ASSIGNMENT BASED SUBJECTIVE QUESTIONS

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

A.

- Clean and clear weather is the most suitable weather for renting bikes .People generally avoid thunderstorm and Misty climate.
- Median is the highest for fall season which is the most optimal one for rental bike.
- Bike Rent preference is more on non holidays than the holidays. On working days workers prefer to hire rented bikes for their intercity travels for work purpose which is not required on holidays.
- Median is also high for the month of September and October, hence the demand is high in these months and that too when the sky is clear which is beginning of winter and end of monsoon.

2. **Why is it important to use drop_first=True during dummy variable creation?**

A.

- Drop_first = True is used to reduce the extra column while creating dummy variables.
- No of levels in Dummy variables will be n-1 where 'n' is the number of levels in categorical variables.
- For example if there are three categories : Literate , Illiterate and 10^{th} Pass. We don't need 10^{th} pass as we can consider them as Illiterate. Hence there will be only two variables Literate and Illiterate.

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

A.

- There is a multicolinearity between 'temp' and 'atemp' variable with target variable 'cnt'.
- However these variables are not used to build model before checking p value and VIF.

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

A.

- Coefficient is not equal to zero and hence we can reject the null hypothesis
- F statistics : 128.7 which means model very much significant
- The multicolinearity between predictors is very low and p value is significant.

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

A.

- Looking to the final model the weather conditions play an important role in contributing the significance towards explaining the demand of shared bikes.
- Thunderstorm variable has a coefficient value of -0.4136 which drastically decreases the demand
- The coefficient value of variable 'yr' is 0.2376 hence there is a considerable rise in the demand as compared to the previous year.
- The coefficient value of 9^{th} and 10^{th} month is 0.1832 and 0.1477 respectively. Hence maximum demand is in the month of September and October

GENERAL SUBJECTIVE QUESTIONS

1. **Explain the linear regression algorithm in detail.**

Regression showcases a relationship between two or more variables. The relationship is shown by an algebraic equation. It is actually used to predict the value of a variable based the value of other variable using algebraic equation.

LINEAR REGRESSION a type of Regression predicts a dependent or a target variable based on the values of independent variables. The dependent value is generally denoted by 'y' and independent by 'X'.

There are two types of Linear regression:

1. Simple linear regression

   It is used to find out relationship between 'y' and one independent 'X'

   Algebraic equation : $y = \beta0 + \beta1*X$

   $\beta0$ is the intercept coefficient

   $\beta1$ is the coefficient of X

   The main aim is to obtain line that best fits with the data

2. Multivariate linear regression

   It is used to find out relationship between 'y' and multiple independent 'X'

   Algebraic equation : $y = \beta0 + \beta1*X1+\ldots..+ \beta n*Xn + \sigma(y)$

   $\beta0$ is the intercept coefficient

   $\beta1$ βn is the regression coefficient of X

   σ is the Residual standard deviation

There are different types of parameters in linear regression algorithm which is used to predict the degree affection of the variables.

1. Residuals: The difference between the observed value of a variable and value predicted from the regression line. It is generally denoted by R SQUARE
2. VIF (VARIANCE INFLATION FACTOR) : It is used to measure the multicolinearity in a set of regression variables
   VIF = 1/1-R^2 where R is the residual

For example from a given data of detailed diet plan for set of population we can predict the major variables or factors which leads to Diabetes or sugar problems. Given the measure sugar intake for a particular meal is provided.

Hence Linear Regression is used real time in many companies to predict the demand of customers which changes based on various factors. Thus it is a very good tool to increase the profitability of the company.

2 Explain the Anscombe's quartet in detail.

- Anscombe's quartet is a model used for data visualization .
- It explains how much data plotting before analyzing the statistical properties are necessary.
- It generally comprises of 4 data sets with (x,y) data points. All the data sets share the same descriptive statistics i.e the mean and standard deviation but different graphical representation.

Data set 1 – Represents linear relationship with some variance.

Data set 2 – Curve shape doesn't show linear relationship

Data set 3- Strong Linear relationship with outlier.

Data set 4- x remains constant except the outlier.

3. **What is Pearson's R?**

Pearson's R is a Correlation measure which shows a linear relationship between two sets of data.

The strength of the relationship between data is defined by the Pearson's R

-1 : Indicates strong negative relationship

1 : Indicates strong positive relationship

0 : No correlation

Hence it is just a numerical indicator of the strength of the linear association between variables

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

- Scaling is a preprocessing of independent variables to normalize the data within a required range

- If scaling is not done for a data with high magnitude and units. Then algorithm only takes magnitude and doesn't consider units which results to incorrect modeling. Hence scaling is required. Scaling only affects the coefficients and not Rsquare or F statistics.

There are two types of scaling

1 Normalization/MinMax Scaling

It brings all the data in the range of 0 and 1 .

2 Standardization Scaling

It brings all the data into standard normal distribution which has mean as 0 and SD as 1

The major difference is Standardization Scaling loses some information about Outliers which is not occurred in Normalization.

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

VIF is the Variance Inflation Factor .

The formula for VIF = 1/1-R^2

Where R is the residual

Hence is if value of R comes as 1 then VIF=1/0 i.e = INFINTIY .

That means VIF comes as infinity in case of perfect correlation between the variables when R is one.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks**

Q-Q Plots are Quantile-Quantile plots which plot two quantiles against each other.

The Q-Q Plots whether there is a common distribution for a given 2 data sets population.

The slope determines the size of the steps of the data. If we have N observations then each steps have 1/ (N-1) of the data. So step sizes can determine the comparison between data and the normal distribution.

A steepy slopes suggest large amount of spread of data than we expect to be normally distributed.

For example in a median 50 percent data lie above the plot and rest lies below the plot. The purpose of Q-Q is to find out two sets of data coming from same distribution.