EXPLORATORY DATA ANALYSIS OF INTERNSHIPS IN INDIA

Team Members

ARUL R & SARANYA S

5/7/2025

PROBLEM STATEMENT

To analyze internship opportunities across locations, stipends, roles, and job offers using data from ***Internshala.com***.

OBJECTIVES

Understand key trends in stipend, location, hiring type, and early application behavior.

TOOLS USED FOR EDA

Python (3.12)+ Jupyter Notebook

Python libraries **,**

For data extract : BeautifulSoup (Web Scraping), Random, Request, Time, numpy , Regular Expressions.

For data analysis : Pandas, Datetime

For data visualization : Matplotlib, Seaborn.

DATASET OVERVIEW

DATASET SOURCE

Scraped from Internshala using BeautifulSoup (web scraping)

DATASET SHAPE

6050 rows x 10 columns

KEY COLUMNS

intern_role, company_name, hiring_type, location,job_offer(LPA), stipend_per_month, duration(months), posted_time, early_applicant, apply_links

ANALYZING DATA

COLUMN CLEANING

1. Column name : **job_offer(LPA)**

After removing the currency symbol from the **job_offer(LPA)** column, the values changed from **₹ 3.5** to **3.5**, and the data type converted from **objec**t to **float**.

2. Column name : posted_time

The **posted_time** column had values like **"just now", "1 day ago", etc.,** stored as text. We converted them to numbers **(e.g., "today" → 0, "1 day ago" → 1)**, then subtracted these from the data extraction date **(2025-06-23**) to get the exact **posted_date.** The data type was changed from **object** to **datetime64[ns].**

3. Column name : hiring_type

The **hiring_type** column had values like **“actively_hiring"** and **NaN**. We cleaned it by converting **“actively_hiring"** to **"yes"** and missing values **(NaN)** to **"no",** assuming companies not listed as actively hiring are treated as "no".

4. Column name : early_applicant

The **early_applicant** column had values like **"Early Applicant"** and **NaN.** We converted **"Early Applicant"** to **"yes"** and missing values to **"no",** assuming no info means not early. This cleaning was done to improve visualization.

REMOVING DUPLICATES

- Duplicate rows are exact copies of other rows in the dataset.
- First, check duplicates.
- If duplicates are found, delete the rows.
- In my dataset, found **“26 rows × 10 columns”** duplicates and deleted permanently.
- Removing duplicates helped improve the quality of analysis and ensured each internship record was unique.

HANDLING NAN VALUES

Checked null values :

code for check null or nan values with count,

df.isna().sum()

Column	Missing Values
duration(months)	477
Other columns	0

Filled missing values :

Calculated the mode for the numeric column **duration(months)** and replaced missing **(NaN)** values with the **mode**.

After identifying the mode, we used it to fill the missing values.

NUMERIC ESTIMATION/OUTLIER REMOVAL

What is outliers ?

Outliers are data points that are significantly higher or lower than the rest of the dataset. They can distort analysis and lead to misleading insights.

Column name: stipend_per_month

Ex: stipend_per_month – 5 Too low

stipend_per_month – 300000 Too high

Method used to handle : IQR Method (Interqurtile Range)

```
Q1 = df['stipend_per_month'].quantile(0.25) # 0.25 → 25% of data points
Q3 = df['stipend_per_month'].quantile(0.75) #0.75 → 75% of data points
IQR = Q3 - Q1
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
```

Filter out outliers

```
df = df[(df['stipend_per_month'] >= lower_bound) & (df['stipend_per_month'] <= upper_bound)]
```

In my dataset, I found many outliers. For those columns with outliers, I replaced the NaN values for the outliers.

- Median or Mode ?

  If outliers are found in a column, fill the NaN values with the median. If there are no outliers, fill the NaN values with the mode.

- Replaced the NaN values with the **median** in the **stipend** column.
- Necessary to clean the outliers before modeling or visualization.

Delete the unwanted Columns:

apply_links, and potsed_time

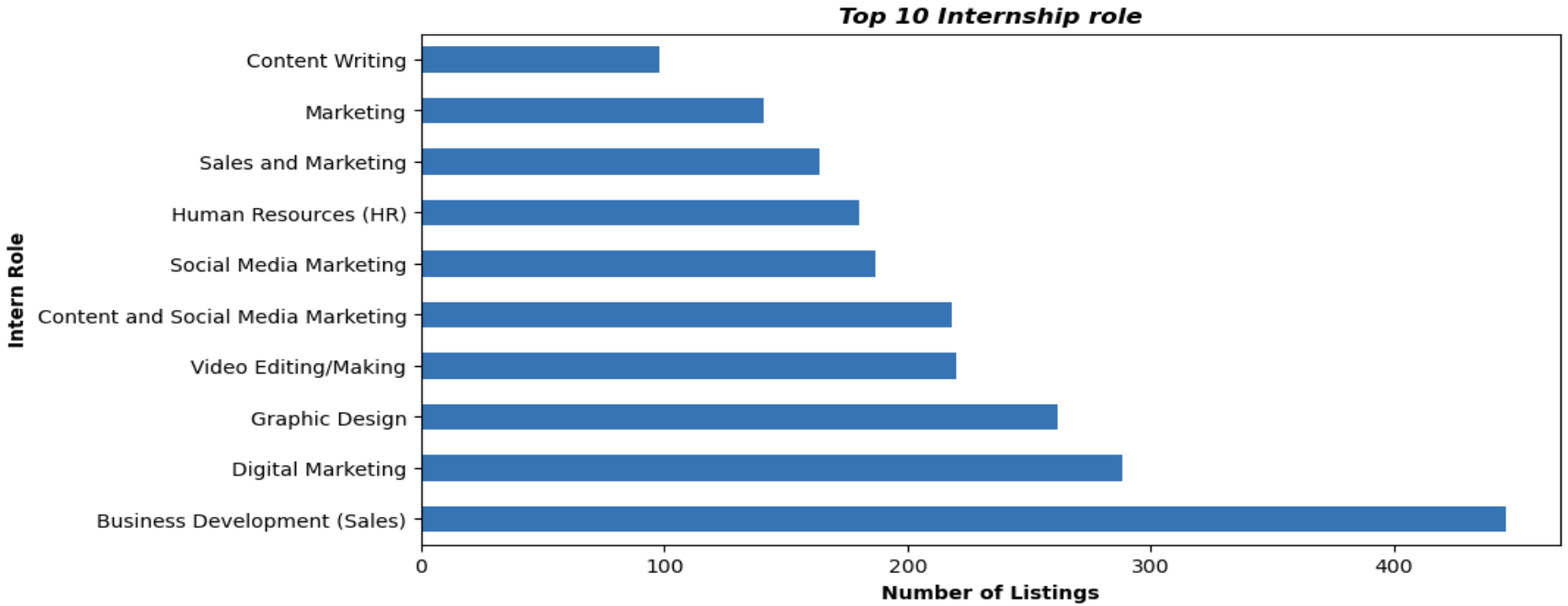
Verified types using: df.info()

#	Column Name	Non-Null Count	Data Type
1	intern_role	6024	object
2	company_name	6024	object
3	hiring_type	6024	object
4	location	6023	object
5	job_offer(LPA)	6024	float64
6	stipend_per_month	6024	int32
7	duration(months)	6024	float64
8	early_applicant	6024	object
9	posted_date	6024	datetime64[ns]

