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)**

1. Season: Count increases in fall and summer while it decreases in winter and it's the least in spring.
2. Month: This follows the same trend as that of seasons.
3. Holiday: Count is more during non-holidays, this has a negative impact.
4. Day of the Week: This is the only categorical variable which has almost negligible impact on the dependent variable.
5. Working Day: This has minor impact, count is a bit lower on the non working days.
6. Weather: This also has an impact, during snow, the count decreases while its the max when the weather is clear.

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 when creating dummy variables is essential to avoid the "dummy variable trap.".Its important so that we can avoid having a variable with a high VIF and correlation with other vars and rather than removing it in the end using either RFE or manually, we should remove it before hand.

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)**

atemp(0.65) or temp(0.64) are highly correlated with the 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)**

The adjusted R squared is 0.829 and the p score is 0 or close to 0 for all the independent variables. Normality of error terms is found. Y_pred and y_test were close which can be seen by residuals plot.(it has a normal distribution).

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)**
1. Temp = 0.440882
2. Snow = -0.255793, ie. negative correlation.
3. Year = 0.231039

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 algorithm that models the relationship between a dependent variable (target) and one or more independent variables (predictors) by fitting a linear equation to observed data. Here's a detailed breakdown of how the algorithm works:

Formulating the Linear Equation:

In simple linear regression (with one predictor), the relationship between the predictor

X X and the target

Y Y is represented as:

Y = β0 + β1*X + ε

β0 is the intercept (the value of Y Y when X X is 0),

β1 is the slope (the change in Y Y for each unit increase in X X),

ε is the error term, representing the difference between the observed and predicted values.

For multiple linear regression (with multiple predictors), the equation expands to:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n * X_n + \varepsilon$

Here, X1, X2, ..., Xn are different predictors.

Objective of Linear Regression:

The goal of linear regression is to find the values of β0, β1, ..., βn that minimize the difference between the predicted values and the actual values of Y. This difference is often measured using a cost function called Mean Squared Error (MSE), which calculates the average of the squared residuals (differences between actual and predicted values).

Optimization with Least Squares:

Linear regression typically uses the Ordinary Least Squares (OLS) method to find the best-fit line. OLS minimizes the sum of the squared residuals to find the optimal values for β. Mathematically, it minimizes:

$MSE = (1/N) * \Sigma (Yi - \hat{Y}i)^2$

where N is the number of observations, Yi are the actual values, and $\hat{Y}i$ are the predicted values based on the linear equation.

Assumptions of Linear Regression:

To ensure accurate results, linear regression makes several key assumptions:

Linearity: The relationship between predictors and the target variable is linear.

Independence: The residuals (errors) are independent.

Homoscedasticity: The residuals have constant variance across all levels of the predictor.

Normality: The residuals are normally distributed, especially important for inference.

Evaluating the Model:

After fitting the model, its effectiveness is evaluated using metrics like R-squared (indicating how much of the variance in Y is explained by the predictors) and Adjusted R-squared (which adjusts for the number of predictors in the model). Other error metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) provide further insights into the accuracy of the model.

