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: From analysing categorical variables we can infer that:

- Season variables play a very important role in the demand for bikes, most demand is during Fall and decreases in winter
- Demand for bikes has increased exponentially from 2018 to 2019
- September has the most demand for renting Bikes, January being the lowest.
- Demand is decreased during holidays
- Weekdays have the most demand for bikes compared to weekends.
- Most bookings for bike rentals are received in clear weather.

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

Ans: It is important to use drop_first+True, to avoid multicollinearity issues in the data, and to have a stable model.

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

Ans: atemp and temp(temperature) variables have a high correlation with target variable 'cnt'

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

training set? Ans: After building the model: I validated the model using VIF to check for multicollinearity between independent variables, a VIF greater than 5 indicates high multicollinearity. Used residual analysis against the predicted values to check if the errors are normally distributed and have constant variance. A horizontal line with constant variance indicated homoscedasticity.

5. Based on the final model, which are the top 3 features contributing significantly towards

explaining the demand for shared bikes?

Ans: Based on the final model, the top 3 features contributing significantly towards explaining the demand for shared bikes are: Year, Temperature and Season.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans:-->Linear regression is a type of algorithm used in machine learning to predict the relationship between a dependent variable (also known as the target variable) and one or more independent variables (also known as the predictor variables).

-->The goal of linear regression is to find the best-fit line that can explain the relationship between the dependent variable and independent variables. This line is represented by a linear equation of the form y = mx + b, where y is the dependent variable, x is the independent variable, m is the slope of the line, and b is the y-intercept.

-->The linear regression algorithm tries to find the best values of m and b that minimize the difference between the predicted values of y and the actual values of y. This difference is known as the error or residual. The algorithm uses a method called least squares regression to find the values of m and b that minimize the sum of the squared residuals.

-->Once the best-fit line has been determined, the algorithm can be used to make predictions about new data. By plugging in the values of the independent variables, the algorithm can predict the corresponding value of the dependent variable.

2. Explain Anscombe's quartet in detail.

Ans:-->Anscombe's quartet is a set of four datasets that have nearly identical statistical properties, including the mean, variance, correlation, and linear regression line. Despite the similarities in statistical properties, the datasets look very different when graphed, highlighting the importance of data visualization in data analysis.

The quartet was created by the statistician Francis Anscombe in 1973 to illustrate the importance of graphing data before concluding. Each dataset consists of eleven x-y pairs and can be represented by a scatter plot.

-->The first dataset is a simple linear relationship, where y increases linearly as x increases. The second dataset has the same regression line as the first but with one outlier that has a large effect on the correlation coefficient. The third dataset is a non-linear relationship, where the x-y pairs follow a quadratic curve. The fourth dataset has a perfect linear relationship, except for one outlier that completely changes the regression line and correlation coefficient.

--> Anscombe's quartet highlights the importance of visualizing data in addition to computing statistical measures. While the four datasets have nearly identical statistical properties, they have very different patterns that would be missed without visual examination. This emphasizes the need to use graphs to check assumptions and to look for patterns and outliers in data analysis.

3. What is Pearson's R?

Ans:-->Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that shows the strength and direction of the linear relationship between two continuous variables. It is denoted by 'r' and ranges from -1 to 1, where -1 indicates a perfect negative linear relationship, 1 indicates a perfect positive linear relationship, and 0 indicates no linear relationship between the variables.

-->Pearson's R is calculated by dividing the covariance between two variables by the product of their standard deviations.

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

Ans:-->Scaling is the process of transforming numerical data so that it can be compared on a common scale. Scaling is performed to bring all the variables to the same range so that one variable does not dominate the other in terms of its influence on the model.

-->Normalized scaling and standardized scaling are two common scaling techniques used in machine learning. Normalized scaling scales the data between 0 and 1, while standardized scaling scales the data to have a mean of 0 and a standard deviation of 1. Normalization is useful when the range of values is not known, or when the range is very large. Standardization is useful when the range of values is known and is relatively small, and when the data has a Gaussian distribution.

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

Ans:-->VIF (Variance Inflation Factor) can be infinite when there is perfect multicollinearity between the predictor variables in a linear regression model. Perfect multicollinearity occurs when one predictor variable can be expressed as a linear combination of other predictor variables with perfect accuracy. This means that the variance of the coefficient for that predictor variable cannot be estimated, resulting in an infinite VIF value.

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

Ans:-->A Q-Q plot (Quantile-Quantile plot) is a graphical technique used to check if a set of data follows a particular distribution. In linear regression, a Q-Q plot is used to check if the residuals (difference between predicted and actual values) follow a normal distribution.

-->The Q-Q plot compares the theoretical quantiles of a normal distribution against the actual quantiles of the sample data. If the residuals follow a normal distribution, the points on the Q-Q plot should roughly follow a straight line.

-->The importance of a Q-Q plot in linear regression lies in its ability to help detect deviations from normality assumptions. Deviations from normality can affect the accuracy of statistical tests and confidence intervals, leading to incorrect inferences. Therefore, a Q-Q plot can help determine if any transformations or modifications are needed to satisfy normality assumptions and improve model accuracy.