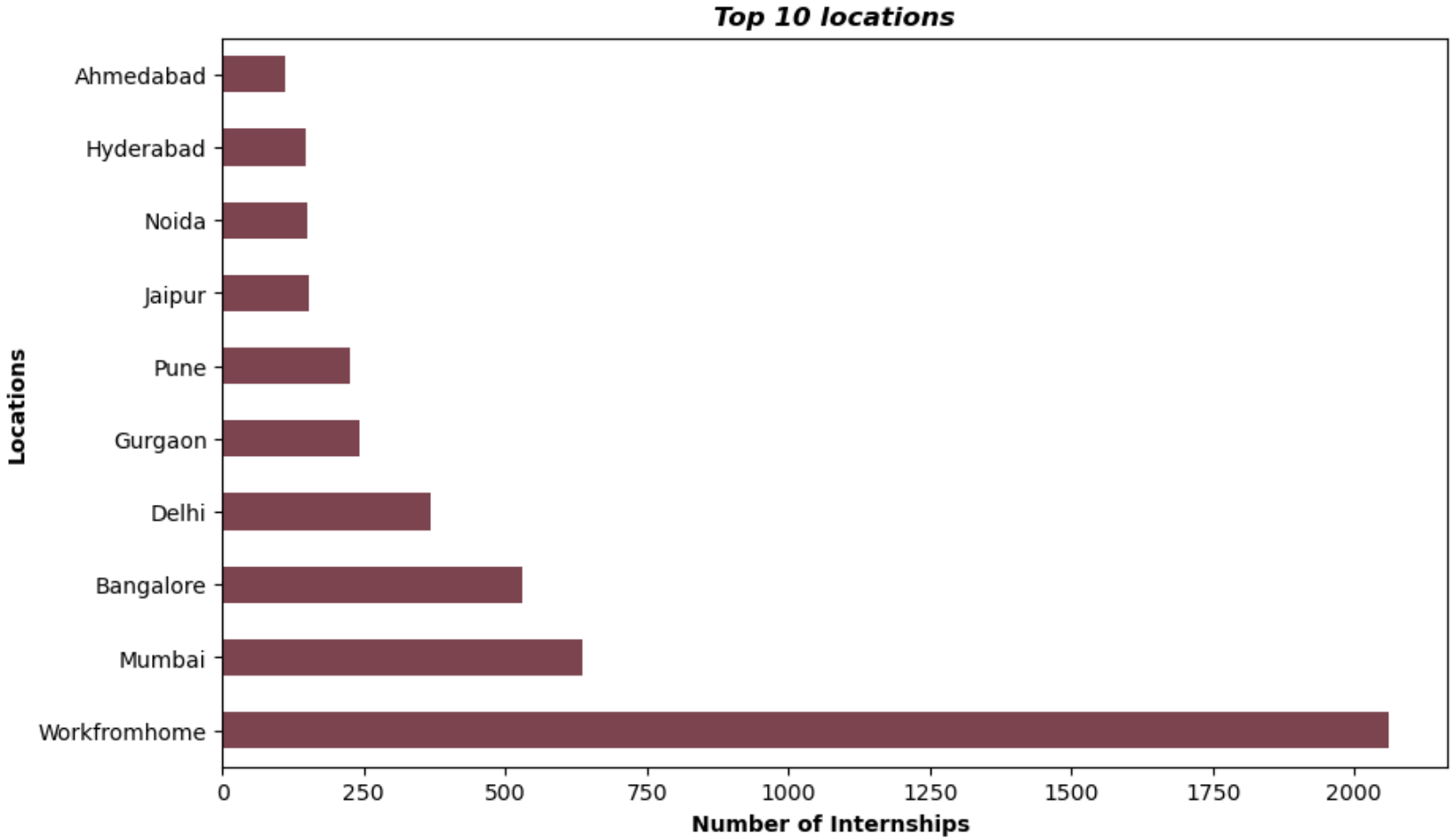
Ensured no NaNs remained before analysis

DATA VISUALIZATION

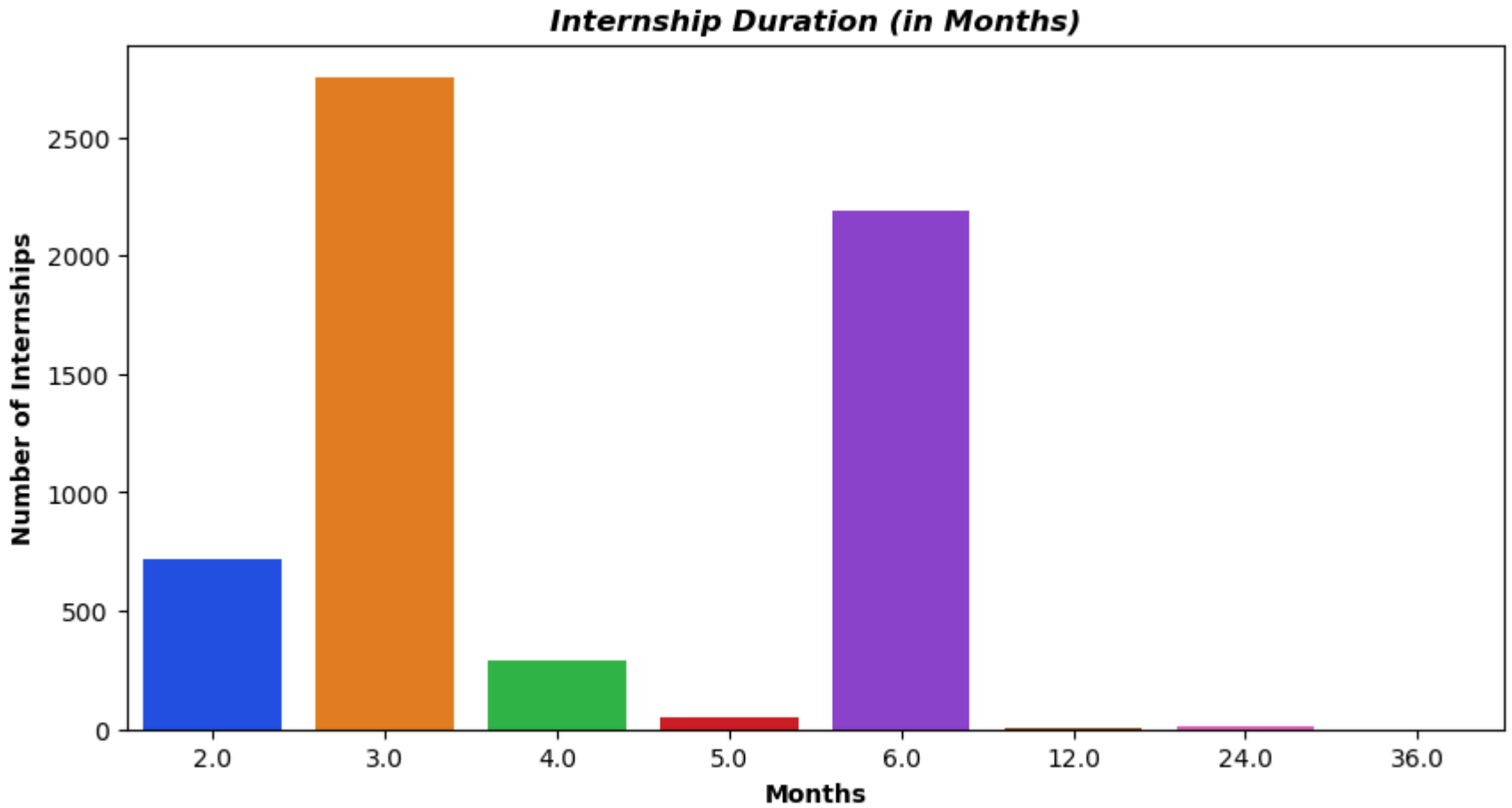
UNIVARIATE ANALYSIS:



