Incremental Capstone - Session 1

Theme: Exploratory Data Analysis

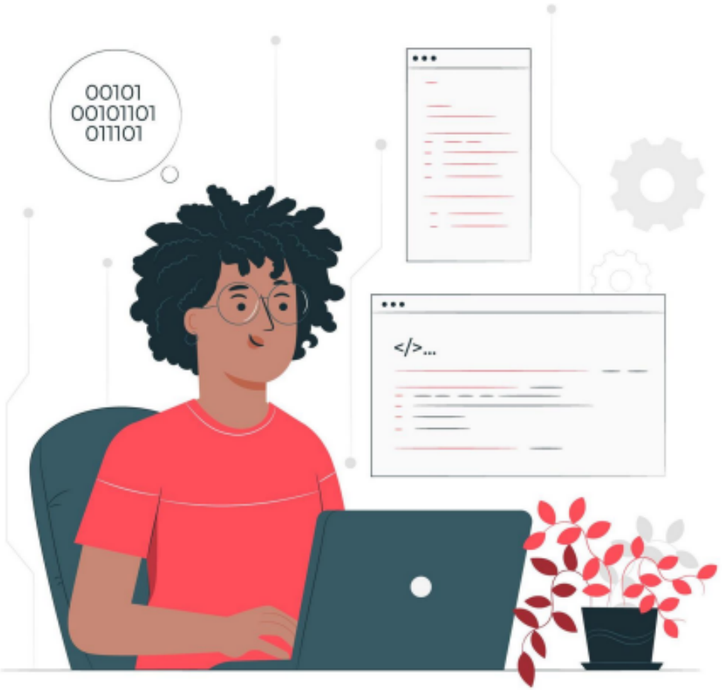
Exploratory Data Analysis (EDA) is a crucial first step in any data analysis or data science project, involving the use of visualizations and summary statistics to understand the characteristics and relationships within a dataset. EDA aims to explore, investigate, and learn about the data to uncover patterns, identify potential issues, and generate hypotheses for further analysis, rather than confirming specific statistical hypotheses.

Goal: To mimic a real world *Data Analyst* team

- **How:** Given real data set, a team member(s), and expected to generate actionable results
- **Who:** Teams of 2 - (One team will have 3)
- **When:** *Right now!!*

Project Statement

Your team has been contacted by a hospital to extract insights from some data.



- You are given a csv file with the hospital's raw data. You are expected to perform data aggregation, wrangling, and visualization using the healthcare dataset.
- The first member of the team will perform the coding and share their screen with the other member(s), the second member will thoroughly document your team's procedures. If there is a third person they should be helping research code from documentation.
- The hospital's goal is to effectively manage and process complex healthcare data to enable insightful analysis and enhance **data-driven** decisions.

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)

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?		

Data Processing and Statistical Analysis

Task A - 1 hour

- Import Python libraries.
- Import the CSV file → NSMES1988.csv into a dataframe.
- Check for missing values in the data.
- Perform memory analysis of the dataframe and mark your comments.
- Perform the following operations on the columns.
    - Rename any nondescript columns to be more specific.
    - Multiply age by 10 and income by 10000.
    - Would any column benefit from changing data types?
        - Indicate possible data type changes, in the report.
- Save the dataframe as ‘NSMES1988_updated.csv’ file in the local space for future use.
- Perform basic statistical analysis on your dataframe and generate a brief report on the outcome.
- Invoke describe command on the dataframe and compare that with the basic statistical analysis done in the previous step.
- Indicate which of the columns are not eligible for statistical analysis.
- Prepare a brief report and enter it in the markdown cells of a JupyterLab Notebook.

Data Visualizations

Task B - 1 hour

- Import Data-Viz libraries
- Generate a plot depicting the relative number of people who were insured vs not insured
  - Make sure to document your findings
- Generate plots depicting the correlation (if any) between a given type of hospital visit and the health of individuals.
  - I.e. Do people with health category, x, visit the hospital in a given manner more, the same, or less often than people in health category, y, typed people?
  - Make sure to document your findings
- Is there any relationship between income and private insurance? Between income and Medicaid?
  - Make sure to document your findings
- Create 2 or more plots of your choosing that help you tell a story with the data.
  - (*Experimentation may be required to find something worth reporting on, or just go with your instincts*)
  - This is completely up to your Team
  - Make sure to document your findings
  - Make sure there are insights to be learned with your plots

Report Generation

- **Task C - 30 Minutes**
- Generate a final report of your team's EDA for the Hospital
- Ensure that you detail →
  - All actions your Team performed on the data
  - Any insights you had while working with the data
  - Is there any recommendations to the hospital for future data collection, (What could they do better)
- Create a story with the data, and use your graphs to tell that story -
  - Make sure that your description is through
  - What information is being learned from this analysis?
  - What questions are still unknown?