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 variables inferred from the visualizations created using boxplots and bar charts.
- Most daily rentals fall within the range of 2,000 to 7,000, with a higher concentration between 4,000 and 5,500.
- Demand has increased significantly from 2018 to 2019
- Season of Fall has highest bike rental demands for 2018 and 2019, Season of Summer and Winter have moderate and equal bike rental demands for 2018 and 2019
- From May to Oct Demands have increased compared to other months, September has the highest demand followed by June
- Demand gradually increases from Wednesday to Saturday, with a peak on Saturday, followed by a decline in demand on Sunday compared to the other weekdays.
- Working days have the highest demand, though the difference compared to non-working days is not very significant.
- It is clear that "Bad" weather significantly reduced the demand
- During Holidays demand has significantly reduced
- Temperature affects the demand for bike rentals, with higher temperatures leading to increased demand.
- The count of rentals shows an increase and is right-skewed towards higher demands per day from 2018 to 2019 (as seen in the KDE).
- Variables other than "temp" and "feels_temp" do not have a significant impact on "cnt.", However, "temp" and "feels_temp" are redundant variables having high corelation.

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)

drop_first=True: Avoids the dummy variable trap by dropping the first category (e.g., if season has categories 1, 2, 3, 4, it will only create dummy variables for 2, 3, 4).

Note: Since season variable has only 3 distinct categories due to the absence of one category in the dataset, if we consider the missing category may appear in the future then we do not need to use drop_first=True when applying get_dummies.

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 "feels_temp" has the highest correlation.

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. Distribution of error terms – is it normally distributed
2. Residual analysis on the training set has any visible pattern using scatter plot
Homoscedasticity – If there If the spread of residuals increases or decreases as the fitted values increase
3. Linearity - The relationship between the independent variables and the dependent variable is linear identified using regplot
4. Multicollinearity - The independent variables should not be highly correlated with each other, which was identified by using RFE and VIF and dropped from the model.

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)
Top 3 feature contributing significantly towards explaining the demand of the shared bikes
1. year
2. weather_bad
3. feels_temp

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)

Linear regression is a fundamental statistical and machine learning technique used for

modeling the relationship between a dependent variable (often called the target or response variable) and one or more independent variables (often called features, predictors, or explanatory variables).

Basic Concept:
Simple Linear Regression: *When there is only one independent variable, the relationship can be visualized as a straight line on a 2D plane. The equation for this line is:*
$y = c + mX$
where:
y
is the dependent variable.
X
is the independent variable.
c
is the y-intercept (the value of y when X is 0).
m
is the slope of the line, indicating how much
y changes for each unit change in X.

Multiple Linear Regression: *When there are multiple independent variables, the equation becomes:*
$y = c + m_1 \cdot x_1 + m_2 \cdot x_2 + \ldots + m_n \cdot x_n$
Here, each
x_i
represents a different independent variable, and each
m_i
its corresponding coefficient.

Assumptions:
1. Linearity: The relationship between the independent and dependent variables should be linear.
2. Independence: Observations should be independent of each other.
3. Homoscedasticity: The variance of residual (prediction errors) should be constant for all values of independent variables.
4. Normality: The residuals should be normally distributed. This is particularly important for making inferences about population parameters.
5. No or little multicollinearity: In multiple regression, the independent variables should not be too highly correlated with each other.

Method of Estimation:
Ordinary Least Squares (OLS): This is the most common method for estimating the coefficients mi. The goal is to minimize the sum of the squared residuals (the differences between observed and predicted values)

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)

Anscombe's quartet is a collection of four datasets that share nearly identical basic statistical properties, but when visualized, they reveal very different patterns. Created by Francis Anscombe, this quartet highlights the importance of visualizing data before making conclusions based solely on numerical summaries.

