Bike Sharing Linear Regression Assignment

Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer:

- Bike demand is higher in 2019 compared to 2018.
- Bike demand is high in the months of May to September with some drop in the months of November & December.
- Bike demand is high if weather is clear or with mist cloudy while it is low when there is light rain or light snow.
- The demand for bikes is almost similar throughout the weekdays.
- Bike demand is higher when there is no holiday

Q2. Why is it important to use drop_first=True during dummy variable creation?

Answer:

- It is important in order to achieve k-1 dummy variables as it can be used to delete extra columns while creating dummy variables.
- It is also used to reduce the collinearity between dummy variables

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer:
atemp and **temp** both have the same correlation with the target variable of **0.63** which is the highest among all numerical variables.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer:
One of the fundamental assumptions of a linear regression model is that the error terms should correspond to a normal curve, when plotted on a histogram. Hence, this assumption is validated.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

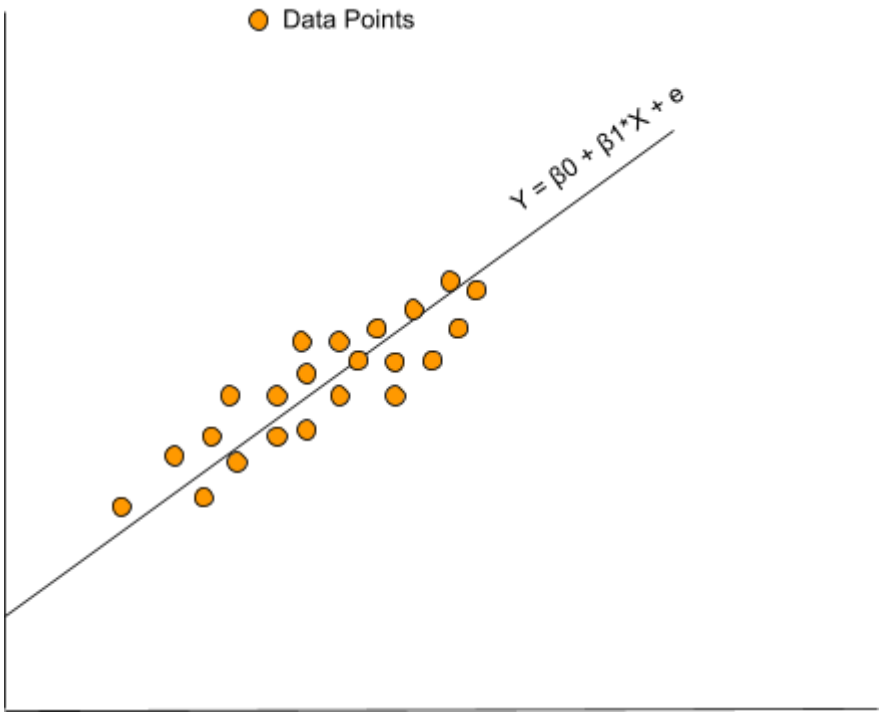
Answer:
The top 3 features which have the highest coefficients:
These are: **temp**, **yr** (positively influencing) and **light snow and rain** (negatively influencing).

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Answer:

- Linear regression is an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events.
- It is a statistical method used in data science and machine learning for predictive analysis.
- It is used for supervised learning



Pros:

- Linear Regression is simple to implement.
- Less complexity compared to other algorithms.
- Linear Regression may lead to overfitting but it can be avoided using some dimensionality reduction techniques, regularization techniques, and cross-validation.

Cons:

- Outliers affect this algorithm badly.
- It over-simplifies real-world problems by assuming a linear relationship among the variables, hence not recommended for practical use-cases.

When to use Linear Regression:
You can use linear regression when you want to predict a continuous dependent variable from a scale of values

Q2. Explain the Anscombe's quartet in detail.

Answer:-

Anscombe's quartet is a set of **four datasets** that have nearly **identical statistical** properties but vastly **different graphical** representations. It was created by the statistician Francis Anscombe in 1973 to emphasize the importance of graphically exploring data before drawing conclusions based solely on statistical summary measures.
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**.

