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?

- Year (yr): On average, there were 1,981.86 more bike rentals in 2019 than in the previous year, suggesting an increase in the service's popularity or availability.
- Working day: There was a slight increase of 168.09 more rentals on working days, indicating a small boost in commuter usage
- Holiday: On average, there are 858.68 fewer bike rentals during holidays, may be due lack of commuters on holidays
- Winter season (season_4): Surprisingly, winter rentals increase by 1,026.79
- September (mnth_9): September sees an increase of 564.31 rentals, possibly due to pleasant weather conditions.
- Rainy/snowy weather (weathersit_3): Bad weather significantly decreases rentals by 2,420.16, which is expected as fewer people would cycle in rain or snow.

Temperature, though not a categorical variable, also has a notable positive effect, with each unit increase in temperature associated with 1,224.32 more rentals.

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' and 'atemp' 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?

1. 'Weathersit_3'
   - We see a large drop in rentals when it is snowing or raining. Highly negatively correlated to 'cnt'
2. 'Temp'

    - For every 1 unit increase in Temp, we see an increase in rental by around 1200
3. 'Season_4'
    - There is an increase of around 1000 rentals on winter.

NOTE: 'yr' is more significant that 'Temp' and 'Season_4', but I haven't added it as 'yr' won't be helpful to come up with any actionable insights.

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.