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?

A. The variable yr has maximum correlation with the target variable cnt

We have seen that the month Aug and Sept has also shown an increase in cnt count with holiday

and rain having a negative coefficient.

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

A. If we have small number of dummies then we can remove the first dummy

like, we don’t need separate dummies for yes and no

we can put 1 and 0 is same dummy.

if we have more categorical values the not dropping is good option.

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

A. temp has the highest correlation followed by yr column.

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

A. By checking if the error terms are also normally distributed and by plotting y_test and y_pred to understand the spread and calculating the r2_score

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

A. Yr and atemp explains the demand of the shared bikes significantly along with rain with negative coefficient.

General Subjective Questions

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

A.

linear regression algorithm predicts a dependent variable value based on a given independent variable

Linear regression models can be classified into two types depending upon the number of independent variables:

Simple linear regression: This is used when the number of independent variables is 1.

Multiple linear regression: This is used when the number of independent variables is more than 1.

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

A.

Anscombe’s quartet consists of four datasets

It has nearly same descriptive statistics

It is used to plotting the graphs before analysing and model building.

3. What is Pearson’s R? (3 marks)

A.

It’s a covariance of two variables divided by product of their standard deviations.

$PX,Y = cov(X,Y)/\sigma X\sigma Y$

where:

cov is the covariance

σX is the standard deviation of X

σY is the standard deviation of Y

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

It is used to normalize the independent variables

Normalized ranges between 0 and 1.

In Standardization the values are around the mean with a unit standard deviation.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

VIF can be infinite only if R-squared value is 1

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another.

It can be used with sample sizes also

Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

For two data sets also it can be used