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. dependent variable cnt is comparatively less in spring season, there is not much difference among the medians for other three seasons
2. dependent variable cnt is comparatively higher in 2019 as it has become more popular compared to 2018
3. dependent variable cnt is comparatively high in month 9(September) and 10(October). bike usage is comparatively less in month 1,2,11,12.
4. dependent variable cnt is high in non-holiday days
5. median is almost same for all the weekdays. For workingday column, the difference between median is not much between 0 and 1.
6. people usually using the bikes during weathersit 1(Clear, Few clouds, Partly cloudy, Partly cloudy) and 2(Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist)

and very less in 3(Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) and none in 4 (Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog)

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)

To encode categorical data, one hot encoding is done, where a dummy variable is to be created for each discrete categorical variable for a feature. This can be done by using pandas.get_dummies() which will return dummy-coded data. Here we use parameter drop_first = True, this will drop the first dummy variable, thus it will give n-1 dummies out of n discrete categorical levels by removing the first level.

If we do not use drop_first = True, then n dummy variables will be created, and these predictors(n dummy variables) are themselves correlated which is known as multicollinearity and it, in turn, leads to Dummy Variable Trap.

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)

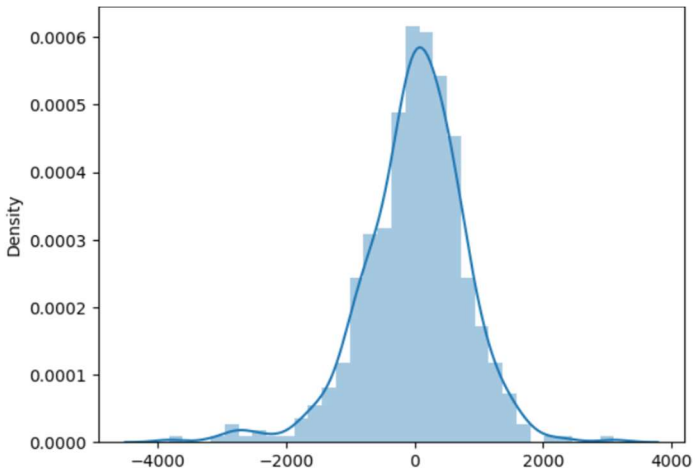
Temp has the highest correlation with the target variable.

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)

Performed residual analysis to validate the assumptions of Linear Regression.
Steps followed:
1. y_train_pred = lm.predict(X_train_rfe)
2. res = y_train - y_train_pred
   sns.distplot(res)



So, from the above plot it is evident that the distribution of the residual values/errors is normal and centered around 0

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)

we can see that equation of best fitted line is:
cnt = 2553 + 2053 x yr -900 x holiday + 3369 x temp - 1262 x windspeed - 918 x spring + 374 x winter - 432 x January + 594 x September - 431 x Sunday - 2558 x Light Snow - 680 x Mist

Top 3 features: yr, temp and September

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 learning algorithm that models the relationship between a dependent variable (target) and one or more independent variables (features) using a linear equation. It is one of the simplest and most widely used regression techniques, used primarily to predict continuous values. Here's an in-depth look at how it works:

1. Objective of Linear Regression
The goal of linear regression is to find a line that best fits the data. This line (or hyperplane in

higher dimensions) is called the regression line, which minimizes the error between the predicted values and the actual values. In mathematical terms, this line aims to minimize the sum of squared errors between the actual and predicted values.

2. Equation of the Regression Line

For a simple linear regression (one independent variable), the relationship between the dependent and independent variable can be represented as:

y = mx + b

where:

y is the dependent variable (target or output).

x is the independent variable (feature or input).

m or theta_1 is the slope of the line, representing the weight assigned to x.

b or theta_0 is the y-intercept, representing the value of y when x = 0.

In multiple linear regression (more than one independent variable), this equation generalizes to:

y = theta_0 + theta_1 * x1 + theta_2 * x2 + …… + theta_n * xn

where n is the number of features, and each theta (parameter) represents the weight associated with each feature.

3. Cost Function

The cost function, often referred to as the Mean Squared Error (MSE), quantifies the error between the predicted and actual values. It's calculated as:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x_i) - y_i)^2$$

where:

- m is the number of data points.
- $h_\theta(x_i)$ is the predicted value for the i-th data point.
- y_i is the actual value for the i-th data point.
- The factor $\frac{1}{2}$ is added for convenience during differentiation.

The cost function's goal is to minimize the sum of squared differences between the predicted and actual values, ensuring the line fits the data as closely as possible.

4. Gradient Descent Optimization

To find the optimal values of theta_0 and theta_1 (in simple linear regression) or theta parameters in general, linear regression often uses Gradient Descent. This optimization algorithm iteratively adjusts the parameters to reduce the cost function:

Step 1: Start with initial guesses for the parameters (often set to 0).

