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?**

Answer:
Overall, the categorical variables show a seasonal trend on the target variable.
- mnth and season show the demand is highest during Fall/ September and lowest during Spring/ December, January .
- Year 2019 show a growth in demand compared to 2018
- Usage is lower on holiday.
- Total usage is similar for all weekdays, but the ratio of casual:registered users is higher on Sunday and Monday.

In term of weather: Clear weather leads to higher demand while snow/rain weather leads to declined usage.

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

Answer:
It helps in reducing the extra column created during dummy variable creation. If we have n category, we only need to create n-1 columns. Have more columns than we need can create correlations among dummy variables.

However, by default, pandas drops the first label alphabetically. If we want to specify a category to drop (usually the one with the most occurrence), then manual dropping is required. The key takeaway here is to ensure dropping at least one reference category.

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

Answer:
Temp (Pearson correlation coefficient 0.63) has the highest correlation to the target variable “cnt”. casual/registered actually have higher correlation, but they are part of the dependent variable, not predictor variables.

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

Answer:
For multicollinearity:
- Variance Inflation Factor (ensure no variable have VIF > 5)

For mean of residuals = 0:
- calculate the mean of residuals.

For homoscedasticity:
1. Use Goldfeld-Quandt test, with null hypothesis is that the residuals are homoscedastic.
2. Plotting the actual values against the predicted values and observe the distribution.

For linearity of variable:
- Plot the residuals and the fitted values, ensure that they are randomly and uniformly scattered on the x-axis.

For normality of error term:
- Plot the histogram of residuals and observe the distribution
- Q-Q plot should form a straight line.

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

Answer:

The answer to this depends on how we choose to quantify "the contribution towards explaining the demand".

Based on the values of coefficients, the top 3 variables are **temperature**, **year 2019** and **humidity**, as one (normalized) unit increase of these 3 will affect the demand most significantly.

An alternative way is calculating the change in R-squared after removing a variable. The change in R-squared represents the variance that a particular variable explained which other variable could not. In other words, it is the amount of unique variance that a variable explains above the other variables. The pros in this approach is that it will remain unchanged, regardless of the unit of measurement or the scaling method (or lack thereof) used on the data. Based on this approach, the top 3 variables are **year2019**, **temp** and **Winter**, as they have the highest amount of unique explained variance.

Variable removed	R2 after removed	Change in R2 by variable
yr_2019	0.570416	0.274971
temp	0.610997	0.234391
Winter	0.784499	0.060888
Summer	0.794658	0.050730
humidity	0.797581	0.047807

General Subjective Questions

1. Explain the linear regression algorithm in detail

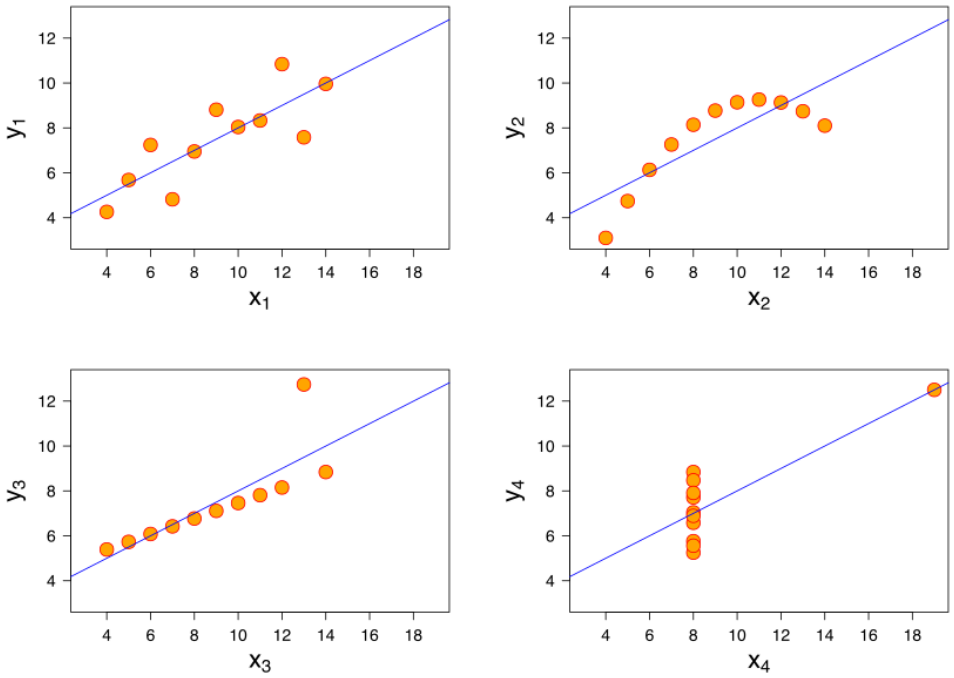
Answer:
Linear regression is a supervised learning algorithm. Linear regression is used to predict the value of a variable (dependent variable) based on the value of another variable or multiple variables (independent variables). This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual values.

2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.

Mean of x: 9
Sample variance of x: 11
Mean of y: 7.5
Sample variance of y: 4.125
Correlation between x and y: 0.816
Linear regression line: y = 3 + 0.5x
R-squared: 0.67



- The first scatter plot (top left) appears to be a simple linear relationship.
- The second graph (top right); while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the relationship is linear, but should have a different regression line. The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

3. What is Pearson's R?

Answer:

Pearson's correlation coefficient is a measure of linear correlation between two sets of data. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

- The strength of the correlation increase as it approach -1 or 1, while 0 indicates no correlation.
- Positive value indicates the two variables move in the same direction, while negative value indicates they move in opposite direction.

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

Answer:

Feature scaling is a method used to normalize the range of independent variables or features of data. When we have different numerical variables with vastly different ranges, without scaling, it is very hard to interpret coefficients of regression models correctly. Scaling ensure all the variables have the same range (0-1) for min-max scaling or same standard deviation of 1 and mean of 0 for standardized scaling. Standardized scaling does not compress the data in a particular range as min-max scaling, so it is useful when there is outliers.

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

Answer:

If value of VIF can be infinity when there is 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.

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

Answer:

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

In linear regression, q-q plot is used to check the assumption of normality of error term. We plot the quantile of error term against a normal distribution. If the error term follows a normal distribution, the plot should form a straight line.