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)

Equation is
cnt = 4491.30 + 1069.78 x yr + 370.72 x september -257.79 x november - 263.97 x may + 318.96 x june -344.55 x Cloudy - 367.86 x january -216.54 x february - 265.03 x december + 260.30 x july + 357.72 x winter + 365.59 x august - 454.45 light snow/rain - 391.91 sprinng

The categorical variables (months, seasons, and weather) have a significant impact on cnt. The model highlights the importance of understanding seasonal trends, weather effects, and yearly growth to optimize bike rental operations and strategies.

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)

When creating dummy variables (also known as one-hot encoding) for categorical variables, it's important to use the drop_first=True parameter in certain cases to avoid multicollinearity and to ensure our regression model works correctly.

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)
Year(yr)

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 a Linear Regression model, it's essential to validate the assumptions to ensure the model's reliability and accuracy. The key assumptions of Linear Regression are:

Linearity: The relationship between the independent variables and the dependent variable is linear.
Independence: The residuals (errors) are independent of each other.
Homoscedasticity: The variance of the residuals is constant across all levels of the independent variables.
Normality: The residuals are normally distributed.
No Multicollinearity: The independent variables should not be 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)

yr 1069.78
september 370.72
light snow/rain -454.45
spring -391.91

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 one of the simplest and most widely used statistical models. It is used to model the relationship between a dependent variable (target) and one or more independent variables (predictors). The goal of linear regression is to find the best-fit line (or hyperplane in the case of multiple features) that minimizes the difference between the predicted values and the actual values.
1. Overview of Linear Regression
In its simplest form (single variable or simple linear regression), the model predicts the dependent variable yyy as a linear function of one independent variable xxx:
y=β0+β1x+ϵy = \beta_0 + \beta_1 x + \epsilony=β0+β1x+ϵ
Where:
yyy is the dependent variable (the value you are trying to predict).
xxx is the independent variable (the feature you are using to predict yyy).
β0\beta_0β0 is the intercept (constant term).
β1\beta_1β1 is the coefficient or slope (the effect of the independent variable on the dependent variable).
ϵ\epsilonϵ is the error term or residual (the difference between the predicted and actual values).
In multiple linear regression, there are multiple independent variables, and the relationship becomes:
y=β0+β1x1+β2x2+⋯+βpxp+ϵy = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilony=β0+β1x1+β2x2+⋯+βpxp+ϵ
Where:
x1,x2,…,xpx_1, x_2, \dots, x_px1,x2,…,xp are the independent variables (features).
β1,β2,…,βp\beta_1, \beta_2, \dots, \beta_pβ1,β2,…,βp are the coefficients corresponding to each feature.
2. Objective of Linear Regression
The main objective of linear regression is to find the optimal values for the coefficients β0,β1,…,βp\beta_0, \beta_1, \dots, \beta_pβ0,β1,…,βp such that the model minimizes the difference between the predicted values and the actual values. This is typically done using Least Squares Estimation (LSE).

3. Steps Involved in Linear Regression

Step 1: Model Representation

The linear regression model can be represented as:

Y=Xβ+ϵ\mathbf{Y} = \mathbf{X} \beta + \epsilonY=Xβ+ϵ

Where:

Y\mathbf{Y}Y is the vector of observed values of the dependent variable.

X\mathbf{X}X is the matrix of independent variables (design matrix).

β\betaβ is the vector of coefficients [β0,β1,…,βp][\beta_0, \beta_1, \dots, \beta_p][β0,β1,…,βp].

ϵ\epsilonϵ is the error term (residuals).

Step 2: Least Squares Method

The primary goal is to minimize the sum of squared errors (residuals). The residual for each observation is the difference between the actual value and the predicted value:

ϵi=yi−y^i\epsilon_i = y_i - \hat{y}_iϵi=yi−y^i

Where yiy_iyi is the actual value and y^i\hat{y}_iy^i is the predicted value.

To find the best-fitting line, we minimize the sum of squared residuals (RSS):

RSS=∑i=1n(yi−y^i)2RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2RSS=i=1∑n(yi−y^i)2

The solution to this minimization problem (i.e., finding the coefficients that minimize the residual sum of squares) can be derived using the Normal Equation:

β^=(XTX)−1XTY\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}β^ =(XTX)−1XTY

Where:

β^\hat{\beta}β^ is the vector of estimated coefficients.

