DSE 2151 DATA ANALYTICS ANSWER SCHEME

Type: MCQ

Q1. 2. A variable called 'Grade in exam' with values coded as : A+, A, B, C ... is a _____ variable (0.5)

1. Ordinal
2. ratio
3. continuous
4. Dichotomous

Q2. 3. _____ study is conducted, when it is impossible, on either logistical or ethical grounds, to conduct a controlled experiment. (0.5)

1. An Experimental
2. An Observational
3. A double blinded
4. Current

Q3. The best fit graph for a single categorical variable is _____ (0.5)

1. Histogram
2. Box Plot
3. Scatter Plot
4. Bar chart

Q4. 10. For the following set of values : 3,4,7,2,3,7,4,2,4,7,4, The Q3 value is _____ (0.5)

1. 4
2. 7
3. 3
4. 26

Q5. _____ is a measure that quantifies the lack of symmetry in a data distribution. (0.5)

1. Kurtosis
2. Mean
3. Skewness
4. Median

Q6. A member of the Data Analytics project who has specific knowledge of the subject or business problem is called a _____ (0.5)

1. Consumer
2. IT Expert
3. Subject Matter Expert
4. Supplier

Q7. _____ is the process where an estimate is calculated for some variable that is unknown. (0.5)

1. Prediction
2. Summarization
3. Exploration
4. Association

Q8. If the frequency distribution is approximately normal, approximately 95% of all observations fall within _____ standard deviations of the mean. (0.5)

1. One
2. Two
3. Three
4. Four

Q9. What is the objective of a hypothesis test ? (0.5)

1. To make some assumptions about the population.
2. To determine if change in one variable directly causes a change in another variable.
3. To generalize our sample data to suitable situations or population.
4. To determine if change in one variable indirectly causes a change in another variable.

Q10. A variable called 'Color of Car' with values coded as : 5-Black, 4-Brown, 3-Grey is a _____ variable (0.5)

1. Continuous
2. Ordinal
3. Discrete
4. Dichotomous

Type: DES

Q11. How does a data analyst identify noisy data? What strategy can be adopted to treat the following (Mention atleast 1 strategy for each sub division):

i. noisy numeric data

To treat noisy numeric data, Strategies include

- *Ignore tuples if they are less*
- *detect noisy data using methods like IQR or Std Dev, replace with mean/median etc.*

ii. inconsistent categoric data. (2)

- *use meta data to replace with mode*

- *•Use another attribute to find nominal label*
- *•Ignore the tuple*

Scheme - Mention of atleast 1 strategy for each sub division (1 mark)

Q12. Perform the chi-square test fro the following data and provide an inference on whether the two variables (Gender and Choice of Pet) are associated to each other w.r.t to model acceptance or rejection. For degrees of freedom of 2 and confidence level of 95%, the critical chi-square value is 5.991 .

	dog	cat	bird	total
men	207	282	241	730
women	234	242	232	708
total	441	524	473	1438

The expected values table : $\frac{row\ total * column\ total}{grand\ total}$ *[0.5 marks]*

	dog	cat	bird	total
men	223.87343533	266.00834492	240.11821975	730
women	217.12656467	257.99165508	232.88178025	708
total	441	524	473	1438

Chi-square table: $\frac{(Observed\_value - Calculated\_value)^2}{Calculated\_value}$ *[0.5 marks]*

observed (o)	calculated (c)	(o-c)^2 / c
207	223.87343533	1.2717579435607573
282	266.00834492	0.9613722161954465
241	240.11821975	0.003238139990850831
234	217.12656467	1.3112758457617977
242	257.99165508	0.991245364156322
232	232.88178025	0.0033387601600580606
Total		**4.542228269825232**

$critical\ value\ of\ \chi^2\ >=\ calculated\ value\ of\ \chi^2$ *[0.5 marks]*

Defining Null Hypothesis and Alternative hypothesis [0.5 marks]

. (2)

Q13. Consider the following data set CARS:

Names	Cylinders	Displacement	Horse-power	Weight	Acceleration	Model Year	Country of Origin	MPG
Chevrolet Chevelle	8	307	130	3504	12	1970	1	18
Plymouth Duster	6	198	95	2833	15.5	1978	1	20
Chevrolet Vega (SW)	4	140	72	2408	19	1971	1	22
Fiat 124B	4	88	76	2065	14.5	1971	2	30
Datsun 1200	4	72	69	1613	18	1975	3	35
Buick Skylark 320	8	350	165	3693	11.5	1972	1	15
Ford Maverick	6	200	85	2587	16	1975	1	21
Volkswagen 1131	4	97	46	1835	20.5	1970	2	19
Toyota Corolla	4	71	65	1773	19	1973	3	31
Ford Torino	8	302	140	3449	10.5	1970	1	17

Considering the dataset describing CARS, Answer the following:

i. Create a summary table, grouping by cylinders and display count of cars, average MPG.

0.5 Mark

Cylinders	*Count of Cars*	*Avg MPG*
4	*5*	*27.4*
6	*2*	*20.5*

8	*3*	*16.67*

ii. Create a contingency table to tabulate the Country of Origin and Number of Cylinders.

0.5 mark

Country of Origin	***Number of Cylinders 4***	***Number of Cylinders 6***	***Number of Cylinders 8***	***Total***
1	*1*	*2*	*3*	***6***
2	*2*	*0*	*0*	***2***
3	*2*	*0*	*0*	***2***
Totals	***5***	***2***	***3***	***10***

iii. Find the correlation between Horse power and Weight and comment on the relationship between the variables.

iv. Visualize the relationship between Horse power and Weight using a scatter plot.

0.5 mark for calculation of Mean 94.3, 2576, Std Deviation 38.17, 77.84

1 mark For correct computation of numerator , Correlation Cofficient 0.939

0.5 mark for inference(strong positive correlation) & scatter plot

. (3)

Q14. Consider a data set with the values 250,370, 420, 605, 1100. Perform Data transformation on each of the above values with the :

i. Mean Normalization method

-0.3517, -0.2015, -0.1517, 0.06588, 0.648

ii. Min-max normalization method by setting min = 1 and max = 10

1,2.27,2.8, 4.758, 10

iii. decimal scaling method

0.025, 0.0370, 0.042, 0.0605, 0.11

1 mark each. (3)