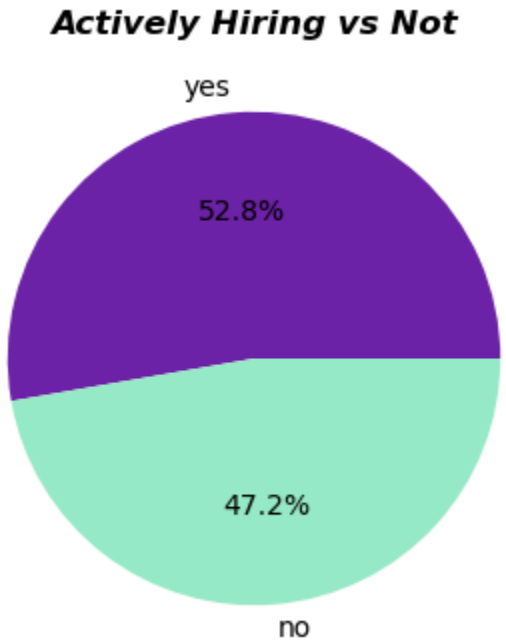
- In this graph ,the most wanted internship roles are listed.
- Business Development and Digital Marketing roles are the most in-demand internships , highlighting strong industry focus on sales and digital skills.
- Business Development is the most available internships



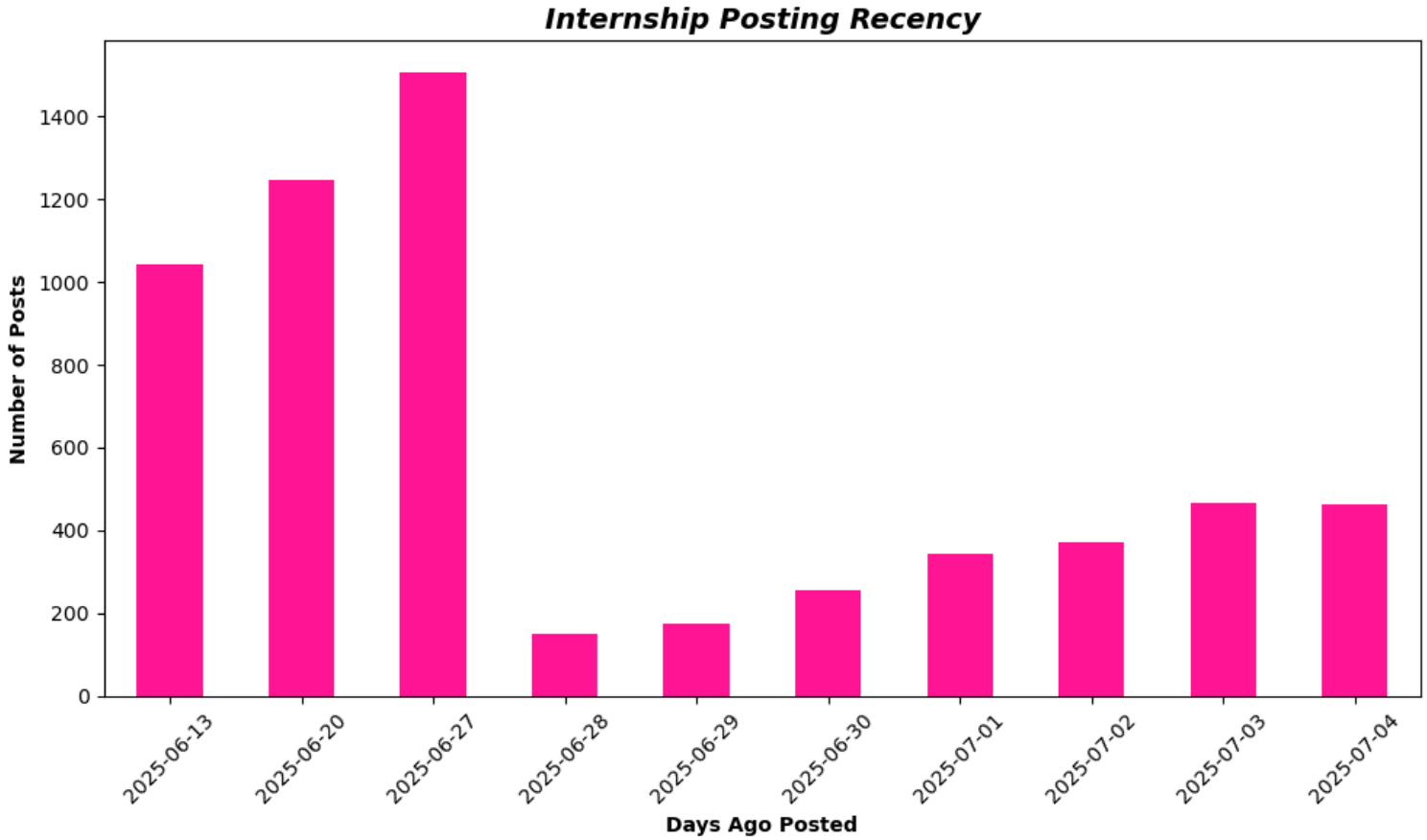
- Work-from-home internships dominate the listings, followed by major cities like Mumbai, Bangalore, and Delhi, reflecting a strong shift toward remote opportunities.



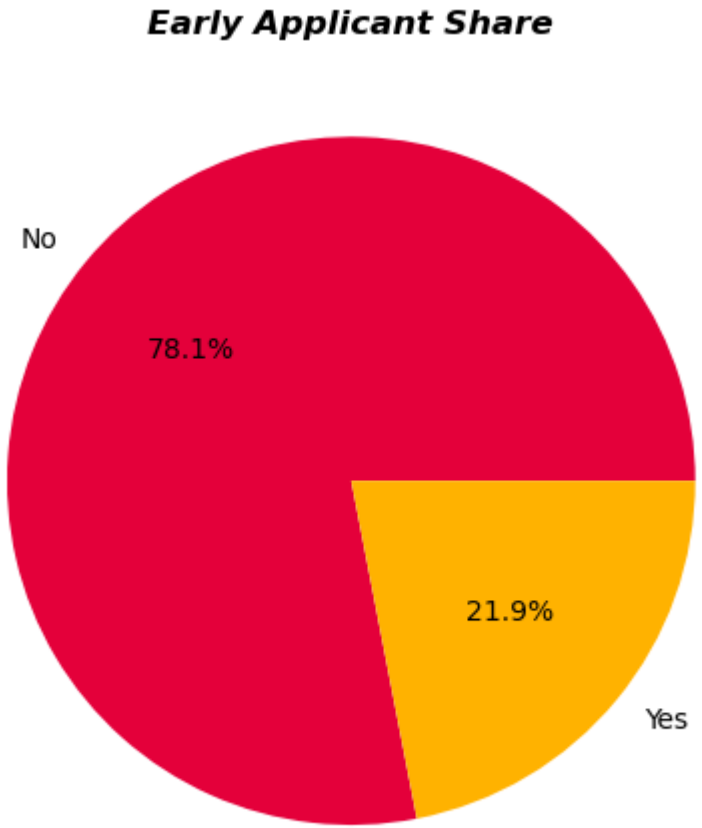
- Most internships have a duration of 3 to 6 months, indicating that companies prefer medium-term internship commitments.



