

pandas

.[RegEx]*

matplotlib

NumPy

seaborn

Dataset

Uncleaned data

	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

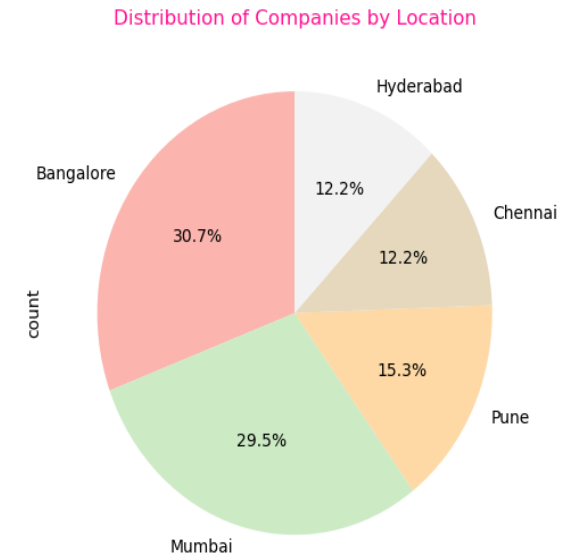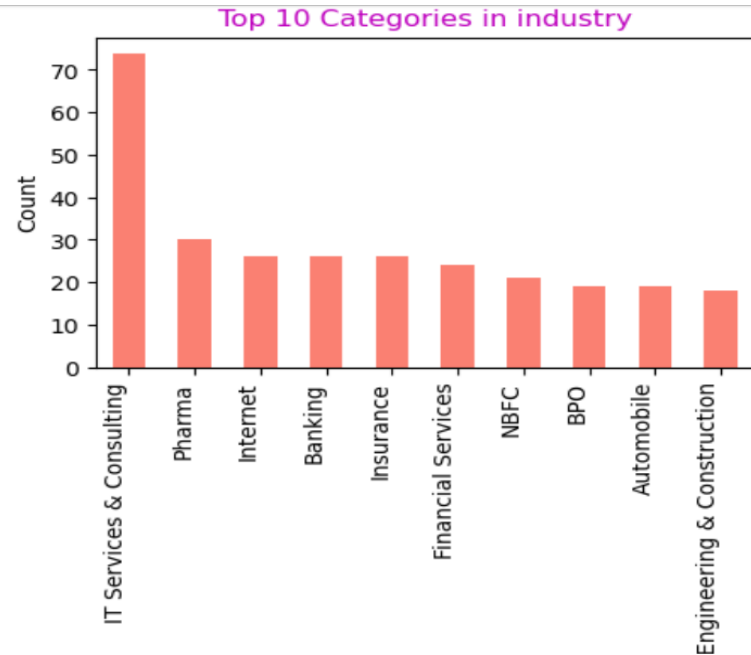INNOMATICS RESEARCH LABS

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	ratings	types	location	reviews	salaries	interviews	jobs	benefits	photos	high	low	tots_oper
0	TCS	3.4	IT Services & Consulting	Bangalore	110000.0	950000.0	11600.0	1100.0	10600.0	87.0	Job Security	Promotions / Appraisal, Salary & Benefits, Wor...	434
1	Accenture	3.7	IT Services & Consulting	Bangalore	68400.0	630000.0	9000.0	27000.0	6700.0	48.0	NaN	Promotions / Appraisal, Salary & Benefits, Wor...	249
2	Wipro	3.7	IT Services & Consulting	Hyderabad	61300.0	470000.0	6400.0	4400.0	4600.0	99.0	Job Security	Promotions / Appraisal, Salary & Benefits, Wor...	368
3	Cognizant	3.7	IT Services & Consulting	Hyderabad	58200.0	590000.0	6200.0	778.0	5500.0	82.0	NaN	Promotions / Appraisal, Salary & Benefits, Wor...	229
4	Capgemini	3.7	IT Services & Consulting	Bangalore	49900.0	470000.0	5300.0	1900.0	3600.0	41.0	Work Life Balance, Job Security	Promotions / Appraisal, Salary & Benefits, Wor...	182
5	HDFC, Bank	3.8	Banking	Mumbai	48300.0	150000.0	2800.0	173.0	3200.0	72.0	Job Security, Skill Development / Learning	Promotions / Appraisal	1787

