**Task Four Work Description**

**Data Cleaning and Insight Generation from Survey Data**

In this task, the *Kaggle Data Science Survey (2017–2021)* dataset was analyzed to extract meaningful insights about respondent demographics, skills, and professional characteristics. The process began with data loading and inspection to understand the structure, missing values, and duplicates. Missing values in categorical features were replaced with the placeholder "Unknown", while numerical fields were imputed using median values. Duplicate entries were removed, irrelevant columns were dropped, and categorical variables were standardized to lowercase with whitespace removed. Label Encoding was applied to transform categorical variables into numerical form for further analysis.

Key insights were generated, including the top 5 countries of respondents, most common job titles, most frequently used programming languages, and the distribution of education levels. Visualizations were created using Matplotlib and Seaborn to present these findings. Additionally, an interactive dashboard was built using Plotly and Dash to allow dynamic exploration of results, featuring bar charts, pie charts, and treemaps for the top categories. This workflow ensured that the dataset was clean, consistent, and ready for effective insight generation.

**Results & Interpretation**

* **Top 5 Countries:** The largest respondent group came from the United States, followed by India, other major tech hubs, and regions with active data science communities. This reflects the dominance of certain countries in the global data science workforce.
* **Job Titles:** Data Scientist, Data Analyst, and Software Engineer emerged as the most common roles, indicating the survey’s focus audience and the field’s professional landscape.
* **Programming Languages:** Python was the most frequently used programming language, with R and SQL also appearing prominently, showing the core tools favored by data professionals.
* **Education Levels:** A large proportion of respondents held a Master’s or Bachelor’s degree, reflecting the educational background typical in data science roles.
* **Coding Experience:** Many respondents had between 3–5 years of programming experience, suggesting a strong base of mid-level professionals in the dataset.

The interactive dashboard further allowed filtering and comparison of these trends, making the analysis useful for identifying patterns in global data science demographics and skill distribution.