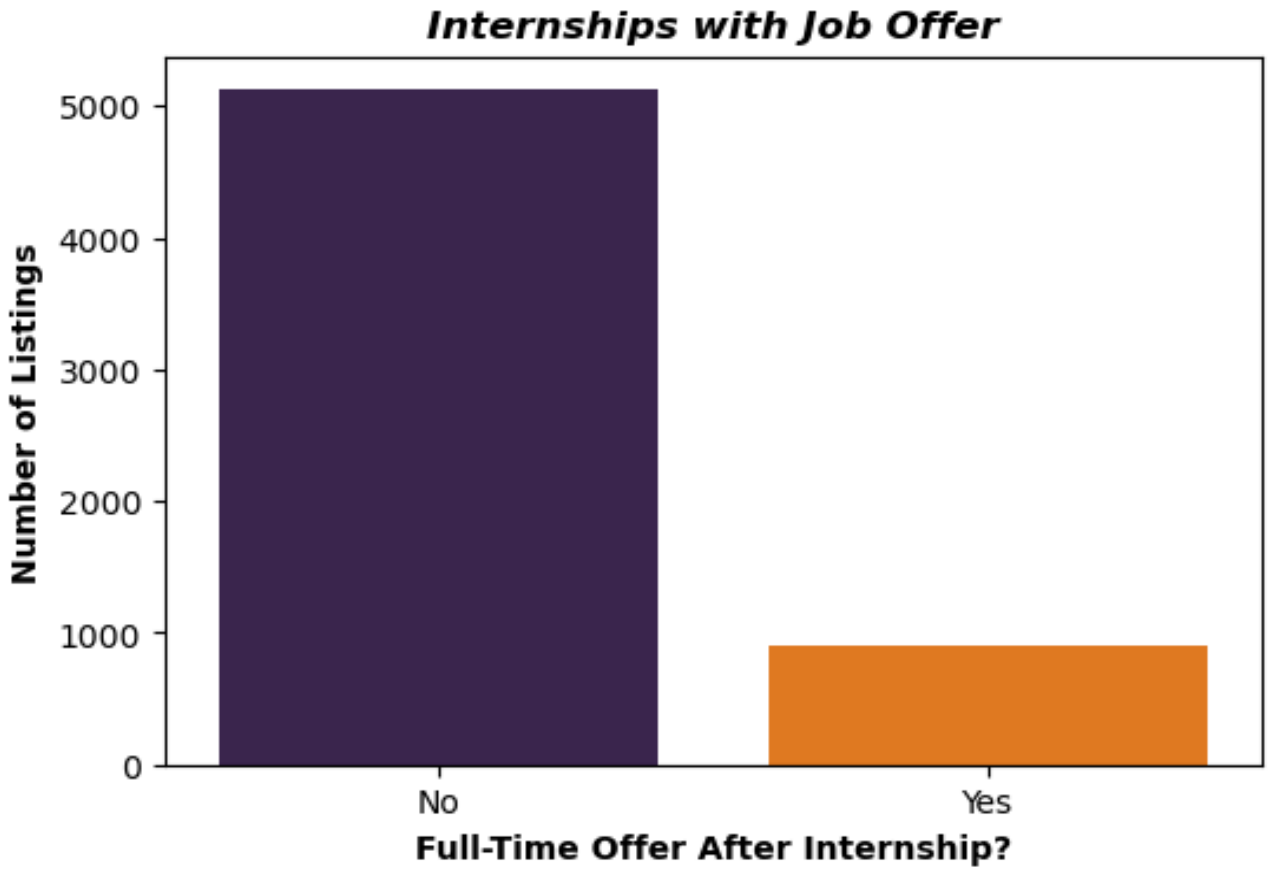
Over half (52.8%) of the internships are from actively hiring companies, indicating strong ongoing recruitment demand.



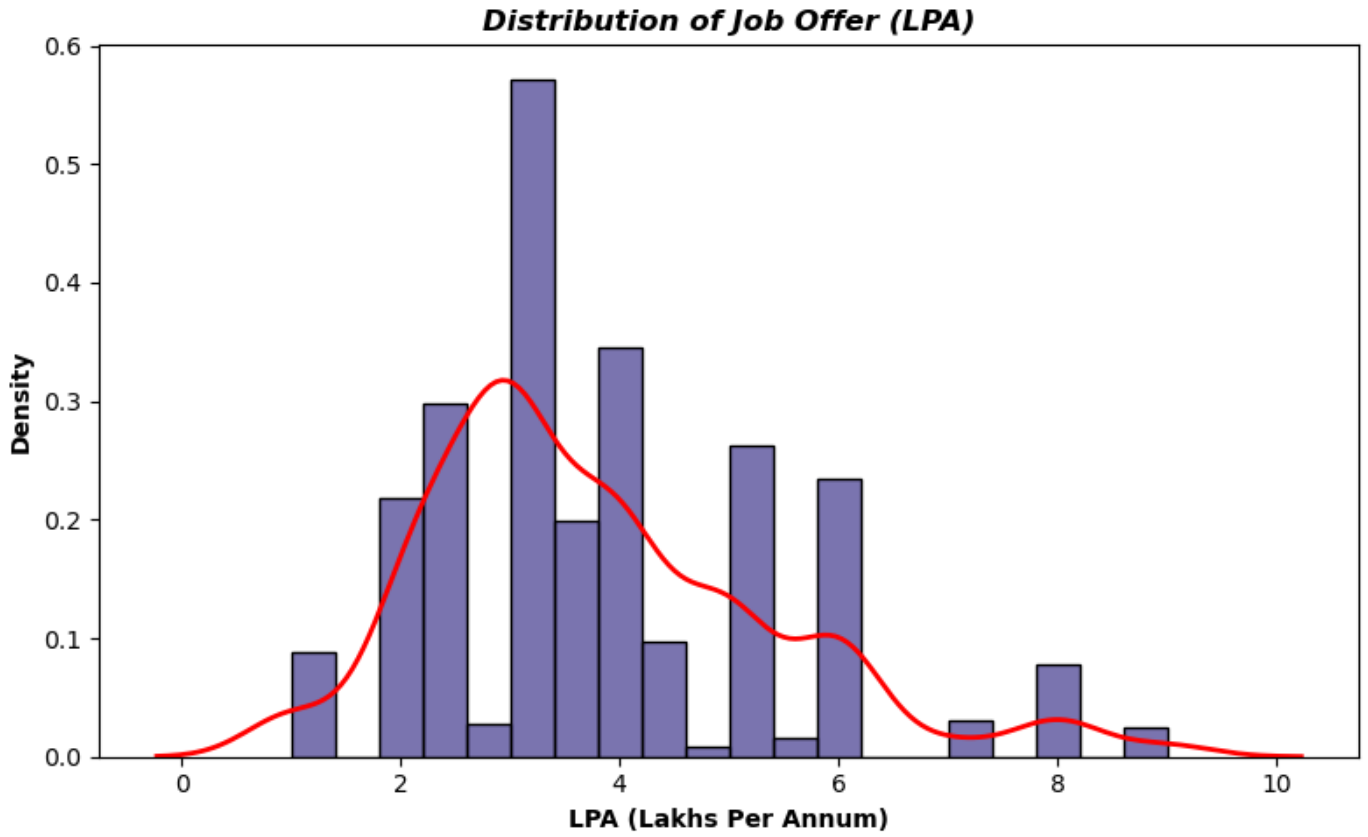
- Most internships were posted between **June 13 and June 27**, especially on **June 27** which saw the highest number of posts.
- After that, postings dropped and slowly increased again by **July 3–4**.