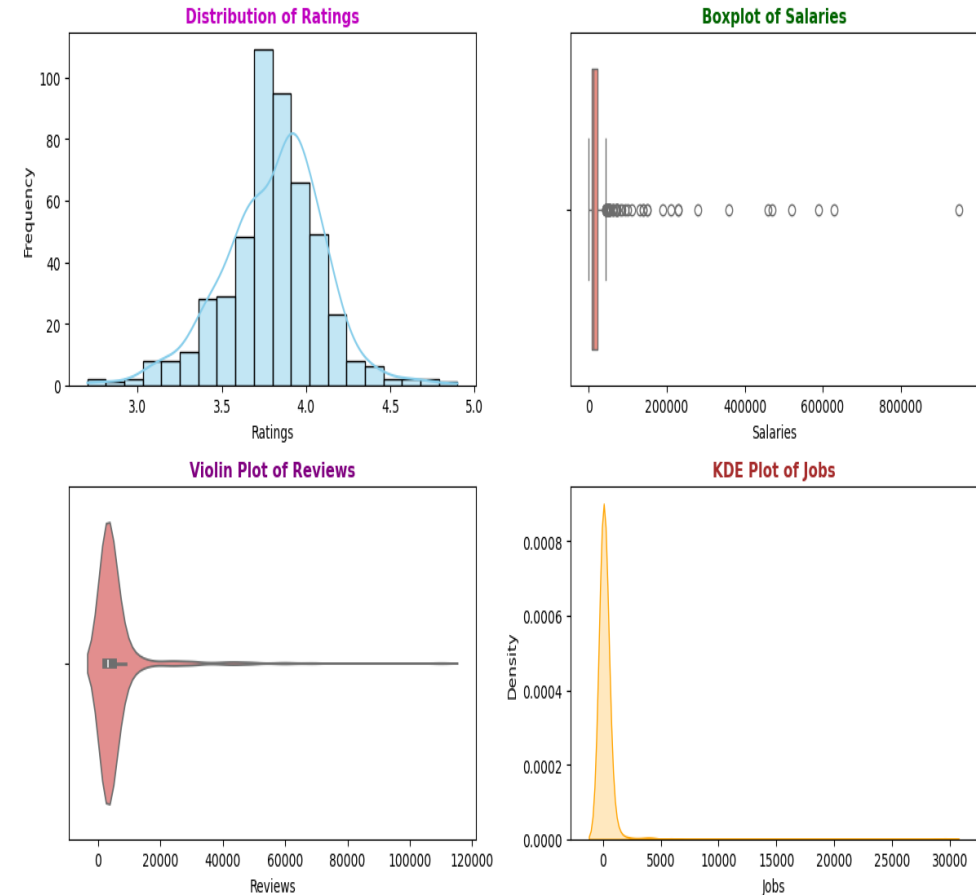# Data Visualization: Univariate Analysis (Categorical)

- The **bar chart** for Top 10 Categories in Industry shows that IT Services & Consulting dominates, followed by Pharma, Internet, and Banking, highlighting the strong demand in the IT sector.

- The **pie chart** for Distribution of Companies by Location indicates that Bangalore (30.7%) and Mumbai (29.5%) are major hubs, with Pune, Chennai, and Hyderabad also contributing significantly.

- The **horizontal bar chart** for Criticality Rates reveals that Promotions/Appraisals are most important to employees, followed by Salary & Benefits and Work Satisfaction.

