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 datasets are month,yr, weather situation if the weather is mild or we can say pleasant it leads to increase in rental of bikes.

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)

It is important because it avoid perfect multicolinearity, improves interpretability and reduces complexity of system

---

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)

In Pairplot cnt and registered users are highly related due to registered is subpart of 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)

1. linerity- i checked this from scatter plot before training dataset because it remains same after or before
2. Multicolinearity-There's no multicolinearity in dataset we can check through vif
3. Homoscedasticity (Constant Variance)-we'll check this
4.Normality of errors- checked by plotting histplot which should be normal or approx. normal distribution 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)

Top 3 features are below mentioned
1. Casual
2. Year hence which is showing it's growing market when lockdown will remove and situation will normalized people will use them definitely
3. Windspeed

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)

<Your answer for Question 6 goes here>

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables.

Simple Linear Regression: In simple linear regression, we have one independent variable (X) and one dependent variable (Y). The model can be represented as:

$Y = \beta_0 + \beta_1 X + \varepsilon$

Where:

Y: Dependent variable

X: Independent variable

β_0: Intercept

β_1: Slope

ε: Error term

The goal is to find the best values for β_0 and β_1 that minimize the error between the predicted values and the actual values. This is often done using the least squares method.

Multiple Linear Regression: Multiple linear regression extends the concept to multiple independent variables:

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

Where:

$X_1, X_2, ..., X_n$: Multiple independent variables

The Least Squares Method: The least squares method is a common technique used to find the best-fitting line. It minimizes the sum of the squared residuals, which are the differences between the observed values and the predicted values.

Steps Involved:

1. **Data Collection:** Gather data for the dependent and independent variables.
2. **Data Preparation:** Clean the data, handle missing values, and normalize or standardize the features if necessary.
3. **Model Training:**
    - **Split the data:** Divide the data into training and testing sets.
    - **Fit the model:** Use the training set to estimate the coefficients β_o and β_1 (or β_o, β_1, ..., β_n in multiple linear regression).
4. **Model Evaluation:**
    - **Make predictions:** Use the trained model to predict the dependent variable for the testing set.
    - **Evaluate performance:** Calculate metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared to assess the model's accuracy.

Key Assumptions of Linear Regression:

1. **Linearity:** The relationship between the dependent and independent variables should be linear.
2. **Independence of Errors:** The errors should be independent of each other.
3. **Homoscedasticity:** The variance of the errors should be constant across all values of the independent variable(s).
4. **Normality of Errors:** The errors should be normally distributed.

Applications of Linear Regression:

- **Predicting house prices:** Based on factors like square footage, number of bedrooms, and location.
- **Sales forecasting:** Predicting future sales based on historical data and other factors.
- **Financial modeling:** Analyzing stock prices or predicting economic trends.
- **Medical research:** Modeling the relationship between drug dosage and patient response.

By understanding the principles of linear regression and its assumptions, we can effectively apply it to various real-world problems.

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)

<Your answer for Question 7 goes here>

Anscombe's Quartet is a set of four datasets, each containing 11 (x, y) pairs of data points. These datasets were created by statistician Francis Anscombe in 1973 to demonstrate the importance of data visualization and the limitations of relying solely on summary statistics.

Key Characteristics of Anscombe's Quartet:

Despite having nearly identical summary statistics, including mean, variance, correlation, and linear regression line, the four datasets have vastly different visual representations when plotted.

- **Dataset I:** This dataset shows a clear linear relationship between x and y, making it suitable for linear regression.
- **Dataset II:** This dataset shows a quadratic relationship, indicating that linear regression might not be the best fit.
- **Dataset III:** This dataset is almost linear, except for one outlier point that significantly influences the regression line.
- **Dataset IV:** This dataset has a constant x-value for all but one point, making the regression line highly sensitive to this outlier.

The Importance of Data Visualization:

Anscombe's Quartet highlights the importance of visualizing data before drawing conclusions. While summary statistics can provide a quantitative overview, they may not capture the underlying patterns and anomalies in the data. By visualizing the data, we can gain valuable insights that might be missed by relying solely on numerical summaries.

