**Birla Institute of Technology & Science, Pilani**

**Work-Integrated Learning Programmes Division**

**First Semester 2021-2022**

**End-Semester Test (EC-3 Makeup)**

Course No. : DSECL ZG555

Course Title : DATA VISUALIZATION & INTERPRETATION

Nature of Exam : Open Book

Weightage : 40%

Duration :

Date of Exam :

Note:

1. Please follow all the *Instructions to Candidates* given on the cover page of the answer book.
2. All parts of a question should be answered consecutively. Each answer should start from a fresh page.
3. Assumptions made if any, should be stated clearly at the beginning of your answer.

| 1 | Given the following scatter plot - Redraw the plot with some reference line or pre-attentive attributes to highlight and answer following questions for audience. Create different plots as needed to avoid clutter and effectively answer the following questions.    A) Which countries are happier than India [2M]  B) How rich are countries that are as happy as India [2M]  C) Highlight if there are any relationships that exist between rich countries and happier citizens. [2M]  D) Highlight any outlier countries that do not fall on any existing relation identified in question C [2M]    The vertical axis shows the national average of the self-reported life satisfaction on a scale ranging from 0-10, where 10 is the highest possible life satisfaction (**Happy**). The horizontal axis shows GDP per capita adjusted for inflation and cross-country price differences (**Rich**)    **Solution-**  **1.5 marks for visual**  **0.5 marks for explanation**  A)  B)  C) & D) | 8M |
| --- | --- | --- |
| 2 | Your team is building warplanes. When making warplanes, you need to consider armour. But armour is heavy. And heavier planes are slower and less fuel efficient. Not having enough armour and having too much armour are both problems. As a data analyst of your team, you went ahead and reviewed the recorded data from various plane that returned from combat.  The recorded shots and damage severity are plotted below. Assume that its equally likely to get shot at any part of the plane. Red highlights more severe compared to green. Analyse the data and draw a warplane with armour with justification.    **Solution –**    Black - Represents the most critical and vulnerable region of the plan that must be armoured.  Full credits to students if they only mark the black region with justification. Saying that black region was surrounded by regions that got severely impacted by bullets. And the black region is most vulnerable as flights did not return from war if they got hit in those areas and therefore no data is available for those regions. A classic example of - "Sometimes most important is the data that is not available".  Grey is 2nd most vulnerable. If students mark both black and grey with correctly mentioning black as the most vulnerable in their justification. Still 2 marks can be deducted as that will not be optimal and efficient  If students mark only the grey region - 3 marks can be deducted as they missed the most vulnerable region. Identifying the black region is a key part to this question.  Anything else can be given 0 marks. | 4M |
| 3 | Assume a CSV file (filename.csv) with multiple rows and columns.  Data has both continuous and categorical data.  Below is some sample data -    Column Name has following abbreviations –  Num – Numerical  Cont – Continuous  Cat - Categorical  Use Matplotlib or Seaborn code to plot following -  A) Read the data in a panda Dataframe. Create a single figure with multiple axes [1M]  B) axis-1 : Create histograms with columns Col-Num-Cont-1 and Col-Num-Cont-2 in a single figure. Avoid any obtrusion in the plot. [1M]  C) axis-2: A bar plot on Col-Num-Cat-1. Plot the bar height text at center of the bar. [1M]  D) axis-3: A scatter plot on Col-Num-Cont-2, Col-Num-Cont-3 to identify presence of outliers and number of modes in the data [1M]  E) axis-4 : A plot to identify relation or correlation between : A scatter plot on Col-Num-Cont-2, Col-Num-Cont-3 and group the scatter plot on Col-Num-Cat-1. Fit a regression line for all the categories [1M] | 5M |
| 4 | Study the dashboard below and identify the mistakes in dashboard designs with justification. Choose any 4 from the 13 common mistakes in dashboard design discussed in the class.    **Solution-**   * Supplying inadequate context for the data * Choosing a deficient measure * Introducing meaningless variety * Arranging the data poorly * Highlighting important data ineffectively or not at all * Cluttering the display with useless decoration * Misusing or overusing color * Designing an unattractive visual display | 4M |
| 5 | Corona virus is mutating and we have seen various variants in last couple of months.  Chronological order of the variants shown below are **Ancestral -> Alpha -> Delta -> Omnicron**. Covid cases are declining. Study the trend of fatality and spread rate for various variant shown below. Predict and plot a visual with all possible areas where the new variant could fall (Assume only the below data as source of truth and there are no other external influences.). Provide justification for your analysis and prediction. Can this pandemic be called as an endemic with the data and visual provided below. Give justification.    **Solution –**  Visual-2M + Justification 1M  Can be called endemic + justification - 1M | 4M |
| 6 | Your team has developed a video and chat messenger. This messenger has become popular. Design a dashboard to monitor and show aspect-based **sentiment analysis** of customer reviews. The reviews are categorized by aspect **(Support, Reliability, Usability, etc.)**, then by sentiment, so team understands which aspects are performing positively and which negatively:  Draw a sample dashboard making use of best practices like (**sparklines, bullet bar chart for compact design**) and avoiding 13 common design mistakes discussed in the class. Assume sample data and various context as needed to create the dashboard.  **List down various best practices used (atleast 2) & few common design mistakes avoided in your dashboard(atleast 2)**.  **Solution-**  Dashboard - 3M  List best practices - 2M | 5M |
| 7 | You want to develop book reading habits. And are not sure which book to start with. You have some data on various books.   1. Read the CSV file from github url – [1M]   <https://raw.githubusercontent.com/AIP-codedb/public/main/data/books.csv>  \*You might see some error while reading the data from the link. That is expected as some erroneous lines are introduced in the data. You need to programmatically ignore erroneous lines from the csv file and read it in pandas dataframe.   1. Do any pre-processing on the data as needed – [1M]  * Drop un-necessary columns that might not be relevant for data-analysis * Remove Nulls and outliers if any * Remove and duplicate rows  1. Plot a visual to highlight following – [6M]  * Number of books in all the different languages * Top 10 most rated books * Top 10 authors with the greatest number of books published * Average rating distribution of all books * Top 10 books with the greatest number of pages * Top 10 highly rated authors  1. Perform any other analysis of choice and select books that you should start reading with. Give proper justification. [2M]   \*\* Add all code snippets and plotted visuals from your Analysis | 8M |