Statistics Worksheet-1

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

10. Normal Distribution: Normal distribution is a probability distribution symmetric at the mean, which shows frequent data at the mean position then the data away from the mean position.

11. Handling Missing Data:
    Data Cleaning is a process of finding and correcting the incorrect/null values that are present in the data. We can use the following methods for removing the incorrect values.
    a. Delete the column with incorrect values.
    b. Delete the row with incorrect values.
    c. Fill the incorrect data with the values

12. A/B Testing: A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment. It includes two randomized experiments includes "statistical hypothesis testing" or "two-sample hypothesis testing". A/B testing is a way to compare two versions of a single variable, typically by testing a subject's response to variant A against variant B, and determining which of the two variants is more effective.

13. Mean Imputation: process of replacing the null values with mean values is called Mean Imputation.
    Is mean Imputation acceptable: NO, as the mean value replace the null values when the accurate value may be less than the mean value.

14. Linear Regression: linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression, for more than one the process is called multiple linear regression

15. Statistics: Statistics is a form of mathematical analysis that uses quantitative models to give a set of experimental data or studies of real life.

Statistics examine the methodology for collecting, reviewing, analyzing, and making data conclusions.

The two branches of Statistics are as follows.

a. **Descriptive Statistics**: Descriptive statistics are brief descriptive coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of a population.

MEAN, MEDIAN, MODE

b. **Inferential Statistics:** Inferential statistics can be defined as a field of statistics that uses analytical tools for drawing conclusions about a population by examining random samples.

Hypothesis testing and Regression Analysis

Machine Learning Assignment 1

1. A
2. A
3. B
4. C
5. C
6. B
7. D
8. A
9. A
10. A
11. B
12. B C

13. Regularization: Regularization refers to techniques that are used to calibrate machine learning models to minimize the adjusted loss function and prevent overfitting or underfitting. It mainly regularizes or reduces the coefficient of features toward zero. In simple words, "In regularization technique, we reduce the magnitude of the features by keeping the same number of features." There are two main types of regularization techniques:
Ridge Regularization and Lasso Regularization.

14. Algorithms used for Regularization: Ridge and Lasso regression are some of the simple techniques to reduce model complexity and prevent over-fitting which may result from simple linear regression
Ridge regression: Ridge regression is one of the types of linear regression in which a small amount of bias is introduced so that we can get better long-term predictions. Ridge regression is a regularization technique, which is used to reduce the complexity of the model. It is also called as L2 regularization. In this technique, the cost function is altered by adding the penalty term to it. The amount of bias added to the model is called Ridge Regression penalty. We can calculate it by multiplying with the lambda to the squared weight of each individual feature.

Lasso regression: Another regularization technique for reducing model complexity is lasso regression. Least Absolute and Selection Operator is what it stands for.
It's similar to the Ridge Regression, but instead of a square of weights, the penalty term simply contains absolute weights. Because it uses absolute data, it can reduce the slope to zero, whereas Ridge Regression can only go close. L1 regularization is another name for it.

15. In most cases, mean-square error (MSE) is used to calculate the model's error in linear regression. MSE is calculated as follows:
calculating the difference between observed and expected y-values for each x value;
each of these distances is squared;
determining the average of each squared distance
Linear regression finds the regression coefficient that produces the smallest MSE and uses it to fit a line to the data.

Python Worksheet-1

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