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? (3 marks)**

ANSWER : There are 7 categorical variables in the data-set and the inferences that we could derive after making a boxplot are :

- **SEASON :-** The season boxplot indicates that more bikes are rented during fall season and least bikes are rented during spring season.
- **YEAR :-** The year boxplot indicates that more bikes were rented in year 2019 as compared to year 2018.
- **MONTH :-** The month boxplot indicates that most bikes were rented between the month of May to October and between them September has the highest count.
- **HOLIDAY & WORKING-DAY :-** The holiday and working day boxplot indicates that most bikes were rented during normal working days rather than on week-ends or holidays.
- **WEEKDAY :-** The week-day boxplot show very close trend and looks like constant every day means every day bikes were rented.
- **WEATHERSIT :-** The weather situation boxplot indicates that most bikes are rented during clear weather(few clouds) & no bikes were rented during heavy rain /snow.

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

ANSWER : **drop_first=True** is important to use , as it helps in reducing the extra column created during dummy variable creation. If we do not use **drop_first=True**, then n dummy variables will be created, and these predictors(n dummy variables) are themselves correlated which is known as multicollinearity and it, in turn, leads to Dummy Variable Trap.

Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

ANSWER : By looking at the pair-plot among numerical variables, both Temp & Atemp has the highest correlation with the target variable('cnt').

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

ANSWER : Very low Multicollinearity between the predictors and the p-values for all the predictors seems to be significant. For now, we will consider this as our final model (unless the Test data metrics are not significantly close to this number). The Coefficient values from the model of all the variables are not equal to zero which means we are able to reject Null Hypothesis F-Statistics is used for testing the overall significance of the Model: Higher the F Statistics, more significant the Model is.

F-statistic: 230.4
Prob (F-statistic): 2.40e-187
The F-Statistics value of 230.4 (which is greater than my critical value) states that the overall model is significant.
The Residuals were normally distributed after plotting the distplot . Hence our assumption for Linear Regression is valid.
VIF calculation we could find that there is no multicollinearity existing between the predictor variables, as all the values are within permissible range of below 5 except 'temp' column which has VIF of '5.09' which is permissible because it is a good predictor for bike rents.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

ANSWER : The top 3 features contributing significantly towards explaining the demand of the shared bikes are :-

- **TEMPERATURE(temp) :-** Temperature shows positive correlation and has a coefficient value of '0.491508' which indicates that a unit increase in temperature variable increases the bike rental count by 0.491508 units.
- **YEAR(2019) :-** Year 2019 shows a positive correlation and has a coefficient value of '0.233482' which indicates that a unit increase in year 2019 variable increases the bike rental count by 0.233482 units.
- **WEATHER-SITUATION (Light-Rain) :-** Light rain weather situation shows a negative correlation and has a coefficient value of '-0.285155' which indicates that a unit increase in light rain weather situation variable decreases the bike rental count by 0.285155 units and vice-versa.

GENERAL SUBJECTIVE QUESTIONS

1. Explain the linear regression algorithm in detail. (4 marks)

ANSWER : Linear Regression is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable i.e. it finds the linear relationship between the dependent(y) and independent variable(x). It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

Linear regression is a powerful tool for understanding and predicting the behaviour of a variable but it has some limitations. One limitation is that it assumes a linear relationship between the independent variables and the dependent variable, which may not always be the case. In addition, linear regression is sensitive to outliers, or data points that are significantly different from the rest of the data. These outliers can have a disproportionate effect on the fitted line leading to inaccurate predictions.

Linear Regression is of two types: Simple and Multiple.

Simple Linear Regression is where only one independent variable is present and the model has to find the linear relationship of it with the dependent variable

Whereas, In **Multiple Linear Regression** there are more than one independent variables for the model to find the relationship.

Equation of Simple Linear Regression, where b_0 is the intercept, b_1 is coefficient or slope, x is the independent variable and y is the dependent variable.

$$y= b_o + b_1x$$

Equation of Multiple Linear Regression, where b_0 is the intercept, b_1, b_2, b_3, b_4…, b_n are coefficients or slopes of the independent variables x_1, x_2, x_3, x_4…, x_n and y is the dependent variable.

$$y= b_o + b_1x_1 + b_2x_2 + b_3x_3 \ldots. +b_nx_n$$

A Linear Regression model's main aim is to find the best fit linear line and the optimal values of intercept and coefficients such that the error is minimized.

Error is the difference between the actual value and Predicted value and the goal is to reduce this difference.

