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?

a) Seasonal Demand: The demand for bikes varies significantly across different seasons, with some seasons showing higher median demand than others, higher in summer and fall.

b) Yearly Trend: There is a noticeable increase in bike demand from one year to the next, indicating a growing trend in bike usage.

c) Monthly Variation: There are clear seasonal patterns in bike sharing counts. Both years show higher counts during the warmer months (May to September) and lower counts during the colder months (November to February).

d) Holiday Impact: Bike demand tends to be lower on holidays compared to non-holidays.

e) Weekday vs. Weekend: Fridays indicating higher usage as people might be planning for weekend activities or commuting home. There is a dip in usage on Sundays and, to a lesser extent, on Saturdays, indicating reduced commuting and possibly more recreational usage.

f) Working Days: There is a difference in bike demand between working days and non-working days, with working days generally showing higher demand.

g) Weather: Bike demand increases when the weather is clear and decreases during light rain.

h) The ideal biking conditions are moderate temperatures, moderate humidity, and moderate wind speeds.

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

When you create dummy variables for a category, you turn each category into a separate column with 0s and 1s. If you have three categories, you get three columns. But, having all three columns can cause problems because one column can be predicted from the others.

By using drop_first=True, you drop one of these columns. This makes your data simpler and avoids confusion in your model. Basically, avoiding multicollinearity and reducing dimensionality.

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

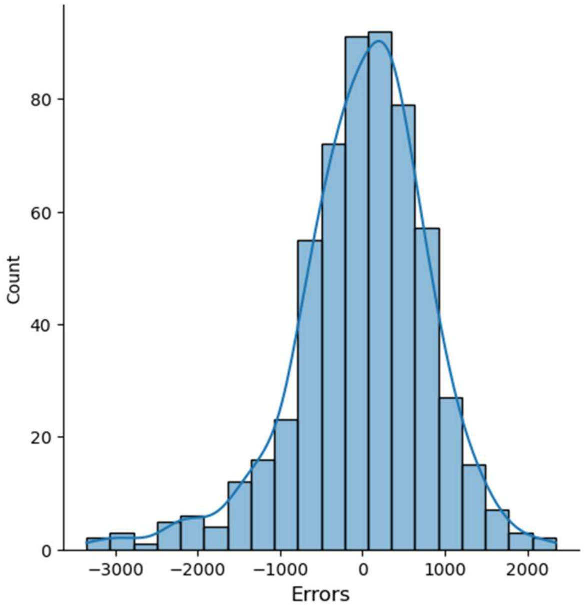
Temp(temperature) and atemp has the highest correlation with the target variable.

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

a) Normality

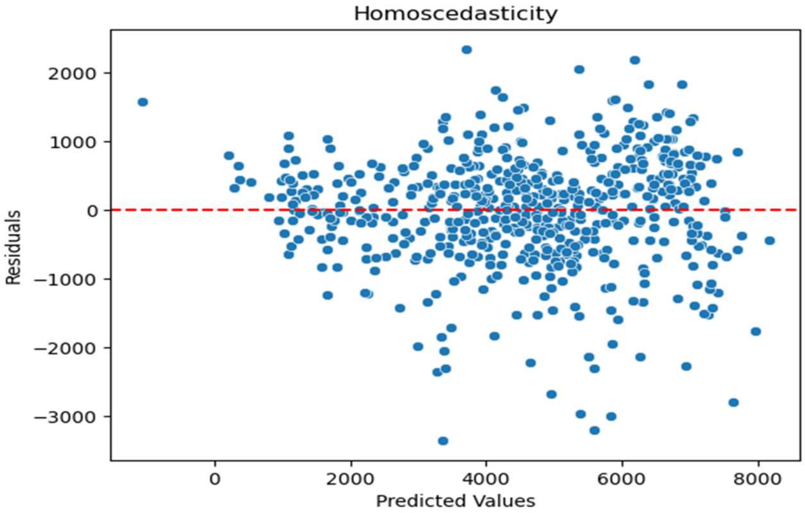
Residual Analysis

Residual analysis is performed to assess the goodness of fit and identify patterns or anomalies in the model's predictions.



