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? (3 marks)

In year 2019 the demand of bike is more than 2018
Demand of bike rises in Clear weather situation
Demand of Bike is high in Fall and Summer season
Workingday has higher demand of bike than holiday.

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

Actually we can drop any column from created dummy variable and we drop 1 column so that we are left with n-1 dummy variables.
It simply means if any of remaining dummy columns doesn't have value as 1 then it means value belongs to dropped dummy variable.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

atemp has the highest correlation

4. How did you validate the assumptions of Linear Regression after building the model on the training set ? (3 marks)

By checking p-values are below 0.05.
By checking VIF of the selected features are in 0-5 range.
We have high value of r2square which says our model explains 82% variance in target variable.

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

year ,temp and Winter season.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear regression alogrithm explains the correlation dependent variable(y) and independent variable(x).
It is done by creating best fit line by using Least square method.
Then we can divide data into tarin and test data.We create the linear model on train data.
We do Residual Analysis by plotting a histogram for the error terms and check error terms are normally di stributed and are independent of each other.
Make predictions and Evaluate the model on test data using r_score.

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet comprises a set of four dataset, having identical descriptive stastical properties in ter ms of meas, variance, R-squared,correlations,and linear regression lines.
Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statics.
It also emphasizes the importance of using data visulaization to stop trends,outliers, and other crucial det ails that might not be obvious from summary statics alone.

3. What is Pearson's R? (3 marks)

Pearson's correlation coefficient is the test sttics that measures the statistical relationship or association b etween two continuous variables.
It is the most common method to use the numerical variables; it assigns a value between -1 and 1 , where 0 is no correlation , 1 is total positive correlation, -1 is total negative correlation.
Positive correlation signifies that if variable A goes up, then B also go up, whereas if the value of the corre lation is negative, then if A increases, B decreases.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and sta ndardized scaling? (3 marks)
It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range so that outliers don't effect the model.
Collected data great ranges and outliers impact the model significantlty
NOrmalized scaling t brings all of the data in the range of 0 and 1.
Standardization Scaling brings all of the data into a standard normal distribution which has mean ( ) zero a nd standard deviation one (σ).

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
In case of perfect correlation between two independant variables VIF becomes infinite and we have to dro p one of the variable in that case.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.(3 marks)
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 and the second one is the actual distribution you are testing it against.
plot helps us compare the sample distribution of the variable at hand against any other possible distributio ns graphically.