Module 2 assignment document

Healthcare Usage Analysis

This notebook performs an exploratory data analysis on a healthcare dataset to understand various factors influencing healthcare usage.

Dataset

The dataset contains information on individuals, including demographic details, insurance information, illness details, and healthcare usage metrics.

Analysis Steps

1. **Data Loading and Inspection:**
    - The dataset was loaded into a pandas DataFrame.
    - Initial inspection of the data was performed using head() and info() to understand the structure, data types, and identify missing values.
2. **Data Cleaning and Transformation:**
    - Categorical features ('insurance' and 'chcond') were converted to the 'category' data type for efficient storage and analysis.
    - 'age' and 'income' columns were converted to 'int64' data type.
3. **Descriptive Statistics:**
    - The sum and mean of several healthcare usage features were calculated to get an overview of the total and average usage.
    - The median values for numerical columns were calculated to understand the central tendency.
4. **Grouped Analysis:**
    - The data was grouped by 'sex' to analyze the average healthcare usage for different genders.
5. **Data Visualization:**
    - Histograms were plotted for numerical features ('age', 'income', 'illness', 'doctorco', 'hospdays') to visualize their distributions.
    - A pair plot was created to visualize the relationships between selected numerical features.
    - A box plot was generated to compare the distribution of 'illness' across different sexes.

Key Findings (Based on the executed code)

- The dataset contains no missing values.

- The summary_by_sex DataFrame shows the average usage of various healthcare services and conditions broken down by sex. (Further interpretation of these values can be added based on domain knowledge).
- The histograms and box plots provide visual insights into the distribution and relationships of key variables.

This notebook provides a foundational analysis of the healthcare usage data. Further analysis could involve more in-depth statistical testing, machine learning models for prediction, and more sophisticated visualizations.