SQL Worksheet Answers

1. d) All of the mentioned
2. a) Discrete
3. a) pdf
4. c) mean
5. c) empirical mean
6. a) variance
7. c) 0 and 1
8. b) bootstrap
9. a) frequency

10. A boxplot and a histogram are both used to visually represent data, but they have different purposes. A histogram is a graphical representation of the distribution of a dataset, showing the number of observations that fall within each given range of values (bins). It is used to show the distribution of continuous data. On the other hand, a boxplot is a standardized way of displaying the distribution of data based on five number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum"). It is used to show the distribution of continuous or discrete data.

11. To select metrics, it is important to consider the goals and objectives of the analysis, as well as the type and structure of the data. One should start by identifying the key variables and factors of interest, and then determining which metrics will best measure and describe these variables. Additionally, it is important to consider the level of measurement (nominal, ordinal, interval, ratio) and the type of data (continuous, categorical, count) when selecting metrics.

12. To assess the statistical significance of an insight, one can use statistical hypothesis testing, specifically p-value. The p-value is a measure of the evidence against a null hypothesis. A small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, so you reject the null hypothesis and conclude that your sample provides enough evidence that the population parameter differs from the value specified in the null hypothesis.

13. Examples of data that doesn't have a Gaussian distribution, nor log-normal are: the data following a logistic distribution, the data following a Poisson distribution, the data following a binomial distribution.

14. An example where median is a better measure than mean is when there are outliers present in the data. Outliers are extreme values that deviate significantly from the rest of the observations, and they can greatly affect the mean, but not the median. Therefore, when outliers are present, the median gives a better representation of the central tendency of the data.

15. Likelihood is a function that describes the probability of obtaining a certain set of data given a set of parameters. It is used in statistical modeling to estimate the probability of a model given a set of observations. The likelihood function is used to update the estimates of the parameters in the model through an optimization process, such as maximum likelihood estimation (MLE).