Capstone Session 3

Python for Data Analysis

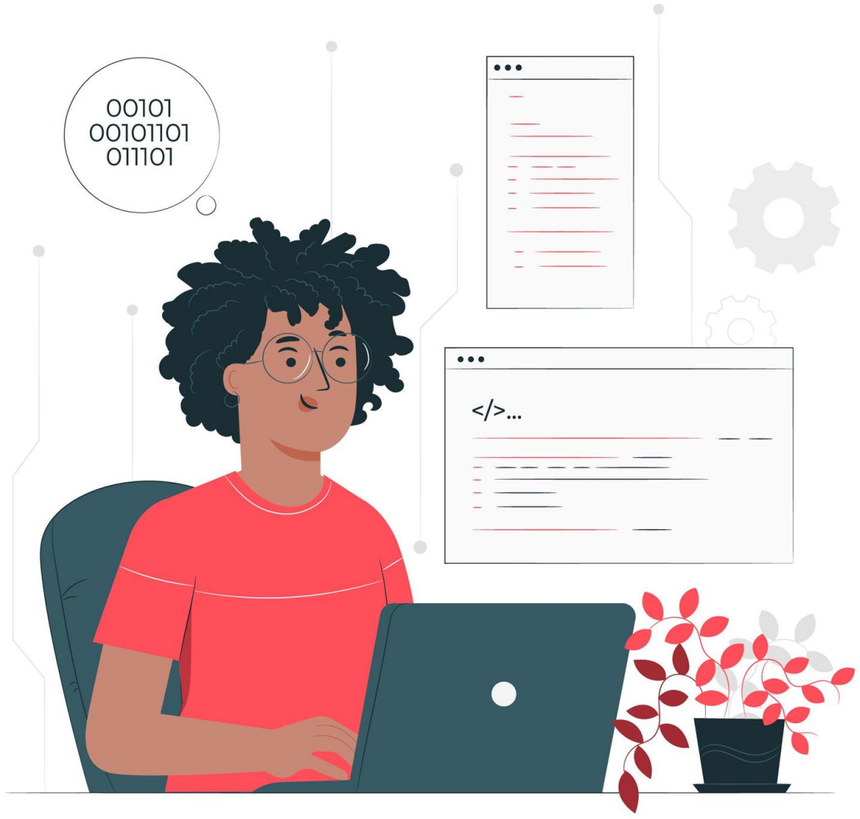
Data Preprocessing with Python

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 3: Data Analysis with Pandas

Task: Working with Pandas

- Import relevant Python libraries necessary for Python and Pandas analysis.
- Import the CSV file NSMES1988updated.csv file and create a new dataframe for working with Pandas.
- Identify different types of data and report it.
- Identify Categorical types in the data.
- Perform a detailed Data pivoting on the dataframe and report it.
- Include the following categorical data in your analysis
  - Health
  - Region
- Prepare a detailed report on your analysis and observations.

Session 3: Data Analysis with Pandas

Task: Analyze and Cleanse data

- Import relevant Python libraries necessary for Python and Pandas analysis.
- Import the CSV file NSMES1988updated.csv file and create a new dataframe for working with Pandas.
- Perform analysis based on the following criteria: Different types of visits, Gender, Marital Status, School, Income, Employment Status, Insurance, Medical Aid
- Explore and analyze the dataset to gain insights into how different factors relate to each other within the dataset. By grouping the data according to specific demographic and economic criteria, create a series of distribution tables by considering the below instructions:
- **Age and Gender Distribution:** Generate a table to view the number of individuals within each age group, separated by gender.
- **Health Status by Gender:** Create a distribution table that categorizes individuals by their health status, differentiated by gender.

Session 3: Data Analysis with Pandas

- **Income Distribution by Gender:** Compile a table to examine the income distribution across genders.
- **Regional Income Distribution:** Prepare a table to display the income distribution across various regions.
- **Age-wise Income Analysis:** Develop a table to analyze the relationship between age and income.

Report your findings.

Thank You