**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?

Solution:

Considering ‘Cnt’ as the dependent variable, we clearly see that there is a linear relationship between the independent variable and the dependent variable. For example, when we see the boxplot of cnt v/s season then we clearly see that during summer season there is a spike in the number of sales and it gradually increases in the season of fall, hence we can infer that dependent variable ‘cnt’ is dependent on season. There are couple of more things on which it is dependent and those factors are:

* Year
* Month
* Weekday
* Holiday
* Working day

These factors create an effect on the dependent variable by spiking the number of counts in particular season yielding to high rate of ‘cnt’ whereas in some season or month it results in low count. Hence it influences the sale of the bikes according to the month, season, day etc.

2. Why is it important to use drop\_first=True during dummy variable creation?

Solution:

When we create the dummy variables, we create exactly that number of dummy variables for which we have the levels. Hence, according to the formula we are supposed to create (n-1) dummy variables to create an effective model, by removing the first column of all the dummy variable we provide lesser columns to model to train and evaluate. For example, if we have 3 variables then we know that if other two variable carries 0 as the value, then it must be the 1st variable, hence it reduces the correlations created among dummy variables.

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

Solution:

Looking at pair-plot among the numerical variables, we can say that ‘temp’ and ‘atemp’ has highest correlation with the target variable. It has a correlation of 0.65 with the target variable.

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

Solution:

Validated the Linear regression by plotting:

- **Linear Relationship**: We see a linear graph for the dependent and independent variable.

- **Homoscedasticity**: The residuals have constant variance with respect to the dependent variable

- **Absence of Multicollinearity**: We removed all the columns with p-value greater than 0.05 and VIF value greater than 5.

- Independence of residuals (absence of auto-correlation)

- **Normality of Errors**: we see a bell-shaped curve.

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

Solution:   
Since the bookings increase on **good weather** days with **hotter temperature,** the company must increase their bike availability and promotions during the summer months to further increase their booking count.

**General Subjective Questions**

1. Explain the linear regression algorithm in detail.

Solution:

Linear Regression is very basic form of machine learning where we train a model to predict the behaviour of our data based on some variables. In case of linear regression, the variables on x-axis and y-axis must be correlated to each other linearly.

It is also termed as labelled machine learning as we use the past history of particular data and train the model based on the data and is corresponding label. By this way we teach the machine regarding what to predict if certain label comes while evaluating.

The algorithm of machine learning is simply based on the mathematical formula of slope i.e., y=mx+c, where ‘y’ is a dependent variable and is dependent on ‘x’ and ‘m’ is slope and ‘c’ is intercept. Linear regression is used to predict a quantitative response Y from the predictor value X.

We try to find the best fit line and for this we perform particular steps:

1. Plot a scatter plot to check the relationship.
2. Found residuals and the RSS for any given line passing through the scatter plot.
3. Then you found the equation of the best line by minimising the RSS and also find the optimal value for ‘m’ and ‘c’.

2. Explain the Anscombe’s quartet in detail.

Solution:

**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. When these models are plotted on a scatter plot, all datasets generates a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities.

3. What is Pearson’s R?

Solution:

When we try to evaluate between 2 variables to find the relationship between them then there are couple of things that we do to find the statistical relationship between them to understand it. One such measurement is concept of correlation. Correlation measures the strength of association between two variables as well as the direction. Pearson’s R is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0. The stronger the association of the two variables, the closer the Pearson correlation coefficient, r, will be to either +1 or -1 depending on whether the relationship is positive or negative, respectively. Achieving a value of +1 or -1 means that all your data points are included on the line of best fit.

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

Solution:

Scaling is a crucial process in machine learning where we try to fit all the values of the columns under some range so that there doesn’t come disparity while running the model which would result in high coefficient of few columns. There are couple of methods to perform the scaling.

* **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.**
* **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.**

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

Solution:

If there is perfect correlation, then VIF= infinity. This shows a perfect correlation between two independent variables. In the case of the perfect correlation, we get R^2=1, which led to 1/(1-R^2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicated that the corresponding variable may be expressed exactly by a linear combination of other variables.

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

Solution:

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. 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.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.