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 demad of bike is less in the month of spring when compared with other seasons
* The demand bike increased in the year 2019 when compared with year 2018.
* Month Jun to Sep is the period when bike demand is high. The Month Jan is the lowest demand month.
* Bike demand is less in holidays in comparison to not being holiday.
* The demand of bike is almost similar throughout the weekdays.
* There is no significant change in bike demand with workign day and non working day.
* The bike demand is high when weather is (Clear, Few clouds, Partly cloudy, Partly cloudy) however demand is less in case of (Light_Snow, Light_Rain_Thunderstorm_Scattered clouds, Light_Rain_Scattered_clouds). We do not have any demand for (Heavy Rain + Ice * Pallets + Thunderstorm + Mist, Snow + Fog) , so we can not derive any conclusion. May be the company is not operating on those days or there is no demand of bike.

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)

Using drop_first=True during the creation of dummy variables is crucial for avoiding multicollinearity, simplifying the model interpretation, and allowing for a more efficient and effective statistical analysis.

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)

From the above pairplot we could observe that, temp has highest positive correlation with 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)

Validating the assumptions of linear regression is essential to ensure that the model's results are reliable and that the conclusions drawn from it are valid.

1. The relationship between the independent and dependent variables should be linear.
2. The residuals (errors) should be independent; there should be no correlation between them.
3. The residuals should have constant variance (homoscedasticity) across all levels of the independent variables.
4. The residuals should be approximately normally distributed, especially important for hypothesis testing.
5. The independent variables should not be too highly correlated with each other.

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)

Based on final model top three features contributing significantly towards explaining the demand are:

* Temperature (0.581)
* weathersit_Light_Snow, Light_Rain_Thunderstorm_Scattered clouds, Light_Rain_Scattered_clouds (-0.226)
* year (0.256)

So it recomended to give these variables utmost importance while planning to achieve maximum demand.

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 statistical method used to model the relationship between one or more independent variables (predictors) and a dependent variable (outcome) through a linear equation. Below is a detailed explanation of the linear regression algorithm, its components, steps involved in implementation, and its evaluation.

1. Concept of Linear Regression

The core idea of linear regression is to find the best-fitting linear relationship between the independent variables (X) and the dependent variable (Y). The relationship is modeled by the following equation:

[

Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon
]

(Y) is the dependent variable (output),
(\beta_0) is the intercept (bias),
(\beta_1, \beta_2, ..., \beta_n) are the coefficients (weights) associated with the independent variables (X_1, X_2, ..., X_n),
(\epsilon) is the error term (the difference between the predicted and actual values).

2. Types of Linear Regression

Simple Linear Regression: One independent variable and one dependent variable.
Multiple Linear Regression: Multiple independent variables and one dependent variable.

3. Steps in the Linear Regression Algorithm

Step 1: Data Preparation

Collect and preprocess data:
Handle missing values.
Encode categorical variables (if any).
Normalize or standardize features if required.

Step 2: Splitting the Dataset

Divide the dataset into training and test sets (commonly a 70-30 or 80-20 split). The training set is used to fit the model, while the test set is used to evaluate its performance.

Step 3: Fitting the Model

Using Ordinary Least Squares (OLS): The coefficients are calculated to minimize the sum of the squared differences between the observed values and the values predicted by the model. This leads to the following formulation:
[
\text{Minimize} ; \sum (Y_i - \hat{Y}_i)^2
]

Where:

(Y_i) = actual values
(\hat{Y}_i) = predicted values
The formula to compute the coefficients (for multiple variables) is:

[
\beta = (X^T X)^{-1} X^T Y
]

Where:

(X) is the matrix of independent variables,

(Y) is the vector of dependent variable values,

(\beta) is the vector of coefficients.

Step 4: Making Predictions

Use the fitted model to make predictions on the training set and the test set using the equation:

[
\hat{Y} = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n
]

Step 4: Model Evaluation

Evaluate the performance of the model using metrics such as:

Mean Absolute Error (MAE): Average of absolute differences between predicted and actual values.

[
\text{MAE} = \frac{1}{n} \sum |Y_i - \hat{Y}_i|
]

Mean Squared Error (MSE): Average of the squares of the differences between predicted and actual values.

