PROJECT PRESENTATION

SAL_BW_Project
Analyzing Data Roles & Trends

Start Date : 15-04-2025

End Date : 20-04-25

TEAM MEMBERS

Er. Manish Debnath

Er. Bhupendra Shivhare

Er. Ashwin Kumar

INTRODUCTION

The SAL_BW_Project – Analyzing Data Roles & Trends focuses on examining job-related data to identify patterns and insights within the job market. The dataset includes over 1.6 million job listings with details such as job titles, companies, locations, average salaries, and posting dates. The project began by cleaning and preparing the raw dataset to ensure consistency and accuracy. The cleaned dataset was then used for further analysis, including the representation of job trends over time, salary distributions, and geographic hiring patterns. This structured approach enables a deeper understanding of how job roles and compensation vary across different companies and regions.

BACKGROUND

With the rapid expansion of the digital economy, the demand for skilled professionals across various job roles has seen significant growth. Companies across industries are increasingly relying on data to drive decision-making, optimize operations, and improve customer experiences. This shift has led to a surge in hiring for diverse roles ranging from marketing and web development to operations and data analysis. To better understand the dynamics of the job market, this project utilizes a large dataset of job listings, capturing key attributes such as job titles, company names, job locations, average salaries, and the dates when jobs were posted. Analyzing this data helps in identifying hiring trends, salary expectations, and regional employment patterns. The motivation behind this project is to extract meaningful insights from job description data, offering a clearer view of current employment trends and aiding individuals and organizations in making informed decisions based on real-world job data.

PROBLEMS

- The original dataset contained inconsistencies, missing values, and unstructured data that required cleaning.
- Lack of clarity in role categorization made it difficult to analyze trends across similar job titles.
- Variations in location names and salary formats posed challenges for comparison.
- Identifying meaningful insights from a large volume of unorganized job data was not possible without preprocessing and analysis.

THEORY

To build a project, the first step is to clearly define the objective. Next, data should be collected from appropriate sources. Once the data is collected, data cleaning and preprocessing should be performed to ensure the data is ready for analysis. After cleaning, exploratory data analysis (EDA) is conducted to uncover patterns and insights. Depending on the project, machine learning or statistical models may be applied. The results should then be visualized to extract key insights. Finally, the findings are communicated through a report or a dashboard. Common tools and libraries used include Pandas, Numpy, Matplotlib or Seaborn, and Scikit-learn.

THEORY

The main objective is to understand how to load, clean, and represent data using Python libraries like Pandas. Data can be loaded using functions like read_csv() or read_excel(), with various parameters like header, column names, and index columns to structure it properly. Cleaning the data includes handling missing values using methods like dropping or filling them. Data types can be converted using the astype() function. Duplicate records can be removed using drop_duplicates(). For further transformation, functions like apply(), map(), and replace() are helpful. Once cleaned, data needs to be represented in a structured form for analysis. This usually means working with DataFrames. Indexing and slicing help in accessing specific data points. Aggregation methods like groupby() and pivot_table() allow summarizing and analyzing patterns in the data.

THEORY

The goal of EDA is to summarize the main characteristics of a dataset and identify any patterns, trends, or anomalies. This often starts with basic descriptive statistics such as mean, median, and standard deviation using functions like describe(). For visualization, libraries like Matplotlib and Seaborn are used. Histograms help understand the distribution of data. Boxplots are useful to detect outliers. Heatmaps help identify correlations between variables. Pairplots are used to visualize relationships between pairs of features. The overall aim of EDA and visualization is to better understand the data, highlight important patterns, identify missing values or outliers, and support decision-making or model building.

THEORY

- The following commands are used to install required Python libraries:
- pip install xlrd: Installs the xlrd package to read .xls Excel files.
- pip install mysql-connector-python: Installs the MySQL connector library for Python to connect and interact with MySQL databases.
- pip install SQLAlchemy: Installs SQLAlchemy, a library for SQL toolkit and ORM (not used in the code directly).
- pip install pandas: Installs the Pandas library for data manipulation and analysis.

