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 i came up with below conclusion

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

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

By default, this is set to drop_first = False .

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 finalLR model, i worked on the below factors

1. By checking the Error Terms - Normality of error terms
2. Also by checking the multicollinearity
3. Also using 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. Year
2. temp
3. winter

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is defined as a machine learning algorithm which analyses how a model is fitting into a linear relationship between one dependent variable (example y) and one or more independent variables (like X1, X2, X3…. Xn). The target variable will depend on how the independent variables change and not the other way round.

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.

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

4. 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

5. 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.

6. 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.