Top 10 Categories in industry

Distribution of Companies by Location

Most Common critically rates

Univariate Analysis(Numerical)

- The **histplot** was used to analyze the *ratings* variable. The histogram with KDE overlay shows that most companies have moderate to high ratings, reflecting overall positive employee experiences.
- The **boxplot** for *salaries* highlights the median, quartiles, and outliers, allowing for comparison of pay scales. It reveals that while many companies offer average pay ranges, a few provide significantly higher salaries.
- The *reviews* column was visualized using a **violin plot**, showing the spread and density of review counts. It indicates that most companies receive a moderate number of employee reviews, while a select few stand out with extensive feedback.
- The *jobs* variable was explored using a **KDE plot**, which shows the distribution of available job positions. The plot suggests that most companies have limited openings, but some larger firms post a notably higher number of jobs.
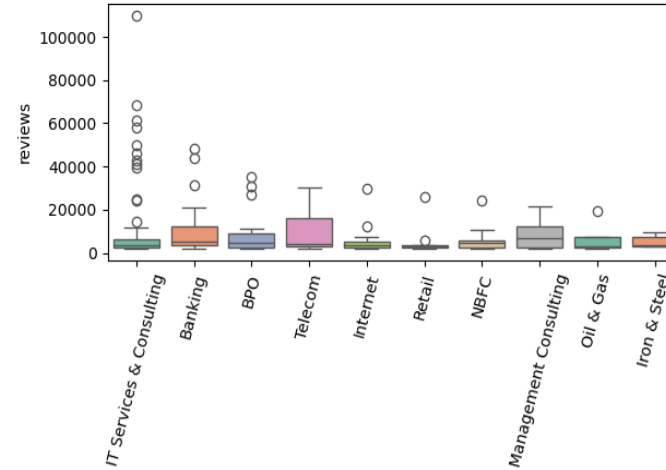

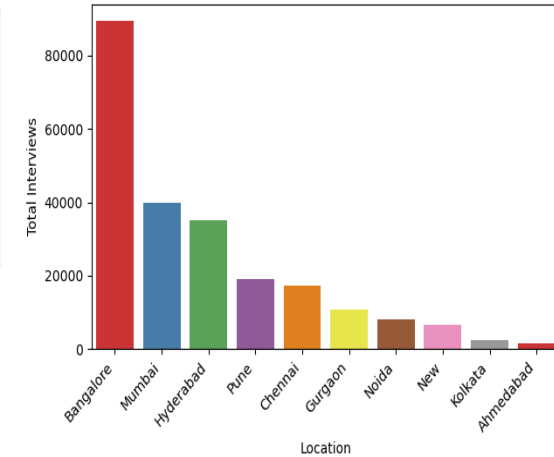Univariate Analysis – Numerical Variables

Bivariate Analysis(Categorical vs Numerical)

- The **boxplot** for reviews by company type shows that IT Services & Consulting firms receive the highest reviews, while sectors like Retail and NBFC have fewer.

- The **bar chart** for top interview locations highlights Bangalore as the major hub, followed by Mumbai and Hyderabad.

