1. ***Explaining the reasoning behind grading criteria.***

Problem Statement:

About data:

The insurance.csv dataset contains 1338 observations and 7 attributes. Context: The data contains medical costs of people characterized by certain attributes. Let's see if we can dive deep into this data to find some valuable insights.

Attributes:-

- age: age of primary beneficiary
- sex: insurance contractor gender, female, male
- bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
- children: Number of children covered by health insurance / Number of dependents
- smoker: Smoking
- region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- charges: Individual medical costs billed by health insurance.

Tasks to perform:

1. Import the necessary libraries (2.5 Marks)

2. Read the data as a data frame (2.5 Marks)

3. Perform basic EDA which should include the following and print out your insights at every step. (25 Marks)

a. Shape of the data (2.5 Marks)

b. Data type of each attribute (2.5 Marks)

c. Checking the presence of missing values (2.5 Marks)

d. 5-point summary of numerical attributes (2.5 Marks)

e. Distribution of 'bmi', 'age' and 'charges' columns. (5 Marks)

f. Measure of skewness of 'bmi', 'age' and 'charges' columns - optional

g. Checking the presence of outliers in 'bmi', 'age' and 'charges columns - optional

h. Distribution of categorical columns (including children) (5 Marks)

i. Pair plot that includes all the columns of the data frame (along with explaining inferences). (5 Marks)

4. Answer the following questions with statistical evidence (30 Marks)

a. Do charges of people who smoke differ significantly from the people who don't? (7.5 Marks)

b. Does bmi of males differ significantly from that of females? (7.5 Marks)

c. Is the proportion of smokers significantly different in different genders? (7.5 Marks)

d. Is the distribution of bmi across women with no children, one child and two children, the same? (7.5 Marks)

2. ***What will be the correct approach for doing the assignment?***

Following all the listed steps as specified in the original project assignment with elaboration on each task as per allocated marks.

1. Import the necessary libraries (2.5 Marks)

2. Read the data as a data frame (2.5 Marks)

3. Perform basic EDA which should include the following and print out your insights at every step. (25 Marks)

a. Shape of the data (2.5 Marks)
b. Data type of each attribute (2.5 Marks)
c. Checking the presence of missing values (2.5 Marks)
d. 5-point summary of numerical attributes (2.5 Marks)
e. Distribution of 'bmi', 'age' and 'charges' columns. (5 Marks)
f. Measure of skewness of 'bmi', 'age' and 'charges' columns - optional
g. Checking the presence of outliers in 'bmi', 'age' and 'charges columns - optional
h. Distribution of categorical columns (including children) (5 Marks)
i. Pair plot that includes all the columns of the data frame (along with explaining inferences). (5 Marks)

4. Answer the following questions with statistical evidence (30 Marks)

a. Do charges of people who smoke differ significantly from the people who don't? (7.5 Marks)
b. Does bmi of males differ significantly from that of females? (7.5 Marks)
c. Is the proportion of smokers significantly different in different genders? (7.5 Marks)
d. Is the distribution of bmi across women with no children, one child and two children, the same? (7.5 Marks)

3. ***What most of the students have missed in the assignment? / Common mistakes made by students (Wrong Assumptions)***

Overall everyone did their best in approaching the problem and solving the solution except with few gaps as listed below.

1. EDA: Majority of students have missed to provide meaningful insights from Pair plot analysis e.g. a. Presence of highly correlated variables. b. Inferences from bi-variate analysis
   c. Presence of multiple gaussians can signify a possibility of mixing up of data from different processes into one single data file or this could be even valid if this nature of the data is just due to a statistical fluke, which can only be confirmed true or false by applying Statistical sampling tests such as 2-way T test or 2-Sample T-test.

2. EDA: Inferences about at least one of listed below listed questions are missing by more than half of the students.

   3.d. 5-point summary of numerical attributes e.g. AGE and BMI are fairly normally distributed because Q2 or 50th percentile value is closer to Mean value. Presence of no missing values among the Numerical Features.

   3.h. Distribution of categorical columns (including children)

   3.i. EDA: Providing inferences from Pair plot analysis is missing. e.g a. Presence of highly correlated variables. b. Inferences from bi-variate analysis.

3. Few of the students have missed to construct Null & Alternative Hypothesis Statements (H0 and H1) for at least one of the four listed tasks in the question Q4.

4. Nearly half of the students have missed to apply statistical analysis while solving the following 4 questions, as it's clearly indicated in the question to solve using "statistical evidence" but most of them must have been assumed that it would be sufficient to answer by using visual analysis (such as bar plots or distribution plots or bi-variate analysis) without even trying statistical tests.

   4. Answer the following questions with statistical evidence
   a. Do charges of people who smoke differ significantly from the people who don't?
   b. Does bmi of males differ significantly from that of females?
   c. Is the proportion of smokers significantly different in different genders?
   d. Is the distribution of bmi across women with no children, one child and two children, the same?

5. Hypothesis and Null Hypothesis statements formulation are not very meaningful by most of the students, though their overall hypothesis testing validation approach is correct.

4. *How this assignment will add value to their perspective?*

In this assignment we are trying to provide valuable business insights using EDA (i.e. using both Uni-Variate and Bi-variate Analysis) and also tries to answer about what factors are likely to influence the Medical costs of the patients (which is our Target variable: "**charges**") i.e. What is the list of input variables that are most likely influencing the Target variable or at least helps me to determine medical costs on any given patient? Also, we try to answer the various types of research questions **with statistical evidences** by using various Inferential statistical tests such as Hypothesis testing techniques such as T-test, ANOVA, Chi-square test of independence, Test of proportions.

Skills in EDA helps one to find hidden patterns in the data that may or may not relevant to existent business problem but still can benefit business in driving decisions to create a positive impact. Whereas the statistical validations such as Null Hypothesis testing helps to confirm the assumption of the business to go ahead on what they are seeing from EDA about data is likely to be true or just a statistical fluke(untrue) i.e. is it statistically significant(to help

business to drive decisions) or just a statistical fluke(alerting business not to go ahead on what they have seen might have happen just by accident or by chance).