Assignment-based Subjective Questions

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

- Answer : Observations from above boxplots for categorical variables:
- The year box plots indicates that more bikes are rent during 2019.
- The season box plots indicates that more bikes are rent during fall season.
- The working day and holiday box plots indicate that more bikes are rent during normal working days than on weekends or holidays.
- The month box plots indicates that more bikes are rent during september month.
- The weekday box plots indicates that more bikes are rent during saturday.
- The weathersit box plots indicates that more bikes are rent during Clear, Few clouds, Partly cloudy weather.

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

Answer. drop_first=True is important to use, as **it helps in reducing the extra column created during dummy variable creation**. Hence it reduces the correlations created among dummy variables

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

Ans.) By looking at the pair plot temp variable has the highest (0.63) correlation with target variable 'cnt'.

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

Ans:-

1. There should be a linear and additive relationship between dependent (response) variable and independent (predictor) variable(s). ...
2. There should be no correlation between the residual (error) terms. ...
3. The independent variables should not be correlated. ...
4. The error terms must have constant variance.

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

Ans:-

- The Top 3 features contributing significantly towards the demands of share bikes are:
- weathersit_Light_Snow(negative correlation).
- yr_2019(Positive correlation).
- temp(Positive correlation).

General Subjective Questions

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

Ans:- Linear Regression is **a machine learning algorithm based on supervised learning**. 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.

2. **Explain the Anscombe's quartet in detail.**

Ans:- 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.

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

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. ... The correlation coefficient is positive and height and weight tend to go up and down together.

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

Ans:- The most common techniques of feature scaling are Normalization and Standardization. Normalization is **used when we want to bound our values between two numbers, typically**, between [0,1] or [-1,1]. While Standardization transforms the data to have zero mean and a variance of 1, they make our data unitless.

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

Ans:- **If there is 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 due to the presence of multicollinearity.

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

Ans:- 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. ... This particular type of Q Q plot is called a normal quantile-quantile (QQ) plot