- The **horizontal bar plot** for top companies by positive reviews shows TCS, Accenture, and Wipro leading with strong employee feedback.
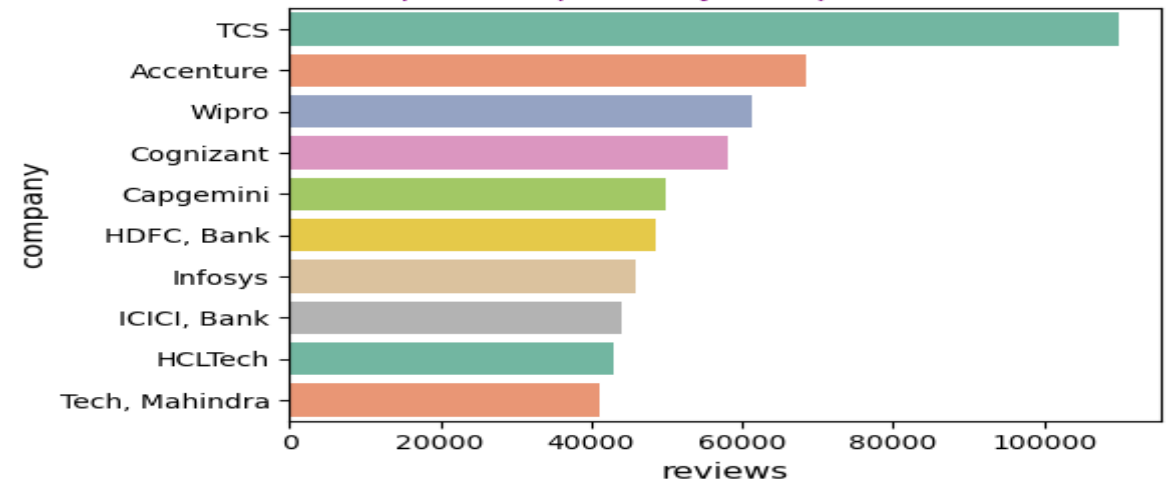

Box Plot — Reviews by Top 10 Company Types

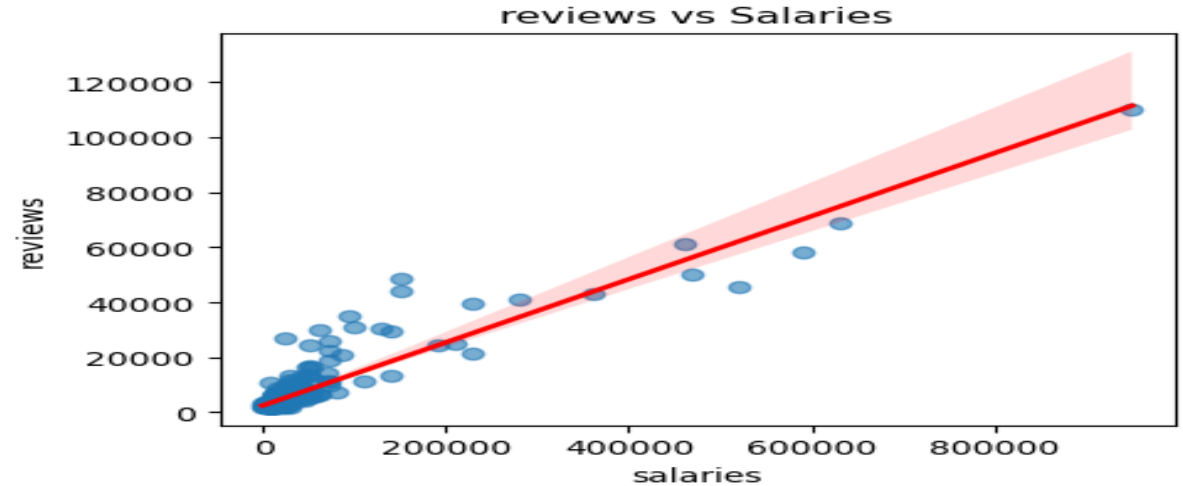Top 10 Locations with Highest Number of Interviews

