ANSWERS to Worksheet_Set1

PYTHON ASSIGNMENT

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

STATISTICS ASSIGNMENT

QNo.	ANSWER
1	a
2	a
3	b
4	d
5	c
6	b
7	b
8	a
9	c
10	Normal Distribution is also called as Gaussian Distribution. It is a probability distribution that is symmetric about the mean. It shows that data near the mean are more frequent in occurrence than the data far from the mean. In graphical form the normal distribution will appear as a bell curve. In a normal distribution mean is 0 and standard deviation is 1.
11	The real-world data often has a lot of missing values. The cause of missing values can be data corruption or failure to record data. There are some imputation techniques to remove missing data in a dataset- 1. Deleting rows with missing values 2. Imputation using (mean/median) values 3. Imputation using (most frequent) or (constant/zero) values. 4. Imputation using K-NN method 5. Imputation using Multivariate Imputation by Chained Equation (MICE) 6. Imputation using Deep learning(DataWig)

12	A/B testing also called split testing**,** originated from the randomized control trials in Statistics, is one of the most popular ways for Businesses to test new UX features, new versions of a product, or an algorithm to decide whether your business should launch that new product/feature or not. The idea behind A/B testing is that you show the variated version of the product to a sample of customers (the experimental group) and the existing version of the product to another sample of customers (the control group)**.** Then, the difference in product performance in experimental/treatment versus the control group is tracked, to identify the effect of this new version(s) of the product on the performance of the product. So, the goal is then to track the metric during the test period and find out whether there is a *difference* in the performance of the product and what type of difference is it.
13	No, mean imputation of missing data is not an acceptable practice. It is a simple solution to remove missing data but there are some disadvantages of using it – 1. Mean imputation does not preserve the relationship among variables. 2. Mean imputation leads to an underestimate of standard errors.
14	Linear Regression is a basic and commonly used type of predictive analysis. The main 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 impact the outcome variable. Simple formula for regression is y = c + bx,

	Where: y = estimated dependent variable score c = constant b = regression coefficient and x = score on the independent variable
15	There are main two types of Statistics – A. Descriptive Statistics B. Inferential Statistics <u>Descriptive Statistics</u> The branch of statistics that focuses on collecting, summarizing and presenting a set of data. Ex. Average age of student in a class, height, weight of students in a particular section etc.. <u>Inferential Statistics</u> The branch of statistics that analyzes sample data to draw conclusions about a population. Ex. A sample survey of people in a city who voted in a particular election, and it can be considered as a population.

MACHINE LEARNING ASSIGNMENT

QNo.	ANSWER
1	A
2	A
3	C
4	B
5	C
6	B
7	D
8	D
9	A
10	B
11	B
12	A B
13	**REGULARIZATION**- Regularization is one of the most important concepts in Machine Learning. It is a technique to prevent the model from overfitting by adding an extra information to it. Sometimes the machine learning model performs well with the training data but does not perform well with the test data. It means the model is not able to predict the output when deals with unseen data by introducing noise in the output, and hence the model is called overfitted. This problem can be deal with the help of a regularization technique. This technique can be used in such a way that it will allow to maintain all variables or features in the model by reducing the magnitude of the variables. Hence, it maintains accuracy as well as a generalization of the model. The linear regression equation is – $y= \beta 0+\beta 1x1+\beta 2x2+\beta 3x3+\cdots+\beta nxn +b$

	In the above equation, Y represents the value to be predicted X1, X2, …Xn are the features for Y. β0,β1,…..βn are the weights or magnitude attached to the features, respectively. Here represents the bias of the model, and b represents the intercept.
14	There are two main techniques or algorithms used for regularization- **1. <u>Ridge Regression</u>** It is also called as L2 regularization. It is used to reduce the complexity of the model. 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. A general linear or polynomial regression will fail if there is high collinearity between the independent variables, so to solve such problems, Ridge regression can be used. **2. <u>Lasso Regression</u>** This type of regression is also called L1 regularization. Lasso regression is another regularization technique to reduce the complexity of the model. It stands for *Least Absolute and Selection Operator*.

	It is similar to the Ridge Regression except that the penalty term contains only the absolute weights instead of a square of weights. Lasso regression can help us to reduce the overfitting in the model as well as the feature selection. **<u>Key Difference between Ridge Regression and Lasso Regression</u>** ○ **Ridge regression** is mostly used to reduce the overfitting in the model, and it includes all the features present in the model. It reduces the complexity of the model by shrinking the coefficients. ○ **Lasso regression** helps to reduce the overfitting in the model as well as feature selection.
15	An error term represents the margin of error within a statistical model; it refers to the sum of the deviations within the regression line, which provides an explanation for the difference between the theoretical value of the model and the actual observed results. The regression line is used as a point of analysis when attempting to determine the correlation between one independent variable and one dependent variable. **Error Term Use in a Formula** An error term essentially means that the model is not completely accurate and results in differing results during real-world applications. For example, assume there is a multiple linear regression function that takes the following form: $Y=\alpha X+\beta\rho+\epsilon$

	where: α,β=Constant parameters X,ρ=Independent variables ϵ=Error term