**Objective** – Statistical Analysis of Business Data. Explore the dataset and extract insights from the data. The idea is for you to get comfortable with doing statistical 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 is greater than those who don't? [Hint- Formulate a hypothesis and prove/disprove it]
3. Prove (or disprove) with statistical evidence that the BMI of females is different from that of males.
4. Is the proportion of smokers significantly different across different regions? [Hint : Create a contingency table/cross tab, Use the function : stats.chi2\_contingency()]
5. Is the mean BMI of women with no children, one child, and two children the same? Explain your answer with statistical evidence.

\*Consider a significance level of 0.05 for all tests.

**Context** - Leveraging customer information is of paramount importance for most businesses. In the case of an insurance company, attributes of customers like the ones mentioned below 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.

**Data Dictionary** –

1. Age - This is an integer indicating the age of the primary beneficiary (excluding those above 64 years, since they are generally covered by the government).
2. Sex - This is the policy holder's gender, either male or female.
3. BMI - This is the body mass index (BMI), which provides a sense of how over or under-weight 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

**Submission Guidelines:**

Please note the following:

1. There are two parts to the submission:
   1. A well commented Jupyter notebook [format - .ipynb]
   2. A presentation as you would present to the top management [format - .ppt /.pptx]
2. Any assignment found copied/ plagiarized with other groups will not be graded and awarded zero marks
3. Please ensure timely submission as a post-deadline assignment will not be accepted
4. Your submission will not be evaluated if it contains more than 2 files.

Happy Learning!!

**Scoring guide (Rubric) - Axis Insurance Project**

| **Criteria** | **Points** |
| --- | --- |
| **Explore the dataset and extract insights using Exploratory Data Analysis.** | 20 |
| **Prove (or disprove) that the medical claims made by the people who smoke is greater than those who don't?** | 10 |
| **Prove (or disprove) with statistical evidence that the BMI of females is different from that of males.** | 10 |
| **Is the proportion of smokers significantly different across different regions?** | 10 |
| **Is the mean BMI of women with no children, one child and two children the same? Explain your answer with statistical evidence.** | 10 |
| Points | 60 |