Top 10 companies by total positive reviews

INNOMATICS
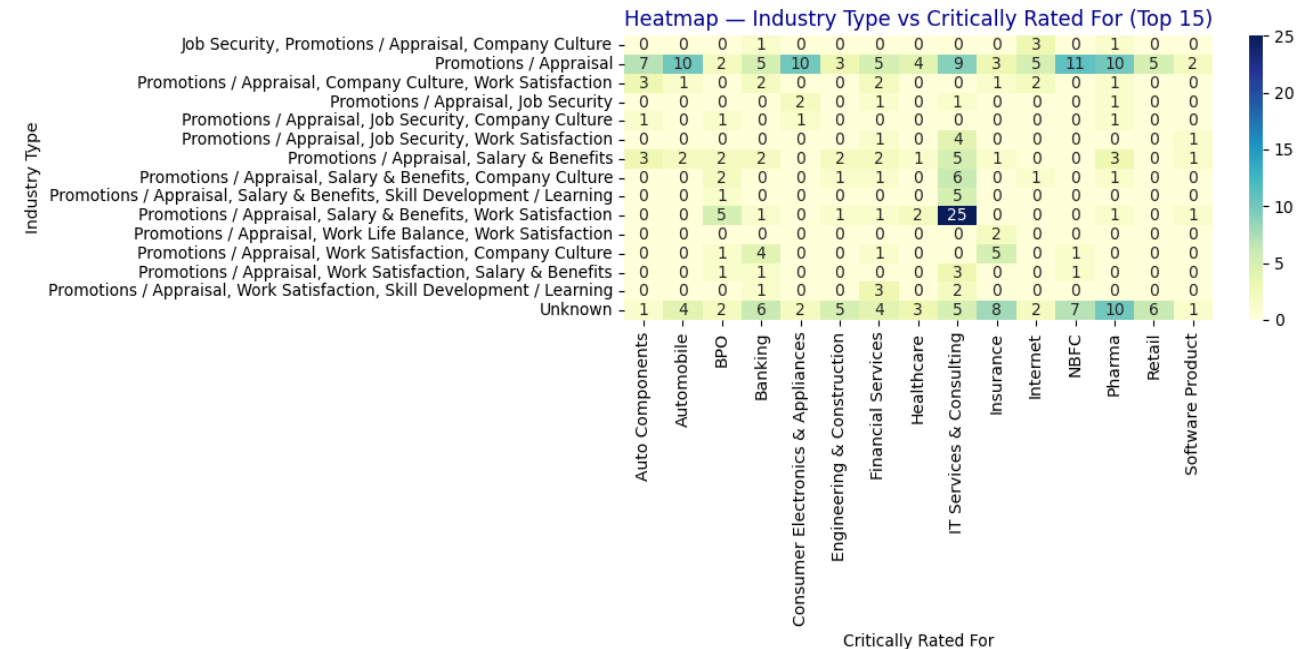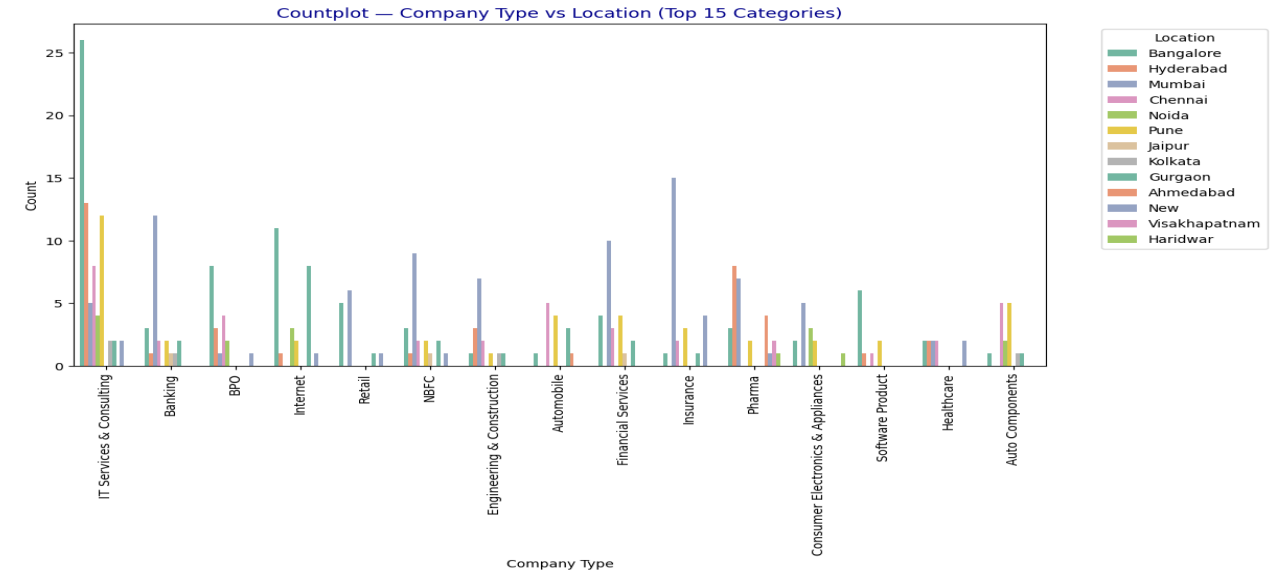RESEARCH LABS

