Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
Ans: - Here are some of the inferences on the analysis of the categorical variables and their effect on the dependent variable.

- The season of Fall has the highest median followed by summer as they have the best weather conditions.
- The median bike rentals have increased in the year 2019 compared to the year2018. This may be due to the people getting conscious about the environment.
- The bike rentals are more on non-holiday days compared to holiday. This indicates that people prefer to spend time at home during the holidays.
- The months of Fall - June to October have a higher median value.
- The overall median for the weekdays and working-days are the same.
- The Clear weather situation has the highest median while the weather situation of Light snow has the least. The count of bike sharing is Zero for the weather situation - 4 'Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog".

Q2. Why is it important to use **drop_first=True** during dummy variable creation?
Ans: - 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.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Ans:- temp and atemp is highly correlated and both have similar correlation with target variable.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?
Ans:- For validating the assumptions of linear regression we have taken following steps,

1. we plotted displot for residual errors and found that they are normally distributed along zero (0).
2. Then plotted regplot for finding that assumption of Error Terms Being Independent.
3. Also plotted regplot for checking the Homoscedasticity of the predicted values.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Ans:- Following are the top 3 Features contributing significantly towards explaining the demand of the shared bikes.

1. Temp
2. casual
3. registered

Assignment-based Subjective Questions

Q1. Explain the linear regression algorithm in detail.
Ans:- Linear regression is an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. It is a statistical method used in data science and machine learning for predictive analysis.

Q2. Explain the Anscombe's quartet in detail.
Ans :- Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

Q3. What is Pearson's R?
Ans :- The **Pearson correlation coefficient (*r*)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Ans:-

Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

1. Normalized scaling:-

It brings all of the data in the range of 0 and 1.
sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

2. Standardized scaling

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

Q.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. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) 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 indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans:- 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.