Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

- **Season (season)**:
  - Season 2 (summer) and Season 4 (winter) show higher coefficients, indicating increased bike rentals compared to other seasons (likely spring and fall). This suggests that summer and winter seasons have more favorable conditions for bike rentals.
- **Month (mnth)**:
  - Specifically, August (mnth_8) and September (mnth_9) have positive coefficients, indicating higher bike rentals during these months. This aligns with the warmer weather typically experienced in late summer and early fall.
- **Holiday (holiday)**:
  - The coefficient for the holiday variable is negative, indicating that bike rentals tend to decrease on holidays compared to non-holiday days. This could be due to people having different activities or travel patterns on holidays, reducing the demand for bike rentals.
- **Weather Situation (weathersit)**:
  - Weather situations 2 (misty conditions) and 3 (light rain/snow) show negative coefficients, indicating a decrease in bike rentals during these weather conditions. On the other hand, weather situation 1 (clear skies) likely corresponds to higher bike rentals, though this is inferred indirectly from the contrast with other weather situations.

2. Why is it important to use drop_first=True during dummy variable creation?

By providing '*drop_first=True',* the numpy library creates (n-1) dummy variables for a Categorical variable with 'n' categories. We only need (n-1) variables to represent the 'n' categories of a specific Categorical variable. This is because having a value of 0 on all the (n-1) variables can represent one of the Categories. So, if we have a total of 'n' variables, then 1 of them will be redundant.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

'temp' variables have the highest correlation with 'cnt' variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?

1. The total sum of residual errors is very close to zero, thereby confirming the assumption that the sum of residual errors should be zero

2. I created a ‘histogram’ to see the distribution of residual errors in the model. It looked like a Bell curve with the center at 0. This confirmed the assumption that the distribution of residuals should be Normal
3. Later, I plotted a scatterplot with Y and Residual. The plot shows that for larger values of 'cnt', the residual is skewed to negative; hence, it is `heteroscedastic`. This fails our assumption that the residuals should be `heteroscedastic`.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

- **yr (year)**:

  - Coefficient: 2025.0362
  - Interpretation: For every unit increase in the year (from 2018 to 2019), bike rentals increase by approximately 2025 units. This suggests a strong positive trend over time.

  (Even though year is high positive trend I don’t think it’s a very good insight for prediction)

- **weathersit (weather situation)**, specifically:

  - **weathersit_3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)**:
    - Coefficient: -2453.7785
    - Interpretation: On days with weather situation 3, bike rentals decrease by approximately 2454 units compared to clear weather (weathersit_1). This indicates a significant negative impact of adverse weather conditions on bike rental demand.

- **temp (temperature)**:

  - Coefficient: 4508.1121
  - Interpretation: For every one-degree Celsius increase in temperature, bike rentals increase by approximately 4508 units. Warmer temperatures have a strong positive effect on bike rental demand, suggesting that people are more likely to rent bikes in pleasant weather conditions.

season (season_4):

- Coefficient: 1189.2003
- Interpretation: During Season 4 (winter), bike rentals increase by approximately 1189 units compared to the base season (likely spring). This indicates that winter has a positive influence on bike rental demand, possibly due to favorable weather conditions for biking or seasonal activities.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a statistical method to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between these variables. The algorithm works as follows:

1. It finds the best-fitting straight line or hyperplane through the data points.
2. This plane is determined by minimizing the SSR
3. The result is an equation of the form: y = B0 + B1x1 + B2x2 + ... + Bnxn, where y is the dependent variable, x1, x2, ..., xn are independent variables, B0 is the y-intercept, and B1, B2, ..., Bn are the coefficients that represent the change in y for a one-unit change in the corresponding x, holding other variables constant.
4. The algorithm calculates these coefficients using methods like Gradient Descent
5. Once the model is fitted, it can be used to make predictions or interpret the relationships between variables.

2. Explain the Anscombe's quartet in detail.

These are four graphs given as an example to show the importance of visualizing data and not relying solely on summary statistics. Francis Anscombe demonstrated this in 1973.

The four graphs have very similar static measures, such as mean, median, standard deviation, and correlations, but when looked at graphs, they reveal entirely different patterns.

The main takeaways from this is that,

1. It is important to visualize data before drawing conclusions.
2. Highlights that summary statistics alone can be misleading 3. The same summary statistics can arise from very different datasets

3. What is Pearson's R?

Pearson's R is a statistical measure quantifying the linear relationship between two continuous variables. It ranges from -1 to +1, where -1 indicates a perfect negative correlation, +1 is a perfect positive correlation, and 0 is no linear correlation. The coefficient measures both the strength and direction of the relationship. Pearson's R has limitations: it only detects linear relationships and is sensitive to outliers.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling is a crucial data preprocessing technique used in various data analysis and machine learning fields. It involves transforming the features of a dataset to a common scale. The main purposes of scaling are to improve the convergence of many machine learning algorithms and to make features more comparable and interpretable.

There are two main types of scaling methods:

1. Normalized scaling (Min-Max scaling): transforms the data to fit between 0 and 1. Formula is `X_norm = (X - X_min) / (X_max - X_min)`

2. Standardized scaling: transforms the data to have a mean of 0 and a standard deviation of 1. Formula is `X_stand = (X - μ) / σ`

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

The formula for VIF = 1/(1-R^2), where R is Pearson's R.

Based on this formula, if R is +/-1, the denominator becomes 0, and the value of VIF becomes infinite. R having a magnitude of 1 means that those variables are perfectly linear. If those variables are feature variables for a model (not dependent variable), it results in high multicollinearity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A Q-Q plot ("quantile-quantile" plot), is a graph used to compare two distributions by plotting their quantiles against each other. The more similar they are, the more aligned points on the plot to a straight line. If the points deviate significantly from the straight line, the distributions differ.

In the case of Linear regression, we can use a Q-Q plot to verify our assumption that Residuals are normally distributed.