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 independent variables are:
'Season', 'Mnth', 'Weekday', 'Weathersit', 'Yr', 'Holiday' and 'Workingday'

From the analysis of the categorical independent variables, the following inferences can be made about their effect on the dependent variable (cnt), which represents the total count of rental bikes:

Season:

Bike demand is strongly influenced by the season. The highest demand occurs during the FALL season, followed by SUMMER and WINTER, while SPRING experiences a significant drop. This indicates that seasonal weather conditions play a crucial role in driving bike rentals.

Month (Mnth):

The months of August, June, and September show the highest demand, followed closely by July, May, and October. This six-month high-demand period (May to October) highlights the importance of seasonality in bike usage patterns, correlating with warmer, pleasant months.

Weekday:

Demand is strongest on Fridays, Saturdays, Sundays, and Thursdays, suggesting that bike rentals cater to a mix of commuting (weekdays) and leisure activities (weekends). This trend indicates consistent usage across different purposes.

Weather Situation (Weathersit):

Clear weather days see peak bike demand, showcasing a strong correlation between favorable weather conditions and bike usage. Poor weather conditions (e.g., heavy rain or snow) lead to a decline in demand.

Year (Yr):

There is a noticeable growth in demand from 2018 to 2019, indicating that bike rentals are becoming more popular, either due to increased adoption or growing awareness of the service.

Holiday:

Bike usage is slightly lower on holidays compared to weekdays, possibly because regular commuting needs are reduced. However, leisure usage compensates for some of the decline, maintaining relatively stable demand.

Working Day (Workingday):

Bike usage remains consistent regardless of whether it's a working day or not, reflecting balanced demand between commuting and leisure purposes.

Overall Impact:

The categorical variables such as Season, Month, Weekday, and Weathersit significantly affect bike demand, driven by seasonal trends, weather conditions, and day-specific usage patterns. Year highlights business growth, while Holiday and Workingday have a milder impact, suggesting that bike rentals cater to both consistent commuter and leisure markets.

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 dummy variable creation avoids the dummy variable trap, which occurs when all categories of a categorical variable are included, leading to **multicollinearity** in regression models. By dropping the first category, one category acts as a reference, and the remaining dummies represent the relative impact of other categories. This ensures the model remains interpretable and computationally efficient without redundant information.

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)

Highest Correlation:

The variable registered shows the highest correlation with the target variable cnt. This is evident from the strong direct relationship observed in the pairplot.

Direct Relationship:

As the number of registered users increases, the overall bike rental count (cnt) also increases significantly, indicating that registered users are the primary contributors to the total demand.

Close Second:

The variable casual also exhibits a strong positive correlation with cnt. Although its impact is slightly less than registered, it still contributes notably to the overall demand.

Temperature Variables:

temp and atemp (feeling temperature) are also positively correlated with cnt. Warmer temperatures encourage higher bike rentals, reflecting seasonal demand patterns.

Inverse Relationships:

humidity and windspeed show an inverse correlation with cnt. Higher humidity and windspeed deter bike rentals, though their impact is less significant compared to registered and casual.

Conclusion:

Among all numerical variables, registered has the strongest influence on cnt, making it the most significant predictor of total bike demand based on the pairplot.

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)

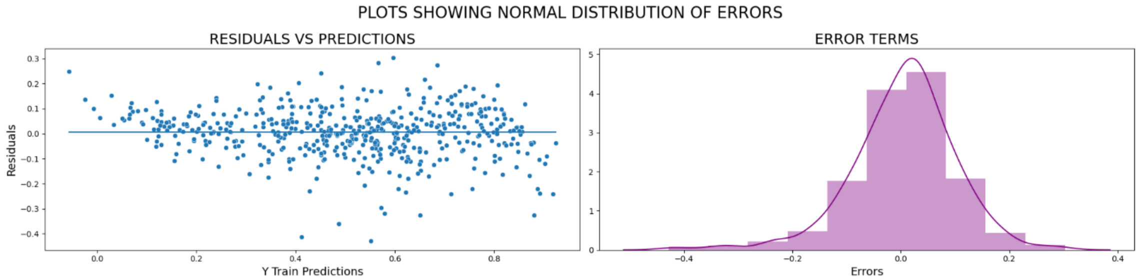
To validate the assumptions of Linear Regression after building the model on the training set, the following steps were taken based on the provided observations:

Mean of Residuals:

- Observation: The mean of residuals is 0.00740, which is very close to zero.
- Validation: This suggests that the model has no significant bias in its predictions, supporting the assumption that the residuals are centered around zero. The residuals should ideally have a mean of zero, which indicates the model is unbiased.

