Project Report

Data Cleaning and Insight Generation from Survey Data

---

1. Introduction

Survey data often contains missing values, duplicates, and inconsistent formatting, making it difficult to analyze directly. This project uses the **Kaggle Data Science Survey (2017–2021)** dataset to demonstrate data cleaning, categorical handling, and insight generation. The aim is to extract meaningful findings about respondents' demographics, tools, and preferences.

---

2. Dataset Description

The dataset is compiled from Kaggle's annual Data Science and Machine Learning surveys conducted between 2017 and 2021. It contains **multiple-choice responses** from thousands of data professionals across the world.

- **Columns:** 40+ survey questions labeled as `Q1`, `Q2`, … `Q42`.
- **Types of Data Collected:**
  - Demographics (age, gender, country)
  - Employment and job titles
  - Programming languages used
  - Machine learning frameworks and tools
  - Salary ranges
  - Career preferences and education

---

3. Methodology

3.1 Data Cleaning

- Removed duplicate rows to avoid bias in insights
- Handled missing values by:
  - Filling categorical NaNs with `"Not Specified"`
  - Dropping columns with excessive null values (e.g., > 50%)
- Standardized inconsistent entries (e.g., formatting of country names)

3.2 Handling Categorical Variables

- Applied **Label Encoding** and **Mapping** to categorical responses
- Converted survey answers into numerical codes for analysis

3.3 Insight Generation

- Generated summary statistics on demographics and tools
- Grouped responses by gender, job role, and country to compare trends
- Visualized correlations between job title and programming language usage

---

4. Key Insights

1. **Python** emerged as the most popular programming language across all survey years.
2. **Data Scientists and Analysts** form the largest group of respondents.
3. **Cloud platforms** such as AWS, Azure, and GCP have shown rapid adoption after 2019.
4. Respondents from the **United States and India** represent the majority of participants.
5. Visualization libraries like **Matplotlib** and **Seaborn** are consistently the most widely used.

---

5. Visualizations

- **Bar Charts**: To compare top programming languages and tools.
- **Heatmaps**: To show correlations between job roles, salaries, and tool usage.
- **Dashboard-style plots**: Summarizing top 5 insights from the survey.

---

6. Conclusion

This project demonstrates how **data cleaning** and **categorical handling** can transform raw survey data into meaningful insights. By applying these techniques, we identified trends in tools, job roles, and preferences among data science professionals worldwide.

The insights generated can help organizations and educators understand how the data science ecosystem evolves over time and where professionals focus their skills.