Bivariate Analysis(Numerical vs Numerical)

- 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.

reviews vs Salaries

rating vs salaries

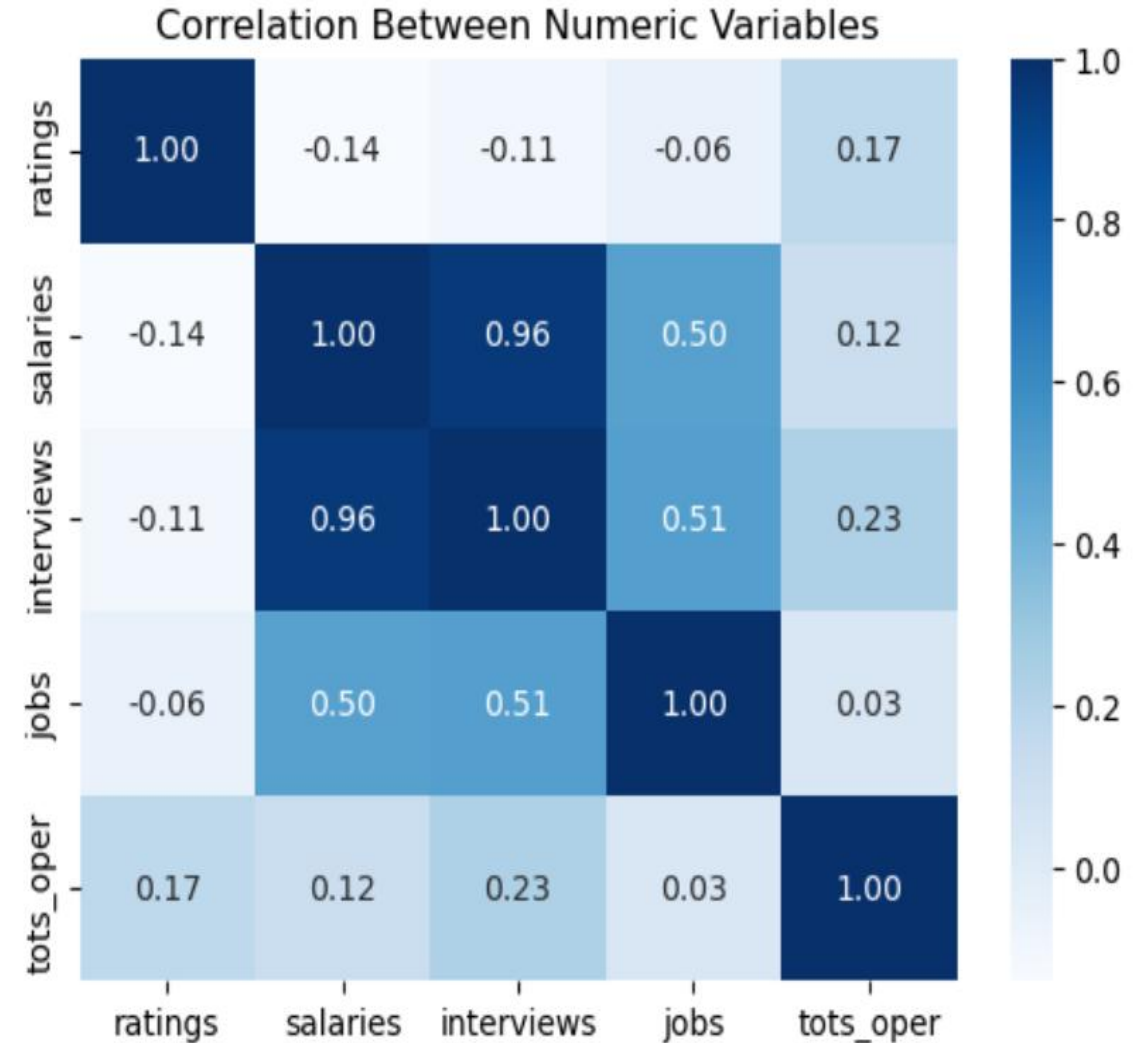# Bivariate Analysis (Categorical vs Categorical)

