Kaggle Survey Data Analysis Report

Author: Rameez

Date: 1-Aug-2025

---

1. Project Overview

This project analyzes the Kaggle Survey (2017-2021) dataset to extract insights about data science professionals' demographics, skills, and preferences. The dataset contains real-world survey responses with missing values, duplicates, and inconsistent formatting, requiring thorough cleaning and preprocessing before analysis.

Objectives:

Clean the dataset (handle missing values, duplicates, and formatting issues)
Apply label encoding for categorical variables (age, gender, country, etc.)
Extract meaningful insights about respondent behavior and preferences
Create a summary dashboard highlighting the top 5 insights

---

2. Tools & Libraries Used

- Python (Primary tool for data cleaning and analysis)
- Pandas (Data manipulation)
- NumPy (Numerical operations)
- Matplotlib & Seaborn (Data visualization)

---

3. Step-by-Step Execution

Step 1: Data Loading & Initial Exploration

- Loaded the dataset using `pd.read_csv()`.
- Checked dimensions (`df.shape`) and column names.
- Identified key columns:
    - `Age`
    - `Gender`
    - `Country`
    - `Education Level`
    - `Job Role`

Step 2: Data Cleaning

✔ Handled Missing Values:

- Dropped rows with missing critical data (age, gender, country).
- Removed columns with >70% missing values.

✔ Removed Duplicates:

- Used `df.drop_duplicates()` to ensure unique responses.

✔ Standardized Text Formatting:

- Stripped whitespace and special characters from categorical columns.
- Renamed columns for consistency (e.g., `What is your age?` → `age`).

Step 3: Categorical Variable Encoding

✔ Age (Ordinal Encoding):

- Categories: `18-21`, `22-24`, ..., `70+`

- Mapped to numerical values (0-10).

✔ Gender (Label Encoding):

- `Man` → 0, `Woman` → 1, `Nonbinary` → 2, etc.

✔ Country (Grouping):

- Kept top 10 countries, grouped others as "Other".