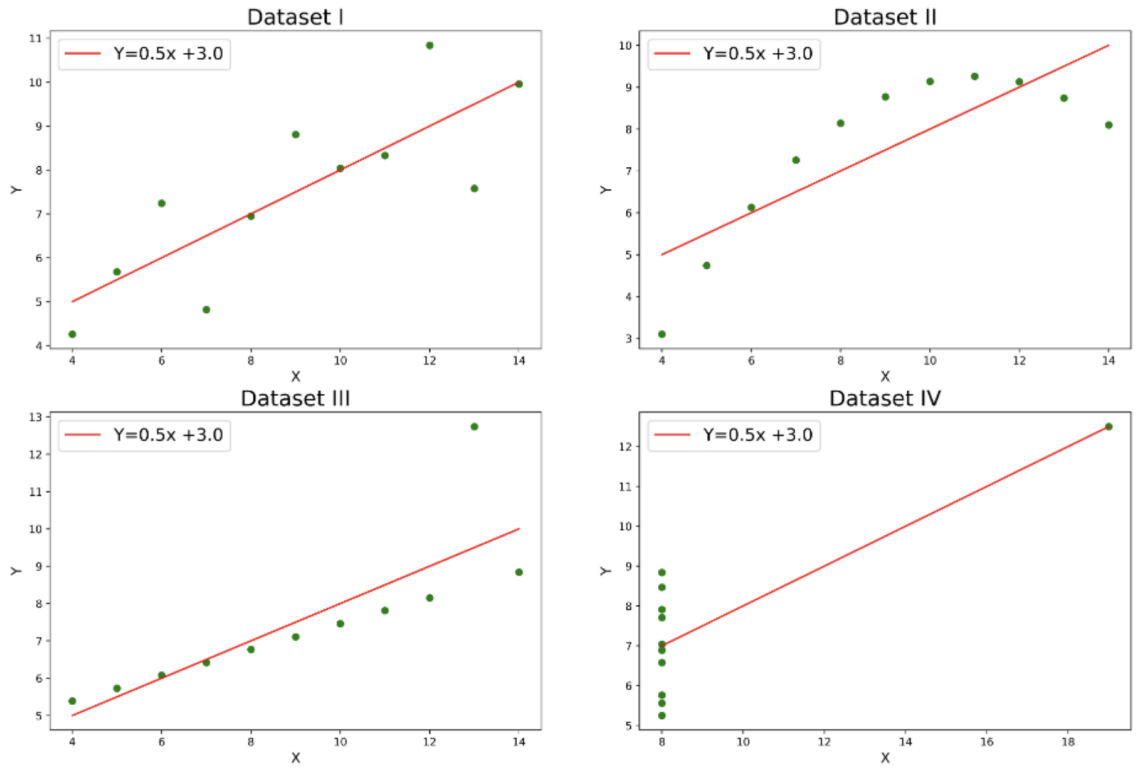
Anscombe's quartet dataset.

x1	x2	x3	x4	y1	y2	y3	y4
10	10	10	8	8.04	9.14	7.46	6.58
8	8	8	8	6.95	8.14	6.77	5.76
13	13	13	8	7.58	8.74	12.74	7.71
9	9	9	8	8.81	8.77	7.11	8.84
11	11	11	8	8.33	9.26	7.81	8.47
14	14	14	8	9.96	8.1	8.84	7.04
6	6	6	8	7.24	6.13	6.08	5.25
4	4	4	19	4.26	3.1	5.39	12.5
12	12	12	8	10.84	9.13	8.15	5.56
7	7	7	8	4.82	7.26	6.42	7.91
5	5	5	8	5.68	4.74	5.73	6.89

All four datasets share the same:

	Data Set 1	Data Set 2	Data Set 3	Data Set 4
Mean_x	9.000000	9.000000	9.000000	9.000000
Mean_y	7.500909	7.500909	7.500000	7.500909
Variance_x	11.000000	11.000000	11.000000	11.000000
Variance_y	4.127269	4.127269	4.122620	4.123249
Correlation	0.816421	0.816237	0.816287	0.816521
Linear Regression slope	0.500091	0.500000	0.499727	0.499909
Linear Regression intercept	3.000091	3.000909	3.002455	3.001727

Linear Regression Equation: All datasets have nearly identical regression lines:
$y=0.5x+3$



Anscombe's quartet Plot

Four datasets in Anscombe's quartet:

- **Data Set 1:** (In Figure - Dataset 1) - fits the linear regression model pretty well.
- **Data Set 2:** (In Figure - Dataset 2) - cannot fit the linear regression model because the data is non-linear.
- **Data Set 3:** (In Figure - Dataset 3) - shows the outliers involved in the data set, which cannot be handled by the linear regression model.
- **Data Set 4:** (In Figure - Dataset 4) - shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

As you can see, Anscombe's quartet helps us to understand the importance of data visualization. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

Q3. What is Pearson's R?