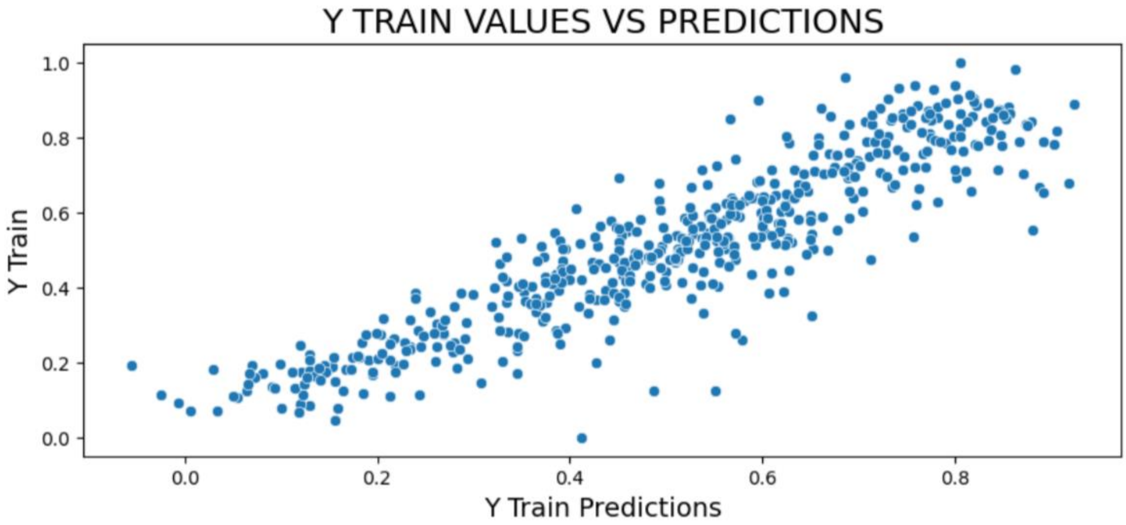
Normality of Residuals:

- Observation: The residuals are greater and concentrated above the mean for smaller predicted values.
- Validation: This indicates that errors are more pronounced at lower predicted values, and it supports the hypothesis that the residuals follow a normal distribution. A more detailed analysis, such as a Q-Q plot, could also confirm this.



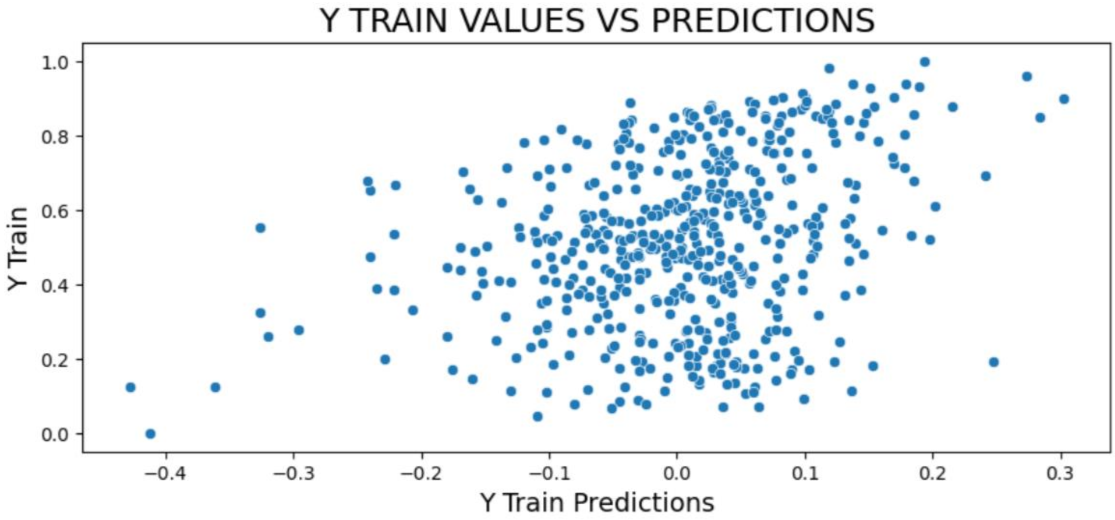
Homoscedasticity:

