**STATISTICS WORKSHEET-6**

Answer No. 1: D

Answer No. 2: A

Answer No. 3: A

Answer No. 4: C

Answer No. 5: A

Answer No. 6: A

Answer No. 7: C

Answer No. 8: B

Answer No. 9: B

Answer No. 10: Histograms and box plots allow to visually assess the central tendency, the amount of variation in the data as well as the presence of gaps, outliers or unusual data points. histograms and box plots are used to explore and present the data in an easy and understandable manner. Histograms are preferred to determine the underlying probability distribution of a data. Box plots on the other hand are more useful when comparing between several data sets. They are less detailed than histograms and take up less space. Histograms are better in displaying the distribution of data, you can use a box plot to tell if the distribution is symmetric or skewed.

Answer No. 11:  Metrics refer to a wide variety of data points generated from a multitude of methods. Every business executive, analyst, portfolio manager, and the project manager have a range of data sources available to them for building and structuring their own metric analysis. This can potentially make it difficult to choose the best metrics needed for important assessments and evaluations. Generally, managers seek to build a dashboard of what has come to be known as key performance indicators.  A final step is also setting goals and targets for KPI metrics that are integrated with business decisions. While there are a wide range of metrics, below are some commonly used tools:

Gross domestic product

Inflation

Unemployment rate

Answer No. 12: Statistical significance refers to the likelihood that a relationship between two or more variables is not caused by random chance. In essence, it's a way of proving the reliability of a certain statistic. Its two main components are sample size and effect size. Here are the steps for calculating statistical significance:

1. Create a null hypothesis.
2. Create an alternative hypothesis.
3. Determine the significance level.
4. Decide on the type of test you'll use.
5. Perform a power analysis to find out your sample size.
6. Calculate the standard deviation.
7. Use the standard error formula.
8. Determine the t-score.
9. Find the degrees of freedom.
10. Use a t-table.

Answer No. 13: Many random variables have distributions that are *asymptotically*Gaussian but may be significantly non-Gaussian for small numbers. For example the Poisson Distribution, which describes (among other things) the number of unlikely events occurring after providing a sufficient opportunity for a few events to occur. It is pretty non-Gaussian unless the mean number of events is very large. The mathematical form of the distribution is still Poisson, but a histogram of the number of events after many trials with a large average number of events eventually looks fairly Gaussian.

Answer No. 14: Assume we have x=(30,25,40,41,30,41,50,33,40,1000)x=(30,25,40,41,30,41,50,33,40,1000), what would be the mean and median.

x.mean <- sum(x) / length(x)

print(x.mean)

mean= 133

x.median <- median(x)

print(x.median)

median= 40

After defining both types of averages, we can now look into the difference between the two. While the arithmetic mean considers **all** the values in a vector, the median value only considers a **subset** of values.

This is because the median basically discards all vector elements except for the most central value(s). This feature of the median can make a big difference. As we have seen in our example, the mean of xx (133) was much larger than its median (40). In this case, this is because the median discards the value 1000 in xx, while the arithmetic mean considers it.

*If your data contains outliers such as the 1000 in our example, then you would typically rather use the median because otherwise the value of the mean would be dominated by the outliers rather than the typical values.*

Answer No. 15: In statistics, the *likelihood function*(often simply called the **likelihood**) measures the goodness of fit of a statistical model to a sample of data for given values of the unknown parameters.  But in both frequentist and Bayesian statistics, the likelihood function plays a fundamental role.

**WORKSHEET 6 SQL**

Answer No. 1: C & D

Answer No. 2: A, C & D

Answer No. 3: B

Answer No. 4: C

Answer No. 5: C

Answer No. 6: B

Answer No. 7: A

Answer No. 8: C

Answer No. 9: D

Answer No. 10: A

Answer No. 11: Denormalization is a database optimization technique in which we add redundant data to one or more tables. This can help us avoid costly joins in a relational database. Note that denormalization does not mean not doing normalization. It is an optimization technique that is applied after doing normalization. In a traditional normalized database, we store data in separate logical tables and attempt to minimize redundant data

Answer No. 12: A SQL cursor is a database object that is used to retrieve data from a result set one row at a time. A SQL cursor is used when the data needs to be updated row by row.

Answer No. 13: There are 5 types of SQL queries: -

1) Data Definition Language (DDL)

2) Data Manipulation Language (DML)

3) Data Control Language (DCL)

4) Transaction Control Language (TCL) and,

5) Data Query Language (DQL)

Answer No. 14: SQL constraints are used to specify rules for the data in a table.