SAMPLE CODE

```
import mysql.connector
conn = mysql.connector.connect(
host="localhost",
user="root",
password="your_password",
 database="buildweek"
)
cursor = conn.cursor()
```

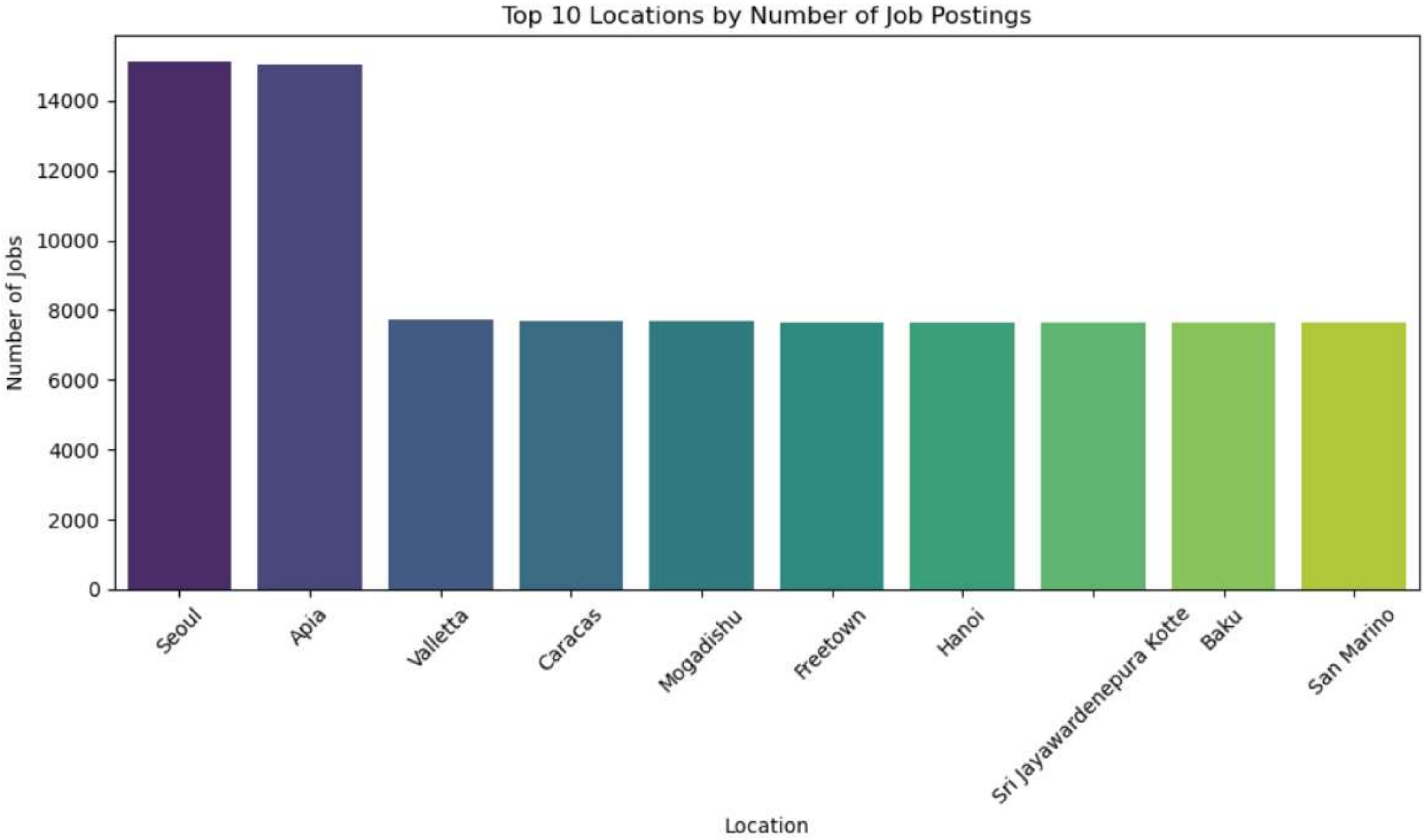
SAMPLE CODE

```
for _, row in df.iterrows():
    cursor.execute("""INSERT INTO clean_job_descriptions(`Title`, `Company`, `Location`, `Average Salary`, `Date Posted`)
    VALUES(%s, %s, %s, %s, %s)""",
                (row["Title"], row["Company"], row["Location"], row["Average Salary"], row["Date Posted"]))
conn.commit()
print("Data inserted successfully")
```

