1. What are the key tasks involved in getting ready to work with machine learning modelling?

Data ingestion, EDA, pre-processing and preparing data, model selection, Train the model, model evaluation, hyperparameter tuning

2. What are the different forms of data used in machine learning? Give a specific example for each of them.

Categorical data such as gender, education

Ordinal such as education level and Nominal such as gender

Numerical data such as age, temperature

This is divided to continuous such as weight and discrete such as zip code

3. Distinguish:

1. Numeric vs. categorical attributes

Numerical value can be known as a number and they are qualitative features in form of continuous or discrete number while categorical data can be known by their name and they are quantitative features.

2. Feature selection vs. dimensionality reduction

Feature selection return a subset of features from the original set of features, which are best representatives of the data which help to improve performance and reduce computational process while dimensionality reduction is generic and only depends on the data and not on what you plan to do with it and just projects the data into a lower dimensionality space, this may increase the performance or not because

4. Make quick notes on any two of the following:

1. The histogram

A histogram is a method of grouping data into logical range which calls bins and sum number of evidences in each group or bin. It has a graphical representation which is similar to bar chart.

2. Use a scatter plot

Scatter plots presents the relationship between two variables in a two-dimensional plot. The independent variable is plotted on the X-axis and the dependent variable is plotted on the Y-axis.

3.PCA (Personal Computer Aid)

5. Why is it necessary to investigate data? Is there a discrepancy in how qualitative and quantitative data are explored?

By investigating data, we can understand type of data and based on this we can plan for filling missing values or cleaning data, also for transformation we need to know data and whether transformation needed or not. also, by plotting and analysing data we can find distribution and relation between them which facilitate the pre-processing and model selection phase.

Always, We cannot distinguish between qualitative and quantitative data based on their shape but qualitative data is descriptive while quantitative is numerical then data investigation is mandatory to recognize them.

6. What are the various histogram shapes? What exactly are ‘bins’?

Bell shape for normal distribution of data, right skewed and left skewed, uniform distribution and bimodal distribution that can be symmetric or non-symmetric.

Bins is number of equal intervals that we plan to divide our data.

7. How do we deal with data outliers?

It is depended on variables if outliers are in target variables, we have to release them but if they are in other variables, we can handle them in different way included:

- replacing them with other fix values such as mean or median or any fix value

- replacing them with the help of quantiles of data

- select the model which are not sensitive to outliers

- scaling data

8. What are the various central inclination measures? Why does mean vary too much from median in certain data sets?

There are three main central tendency: mean, median and mode

If there are outliers in data then the mean and median will be far from each other. In other word if the variance of data is high then the mean and median will be far from each other

9. Describe how a scatter plot can be used to investigate bivariate relationships. Is it possible to find outliers using a scatter plot?

The scatter plot shows the relation between two variables, the value of one variable come in axis x and the value of other variable come in axis y and present a point. If by increasing value of x axis the value of y axis also increases the variables have positive linear relation and vs.

If a point be far from other points it means that is an outlier.

10. Describe how cross-tabs can be used to figure out how two variables are related.

Cross tab is used to compute a simple cross-tabulation of two or more variables by computes a frequency table of the variables. This relation can be calculated in different way in python which need to specify the aggregation function.