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?

Total Marks: 3 marks (Do not edit)

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

I have used boxplot for analysis of categorical variables
Season - Summer and fall seasons have higher median rental counts compared to spring and winter.

Month- Peak demand occurring during the summer months.
Holiday- The box plot suggests that bike rentals might be slightly lower on holidays compared to non-holidays.
Weekday-Weekdays might have higher rental counts, dip on weekends could be because of reduced commuting needs.
Weather-Weather conditions impact bike rentals. "Clear" weather appears to be most favorable for bike rentals, followed by "Misty" conditions. "Light_snowrain" conditions have the lowest median rental count.
Bike rentals are more in the year 2019 as compared to 2018.

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 *n* categories, only *n-1* dummy variables are needed (last dummy variable can be predicted from the others). Pd.get_dummies with drop_first=True will give n-1 variables for n categories and reduce correlation among dummy variables.

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.

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)

Residual Plot: Create a scatter plot of the residuals (actual values - predicted values) against the predicted values. Ideally, you should see a random scatter of points around the zero line. Any patterns or trends (e.g., a curve, funnel shape) suggest violations of assumptions.

Variance Inflation Factor (VIF): Calculate the VIF for each predictor variable.[6] High VIF values (typically above 5 or 10) indicate high multicollinearity, which can affect coefficient estimates and model stability.[7]

Durbin-Watson Test: This test checks for autocorrelation in the residuals, which can occur when the errors in one observation are correlated with the errors in subsequent observations

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, windspeed, year

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 is also a type of supervised machine-learning algorithm that learns from the labelled datasets and maps the data points with most optimized linear functions which can be used for prediction on new datasets. It computes the linear relationship between the dependent variable and one or more independent features by fitting a linear equation with observed data. It predicts the continuous output variables based on the independent input variable.

For example, if we want to predict house price we consider various factors such as house age, distance from the main road, location, area and number of rooms. Linear regression uses all these parameters to predict house price as it considers a linear relation between all these features and price of house.

Assumptions are:
Linearity: It assumes that there is a linear relationship between the independent and dependent variables. This means that changes in the independent variable lead to proportional changes in the dependent variable.

Independence: The observations should be independent from each other, that is the errors from one observation should not influence the other.

Types of Linear Regression
Linear regression can be further divided into two types of the algorithm:
Simple Linear Regression:If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

Multiple Linear regression:If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Finding the best fit line:
When working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized. The best fit line will have the least error.
The different values for weights or the coefficient of lines (a0, a1) gives a different line of regression, so we need to calculate the best values for a0 and a1 to find the best fit line, so to calculate this we use cost function.
Residuals: The distance between the actual value and predicted values is called residual. If the observed points are far from the regression line, then the residual will be high, and so cost function will be high. If the scatter points are close to the regression line, then the residual will be small and hence the cost function.

Gradient Descent:
Gradient descent is used to minimize the MSE by calculating the gradient of the cost function.A regression model uses gradient descent to update the coefficients of the line by reducing the cost function.It is done by a random selection of values of coefficient and then iteratively update the values to reach the minimum cost function.

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>
Anscombe's Quartet is a famous set of four datasets.1 Each dataset has the same:

Mean of x: 9

Mean of y: 7.52

Variance of x: 113

Variance of y: 4.124

Correlation between x and y: 0.8165

Despite these identical summary statistics, when you plot each dataset, you'll discover vastly different relationships between the variables:

Dataset 1: A roughly linear relationship between x and y.

Dataset 2: A clear non-linear, curved relationship.

Dataset 3: A linear relationship, but with an outlier that significantly influences the regression line.

Dataset 4: A horizontal line with one outlier point that creates a spurious correlation.

The Importance of Anscombe's Quartet
Visualizations are crucial: Anscombe's Quartet demonstrates the importance of visualizing data before jumping to conclusions based solely on summary statistics.
Limitations of summary statistics: The quartet highlights the limitations of relying solely on summary statistics like mean, variance, and correlation. These statistics can mask underlying patterns or be heavily influenced by outliers.
Importance of data exploration: It emphasizes the importance of thoroughly exploring and visualizing your data before performing any statistical analysis.

