Capstone 1

Pandas Exercise

Bassel Dakhlallah

Data Dictionary

Variable	Description	Variable	Description
visits	Number of physician office visits	**health**	Factor indicating self-perceived health
nvisits	Number of non-physician office visits	**chronic**	Number of chronic conditions
ovisits	Number of physician hospital outpatient visits	**adl**	Factor indicating whether the individual has a condition that limits activities of daily living
novisits	Number of non-physician hospital outpatient visits	**region**	Factor indicating region
emergency	Emergency room visits	**age**	Age in years (divided by 10)

Data Dictionary

Variable	Description	Variable	Description
hospital	Number of hospital stays	**afam**	Factor. Is the individual African-American?
gender	Factor indicating gender	**married**	Factor. Is the individual married?
school	Number of years of education	**income**	Family income in USD 10000
employed	Factor. Is the individual employed?	**insurance**	Factor. Is the individual covered by private insurance?
medicaid	Factor. Is the individual covered by Medicaid?		

Section 1

Data Import, Inspection, and Cleaning

Please provide your observations for each output.

1. Import Libraries: Begin by importing the necessary Python libraries for programming and numerical operations, such as pandas and numpy.
2. Load CSV Data: Import the **NSMES1988.csv** file into a Pandas DataFrame.
3. Inspect the Data: Use Pandas functions to examine the dataset, including checking the number of rows and columns, data types, duplicates, nulls, and a general overview of the data.
4. Memory Optimization: Analyze the memory usage of the DataFrame and suggest appropriate non-default data types to optimize memory.
5. Data Cleaning: drop any unnecessary columns, such as those labeled "unnamed."

Section 2

Data Transformation and Aggregation

Please provide your observations for each output.

1. Column Updates: Multiply the values in the 'Age' column by 10. Also, multiply the values in the 'Income' column by 10,000.
2. Value Ranges: Calculate the counts of values for different age and income ranges.
3. Aggregation:
    1. Calculate the total number of visits by gender.
    2. Compute the average income by health status.
    3. Calculate the average number of visits by region.
4. Crosstab and Pivot Table:
    1. Build a crosstab between two categorical columns (e.g., health and insurance).
    2. Create a pivot table for health and gender, with hospital values and aggregate both sum and mean.

Section 3

Data Export and Visualization

Please provide your observations for each output.

1. Export to CSV: Export the cleaned DataFrame to a CSV file with pipe (|) delimiters and ensure the index is excluded.
2. Scatter Plot: Create a scatter plot to visualize the relationship between income and age, and provide your observation based on the plot.
3. Advanced Aggregation: Use groupby() and agg() functions to perform more advanced aggregation, such as calculating total visits or average income by specific categories (e.g., by region and gender).