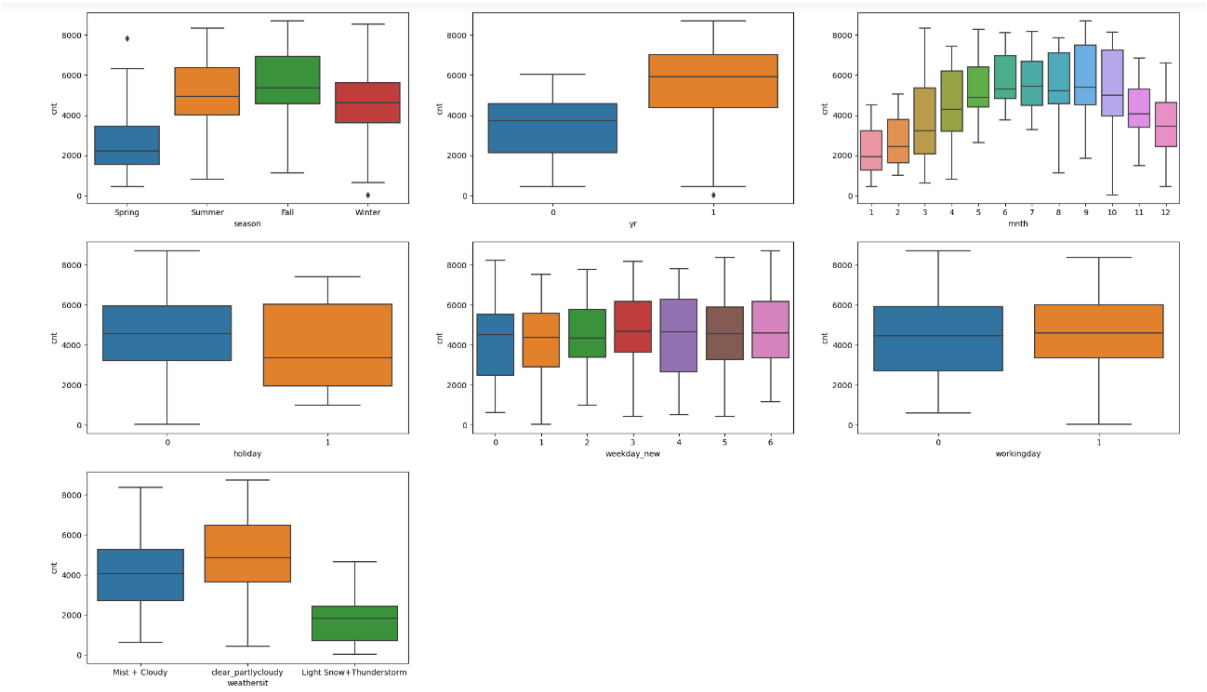
Assignment Based Subjective Questions:

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

   Answer:



The categorical variable used in the dataset: season, yr(year), holiday, weekday_new, workingday, and weathersit (weather situation) and mnth(month) . These were visualized using a boxplot. These variables had the following effect on our dependent variable:

Season - For the variable season, we can clearly see that the category 3: Fall, has the highest median, which shows that the demand was high during this season. It is least for 1: spring.

Yr - The year 2019 had a higher count of users as compared to the year 2018.

Holiday - rentals reduced during holiday.

Weekday - The bike demand is almost constant throughout the week.

Workingday – From the "Workingday" boxplot we can see those maximum bookings happening between 4000 and 6000, that is the median count of users is constant almost throughout the week. There is not much of difference in booking whether its working day or not.

Weathersit - There are no users when there is heavy rain/ snow indicating that this weather is quite adverse. Highest count was seen when the weather situation was Clear, Partly Cloudy.

Mnth - The number of rentals peaked in September, whereas they are less in December. As a result of the typical substantial snowfall in December, rentals may have declined.