Step 4: Insight Extraction

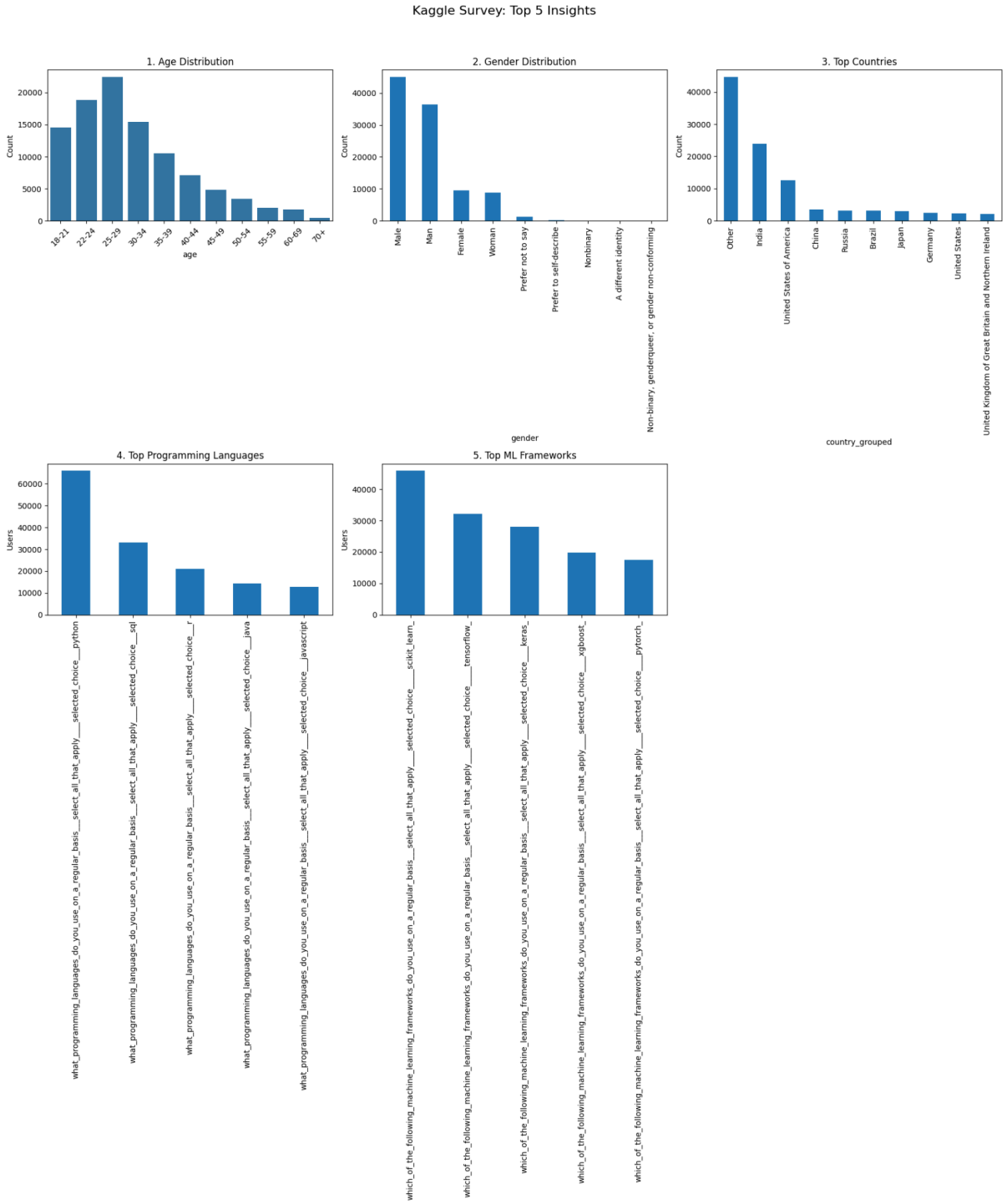
Extracted 5 key insights from the data:

1. Age Distribution
    - Most respondents are aged 25-34.
    - Fewer respondents in the 55+ age group.
2. Gender Distribution
    - Male-dominated field (~75% male, ~20% female).
    - Small representation of nonbinary and other genders.
3. Top Countries
    - India, USA, and Brazil have the highest respondents.
    - Reflects Kaggle's global user base.
4. Top Programming Languages
    - Python dominates (~85% usage).
    - Followed by SQL and R.
5. Top Machine Learning Frameworks
    - Scikit-learn is the most popular.
    - TensorFlow and PyTorch follow closely.

---

4. Summary Dashboard (Visualization)

Created a dashboard summarizing the top insights



(Generated using Matplotlib/Seaborn)

Dashboard Breakdown:

📊 Age Distribution (Bar Chart)

🚻 Gender Distribution (Bar Chart)

🌍 Top Countries (Bar Chart)

💻 Top Programming Languages (Bar Chart)

🤖 Top ML Frameworks (Bar Chart)

---

5. Conclusion & Key Takeaways

- 🔹 Python is the dominant language in data science (used by ~85%).
- 🔹 Gender imbalance persists (majority male respondents).
- 🔹 India & USA lead in participation, reflecting Kaggle's user base.
- 🔹 Scikit-learn is the go-to ML framework, followed by TensorFlow.
- 🔹 Most respondents are early/mid-career professionals (25-34 age group).

Future Work

- Analyze salary trends by country/experience.
- Compare tool preferences across different job roles.
- Track changes in trends over survey years (2017-2021).

---

6. Appendix: Code Repository

📂 GitHub Link: https://github.com/Insight_Generation_from_Survey

📄 Dataset Source: https://www.kaggle.com

Final Thoughts

This project successfully cleaned, analyzed, and visualized Kaggle survey data to extract meaningful insights. The summary dashboard effectively highlights key trends in the data science community.

Next Steps: Expand analysis to include salary trends, job role comparisons, and year-over-year changes.