Assignment-based Subjective Questions

Question 1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer: Categorical variables like “Spring season”, “July month”, “Snowy weather” and “Misty weather” have a good negative correlation on the dependant variable. “October month” and “September month” have a good positive correlation on the dependent variable.

Question 2: Why is it important to use **drop_first=True** during dummy variable creation?

Answer: When we create dummy variable for a categorical column which has n levels, we ideally need n-1 columns which will suffice i.e. in order to get rid of an extra column we use drop_first = True. For example, we have a categorical column “transport_type” which has 3 values: air, road and water. If one variable is not air and road, then obviously it is water. So, we do not need the third variable to show water.

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

Answer: Actually casual and registered have highest correlation with the target variable but since the sum of casual and register is the target variable then we can ignore them and keep them aside hence temp and atemp have the highest correlation with the target variable “cnt”.

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

Answer: We observed the error terms i.e. residuals i.e. difference between actual and predicted values are following a normal distribution. We observed a linear relationship between target variable and independent variables which we found from model and the residuals i.e. error terms are independent of each other.

Question 5: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer: Based on the model obtained, the top 3 features contributing towards explaining the demand of the shared bikes are “temp”, “yr” and “september month”.

General Subjective Questions

Question 1: Explain the linear regression algorithm in detail.

Answer: Linear regression algorithm is a type of supervised machine learning algorithm which is used to establish a linear relationship between one target variable and one or more independent variables. There are 2 types of linear regression i.e. simple and multiple where simple denotes a linear relationship between 1 target variable and 1 independent variable and multiple denotes a linear relationship between 1 target variables and multiple independent variables. The goal is to find a best fit line which best describes the relationship between a target variable and one or more independent variables. There are certain assumptions in linear regression model building is that the residuals i.e. difference between actual and predicted values follows a normal distribution and the error terms independent of each other and have constant variance.

Question 2: Explain the Anscombe's quartet in detail.

Answer: Anscombe's quartet contains 4 datasets and each data set contains 11 x and y data pairs. These 4 datasets have same statistical features like mean, variance, r-squared, etc but upon plotting them on scatter plot on a graph they have different representations. It also emphasizes the importance of exploratory data analysis and the drawbacks of depending only on summary statistics.

Question 3: What is Pearson's R?

Answer: Pearson's correlation i.e. R is one of the most popular formulae for calculating correlation coefficient. It is used to measure the strength of relationship between 2 variables. The correlation coefficient formula returns a value between 1 and -1 where 1 indicates a strong positive relationship, 0 indicated no relationship at all and -1 indicates strong negative relationship.

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

Answer: Scaling is nothing but normalizing the data i.e. it is used for bringing all the data into the same scale so that there are no extreme results. We have 2 ways of scaling data i.e. normalized scaling and standardized scaling. By using normalized scaling, we can bring the data into the range of -1 to 1 and it is also known as MinMax scaling and coming to standardized scaling – it transforms the data in such a way that the mean in 0 and standard deviation is 1.

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

Answer: When we get the value of VIF as infinite it indicates that a independent variable is perfectly predicted by other variables. This happens when R2 score approaches 1 where it is a perfect line. This is a case of perfect multicollinearity so in order to rectify this we will have to figure out and drop one of the variables which is causing this.

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

Answer: A Q-Q plot is short for Quantile-Quantile plot. It is used for determining whether 2 samples of data came from the same population or not. If points in the Q-Q plot lies close to a straight line, it means that the two distributions are similar.