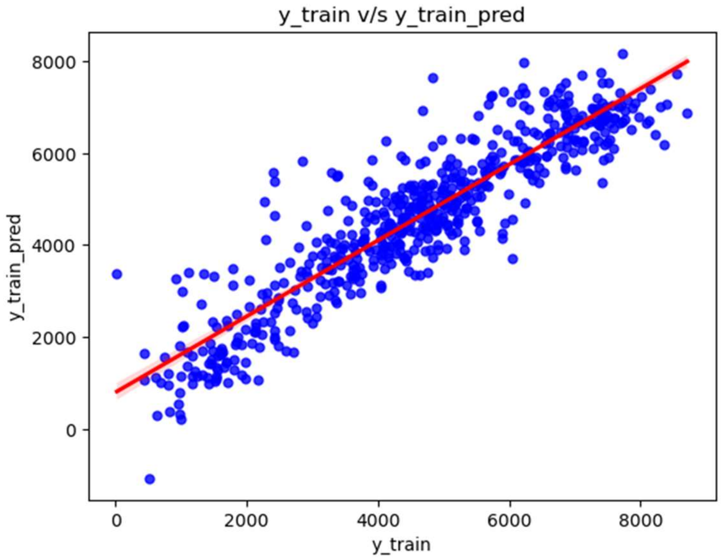
The error terms follow the principle of a normal distribution curve.

b) Homoscedasticity



The variance of residuals is constant across all levels of the independent variables, indicating consistent levels of variability.

c) Linearity



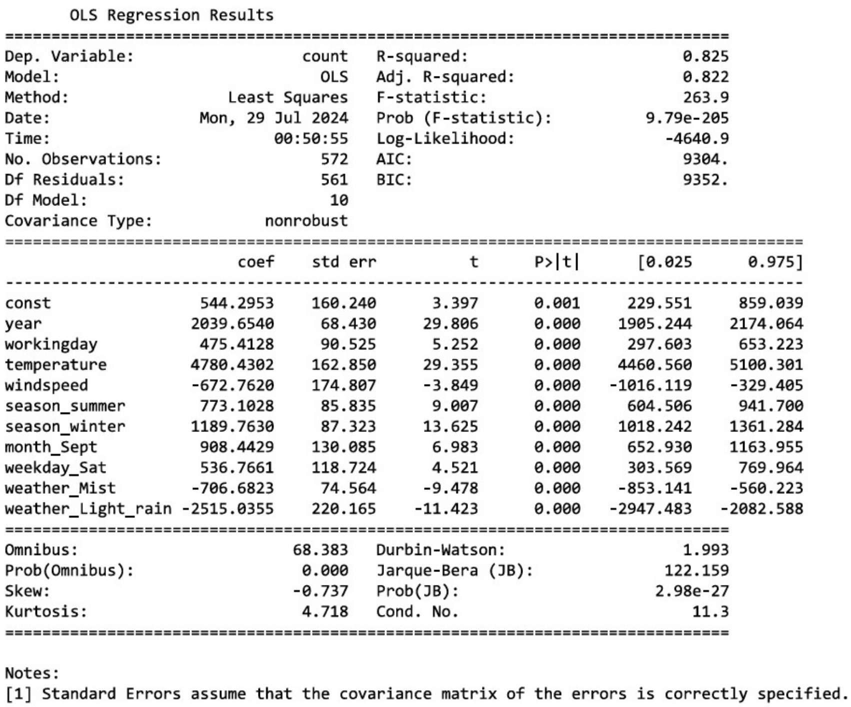
The relationship between the actual values (y_train) and the predicted values (y_train_pred) should be linear. The scatter plot shows a straight-line pattern. The red line in plot helps to visualize this. The most points are close to this line, the linearity assumption is likely satisfied.

