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 variables like 'season', 'yr', 'weathersit', 'month', 'holiday' effect the dependent variable and have high significance.

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

Ans: If we do not use drop_first=True, it will create extra columns during dummy variable creation. And these dummy variables are themselves correlated. Hence it increases the multicollinearity and in turn will cause Dummy Variable Trap.

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

Ans: 'temp' has the highest

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

Ans:

- Dist. plot - Residual Analysis on the y_train and y_train_pred. To check, plot a histogram of the residuals. See if it's closer to normal distribution.
- Using the Q-Q Plot we can infer if the data comes from a normal distribution. If yes, the plot would show straight line. Absence of normality in the errors can be seen with deviation in the straight line
- If the mean of error terms is significantly away from zero, it means that the features we have selected may not actually be having a significant impact on the outcome variable
- By seeing the distribution plot of y_test and y_pred to understand the slope. Here it looks kind of linear, which supports the assumptions of Linear Regression

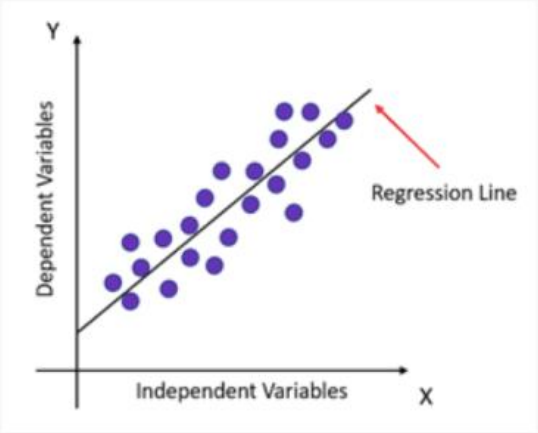
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans: 'temp', 'windspeed' and , 'yr'

General Subjective Questions

1. Explain the linear regression algorithm in detail

Ans: Linear Regression is an algorithm that belongs to supervised Machine Learning. It tries to apply relations that will predict the outcome of an event based on the independent variable data points. The relation is usually a straight line that best fits the different data points as close as possible. The output is of a continuous form.



Linear regression can be expressed mathematically as:

$y = \beta 0 + \beta\ 1x + \varepsilon$

Here,

- Y= Dependent Variable
- X= Independent Variable
- β 0= intercept of the line
- β1 = Linear regression coefficient (slope of the line)
- ε = random error

The last parameter, random error ε, is required as the best fit line also doesn't include the data points perfectly.

2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph. The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

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 coefficient is a value that indicates statistical relationship/association between two continuous variables. It is considered as the best method to determine association as it is based on the method of covariance. It gives information about the magnitude of correlation and the direction of relationship between two variables.

Pearson correlation coefficient (r)	Pearson correlation coefficient (r)	Interpretation	Example
Between 0 and 1	Positive correlation	Both variables change in same direction	Baby's length and weight
0	No correlation	To relationship between values	Car's price and size
Between 0 and -1	Negative correlation	One rises, another goes down. And, vice-versa	Elevation and airpressure

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

Ans:

- Scaling is a technique to make the variables closer to each other or in simpler words, we can say that the scaling is used for making data points generalized so that the distance between them will be lower.

- We rescale to increase interpretability of the attributes. By rescaling, we bring the variables to the same scale. After which coefficient become significantly considerable.
- At different scale, coefficient cannot show significance, At the same scale, behind-the-scene optimizations also improve. If not scale, the feature with a higher value range starts dominating when calculating distances
- Two ways of scaling –
    - 1. Min-max scaling(Normalization) : Between 0 and 1 2.
    - Standardization (mean-0, sigma-1)

For, x normalization : (x-xmin)/(xmax-xmin) ; for x=xmax , it becomes 1 || for x=xmin, it becomes 0 Standardization for x , (x - mu)/sigma

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

Ans: If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables

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

Ans: QQ-plots are ubiquitous in statistics. Most people use them in a single, simple way: fit a linear regression model, check if the points lie approximately on the line, and if they don't, your residuals aren't Gaussian and thus your errors aren't either. This implies that for small sample sizes, you can't assume your estimator $\hat{\beta}$ is Gaussian either, so the standard confidence intervals and significance tests are invalid. However, it's worth trying to understand how the plot is created in order to characterize observed violations