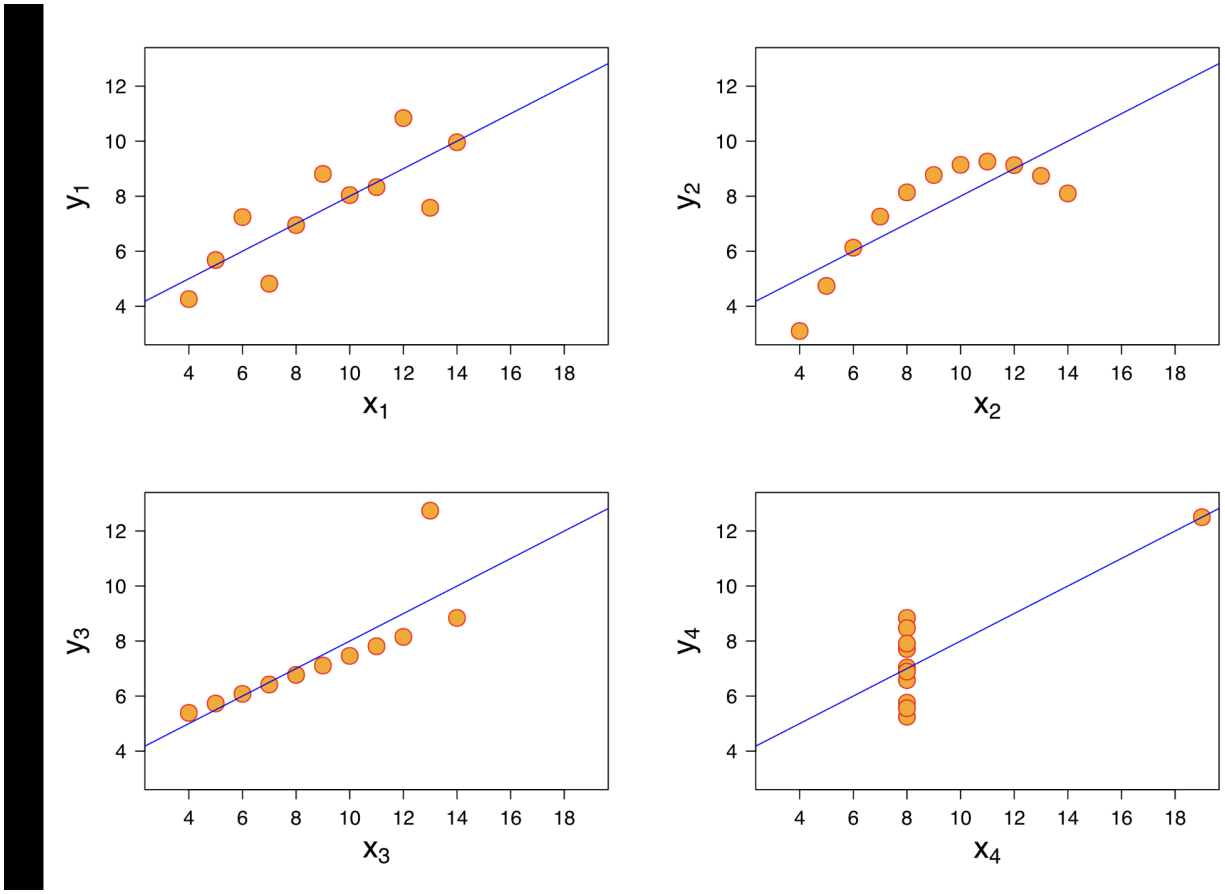
<Your answer for Question 6 goes here>

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 appear similar in basic statistical properties but differ greatly when visualized. Created by statistician Francis Anscombe in 1973, the quartet demonstrates the importance of visualizing data rather than relying solely on summary statistics. Here's a breakdown:

Statistical Properties:

All four datasets in Anscombe's quartet have nearly identical summary statistics:
Mean and variance of the x and y values,
Correlation coefficient between x and y (around 0.82),
Linear regression line with a similar slope and intercept.
Despite these identical metrics, each dataset shows a distinctly different pattern when plotted.
Different Patterns in Data:

Dataset 1: Represents a typical linear relationship, fitting the linear regression model well.
Dataset 2: Appears as a clear nonlinear (curved) relationship, which linear regression fails to capture accurately.
Dataset 3: Contains a linear relationship with one outlier, which heavily influences the regression line, showing how outliers can distort interpretations.
Dataset 4: Shows nearly all points at the same x-value except one, creating an illusion of correlation that does



n’t represent the dataset accurately.
Key Insights:

Anscombe's quartet emphasizes the limitations of summary statistics, underscoring the need to visualize data to detect patterns, outliers, or unusual distributions.
It highlights the role of exploratory data analysis (EDA) in uncovering data characteristics that statistical summaries might overlook, ensuring more accurate modeling and decision-making.
Anscombe’s quartet serves as a powerful reminder that visual context is essential in data analysis, complementing statistical insights to provide a fuller picture.

