“Assignment-based Subjective Questions “

Question 1 - From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans -- Sell has increased in year 2019 compared to 2018.

Question2 -- Why is it important to use **drop_first=True** during dummy variable creation?

Ans – It reduces the correlations among the dummy variables.

Question 3 -- Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans – “Holiday” has the highest correlation with target variable.

Question 4 -- How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans – By distribution of residual against the dependent variables.

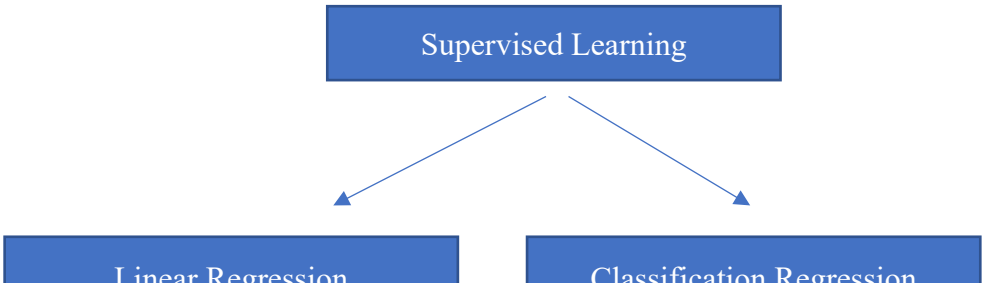
Question 5 -- Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans -- Holiday, year and Seasons.

“Subjective General Questions”

Question 1- Explain the linear regression algorithm in detail.

Ans -- Linear regression is supervised learning, where the predicated output is continuous and has a constant slope. Linear regression can be used for modelling, where machine run from the history set of data and predicate the output. Regression and Classification are type of supervised learning in which previous year data is used to predict the model.



Regression Algo is commonly used in predicative analysis across industry, ex -Banks, IT, Space & Research, share market, economics, education, engineering.

Regression algo can be further divided into Simple Linear Regression and Multi Linear Regression.

A. Simple Linear Regression-

   This is most basic type of regression in which we relationship is establish between dependent variable and one independent variable using straight line.

   Y=Bo+B1x (Where Bo= Intercepts and B1X =Slope.

   The strength of Liner regression can be accessed by

   - R^2

     This signifies if the model strength, ideally it should be in the range of 0-1.

   - Residual Standard Error (RSE)

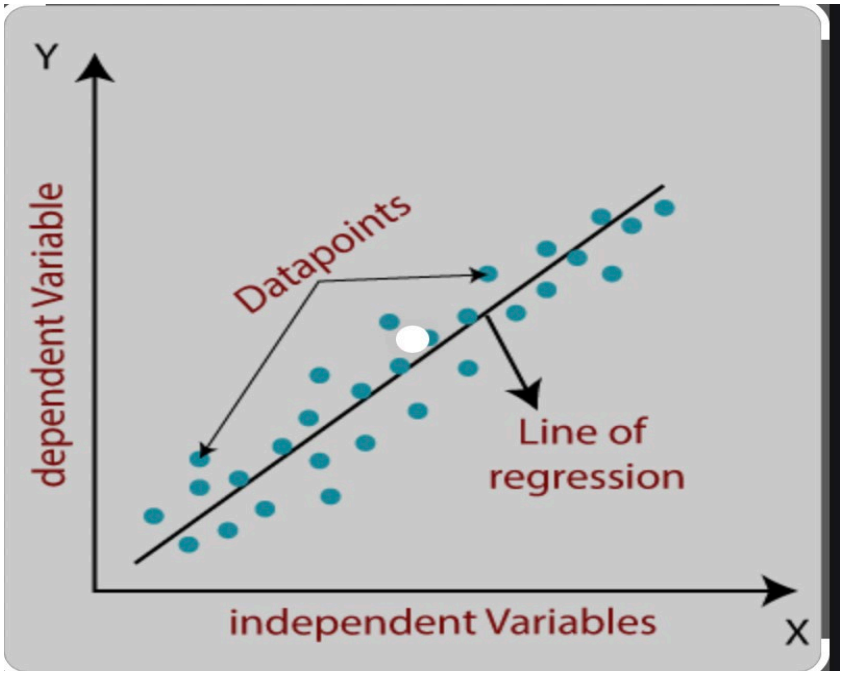
     It is total sum error across the sample, it is difference between expected and actual output. Small RSS indicate the tight and fit model.