XT\mathbf{X}^TXT is the transpose of the design matrix.

(XTX)−1(\mathbf{X}^T \mathbf{X})^{-1}(XTX)−1 is the inverse of the matrix XTX\mathbf{X}^T \mathbf{X}XTX.

This equation gives the optimal values of the coefficients that minimize the sum of squared residuals.

Step 3: Model Fitting

After finding the optimal coefficients, the model is fit to the data, and predictions can be made. For a new observation with features x1,x2,…,xpx_1, x_2, \dots, x_px1,x2,…,xp, the predicted value is:

y^=β0^+β1^x1+β2^x2+⋯+βp^xp\hat{y} = \hat{\beta_0} + \hat{\beta_1} x_1 + \hat{\beta_2} x_2 + \dots + \hat{\beta_p} x_py^=β0^+β1^x1+β2^x2+⋯+βp^xp

4. Assumptions of Linear Regression

For the model to produce reliable results, several assumptions must hold:

Linearity: The relationship between the dependent variable and the independent variables is linear.

This is the fundamental assumption of linear regression, which is why we use a straight line (or hyperplane in the case of multiple predictors).

Independence of Errors: The residuals (errors) should be independent of each other.

This assumption is critical for the statistical tests used to evaluate the model.

Homoscedasticity: The variance of the residuals should be constant across all levels of the independent variables.

If the variance of the errors is not constant (heteroscedasticity), it can lead to inefficient estimates of the coefficients.

Normality of Errors: The residuals should follow a normal distribution.

This assumption is important for hypothesis testing and confidence intervals of the regression coefficients.

No Multicollinearity: The independent variables should not be highly correlated with each other.

Multicollinearity can make the estimates of the regression coefficients unstable.
5. Model Evaluation
After fitting the model, several metrics are used to evaluate the performance of the linear regression model:
R-squared: This is the proportion of the variance in the dependent variable that is explained by the independent variables. It ranges from 0 to 1, with 1 indicating that the model explains all the variance.
R2=1−SSresidualSStotalR^2 = 1 - \frac{SS_{residual}}{SS_{total}}R2=1−SStotalSSresidual
Where SSresidualSS_{residual}SSresidual is the sum of squared residuals, and SStotalSS_{total}SStotal is the total sum of squares (variance of the target).
Mean Squared Error (MSE): This measures the average squared difference between the actual and predicted values. The smaller the MSE, the better the model.
MSE=1n∑i=1n(yi−y^i)2MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2MSE=n1i=1∑n(yi−y^i)2
Root Mean Squared Error (RMSE): This is the square root of MSE and represents the average error in the units of the dependent variable.
p-value: In hypothesis testing, p-values for each coefficient are used to determine if the corresponding feature significantly contributes to the model.
6. Advantages and Disadvantages of Linear Regression
Advantages:
Simple to implement: Linear regression is one of the easiest models to understand and implement.
Interpretability: The model coefficients have straightforward interpretations.
Efficiency: Linear regression is computationally efficient and fast to train on large datasets.
Disadvantages:
Assumptions: Linear regression makes several assumptions (linearity, normality, independence, etc.) that might not hold in real-world data, potentially leading to biased or incorrect results.
Sensitive to outliers: Linear regression is sensitive to outliers, which can disproportionately affect the model's predictions and coefficients.
Limited to linear relationships: If the true relationship between the features and target is non-linear, linear regression will not perform well.
7. Extensions of Linear Regression
Multiple Linear Regression: When there are more than one independent variable, linear regression extends to multiple linear regression.
Ridge and Lasso Regression: These are regularization techniques to prevent overfitting by adding a penalty term to the cost function.
Polynomial Regression: This extends linear regression by including polynomial terms (e.g., x2,x3x^2, x^3x2,x3) to model non-linear relationships.
>

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 created by the statistician **Francis Anscombe** in 1973. These datasets are famous for demonstrating how datasets with nearly identical statistical properties (e.g., mean, variance, correlation, and regression line) can exhibit vastly different

patterns when visualized.
The quartet underscores the importance of visualizing data rather than relying solely on summary statistics.

