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

A: Data collection, data preparation, model training, evaluation, model tuning, deployment. Data preprocessing includes data cleaning, data splitting, standardization and any other data operations that happen before training a model.

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

A: Quantitative data uses numbers, counts as information while qualitative data often uses languages to convey information.

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

A:

Name	Age	Gender	Date of Birth	Rating
Newton	20	Male	20-08-2002	2
Susan	30	Female	21- 09-1992	5

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

A: There are a variety of issues that are caused by human errors, sometimes collected data is not enough. In both cases the model loses generalisation.

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

A: Checking distinct values is need to see if the same values have different cases or spellings. Eg: Sometimes both Male and M are used in the same column

Bar graph can be used to visualise the frequency of each category

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

A: Absence of data generally reduces the predictability of a model. They can be treated in many ways including deletion, updation and so on

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

A: Deletion of rows: When the number of missing values is very low, rows can be deleted

Deletion of columns: When there are a lot of missing values in a given column, the column won't be useful for analysis. So, deletion is an option

Updation of missing values: Sometimes missing values can be updated based on other columns. Eg: Date of week can be updated using date or age can be derived using date of birth. Sometimes the mean or mode used to replace null values.

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

A: There are a lot of data pre-processing techniques depending on the data and the decided algorithm. Techniques like data cleaning, data standardization are done for every ML problem while techniques like feature selection, dimensionality reduction are done depending on the number of features.

Dimensionality reduction: It is a process through which the number of variables is reduced in such a way that relations between these variables are stored.

Feature Selection: It is a process through which useful variables are selected so that the model gives the best results

9.

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

A: IQR is interquartile range. It is the difference between 75th percentile and 25th percentile. Higher the IQR, the more spread the data is.

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: Median and quartiles are represented by the box. Whiskers represent the minimum and maximum values excluding outliers. Outliers are represented by dots.

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

1. Data collected at regular intervals

A: Data collected only after a defined interval in a few cases. For example: Integers

2. The gap between the quartiles

A: Quartiles are 0th, 25th, 75th, 100th percentiles respectively. Gap between them represents spread on data

3. Use a cross-tab

1. Make a comparison between:

1. Data with nominal and ordinal values

A: Nominal values do not have order while ordinal values have order.

2. Histogram and box plot

A: Histogram shows frequency of numeric values while box plots show the percentiles.

3. The average and median