2. **Why is it important to use drop_first=True during dummy variable creation?**
   Answer:
   When creating dummy variables for categorical features in linear regression, it is crucial to use drop_first=True for these reasons:

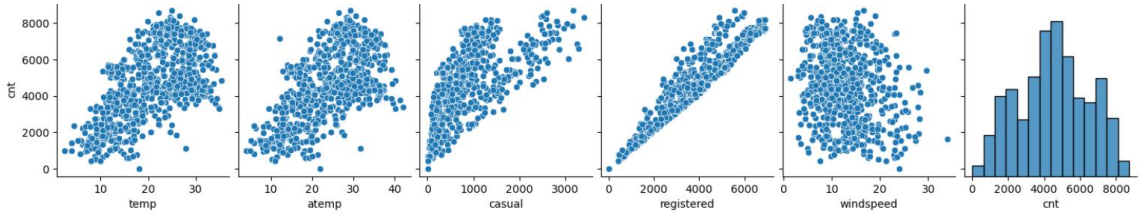
- **Reduce Model Complexity**: Dropping one dummy variable reduces the number of columns, simplifying the model and making it easier to interpret which variables significantly impact the dependent variable.
- **Avoid Multicollinearity**: Including all dummy variables can lead to multicollinearity, where dummy variables are highly correlated with each other. Using drop_first=True prevents this issue by removing one variable and avoiding perfect correlation.
- **Clarify Variable Impact**: By designating the dropped dummy variable as a baseline, the remaining variables' coefficients reflect their impact relative to this reference category, aiding in clearer interpretation.
    - If we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

For an instance, a categorical column consists of 3 types of values, we need to create the dummy variable to that column. If one variable is neither furnished nor semi_furnished, then, it is obvious unfurnished. So, we do not need 3rd variable to identify the unfurnished.

Value	Indicator Variable	
Furnishing Status	furnished	semi-furnished
furnished	1	0
semi-furnished	0	1
unfurnished	0	0

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

   Answer: Following is the pair-plot of the dependent variable vs other numeric variables:



Here the variable “registered” has the highest correlation with the target variable. However, the target variable is actually the sum of casual and registered users,

hence it would make sense to exclude them from further analysis .Additional to "casual" and "registered" , the variables "temp" and "atemp" have the highest correlation with the target variable.

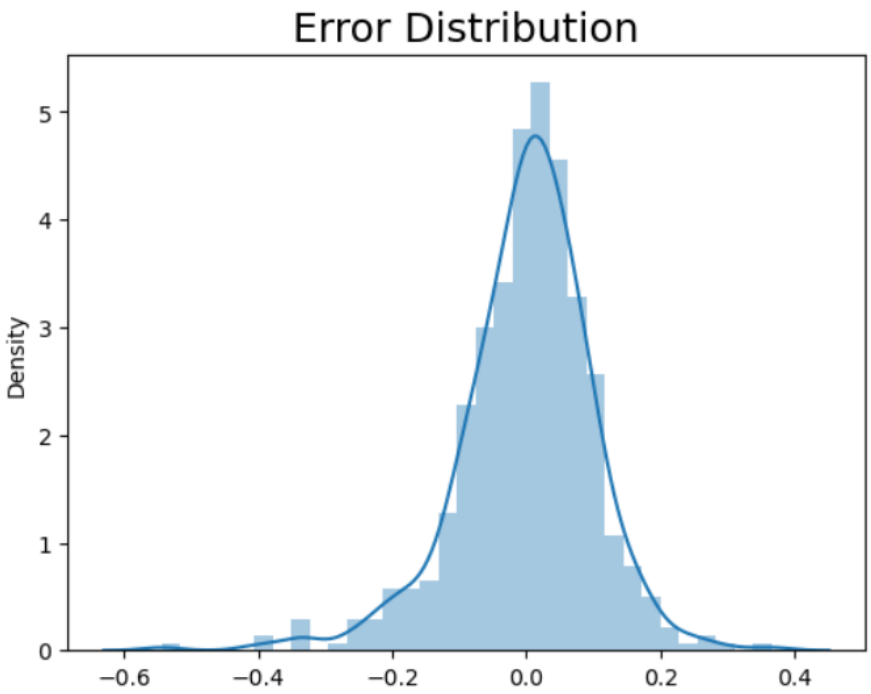
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

