Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

Categorical columns are weathersit, season, yr.

Rental is highest in summer. Rainy weather reduces rentals

year 2019 saw demand for bike rentals higher by 2084 units.

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

If we don't use the drop_first=True it will cause multicolinearity which in turn is going to impact the model. Also it will eliminate redundancy.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer: <Your answer for Question 3 goes below this line> (Do not edit)

temp has the highest correlation with the target variable (cnt)

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 4 goes below this line> (Do not edit)

Using residual analysis I tried to validate the assumption of linear regression. Also by calculating the VIF tried to validate the no multicoliearity assumption.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 5 goes below this line> (Do not edit)

temp, yr

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)

Total Marks: 4 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 6 goes here>

It is basically a machine learning algorithm used to model relationship between an dependent variable(target) and one or many independent variables(Predictors) by forming a lienar equation. main objective is to find the target variabled based upon the values of predictors.

Mathematically represented as: $Y = ß_0 + ß_1 X_1 + ß_2 X_2 \ldots + \varepsilon$

Y = Target variable

ß0 = Intercept i.e value of y when x = 0

$ß_1$ = Slope of line (Change in y for one unit change in x)

ε: Error term

Goal of linear regression is to minimize the error between predicted and actual values. This is done by minimizing the Mean square error function.

MSE=n1i=1∑n(yi—y^i)2

To calculate the co-efficients Ordinary Least Square method is used.

$ß = (X^T X)^{-1} X^T y$

Where:

X: Matrix of independent variables.

y: Vector of the dependent variable.

β: Vector of coefficients.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>

It's group of 4 datasets designed to show the importance of visualizing data before analyzing data.

Even though mean of all the dataset is approximately same, variance is same for x and y
and r is similar but while plotting the plots appear to be completely different.

So, from this anscombe's quartet it is important to note that even though stastical metrics are almost similar but visualization is differnt so need to visualise before building the model

Question 8. What is Pearson's R? (Do not edit)

Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>
It signifies the strengh and direction of relationship of two variables.
Formula:
$r=\sum(xi-\bar{x})2\sum(yi-\bar{y})2\sum(xi-\bar{x})(yi-\bar{y})$

r=1: Perfect positive linear correlation.
r=-1: Perfect negative linear correlation.
r=0: no linear corelation

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>
Scaling is transforming the range of predictors in a dataset if the predictos have wide range of magnitude.

In the assignment we scaled temp, humidity and windspeed to be in similar magnitude.

In normalized scaling value is adjusted within a specific range 0 to 1 where as in standardized scaling centers data around 0 and have standard deviatin 1.
Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 10 goes here>

VIF can be infinite if there is multicolinearity between predictors. That means two or more columns may be duplicate and one more thing is all dummy variables are included without dropping one.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>

It is used to chek if a dataset is normally distributed or not. In Linear regression it is used to check if residuals are normally distributed or not.