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

The various steps in preparing data ready to work in machine learning modelling are :-

**Feature Engineering**

Identifying numerical and categorical features

Handling null values

Handling imbalanced datasets

Outliers handling

Scaling the data{ Standardisation & Normalisation }

Converting the categorical features into numerical features

**Feature Selection**

PCA

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

The various forms of data are:-

1) Numerical – These types of data can be measured and can be represented in the form of a number. These are further divided into discrete and continuous.

2) Categorical – These type of data are divided in categories . They do not have numerical relationship between them i.e. they cannot be put in an order based on their values. For example, Asia, Africa, Antartica etc. are the continents and they can easily be categorised. But we can’t put it in an order as it don’t have a numerical relationship.

3) Ordinal - Ordinal data is a kind of categorical data with a set order or scale to it. These data have numerical relationship among them. For example, ordinal data is said to have been collected when a responder inputs his/her financial happiness level on a scale of 1-10.

**3. Distinguish:**

**1. Numeric vs. categorical attributes**

Numerical attributes are represented in numbers and can be put in a logical order

Categorical attributes are represented in the form of categories and can’t be put in a logical order.

**2. Feature selection vs. dimensionality reduction** –

Feature selection involves selecting only those features that are relevant to our usecases out of all the given features. Some of the information gets lost because of their dropping.

Dimensionality reduction transforms feature into lower dimension. The whole of the information is representd by a smaller number of features and remains retained. PCA (Principal Component Analysis) is a popular example for it.

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

**1. The histogram**

Histogram represents the frequency distribution of data along given class intervals. These class intervals are also known as bins. Histograms are commonly used in [statistics](https://www.investopedia.com/terms/s/statistics.asp) to demonstrate how many of a certain type of variable occurs within a specific range. For example, a census focused on the [demography](https://www.investopedia.com/terms/d/demographics.asp) of a country may use a histogram to show how many people are between the ages of 0 - 10, 11 - 20, 21 - 30, 31 - 40, 41 - 50, etc.

2. Use a scatter plot

**3.PCA (Principal Component Analysis)**

PCA is a dimensionality reduction technique. In this, The higher dimensions are represented by squeezing them into lower dimesions. No information is lost as no features are dropped.

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

The data needs to be cleaned and relevant features must be selected in order to get a predictive and accurate model. Analysing the data helps us in creating a meaningful data visualisation and predict future trends from the data.

Yes, there are differences in how qualitative and quantitative data to be explored. Qualitative data research uses experiments, observations & surveys with close ended questions. While Qualitative data research adopts interview with open ended questions, observations described in words and literature reviews concepts & theories.

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

Various shapes of histogram are:-

1) Bell Shaped :- This kind of distribution contain one single peak in the middle of distribution. The popular example of this is Normal Distribution.

2) Uniform :- This distribution have same number of values in dataset occurring in each interval of distribution.

3) Bimodal :- This type of distribution have two distinct peaks.

4) Multimodal :- This type of distribution have more than two distinct peaks.

5) Left skewed :- This type of shape have a tail on the left side of distribution. This is also called as negatively skewed.

6) Right skewed :- Shapes have a tail tilted on the right side of the distribution. Also known as positively skewed.

7) Random :- We don’t have a clear pattern for this type of distribution.

We divide the entire range of values present in the dataset into a series of intervals. These intervals are called as Bins.

**7. How do we deal with data outliers?**

Data outliers can be treated in a number of ways mentioned below

1) Trimming/Removing the outliers :- In this method, Outliers are removed. Although it is not considered a good approach.

2) Quantile based flooring & Capping :- In this technique, outliers are capped at a value above the 90th percentile & floored at below 10th percentile. The data points above 90th percentile value are replaced by 90th percentile value.

3) Mean/Median Imputation :- As mean value is highly influenced by the outliers, it is always advised to replace the outliers with median value.

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

Various central inclination measures or measure of central tendency features are:-

1) Mean

2) Median

3) Mode

As mean is summation of the values divided by the number of values, it gets easily impacted by the outliers. Whereas median is taking the central term as its values. Hence sometimes mean vary too much from median in certain data sets

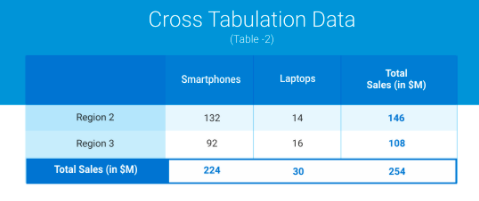
**9. Describe how a scatter plot can be used to investigate bivariate relationships. Is it possible to find outliers using a scatter plot?**

A scatterplot is a type of data display that shows the relationship between two numerical variables. One of the variable is plotted on the x-axis while second variable is plotted on y-axis. Now each dataset is plotted in x-y axis as a point.

Outliers can easily be identified as it lies far away from thick distribution of datasets.

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

A cross tab is a type of table in matrix format used for reflecting the frequency distribution of the variable. Consider an example



This table basically shows the frequency distribution by plotting gadgets as columns and region as rows. The total number of sales in accordance to the region are reflected accordingly. Some observations may be drawn using this like laptop sales are almost equal in both the regions. Accordingly market strategy may be drawn.