We have done following tests to validate assumptions of Linear Regression:

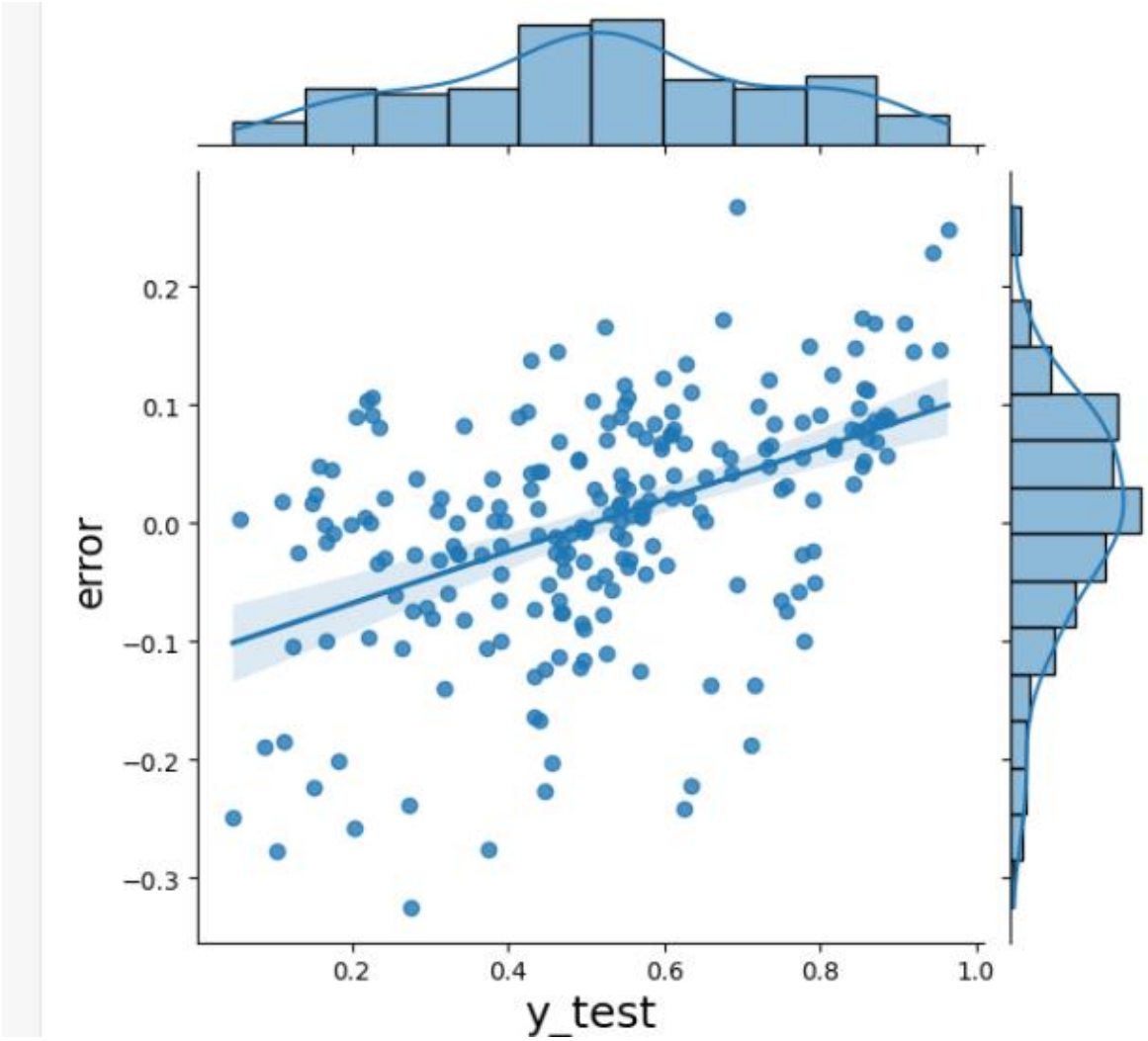
a. There should be linear relationship between independent and dependent variables. We visualised the numeric variables using a pairplot to see if the variables are linearly related or not.

b. Residuals distribution should follow normal distribution and centred around 0 (mean = 0). We validated this assumption about residuals by plotting a distplot of residuals and saw if residuals are following normal distribution or not.

