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?

   Answer: The outcome is dependent on holiday , weather situation and season. Demand decreases during holidays and on misty weather situation. Demand is more during winter and summer while it reduces in spring.

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

   Answer: This is required to reduce the number of variables as (n-1) dummy variables can is sufficient to describe n categorical features.

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

   Answer: Temperature has the highest correlation with cnt

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

Answer: 1) All the independent variables are significant.(p values < 0.05)

2) VIF values of all independent variables are less than 5.So there is no significant correlation between the independent variables

3)R squared and Adjusted R squared of the model is 80%.There is significant confidence that the predictor variables can effectively determine the dependent variable.

4)Probability of F-statistic is nearly 0.

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

Answer: 1)Top three features are Temperature, Humidity and Windspeed . Humidity and Windspeed negatively impacts the demand.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Answer: Linear regression is a machine learning algorithm that computes the linear relationship between the dependant variable and one or more independent features by fitting a linear equation to the observed data.Gradient Descent is one of the technique used to perform Linear regression analysis by iteratively modifying model parameters to reduce the mean squared error of the model on the training data set.

2. Explain the Anscombe’s quartet in detail.

Answer: It is a group of dataset (x,y) (four) that have same mean and standard deviation and regression line but are qualitatively different. So if the things are looked at statistically all the data sets will look similar only when the data is viewed graphically the differences can be observed.

3. What is Pearson’s R?

Answer: This is a correlation coefficient statistic which used to determine the correlation between 2 independent variables. A person coefficient value of 0 determines no correlatin between 2 variables while 0 to 1 determines positive correlation and 0 to -1 determines negative correlation. A correlation value > mod(0.5) is considered as strong correlation.

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

Answer: Since the independent values of numerical variables might have different scale (for example temperature and humidity values are different) the linear regression model might compute irrelevant coefficients .The solution for this is to scale all numeric variables to standard values so that the model computes the coefficients of the variables correctly.
Normalized scaling changes the variables to a standard normal distribution with mean as 0.Standardized scaling changes the values to a range from 0-1 using the max and min of the dataset

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

Answer: VIF is given as 1/(1-R^2). So VIF is infinite means perfect correlation among 2 or more independent variables. Infininte VIF is not good news for variables involved as predictor in Linear regression hence needs to be dropped one by one.

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 a plot of quantiles of the first data set against the quantiles of second data set.In case of Linear regression the distribution of the error terms or prediction error can be checked using Q-Q plots.If there is significant deviation from the mean the distribution of the feature variables should be looked into while reavaluating the model