Statistics Assignment Answers

Question Number	Answers
1	A
2	A
3	B
4	D
5	C
6	B
7	B
8	A
9	C

Q10.

The normal distribution, also known as the Gaussian or standard normal distribution, is the probability distribution that plots all its values in a symmetrical fashion, and most of the results are situated around the probability's mean. Values are equally likely to plot either above or below the mean.

The normal distribution is a probability distribution that (roughly) describes many common datasets in the real world. It is the most common type of distribution, and it arises naturally in statistics through random sampling techniques.

Nowadays, it is more common to show up as a model for the "lifespan" of a product, like a lightbulb, or the outcome of standardized tests, like IQ. Biological measurements, like height or weight, are often estimated with normal distributions.

Q11.

Handling Missing data

After classified the patterns in missing values, it needs to treat them.

Deletion:

The Deletion technique deletes the missing values from a dataset. followings are the types of missing data.

Listwise deletion:

Listwise deletion is preferred when there is a Missing Completely at Random case.
In python we use **dropna()** function for Listwise deletion.

Pairwise Deletion:

Pairwise Deletion is used if missingness is missing completely at random i.e MCAR.
Pairwise deletion is preferred to reduce the loss that happens in Listwise deletion. It is also called an available-case analysis as it removes only null observation, not the entire row.
All methods in pandas like mean, sum, etc. intrinsically skip missing values.

Dropping complete columns

If a column holds a lot of missing values, say more than 80%, and the feature is not meaningful, that time we can drop the entire column.

Imputational Techniques

Imputation using Statistics:

The syntax is the same as imputation with constant only the SimpleImputer strategy will change. It can be “Mean” or “Median” or “Most_Frequent”.

Advanced Imputation Technique:

K_Nearest Neighbour Imputation:

The KNN algorithm helps to impute missing data by finding the closest neighbours using the Euclidean distance metric to the observation with missing data and imputing them based on the non-missing values in the neighbours.

Q12.

A/B testing (also known as bucket testing or split-run testing) is a user experience research methodology. A/B tests consist of a randomized experiment with two variants, A and B. It includes application of statistical hypothesis testing or "two-sample hypothesis testing" as used in the field of statistics. 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.

Q13.

The process of replacing null values in a data collection with the data's mean is known as mean imputation.

Mean imputation is typically considered terrible practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does.

Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

Q14. Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

Naming the Variables. There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand. The independent variables can be called exogenous variables, predictor variables, or regressors. Three major uses for regression analysis are (1) determining the strength of predictors, (2) forecasting an effect, and (3) trend forecasting.

Q15.

The two main branches of statistics are:-

Descriptive Statistics

CONCEPT The branch of statistics that focuses on collecting, summarizing, and presenting a set of data.

EXAMPLES The average age of citizens who voted for the winning candidate in the last presidential election, the average length of all books about statistics, the variation in the weight of 100 boxes of cereal selected from a factory's production line.

INTERPRETATION You are most likely to be familiar with this branch of statistics, because many examples arise in everyday life. Descriptive statistics forms the basis for analysis and discussion in such diverse fields as securities trading, the social sciences, government, the health sciences, and professional sports. A general familiarity and widespread availability of

descriptive methods in many calculating devices and business software can often make using this branch of statistics seem deceptively easy.

Inferential Statistics

CONCEPT The branch of statistics that analyzes sample data to draw conclusions about a population.

EXAMPLE A survey that sampled 2,001 full-or part-time workers ages 50 to 70, conducted by the American Association of Retired Persons (*AARP*), discovered that 70% of those polled planned to work past the traditional mid-60s retirement age. By using methods discussed in Section 6.4, this statistic could be used to draw conclusions about the population of all workers ages 50 to 70.

INTERPRETATION When you use inferential statistics, you start with a hypothesis and look to see whether the data are consistent with that hypothesis. Inferential statistical methods can be easily misapplied or misconstrued, and many inferential methods require the use of a calculator or computer.