B. Multi Linear Regression-

   This regression helps us to stablish relationship between one dependent variable and several independent variables. The objective of multiple regression variable is to find linear regression equation that can establish the relationship between dependent and independent variable.

   In short Multiple regression is extension to Linear Regression.

Y = Bo+B1X1+B2X2+B3X3……+ BpXp+



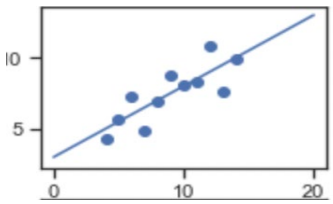
Question 2 -- Explain the Anscombe's quartet in detail.

Ans -- *Anscombe's Quartet* is the modal example to demonstrate the importance of data visualization which was developed by the statistician *Francis Anscombe* in 1973 to signify both the importance of plotting data before analysing it with statistical properties.

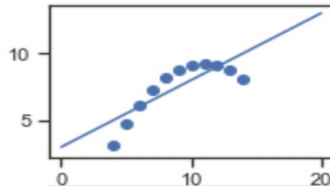
It comprises of four data-set and each data-set consists of eleven (x,y) points.
The basic thing to analyse about these datasets is that they all share the same descriptive statistics (mean, variance, standard deviation) but different graphical representation.

However, the statistical analysis of these four data-sets are pretty much similar. But when we plot these four data-sets across the x & y coordinate plane, we get the following results & each pictorial view represent the different behaviour.

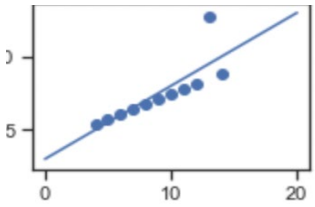
A. Data-set I , consists of a set of (x,y) points that represent a linear relationship with variance.



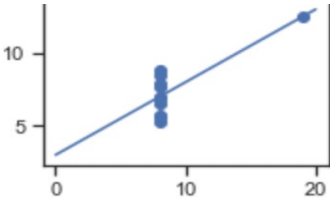
B. Data-set II, shows a curve shape but doesn't show a linear relationship.



c. Data-set III — looks like a tight linear relationship between x and y, except for one large outlier.



d. Data-set IV — looks like the value of x remains constant, except for one outlier as well.



Question3 - What is Pearson's R?

Ans – Pearson's R is also known as Pearson correlation aka Correlation Coefficient. The range of Pearson's R (-1 to 1).

It help to measure the relationship between 2 variable.

Pearson's r is a numerical summary of the strength of the linear association between the variables.

a. If the variables tend to go up and down together the correlation coefficient will be positive.
b. If the variables tend to go opposite side together the correlation coefficient will be negative.

Question4- What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans -- Scaling is important aspects of modelling. It help to streamline the data properly. When we have lot of independent variable in a model at lot of them might be on different level/scale which will lead a model with very weird co-efficient which might be impossible to predict. It's important to note the scaling only impacts the coefficient values other stat parameter remains same.

Therefore we need scaling manly for two purposes –

a. Ease of Interpretation.
b. Faster convergence for gradient descent method.

Standardized Scaling –

The variables are scaled in such a way that their mean is zero and standard deviation is one.

Normalized Scaling –

The variables are scaled in such a way it lies between 0 and 1 using the maximum and minimum values in the data.

Question5 - You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans- Yes, when there is perfect correlation between two independent variables, we observe VIF = Infinite.
An Infinite VIF indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

Question6 -- What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans – Q-Q is also known as Quantile-Quantile plot, which help us to access the distribution of random variable whether it is Normal, Exponential or Uniform distribution.

It is important as it help to determine two datasets with common background. Also help us to understand common location, scale, distribution shapes, distribution behaviour.