[
\text{MSE} = \frac{1}{n} \sum (Y_i - \hat{Y}_i)^2
]

Root Mean Squared Error (RMSE): Square root of MSE, providing error in the same units as the dependent variable.

[
\text{RMSE} = \sqrt{\text{MSE}}
]

R-squared: Proportion of the variance in the dependent variable that is predictable from the independent variables.

[
R^2 = 1 - \frac{\sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \bar{Y})^2}
]

Where (\bar{Y}) is the mean of the actual 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 famous dataset introduced by the statistician Francis Anscombe in 1973. It consists of four different datasets that have nearly identical simple descriptive statistics, yet each dataset has a distinctly different distribution and relationships between the variables. The purpose of Anscombe's Quartet is to demonstrate the importance of visualizing data before performing statistical analyses, particularly linear regression.

Structure of Anscombe's Quartet
The quartet consists of four datasets (denoted as A, B, C, and D), each containing 11 pairs of ((x, y)) values. The key characteristics of these datasets are:

All datasets have:
The same mean of (x) (average of the x-values).
The same mean of (y) (average of the y-values).
The same correlation coefficient between (x) and (y).
The same linear regression line (slope and intercept).

Datasets Overview

1. Dataset A

Data Points:
text
Copy
x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
y: 8.04, 6.58, 5.76, 6.58, 7.24, 5.25, 12.74, 12.74, 5.33, 4.74, 5.25
Visualization: The points form a linear relationship with a slight positive slope.

2. Dataset B

Data Points:
text
Copy
x: 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8
y: 6.58, 5.76, 7.24, 6.58, 5.25, 12.74, 12.74, 5.33, 4.74, 5.25
Visualization: All x-values are identical, resulting in a vertical line of points with an outlier that drastically affects the regression result.

3. Dataset C

Data Points:
text
Copy
x: 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8
y: 6.58, 5.76, 7.24, 6.58, 5.25, 12.74, 12.74, 5.33, 4.74, 5.25
Visualization: The points illustrate a curve, showing that the relationship between x and y is not linear.

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 correlation coefficient, is a statistical measure used to quantify the strength and direction of the linear relationship between two continuous variables. It is one of the most widely used correlation coefficients and is denoted by the symbol (r).

Key Features of Pearson's R
Range of Values:

The value of Pearson's R ranges from -1 to +1:
(r = +1): Perfect positive correlation, indicating that as one variable increases, the other variable also increases in a perfectly linear fashion.
(r = -1): Perfect negative correlation, indicating that as one variable increases, the other variable decreases in a perfectly linear fashion.
(r = 0): No correlation, indicating that there is no linear relationship between the variables.

Interpretation:

The closer the value of (r) is to +1 or -1, the stronger the linear relationship between the variables.
A positive value indicates a positive relationship, while a negative value indicates a negative relationship.

The formula for calculating Pearson's R is:

[
r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum

y)^2]}}
]

Where:

(n) = number of data points
(x) and (y) are the individual sample points from the two variables.

Assumptions:

Linearity: The relationship between the two variables should be linear.
Normality: The distributions of the variables should be approximately normally distributed, especially in smaller samples.
Homoscedasticity: The variance of the residuals should be constant across the range of values.
Applications:

Pearson's R is commonly used in a variety of fields, including psychology, biology, economics, and social sciences, to determine relationships between variables, assess the strength of these relationships, and inform analytical models.

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)

Scaling: Definition and Purpose
Scaling refers to the process of transforming the features (variables) of datasets, typically in machine learning or statistical analysis, so that they conform to a particular scale or range. This transformation is crucial when working with data as it can impact the performance of machine learning algorithms, especially those sensitive to the scale of data, such as gradient descent-based algorithms and distance-based algorithms (e.g., k-nearest neighbors and clustering).

Why is Scaling Performed?

Equal Contribution: Features with larger ranges can dominate the distance measures used in certain algorithms and skew results. Scaling ensures that all features contribute equally to the distance calculations and model training.