Key Takeaways:

- **Data visualization is essential:** Always plot your data to understand its underlying patterns and anomalies.
- **Summary statistics can be misleading:** Don't rely solely on summary statistics to draw conclusions.
- **Outliers can significantly impact regression analysis:** Identify and handle outliers appropriately.
- **The choice of statistical model depends on the data:** Select a model that best fits the underlying relationship between the variables.

By understanding the lessons from Anscombe's Quartet, we can make more informed decisions in data analysis and avoid drawing incorrect conclusions.

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 is a statistical measure that quantifies the linear relationship between two variables. It ranges from -1 to +1:

- -1: Perfect negative correlation (as one variable increases, the other decreases)
- 0: No correlation (no linear relationship)
- +1: Perfect positive correlation (as one variable increases, the other also increases)

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)

<Your answer for Question 9 goes here>

Scaling is a data preprocessing technique that involves transforming numerical features to a common scale.

Importance?

- **Improves Model Performance:** Many machine learning algorithms, especially distance-based algorithms like K-Nearest Neighbors (KNN) and clustering algorithms, are sensitive to feature scales. Scaling ensures that features with larger ranges don't dominate those with smaller ranges.
- **Faster Convergence:** Gradient-based optimization algorithms converge faster when features are on a similar scale.

Normalized Scaling vs. Standardized Scaling:

- **Normalized Scaling (Min-Max Scaling):**
  - Rescales features to a specific range, typically between 0 and 1.
  - Formula: x_scaled = (x - min(x)) / (max(x) - min(x))
  - Preserves original data range.
  - Sensitive to outliers.
- **Standardized Scaling (Z-score Scaling):**
  - Rescales features to have a mean of 0 and a standard deviation of 1.
  - Formula: x_scaled = (x - mean(x)) / std(x)
  - Less sensitive to outliers.
  - Can introduce negative 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)

<Your answer for Question 10 goes here>

A VIF value of infinity indicates perfect multicollinearity among the independent variables in a linear regression model.

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, or Quantile-Quantile plot, is a graphical tool used to compare two probability distributions. In the context of linear regression, it's primarily used to assess whether the residuals of a model are normally distributed.

Why Normality is Important in Linear Regression:

- **Confidence Intervals and Hypothesis Testing:** Many statistical tests, including those used to construct confidence intervals and perform hypothesis tests, rely on the assumption of normality.
- **Model Assumptions:** Linear regression models assume that the error terms are normally distributed.
- **Outlier Detection:** Non-normality can sometimes indicate the presence of outliers that may be influencing the model's performance.

How to Interpret a Q-Q Plot:

1. **Perfect Normality:** If the data is perfectly normally distributed, the points on the Q-Q plot will fall along a straight line.
2. **Positive Skewness:** If the points deviate from the line and curve upwards, the data is positively skewed.
3. **Negative Skewness:** If the points deviate from the line and curve downwards, the data is negatively skewed.
4. **Heavy Tails:** If the points deviate from the line at the tails, the data has heavier tails than a normal distribution.
5. **Light Tails:** If the points are closer to the line at the tails, the data has lighter tails than a normal distribution.

Using Q-Q Plots in Linear Regression:

1. **Check Residual Normality:**
    - Calculate the residuals (the difference between the predicted and actual values).
    - Create a Q-Q plot of the residuals.
    - If the points on the plot roughly follow a straight line, it suggests that the residuals are normally distributed.
2. **Identify Outliers:**
    - Outliers can significantly impact the normality assumption.
    - Look for points that deviate significantly from the straight line on the Q-Q plot.
3. **Consider Transformations:**
    - If the residuals are not normally distributed, you might consider transforming the response variable or independent variables. Common transformations include log, square root, or Box-Cox transformations.

In conclusion, Q-Q plots are a valuable tool for assessing the normality assumption in linear regression.