Step 2: Calculate the cost function with these parameters.

Step 3: Update each parameter using the gradient of the cost function:

$$\theta_j = \theta_j - \alpha \cdot \frac{\partial J(\theta)}{\partial \theta_j}$$

where:

- α is the learning rate, controlling how big a step is taken in each iteration.
- $\frac{\partial J(\theta)}{\partial \theta_j}$ is the partial derivative of the cost function with respect to θ_j.

Step 4: Repeat steps 2 and 3 until the cost function converges (i.e., changes minimally between iterations).

5. Assumptions of Linear Regression

For linear regression to perform effectively, a few assumptions should be met:

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

Independence: Observations are independent of each other.

Homoscedasticity: The variance of error terms is constant across all values of the independent variables.

Normality: The residuals (differences between observed and predicted values) are normally distributed, especially important for hypothesis testing.

6. Metrics to Evaluate Linear Regression

The quality of a linear regression model can be assessed using various metrics:

Mean Absolute Error (MAE): The average absolute difference between actual and predicted values.

Mean Squared Error (MSE): The average of squared differences between actual and predicted values.

Root Mean Squared Error (RMSE): The square root of MSE, providing an error metric in the same units as the target variable.

R2_Score: Also known as the coefficient of determination, it indicates how well the independent variables explain the variance in the dependent variable, where 1 means perfect prediction and 0 means no predictive power.

7. Regularization in Linear Regression

Regularization techniques like Lasso (L1 regularization) and Ridge (L2 regularization) are sometimes applied to linear regression to avoid overfitting, especially with high-dimensional data. These methods penalize large coefficients in the model:

Lasso (L1 regularization): Adds a penalty proportional to the absolute values of the coefficients.

Ridge (L2 regularization): Adds a penalty proportional to the square of the 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 collection of four datasets created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data before analyzing it solely through summary statistics. Each dataset in the quartet has nearly identical summary statistics (mean, variance, correlation, and linear regression line), but when plotted, the datasets show distinctly different

patterns. This highlights how relying only on summary statistics can be misleading without visual inspection of the data.

Overview of Anscombe's Quartet

The four datasets in Anscombe's quartet consist of an x variable and a y variable, structured as follows:

1. Dataset 1: A straightforward linear relationship.
2. Dataset 2: A nonlinear (quadratic) relationship.
3. Dataset 3: A linear relationship with an outlier affecting the regression.
4. Dataset 4: A dataset with nearly constant x values but with an outlier impacting the statistics.

Each dataset has approximately the same:

- Mean of x and y.
- Variance of x and y.
- Correlation between x and y.
- Linear regression line equation (almost identical slope and intercept).

Detailed Explanation of Each Dataset

Here's a breakdown of each dataset and what its plot reveals:

Dataset 1

Description: A classic, near-perfect linear relationship between x and y.

Plot: The data points form a straight line, with points clustered closely around the regression line.

Insights: This dataset represents a well-behaved linear relationship, with the summary statistics accurately reflecting the linear trend.

Dataset 2

Description: A nonlinear, parabolic (or quadratic) relationship between x and y.

Plot: When plotted, the data points show a clear curve rather than a straight line.

Insights: The correlation and linear regression line do not capture the nonlinearity, demonstrating that relying only on these metrics would incorrectly suggest a linear relationship.

Dataset 3

Description: A linear relationship, but with one outlier.

Plot: Most points lie close to the regression line, but an outlier significantly deviates from this trend.

Insights: The outlier heavily influences the regression line, skewing the summary statistics. Without plotting, one would miss the presence and impact of the outlier.

Dataset 4

Description: A dataset where x values are nearly identical, with one point that acts as an outlier in y.

Plot: The points are essentially in a vertical line, but one outlier pushes the regression line to fit the single point rather than the data distribution.

Insights: This dataset shows that with limited variability in x, even a single outlier can drastically affect the regression, yielding misleading interpretations.

Key Lessons from Anscombe's Quartet

1. Visualize Data: Visualizing data provides insight that summary statistics alone cannot capture.

Patterns, clusters, trends, and outliers are more easily identified when data is plotted.

2. Limitations of Summary Statistics: Summary statistics (mean, variance, correlation, regression) can sometimes mask the actual data distribution. Different datasets can yield similar statistics despite having very different structures.

3. Influence of Outliers: Outliers can disproportionately affect summary statistics, leading to misleading interpretations. Identifying outliers often requires plotting the data.

4. Nonlinear Relationships: Not all data relationships are linear; many real-world datasets exhibit nonlinear patterns that a linear regression line cannot adequately model.

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 the Pearson correlation coefficient, is a measure of the linear relationship between two variables. It quantifies the strength and direction of a linear association between these variables, commonly denoted as x and y. Pearson's r ranges from -1 to 1, where:

+1 indicates a perfect positive linear relationship (as one variable increases, the other also increases proportionally).
-1 indicates a perfect negative linear relationship (as one variable increases, the other decreases proportionally).
0 suggests no linear relationship (the variables are uncorrelated).