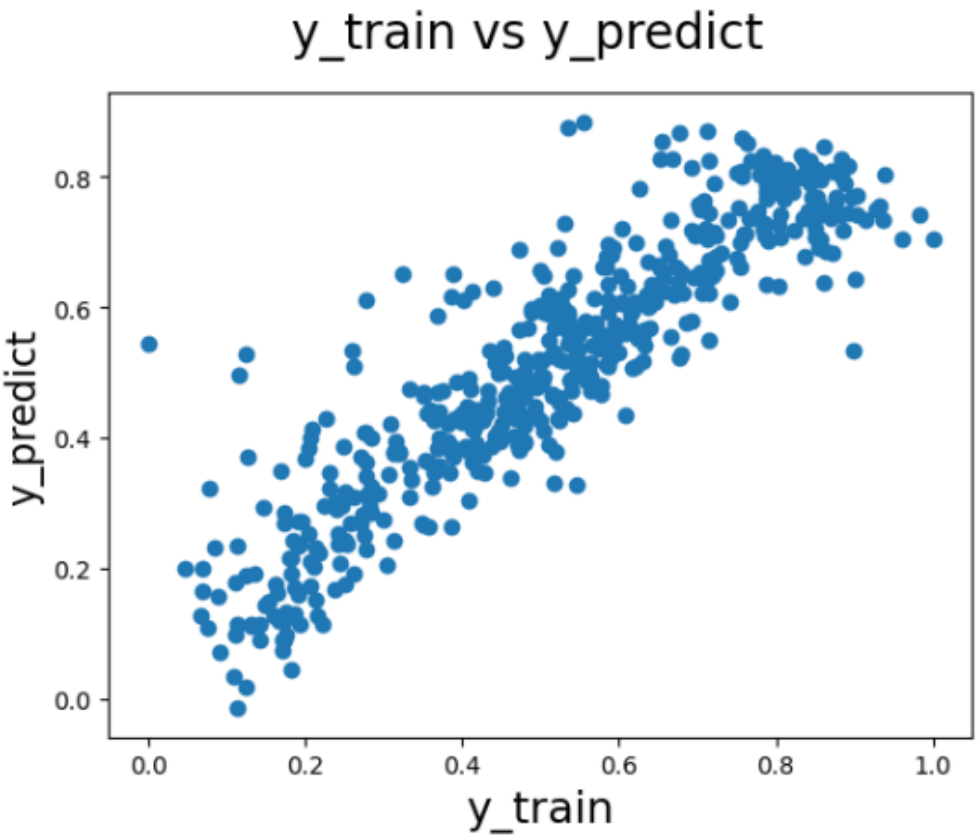
Assumption 1) Normal distribution of error terms are observed for the training dataset. Following figure shows the distribution of error terms-



Assumption 2) The error terms are independent i.e. they are not following a pattern. The error distribution does not seem independent. A higher order model might be necessary to ascertain better information for this problem. Following figure demonstrates this distribution,



Assumption 3) There is constant variance in the distribution of error terms (homoscedastic), following figure shows that



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

   Answer:
   Following is the equation mentioned below:
   cnt = 0.135 + 0.237(yr) - 0.0817(holiday) + 0.467(temp) - 0.177(windspeed) - 0.077(Spring) + 0.043(Summer) + 0.069(Winter) + 0.091(clear _partlycloudy)

   Top 3 features-
   "temp", "yr" and "clear_partlycloudy"