Answer:-

Pearson's r (also known as Pearson correlation coefficient) is a statistical measure that quantifies the **strength** and **direction** of the linear relationship between **two continuous variables**. It's named after Karl Pearson, who developed the coefficient.

Pearson's r ranges from -1 to 1, where:

- **r = 1:** Indicates a perfect **positive** linear relationship. As one variable increases, the other variable also increases proportionally.
- **r = -1:** Indicates a perfect **negative** linear relationship. As one variable increases, the other variable decreases proportionally.
- **r = 0:** Indicates **no linear** relationship between the variables.

The formula to calculate Pearson's r is:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

Pearson Correlation Coefficient Interpretation

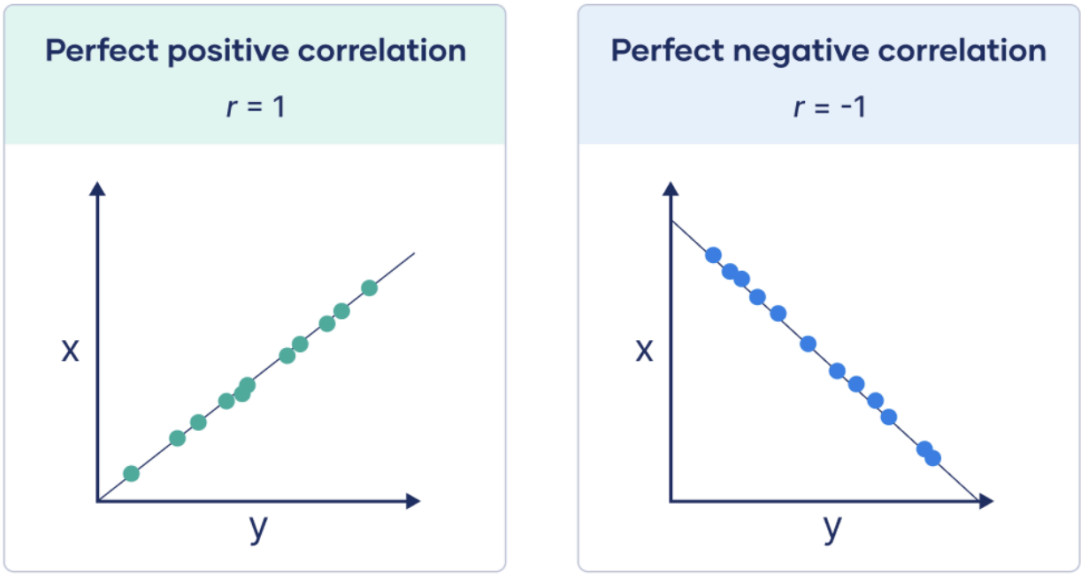
Pearson correlation coefficient (*r*) value	Strength	Direction
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None

Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

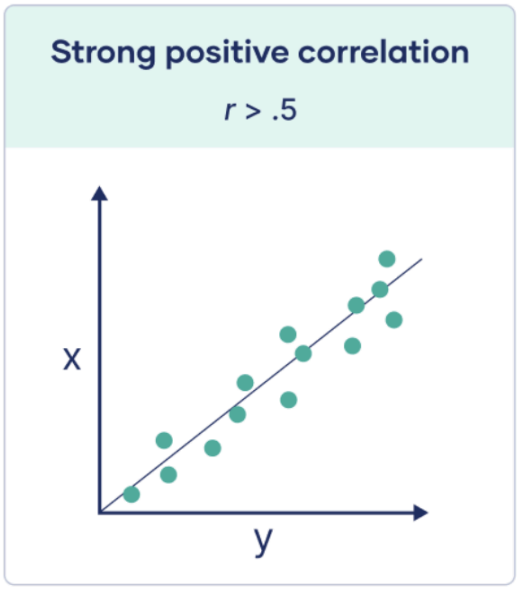
Visualizing the Pearson correlation coefficient

