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)

Ans :

- Spring have lowest 'cnt' but fall have the highest 'cnt' value.
- Heavy rain/snow lowest 'cnt' but clear or cloudy have highest 'cnt' value.
- Low 'cnt' on Holidays.
- September have highest 'cnt' value and December have lowest 'cnt' value.

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

Ans : drop_first=True helps in reducing the extra column created during the dummy variable creation , which will be highly correlated.

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

Ans : *atemp* and *temp* has the highest correlation with the target variable,

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

Ans : The distribution of the residuals was normal and centred around 0.

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

Ans :

Top 3 features are :

i. temp
ii. Year
iii. Light_rain_snow_Thunderstorm

General Subjective Questions

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

Ans :

• Linear Regression Algorithm is a machine learning algorithm based on supervised learning. Linear regression is a part of regression analysis. Regression analysis is a technique of predictive modelling that helps you to find out the relationship between Input and the target variable.

• Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

- Example for that can be let's say you are running a sales promotion and expecting a certain number of count of customers to be increased now what you can do is you can look the previous promotions and plot if over on the chart when you run it and then try to see whether there is an increment into the number of customers whenever you rate the promotions and with the help of the previous historical data you try to figure it out or you try to estimate what will be the count or what will be the estimated count for my current promotion this will give you an idea to do the planning in a much better way about how many numbers of stalls maybe you need or how many increase number of employees you need to serve the customer. Here the idea is to estimate the

future value based on the historical data by learning the behaviour or patterns from the historical data.

- In some cases, the value will be linearly upward that means whenever X is increasing Y is also increasing or vice versa that means they have a correlation or there will be a linear downward relationship.
- One example for that could be that the police department is running a campaign to reduce the number of robberies; in this case, the graph will be linearly downward.
- Linear regression is used to predict a quantitative response Y from the predictor variable X.
- Mathematically, we can write a simple linear regression equation as follow y ~ b0 + b1*x Where y is the predicted variable (dependent variable), b1 is slope of the line, x is independent variable, b0 is intercept(constant). It is cost function which helps to find the best possible value for m and c which in turn provide the best fit line for the data points.

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

Ans :

Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

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

Ans :

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.
Pearson's r measures the strength of the linear relationship between two variables. Pearson's r always between -1 and 1.
If data lie on a perfect straight line with negative slope, then r = -1.

• Positive correlation indicates the both the variable increase and decrease together. Negative correlation indicates the one the variable increase and the other variable decrease and vice versa.

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

Ans : Scaling is a technique performed in pre-processing during building a machine learning model to standardize the independent feature variables in the dataset in a fixed range. The dataset could have several features which are highly ranging between high magnitudes and units. If there is no scaling performed on this data, it leads to incorrect modelling as there will be some mismatch in the units of all the features involved in the model.
The difference between normalization and standardization is that while normalization brings all the data points in a range between 0 and 1, standardization replaces the values with their Z scores.

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

Ans : In cases where we have perfect correlation between the dependent variable and independent variable(s), the R-squared value comes out to be 1. Hence VIF, which is (1/(1-R^2)) turns out to approach infinity.

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

Ans: Q-Q plot is a graphical tool to assess if sets of data come from the same statistical distribution. It is particularly helpful in linear regression when we are given testing and training datasets differently. In this scenario, it becomes important to check whether both the data comes from the same background, in order to maintain the sanity of the model.