Around 21.9% of companies posted only posted with the early applicant.

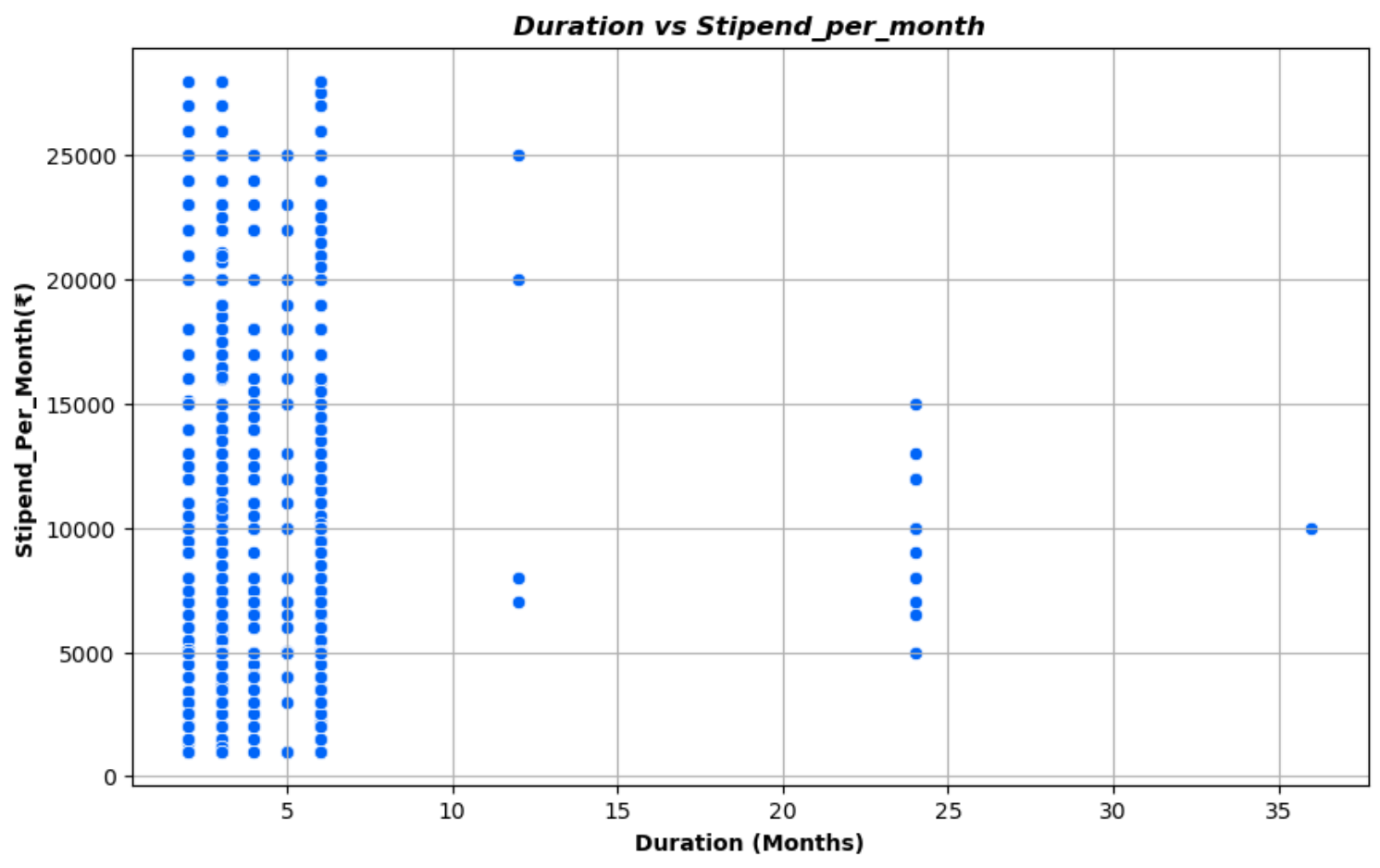


From 6024 listings, there are only have less than 1000 full-time job offers.