- Observation: The below plot of residuals display constant variation across the range of predictions, implying homoscedasticity.
- Validation: Homoscedasticity is confirmed if the residual plot shows no pattern or funnel-shaped distribution, meaning the variance of residuals remains consistent across different levels of the predicted values. The observation supports this assumption of constant variance of errors.



No Autocorrelation:

Although not explicitly mentioned in the provided observations, checking for autocorrelation using the Durbin-Watson test or visualizing residuals over time would help further validate this assumption.



Conclusion:
The provided observations suggest that the model meets the key assumptions of Linear Regression, including unbiasedness (mean residual = 0), normality, and homoscedasticity, which are essential for reliable and valid model predictions.

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 the final model and the updated feature set, the top 3 features contributing significantly towards explaining the demand for shared bikes in a positive way are:

Temperature (temp):

- Impact: Highly positive (0.64735). Warmer temperatures significantly drive bike demand, as people are more likely to use bikes in favorable weather.

Year (yr):

- Impact: Positive (0.24025). The demand for shared bikes is increasing over time, showing a growing adoption of bike-sharing services.

Season (Winter):

- Impact: Slightly positive (0.10697). While the impact is less significant than other factors, there is still some demand in the winter months, indicating that people are willing to use bikes even in colder weather.

These features highlight the importance of temperature and the growing trend in demand over time, while also acknowledging that even during winter, there is a positive but smaller demand for bike-sharing.

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 algorithm used for modeling the relationship between a dependent variable (target) and one or more independent variables (predictors). It assumes a linear relationship, where the target variable is a weighted sum of the predictors.

For Simple Linear Regression, the equation is:

$$y = \beta_0 + \beta_1 \cdot x + \epsilon$$

Where:
y is the target variable, x is the predictor, β 0 is the intercept, β1 is the slope, and ϵ represents the error term

For Multiple Linear Regression, the equation extends to multiple predictors:

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_n \cdot x_n + \epsilon$$

The goal of Linear Regression is to find the values of the coefficients 𝛽0, 𝛽1,…,βn that minimize the residual sum of squares, typically using Ordinary Least Squares (OLS).

Key assumptions include linearity, independence of residuals, homoscedasticity (constant variance), and normality of residuals. If these assumptions hold, Linear Regression can be a powerful tool for prediction.

Performance is evaluated using metrics such as R-squared (which measures the proportion of variance explained by the model), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

Linear Regression is simple, interpretable, and effective when the relationship between variables is linear, though it may struggle with complex relationships or multicollinearity.

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 group of four datasets devised by Francis Anscombe in 1973 to highlight the significance of data visualization in statistical analysis. Despite sharing identical summary statistics—such as mean, variance, and correlation—these datasets exhibit vastly different patterns when plotted, underscoring the potential pitfalls of relying solely on numerical summaries.

Features of Anscombe’s Quartet:

- Identical Statistics:

    1. Each dataset has the same mean and variance for both x and y.
    2. The correlation coefficient between x and y is identical.
    3. The regression line for all datasets is the same.

- Distinct Visual Patterns:

    1. Dataset 1: Displays a strong linear relationship between x and y, fitting the regression line well.
    2. Dataset 2: Demonstrates a clear non-linear (quadratic) relationship, indicating that a straight line is not an appropriate fit.
    3. Dataset 3: Contains an outlier that strongly influences the regression line, despite an otherwise linear pattern.
    4. Dataset 4: Shows no meaningful relationship between x and y, with points scattered randomly.

Anscombe’s Quartet serves as a compelling reminder to complement statistical analysis with visual exploration to avoid incorrect interpretations.

