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?

After evaluation of the model I came up with below conclusion

- Most of the bikes are rented on a holiday
- Most of the bikes rented in 2019 in comparison to the year 2018
- Most of the bikes are rented during fall season
- Fall season from September to November, bikes rented are high in September
- Bikes are mostly rented in Good weather
- Most of the people opted rent bikes during working days
- Bike rental increases from Jan to July

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

drop_first is set to false by default. drop_first = True, it help to have the duplicate column after creating the dummy columns

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

Temp and atemp has the highest correlation with target cnt variable

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

After creating the final LR model,

1. By checking the Error Terms - Normality of error terms
2. Multicollinearity
3. Residual analysis

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

After seeing the final model variables, i found the below 3 features had significant impact

1. temp
2. weather
3. 3. Year

General Subjective Questions

1. Explain the linear regression algorithm in detail.
   Linear Regression is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

   Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.
   Hypothesis function for Linear Regression :

   $$y = \theta 1 + \theta 2 x$$

   While training the model we are given:
   x: input training data (univariate – one input variable(parameter))
   y: labels to data (supervised learning)
   When training the model – it fits the best line to predict the value of y for a given value of x.
   The model gets the best regression fit line by finding the best θ1 and θ2 values.
   θ1: intercept
   θ2: coefficient of x
   Once we find the best θ1 and θ2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.
   There are 2 types of linear regression:
   1. Simple Linear Regression
   2. Multiple Linear Regression

   Simple Linear Regression: It is a type of linear regression model where there is only one independent or explanatory variable.
   Multiple Linear Regression: It is similar to simple linear regression but here we have more than one independent or explanatory variable.
   Linear Regression can be written mathematically as follows: Y =
   mx +c
   Y - dependent variable X
   - independent variable C -
   is constant

2. Explain the Anscombe's quartet in detail.
   Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed

3. What is Pearson's R?

   - Pearson's r is a measure of the strength of the linear association between the variables.
   - The Pearson coefficient is a mathematical correlation coefficient representing the relationship between two variables, denoted as X and Y.

- Pearson coefficients range from +1 to -1, with +1 representing a positive correlation, -1 representing a negative correlation, and 0 representing no relationship.
- The Pearson coefficient shows correlation, not causation

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

- Scaling is a process which is applied to independent variables which has the high values unlike the dummy values.It helps to speed the linear regression algorithm

- 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.

- We have like Normalization / Min MAx scaling - It scales in a way that all the values lie between 0 & 1
- We have standardization scaling which replaces the values by Z scores

4. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

5. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
   Quantile-Quantile plot or Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other.
   - The first quantile is that of the variable you are testing the hypothesis for
   - The second one is the actual distribution you are testing it against.