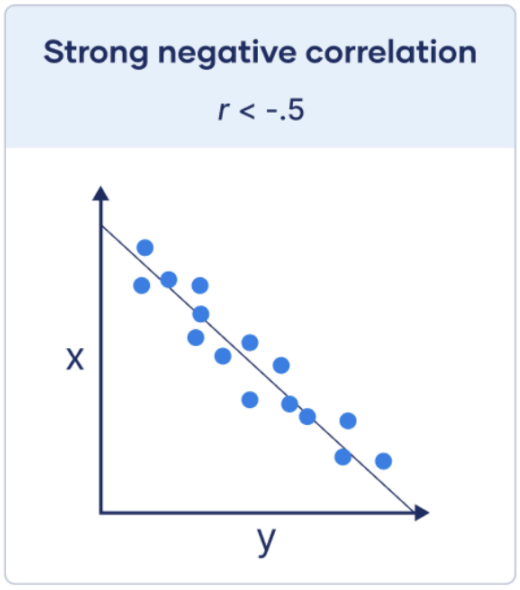
- Visualizing is another way to think of the Pearson correlation coefficient (*r*) as a measure of how close the observations are to a **line of best fit.**
- The Pearson correlation coefficient also tells you whether the slope of the line of best fit is **negative** or **positive**. When the slope is negative, *r* is negative. When the slope is positive, *r* is positive.

When *r* is 1 or –1, all the points fall exactly on the line of best fit:

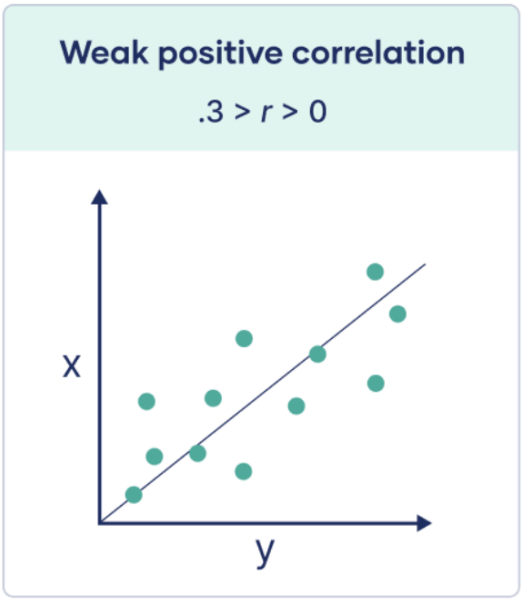


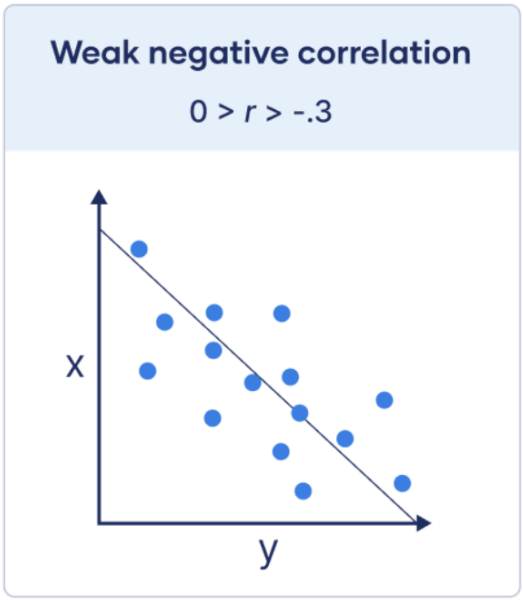
When *r* is greater than .5 or less than –.5, the points are close to the line of best fit:



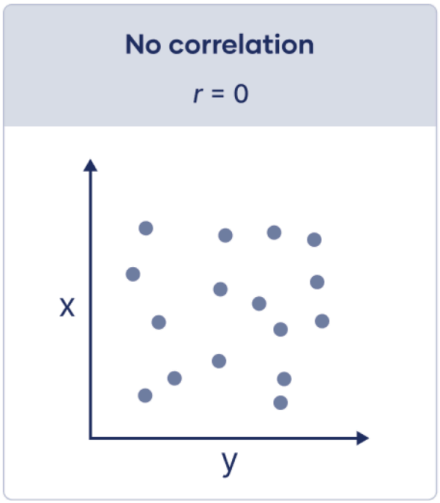


When *r* is between 0 and .3 or between 0 and –.3, the points are far from the line of best fit:





When *r* is 0, a line of best fit is not helpful in describing the relationship between the variables:



Bivariate Correlation

Pearson's correlation coefficient is a statistical tool used to measure **bivariate correlation**. This refers to the strength and direction of the linear relationship between **two** variables. It assesses how much one variable tends to change along with the other.