THEORY

- Database Creation and Setup:
- A new database named buildweek is created using the CREATE DATABASE statement.
- The USE buildweek command sets this database as the active one for subsequent operations.
- Table Creation:
- A table named clean_job_descriptions is created with the following columns:
- Title: Job title (string, up to 400 characters)
- Company: Name of the hiring company (string, up to 400 characters)
- Location: Job location (string, up to 400 characters)
- Average Salary: Monthly average salary (integer)
- Date Posted: The date when the job was posted (date format)
- Initial Data Display:
- The command SELECT * FROM clean_job_descriptions; is used to display all the records from the table.

BAR CHART



Top 10 Locations by Number of Job Postings
Number of Jobs
0
2000
4000
6000
8000
10000
12000
14000
Seoul
Apia
Valletta
Caracas
Mogadishu
Freetown
Hanoi
Sri Jayawardenepura Kotte
Baku
San Marino
Location

INSIGHTS

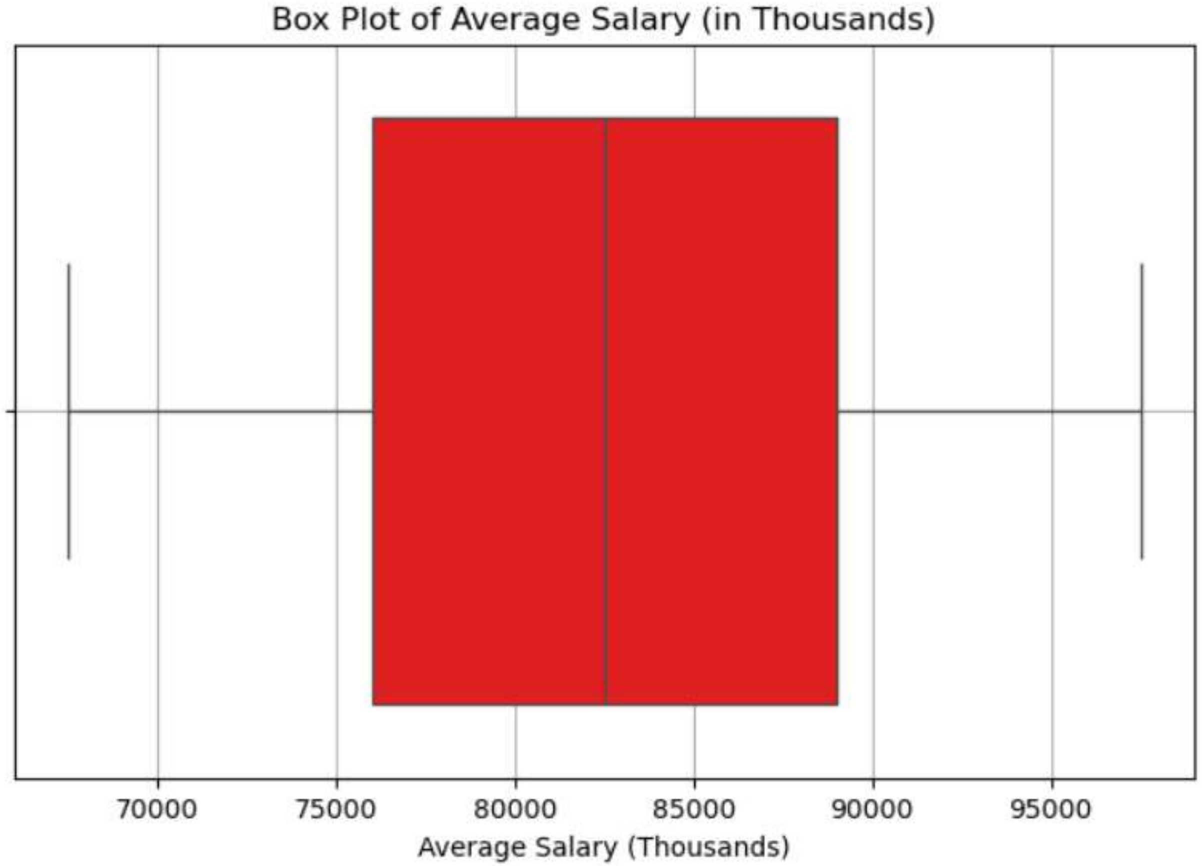
•**Job Hotspots:**
The chart helps identify **which locations have the highest demand** for jobs.
• If you're looking for a job or analyzing hiring trends, these are the places to watch.