Overview of the Datasets
Anscombe's quartet consists of four datasets, each with 11 data points. ***It serves as a reminder that visualizing data can uncover insights that numbers alone may not reveal.****. Key statistics for all four datasets are:*

Mean of X: 9
Variance of X: 11
Mean of Y: 7.50
Variance of Y: 4.123
Correlation between X and Y: 0.816
Linear regression line: y=3+0.5x
Despite these similarities, the visual structures of the datasets differ significantly:

Dataset 1:
Description: This dataset is relatively straightforward, with points scattered in a way that follows the linear trend defined by the regression line.
Visual: A scatter plot would show the points forming a line with some spread around it.

Dataset 2:
Description: Most points in this dataset lie along a straight line, but one outlier significantly affects the statistics.
Visual: The scatter plot would show a line of points with one extreme outlier disrupting the pattern.

Dataset 3:
Description: The points are tightly clustered near $x=8, y=6$
x=8,y=6, except for one outlier that skews the statistical properties.
Visual: The plot would appear as a dense cloud near a single spot, with one point far away from the cluster.

Dataset 4:
Description: This dataset forms a perfect quadratic relationship. However, the linear regression line is applied due to the specific arrangement of the points.
Visual: A scatter plot would show the points arranged in a parabolic curve.

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)

Pearson's R, also known as Pearson's correlation coefficient or Pearson product-moment correlation coefficient, is a measure of the strength and direction of the linear relationship between two continuous variables.

Definition:
Pearson's R is denoted by r and can take values between -1 and +1:

+1 indicates a perfect positive linear relationship: as one variable increases, the other increases proportionally. When r is positive, it suggests that as one variable increases, the other tends to increase as well.

-1 indicates a perfect negative linear relationship: as one variable increases, the other decreases proportionally. When r is negative, it implies that as one variable increases, the other tends to decrease.

0 signifies no linear correlation; the variables do not have a linear relationship. The absolute value of r tells you the strength of the relationship. Closer to 1 (or -1) means a stronger relationship, while closer to 0 indicates a weaker linear relationship.

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)

What is Scaling?
Scaling is the process of bringing the range of data features to a common scale, ensuring that no single feature dominates due to its numerical range. It is important for algorithms that are sensitive to the scale of input features, like those that use distance measures or gradient descent. Popular methods include normalization (min-max scaling) and standardization (z-score scaling).

Why is Scaling Performed?
Scaling helps improve model performance by:
- Ensuring all features contribute equally to the model.
- Speeding up the optimization process in algorithms like gradient descent.
- Avoiding bias caused by features with larger numerical ranges.

- It also makes data easier to interpret in some cases.

Difference Between Normalized Scaling and Standardized Scaling:

Normalized Scaling (Min-Max)	***Standardized Scaling (Z-Score)***
Scales data to a fixed range (e.g., [0, 1]).	*Centers data around 0 with a standard deviation of 1.*
Sensitive to outliers.	*Less sensitive to outliers.*
Preserves the original data distribution's shape but adjusts scale.	*Focuses on making the data follow a normal distribution.*

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)

If VIF as infinite, it means one of that variables is just a repeat of another in the dataset, like having the same information twice but labeled differently. This perfect match causes the math used to calculate VIF to try to divide by zero, because there's no new information being added by that variable. It's like saying the same thing twice; it doesn't help in understanding or predicting anything new.

e.g temp and feels_temp

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)

A Q-Q plot, short for Quantile-Quantile plot, is used to compare the distribution of your data with a theoretical normal distribution by plotting their quantiles against each other. In linear regression, it's particularly useful for checking if the residuals (the differences between observed and predicted values) follow a normal distribution, which is one of the key assumptions for making valid statistical inferences. If the points in the plot align closely along a straight line, it suggests that the residuals are normally distributed.

The importance of the Q-Q plot in linear regression lies in its ability to visually diagnose model fit and assumptions. Deviations from the line can indicate non-normality, skewness,

outliers, or other issues with the model's residuals, suggesting potential model improvements or data issues.