'he *x* values are the same for the first three datasets.[1]

Anscombe's quartet

Dataset I		Dataset II		Dataset III		Dataset IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

<Your answer for Question 7 goes here>

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 R, or the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two

continuous variables. It is a widely used method in data analysis to assess how strongly two variables are linearly related.

Key Points:
Range: The value of Pearson's R ranges from -1 to +1:

+1 indicates a perfect positive linear correlation, where increases in one variable consistently lead to proportional increases in the other.
-1 indicates a perfect negative linear correlation, meaning an increase in one variable results in a proportional decrease in the other.
0 indicates no linear correlation between the variables.
Interpretation:

Values near +1 suggest a strong positive relationship; as one variable increases, the other tends to increase.
Values near -1 indicate a strong negative relationship; as one variable increases, the other tends to decrease.
Values around 0 suggest little to no linear relationship.
Formula:

Pearson's R is calculated as:
$R = \Sigma((Xi - \bar{X})(Yi - \bar{Y})) / \sqrt{(\Sigma(Xi - \bar{X})^2 * \Sigma(Yi - \bar{Y})^2)}$
where Xi and Yi are individual values of variables X and Y,
$\bar{X}$ and $\bar{Y}$ are the means of X and Y, respectively.
The formula essentially standardizes the covariance of the variables, ensuring R remains between -1 and +1.
Limitations:

Linearity: Pearson's R only captures linear relationships and doesn't work well with nonlinear data.
Sensitivity to Outliers: Outliers can distort the correlation, making it appear stronger or weaker than it truly is.
Pearson's R provides a quick way to quantify relationships, but it's best to use it alongside visual tools like scatter plots to confirm linearity and check for outliers.

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 is the process of adjusting the range and distribution of features in a dataset. This is often necessary in machine learning and data preprocessing, especially when features have different units or vastly different ranges.

Why Scaling is Performed:
Improving Model Performance: Many algorithms, such as those using distance metrics (e.g., K-Nearest Neighbors, SVM) or gradient-based optimization (e.g., neural networks), work better and converge faster when features are on a similar scale.
Ensuring Balanced Feature Contribution: Scaling prevents features with larger ranges from disproportionately influencing the model, allowing all features to contribute meaningfully.
Types of Scaling:
Normalized Scaling:

Definition: Also known as Min-Max scaling, normalization transforms values to a fixed range, typically [0, 1].
Formula: $X_{norm} = (X - X_{min}) / (X_{max} - X_{min})$, where X_{min} and X_{max} are the minimum and maximum values of the feature.
Use Cases: Effective when the data distribution is not Gaussian or when features have a limited range, such as pixel intensities or ratings.
Standardized Scaling:

Definition: Standardization (or Z-score scaling) transforms data so that it has a mean of 0 and a standard deviation of 1, creating a standard normal distribution.
Formula: $X_{std} = (X - \mu) / \sigma$, where μ is the mean and σ is the standard deviation of the feature.
Use Cases: Often applied in algorithms that assume normally distributed data or are sensitive to feature distribution, such as linear regression, PCA, and logistic regression.
Key Difference:
Normalized Scaling adjusts the range between fixed minimum and maximum values, while Standardized Scaling centers the data around the mean with unit variance, preserving relative distances among data points.

<Your answer for Question 9 goes here>
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>

The Variance Inflation Factor (VIF) becomes infinite when there is perfect multicollinearity between predictor variables in a regression model. This happens because, with perfect multicollinearity, one predictor variable is an exact linear combination of another (or a combination of other predictors), resulting in a denominator of zero in the VIF formula:

VIF(i) = 1 / (1 - R(i)^2)

Here, R(i)^2 is the coefficient of determination of the i-th predictor when regressed on all other predictors. When R(i)^2 = 1, indicating perfect correlation, the denominator becomes zero, making the VIF “infinite.”

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)
<Your answer for Question 11 goes here>
A Q-Q plot is a scatterplot used in statistics to check if a dataset follows a specific theoretical distribution, like the normal distribution. In linear regression, it's crucial for evaluating the normality of residuals (differences between observed and predicted values). A straight Q-Q plot line indicates well-behaved residuals, confirming that key assumptions of linear regression are met. Deviations may prompt further investigation or the need for alternative regression methods.