When to use the Pearson correlation coefficient

The Pearson correlation coefficient (*r*) is one of several correlation coefficients that you need to choose between when you want to measure a correlation. The Pearson correlation coefficient is a good choice when **all** of the following are true:

- **Both variables are quantitative:** You will need to use a different method if either of the variables is qualitative.
- **The variables are normally distributed:** You can create a histogram of each variable to verify whether the distributions are approximately normal. It's not a problem if the variables are a little non-normal.
- **The data have no outliers:** Outliers are observations that don't follow the same patterns as the rest of the data. A scatterplot is one way to check for outliers—look for points that are far away from the others.
- **The relationship is linear:** Linear means that the relationship between the two variables can be described reasonably well by a straight line. You can use a scatter plot to check whether the relationship between two variables is linear.

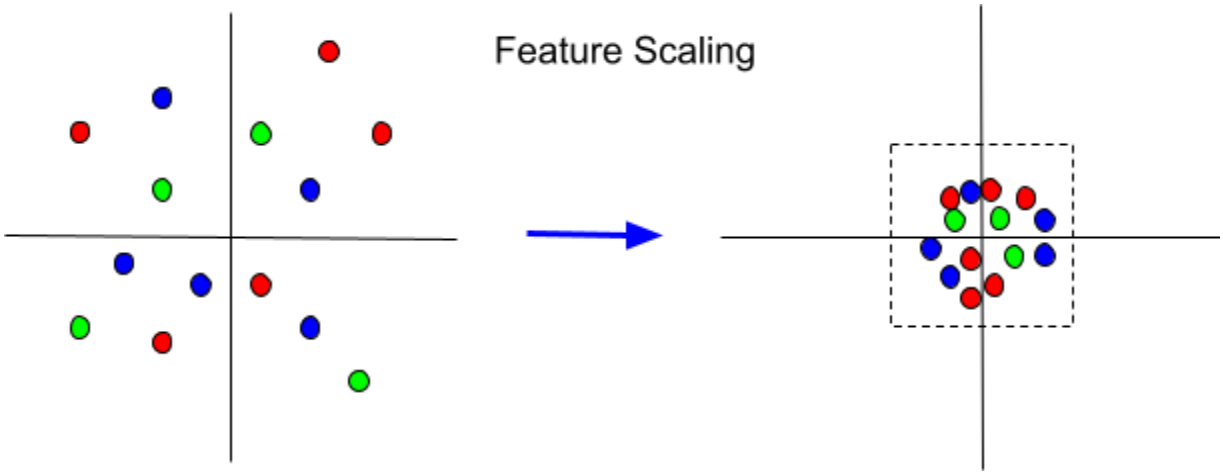
In summary, Pearson's R is a valuable tool for understanding the linear relationships between continuous variables. By calculating and interpreting its value, you can gain insights into the data and make more informed decisions.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer:
Scaling in General: The process of **adjusting the size** or **magnitude** of something relative to a reference point. Scaling in Machine Learning is called Feature Scaling.

Feature scaling is a data preprocessing technique used to transform the values of features or variables in a dataset to a similar scale. The purpose is to ensure that all features contribute equally to the model and to avoid the domination of features with larger values.

Feature scaling becomes necessary when dealing with datasets containing features that have different ranges, units of measurement, or orders of magnitude. In such cases, the variation in feature values can lead to biased model performance or difficulties during the learning process.



There are several common techniques for feature scaling, including **standardization**, **normalization**, and **min-max scaling**. These methods adjust the feature values while preserving their relative relationships and distributions.

Why is scaling performed?

Many machine learning algorithms use **distance-based** calculations to make predictions. If the features are not scaled, those with larger values can have a disproportionate impact on the results.

This helps in handling skewed data and outliers, which can influence the model's behavior.