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)

A statistical tool for determining the direction and strength of a linear relationship between two continuous variables is Pearson's R, sometimes referred to as the Pearson correlation coefficient (PCC). Karl Pearson, who created it at the beginning of the 20th century, is honored by its name.

The range of values for r is -1 to 1, where:

- A perfect positive linear relationship is represented by a value of 1,

- a perfect negative linear relationship by a value of -1, and

- No linear relationship between the variables is indicated by a value of 0.

Formula: Pearson's R is calculated using:

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

Components:

1. n: The total number of data points (sample size).
2. $\sum xy$: The sum of the product of paired values for x and y (i.e., x1·y1+x2·y2+...+xn·yn).
3. $\sum x$: The sum of all x values.
4. $\sum y$: The sum of all y values.
5. $\sum x^2$: The sum of the squares of each x value (i.e., $x_1^2 + x_2^2 + \dots + x_n^2$).
6. $\sum y^2$: The sum of the squares of each y value (i.e., $y_1^2 + y_2^2 + \dots + y_n^2$).

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)

The process of modifying a dataset's feature value range to make it similar or fall inside a predetermined range is known as scaling. This is crucial because a lot of machine learning methods, especially those that use distance or gradient descent optimization, are sensitive to the size of the data.

Why is Scaling performed:

- **Equal Weightage:** It ensures all features contribute equally to the model and are not dominated by features with larger magnitudes.
- **Faster Convergence:** Scaling accelerates the convergence of gradient descent by avoiding large variations in feature values.
- **Improved Performance:** Some models, like SVMs, k-NN, and PCA, require scaled data for better performance, as they are distance-based.
- **Eliminate Bias:** Scaling prevents bias in models where certain features might dominate due to their larger values.

Difference Between Normalized Scaling and Standardized Scaling:

Aspect	**Normalized Scaling**	**Standardized Scaling**
Definition	Rescales data to a specific range, typically [0, 1] or [-1, 1].	Centers the data around the mean with a standard deviation of 1.

Formula	$\mathrm{X}_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$	$\mathrm{X}_{\text{std}} = \frac{X - \mu}{\sigma}$
Range of Values	[0, 1] (or [-1, 1] for some cases).	Mean = 0, Standard Deviation = 1.
Use Case	Useful for algorithms like k-NN or k-Means, where relative distances matter.	Preferred for algorithms like logistic regression, SVM, and PCA.
Effect on Outliers	Outliers can significantly affect the range.	Outliers have less impact since it uses standard deviation.
Example	Min-Max Scaling	Z-Score Normalization

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)

The value of the Variance Inflation Factor (VIF) can become infinite when there is perfect multicollinearity among the independent variables in a dataset.

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

If $R_i^2 = 1$, then $1 - R_i^2 = 0$, making VIF infinite. This occurs because the regression matrix becomes singular, preventing reliable coefficient estimation. Resolving this requires identifying and removing or combining collinear variables.

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)

An indicator of whether a dataset adheres to a particular theoretical distribution—usually the normal distribution—is a graphical tool called a Q-Q plot, or quantile-quantile plot. It plots a theoretical distribution's quantiles versus the quantiles of the observed data. The data is in good agreement with the expected distribution if the points roughly fall on a 45-degree line.

Use in Linear Regression:
The main purpose of a Q-Q plot in linear regression is to assess the assumption of residual normality, which is a crucial prerequisite for reliable hypothesis testing and confidence intervals. A normal distribution should be followed by residuals; departures from this rule suggest possible problems such as outliers or model misspecification.

Importance of a Q-Q plot in Linear Regression:

Residual Normality: A Q-Q plot aids in confirming the normality assumption, which is necessary for reliable hypothesis testing and regression analysis confidence intervals.

Spotting Deviations: Non-normality, such as skewness or heavy tails in the residuals, can be indicated by deviations from the straight line in the Q-Q plot. This could imply that the model would benefit from transformations (such as log or square root) or an other modeling strategy.

Model Hypotheses: The Q-Q plot is an essential diagnostic tool in regression modeling since it ensures normal residuals, which enhances the validity and reliability of statistical inferences.