Statistics Worksheet1

1) A
2) A
3) B
4) D
5) C
6) B
7) B
8) A
9) C

10) Normal Distribution in graphical representation is a bell shape curve that implies the mean, median and mode to be converging at the same point. If we see any kind of skewness, be it a left or a right one then the data at hand consists of outliers which can affect the outcome of a model built on such a data. There are various ways to make skewed data into a normal distribution data but there are times when an approximate bell curve is formed instead of a perfect one on real time data. The peak of a normal distribution consists of the maximum data points while the bottom part reduces in terms of frequency of data.

11) To handle missing data first I will use the is null to identify how many null values are present in a table. Then I will check the percentage of missing data because if more than 50 percent of the data in particular row is missing then we will need to accumulate proper information for that particular column else deleting the entire column would be logical instead of treating it manually and giving incorrect data for the model to be trained and tested upon. If there are very few data missing then depending on the data can use mean, median and mode options to fill the correct data. The imputation techniques that I will be using are mean imputation, simple imputer, iterative imputer and knee imputer.

12) A/B testing is mainly used when trying out a new feature on an existing product. It is similar to the concept of main and branch used in GitHub where we create branch for new feature changes and then merge them into the main section if things go well or keep the main untouched. In A/B testing we create sample of an entire population and then use the Hypothesis testing mechanism to check if our Null Hypothesis is correct or our Alternative Hypothesis is right. We need to check where we are able to reject the Null Hypothesis or whether we fail to reject the Null Hypothesis keeping in the mind the Type 1 and Type 2 errors that can be checked via the confusion matrix. In Null Hypothesis the important values considered are the alpha (allowable percentage of error), p value and the confidence level of the test model.

13) I feel mean imputation of missing data is not accurate even if it is popular and commonly used technique the logic behind filling missing data with the mean of all the other observations of a column does not always provide the correct information plus increases the number of same data in a single column making the model biased towards the usage of the

same mean value over other legitimate observations secured from proper data collection channels. There are other better imputation methods which provide accurate data for filling the missing value. However, if every other technique fails then mean imputation can be used as a last resort instead of deleting null values especially for a smaller data set.

14) Linear Regression is a supervised machine learning algorithm that can be used to predict continuous data. The underlying equation used by linear regression model is $y=mx+c$ where m is the slope and c is the intercept of the best fit line. There are 2 types of linear regression and they are Simple linear regression and Multiple linear regression techniques. Linear Regression is majorly used to predict the labels with the help of one or more feature variables ensuring that the features selected in the equation provide the necessary input to obtained the desired labels without creating an overfitting or underfitting model that provides a reasonable accuracy on future predictions.

15) The 2 main branches of Statistics are Descriptive statistics and Inferential statistics. Descriptive statistics as the name suggests focuses more on the description of how the data is collected (via primary or secondary sources) and how it is presented to get meaningful insights from it without any biases. Inferential statistics on the other hand improvises the descriptive statistics information to extract/draw conclusions relevant to the problem or request at hand. Most of these experiments are performed on sample variables that are a part of a huge population to build a proper result without wasting much of the computational resources on the entire data set