1. What are the key tasks that machine learning entails? What does data pre-processing imply?

Machine learning involves tasks like data collection, data pre-processing, model selection, training, evaluation, and deployment. Data pre-processing involves cleaning, transforming, and encoding data to ensure it is in the correct format and free of errors. This step is crucial for enhancing the quality and accuracy of the machine learning models.

2. Describe quantitative and qualitative data in depth. Make a distinction between the two.

Quantitative data is numerical and can be measured, like height or age. Qualitative data is descriptive and categorical, such as colors or names. The distinction lies in that quantitative data supports mathematical operations, while qualitative data is used to categorize and describe characteristics without numerical values.

3. Create a basic data collection that includes some sample records. Have at least one attribute from each of the machine learning data types.

| ID | Name | Age | Height (cm) | Gender | Purchase Amount ($) |

|----|--------|-----|-------------|--------|---------------------|

| 1 | Alice | 30 | 165 | Female | 250.75 |

| 2 | Bob | 22 | 175 | Male | 150.50 |

| 3 | Charlie| 28 | 180 | Male | 300.20 |

```

4. What are the various causes of machine learning data issues? What are the ramifications?

Causes of data issues include missing values, noise, outliers, and incorrect data entries. Ramifications include reduced model accuracy, biased predictions, and unreliable insights. Addressing these issues is crucial for maintaining the integrity and reliability of machine learning models.

5. Demonstrate various approaches to categorical data exploration with appropriate examples.

Approaches to categorical data exploration include frequency distribution tables, bar charts, and pie charts. For example, using a bar chart to visualize the distribution of a "Gender" attribute in a dataset can reveal the proportion of male and female entries, aiding in understanding the dataset composition.

6. How would the learning activity be affected if certain variables have missing values? Having said that, what can be done about it?

Missing values can lead to biased models and reduced accuracy. To address this, techniques like imputation (filling missing values with mean, median, or mode), using algorithms that handle missing values, or removing affected records can be applied to ensure the dataset's completeness and integrity.

7. Describe the various methods for dealing with missing data values in depth.

Methods for handling missing data include:

- Deletion: Removing records with missing values.

- Imputation: Replacing missing values with statistical measures like mean, median, or mode.

- Predictive Modeling: Using algorithms to predict and fill in missing values.

- Using Algorithms that Handle Missing Data: Employing models that can manage missing values directly.

8. What are the various data pre-processing techniques? Explain dimensionality reduction and function selection in a few words.

Data pre-processing techniques include normalization, standardization, encoding categorical variables, handling missing values, and outlier detection. Dimensionality reduction reduces the number of features while retaining essential information, and feature selection involves choosing relevant features that contribute most to the predictive power of the model.

9. i. What is the IQR? What criteria are used to assess it?

The Interquartile Range (IQR) measures statistical dispersion and is calculated as the difference between the third quartile (Q3) and the first quartile (Q1). It assesses data spread and identifies outliers by highlighting the middle 50% of data points.

ii. Describe the various components of a box plot in detail? When will the lower whisker surpass the upper whisker in length? How can box plots be used to identify outliers?

A box plot includes the minimum, first quartile (Q1), median, third quartile (Q3), and maximum. Whiskers extend from Q1 to the lowest value within 1.5 IQR of Q1, and from Q3 to the highest value within 1.5 IQR of Q3. The lower whisker surpasses the upper whisker in length when the lower half of the data is more dispersed. Outliers are points outside the whiskers.

10. Make brief notes on any two of the following:

1. Data collected at regular intervals

Regular interval data, or time-series data, is gathered consistently over time (e.g., daily temperature readings). It's crucial for identifying trends, seasonal patterns, and making forecasts.

2. The gap between the quartiles

The gap between quartiles, or IQR, measures data variability. It highlights the range within which the central 50% of data points lie, offering insights into data spread and potential outliers.

11. Make a comparison between:

1. Data with nominal and ordinal values

Nominal data categorizes without order (e.g., colors), while ordinal data categorizes with a defined order (e.g., rankings). Nominal values can't be ranked, whereas ordinal values have a meaningful sequence.

2. Histogram and box plot

Histograms display data distribution over intervals, showing frequency of data within each bin. Box plots summarize data distribution with quartiles, highlighting median, range, and outliers. Histograms are better for distribution shape, while box plots excel in identifying outliers and variability.

3. The average and median

The average (mean) is the sum of all values divided by their count, sensitive to outliers. The median is the middle value in a sorted dataset, less affected by outliers. Median provides a better central tendency measure in skewed distributions.