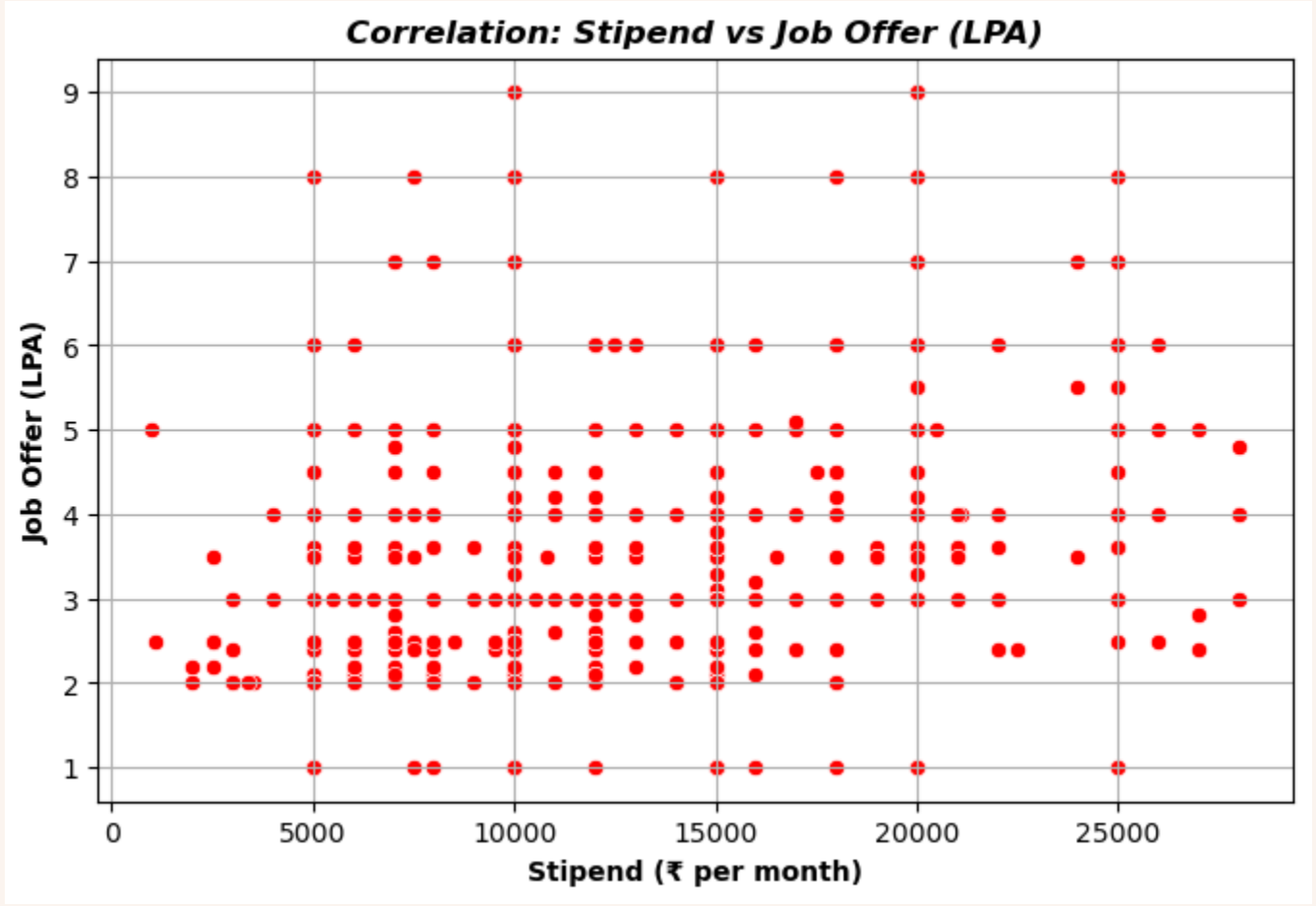
- Most job offers fall between 2 to 4 LPA.
- The curve is right-skewed, meaning fewer high-paying offers (above 6 LPA) exist.
- Very few internships offer job offers greater than 7 LPA.

BI-VARIATE ANALYSIS/MULTIVARIATE



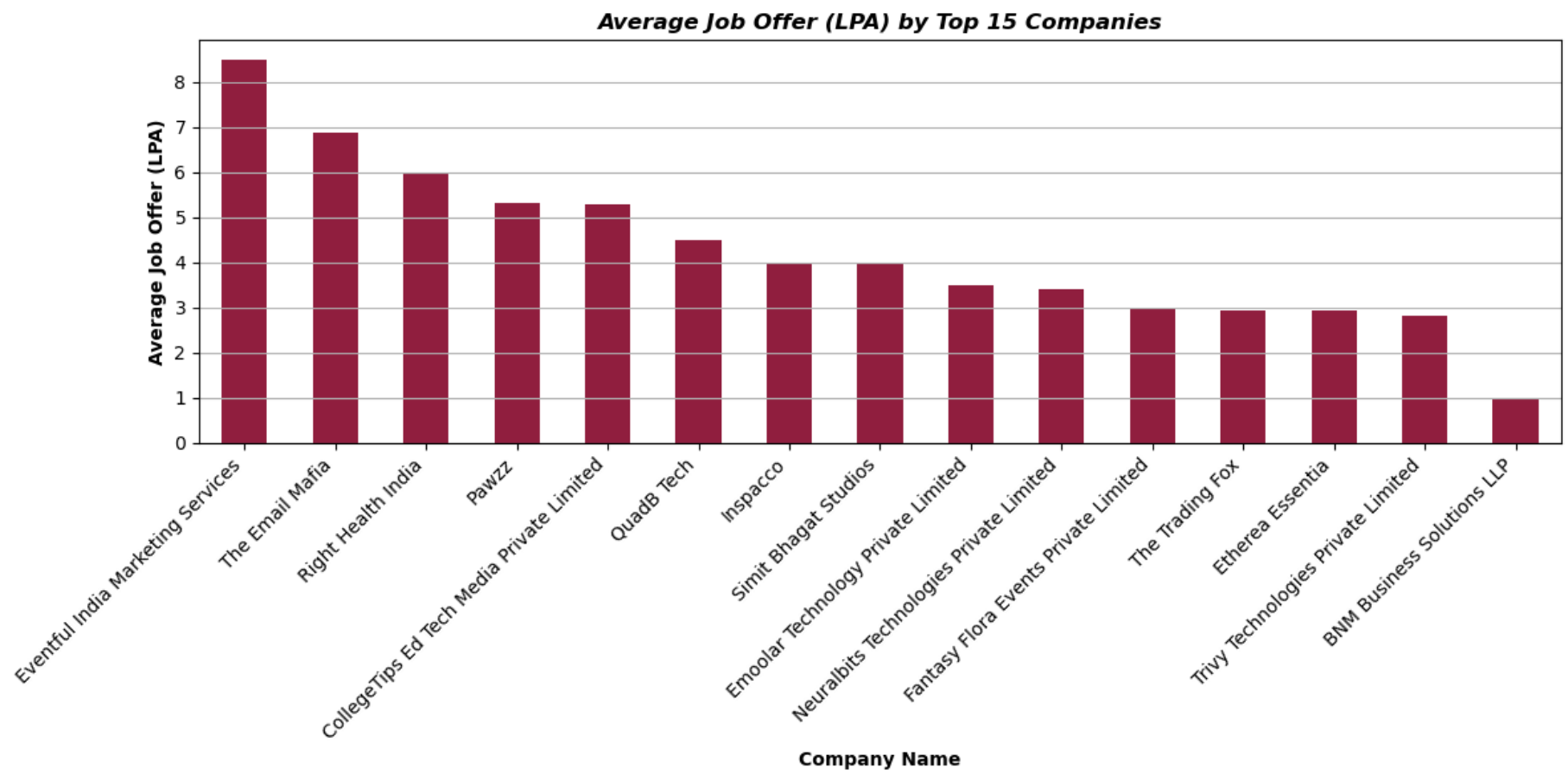
Continues VS Continues

Most internships last 2–6 months and offer variable stipends, showing no strong link between longer duration and higher pay.

