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

Ans: Following are the tasks involved:

1.Data Collection

2.Data Preprocessing

3. Feature Engineering

4.Data Normalization and scaling

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

Ans: Following are the forms of data used in machine learning:

1.Numerical Data:1] data points are exact numbers

2] data is information of some measurements,recordings etc

3] data can be continuous or discrete

4] Discrete data ex: no of students,no of employees etc

5] Continuous data ex:height,weight of students

2. Categorical Data:1] It represents characteristics

2] It can be numerical data

3] It is a class label in case of classification

4]ex: stage of cancer is first,second or third

3.Time Series Data:1] It is a sequence of numbers collected at regular intervals

2] It is useful in the field of finance

3]ex: Temperature recorded for past one month

4. Text :1] These are words they are converted into numbers using some functions

3. Distinguish:

1. Numeric vs. categorical attributes

**Ans:** 1. Numerical Data:1] data points are exact numbers

2] data is information of some measurements,recordings etc

3] data can be continuous or discrete

4] Discrete data ex: no of students,no of employees etc

5] Continuous data ex:height,weight of students

2. Categorical Data:1] It represents characteristics

2] It can be numerical data

3] It is a class label in case of classification

4]ex: stage of cancer is first,second or third

1. Feature selection vs. dimensionality reduction

**Ans:** 1] In feature selection we select a subset of the original feature set.

2] We do not perform any manipulation of the data

3] In dimensionality reduction we choose a new representation within which we can describe most but not all of the variance within your data.

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

1. The histogram: It represents distribution of numerical data

It organizes a group of data points which are called bins into user-specified ranges which are decided by observing the data.

Frequencies are shown by rectangles

1. Use a scatter plot: It pairs of numerical data, with one variable on each axis, to look for a relationship between them.

If the variables are correlated, the points will fall along a line or curve.

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?

**Ans:** Data investigation is necessary for following reasons:

1. Data comes from various sources, so it is in different formats
2. It contains duplicates, null values
3. It contains unnecessary columns which are not important for our analysis

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

**Ans:**1.Normal distribution

2.Skewed distribution

3. Double-Peaked or Bimodal

4.Multimodal distribution

5.Edge peak distribution

6. Comb distribution

7.Truncated distribution

**Bins:** The data is separated by using some interval. Each group of data is called a bin.The number of values in each bin are represented by bars on the graph

7. How do we deal with data outliers?

**Ans**: There are following methods of dealing with outliers:

1.Univariate method: In this method the extreme values in one variable are identified.

Using box plot we can detect the outlies.

2.Multivariate method: Here we look for relationships or combinations of variables which are unusual. We can use neural networks, linear regression in this method

### 3. Minkowski error: it reduces the impact that outliers will have in the model.

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

**Ans:** Mean, median, mode are three measures of central tendency of data.

1.The mean is normally used for [normal distribution](https://www.wallstreetmojo.com/normal-distribution/)s.The median is used for the skewed distributions data set.

2.The mean is not robust. Median is robust.

3.Median is not affected by strong outlier values.

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

**Ans:**1.In scatter plot x-axis represents the value of predictor variable and y-axis represents the scores of the predicted variable.

2.We can draw a straight line in between the points.

3. The distance of the points to the line is called "scatter".

4.A large amount of scatter around the straight line indicates a weak relationship between the variables. Small scatter represents a strong relationship.

5.If all points fall directly on a straight line, we have a perfect linear relationship between our two variables.

Yes, it is possible to find outliers using the scatter plot. The points which are far from the straight line are considered as outliers.

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