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

Based on the OLS regression results, the top 3 features contributing significantly towards explaining the demand for shared bikes are:

1. **temperature (coef: 4780.43)** - This feature has the highest coefficient, indicating that temperature has a strong positive effect on bike demand. Higher temperatures increase bike usage.
2. **year (coef: 2039.65)** - This feature also has a significant positive coefficient, suggesting that bike demand tends to increase over the years.
3. **season_winter (coef: 1189.76)** - This feature has a substantial positive coefficient as well, indicating that bike demand is notably higher in winter.

These features have the highest coefficients, indicating their strong influence on bike demand according to the model.



```
       OLS Regression Results
==============================================================================
Dep. Variable:                  count   R-squared:                       0.825
Model:                            OLS   Adj. R-squared:                  0.822
Method:                 Least Squares   F-statistic:                     263.9
Date:                Mon, 29 Jul 2024   Prob (F-statistic):          9.79e-205
Time:                        00:50:55   Log-Likelihood:                -4640.9
No. Observations:                 572   AIC:                             9304.
Df Residuals:                     561   BIC:                             9352.
Df Model:                          10
Covariance Type:            nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
const                544.2953    160.240      3.397      0.001     229.551     859.039
year                2039.6540     68.430     29.806      0.000    1905.244    2174.064
workingday           475.4128     90.525      5.252      0.000     297.603     653.223
temperature         4780.4302    162.850     29.355      0.000    4460.560    5100.301
windspeed           -672.7620    174.807     -3.849      0.000   -1016.119    -329.405
season_summer        773.1028     85.835      9.007      0.000     604.506     941.700
season_winter       1189.7630     87.323     13.625      0.000    1018.242    1361.284
month_Sept           908.4429    130.085      6.983      0.000     652.930    1163.955
weekday_Sat          536.7661    118.724      4.521      0.000     303.569     769.964
weather_Mist        -706.6823     74.564     -9.478      0.000    -853.141    -560.223
weather_Light_rain -2515.0355    220.165    -11.423      0.000   -2947.483   -2082.588
==============================================================================
Omnibus:                       68.383   Durbin-Watson:                   1.993
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              122.159
Skew:                          -0.737   Prob(JB):                     2.98e-27
Kurtosis:                       4.718   Cond. No.                         11.3
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a statistical method used for modelling the relationship between a dependent variable and one or more independent variables. It is widely used for predicting the value of the dependent variable based on the values of one or more independent variables. The basic idea is to find the best-fitting line (or hyperplane in the case of multiple independent variables) that minimizes the sum of the squared differences between the observed and predicted values of the dependent variable.