•**Geographical Trends:**
You can assess whether jobs are **concentrated in metropolitan areas**, tech hubs, or spread out more evenly.

•**Comparison Across Locations:**
You can quickly see **how one location compares to another**—some may have significantly higher postings, indicating a more active job market.

•**Decision Making:**
If you're a job seeker, this chart tells you **where to focus your applications**. If you're a recruiter or analyst, it tells you **where job demand is highest**.

BOXPLOT CHART



INSIGHTS

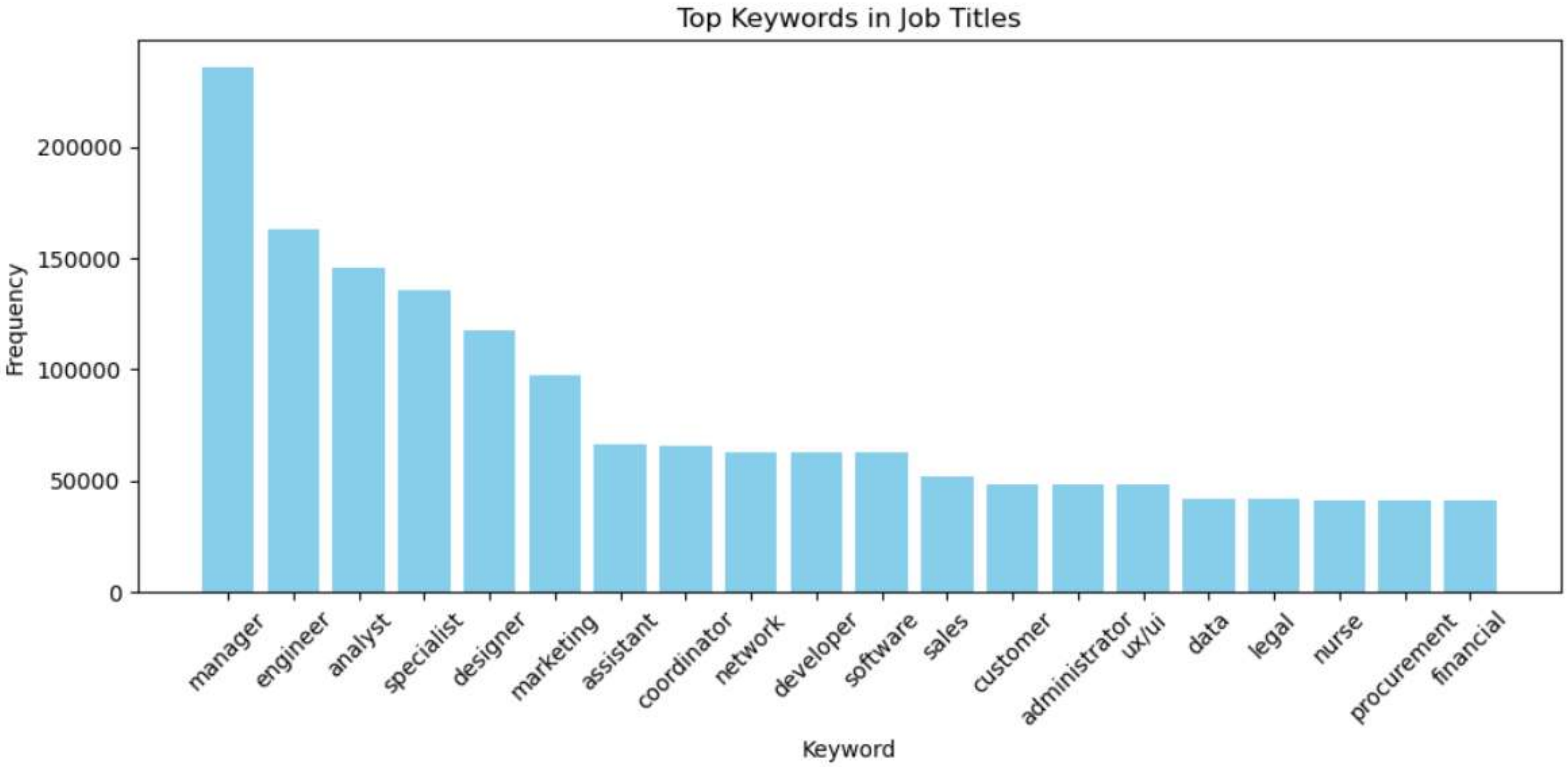
- **Salary Range & Spread:**
- Helps you see the overall **spread of salaries**, whether they’re tightly packed or widely varied.
- A wider box means more variation in the middle 50% of salaries.

