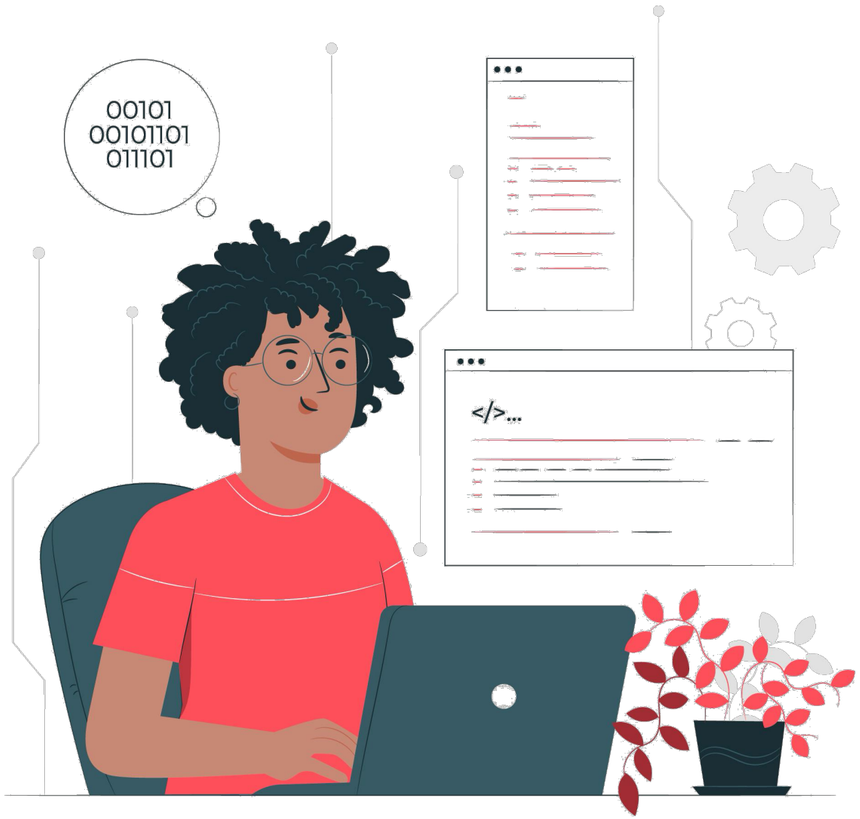
DATA AND
ARTIFICIAL INTELLIGENCE

Capstone Session
EDA and Visualization

Project Statement

Build necessary data aggregation, wrangling and visualization modules for your project using the Healthcare dataset.



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	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?		

Working with Pandas and Non-graphical EDA

Tasks:

- Preparation:
  - Import python libraries necessary for the analysis.
  - Import the CSV file NSMES1988.csv file as a dataframe df
  - Convert age and income to their proper values based on the data dictionary
- Identify different data types
- Identify Categorical types in the data.
- Get basic statistical measures for the numerical columns in the dataset
- Non-graphical analysis using data aggregation:
  - What is the total number of hospital stays for different employment statuses?
  - Build a pivot table that shows all numerical measures with health column as an index
  - What is the median number of emergency room visits for each region and gender combination?
  - Get the number of individuals covered by Medicaid and its effect on the number of hospital stays

NOTE: Use markdown to provide your comments and observations for each ask.

Working with Pandas and Data Visualization

Task: Visualize and Analyze data

Visualization tips:

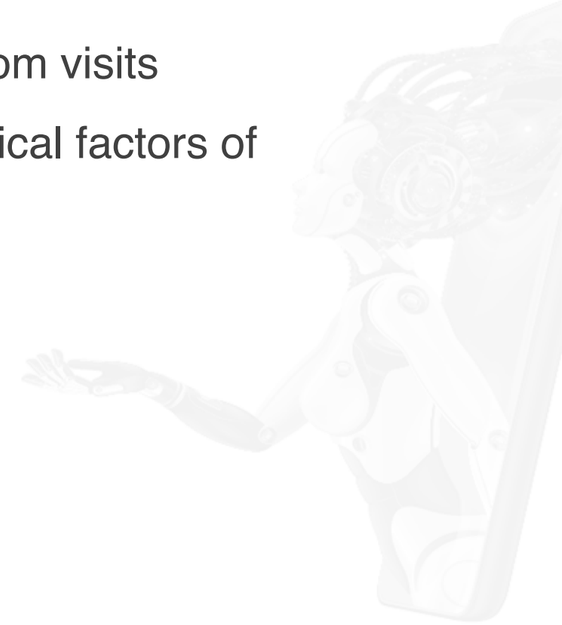
- Choose proper plot size to demonstrate data points clearly
- Add plot and axis titles
- Use bmh or ggplot matplotlib style (hint: **plt.style.use('ggplot'))**

- Build a histogram for every numerical column using subplots and report the columns that have a lot of outliers
- Perform box and whisker analysis based on the following categories:
  - Different types of visits
  - Gender
  - Marital Status
  - Employment Status
- Generate a plot to view the number of individuals within each age group, separated by gender.
- Build a plot to check if there's a relationship between Age and number of physician's office visits.

NOTE: Use markdown to provide your comments and observations for each ask.

Working with Pandas and Data Visualization

- Check the correlation between having chronic conditions and emergency room visits
- Regional Income Distribution: build a violin plot to display the income distribution across various regions.
- Build a correlation heatmap for the columns that highly correlate with emergency room visits
- Build a scatterplot to analyze the relationship between age and income with categorical factors of your choice. *Hint: use size, hue, and style.*
- Upload your work to slack when done.



NOTE: Use markdown to provide your comments and observations for each ask.

DATA AND
ARTIFICIAL INTELLIGENCE

Thank You

simplilearn