- The **countplot** compares *company types across locations*, showing that IT Services, Consulting, and Internet-based firms dominate in metro cities like **Bangalore, Hyderabad, and Pune**.

- It reveals that certain industries are highly concentrated in specific regions, indicating regional specialization.

- The **heatmap** highlights relationships between *industry types* and *critical ratings*, allowing easy identification of which industries face more negative feedback.

- Together, they help understand how **location and industry type influence employee experiences** and perceptions across companies.

Countplot — Company Type vs Location (Top 15 Categories)

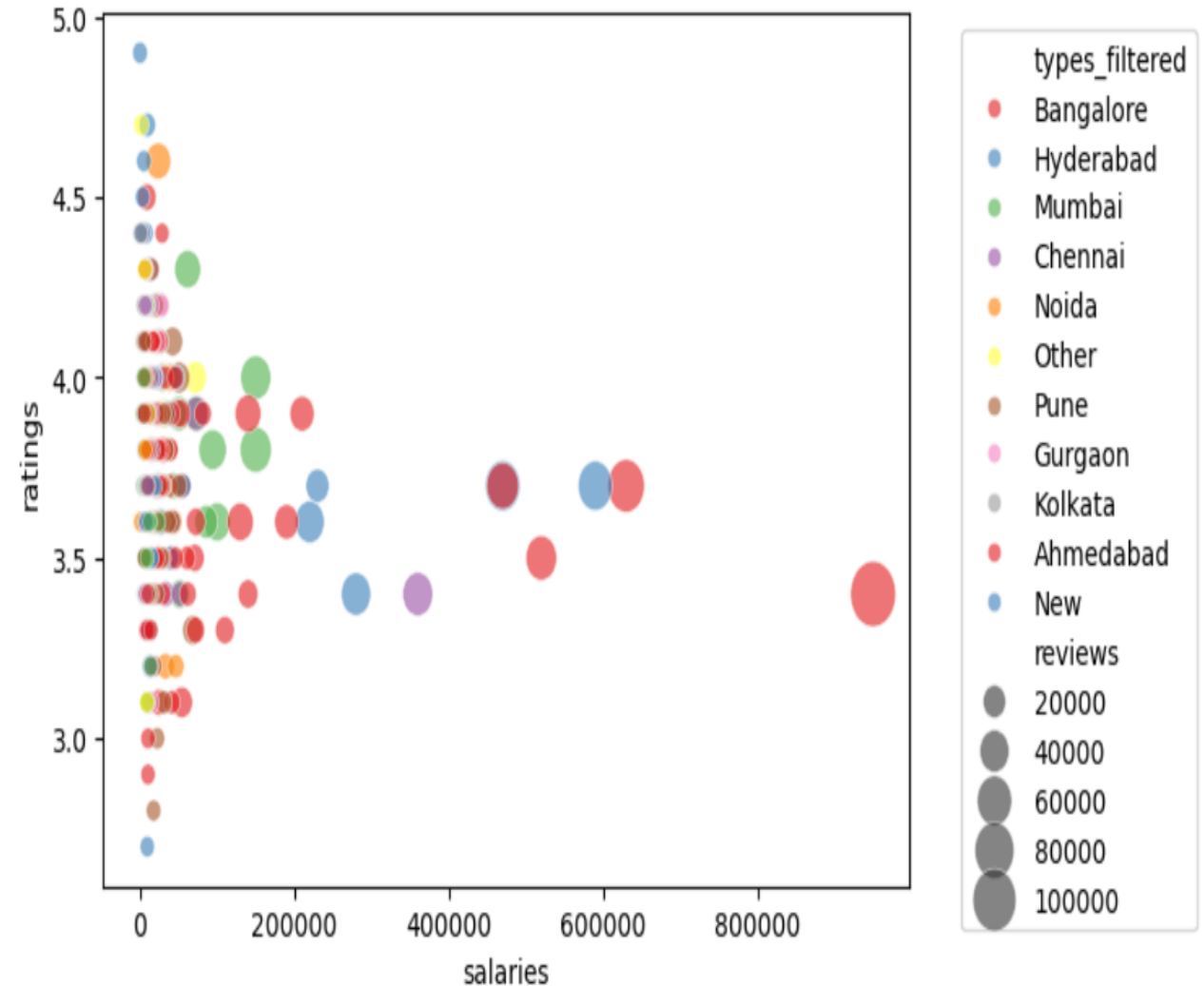Heatmap — Industry Type vs Critically Rated For (Top 15)