Properties of Anscombe's Quartet
Each of the four datasets in Anscombe's Quartet has the following nearly identical statistical properties:
Mean of x and y:
Mean of x=9\text{Mean of } x = 9Mean of x=9
Mean of y=7.50\text{Mean of } y = 7.50Mean of y=7.50
Variance of x and y:
Variance of x=11\text{Variance of } x = 11Variance of x=11
Variance of y=4.12\text{Variance of } y = 4.12Variance of y=4.12
Correlation between x and y:
r=0.816r = 0.816r=0.816
Regression line:
The regression equation for all datasets is approximately: y=3+0.5xy = 3 + 0.5xy=3+0.5x
Despite these similar numerical summaries, when the datasets are plotted, their differences become clear.
The Four Datasets in Anscombe's Quartet
Dataset 1:
A typical dataset where the points form a linear relationship.
The data aligns closely with the regression line.
Dataset 2:
A dataset with a perfect quadratic relationship.
While the linear regression equation is the same, the data clearly follows a curve.
Dataset 3:
A dataset where all data points except one form a perfect linear relationship.
A single outlier significantly affects the regression line and correlation.
Dataset 4:
A dataset where almost all points have the same xxx-value except for one.
The regression line is heavily influenced by this single influential data point.

Lessons from Anscombe's Quartet
The Importance of Visualization:
Summary statistics like mean, variance, and correlation can be misleading.
Visualization reveals the underlying patterns, relationships, and anomalies in the data.
Identifying Outliers and Influential Points:
Outliers and influential points can disproportionately affect regression lines and correlation coefficients.
Plotting the data can help identify these points.
Context in Statistical Analysis:
Data analysis must consider the context and structure of the data rather than relying solely on numbers.
Understanding the real-world implications of patterns is crucial.
Misleading Models:
The regression line might fit well statistically but fail to capture the true relationship in the data.
>

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)

< The Pearson's R (also known as Pearson's correlation coefficients) measures the strength between the different variables and the relation with each other. The Pearon's R returns values between -1 and 1. The interpretation of the coefficients are:
-1 coefficient indicates strong inversely proportional relationship.
0 coefficient indicates no relationship.
1 coefficient indicates strong proportional relationship.

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

Where:
N = the number of pairs of scores
Σxy = the sum of the products of paired scores
Σx = the sum of x scores
Σy = the sum of y scores
Σx2 = the sum of squared x scores
Σy2 = the sum of squared y scores
>

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)

< What - The scaling is the data preparation step for regression model. The scaling normalizes these varied datatypes to a particular data range.
Why – Most of the times the feature data is collected at public domains where the interpretation of variables and units of those variables are kept open collect as much as possible. This results in to the high variance in units and ranges of data. If scaling is not done on these data sets, then the chances of processing the data without the appropriate unit conversion are high. Also the higher the range then higher the possibility that the coefficients are impaired to compare the dependent variable variance.
The scaling only affects the coefficients. The prediction and precision of prediction stays unaffected after scaling.
Normalization/Min-Max scaling – The Min max scaling normalizes the data within the range of 0 and 1. The Min max scaling helps to normalize the outliers as well.

$$\boldsymbol{MinMaxScaling}: \boldsymbol{x} = \frac{\boldsymbol{x} - \mathbf{min}(\boldsymbol{x})}{\mathbf{max}(\boldsymbol{x}) - \mathbf{min}(\boldsymbol{x})}$$

Standardization converges all the data points into a standard normal distribution where mean is 0 and standard deviation is 1.

$Standardization: x = \frac{x - mean(x)}{sd(x)}$ >

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)

< $VIF = \frac{1}{1-R^2}$
The VIF formula clearly signifies when the VIF will be infinite. If the R2 is 1 then the VIF is infinite. The reason for R2 to be 1 is that there is a perfect correlation between 2 independent 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)

< Q-Q plots are the quantile-quantile plots. It is a graphical tool to assess the 2 data sets are from common distribution. The theoretical distributions could be of type normal, exponential or uniform. The Q-Q plots are useful in the linear regression to identify the train data set and test data set are from the populations with same distributions. This is another method to check the normal distribution of the data sets in a straight line with patterns explained below
Interpretations
Similar distribution: If all the data points of quantile are lying around the straight line at an angle of 45 degree from x-axis.
Y values < X values: If y-values quantiles are lower than x-values quantiles.
X values < Y values: If x-values quantiles are lower than y-values quantiles.
Different distributions – If all the data points are lying away from the straight line.
Advantages
Distribution aspects like loc, scale shifts, symmetry changes and the outliers all can be daintified from the single plot.
The plot has a provision to mention the sample size as well >
