F/E/TX

Reg. No:

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of the UGC Act, 1956)

Final Assessment Test – May 2024

Course: **CSI3005 - Advanced Data Visualization Techniques**

Class NBR(s): **2470/2495** Slot: **A1+TA1**

Time: **Three Hours** Max. Marks: **100**

- **KEEPING MOBILE PHONE/ELECTRONIC DEVICES EVEN IN 'OFF' POSITION IS TREATED AS EXAM MALPRACTICE**
- **DON'T WRITE ANYTHING ON THE QUESTION PAPER**

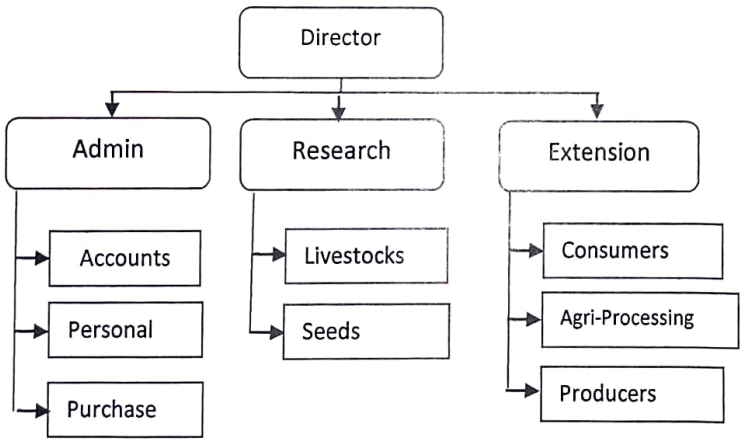
Answer any TEN Questions

(10 X 10 = 100 Marks)

1. Explore the concept of task abstraction in data visualization, specifically examining how actions and targets are represented using appropriate idioms.

2. With the onset of the southwest monsoon impacting the southern regions of India, could you elucidate the concept of divergence using appropriate idioms and illustrate its visualization through different glyphs?

3. A bureaucratic organization is directing and coordinating the Finance, Research & Development and Extension activities as illustrated in the below figure.



Visualize the organization tree structure data into various visual encoding idioms (any 5).

4. Consider the diamonds dataset *diamonds.csv* with the dimension 15044x10. The attributes are carat, cut, color, clarity, depth table, price, x, y, z axis as shown below

Carat	cut	color	clarity	depth	table	price	x	y	z
00.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
10.21	Premium	E	SI1	59.8	NA	326	3.89	3.84	2.31
20.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
30.29	VGood	I	VS2	62.4	58.0	334	4.20	4.23	2.63
40.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75
	...	...	...	...					

From the above dataset write the code to

a) Extract all the rows where the *color* is either E or J.

b) Extract all the rows where the *depth* is between 50 and 60.

c) Extract all the rows other than the *cut* 'Ideal' and 'Good'.

d) Extract the column from *cut* to *price*.

e) Calculate the average for the feature *table*.

5. A large healthcare organization tasked with creating a comprehensive dashboard for healthcare analytics. The organization aims to leverage data visualization to improve decision-making processes, optimize patient care, and enhance operational efficiency. The main role involves designing a dashboard that provides actionable insights for various stakeholders within the organization, including hospital administrators, healthcare providers, and medical staff. Using your expertise in Tableau and knowledge of healthcare analytics, outline how you would design a dashboard.

6. The dataset contains information on crime incidents in a metropolitan area over a period of one year. Each record includes the type of crime, the location (latitude and longitude) where the crime occurred, and the date/time of the incident around the globe of 10,000 records.

Crime Type	Latitude	Longitude	Date/Time
Burglary	40.7128	-74.006	01-03-2024 08:30
Assault	34.0522	-118.2437	15-05-2024 17:45
Theft	51.5074	-0.1278	10-07-2024 12:00
Robbery	-33.8688	151.2093	20-09-2024 21:00
Vandalism	39.9042	116.4074	03-11-2024 10:20
...	...	...	...

By using the given dataset on crime incidents in the metropolitan area, propose a comprehensive analysis plan to visualize crime patterns using spatial Chloropleth map and Hexagonal Binning data visualization. Describe how each technique can effectively represent the distribution of crime incidents and highlight areas of high activity or concentration. Discuss the advantages and limitations of each visualization method in identifying crime hotspots and informing policy decisions. Additionally, recommend potential interventions or strategies that could be implemented based on the insights gained from your analysis. Support your recommendations with visualizations generated from the dataset and considerations for data preprocessing and interpretation.

7. Consider a 'Advanced Data Visualization Techniques' text book with 15 chapters. The entire text book contents are available in the file '*books.dat*'. Using R/Python code

a) convert all non-tidy text to tidy text. **[2]**

b) compute total number of positive and negative words. **[2]**

c) plot top 10 positive words and negative words in descending order. **[3]**

d) use appropriate idioms to display the top 25 negative words. **[3]**

8. A titanic dataset with 708x8 as shown below in the table

Passenger Id	Survived	Pclass	Name	Sex	Age	SibSp	Parch
1	0	3	Braund, Mr. Owen	M	22	1	0
2	1	1	Cumings, Mrs. John	F	38	1	0
3	1	3	Heikkinen, Miss. Laina	F	26	0	0
4	1	2	Futrelle, Mrs. Jacques	F	35	1	0
5	0	3	Allen, Mr. William	M	35	0	0
...	...	...	...	...	...	...	...

a) Construct a mosaic plot illustrating the relationship between Pclass, Survived, and Sex. **[3]**

b) Construct a Circle Segmentation visualization representing the distribution of Pclass, Sex, Age, SilSp, Parch and Survived. **[3]**

c) Construct a Trellis Display visualization representing the relationship between Pclass, Age, Sex and Survived status, with appropriate panel arrangement and visualization elements. **[4]**

9. Elaborate on the integration process of incorporating real-time data streams into various analytical tools.

10. ABC Health Insurance Company has been facing challenges in understanding and optimizing their claims processing system. They have collected extensive data over the years related to claims, including patient demographics, diagnosis codes, treatment procedures, claim amounts, and processing times. However, they lack a comprehensive understanding of their data and struggle to identify patterns, trends, and potential areas for improvement. As a data analyst consultant, you have been hired to assist ABC Health Insurance Company in leveraging their data effectively to enhance their claims processing system and overall operational efficiency.

a) Propose a strategy for abstracting and organizing the data to facilitate analysis and decision-making.

b) Define the specific tasks or objectives that ABC Health Insurance Company aims to accomplish through data analysis and optimization of their claims processing system.

c) Choose four validation methods from the following options and explain how each method can be applied to ensure the accuracy, reliability, and validity of the findings derived from the data analysis.

d) Based on the identified data variables and objectives, recommend suitable visualization tools for effectively visualizing the data and communicating key findings.

e) Justify your choice of visualization tools based on their ability to highlight patterns, trends, and anomalies in the data relevant to claims processing optimization.

11. A *Students* dataset containing information about student performance categorized by test scores and study hours. Using a appropriate idiom, elucidate the significance of various color palettes in both univariate and bivariate visualizations.

12. The River Nile annual streamflow data for the year 1871 to 1970 sample is given below

Time Series: Start = 1871 End = 1970 Frequency = 1

[1] 1120 1160 963 1210 1160 1160 813 1230 1370 1140 995 935 ... ...

Using R/Python code to build a suitable model to predict for the year between 1971 and 1980. (Use the suitable idioms to display the output for each steps)

⇔⇔⇔**F/E/TX**⇔⇔⇔