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?

There are categorical variables such as season, year, month, holiday, weekday, weathersit. each of them has a varied degree of dependency on the target variable.

Year has a positive correlation of +0.57 and **mnth** of +0.28 and **weathersit** has -0.3 and **holiday** also has a -ve correlation -0.069.

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

As we know for a n-level variable, we only need n-1 columns to represent the data. "drop_first=True" drops one of the dummy variables after creation as the dropped variable can be inferred from remaining values of the dummy variables.

Example: Travel by Air/road/sea

Can be represented by 2 variables: 00 is air, 01 is sea and 10 is road

Air	Road	Sea		Road	Sea
0	1	0	ROAD	1	0
1	0	0	AIR(0,0)	0	0
0	0	1	SEA	0	1

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

The dependent variable Temp has the highest correlation with the target variable cnt.

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

1. Zero mean, independent, normally distributed error terms that have constant variance

2. p-values for variables are in acceptable range

3. R-squared and adj R-Squared are close

4. Prob F- Statistic value is in the acceptable range

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

Temp , hum , holiday are the three features.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a method used to define a relationship between 1 or more independent variables and the target/dependent variable. The relationship between them is linear and is proportional to the amount of change in independent variable and the extent/degree of the change is defined by the **slope** of the equation **y = B0(y-intercept) + B1(slope) . x + e (error term).**

This is most frequently used in AI to predict the values of continuous variables with the help of supervised machine learning.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is group of 4 data sets, it is used to explain linear, quadratic, clusters which exhibits different properties than the original data. These data sets are used to identify the effects of present of small subset of dataset causing large variations in the data set summaries.

Example presence of 1 outlier causes large effect on regression line by shifting the slope and y-intercept.

The other dataset summary resembles a linear data set though the actual relationship of the dataset is quadratic.

3. What is Pearson's R?

Pearson's R is a correlation coefficient r which measures linear relationship between two variables X and Y. it ranges between -1 to +1.

1 indicates +ve linear relationship between X and Y

-1 indicated perfect -ve relationship between Xand Y

0 indicates that there is no relationship between the variables.

A +ve value indicates: (both increase or decrease in the same direction).

A -ve value indicates: (one increases the other decreases)

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

Scaling is a process in which the features of the data set are transformed so that they fall within a specific range, and to improve convergence or better gradient decent to use efficient compute resources.

Normalized scaling also known as min-max scaling is done to keep the value range between 0 and 1 , is it achieved by subtracting minimum value from xi and divide it by the difference between xmax-xmin. Xscaled = X-Xmin/Xmax-Xmin.

Standardized scaling uses Mean and standard deviation to achieve a mean of 0 and standard deviation of 1. X-mean(X)/std(X)

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

During my assignment work – I forgot to remove the y_train from the x_train dataset and it caused the result of R-squared to be equal to 1. If the R2 value is 1 VIF -> infinity as VIF=1/1-R2. This occurs as one variable can be perfectly predicted from other or perfect multicollinearity occurs.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A Q-Q plot, or quantile-quantile plot, is a visual tool that compares the quantiles of two data sets.

A quantile is the fraction of points that fall below a given value.

Q-Q plots can be used to:

1. Determine if a dataset follows a specific probability distribution.
2. Determine if two samples of data came from the same population.
3. Assess the similarity between the distribution of one numeric variable and a normal distribution