WORKSHEET 6 STATISTICS

1) D
2) A
3) A
4) C
5) A
6) D
7) C
8) B
9) B
10) Histograms indicate the whole frequency distribution of a variable, whereas the boxplot summarizes its most prominent features. histograms are better in determining the underlying distribution of the data, box plots allow you to compare multiple data sets better than histograms as they are less detailed and take up less space.

11) Based on prerequisites, we need to understand what kind of problems we are trying to solve.

In Classification, algorithm will predict data type from defined data arrays. For example, it may respond with yes/no/not sure. So we use confusion accuracy, recall, precision, f-1 score metrics.

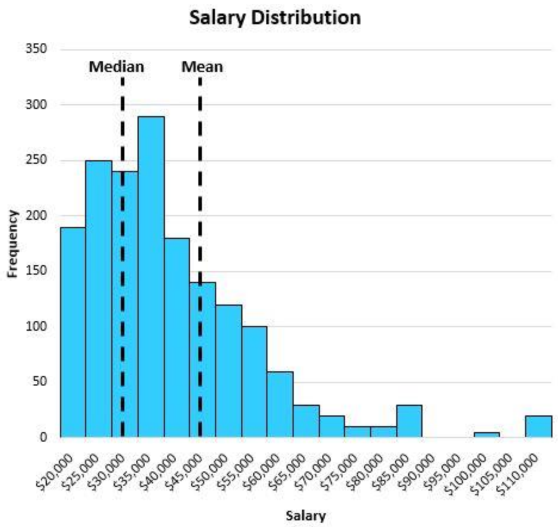
In Regression, the algorithm will predict some values. For example, weather forecast for tomorrow. In this case we use, MAE (Mean Absolute Error), MSE (Mean Squared Error) , RMSE (Root Mean Squared Error),

12) To assess statistical significance, you would use hypothesis testing. The null hypothesis and alternate hypothesis would be stated first. Second, you'd calculate the p-value, which is the likelihood of getting the test's observed findings if the null hypothesis is true. Finally, you would select the threshold of significance (alpha) and reject the null hypothesis if the p-value

is smaller than the alpha — in other words, the result is statistically significant.

13) Exponential distributions do not have a log-normal distribution or a Gaussian distribution. In fact, any type of data that is categorical will not have these distributions as well. Example: Duration of a phone car, time until the next earthquake, etc.

14) It is best to use the median when the distribution is either skewed or there are outliers present.



The median does a better job of capturing the “typical” salary of a resident than the mean. This is because the large values on the tail end of the distribution tend to pull the mean away from the center and towards the long tail.

In this example, the mean tells us that the typical individual earns about $47,000 per year while the median tells us that the typical individual only earns about $32,000 per year, which is much more representative of the typical individual.

15) Likelihood indicates how likely a particular population is to produce an observed sample. Eg: Suppose you have an unbiased coin. If you flip the coin, the probability of getting head and a tail is equal, which is 0.5 Now suppose the same coin is tossed 50 times, and it shows heads only 14 times. You would assume that the likelihood of the unbiased coin is very low. If the coin were fair, it would have shown heads and tails the same number of times.