# 📊 Data Analyst Job Market Analysis (Naukri.com)

## Project Summary:

This project performs a comprehensive market analysis of Data Analyst job listings scraped from Naukri.com to identify key industry trends, most demanded technical skills, experience-level distribution, and compensation insights. The entire data pipeline, from raw data acquisition to interactive visualization, was engineered using a robust set of tools, including **Selenium** for scraping, **Python (Pandas)** for cleaning, **MySQL** for structured querying, and **Power BI** for dynamic reporting.

This analysis provides data-driven recommendations for both job seekers aiming to enter the Data Analyst field and companies looking to hire top talent.

## Project Goals:

The primary objectives of this analysis were:

1. **Understand Demand:** Determine the distribution of Data Analyst job openings across different experience levels (e.g., Fresher, Mid-Level, Senior).
2. **Identify Core Competencies:** Extract and quantify the most frequently required technical skills (e.g., SQL, Python, Power BI, Excel).
3. **Analyze Industry Trends:** Identify which industries are the primary employers of Data Analysts.
4. **Visualize Insights:** Create an interactive dashboard (in Power BI) to communicate the findings clearly and efficiently.

# Technology Used:

| Category | Tool/Language | Purpose in Project |
|---|---|---|
| Data Acquisition | Python, Selenium | Automated web scraping of job details from Naukri.com |
| Data Cleaning | Python (Pandas) | Handling missing values, normalizing text, and structuring data for analysis. |
| Data Storage | MySQL | Used as the centralized, persistent database for structured data. |
| ETL/Loading | SQLAlchemy | Python toolkit used to create the database engine and load cleaned Pandas DataFrames into MySQL. |
| Data Analysis | SQL (Group By, CTEs) | Executing complex, aggregated queries to derive core job market statistics. |
| Visualization | Power BI | Connected to the MySQL dataset to build interactive reports and dashboards. |

# Methodology (The Data Pipeline):

The project followed a standard, Five-phase data analysis lifecycle:

### Phase 1: Data Acquisition (Web Scraping)

- Source: Naukri.com job portal.
- Tool: Python and the Selenium library.
- Action: Developed a script to navigate the job listing pages, extract key information (Job Title, Company, Location, Experience, Key Skills, Salary disclosure, Industry Type), and store the raw data into a CSV file.

### Phase 2: Data Cleaning and Preprocessing (Jupyter Notebook)

- Missing Value Strategy:
    - Identified columns containing null values and applied a combination of row-wise and value-wise treatments.
    - Rows with **more than 50% missing data were dropped** to maintain dataset integrity.
    - For the remaining missing values, imputation was done using **mean, median, or mode**, selected based on the nature of the variable and its distribution.
- Duplicate Removal:
  Identified and removed duplicate job entries based on key identifiers.

- Text Normalization:
  Performed text preprocessing to convert unstructured job information into structured, analysable fields:
  - Parsed the “Experience” column to extract minimum and maximum years from text-based ranges (e.g., *“2–5 years”* → *2 and 5*).
  - Tokenized the “Skills” field to identify individual skills and split combined text entries into separate standardized values.
  - Converted salary information into a binary indicator where *1 represents salary disclosed* and *0 represents salary not mentioned*, enabling further filtering and KPI analysis.

**Phase 3: Data Loading (SQLAlchemy & MySQL)**

- Database Setup:
  A clean schema and table were defined in MySQL.
- Data Transfer:
  Used SQLAlchemy to establish a connection (engine) between the Python environment and the MySQL server. The cleaned Pandas DataFrame was loaded into the MySQL table.