   - temp: For a unit increase in the ambient temperature, the target variable increases by **0.46** times
   - yr: for a unit increase in the feature "yr", the target variable increases by **0.237** times
   - clear_partlycloudy: for a unit increase in this variable, the target variable increases by **0.091** times

General Subject Questions

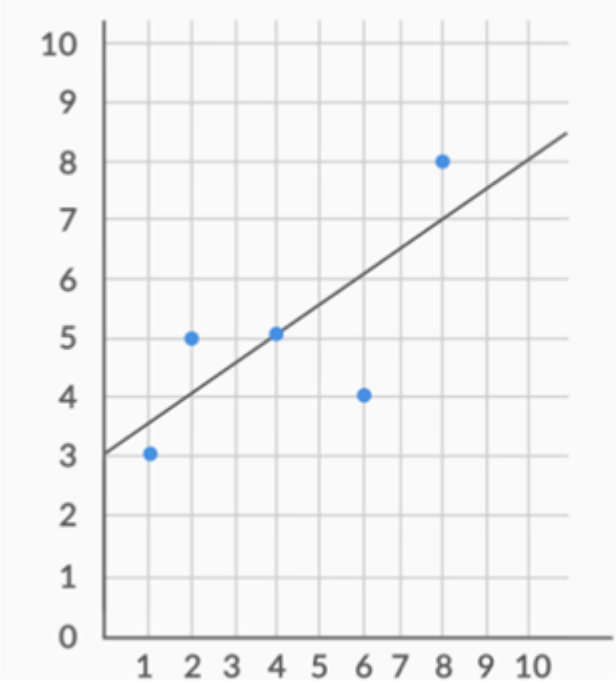
1.Explain the linear regression algorithm in detail.

Answer:

Definition: Linear regression is a supervised learning algorithm used for regression tasks, predicting a continuous output variable (y) based on one or more input variables (x).

Objective: The goal is to find a line that best fits the data points by minimizing the distance between the line and the data points.

Basic Concept: Linear regression is primarily used for understanding the impact of historical data on the target variable. Here the linear regression algorithm can be applied only for continuous data, with the assumption that there is some linear relationship between the target variable and other independent variables. The accuracy of the best fit line is obtained through the least squares method. Following figure shows a straight line fit for a scatter plot-



Equation: Following is the equation of simple linear regression model-

Formula

$$y = \alpha + \beta x$$

β = slope

α = y-intercept

y = y- coordinate

x = x-coordinate

Following is the equation of multiple linear regression model-

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

Cost Function: The Mean Squared Error (MSE) is commonly used to measure the fit of the line:

$$\text{MSE} = \frac{1}{n} \sum (y - y')^2$$

where n is the number of data points, y is the actual value, and y' is the predicted value.

Optimization Methods: To find the best values of

Intercept and slope , optimization techniques are used:

- Gradient Descent**:** Iteratively adjusts the parameters to minimize the cost function.
- Normal Equation**:** Computes the parameters directly using a closed-form solution.
- Libraries: Tools like scikit-learn provide built-in functions to perform linear regression efficiently.

Limitations:

- Assumes a linear relationship between input and output variables, which may not always hold.

- Sensitive to outliers and multicollinearity, which can affect model performance.

2. Explain the Anscombe's quartet in detail.

Answer:

