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?

Ans:

● The fall season has a high number of bike rentals.

● June, Aug, Oct have a high number of bike rentals.

● People prefer bike rental when weather situation is good.

● People do not prefer bike rental on weekends.

● People do not prefer renting bike on holidays.

● Year 2019 has higher Bike Rental than the year 2018.

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

Ans:

● To Avoid multi-collinearity. When we create dummy variables for categorical variables without dropping one category introduces multi-collinearity. By dropping one category, we can avoid multicollinearity, as the information from the dropped category is inherently included in the remaining ones.

● Interpretability: Dropping one category (typically the reference category) makes the interpretation of the model coefficients more intuitive.

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

Ans: registered variable is having high correlation with the target cnt variable.

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

Ans:

● By using Residual Analysis to validate the assumptions of Linear Regression.

● Plotted the histogram of the error terms and found that "Error Distribution" Is Normally Distributed Across 0, which indicates that our model has handled the assumption of Error Normal Distribution properly.

● Assumption of Error Terms Being Independent: we see that there is almost no relation between Residual and predicted Value

● Homoscedasticity: we can see that variance is similar from both ends of the fitted line.

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 are:

(a) season_summer

(b) season_fall

(c) weathersit_bad

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans:

● Linear Regression is a type of Supervised Machine Learning. There are two types of linear regressions.

● Simple linear Regression: where the number of predictors is one.

    ○ ex: y=b0+b1x, where b0 interceptors and b1 is coefficient or slop is x. x is the predictor.

● Multiple Linear Regression where the number of predictors is more than one.

    ○ ex : y= b0+b1x1+b2x2 + ...+bnxn, where b0 is interceptor and b1, b2,b3,...bn is coefficient/slopes of x1, x2, x3 ...xn predictors.

In Linear Regression Target variable is a continuous value. So linear regression is finding a fitted line(the fitted plane in case of multiple linear regression) so that the sum of error between the target value and the predicted value is minimum.

2. Explain the Anscombe's quartet in detail.

Ans:

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

3. What is Pearson’s R?

Ans:

"Pearson's r," also called a Pearson correlation coefficient is a statistic that quantifies the strength and direction of the linear relationship between two continuous variables. It measures how well the data points of two variables fit on a straight line. Pearson's correlation coefficient ranges from -1 to 1.

(a) When r is close to 1, it indicates a strong positive linear relationship. This means that as one variable increases, the other tends to increase as well.

(b) An r value of 0 suggests no linear relationship between the variable

(c) When r is close to -1, it indicates a strong negative linear relationship. This means that as one variable increases, the other tends to 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 preprocessing technique in data analysis and machine learning that transforms the features (variables) of a dataset to a common scale or range. It is primarily done to address issues related to the differing scales of variables, which can affect the performance of various machine learning algorithms.

Variables in a dataset may have different measurement units and scales. Some variables may have values in a small range, while others may have values in a much larger range. Scaling ensures that all variables contribute equally to the analysis or modelling process.

(a) In normalized scaling, the data is scaled to a specified range, typically [0, 1]. This is done by subtracting the minimum value of the variable from each data point and then dividing by the range. normalized scaling is useful when you want to preserve the original range of the data, and you're not concerned about the distribution's shape.

(b) In standardized scaling, the data is transformed to have a mean of 0 and a standard deviation of 1. This is achieved by subtracting the mean of the variable from each data point and dividing by the standard deviation.

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

Ans:

A VIF of infinity can occur when there is perfect multicollinearity in the model. Perfect multicollinearity means that one or more independent variables can be exactly predicted from a linear combination of the other independent variables.

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

Ans:

The Q-Q plot is designed to help you visually compare the quantiles of your data to the quantiles of a theoretical distribution, which can reveal deviations from the expected distribution. a Q-Q plot is a valuable tool for assessing the distribution of data, especially in the context of linear regression. It helps evaluate the normality assumption, detect skewness and outliers, and guide model improvement if deviations are observed.