**Phase 4: Core Data Analysis (MySQL & SQL Queries)**

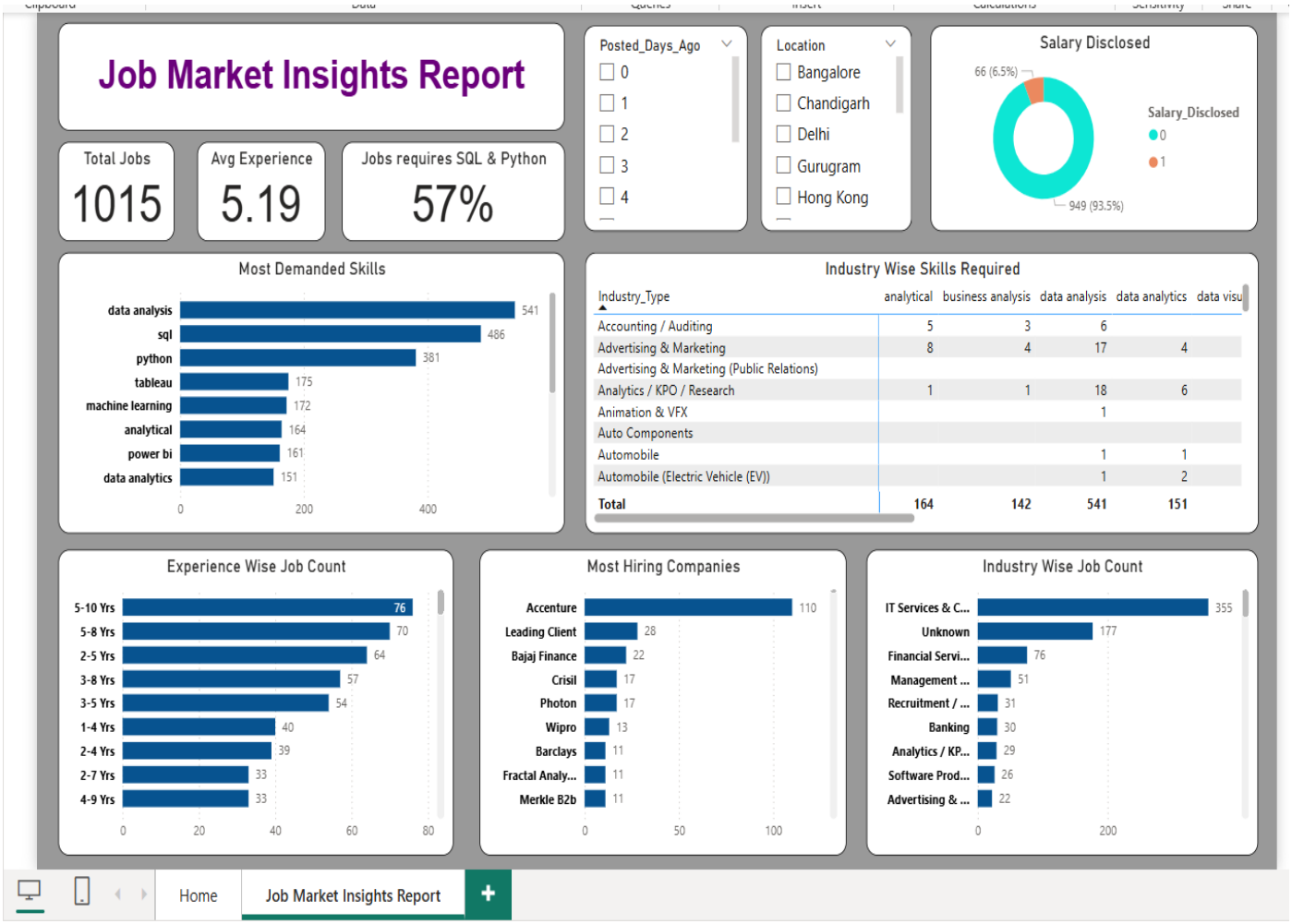
SQL queries were executed against the structured MySQL dataset to derive the primary insights:

- Experience vs. Count:
  Used GROUP BY to count jobs based on required experience levels.
- Industry Breakdown:
  Grouped and counted jobs by Industry/Sector.
- Skill Demand: Employed Common Table Expressions (CTEs) and advanced text parsing logic to accurately count the demand for specific skills (e.g., "Python", "R", "Tableau")
- Top Hiring Companies:
  Aggregated job counts by company name

**Phase 5: Visualization and Reporting (Power BI)**

- Connection:
  Power BI was connected directly to the MySQL database, ensuring the visualization reflects the data structure and analysis performed in SQL
- Key Visualizations:
  Dashboards were created to dynamically display:
  - Job Count by Experience Level.
  - Bar Charts showing the distribution of the Top 10 Most Demanded Skills.
  - A donut chart was used to represent the overall distribution of salary information in job postings, clearly showing the percentage of disclosed vs. non-disclosed salaries.

# Key Insights Generated




## 1.Market Overview and Job Demand

- Total Jobs & Experience:
  The report covers a total of **1015** jobs with an average required experience of **5.19** years.
- Essential Skills Mandate:
  A significant **57%** of all jobs explicitly require proficiency in SQL and Python, highlighting their foundational role in the data domain.

## 2. Most Demanded Skills

The job market heavily prioritizes core technical and programming skills:

- Top 5 Skills:
    - Data Analysis (541 postings)
    - SQL (486 postings)
    - Python (381 postings)
    - Tableau (175 postings)
    - Machine Learning (172 postings)

- Conclusion:
  Companies prioritize **core analytical and programming skills** (Data Analysis, SQL, Python) along with **data visualization tools** (Tableau, Power BI), indicating a need for strong technical and analytical abilities in candidates.

## 3. Experience Requirement

- Highest Demand:
  The highest hiring demand is concentrated in the **mid-level experience** brackets (**2-10 years**).
- Conclusion:
  Hiring demand is highest for mid-level experience. Freshers (0 Yrs) and very senior candidates (e.g., 7+ Yrs) have fewer opportunities, suggesting companies prefer **skilled, hands-on professionals** over entry-level or very senior profiles.

## 4. Industry-wise Job Opportunities

The IT Services & Consulting sector is the largest recruiter by a significant margin.

- Dominant Sectors:
  - **IT Services & Consulting:** 355 jobs
  - **Financial Services:** 177 jobs
  - **Management Consulting:** 76 jobs
- Conclusion:
  The **IT & Consulting** sectors dominate hiring, while industries such as Oil & Gas and Real Estate show significantly lower hiring volumes.

## 5. Compensation Transparency (Salary Disclosed)

- Lack of Disclosure:
  Out of 1015 jobs, only **66 (6.5%)** disclose the salary.
- Conclusion:
  The vast majority of companies (93.5%) do not reveal salary ranges upfront, making **transparency in compensation very limited** in job postings.

## 6. Synthesis of Industry-Wise Skills Required

- IT Services & Consulting:
  Dominates the demand for **analytical, business analysis, data analytics, and data visualization** skills, reflecting a broad need for data-centric expertise.
- Financial Services:
  Emphasizes **Analytical Skills, Excel, and SQL**, suggesting a focus on reporting, risk analysis, and core data processing.