In essence, Anscombe's Quartet serves as a stark reminder that:
Visualizations are essential for understanding data: They reveal patterns and relationships that summary statistics might not capture.
Outliers can have a significant impact: Even a single outlier can dramatically influence summary statistics and the interpretation of the data.
Data exploration is crucial: It's important to go beyond simple descriptive statistics and delve deeper into the data using visualizations and exploratory analysis techniques.

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 correlation coefficient (often denoted as "r" or "Pearson's r") is a statistical measure that quantifies the linear relationship between two continuous variables.

Range: The value of Pearson's r ranges from -1 to 1:
-1: Perfect negative correlation (as one variable increases, the other decreases)
0: No correlation between the variables
1: Perfect positive correlation (as one variable increases, the other also increases

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 a data preprocessing technique used to transform the values of features in a dataset to a similar range or scale. This is often necessary because features in a dataset can have vastly different ranges (e.g., age in years vs. income in dollars), which can negatively impact the performance of some machine learning algorithms.

Feature	Standardized Scaling	Normalized Scaling
Transformation	Zero mean, unit variance	Specific range (e.g., 0 to 1)
Outlier Sensitivity	Less sensitive to outliers	More sensitive to outliers
Data Distribution	Preserves the shape of the original distribution	May change the shape of the distribution
Application	Algorithms sensitive to feature scaling and outliers (e.g., SVM, KNN)	Algorithms that require data within a specific range (e.g., neural networks, image processing)

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 for a predictor variable is calculated as: $VIF_i = 1 / (1 - R_i^2)$
where R_i^2 is the R-squared value from a regression model where the predictor variable i is regressed on all other predictor variables in the model.

Certainly, let's explore why the Variance Inflation Factor (VIF) can sometimes be infinite.

VIF and Multicollinearity:

VIF measures the degree of multicollinearity among predictor variables in a regression model. High VIF values (typically above 5 or 10) indicate a high degree of multicollinearity, meaning that two or more predictor variables are highly correlated with each other.

Calculation: VIF for a predictor variable is calculated as:

$VIF_i = 1 / (1 - R_i^2)$

where R_i^2 is the R-squared value from a regression model where the predictor variable i is regressed on all other predictor variables in the model.

The VIF becomes infinite when the denominator in the equation $(1 - R_i^2)$ becomes zero. If a predictor variable can be perfectly predicted by a linear combination of other predictor variables, then R_i^2 will be 1. In this case, the denominator becomes 1 - 1 = 0, leading to an infinite VIF.

In simpler terms:For predictor variables, X1 and X2, and they are perfectly correlated (like X2 = 2 * X1). In this scenario, one variable provides no additional information beyond what is already captured by the other variable. This perfect linear relationship results in an infinite VIF for both variables.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Total Marks: 3 marks (Do not edit)

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

Q-Q plots are a valuable tool for assessing the normality of residuals in linear regression. By visualizing the distribution of residuals, you can gain insights into the validity of model assumptions and identify potential areas for improvement.
In linear regression, Q-Q plots are primarily used to assess the normality of the residuals.

Use a Q-Q Plot to Check Residual Normality:
Calculate Residuals: Obtain the residuals from your linear regression model.
Create the Q-Q Plot: Plot the quantiles of your residuals against the quantiles of a standard normal distribution.

Interpret the Plot:

Straight Line: If the points on the plot fall roughly along a straight line, it suggests that the residuals are normally distributed. This is generally desirable for valid statistical inference in linear regression.
Deviations from the Line:
S-shaped Curve: Suggests that the residuals are skewed (either positively or negatively).
Curved Line with Tails: Indicates that the residuals have heavier tails than a normal distribution (more outliers).
Points Far from the Line: Suggest potential outliers that may be influencing the model.

Importance of Q-Q Plots in Linear Regression:
Assumption Checking: Q-Q plots provide a visual diagnostic to check the normality assumption of the residuals, which is a key assumption for many statistical inferences in linear regression.

Model Improvement: If the residuals are not normally distributed, it may indicate that the model needs to be improved.