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?

=> From the Categorical Analysis , we can infer that Fall and Summer are the most preferred seasons whereas Friday to Sunday faces a comparative higher demand . More bikes are rented in the period of June to October. Year beginning and ends has the least demand, but more demand on holidays. Significant rise in the demand in the year 2019 as compared to the previous year.

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

=> “drop_first = True” is an important parameter used to control the encoding scheme(n-1) during the time of creating dummy variables. When set to “True” , it drops the first level of each categorical variable , maintain the scheme and also helps in avoiding the multicollinearity issue in a Regression model.

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

=>Temp 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?

=>The assumptions can be validated by checking the VIF scores, p-values, residual analysis , distribution of the error and the linear relationship between the dependant variable and a feature variable.

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

=>Temperature, year and season_winter are the top 3 significant parameters in predicting the demand for Boom Bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

=>Linear Regression is a Machine learning algorithm which falls under supervised learning. It aims to establish a linear relationship between a dependent variable and one or more independent variables. It assumes that the relationship between the variables can be approximated by a straight line. There are two types of linear regression : simple linear regression and Multiple linear regression.

Simple linear regression is used when one independent variable is used to predict the target variable. Equation of the simple linear regression is $y = \beta_o + \beta_1 x + \varepsilon$

Multiple linear regression is used when multiple variables are used to predict the target variable . Equation of the multiple linear regression is $y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \varepsilon$

Linear regression is widely used for various purposes, such as predicting outcomes, understanding the relationships between variables, and identifying the importance of different predictors in explaining the variability in the dependent variable.

2. Explain the Anscombe’s quartet in detail.

=>Anscombe's quartet is a collection of four datasets, designed to have nearly identical summary statistics (such as means, variances, and correlations), but they exhibit distinct patterns when graphed. The purpose of Anscombe's quartet is to emphasize the importance of data visualization in understanding and interpreting statistical analyses.The purpose of Anscombe's quartet is to demonstrate that summary statistics alone cannot capture the nuances and complexities of data. Visualizing the data is essential for understanding the underlying patterns and making accurate interpretations.

3. What is Pearson’s R?

=>Pearson's R, also known as the Pearson correlation coefficient or Pearson's correlation, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It is named after Karl Pearson, who introduced this measure in the late 19th century.

Pearson's correlation coefficient (denoted as "r") is a value that ranges between -1 and 1. The coefficient indicates the extent to which the two variables are linearly related. The sign of the coefficient (+/-) indicates the direction of the relationship, while the magnitude of the coefficient indicates the strength of the relationship.

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

=>Scaling is the process of transforming data to a common scale. It is performed to make variables comparable and address issues caused by differences in magnitudes. Scaling ensures variables with different scales or units can be properly compared. It also avoids dominance of certain variables and helps optimization algorithms converge more effectively. Two common scaling techniques are normalized scaling (bringing data to a specific range) and standardized scaling (adjusting data to have a mean of 0 and a standard deviation of 1). Normalized scaling preserves relative relationships but may alter the original distribution. Standardized scaling maintains the distribution and facilitates comparison based on deviations from the mean. The choice depends on specific requirements: normalized scaling for relative relationships, and standardized scaling for distribution preservation and deviation-based comparison.

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

=>Infinite values in the Variance Inflation Factor (VIF) occur when there is perfect multicollinearity, which means there is an exact linear relationship between independent variables. This causes the VIF calculation to break down because it involves dividing by zero. Perfect multicollinearity undermines the estimation of individual effects, leads to unstable coefficients, and hampers interpretation. Dealing with multicollinearity involves identifying and addressing correlated variables through methods like variable removal or transformation. Infinite VIF values indicate a severe problem requiring immediate resolution.

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

=>A Q-Q plot is a graphical tool used in linear regression to assess the normality assumption of residuals. It compares the observed data's quantiles to the quantiles expected from a normal distribution. By examining the points' alignment with a straight line, we can determine if the residuals follow a normal distribution. Deviations from the line indicate departures from normality, which can highlight issues with the regression model. Q-Q plots help validate model assumptions and guide improvements, such as considering alternative models or variable transformations. They are important for assessing the reliability and appropriateness of linear regression analysis.