10. What is the difference between a boxplot and histogram?

Histograms are a special kind of bar graph that shows a bar for a range of data values instead of a single value. A box plot is a data display that draws a box over a number line to show the interquartile range of the data. The 'whiskers' of a box plot show the least and greatest values in the data set.

11. How to select metrics?

- Classification. This algorithm will predict data type from defined data arrays. For example, it may respond with yes/no/not sure.
- Regression. The algorithm will predict some values. For example, weather forecast for tomorrow.
- Ranking. The model will predict an order of items. For example, we have a student group and need to rank all the students depending on their height from the tallest to the shortest.

12. How do you assess the statistical significance of an insight?

Steps in Testing for Statistical Significance

1. State the Research Hypothesis.
2. State the Null Hypothesis.
3. Select a probability of error level (alpha level)
4. Select and compute the test for statistical significance.
5. Interpret the results.

13. Give examples of data that doesnot have a Gaussian distribution, nor log-normal.

Many random variables have distributions that are asymptotically Gaussian but may be significantly non-Gaussian for small numbers. For example the Poisson Distribution, which describes (among other things) the number of unlikely events occurring after providing a sufficient opportunity for a few events to occur. It is pretty non-Gaussian unless the mean number of events is very large. The mathematical form of the distribution is still Poisson, but a histogram of the number of events after many trials with a large average number of events eventually looks fairly Gaussian.

14. What is the Likelihood?

The likelihood is **the probability that a particular outcome is observed when the true value of the parameter is , equivalent to the probability mass on** ; it is not a probability density over the parameter . The likelihood, , should not be confused with , which is the posterior probability of given the data .