##### **CODE PLATEAU FELLOWSHIP DATA SCIENCE TRACK**

##### **CAPSTONE PROJECT 1: Plateau Insurance Company.**

**Background:**

Leveraging customer information is of paramount importance for most businesses. In the case of an insurance company, the attributes of customers can be crucial in making business decisions. Hence, knowing to explore and generate value out of such data can be an invaluable skill to have.

You are hired as a Data Scientist in an Insurance company in Plateau State. The company wants to have a detailed understanding of the customer base for one of its Insurance Policy 'MediClaim'. The idea is to generate insights about the customers and answer a few key questions with statistical evidence, by using the past dataset. The dataset 'plateau\_Insurance' contains customers' details like age, sex, charges, etc. Perform the statistical analysis to answer the following questions using the collected data.

**Objective:**

Perform an Analysis of the given Data. Explore the dataset and extract insights from the data. The idea is for you to get comfortable with doing analysis in Python.

You are expected to do the following:

1. Explore the dataset and extract insights using Exploratory Data Analysis.
2. Prove(or disprove) that the medical claims made by the people who smoke are greater than those who don't?
3. Prove (or disprove) that the BMI of females is different from that of males.
4. Does the smoking habit of customers depend on their region? [Hint: Create a contingency table using the pandas.crosstab() function]
5. Is the mean BMI of women with no children, one child, and two children the same?

**Data Dictionary** –

1. Age - This is an integer indicating the primary beneficiary's age (excluding those above 64 years, since the government generally covers them).
2. Sex - This is the policyholder's gender, either male or female.
3. BMI - This is the body mass index (BMI), which provides a sense of how over or underweight a person is relative to their height. BMI is equal to weight (in kilograms) divided by height (in meters) squared. An ideal BMI is within the range of 18.5 to 24.9.
4. Children - This is an integer indicating the number of children/dependents covered by the insurance plan.
5. Smoker - This is yes or no depending on whether the insured regularly smokes tobacco.
6. Region - This is the beneficiary's place of residence in the U.S., divided into four geographic regions - northeast, southeast, southwest, or northwest.
7. Charges - Individual medical costs billed to health insurance

**Best Practices for Notebook:**

* The notebook should be well-documented, with inline comments explaining the functionality of code and markdown cells containing comments on the observations and insights.
* The notebook should be run from start to finish in a sequential manner before submission.
* It is preferable to remove all warnings and errors before submission.
* The notebook should be submitted as an HTML file (.html) and NOT as a notebook file (.ipynb).(google how to download jupyter notebooks as html file)

**Submission Guidelines:**

Once completed, you are expected to submit a folder(use your name then project 1 as the folder name), and you are expected to submit 2 files in that folder. Please note the following:

1. There are two parts to the submission:
   1. A well-commented Jupyter notebook [format - .html]
   2. A presentation (Report) as you would present to the top management/business leaders [format - .pdf ] (you have to export/save the .pptx file as .pdf)
2. Any assignment found copied/plagiarised by other people will not be graded and awarded zero marks
3. Please ensure timely submission as a post-deadline assignment will not be accepted
4. Submission will not be evaluated if,
   1. it is submitted post-deadline, or,
   2. more than 2 files are submitted

##### **Scoring guide (Rubric) - Plateau Insurance Project**

| **Criteria** | **Points** |
| --- | --- |
| **Define the problem and perform an Exploratory Data Analysis** - Problem definition, questions to be answered - Data background and contents - Univariate analysis - Bivariate analysis |  |
| **Illustrate the insights based on EDA** Key meaningful observations on individual variables and the relationship between variables |  |
| **Prove (or disprove) that the medical claims made by the people who smoke is greater than those who don't?** Perform the statistical analysis and visual analysis |  |
| **Prove (or disprove) with visuall evidence that the BMI of females is different from that of males.** Perform the statistical analysis and visual analysis |  |
| **Does the smoking habit of customers depend on their region?** Perform the statistical analysis and visual analysis |  |
| **Is the mean BMI of women with no children, one child and two children the same? Explain your answer with statistical evidence.** Perform the statistical analysis and visual analysis |  |
| **Presentation - Overall quality** - Structure and flow - Crispness - Visual appeal - Key insight and recommendation based on EDA |  |
|  |  |