Mini Project

Srushti Shobhane

SUID: 2277366342

Scripting for Data Analysis

Data Analysis Career Transition Analysis

Course number: IST 652

26th November 2024

Data and Source:

The primary dataset used in this analysis is the "Field of Study vs Occupation" dataset from Kaggle (https://www.kaggle.com/datasets/jahnavipaliwal/field-of-study-vs-occupation). This dataset contains information on the relationship between an individual's field of study and their current occupation.

The secondary dataset is the "HR Analytics: Job Change of Data Scientists" dataset, also from Kaggle (https://www.kaggle.com/datasets/arashnic/hr-analytics-job-change-of-data-scientists?select=aug_train.csv). This dataset provides details on the career transitions of data scientists, including information such as their experience, training hours, and whether they transitioned to a new role or not.

Data Exploration and Data Cleaning Steps:

The code first loads both the primary and secondary datasets using the load_data() method. It then performs the following data cleaning and preprocessing steps:

Handling missing values in numeric columns (such as "city_development_index" and "training_hours") by filling them with the median values.

Cleaning the "experience" column by filling in missing values, replacing outliers, and converting the data to a numeric format.

Filling in missing values in categorical columns (such as "gender", "education_level", and "major_discipline") with the "Unknown" label.

Standardizing the column names across the two datasets by converting them to lowercase and removing spaces.

Merging the two datasets based on the common columns of "education_level" and "gender" using the merge_and_clean_data() method.

Comparison Questions with Analysis:

The code addresses the following three analysis questions:

a. **What are the top fields of study that lead to a career in Data Science?**
   The plot_top_fields_for_data_science() method filters the merged dataset to include only data science-related occupations (Data Scientist, Data Analyst, Business Analyst). It then counts the number of people in each field of study and creates a bar plot to visualize the top fields that lead to data science careers.

b. **What is the relationship between educational background and career transitions in data-related occupations?**
 The plot_education_career_transition() method creates a stacked bar chart to show the career transition rates for each education level. This visualization helps understand the relationship between educational background and career transitions in data-related occupations.

c. **How do relevant skills (based on training) influence career progression in the data science domain?**
The plot_training_impact_on_progression() method creates a stacked histogram to show the distribution of training hours for both the transition and non-transition groups. This plot helps identify the impact of training and skill development on career progression in the data science domain.

Program Description:

The code is structured as a CareerTransitionAnalyzer class, which provides methods for loading, preprocessing, and merging the data, as well as generating the various visualizations. The class is initialized, and the data is loaded and processed before the visualization methods are called.

The program includes the following key methods:

load_data(): Loads the primary and secondary datasets.

preprocess_data(): Handles missing values and cleans the data.

merge_and_clean_data(): Merges the datasets based on common columns and standardizes the column names.

plot_top_fields_for_data_science(): Generates a bar plot showing the top fields of study leading to data science careers.

plot_education_career_transition(): Creates a stacked bar chart displaying the career transition rates by education level.

plot_training_impact_on_progression(): Generates a stacked histogram illustrating the impact of training hours on career progression.

Output Files Description:

The code does not generate any output files; it focuses on creating interactive visualizations to explore and analyze the data. The visualizations include:

A bar plot showing the top fields of study leading to data science careers

A stacked bar chart displaying the career transition rates by education level

A stacked histogram illustrating the impact of training hours on career progression

Additional Pointers:

- The code includes error handling and data validation checks to ensure the successful execution of the data loading and preprocessing steps.
- The visualization methods are added to the CareerTransitionAnalyzer class, making them easily accessible and reusable.
- The code could be further extended to include additional analysis, such as examining the impact of other factors (e.g., gender, geographic mobility) on career transitions and progression.
- The visualizations could be customized and refined to enhance their clarity and effectiveness for presenting the findings.