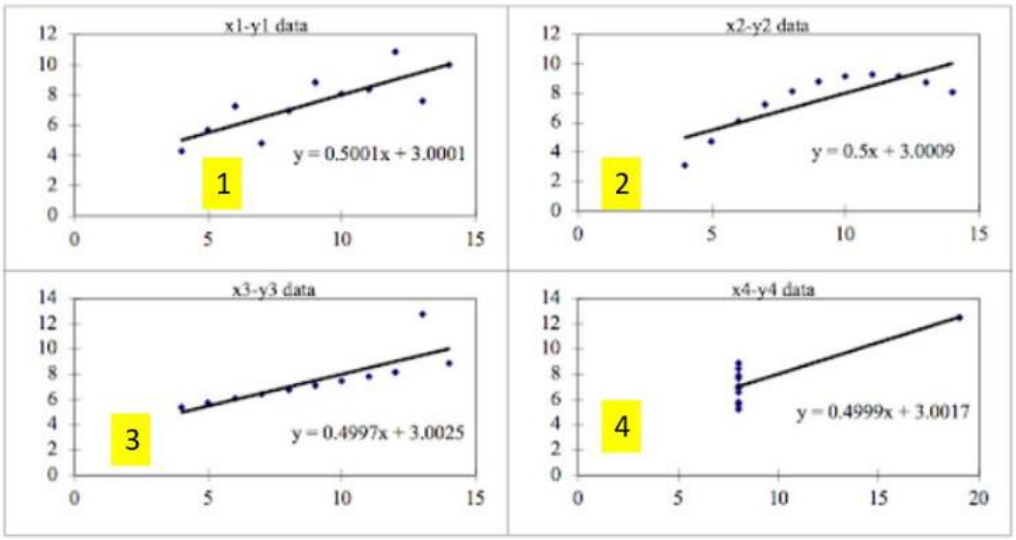
Anscombe's Quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph. It was developed to emphasize both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties.

Datasets and Descriptive Statistics: Despite having the same mean values, variances, and correlation coefficients, each dataset in Anscombe's quartet exhibits different patterns when visualized. The common statistics include:

- Mean of x and y.
- Variance of x and y.
- Correlation between x and y.
- Slope and intercept of the linear regression line.

Plots in Anscombe's Quartet:

Following figure shows the linear regression model fitted for these 4 datasets-



Dataset 1: Here the linear regression model is a good fit .

Dataset 2: The linear regression model cannot capture non-linear distribution.

Dataset 3: Linear regression model is sensitive to the outlier, resulting in an incorrect result.

Dataset 4: Here again the sensitivity towards outliers of the linear regression model is demonstrated .

Anscombe's quartet underscores the necessity of plotting data before analysis to avoid misleading conclusions and ensure accurate understanding of the data's true nature.

3. What is Pearson's R?

Answer:

Definition: Pearson's R, or the Pearson correlation coefficient, measures the strength and direction of the linear relationship between two continuous variables.

Calculation:

- Determined by dividing the covariance of the variables by the product of their standard deviations.
- Formula:

$$\text{Formula: } r = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

.

Range: The coefficient ranges from -1 to 1:

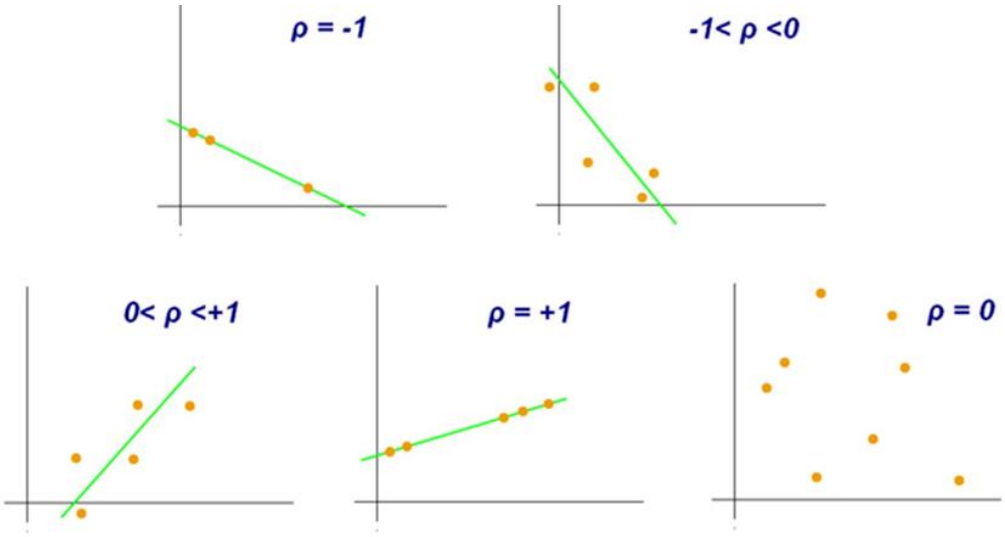
Following table provides Pearson correlation coefficient table range-