Formula for Pearson's r

The formula for calculating Pearson's r is:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

where:

- x_i and y_i are the individual data points of variables x and y,
- $\bar{x}$ and $\bar{y}$ are the mean values of x and y,
- The numerator calculates the covariance between x and y,
- The denominator normalizes this covariance by the standard deviations of x and y, producing a value between -1 and 1.

Interpreting Pearson's r
The absolute value of r indicates the strength of the linear relationship:

|r| = 0.0 - 0.2: Very weak or no correlation.
|r| = 0.2 - 0.4: Weak correlation.
|r| = 0.4 - 0.6: Moderate correlation.

|r| = 0.6 - 0.8: Strong correlation.
|r| = 0.8 - 1.0: Very strong correlation.
Limitations of Pearson's r
Sensitivity to Outliers: Outliers can heavily influence Pearson's r by skewing the correlation.
Only Linear Relationships: Pearson's r only measures linear associations. It does not capture nonlinear relationships.
Assumes Normal Distribution: Pearson's correlation works best when both variables are normally distributed. Non-normal distributions may lead to inaccurate interpretations of r.

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 a data preprocessing technique used in machine learning and data analysis to adjust the values of numeric features so they are on a comparable scale. By scaling, we can ensure that one feature does not disproportionately influence the model due to its larger range or unit differences.

Why Scaling is Performed:
1. Improves Model Convergence: Models often converge faster when the input data is scaled, as the model can update parameters more evenly across features.
2. Reduces Bias: If some features have significantly larger values, models might weigh these more heavily, which could lead to biased results.
3. Enhances Interpretability: Some algorithms interpret data points differently based on their distance (like clustering), so scaling helps provide a more realistic measure of similarity.

Normalized Scaling vs. Standardized Scaling:
- Normalization (Min-Max Scaling):
  - Transforms features to a fixed range, typically [0, 1].
  - Formula:

$$X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

  - Best for: Data without outliers, where a fixed range is desirable (e.g., image pixel intensities).

- Standardization (Z-Score Scaling):
  - Centers features around the mean and scales them by the standard deviation, resulting in a mean of 0 and standard deviation of 1.
  - Formula:

$$X_{\text{std}} = \frac{X - \mu}{\sigma}$$ where μ is the mean, and σ is the standard deviation.

  - Best for: Data with outliers or where Gaussian distribution is assumed, as it preserves the influence of extreme values.

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 Variance Inflation Factor (VIF) becomes infinite when there is perfect multicollinearity between predictor variables in a regression model. This situation occurs when one variable can be expressed as an exact linear combination of one or more other variables.

Understanding VIF: VIF quantifies how much the variance of a regression coefficient is inflated due to multicollinearity with other predictors. Mathematically, for a predictor Xj , VIF is calculated as,

$$VIFj = 1/(1 - Rj\ Square)$$

Where Rj square is the R square value obtained by regressing Xj on all the predictors.

Why VIF Becomes Infinite:
When there is perfect multicollinearity (i.e., Rj square = 1) , the denominator (1 – Rjsquare) becomes zero. This division by zero leads to an infinite VIF, signaling that the predictor Xj is perfectly predicted by other predictors in the model.
Infinite VIF values indicate that the model cannot distinguish the unique contribution of the multicollinear predictor, making it impossible to estimate reliable coefficients.

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 (Quantile-Quantile) plot is a graphical tool used to assess if a dataset follows a particular theoretical distribution, usually the normal distribution. In the Q-Q plot, the quantiles of the sample data are plotted against the quantiles of the theoretical distribution. If the data closely follows the distribution, the points in the Q-Q plot will approximately form a straight line.

Use and Importance of a Q-Q Plot in Linear Regression:
In linear regression, a Q-Q plot is used to check the normality assumption of the residuals, which is important for several reasons:

1. Assessing the Normality Assumption:
- Linear regression assumes that the residuals (errors) are normally distributed, especially for hypothesis testing and confidence intervals to be valid.
- Non-normal residuals can affect the reliability of t-tests, F-tests, and confidence intervals. A Q-Q plot helps visually assess this assumption by showing how closely the residuals follow a normal distribution.

2. Detecting Skewness and Heavy Tails:

- If the Q-Q plot shows a curved pattern, it may suggest skewness, while points that deviate towards the tails could indicate heavy-tailed distributions.
- Skewness or heavy tails can imply that the model may not be capturing the data structure well or that certain transformations might improve model performance.

3. Identifying Outliers:
- Points far from the 45-degree line in the Q-Q plot indicate outliers or points that do not fit the assumed normal distribution.
- Identifying these outliers can help decide if the model needs adjustment, such as handling influential points separately.
