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)

The categorical variable, **weathersit** is significant. Especially the values **1** (*Clear, Few clouds, Partly cloudy, Partly cloudy*) and **3** (*Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds*) has **negative correlation** on the dependent variable, **cnt.**

The categorical variable, **mnth** is significant. While month **Sep** has a positive correlation with **cnt,** the months, **dec, jan** and **nov** have negative correlation with **cnt**.

The categorical variable, **weekday** is significant with only **Sat** has positive correlation with **cnt**.

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

By creating dummy variables we are converting the categorical variables (with categories >2) into numeric variables that can hold 0 or 1 values.

We would need n-1 dummy variables to represent the distribution. All 0 values across n-1 variables indicate what we are observing is outside the distribution across n-1 variables and covers the nth variable as well. It is important to have optimum features in the model and hence the use of drop first=True.

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

with the target variable? (1 mark)

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)

By drawing the pairplot for the numerical variables and box plot for the categorical variables, we test the linear relationship between Xs and Y.

To check that the error terms are normally distributed, plotted a histogram of the error terms from the final model on the train data set to check it is centred around zero and normally distributed.

During the model fitment, using the p value and VIF, removed the insignificant and the independent variables which are highly correlated (multicollinearity) such that the final model on the train dataset only contains error terms that are independent of each other.

To check homoscedasticity, checked from the histogram that the error term has mean of zero and have constant variance.

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

Working day with coeff value of 0.0570, yr with the coeff value of 0.2457 and light rain (dummy variable of weathersit) with coeff value of -0.3207 are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

General Subjective Questions

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

Linear regression is a type of machine learning wherein we determine the linear relationship between the independent variable(s) Xi and the dependant variable, y.

Linear regression can be simple linear regression wherein the linear relationship is between one independent variable x and a dependant variable, y.

Multiple linear regression has multiple independent variables.

Linear regression follows below assumptions:

- There is a linear relationship between X and y
- Error terms are normally distributed with mean zero
- Error terms are independent of each other
- Error terms have constant variance

In the linear regression, we follow the below steps:

- Data understanding and data loading
- Data preprocessing – dropping variables, creating dummy variables for the categorical variables, scaling the variables
- Per form EDA
- Train and Test split of the data
- Feature selection
- Model building with Train data
- Evaluation with Test data using R squared and adjusted R Squared

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

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

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

Scaling is performed to bring all the numerical independent variables on the same scale for better interpretation. In the normalized scaling the values are scaled between 0 and 1. In the standardized scaling, the advantage is we can find the outliers.

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

When there is perfect correlation between the variables and multicollinearity, we witness infinite value for VIF.

(3 marks)

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

(3 marks)