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?
Ans :

The categorical variables such as **season**, **yr**, **mnth**, **holiday**, **weekday**, **workingday**, and **weathersit** have varying levels of influence on bike demand. For instance:

- **Bike rentals** tend to increase during the **autumn** months.
- The months of **June, August, and October** experience a surge in **bike rentals**.
- Favorable **weather conditions** often lead to a higher preference for **renting bikes**.
- **Weekends** see a lower tendency for people to **rent bikes**.
- **Bike rentals** are less popular on **holidays**.
- The year **2019** saw **more bike rentals** compared to the year **2018**

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

Ans:

a) **Avoiding Multicollinearity**:
   - When we create dummy variables for categorical features without dropping one category, it introduces **multicollinearity**.
   - By dropping one category (usually the reference category), we can prevent multicollinearity.
   - The information from the dropped category is inherently included in the remaining dummy variables.

b) **Enhancing Interpretability**:
   - Setting drop_first=True makes the interpretation of model coefficients more intuitive.
   - It ensures that the reference category is excluded, allowing us to focus on the effects of other categories relative to the reference.

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

Ans:
When we look at the pair-plot among the numerical variables, we notice that the **"registered"** variable has a **strong correlation** with the target variable **"cnt"**. It's like they're close friends who always hang out together in the scatter plot!

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

Ans:

1. **Avoiding Multicollinearity**:
   - In statistical modeling, multicollinearity occurs when two or more independent variables are highly correlated with each other. It's like having friends who always agree and say the same things—it doesn't provide new insights.
   - When we create dummy variables for categorical features (like "Neighborhood" with categories such as "Downtown," "Suburb," and "Rural"), we need to drop one category.

Why? Because if we keep all categories, they become linearly dependent, causing multicollinearity.
    - By dropping one category (usually the reference category), we ensure that the information from the dropped category is inherently included in the remaining dummy variables. It's like saying, "Okay, let's focus on the unique aspects of each friend without redundancy."
2. **Enhancing Interpretability**:
    - Imagine explaining your model to someone over coffee. When you set drop_first=True, you're making the conversation smoother.
    - Let's say "Downtown" is our reference category. Now, when you see a positive coefficient for "Suburb," you can confidently say, "Ah, houses in the suburbs tend to have higher prices compared to downtown." It's like decoding a secret message!
    - So, dropping that first category makes the model coefficients more intuitive and easier to explain. It's like translating complex math into everyday language.

Top features that significantly contribute to explaining the demand for shared bikes, presented in a scholarly manner:

- **Temperature (temp):**

The variable 'temp' is indicative of the ambient temperature conditions. It has a substantial impact on bike rental demand, as favorable temperatures are conducive to outdoor activities such as cycling.

- **Weather Situation (weathersit_bad):**

The 'weathersit_bad' variable captures adverse weather conditions. Its significant negative influence on bike rental demand underscores the sensitivity of this activity to weather disruptions.

- **Year (yr):**

The 'yr' variable represents the year of observation. Its inclusion as a top feature suggests a temporal trend in bike rental demand, possibly reflecting an increase in the popularity of bike-sharing services over time.

These features underscore the importance of environmental and temporal factors in the utilization of bike-sharing services.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans :

- Linear Regression is a type of Supervised Machine Learning. There are two types of linear regressions.
- Simple linear Regression: where the number of predictors is one.
    - ex: y=b0+b1x, where b0 interceptors and b1 is coefficient or slop is x. x is the predictor.
- Multiple Linear Regression where the number of predictors is more than one.
    - ex : y= b0+b1x1+b2x2 + ...+bnxn, where b0 is interceptor and b1, b2,b3,...bn is coefficient/slopes of x1, x2, x3 ...xn predictors.

In Linear Regression Target variable is a continuous value. So linear regression is finding a fitted line(the fitted plane in case of multiple linear regression) so that the sum of error between the target value and the predicted value is minimum.

2. Explain the Anscombe's quartet in detail.

Anscombe’s quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

3. What is Pearson’s R?

Ans :

"Pearson's r," also called a Pearson correlation coefficient is a statistic that quantifies the strength and direction of the linear relationship between two continuous variables. It measures how well the data points of two variables fit on a straight line. Pearson's correlation coefficient ranges from -1 to 1.

1. When r is close to 1, it indicates a strong positive linear relationship. This means that as one variable increases, the other tends to increase as well.

2. An r value of 0 suggests no linear relationship between the variable

3. When r is close to -1, it indicates a strong negative linear relationship. This means that as one variable increases, the other tends to decrease, and vice versa.

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

Ans :

Scaling is a preprocessing technique in data analysis and machine learning that transforms the features (variables) of a dataset to a common scale or range. It is primarily done to address issues related to the differing scales of variables, which can affect the performance of various machine learning algorithms.

Variables in a dataset may have different measurement units and scales. Some variables may have values in a small range, while others may have values in a much larger range. Scaling ensures that all variables contribute equally to the analysis or modelling process.

1. In normalized scaling, the data is scaled to a specified range, typically [0, 1]. This is done by subtracting the minimum value of the variable from each data point and then dividing by the range. normalized scaling is useful when you want to preserve the original range of the data, and you're not concerned about the distribution's shape.

2. In standardized scaling, the data is transformed to have a mean of 0 and a standard deviation of 1. This is achieved by subtracting the mean of the variable from each data point and dividing by the standard deviation.

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

Ans :

A VIF of infinity can occur when there is perfect multicollinearity in the model. Perfect multicollinearity means that one or more independent variables can be exactly predicted from a linear combination of the other independent variables.

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

Ans:

The Q-Q plot is designed to help you visually compare the quantiles of your data to the quantiles of a theoretical distribution, which can reveal deviations from the expected distribution.
a Q-Q plot is a valuable tool for assessing the distribution of data, especially in the context of linear regression. It helps evaluate the normality assumption, detect skewness and outliers, and guide model improvement if deviations are observed.