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?

Ans: We can see in fall people are more likely to rent bikes. Bike rental is more in working days compared to holidays. We can see are less likely to rent bike on rainy and more likely to rent in clear whether.

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

Ans: We don't want to add redundant features in the feature set. So, we need to remove one feature while creating dummy variables.

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

Ans: cnt is highly correlated with registered with correlation of 95%. But we are not going to take this as feature variable. So, in the features we have considered, atemp is highly correlated with correlation of 63%.

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

Ans: We can see the probability of f statistics is very low around 6.47e-209, which is a good indicator of the model. Also, we can see the probability of the features are less than 0.05 which is also a good indicator. At the end we have created residual histogram, which seems like normal plot that satisfies the normal linear relationship assumption.

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

Ans: We can see from the coefficient list atemp, yr and mnth.

General Subjective Questions

1.Explain the linear regression algorithm in detail.

Ans: In linear regression model we take multiple features as input and predict the output in numbers. Example predicting sales of a company based on different parameters like year, month, ongoing festivals, investments on in different areas of business and current trend etc.

We take features in X variable and target in Y. The general equation for linear regression is

Y = mx + c

Where,

m is the slope i.e., dy/dx

c is the y intercept

So, with the help of past data, we try to find the best value for m and c, which would help in predicting future outputs.

2. Explain the Anscombe's quartet in detail.
Ans: Anscombe's quartet has four datasets those are seems to be identical in simple descriptive statistics. But have different distributions and appear very different when graphed.
It was constructed by Francis Anscombe in 1973 to demonstrate the importance of plotting graph before analysing and model building.
Because of the different distribution it is not possible to interpret all the 4 datasets with a linear model.

3. What is Pearson's R?
Ans: It is Pearson's correlation coefficient, which is also called Pearson's R. It is the measure of linear correlation between two sets of data. The Pearson's correlation lies between -1 and 1.

r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
r = 0 means there is no linear association
r > 0 < 5 means there is a weak association
r > 5 < 8 means there is a moderate association
r > 8 means there is a strong association

Formula: $\sum((X_i - X_{mean})(Y_i - Y_{mean}))/\surd(\sum(X_i - X_{mean})^2\sum(Y_i - Y_{mean})^2)$

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Ans: Scaling the process in which we scale the features into same levels. So improve the calculation.
Normalization: It is one of the scaling method in which we use min and max value of the feature used to scale the same feature.
Scales value between 0,1 or 1, -1
Useful when we don't know about outliers.

Standardization: It is one of the scaling method in which we use mean and standard deviation value of the feature used to scale the same feature.
It ensures 0 mean and unit standard deviation.
Useful when the distribution is normal.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Ans: When VIF is infinity then it indicates strong collinearity between independent variables. Because in that case we get R2 = 1 and as we know the formula to calculate the VIF is 1/(1-R2), which is infinity in such scenario.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Ans: When the quantiles of two variables are plotted against each other, then the plot obtained is known as quantile – quantile plot or qqplot. This plot provides a summary of whether the distributions of two variables are similar or not with respect to the locations.

It is used to compare the shapes of distributions.