Linear regression algorithm follows following steps:

1. Model Representation:

Simple Linear Regression: In the case of a single independent variable, the model is represented as:

$$y = b_0 + b_1 \cdot x + \varepsilon$$

where:
- y is the dependent variable,
- x is the independent variable,
- b_0 is the y-intercept (constant term),
- b_1 is the slope of the line, and
- ε represents the error term.

Multiple Linear Regression: When there are multiple independent variables, the model is extended to:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

where: (x_1, x_2 ,..., xp) are the independent variables, and (β_1, β_2,... β_p) are the coefficients.

2. Objective Function:

The goal is to find the values of (β_0, β_1, β_2,... β_p) that minimize the sum of the squared differences between the observed and predicted values. This is often expressed as the sum of squared errors (SSE) or mean squared error (MSE):

$$\text{MSE} = \frac{1}{2m}\sum_{i=1}^{m}(y_i - \hat{y_i})^2$$

where (m) is the number of data points, (y_i) is the observed value, and ($\hat{y_i}$) is the predicted value.

3. Minimization: To find the optimal values of the coefficients, the algorithm uses optimization techniques such as gradient descent. The objective is to iteratively update the coefficients in the direction that minimizes the cost function.

4. Training the Model: The model is trained on a dataset, where the algorithm learns the values of the coefficients that best fit the data. This involves feeding the algorithm input-output pairs and adjusting the coefficients until the model produces predictions close to the actual outcomes.

5. Prediction: Once the model is trained, it can be used to make predictions on new, unseen data. The predicted values are obtained by plugging the new input values into the learned regression equation.

6. Evaluation: The model's performance is assessed using metrics such as (R) (coefficient of determination), MSE, or other relevant metrics, depending on the context.

7. Assumptions: Linear regression relies on the assumption of a linear relationship between independent and dependent variables, normally distributed errors, constant error variance (homoscedasticity), and the absence of perfect multicollinearity, ensuring that there is no perfect linear relationship among the predictors.

Linear regression is a versatile and widely used algorithm, but it's important to check whether its assumptions hold in each dataset and consider more advanced techniques when those assumptions are violated.

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression **model** if built. They have very different distributions and appear differently when plotted on scatter plots.

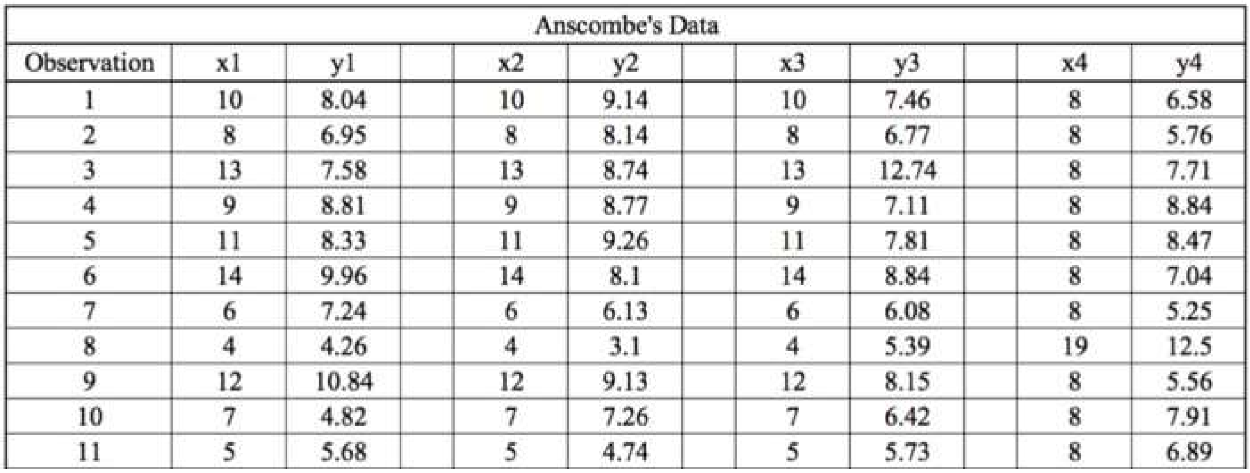
It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.

There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

This tells us about the importance of visualizing the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc.

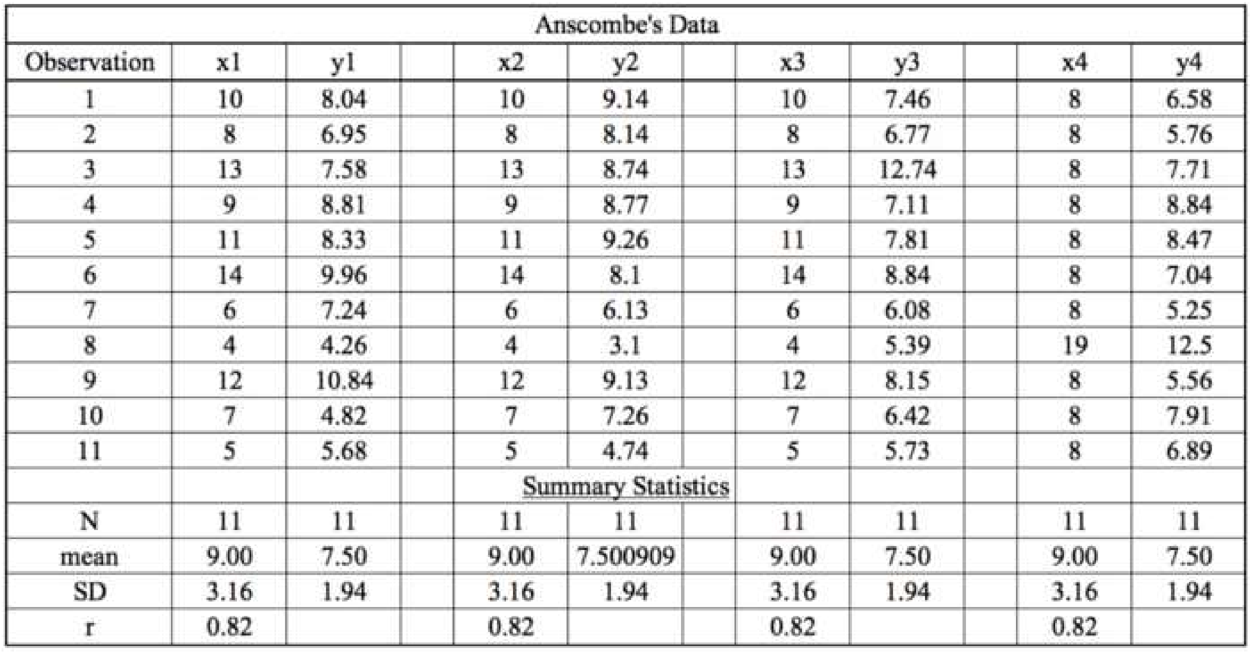
Also, the Linear Regression can be only be considered a fit for the **data with linear relationships** and is incapable of handling any other kind of datasets.

These four plots can be defined as follows:



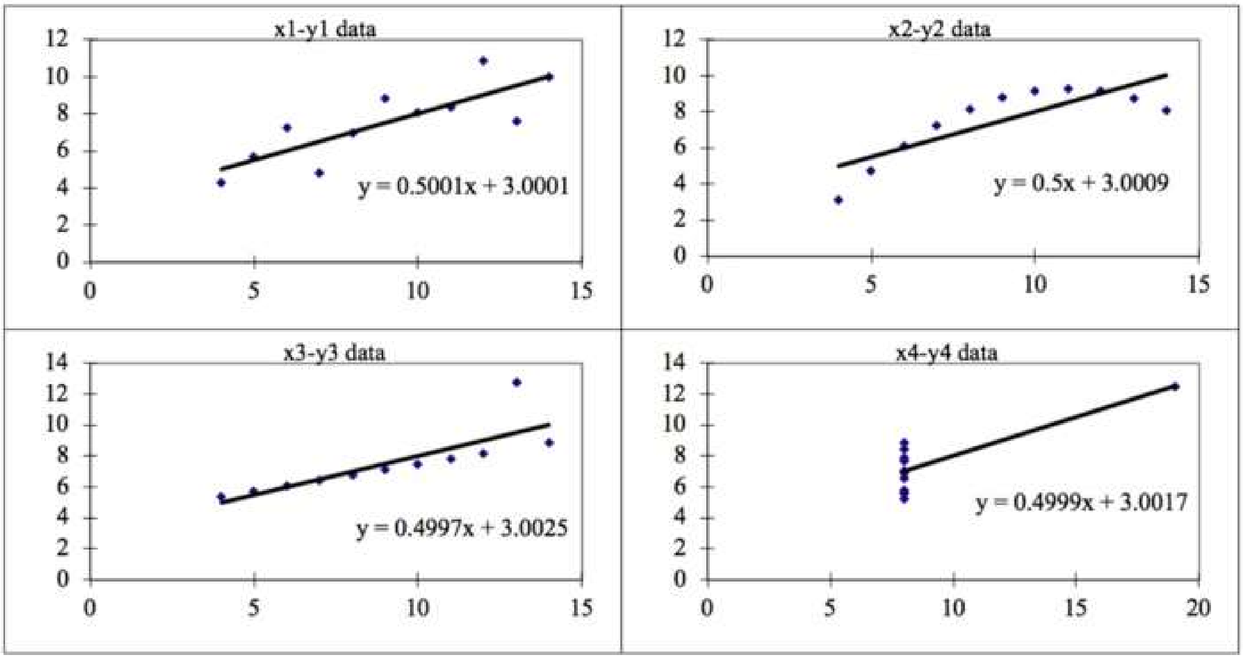