The importance of feature scaling in Machine learning:

Enhancing Model Performance:

Feature scaling can significantly enhance the performance of machine learning models. Scaling the features makes it easier for algorithms to find the optimal solution, as the different scales of the features do not influence them.

It can lead to faster convergence and more accurate predictions, especially when using algorithms like k-nearest neighbors, support vector machines, and neural networks.

Addressing Skewed Data and Outliers:

Skewed data and outliers can negatively impact the performance of machine learning models. Scaling the features can help in handling such cases. By transforming the data to a standardized range, it reduces the impact of extreme values and makes the model more robust.

This is particularly beneficial for algorithms that assume a normal distribution and are sensitive to outliers, such as linear regression.

Faster Convergence During Training:

For gradient descent-based algorithms, feature scaling can speed up the convergence by helping the optimization algorithm reach the minima faster.

Balanced Feature Influence:

When features are on different scales, there is a risk that larger-scale features will dominate the model's decisions, while smaller-scale features are neglected. Feature scaling ensures that each feature has the opportunity to influence the model without being overshadowed by other features simply because of their scale.

Improved Algorithm Behavior:

Certain machine learning algorithms, particularly those that use distance metrics like Euclidean or Manhattan distance, assume that all features are centered around zero and have variance in the same order.

Without feature scaling, the distance calculations could be skewed, leading to biases in the model and potentially misleading results. Feature scaling normalizes the range of features so that each one contributes equally to the distance calculations.

Difference between normalized scaling and standardized scaling

Both normalized scaling and standardized scaling are techniques used in machine learning to transform features within a dataset to a specific range. However, they differ in the way they achieve this transformation:

Normalized Scaling:

Normalization in machine learning is a data preprocessing technique used to change the value of the numerical column in the dataset to a common scale without distorting the differences in the range of values or losing information.

Range: Scales features to a **fixed range**, typically between 0 and 1 (or -1 and 1)

Impact of Outliers: Sensitive to outliers. Extreme values in the original data can significantly affect the scaling factor and compress the range for other data points.

Formula:

$$X_{new} = \frac{X - X_{mean}}{X_{max} - X_{min}}$$

X_{new} - Scaled Feature, X - Original Feature

Standardized Scaling:

Distribution: Transforms features to have a **standard normal distribution** with a mean of 0 and a standard deviation of 1.

Impact of Outliers: Less sensitive to outliers compared to normalized scaling. Outliers have a reduced impact on the mean and standard deviation used for scaling.

Formula:

$$X_{new} = \frac{X - X_{mean}}{\sigma}$$

Here's a table summarizing the key differences:

Feature	Normalized Scaling	Standardized Scaling
Target Range	0 to 1 (or -1 to 1)	Mean: 0, Std. Dev: 1

Underlying Distribution	No specific assumption	Assumes normal distribution
Sensitivity to Outliers	High	Less sensitive
Typical Use Cases	General purpose	Preferred for most ML algorithms

In essence, both techniques aim to bring features to a similar scale, but normalized scaling focuses on a fixed range, while standardized scaling targets a specific statistical distribution.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer:

The VIF (Variance Inflation Factor) becomes infinite in a situation where there is perfect multicollinearity

VIF is calculated as 1 / (1 - R^2), where R^2 is the coefficient of determination between an independent variable and the fitted values of all other independent variables (excluding itself).

VIF = 1 / (1 - R^2)

When R^2 = 1, - In perfect multicollinearity.

Then VIF = 1 / (1 - 1) => VIF is **infinity**, This is a serious issue with your model. The coefficients of the independent variables become unreliable, making it difficult to interpret their significance and the overall model's accuracy. To Handle this infinity issue, If possible, remove one or more highly correlated variables.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answer:

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with the same distributions.

Importance of Q-Q Plots in Linear Regression:

- **Normality Assumption:** Many statistical tests used in linear regression, such as hypothesis testing for coefficients, rely on the assumption of normally distributed errors.
- **Identifying Issues:** A Q-Q plot can help you visually identify if the residuals deviate from normality. This can indicate potential problems like heteroscedasticity (unequal variance of errors) or outliers that might affect the validity of your model's results.
- **Model Improvement:** If the Q-Q plot reveals non-normality, you might need to consider transformations of the data (e.g., log transformation) or explore alternative regression models that don't rely on the normality assumption.