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)

The categorical variables like month, weathersit, holiday and season has significant effect on the dependent varible.

Weekday has the incorrect values in the dataset for 424 records and has not much effect after deriving it again from the date column.

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)

For dummy variables we create 1 column for each unique value and set a 0/1 flag(1-hot encoding). In such scenario when one column has 1 remaining has 0 value. If 1st column has 1 then remaining columns has 0. So, we drop the first column and still retain the intention of the data and by reducing features we improve efficiency too.

For N unique values in category we create N-1 columns.

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 and **atemp** variables are correlated 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)

1. Normal distribution of residuals. using histplot plotted residuals. Normal distribution of residuals is noted.
2. Constant variance of residuals. using scatterplot plotted residuals vs y_test_pred. Residuals are randomly scattered.
3. Independence of residuals. using scatterplot plotted residuals vs features. Residuals are randomly scattered.

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)

Light Snow(weather sit category), yr and atemp are the 3 features contributing significantly towards

the demand of the shared bikes.

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>
Linear regression falls under Supervised Learning.

Equation:
y = β0 + β1.x1 + β2.x2 + β3.x3 + β4.x4 + …….. + βn.xn

Cost function $\boldsymbol{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
For calculating weights/coefficients one among 2 approaches as stated below

1. $\boldsymbol{\beta} = (\boldsymbol{X^T X})^{-1}\boldsymbol{X^T Y}$ – Ordinary Least Squares
2. $\boldsymbol{\beta_j} = \boldsymbol{\beta_j} - \boldsymbol{\alpha}\frac{\partial}{\partial \boldsymbol{\beta_j}}(\boldsymbol{MSE})$ - Gradient Descent

Gradient Descent is the effective way to calculate the weights.

Linear regression needs to follow few assumptions.

1. Relationship between the X and y should be linear.
2. Residuals should be normally distributed.
3. Residuals should be independent
4. Constant variance of residuals.
5. No correlation among the independent variables.

We split our data into 2 parts training and test data and then derive the weights on training data.
After deriving the weights from training data.
We evaluate our model on the test data.
For evaluation we do use few metrics like
R^2 (R-square),
Adj. R^2
F-Statistic,
Prob(F-Statistic)
If the model has **Adj. R^2** closer to 1 we consider the model is good.
Prob(F-Statistic) should be closer to 0 or equal to 0.

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>
In Anscombe's quartet the dataset is categorized into 4 sets. And on these datasets, we are going to derive the statistics summary.
Sometimes the statistics are do not reveal much trends.
So we need to visualize before we start building our model. Anscombe's quartet tries to explain the same.

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>

Pearson's R is otherwise called Pearson's correlation coefficient.
Pearson's R is used to measure linear relationships.

It is used to understand the relationship among the variables.
If r > 0.7 and r < 1 then variables are strong positive correlated
r > 0.3 and r < 0.7 then variables are positive correlated
r = 0 then variables are not correlated
r > -0.7 and r < -0.3 then variables are negatively correlated
r > -1 and r < -0.7 then variables are strong negative correlated.

Note: r value strictly lies between -1 and +1 and does not range beyond.

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 used for bring the data on same scale.
There are 2 types of scaing

1. Normalised scaling
2. Stadardized scaling

For normalized scaling, it scales the data between the range [-1, 1] or [0, 1]. For every value we try to subtract minimum and divide by the difference of Max min values.

Equation: $$\boldsymbol{x = \frac{x - x_{min}}{x_{max} - x_{min}}}$$

Ex: min= 10, max = 100, x = 25,

$$\boldsymbol{x = \frac{25-10}{100-10} = \frac{15}{90} = \frac{1}{6} = 0.166}$$

For standardized scaling, we are going to bring the data around the mean 0 with variance 1.

Equation: $$x_{std} = \frac{x-\mu}{\sigma}$$

Ex: mean = 10, variance = 1, x=11

$$x = \frac{11-10}{1} = \frac{1}{1} = 1$$

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 is variance inflation factor. It is used to identify if a independent variable is highly correlated with the other independent variables.

Equation: $$VIF = \frac{1}{1-R^2}$$

If R2 is closer to 1 or equal to 1 then the denominator becomes 0 and anything divided by 0 is infinite.

Ex: R2=1,

$$VIF = \frac{1}{1-1} = \frac{1}{0} = \infty$$

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>

Q-Q plot stands for Quantile-Quantile plot. It is used to compare the theoretical normal distribution with the residual/error terms distribution.

For the normally distributed residuals we can have a straight-line using Q-Q plot.

Q-Q plot is important for Linear regression as we can visualize the residuals of the model and check if they are normally distributed. Whether there is a curvature is present or not.

We can identify the trend also using the parameters of the Q-Q plot in python statsmodel package.