Improved Convergence: For optimization algorithms (like gradient descent), feature scaling helps speed up convergence. Algorithms may converge faster and require fewer iterations when features are on a similar scale.

Enhanced Performance: Some algorithms perform better when the input features have similar scales. For example, algorithms that rely on distance computations (like k-means clustering) can become inefficient if the features are not scaled.

Model Interpretation: Scaling can facilitate the interpretation of coefficients in linear models, as they indicate the change in the dependent variable per unit change in the independent variable.

Types of Scaling
Two common types of scaling are normalized scaling and standardized scaling.

1. Normalized Scaling (Min-Max Scaling)

Normalization transforms the features to a range between 0 and 1, or sometimes -1 to 1, using the following formula:

[
X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}
]

Where:

(X) is the original value,
(X_{min}) is the minimum value in the feature,
(X_{max}) is the maximum value in the feature.
When to Use:

Normalization is useful when the data needs to be bounded by a specific range, especially for algorithms that are sensitive to the relative scaling of features, such as neural networks or algorithms using distance calculations.

2. Standardized Scaling (Z-score Scaling)
Definition: Standardization transforms the features by removing the mean and scaling to unit variance (standard deviation). The formula for standardized scaling is:

[
X_{std} = \frac{X - \mu}{\sigma}
]

Where:

(\mu) is the mean of the feature,
(\sigma) is the standard deviation of the feature.
When to Use:

Standardization is preferred when the data follows a Gaussian (normal) distribution or when features have different units and variances. It's commonly used in machine learning models that assume a normally distributed dataset.

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)

Variance Inflation Factor (VIF) is a measure used to quantify the extent of multicollinearity in regression analysis. Specifically, it assesses how much the variance of an estimated regression coefficient increases when your predictors are correlated. A high VIF indicates that a predictor is highly correlated with other predictors in the model, which can make the model estimates unstable and difficult to interpret.

When the VIF for a predictor variable is infinite, it typically indicates one of the following scenarios:

Perfect Multicollinearity:

Perfect multicollinearity occurs when one predictor variable is an exact linear combination of one or more other predictor variables in the regression model.
For example, if you have a dataset where one variable is a multiple of another variable (e.g., (X_3 = 2 \times X_1)), the model cannot distinguish between these predictors, leading to an indeterminate situation. As a result, the calculation of VIF will yield an infinite value for the variable that is perfectly correlated.
Redundant Variables:

If two or more predictors give the same information (i.e., they are duplicates), this will also result in perfect multicollinearity. For instance, if you mistakenly include both Income and Income after tax in a model where they are perfectly correlated, this redundancy can lead to infinite VIF.
Data Issues:

Sometimes, data issues such as data entry errors, improper coding, or combining variables incorrectly during preprocessing can inadvertently introduce perfect multicollinearity.

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 (Quantile-Quantile plot) is a graphical tool used to assess whether a dataset follows a specified theoretical distribution, most commonly the normal distribution. It provides a visual means to compare the quantiles of the data against the quantiles of a given distribution.

Use and Importance of Q-Q Plots in Linear Regression

Normality of Residuals:

One of the key assumptions of linear regression is that the residuals (the differences between observed and predicted values) should be normally distributed.

A Q-Q plot of the residuals can visually assess this assumption. If the residuals follow a normal distribution, the points will lie approximately along the diagonal line.

Model Diagnostics:

Using a Q-Q plot helps in diagnosing model fit. If the Q-Q plot indicates non-normality of residuals, it could suggest that the linear regression assumptions are violated, which might lead to less reliable conclusions.

Outlier Identification:

Q-Q plots help to visualize the presence of outliers. Points that fall far from the reference line may indicate outliers in the dataset, which can disproportionately affect the regression model.

Model Improvement:

If non-normality is detected, it may prompt you to consider transformations of the target variable (e.g., logarithmic or square root transformations) or to explore different modeling approaches (e.g., generalized linear models) that better account for the distribution of residuals.

Robustness of Conclusions:

Ensuring that the residuals are normally distributed contributes to the robustness of inferential statistics derived from linear regression (e.g., confidence intervals, hypothesis testing). A Q-Q plot provides an easy way to check this before making statistical inferences.