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 such as season, holiday status, and weather situation significantly influence the demand for shared bikes. For instance, certain seasons may see higher usage due to favorable weather conditions, while holidays might affect daily commuting patterns, leading to variations in bike demand.

Understanding these categorical influences is crucial for accurately modeling and predicting bike-sharing demand.

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 is essential to avoid multicollinearity in regression models. By dropping the first category, we prevent the dummy variables from being perfectly collinear, ensuring that the model can uniquely estimate the effect of each category without redundancy. This practice leads to more stable and interpretable regression coefficients.

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)

Looking at pair plot among numerical variables, casual, and registered columns has highest correlation. But Cnt is sum of casual and registered column if we skip this two columns Temperature looks like a highest correlation variable to target variable.

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)

After building the linear regression model, the following steps are typically taken to validate its assumptions:

* **Linearity:** Checking scatter plots of residuals versus predicted values to ensure no obvious patterns, indicating a linear relationship.
* **Homoscedasticity:** Assessing whether residuals have constant variance across all levels of the independent variables.
* **Normality of Residuals:** Using Q-Q plots or statistical tests to confirm that residuals are approximately normally distributed.
* **Independence:** Ensuring that residuals are independent, particularly in time series data, by analyzing autocorrelation plots.
* **Multicollinearity:** Calculating Variance Inflation Factor (VIF) for predictors to detect

multicollinearity issues.

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 features contributing significantly towards explaining the demand of the shared bikes are as follow:

1. Humidity
2. Temperature
3. Wind Speed

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 *supervised machine learning algorithm* that models relationship between a dependent variable `$y$`and one or more independent variables `$x_1, x_2, ..., x_n$`.

Goal is to find best-fitting straight line (or hyperplane in higher dimensions) that minimizes difference between predicted and actual values of the dependent variable.

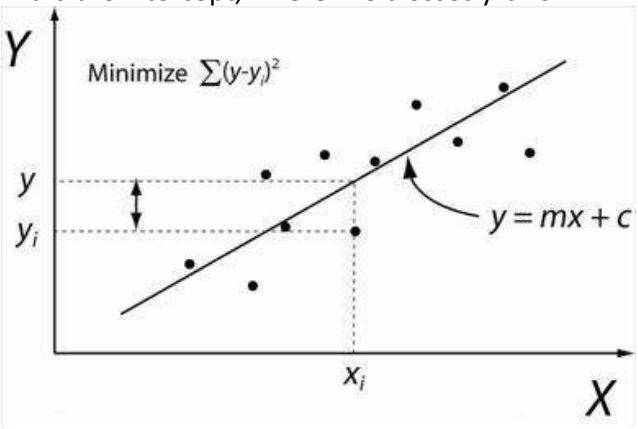
This is achieved by minimizing the Residual Sum of Squares (RSS), difference between observed values and predicted values.

Equation of a Simple Linear Regression is:

$$y=mx+b$$

Where:

m is the slope, representing relationship between x and y,
b is the intercept, where line crosses y-axis.



For multiple variables, model becomes:

$$y=b_0+b_1x_1+b_2x_2+...+b_nx_n$$

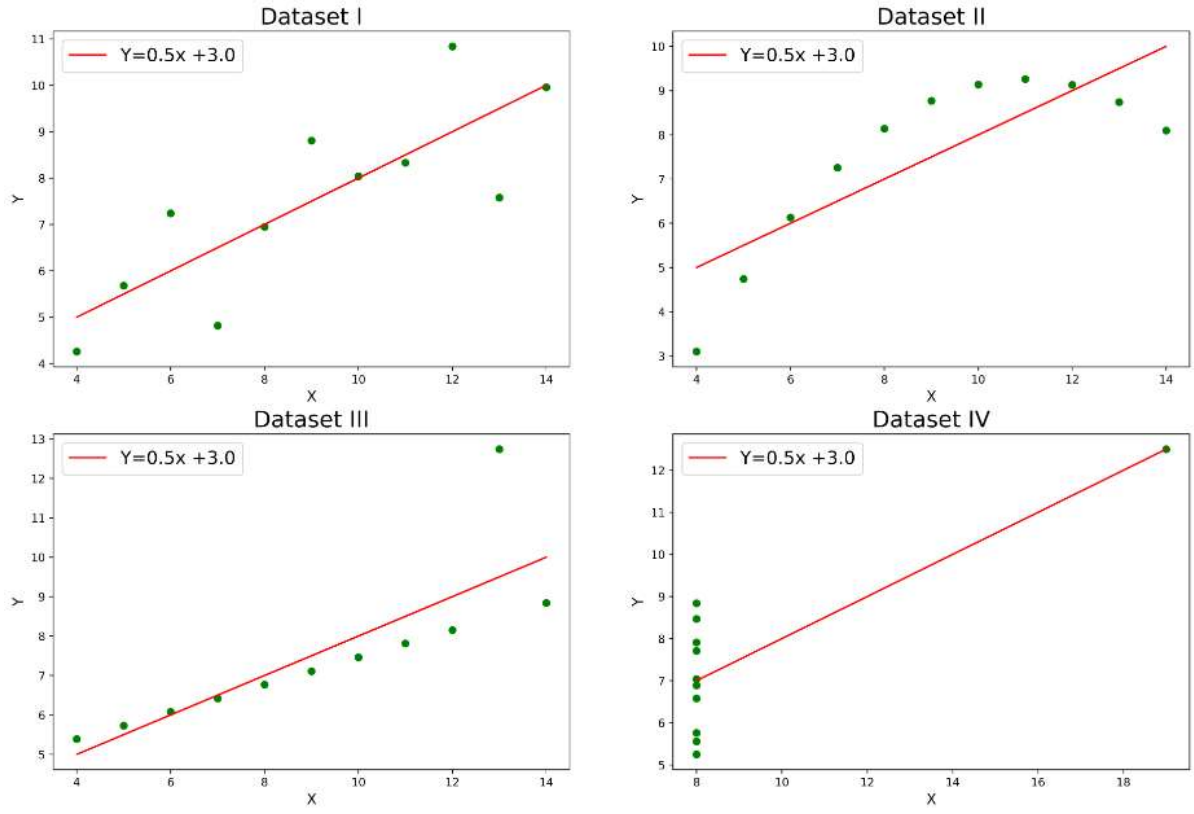
Linear regression uses techniques like Ordinary Least Squares (OLS) to find best values of coefficients.