Mathematical Approach:

Residual/Error = Actual values – Predicted Values

Sum of Residuals/Errors = Sum(Actual- Predicted Values)

Square of Sum of Residuals/Errors = (Sum(Actual- Predicted Values))^2

i.e.

$$\sum {e_i}^2 = \sum (Y_i - \hat{Y}_i)^2$$

There are 5 evaluation metrics :-

- r^2 score = R-square value
- Adj. r^2 score = Adjusted R-square
- MSE = Mean Squared Error
- RMSE = Root Mean Squared Error & **5.** MAE = Mean Absolute Error

2. Explain the Anscombe's quartet in detail. (3 marks)

ANSWER : Anscombe's Quartet can be defined as a group of four data sets which are nearly

identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties. Anscombe's quartet highlights the importance of plotting data to confirm the validity of the model fit.

Let say for example we take a data-set, the Pearson correlation between the x and y values is the same, r = 0.816. In fact, the four different data sets are also equal in terms of the mean and variance of the x and y values. Despite the equivalence of the four data patterns in terms of popular summary measures, the graphical displays when we plot on scatter plot reveal that the patterns are very different from one another, and that the Pearson correlation is only valid for the data set.

3. What is Pearson's R? (3 marks)

ANSWER : Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposite direction with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

It was developed by Karl Pearson from a related idea introduced by Francis Galton in the 1880s, and for which the mathematical formula was derived and published by Auguste Bravais in 1844. The naming of the coefficient is thus an example of Stigler's Law.

The Pearson correlation coefficient (r) is the most widely used correlation coefficient and is known by many names :

- Pearson's r
- Bivariate correlation
- The correlation coefficient
- Pearson product-moment correlation coefficient (PPMCC)

The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics pf a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association

- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association

The Pearson correlation coefficient (r) is one of several correlation coefficient that you need to choose between when you want to measure a correlation. The Pearson correlation coefficient is a good choice when all of the following are true :

- Both variables are quantitative.
- The variables are normally distributed.
- The data have no outliers.
- The relationship is linear.

Pearson r Formula :

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Here,

- r = correlation coefficient
- x_i = values of the x-variable in a sample
- $\bar{x}$ = mean of the values of the x-variable
- y_i = values of the y-variable in a sample
- $\bar{y}$ = mean of the values of the y-variable

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

ANSWER : Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

Feature scaling is one of the most important data pre-processing step in ML. Algorithms that compute the distance between the features are biased towards numerically larger values if the data is not scaled. Feature scaling helps machine learning and deep learning algorithms train and converge faster. There are some feature scaling techniques such as Normalization and Standardization that are the most popular.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

NORMALIZATION :- Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Here’s the formula for normalization:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Here, X_{max} and X_{min} are the maximum and the minimum values of the feature respectively.

- When the value of X is the minimum value in the column, the numerator will be 0, and hence X’ is 0
- On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X’ is 1
- If the value of X is between the minimum and the maximum value, then the value of X’ is between 0 and 1

STANDARDIZATION :- Standardization is another scaling technique where the values are centred around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

Here’s the formula for standardization:

$$X' = \frac{X - \mu}{\sigma}$$

μ is the mean of the feature values and σ is the standard deviation of the feature values.

In this case, the values are not restricted to a particular range.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

ANSWER : VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In other words, VIF is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model. This can adversely affect the regression results. Thus, the VIF can estimate how much the variance of a regression coefficient is inflated due to multicollinearity.

If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is a perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 dur to the presence of multicollinearity. This would mean that the standard error of this coefficient is inflated by a factor of 2. The standard error of the coefficient determines the confidence interval of the model coefficient. If the standard error is large then the confidence intervals may be large and the model coefficient may come out to be non-

significant dur to the presence of multicollinearity. A general rule of thumb is that if VIF > 5 then there is multicollinearity. Note that this is a rough rule of thumb, in some cases we might choose to live with high VIF values if it does not affect our model results.
In general terms,

- VIF equal to 1 = variables are not correlated
- VIF between 1 and 5 = variables are moderately correlated
- VIF greater than 5 = variables are highly correlated

The value of VIF is calculated by the below formula:

$$VIF_i = \frac{1}{1-R_i^2}$$

Where, 'i' refers to the ith variable and R^2 represents the coefficient of determination.

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

ANSWER : The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential. If both sets of quantiles came from the same distribution, we should see the points forming the line that's roughly straight. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

Use of Q-Q plot in Linear Regression : The Q-Q plot is used to see if the points lie approximately on the line. If they don't, it means our residuals aren't Gaussian(Normal) and thus our errors are also not Gaussian.

Importance of Q-Q plot :

- The sample sizes do not need to be equal.
- Many distributional aspects can be simultaneously tested.
- The Q-Q plot can provide more insight into the nature of the difference than analytical methods.
- Skewness of distribution.