Anscombe's Data												
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89

The statistical information for all these four datasets are approximately similar and can be computed as follows:



Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
					Summary Statistics						
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

When these models are plotted on a scatter plot, all datasets generates a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows:



The four datasets can be described as:

- **Dataset 1:** this **fits** the linear regression model pretty well.

- **Dataset 2:** this **could not fit** linear regression model on the data quite well as the data is non-linear.
- **Dataset 3:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model.
- **Dataset 4:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model.

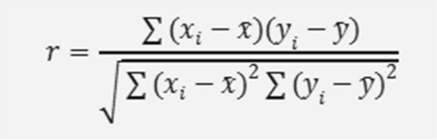
Anscombe’s quartet serves as a powerful reminder that data visualization is a crucial step in data analysis. It helps to uncover patterns, relationships, and anomalies that might not be evident from summary statistics alone.

3. What is Pearson’s R?

Pearson’s R, also known as the Pearson correlation coefficient, is a measure of the linear relationship between two variables. It quantifies the strength and direction of this relationship and is represented by a value between -1 and 1.

Key Points:

1. **Value Range**:
   - **+1**: Perfect positive linear relationship.
   - **0**: No linear relationship.
   - **-1**: Perfect negative linear relationship.
2. **Interpretation**:
   - **Positive Correlation**: As one variable increases, the other also increases.
   - **Negative Correlation**: As one variable increases, the other decreases.
   - **No Correlation**: No predictable relationship between the variables.
3. **Calculation**: Pearson’s R is calculated using the covariance of the variables divided by the product of their standard deviations. The formula is:



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

where (x_i) and (y_i) are the individual sample points, and ($\bar{x}$) and ($\bar{y}$) are the means of the variables.

4. **Use Cases**:

- **Statistics**: To determine the strength and direction of the relationship between two variables.
- **Data Analysis**: To identify patterns and correlations in data sets.
- **Research**: To test hypotheses about relationships between variables.

Example:

If you were analyzing the relationship between hours studied and exam scores, a Pearson's R value of 0.85 would indicate a strong positive correlation, meaning that more hours studied is associated with higher exam scores.

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

Scaling is the process of transforming data to fit within a specific range or distribution, often to make it suitable for analysis, especially in machine learning algorithms. It involves adjusting the values of the dataset so that they fall within a certain range or follow a specific distribution.

Why is Scaling Performed?

1. **Improving Model Performance**: Many machine learning algorithms, especially those based on distance metrics (e.g., K-Nearest Neighbors, Support Vector Machines, and clustering algorithms), perform better when features are on a similar scale. This is because large differences in feature scales can disproportionately affect the distance calculations and, subsequently, the model performance.