- **Median Salary:**
- The **middle point** shows what most people are earning — a good benchmark for comparison.

- **Outliers Detection:**
- High outliers might indicate **very senior or niche roles**.
- Low outliers could signal **internships, entry-level positions**, or possibly data issues.

- **Skewness:**
- If the median is closer to one side, it shows **skewed distribution** (e.g., right skew = a few very high-paying jobs pulling up the average).

BAR CHART


Top Keywords in Job Titles
Frequency
0
50000
100000
150000
200000
manager
engineer
analyst
specialist
designer
marketing
assistant
coordinator
network
developer
software
sales
customer
administrator
ux/ui
data
legal
nurse
procurement
financial
Keyword

INSIGHTS

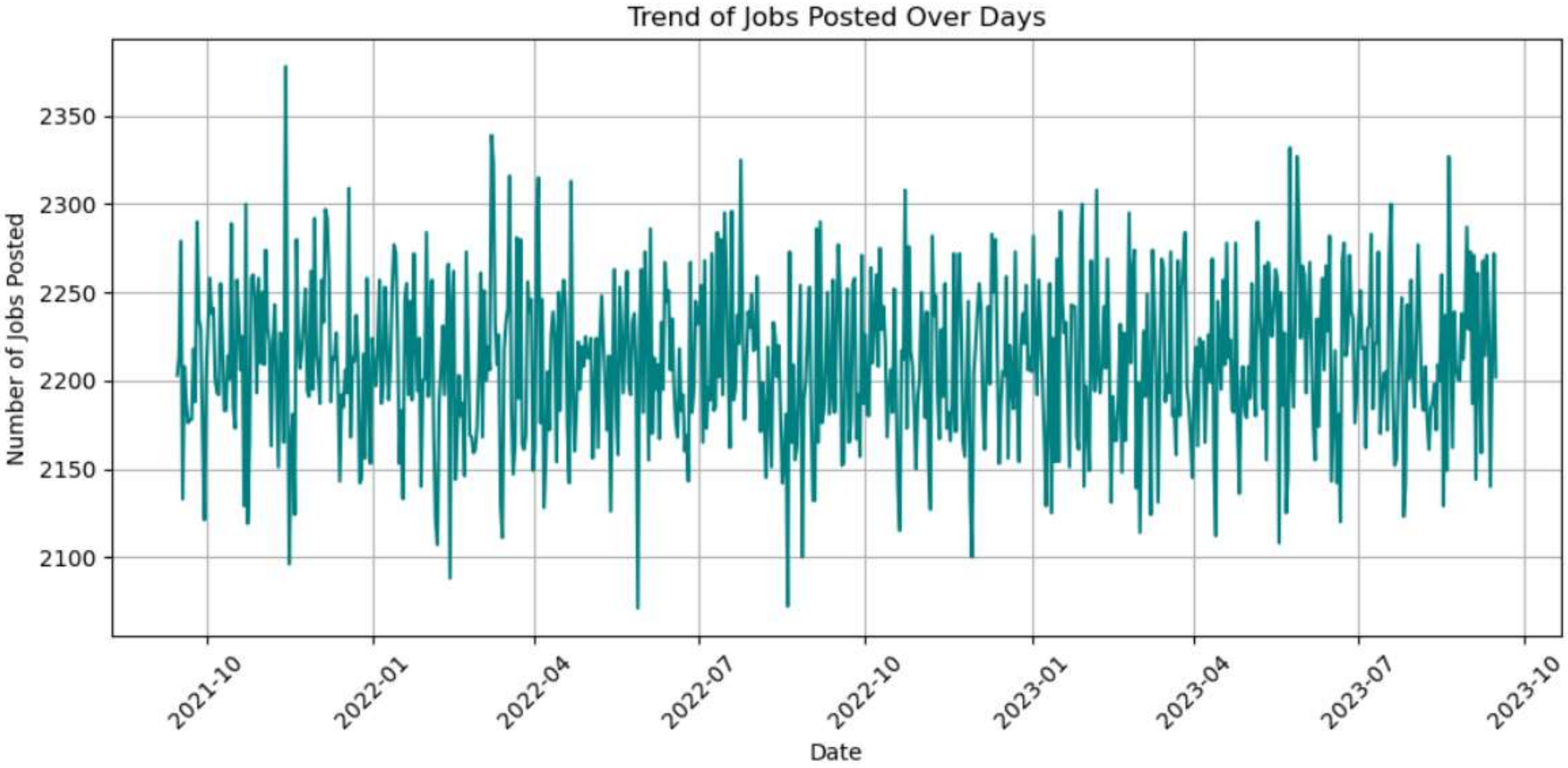
- **Most Common Job Roles or Fields:**
- For example, if "developer", "engineer", "analyst" are among the top words, it shows demand in **tech and analytics roles**.
- Keywords like "sales", "marketing", or "consultant" would indicate other popular domains.

- **Job Title Trends:**
- You can detect naming patterns used by companies. For instance, lots of "senior", "junior", or "lead" roles? That hints at how roles are tiered or labeled.

- **Buzzwords & Skill Focus:**
- Words like “data”, “cloud”, “AI” might reflect industry trends and in-demand skills.

- **Optimization for Job Seekers:**
- If you're writing or optimizing a resume or LinkedIn title, knowing these high-frequency words helps you align with common job titles.

LINE CHART


Trend of Jobs Posted Over Days
Number of Jobs Posted
2350
2300
2250
2200
2150
2100
2021-10
2022-01
2022-04
2022-07
2022-10
2023-01
2023-04
2023-07
2023-10
Date

INSIGHTS

- **Posting Patterns:**
- You can see whether job postings are **increasing, decreasing, or stable** over time.
- Spikes might indicate hiring drives, events, or batch uploads by companies.

- **Seasonal or Weekly Trends:**
- If the dataset spans weeks/months, you might notice patterns:
    - **More jobs on weekdays** vs fewer on weekends.
    - **Monthly cycles** in hiring.

- **Recent Activity:**
- Identify **active periods** when companies are posting frequently.
- See if posting has dropped recently—this could indicate off-seasons or market shifts.

- **Data Quality Checks:**
- Sudden drops to zero might suggest **missing or inconsistent data** for certain dates.

THANK YOU