1) You are asked to plan a data analytics project to *analyse student feedback to DCU in relation to online teaching in 2020 and 2021*. Using the Generic Data Analytics Pipeline discussed in CA682, assign each of the following activities to one of the 5 main categories: Gathering, Processing, Analysing, Presenting and Preserving.

   Liaising with DCU Registry to get datasets from the student registration and results systems. → Gathering,

   Creating a document to share with senior university management summarising the findings. → Presenting,

   Anonymising student comments that include identifying details. → Processing,

   Removing incorrect entries from the student datasets. → Processing,

   Converting student words into sentiment ratings and correlating with field of study. → Analysing,

   Calculating the average satisfaction levels based on the sentiment ratings. → Analysing,

   Documenting the data formats used in the study and saving all of the created datasets. → Preserving,

   Conducting student surveys to answer the key questions about their experience. → Gathering

2) According to the three classical definitions of big data, which of the following datasets is *most likely* to be classified as "big data"?

a. The "Titanic" dataset showing passenger details from the final voyage of the ship.

b. Records from Spotify of the tracks listened to by each user (est. 232M users).

c. Sales records from the DCU merchandise store.

The correct answer is: Records from Spotify of the tracks listened to by each user (est. 232M users).

3) For data that records the "Type of pet (e.g., cat, dog, bird, fish)" choose *all* of the following descriptions that can apply. Marks will be deducted for including wrong choices.

The correct answers are: Qualitative, Nominal

4) For data that records the "Number of pets currently owned" choose *all* of the following descriptions that can apply. Marks will be deducted for including wrong choices.

The correct answers are: Quantitative, Discrete, Ratio

5) For data that records the "Weight of pets (in grams)" choose *all* of the following descriptions that can apply. Marks will be deducted for including wrong choices.

The correct answers are: Quantitative, Continuous, Ratio

6) For data that records the "Happiness of pet owners (self-rated from

1 to 5)" choose *all* of the following descriptions that can apply. Marks will be deducted for including wrong choices.

The correct answers are: Qualitative, Ordinal

7) Which of the following is *not* a valid description of metadata?

Question

a. Metadata is an inferior form of cataloguing.

b. Metadata is data about data.

c. Metadata is information on the organisation of the data, data domains and the relationship between them.

d. Metadata is created by humans and is often incorrect.

The correct answer is: Metadata is created by humans and is often incorrect.

8) Which of the following statements are correct in relation to open data?

Question 8Answer

a. Open data is allowed to contain personal information.

b. Open data may help make governments and corporations more transparent.

c. Open data can be used commercially.

d. Open data is only provided by governments.

The correct answers are: Open data can be used commercially., Open data may help make governments and corporations more transparent.

9) Which of the following statements are correct in relation to data cleaning and data quality metrics?

a. It is possible to perfectly and absolutely measure quality of a dataset to compare performance.

b. Many data quality metrics (accuracy, completeness) are unmeasurable.

c. A good measure of data quality is accuracy and completeness.

d. Data quality metrics can be used for contracts for service delivery.

The correct answers are: Data quality metrics can be used for contracts for service delivery., Many data quality metrics (accuracy, completeness) are unmeasurable.

10) Match the error with the *most likely* phase of the generic data analytics pipeline where it was introduced.

The correct answer is:

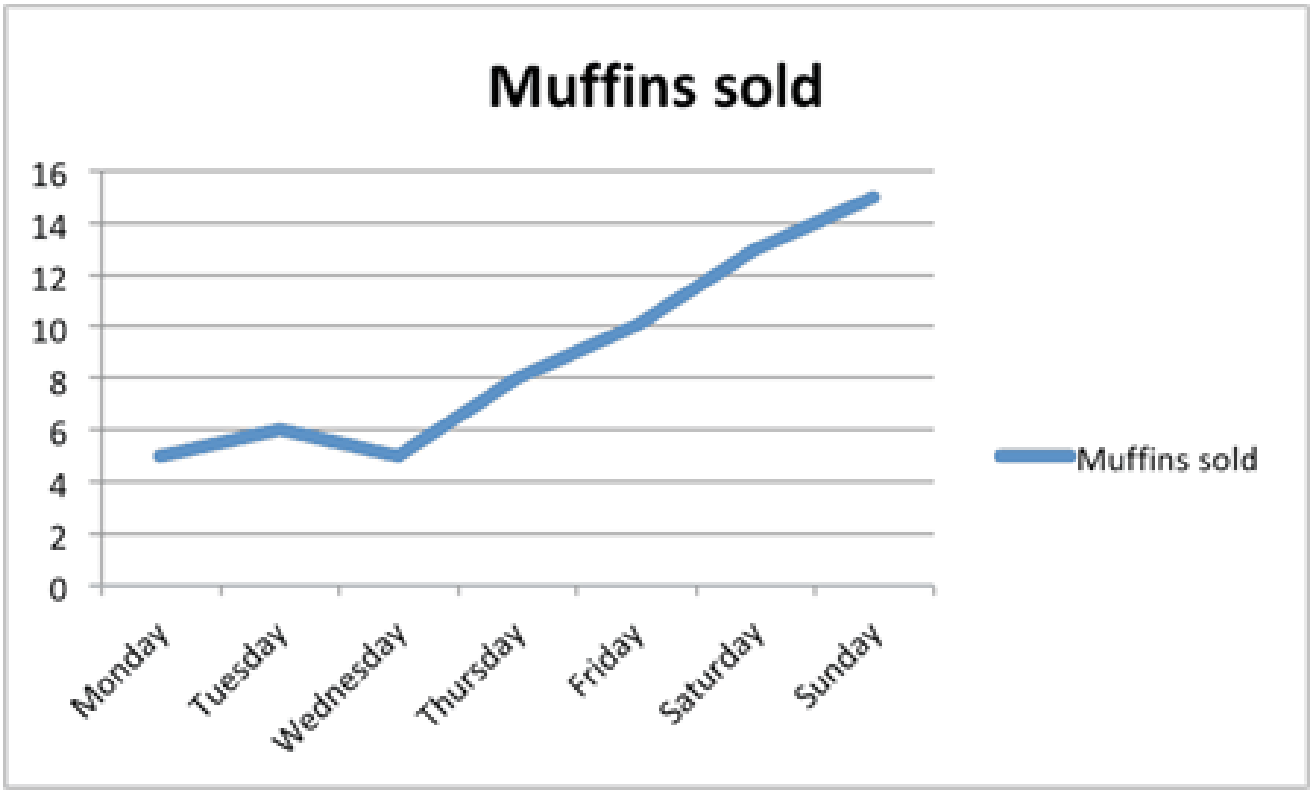
Time synchronization errors resulting in missing values. → Processing,

Unnecessary precision of generated numerical data. → Analysing,

No documentation provided on format for missing values (“”, Nan, -999) → Preserving,

Data delivery issues such as transmission problems that may result in loss of network connectivity, buffer overflows or corruption. → Gathering

11) The graph shown below is a ...



The correct answer is:

line chart

12) In the graph, how many muffins were sold on Thursday?

The correct answer is: 8

13) Identify the major marks and attributes used to encode data in the graph below.

The correct answers are: line, position, slope

14) Match the scenario with the *most appropriate* choice of chart to visualise the given data.

Distribution of grades in CA682 over the past 5 years. → Box plot,

Show the improvement in sales (total profit in €) over each of the past 5 years for your product compared to your competitors. → Line chart,

Understand the relationship between maximum daily temperature (^{0}C) and average daily personal water consumption (Litres) in Ireland. → Scatterplot,

The most popular method of travel to DCU during 2019. → Bar chart