2. **Convergence in Gradient Descent**: For algorithms that use gradient descent (e.g., linear regression, neural networks), scaling can help in faster convergence. Features on different scales can cause the cost function to have an uneven surface, making it harder for the algorithm to find the optimal parameters efficiently.
3. **Interpretability**: Scaling can make the coefficients of a model more interpretable. When features are on similar scales, the magnitudes of the coefficients can be compared to understand the relative importance of each feature.

Types of Scaling

Normalized Scaling (Min-Max Scaling)

Normalized scaling transforms the data to fit within a specific range, typically [0, 1] or [-1, 1]. This is done using the formula:

$X' = (X-X_{min})/(X_{max}-X_{min})$

Where X is the original value, X_{min} is the minimum value of the feature, and X_{max} is the maximum value of the feature. This method preserves the relationships between the original data points but can be sensitive to outliers.

When to use:

- When the data has a known minimum and maximum.
- When the model does not assume normally distributed features.
- When the presence of outliers is minimal.

Standardized Scaling (Z-score Normalization)

X′= (X−μ)/ σ

Where X is the original value, μ is the mean of the feature, and σ is the standard deviation of the feature. This method centers the data around zero and scales it according to the feature's variance, making it less sensitive to outliers compared to normalized scaling.

When to use:

- When the data is assumed to follow a normal distribution.
- When features have different units and scales.
- When the algorithm assumes normally distributed features.

Key Differences

1. **Range**:
   - **Normalized Scaling**: Transforms data to a specific range, typically [0, 1].
   - **Standardized Scaling**: Transforms data to have a mean of 0 and a standard deviation of 1.
2. **Sensitivity to Outliers**:
   - **Normalized Scaling**: More sensitive to outliers since it relies on the min and max values.
   - **Standardized Scaling**: Less sensitive to outliers as it uses mean and standard deviation.
3. **Usage Context**:
   - **Normalized Scaling**: Preferred when data is bounded and outliers are not significant.
   - **Standardized Scaling**: Preferred when data is assumed to be normally distributed or when dealing with features of different units and scales.

By understanding and applying the appropriate scaling technique, one can ensure that the data is in a form that maximizes the performance and interpretability of the machine learning models.

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

An infinite value of the Variance Inflation Factor (VIF) occurs when there is perfect multicollinearity in your dataset. This means that one of the independent variables can be perfectly predicted by a linear combination of the other independent variables. Here are the main reasons why this happens:

1. **Perfect Correlation**: If two or more predictors are perfectly correlated (i.e., the correlation coefficient is 1 or -1), the VIF will be infinite. This is because the denominator in the VIF formula, (1 - R^2), becomes zero when (R^2) is 1, leading to division by zero.
