STATISTICS WORKSHEET 1

1-A

2-A

3-B

4-D

5-C

6-B

7-B

8-A

9-C

10- the normal distribution, also known as the Gaussian distribution, is the most important probability distribution in statistics for independent, random variables. Itis a continuous probability distribution that is symmetrical around its mean, most of the observations cluster around the central peak, and the probabilities for values further away from the mean taper off equally in both directions. Extreme values in both tails of the distribution are similarly unlikely. While the normal distribution is symmetrical, not all symmetrical distributions are normal.

11- The concept of missing data is implied as a data that is not captured for a variable for the observation in question. Missing data reduces the statistical power of the analysis, which can distort the validity of the results.

For missing data, it can be recommended for Missing at Random (MAR) imputation technique as Missing at Random means the data is missing relative to the observed data. The data is not missing across all observations but only within sub-samples of the data. It is not known if the data should be there; instead, it is missing given the observed data the missing data can be predicted based on the complete observed data.

12- A/B testing is one of the most important concepts because it is one of the most effective methods in making conclusions about any hypothesis one may have. Its simplest sense is an experiment on two variants to see which performs better based on a given metric.

13- It is not considered as good practice in general as explained below.

- Mean imputation preserves the mean of the observed data and it Leads to an underestimate of the standard deviation. It distorts the relationships between variables by “pulling” estimates of the correlation toward zero.
- It decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate.

14- Linear regression analysis is used to predict the value of a variable based on the value of another variable. This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values.

15-There are two main branches of statistics as listed below.

- Descriptive statistics-They are brief informational coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread). Measures of central tendency include the mean, median, and mode, while measures of variability include standard deviation, variance, minimum and maximum variables, kurtosis, and skewness..
- Inferential Statistics- Inference statistics are techniques that enable statisticians to use the information collected from the sample to conclude, bring decisions, or predict a defined population. They are of multiple types such as - Regression analysis, Analysis of variance (ANOVA), Analysis of covariance (ANCOVA), Statistical significance (t-test), Correlation analysis.