Constraints are used to limit the type of data that can go into a table. This ensures the accuracy and reliability of the data in the table. If there is any violation between the constraint and the data action, the action is aborted. Constraints can be column level or table level. Column level constraints apply to a column, and table level constraints apply to the whole table.

Answer No. 15:  While creating a table we do not have unique identifier within the table hence we face difficulty in choosing Primary Key. so as to resolve such an issue we’ve to manually provide unique keys to every record but this is often also a tedious task. So, we can use Auto Increment feature that automatically generates a numerical Primary key value for every new record inserted.

In SQL Server, IDENTITY(starting\_value, increment\_value) is used for auto increment feature.

**MACHINE LEARNING**

Answer No. 1: C

Answer No. 2: B

Answer No. 3: A

Answer No. 4: A

Answer No. 5: B

Answer No. 6: A & D

Answer No. 7: B & C

Answer No. 8: D

Answer No. 9: B

Answer No. 10: The adjusted R2 will penalize you for adding independent variables (K in the equation) that do not fit the model. Why? In [regression analysis](https://www.statisticshowto.com/probability-and-statistics/regression-analysis/), it can be tempting to add more variables to the data as you think of them. Some of those variables will be [significant](https://www.statisticshowto.com/what-is-statistical-significance/), but you can’t be sure that significance is just by chance. The adjusted R2 will compensate for this by that penalizing you for those extra variables.

While values are usually positive**,** they can be negativeas well. This could happen if your R2 is zero; After the adjustment, the value can dip below zero. This usually indicates that your model is a poor fit for your data. Other problems with your model can also cause sub-zero values, such as not putting a constant term in your model.

Answer No. 11: Ridge and Lasso regression uses two different penalty functions. Ridge uses l2 whereas lasso go with l1. In ridge regression, the penalty is the sum of the squares of the coefficients and for the Lasso, it’s the sum of the absolute values of the coefficients. It’s a shrinkage towards zero using an absolute value (l1 penalty) rather than a sum of squares (l2 penalty).

As we know that ridge regression can’t zero coefficients. Here, you either select all the coefficients or none of them whereas LASSO does both parameter shrinkage and variable selection automatically because it zero out the co-efficient of collinear variables. Here it helps to select the variable(s) out of given n variables while performing lasso regression.

Answer No. 12: Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. the VIF for a regression model variable is equal to the ratio of the overall model variance to the variance of a model that includes only that single independent variable. This ratio is calculated for each independent variable. A high VIF indicates that the associated independent variable is highly collinear with the other variables in the model.

Answer No. 13: In machine learning algorithms, to bring all features in the same standing, we need to do scaling so that one significant number doesn’t impact the model just because of their large magnitude.

Feature scaling in machine learning is one of the most critical steps during the pre-processing of data before creating a machine learning model. Scaling can make a difference between a weak machine learning model and a better one.

The most common techniques of feature scaling are Normalization and Standardization.

Answer No. 14: Mean Squared Error (MSE)

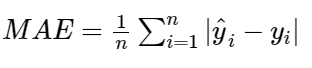
The most common metric for regression tasks is MSE. It has a convex shape. It is the average of the squared difference between the predicted and actual value. Since it is differentiable and has a convex shape, it is easier to optimize.



MSE penalizes large errors.

Mean Absolute Error (MAE)

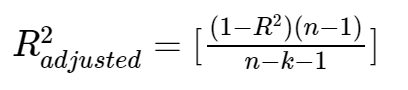
This is simply the average of the absolute difference between the target value and the value predicted by the model. Not preferred in cases where outliers are prominent.



MAE does not penalize large errors.

R-squared or Coefficient of Determination

This metric represents the part of the variance of the dependent variable explained by the independent variables of the model. It measures the strength of the relationship between your model and the dependent variable.



Answer No. 15: Sensitivity = tp/(tp+fn) = 1000/ (1000+250) = 1000/1250

Specificity= tn / (tn+fp) = 120/ (120+50) = 120/170

Precision = tp/ (tp+fp) = 1000/ (1000+50) = 1000/1050

Recall= tp/(tp+fn) = 1000/ (1000+250) = 1000/1250

Accuracy= tp+tn/(tp+tn+fp+fn) = 1000+120 / (1000+120+250+50) = 1120/1420