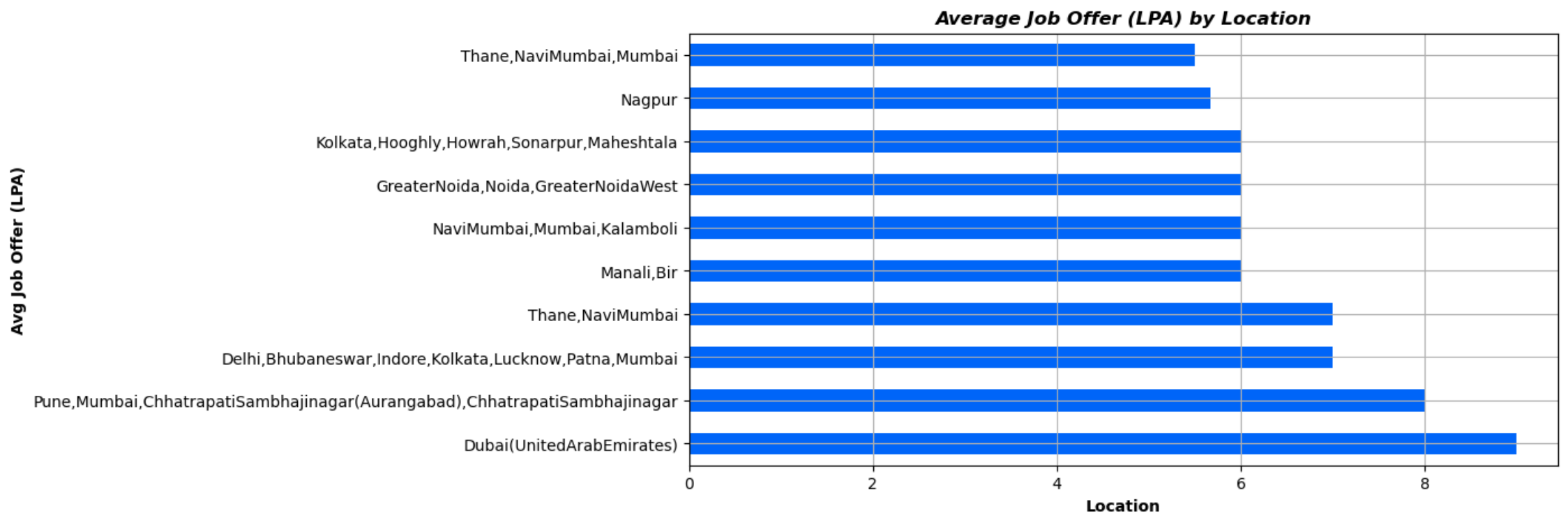


There is **no strong correlation** between stipend and job offer — **higher stipend doesn't guarantee** a higher job offer (LPA). Most offers are around 2–4 LPA regardless of stipend.

NUMERICAL VS CATEGORICAL

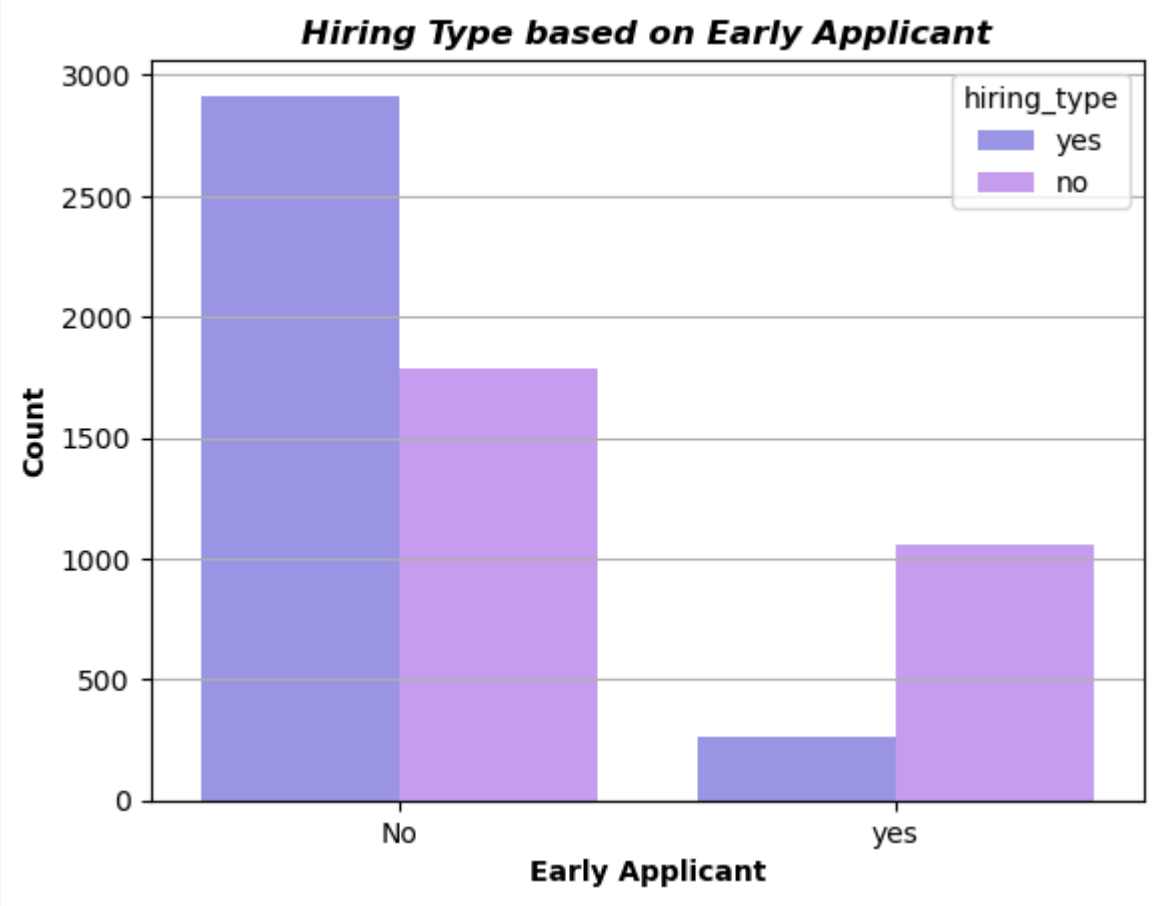


Among the top 15 companies offering internships, those with higher full-time job offers (LPA) indicate that students should also consider the employer.

