Preview Test: Mid Semester MOCK Exam

Test Information

Description	This is a MOCK-UP of the actual exam next week.
Instructions	
Timed Test	This test has a time limit of 45 minutes.This test will save and submit automatically when the time expires. Warnings appear when **half the time**, **5 minutes**, **1 minute**, and **30 seconds** remain. *[The timer does not appear when previewing this test]*
Multiple Attempts	This test allows multiple attempts.
Force Completion	This test can be saved and resumed at any point until time has expired. The timer will continue to run if you leave the test.

QUESTION 1 **3 points** Save Answer

Provide three benefits of approaching data science problem formulation through design thinking.

This question can be answered using knowledge gained from the design thinking lecture as well as the design thinking assignment. Benefits can include but are not limited to business benefits (stakeholder alignment, purposeful analytics, viability, user adoption ...), technical benefits (engineering feasibility, cost-benefit of IT investment etc.) It is expected that any points you make are sufficiently explained in your own words.

QUESTION 2 **1 points** Save Answer

The data science process is a *sequence* of 5 steps

- ○ True
- ○ False

This is false. Although there are 5 steps according to what you have learnt in the course, these will be highly iterative and not sequential in any data science project

QUESTION 3 **1 points** Save Answer

Three characteristics of big data are:

- ○ a. Volume, Variety, and Viscosity
- ○ b. Volume, Variety, and Velocity b is correct
- ○ c. Variety, Vanity, and Volume
- ○ d. Vastness, Variability, and Value

QUESTION 4

1 points Save Answer

Which of the following statements is *not true* for relational database management system (RDBMS):

a. Separates physical and logical strucutres

b. Supports multi-user access

c. Separates data from applications

d. Mainly used for storing unstructured data

Relational databases mainly store structured data so correct answer is d

QUESTION 5

3 points Save Answer

Describe in detail how to take a stratified random sample of total size 12 from a set of 120 observations with three strata (20 observations in the first stratum, 30 observations in the second stratum and 70 in the other stratum).

We keep the proportions of the strata the same: numbers of items to sample from each strata are 12*20/120 = 2, 12*30/120 = 3, 12*70/120 = 7. Then use simple random sampling WR to sample 2, 3, and 7 observations from 1st, 2nd and 3rd strata respectively.

QUESTION 6

3 points Save Answer

The following data has 2-anonymity with respect to the attributes 'Age', 'Gender', and 'State'. Explain how this data is still vulnerable. Suggest one alternative technique for data anonymization.

Age	Gender	State	Disease
[20 – 30]	F	NSW	Cancer
[20 – 30]	F	QLD	Viral Infection
[20 – 30]	F	NSW	TB
[10 – 20]	M	VIC	No illness
[10 – 20]	F	QLD	Heart
[10 – 20]	M	VIC	TB
[20 – 30]	F	QLD	Cancer
[10 – 20]	F	QLD	Heart

[10-20], F, QLD has the same sensitive attribute ie 'Heart'. An adversary dataset could hence identify an individual's sensitive attribute causing an anonymisation failure.

Due to these limitations of k-anonymity, new techniques such as i-diversity and differential privacy have been proposed to overcome these problems

QUESTION 7

3 points Save Answer

With the help of an example, explain the difference between cleaning from rules and cleaning from filter.

It is expected that the examples you provide clearly indicate how the error can be fixed with a rule e.g. inconsistent units of measurement for rules. This is in contrast to when a reliable reference source is needed to fix an error e.g. date of birth of a person

QUESTION 8

3 points Save Answer

Describe how deterministic regression imputation works and describe two problems arise from using deterministic imputation?

Deterministic regression imputation fits a regression model of the missing value against non-missing values and then uses it to predict missing values.

Problems: it can strengthen observed relationships between predictors and response; it can reduce variability in data and estimates.

QUESTION 9

4 points Save Answer

Match the description to one of the following:

a. Supervised learning

b. Density estimation

c. Clustering

b - Given a sample drawn from a Gaussian distribution, determine the parameters of the Gaussian distribution.

a - Given a collection of emails labelled as SPAM or NOT SPAM, build a model to check whether an email is a spam or not.

a - Given a dataset containing the education level, gender, age, and salary for a group of people, build a model to predict the salary of a person given the person's education level, gender, and age.

c - Given a collection of articles about science and sports, divide them into groups about these respective topics.

a. a. Supervised learning
b. Density estimation
c. Clustering

b. a. Supervised learning
b. Density estimation
c. Clustering

c. a. Supervised learning
b. Density estimation
c. Clustering

d. a. Supervised learning
b. Density estimation
c. Clustering

QUESTION 10

2 points Save Answer

For a linear regression model, what are the assumptions on the residuals? Why are these assumptions made?

Assumptions: the residuals are identically and independently drawn from a normal distribution with mean 0. You can also use the terms zero mean, constant variance, normality, and independence to describe the assumptions.

These assumptions are needed so that the true relationship between the predictors and response can be accurately estimated when there is enough data.

QUESTION 11

2 points Save Answer

Recall the importance of understanding the context and consider the following visual. Does this visual persent an effective way of understanding the context ((Y/N)? If so ...

Who is the expected audience?

What action is expected from the audience?

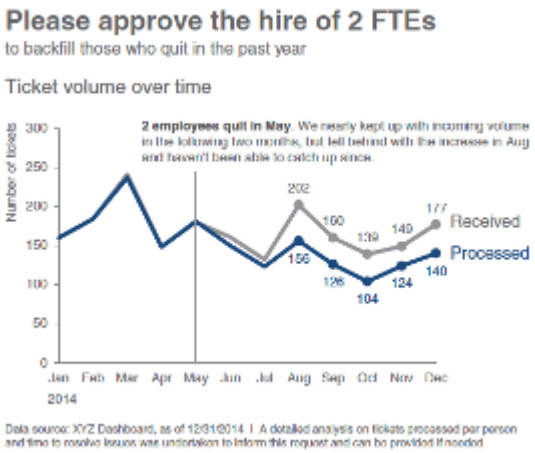
Refer to the summary slides for this module where we discussed the importance of understanding context. This is a good example as the expected audience is clear ie public/ citizens and the expected action is clear too ie cut social interactions/ stay home

QUESTION 12

3 points Save Answer

Recall Gestalt's principles of visual perception: Proximity, Similarity, Enclosure, Closure, Continuity, and Connectedness.

Which of these are evident in the visual below? Briefly explain your answer.



Several principles are evident here e.g. connectedness of the two measurements. It is expected that you provide an explanation of where and how the principle has been applied and why it is working well in this visual

QUESTION 13

1 points Save Answer

Assume that you have access to sale data including: property address (street, suburb, state), property features such as no. of bedrooms, lot size, car spaces etc; sale history (date and price). Describe how a data story using *Drill Down* can help understand the sales data.

This can be provided through location state - suburb - street or time year - month. Make sure you also explain why the drill down example you provide will help to understand sales data for the intended audience