Dear Participants,

Please find below Advanced Statistics Project instructions:

- Submissions: 2 separate files

1. **Business Report: Submit answers to all the questions in a sequential manner.** Your report must **include a detailed explanation of the approach taken, inferences, and insights. Include outputs such as graphs, tables, and all other relevant information.** Business Report must not include any codes. **You will be evaluated based on Business Report only**. Hence please ensure that your Business Report is logical and detailed enough (without any code) for a reader somewhat conversant in analytics to understand the solution mechanism. 6 Marks are allotted for the "Quality of Business Report".
2. **Jupyter Notebook File**: This is a must and will be used for reference while evaluating

- Any assignment found copied/ plagiarized with another person will not be graded and marked as zero.
- Please ensure timely submission as a post-deadline assignment will not be accepted.

Problem 1A:

Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals [SalaryData.csv] are collected and each person's educational qualification and occupation are noted. Educational qualification is at three levels, High school graduate, Bachelor, and Doctorate. Occupation is at four levels, Administrative and clerical, Sales, Professional or specialty, and Executive or managerial. A different number of observations are in each level of education – occupation combination.

[Assume that the data follows a normal distribution. In reality, the normality assumption may not always hold if the sample size is small.]

1. State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually.
2. Perform a one-way ANOVA on Salary with respect to Education. State whether the null hypothesis is accepted or rejected based on the ANOVA results.
3. Perform a one-way ANOVA on Salary with respect to Occupation. State whether the null hypothesis is accepted or rejected based on the ANOVA results.
4. If the null hypothesis is rejected in either (2) or in (3), find out which class means are significantly different. Interpret the result. **(Non-Graded)**

Problem 1B:

1. What is the interaction between two treatments? Analyze the effects of one variable on the other (Education and Occupation) with the help of an interaction plot.[hint: use the 'pointplot' function from the 'seaborn' function]
2. Perform a two-way ANOVA based on Salary with respect to both Education and Occupation (along with their interaction Education*Occupation). State the null and alternative hypotheses and state your results. How will you interpret this result?
3. Explain the business implications of performing ANOVA for this particular case study.

Problem 2:

The dataset Education - Post 12th Standard.csv contains information on various colleges. You are expected to do a Principal Component Analysis for this case study according to the instructions

given. The data dictionary of the 'Education - Post 12th Standard.csv' can be found in the following file: Data Dictionary.xlsx.

- Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?
- Is scaling necessary for PCA in this case? Give justification and perform scaling.
- Comment on the comparison between the covariance and the correlation matrices from this data [on scaled data].
- Check the dataset for outliers before and after scaling. What insight do you derive here? [Please do not treat Outliers unless specifically asked to do so]
- Extract the eigenvalues and eigenvectors.[print both]
- Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features
- Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only).
- Consider the cumulative values of the eigenvalues. How does it help you to decide on the optimum number of principal components? What do the eigenvectors indicate?
- Explain the business implication of using the Principal Component Analysis for this case study. How may PCs help in the further analysis? [**Hint:** Write Interpretations of the Principal Components Obtained]

Please reflect on all that you have learned while working on this project. This step is critical in cementing all your concepts and closing the loop. Please write down your thoughts here.

Scoring guide (Rubric) - Project - New Advanced Statistics (1)

Criteria	Points
1.1 State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually.	4
1.2 Perform one-way ANOVA for Education with respect to the variable 'Salary'. State whether the null hypothesis is accepted or rejected based on the ANOVA results.	4
1.3 Perform one-way ANOVA for variable Occupation with respect to the variable 'Salary'. State whether the null hypothesis is accepted or rejected based on the ANOVA results.	4
1.4 If the null hypothesis is rejected in either (1.2) or in (1.3), find out which class means are significantly different. Interpret the result.	0
1.5 What is the interaction between the two treatments? Analyze the effects of one variable on the other (Education and Occupation) with the help of an interaction plot.	4
1.6 Perform a two-way ANOVA based on the Education and Occupation (along with their interaction Education*Occupation) with the variable 'Salary'. State the null and alternative hypotheses and state your results. How will you interpret this result?	4
1.7 Explain the business implications of performing ANOVA for this particular case study.	2

2.1 Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?	4
2.2 Is scaling necessary for PCA in this case? Give justification and perform scaling.	2
2.3 Comment on the comparison between the covariance and the correlation matrices from this data.[on scaled data]	3
2.4 Check the dataset for outliers before and after scaling. What insight do you derive here?	3
2.5 Extract the eigenvalues and eigenvectors.[print both]	4
2.6 Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features	4
2.7 Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only).	4
2.8 Consider the cumulative values of the eigenvalues. How does it help you to decide on the optimum number of principal components? What do the eigenvectors indicate?	4
2.9 Explain the business implication of using the Principal Component Analysis for this case study. How may PCs help in the further analysis? [Hint: Write Interpretations of the Principal Components Obtained]	4
Quality of Business Report	6
Please reflect on all that you learnt and fill this reflection report - https://forms.gle/2DF5i1W7sUXXmsaw5	0
Points	60