Multivariate Analysis

- Salaries and interviews have the strongest positive correlation (0.96) among all variables.

- Jobs show moderate positive correlation with both salaries (0.50) and interviews (0.51).

- Ratings display very weak or negative correlations with other numeric variables.

- Operational totals have little association with other variables, with the highest correlation at just 0.23.

Correlation Between Numeric Variables

	ratings	salaries	interviews	jobs	tots_oper
ratings	1.00	-0.14	-0.11	-0.06	0.17
salaries	-0.14	1.00	0.96	0.50	0.12
interviews	-0.11	0.96	1.00	0.51	0.23
jobs	-0.06	0.50	0.51	1.00	0.03
tots_oper	0.17	0.12	0.23	0.03	1.00

Multivariate Analysis

- **Salaries vs. Ratings**: Most data points are clustered at lower salaries (0-200000) with ratings between 3.0 and 4.5. Higher salaries aren't strongly associated with higher ratings; in fact, higher salaries seem to have ratings mostly between 3.5 and 4.
- Bangalore (red) has a wide spread of salaries and ratings.
- Other cities like Hyderabad (blue), Mumbai (green), Chennai (purple) have points mostly at lower salaries.
- **Number of Reviews (dot size)**: Bigger dots (more reviews) aren't clearly linked to higher ratings or salaries based on this plot.

INNOMATICS
RESEARCH LABS

Conclusion

1. **Strong Salary–Interview Relationship**

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

2. **Bangalore as the Key Employment Hub**

   Bangalore leads in both salary range and interview volume, followed by Hyderabad and Mumbai as major recruitment centers.

3. **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.

4. **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.

INNOMATICS
RESEARCH LABS

Experiences & Challenges

- Extracted large datasets from dynamic web pages using Python (BeautifulSoup, Requests).

- Faced challenges with inconsistent data formats and missing values during cleaning.

- Automated scraping process to handle multiple pages and reduce time.

- Analyzed salaries, ratings, and job trends to generate meaningful insights.

- Ensured ethical data collection and handled website rate limits effectively.

- Learned end-to-end data workflow: **scraping→ cleaning→ analysis→ visualization.**

THANK YOU