Project : Data jobs survey

Introduction

The 'tech_jobs' dataset contains information related to job roles, salaries, industries, programming languages, and desired job factors. However, before conducting any meaningful analysis on this dataset, it was necessary to perform data cleaning tasks to address issues such as irrelevant columns, inconsistent data formats, and unstandardized information.

Data Cleaning Analysis

This data cleaning analysis focuses on preparing the 'tech_jobs' dataset for further analysis by performing various cleaning and transformation steps. The following steps were carried out:

Drop Irrelevant Columns:

The initial step involves dropping irrelevant columns from the 'tech_jobs' dataset. These columns, such as 'Browser', 'OS', 'Email', etc., are not relevant to the specific analysis being conducted.

Splitting the 'Q3 - Current Yearly Salary (in USD)' Column:

The 'Q3 - Current Yearly Salary (in USD)' column is split into two separate columns, 'Salary1' and 'Salary2', by using the '-' delimiter. This enables the calculation of averages and facilitates a more detailed analysis of salary ranges.

Renaming Long Columns:

Long column names are renamed to shorter, more concise names for better readability and easier referencing during analysis. This step simplifies the dataset and makes it more manageable.

Mapping Job Titles to a Standardized Title:

Various job titles in the 'Current Role Title' column are mapped to a standardized title, specifically 'Data Analyst', based on specific keywords present in the job titles. This process categorizes similar job roles under a common title, aiding comparison and analysis.

Mapping Industries to a Standardized Title:

Industries listed in the 'Industry' column are mapped to a standardized title based on specific keywords present in the industries. This mapping helps group similar industries together, facilitating analysis and comparison within specific industry categories.

Mapping Programming Languages to a Standardized Title:

Programming languages mentioned in the 'ProgLanguage' column are mapped to a standardized title based on specific keywords present in the languages. This mapping process ensures consistency in categorizing programming languages, simplifying language preference analysis.

Consolidating Desired Job Factors:

Desired job factors, stored in the 'Desired Job Factor' column, are consolidated into meaningful categories. This consolidation process helps identify common themes and priorities among respondents, providing insights into their preferences and priorities in job selection.

Consolidating Countries into Continents:

To provide a broader geographical perspective, countries listed in the 'Residence' column are consolidated into continents. This consolidation aids in high-level analysis based on geographical regions, enabling comparisons and insights at a continental level.