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 set of four datasets that demonstrate importance of visualizing data.

Each dataset has nearly identical statistical properties, including the same mean, variance, correlation, and regression line.

	I	II	III	IV
Mean_x	9.000000	9.000000	9.000000	9.000000
Variance_x	11.000000	11.000000	11.000000	11.000000
Mean_y	7.500909	7.500909	7.500000	7.500909
Variance_y	4.127269	4.127629	4.122620	4.123249
Correlation	0.816421	0.816237	0.816287	0.816521
Linear Regression slope	0.500091	0.500000	0.499727	0.499909
Linear Regression intercept	3.000091	3.000909	3.002455	3.001727



However, their scatter plots reveal distinct patterns: one is linear, another is nonlinear, a third shows an outlier influencing the trend, and last has a vertical cluster with an influential point.

This illustrates how relying solely on summary statistics can be misleading, as different datasets

with similar metrics can exhibit vastly different distributions and relationships. Visualization ensures a better understanding of the data's true characteristics.

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)

In linear regression, Pearson's R (Pearson correlation coefficient), quantifies strength and direction of linear relationship between two continuous variables.
It ranges from -1 to +1:
* +1 indicates a perfect positive linear relationship.
* -1 indicates a perfect negative linear relationship.
* 0 means no linear relationship.

In context of linear regression, Pearson's R is closely related to the slope of regression line; a higher absolute value of R signifies a steeper slope, indicating a stronger linear association.
Below is a formula for calculating the Pearson correlation coefficient (r):

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

Additionally, Square of Pearson's R, known as coefficient of determination or R^2, represents proportion of variance in dependent variable that is predictable from independent variable.

Relationship underscores importance of Pearson's R in assessing how well regression model explains variability of the data.

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 adjusts range of independent variables (features) to ensure they contribute equally to the model.
This is crucial when features have different units or magnitudes, as unscaled data can lead to biased coefficient estimates and affect model performance.

Scaling is performed because =>
* **Improved Model Performance:** Scaling ensures that features with larger ranges don't dominate the regression coefficients, leading to a more balanced model.
* **Faster Convergence:** Algorithms like gradient descent converge more quickly when features are on similar scales.
* **Interpretability:** Scaled data allows for more meaningful comparisons between coefficients, aiding in understanding feature importance.

Normalized Scaling vs. Standardized Scaling
Normalization (Min-Max Scaling): This technique rescales features to a specific range, typically [0, 1]. It's useful when the data doesn't follow a Gaussian distribution and is sensitive to outliers.
The formula is:

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Standardization (Z-score Scaling): This method transforms features to have a mean of 0 and a standard deviation of 1, centering the data around zero.
It's effective when the data follows a normal distribution and is less sensitive to outliers.
The formula is:

$$z = \frac{x - \mu}{\sigma}$$

$\mu =$ Mean
$\sigma =$ Standard Deviation

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) measures how much the variance of a regression coefficient is inflated due to collinearity with other predictors.
A VIF value of infinity occurs when a predictor variable is perfectly or nearly perfectly correlated with one or more other predictors in the model.

Mathematically, VIF for a variable is calculated as:

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

Where R_i^2 is coefficient of determination from regressing x_i on the other predictors.

When R_i^2 approaches 1 (i.e., perfect correlation), denominator becomes nearly zero, causing VIF to approach infinity.

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 compare the distribution of a dataset to theoretical distribution, typically normal distribution.

It plots quantiles of observed data against quantiles of theoretical distribution. If data follows theoretical distribution, points will lie along a straight line. If there are deviations from straight line,

it indicates that data does not follow expected distribution.

In the context of linear regression, a Q-Q plot is primarily used to assess assumption that residuals (the differences between observed and predicted values) are normally distributed. This is important because one of the key assumptions in linear regression is that the residuals should follow a normal distribution for valid hypothesis testing and reliable coefficient estimates.

Importance in Linear Regression:

Normality of Residuals: A Q-Q plot helps to visually assess if residuals are normally distributed. If points on plot lie close to a straight line, it indicates that residuals are approximately normal, which is a crucial assumption for valid inference in regression.

Model Validity: Deviations from normality (e.g., skewness or heavy tails) can signal model issues, such as incorrect functional form or outliers, which can lead to misleading conclusions or unreliable statistical tests.

Outlier Detection: Q-Q plots also help identify outliers, as points far from line suggest data points that deviate significantly from expected distribution.