2. **Redundant Variables**: Including redundant variables that do not add new information to the model can cause infinite VIF values. For example, if you include both "total sales" and "average sales per day" in a model where "total sales" is just "average sales per day" multiplied by the number of days, this redundancy can lead to perfect multicollinearity.
3. **Dummy Variable Trap**: When creating dummy variables for categorical data, if you include all categories without dropping one, you can end up with perfect multicollinearity. This is known as the dummy variable trap.

How to Address Infinite VIF Values

1. **Remove Redundant Variables**: Identify and remove variables that are perfectly correlated or redundant.
2. **Drop One Dummy Variable**: When creating dummy variables, always drop one category to avoid the dummy variable trap.
3. **Regularization Techniques**: Use regularization methods like Ridge Regression, which can handle multicollinearity by adding a penalty to the coefficients.

Understanding and addressing infinite VIF values is crucial for building stable and reliable regression models.

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

A Q-Q plot, or quantile-quantile plot, is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, typically the normal distribution. It helps to assess whether the data follows a specific distribution by plotting the quantiles of the data against the quantiles of the theoretical distribution.

How to Interpret a Q-Q Plot

1. **Straight Line**: If the points in the Q-Q plot lie approximately along a straight line, it indicates that the data follows the theoretical distribution.
2. **Deviations from the Line**: Deviations from the straight line suggest departures from the theoretical distribution. For example:
   - **Upward Curve**: Indicates a right-skewed distribution.
   - **Downward Curve**: Indicates a left-skewed distribution.

- **S-shaped Curve**: Indicates heavy tails (leptokurtic distribution).

Use and Importance in Linear Regression

1. **Assessing Normality of Residuals**: In linear regression, one of the key assumptions is that the residuals (errors) are normally distributed. A Q-Q plot of the residuals can help to visually assess this assumption. If the residuals follow a normal distribution, the points will lie on a straight line.
2. **Identifying Outliers**: Q-Q plots can help identify outliers or extreme values that deviate significantly from the theoretical distribution. These outliers can affect the regression model's performance and interpretation.
3. **Model Validation**: By checking the normality of residuals, Q-Q plots help validate the linear regression model. If the residuals are not normally distributed, it may indicate that the model is not appropriate, and transformations or different modeling techniques might be needed

Example

Imagine you have a dataset of residuals from a linear regression model. You create a Q-Q plot to check if these residuals are normally distributed. If the points in the Q-Q plot lie along a straight line, you can be confident that the residuals are normally distributed, validating one of the key assumptions of linear regression.

Q-Q plots are a valuable tool in linear regression for assessing the normality of residuals, identifying outliers, and validating the model. They provide a visual way to check if the data follows a theoretical distribution, which is crucial for making accurate inferences from the model