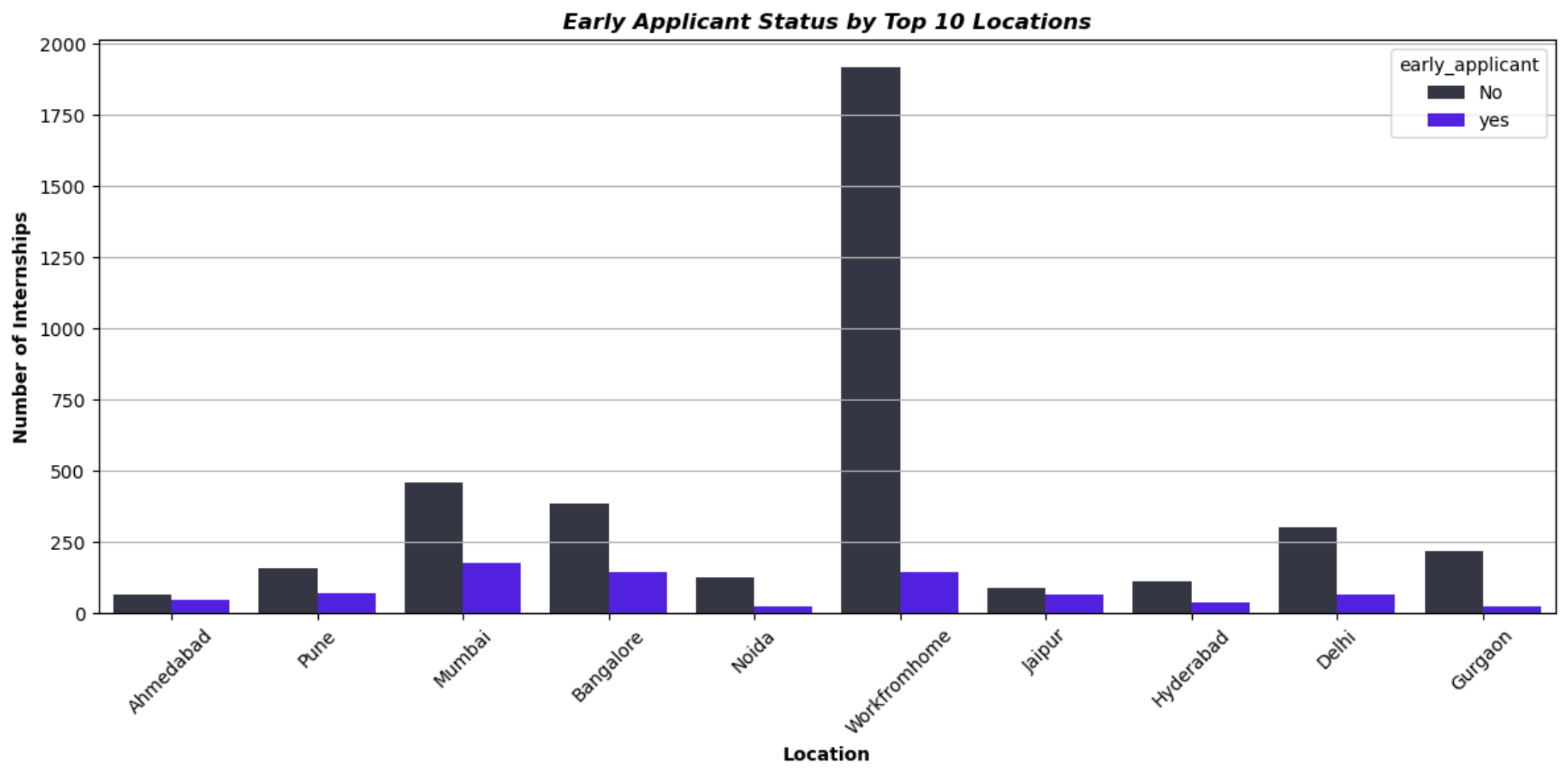


Dubai gives the highest job offer (above 9 LPA), followed by Pune, Mumbai, and Aurangabad with around 8 LPA.Nagpur, Thane, and Kolkata areas have lower job offers (around 5–6 LPA).

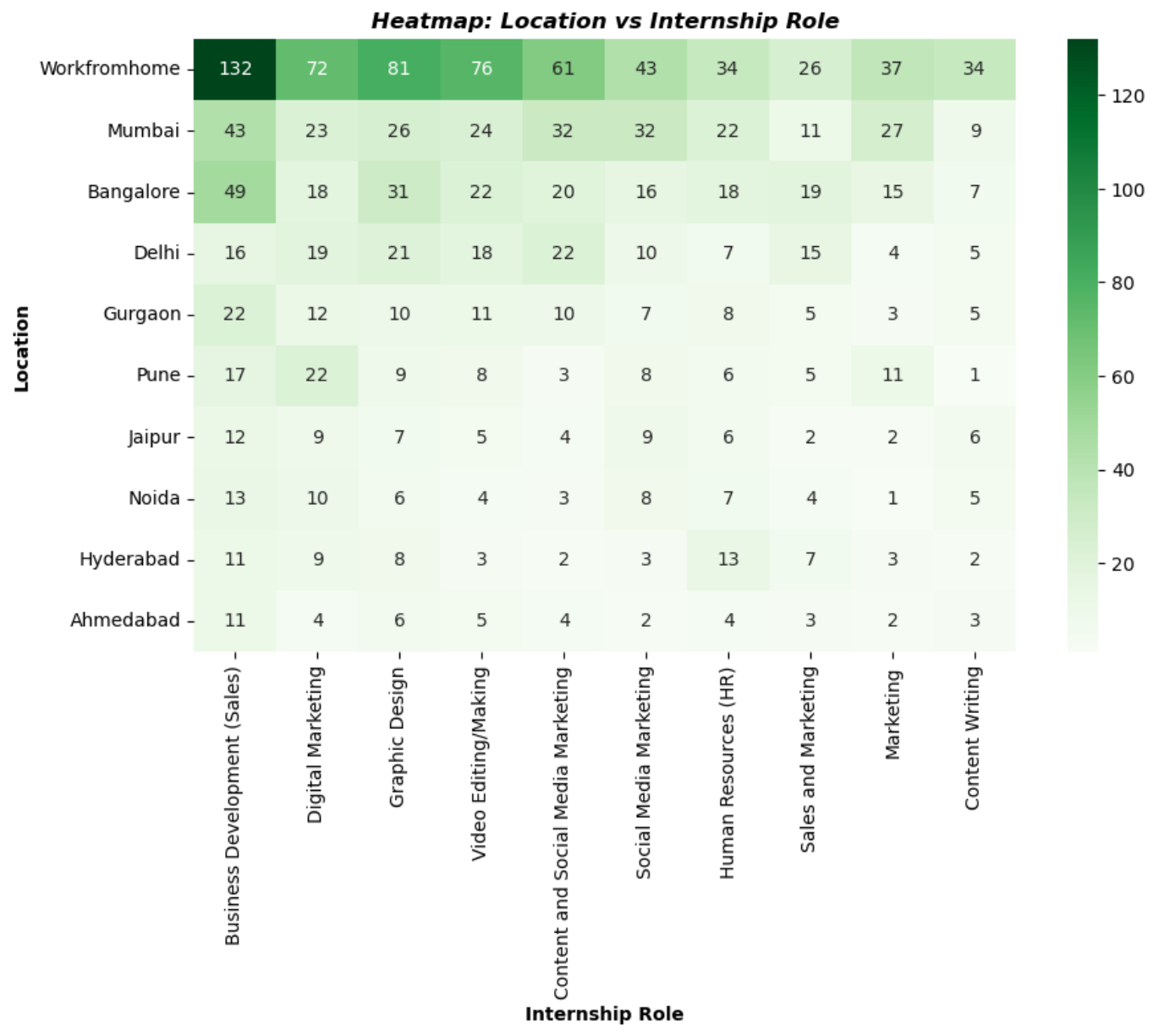
CATEGORICAL VS CATEGORICAL



- Most internships were applied without early applicant status
- In both early and non-early cases, many companies were not actively hiring.
- So, being early doesn't guarantee hiring — but it might still help visibility.



Work from home has the highest number of internships, but most are not early applicants.Mumbai and Bangalore also have many internships with some early applicants.



Work-from-home internships dominate all roles, offering the widest variety. Big cities like Mumbai, Bangalore, and Delhi follow next with decent opportunities across multiple roles, while smaller cities have fewer and more specific roles.

SUMMARY OF OBSERVATIONS

- Work-from-home roles dominate across India
- Most internships are paid (₹5000–₹10000)
- High job offers (LPA) are rare and not tightly linked to stipend
- Business Development & Digital Marketing are most common roles
- Early applicants don't always guarantee job conversion

THIS EDA HELPED UNCOVER:

- Internship trends across India
- Pay structure and job offer chances
- Student behavior around early applications

This can guide:

- Students in internship selection
- Companies in targeting talent

THANK YOU

Any Q/A

Presented by:
Arul R & Saranya S