Capstone Session 2

Python for Data Analysis

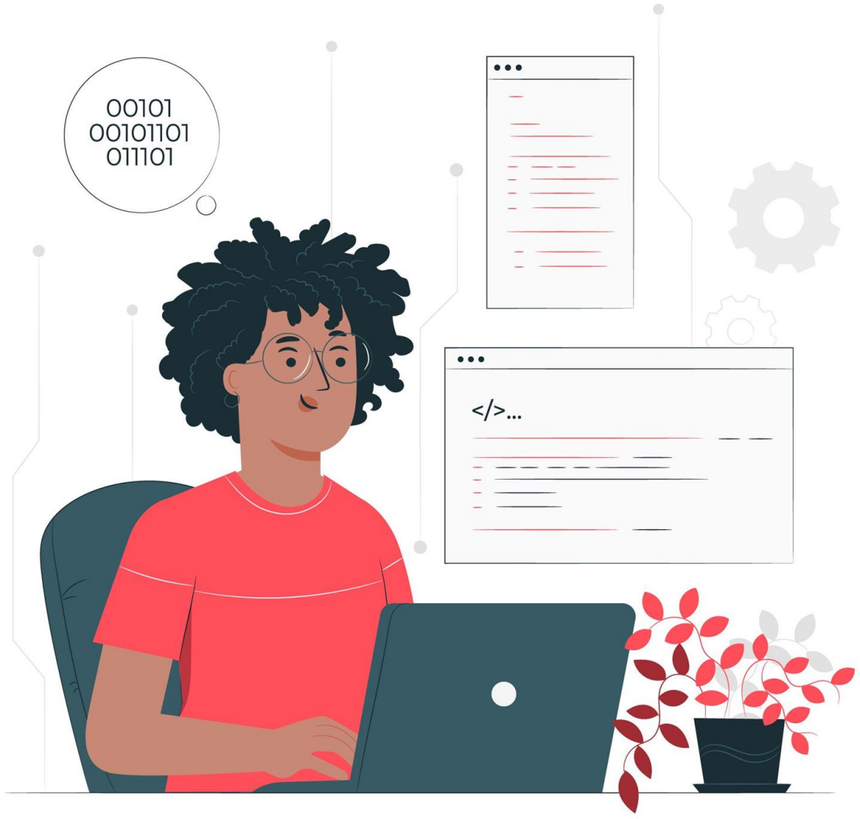
Data Preprocessing with Python

Aura must be built to receive and process marketing campaign and user behavior data from various sources such as healthcare, technology and manufacturing domains.

- Aggregate data
- Shape data
- Understand the statistical analysis of data
- Normalize data

Project Statement

Develop a comprehensive solution for data aggregation, wrangling, and visualization using a healthcare dataset for the Aura platform



. The goal is to effectively manage and process complex healthcare data to enable insightful analysis and enhance data-driven decision-making capabilities within Aura.

Dataset Description

NSMES1988.csv

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)

Dataset Description

NSMES1988.csv

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

Session 2: Data Processing and Statistical Analysis

- **Task: Data Processing and Statistical Analysis**
- Import relevant Python libraries.
- Import the CSV file – NSMES1988new.csv into a dataframe.
- Perform memory analysis of the new dataframe and compare it with the memory of the dataframe in the previous week and mark your comments.
- Perform the following operations on age and income columns. Multiply age by 10 and income by 10000.
- Perform basic statistical analysis on the new dataframe and generate a brief report on the outcome. Save the dataframe as NSMES1988updated.csv file in the local space for possible future use.
- Invoke describe command on the dataframe and compare that with the basic statistical analysis done in the previous step.

Session 2: Data Processing and Statistical Analysis

- Indicate which of the columns are not eligible for statistical analysis and indicate possible datatype changes, and report.
- Make changes to the recommended file from previous step in the previous step, export it as a new .csv file for possible future use (Optional).
- Prepare a brief report and enter it in the mark-up cells of JupyterLab Notebook.

Thank You