Subjective Answer

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

Ans :

The given data set has both categorical and numerical values and both of them has significance in determining the model. If we see the top contributor for the model

Year: which is categorical variable has huge significance

Season : if its snowing then its less likely the bikes will be rented.

So overall categorical variables has huge impact on the bike rents we need to handle them properly by creating dummy variables

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

Ans:

It is important to use drop_first=True when creating dummy variables to avoid creating an extra column. When we have n dummy variables, we need to create only n−1 columns because the extra column can be inferred from the combination of the other columns. This helps in reducing the correlation between the columns.

Example: If we have four seasons—Winter, Snow, Spring, and Fall—we need only the Snow, Spring, and Fall columns. When all of these are false, it implies that the season is Winter.

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

Ans:

Register column has highness correlation about 95%

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

Ans:

We assumed the following things

Linearity : relationship between the independent variables and the dependent variable is linear.

Plot the residuals (the differences between observed and predicted values) against the predicted values or each independent variable. The plot should show no obvious patterns. If you see patterns such as curves, this suggests non-linearity.

The residuals are normally distributed. histogram of the residuals to visually inspect if it resembles a normal distribution.

No Multicollinearity: Calculate VIF for final values has values less than 10

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

Ans:
temperature,
season snow (If its snowing less likely to rent)
Year - whether 2018 or 2019

Explain the linear regression algorithm in detail.
One simple model for predicting a dependent variable using other independent variables is linear regression. As the name suggests, a linear regression model assumes a linear relationship between the target variable and the predictor variables.
y=c+bx

In linear regression, the goal is to find the values of c and b that provide the best fit to the data. "Best fit" means minimising the sum of the errors between the predicted and actual values of the target variable.
To find the best-fitting line, we use methods to minimize the sum of squared errors. Gradient Descent One common technique to find the optimal values for c and b

gradient descent. This iterative method adjusts the coefficients in the direction that reduces the error, when plotted the error decreases as the coefficients are updated.looks like inverse hyper parabola.

When dealing with multiple independent variables, the model generalizes to:

y=c+b1 x1 +b2 x2 +…+bn xn

While multiple variables can be used in the model, it's crucial that these variables are independent of each other. High correlation among predictor variables (multicollinearity) can distort the model's estimates

Explain the Anscombe's quartet in detail.
Anscombe's Quartet majorly talks about how important it is to understand the data and not just numeric of model. Visualising the data can help us a lot in understanding the relationship between the variables. Anscombe's Quartet talks about 4 types of data that has the same numerical values like mean median variance but how does their relationship with target variable can be completely different. Like one might be a proper data set with linear relationship. One can be linear relationship with outliners. One can be completely non-linear relationship and one can be a straight line with one big/far outliner.

What is Pearson's R?
Pearson's R is the statistical measurement of the strength and direction of linear relation between two continuous variable. The value is between [-1 , 1] where 1 indicates strong linear relation when one variable increases the other would also increase .-1 indicate inverse linear when one increases other would decrease. relation and 0 indicate no linear relation. Values nearing -1 0r1 indicate they are highly co related
One thing to remember Pearson's R measures only correlation and not causation.

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

Ans:

Scaling is one of the very important step in data preparation where we adjust the range of the values. This is very important if not done properly this might affect overall model accuracy. One or few variables might dominant the other variables. Variables with higher values might have higher co efficient which would imply that as a top feature but it might actually not be. Example in case of housing problems area values are comparatively large than others if not scaled co efficient of area will be high compared to all others

Two types of scaling

Min-max scaling : scaling all the values between [0,1]

Standardization scaling is such that the mean of the data is 0 and standard deviation is 1

When the data has outliners Min-max scaling is sentive the effect the min and max values were as Standardization is more robust. And generally min-max scaling is used when features have different units or when a bounded range is needed.

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

VIF is a parameter used to measure multicollinearity in regression. multicollinearity when predictor variable are highly correlated with each other, which can make it difficult to determine the individual effect of each predictor on the dependent variable.

The VIF for a particular predictor variable measures how much the variance of that variable's estimated regression coefficient is inflated due to multicollinearity with other predictors. It is defined as:

$VIFi = 1/ 1-Ri2$

A VIF value becomes infinite when the denominator of the formula $1-Ri2$ approaches zero happens when $Ri2$ is equal to If the i-th predictor variable is a perfect linear combination of the other predictor variables, $Ri2$ will be 1. This indicates that the predictor variable is perfectly predicted by the other variables, leading to an infinite VIF. In practical terms, this means that one of your predictor variables is redundant because it can be exactly predicted using other variables in the model.

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 graphical tool used to assess if a dataset follows a particular theoretical distribution, such as the normal distribution.
A Q-Q plot is a scatter plot that compares the quantiles of the sample data against the quantiles of a theoretical distribution.
It create a scatter plot where the x-axis represents the quantiles from the theoretical distribution, and the y-axis represents the quantiles from the sample data

Use: One of the key assumptions of linear regression is that the residuals (the differences between observed and predicted values) are normally distributed. Checking this assumption is important for validating the reliability of the regression model's statistical inferences.