Pearson Correlation Coefficient (r) Range	Type of Correlation	Description of Relationship	New Illustrative Example
$0 < r \leq 1$	Positive	An increase in one variable associates with an increase in the other.	**Study Time vs. Test Scores:** More hours spent studying tends to lead to higher test scores.
$r = 0$	None	No discernible relationship between the changes in both variables.	**Shoe Size vs. Reading Skill:** A person's shoe size doesn't predict their ability to read.
$-1 \leq r < 0$	Negative	An increase in one variable associates with a decrease in the other.	**Outdoor Temperature vs. Home Heating Cost:** As the outdoor temperature decreases, heating costs in the home increase.

Following table illustrates the general rules of thumb followed while interpreting the pearson's correlation coefficient

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

As can be seen from the graph below, r = 1 means the data is perfectly linear with a positive slope r = -1 means the data is perfectly linear with a negative slope r = 0 means there is no linear association



Assumptions:

- The relationship between variables is linear.
- Data is normally distributed.

Applications: Used in various fields like psychology, finance, and social sciences to determine the relationship between variables.

Limitations:

- Sensitive to outliers, which can distort the correlation.
- Only measures linear relationships, not suitable for non-linear data.

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

.Answer:

a. What is scaling? Why is scaling performed?

Scaling of features is an import data pre-processing step, done to ensure that all the values are within the same range of magnitude of 0 to 1. If scaling is not done then it results in incorrect values of coefficients of the respective variables, as a result the model would end up assigning incorrect weightage to some variables (owing to their magnitudes). Hence it is important to normalize the magnitudes of all the continuous variables (dummy variables are excluded).

b. What is the difference between normalized scaling and standardized scaling?

Standard scaling: the feature is scaled by subtracting the mean from all the data points and dividing the resultant values by the standard deviation of the data. Following is the formula-

$$X_{\text{scaled}} = \frac{X_i - X_{\text{mean}}}{\sigma}$$

Normalized scaling (Min-max scaling): Here the data point is subtracted with the minimum value from the datapoint and result is divided by the difference between the maximum and the minimum value. Following is the formula-

$$X_{\text{scaled}} = \frac{X_i - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

Normalized scaling results in the feature values scaled to magnitudes between 0 to 1, while standardized scaling scales the feature in which the datapoints are having a mean of zero and standard deviation of 1 .

Standardized scaling is useful when the distribution of the feature values is not known or when there are outliers in the data.

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

Answer:

Definition: VIF measures how much the variance of a regression coefficient is inflated due to multicollinearity among the independent variables.

Calculation: VIF for a predictor is calculated where , Ri^2 is the R-squared value obtained by

$$VIF = \frac{1}{1 - R^2}$$

regressing the predictor against all other predictors.

Interpretation: A higher VIF indicates a higher degree of multicollinearity. Specifically, it quantifies how much the variance of a coefficient is increased due to collinearity.

VIF = Infinity: If the independent variable can be perfectly explained by other independent variables (perfect multicollinearity

$$R_i^2 = 1$$

leading to VIF being infinite.

Thresholds:

- **VIF = 1**: No multicollinearity (predictors are not correlated).
- **1 < VIF < 5**: Moderate multicollinearity (predictors are moderately correlated).
- **VIF > 5**: High multicollinearity (predictors are highly correlated), often warranting further investigation or corrective action.

Implications of High VIF: High VIF values suggest that the estimated coefficients are unreliable due to multicollinearity, leading to large standard errors and less stable predictions.

Effect on Model Interpretation: Multicollinearity does not affect the predictive power or reliability of the model as a whole but affects the individual coefficient estimates, making them difficult to interpret.

Variance Inflation: The numerical value of VIF indicates the percentage increase in the variance of the coefficient due to multicollinearity. For example, a VIF of 1.9 means the variance is inflated by 90%.

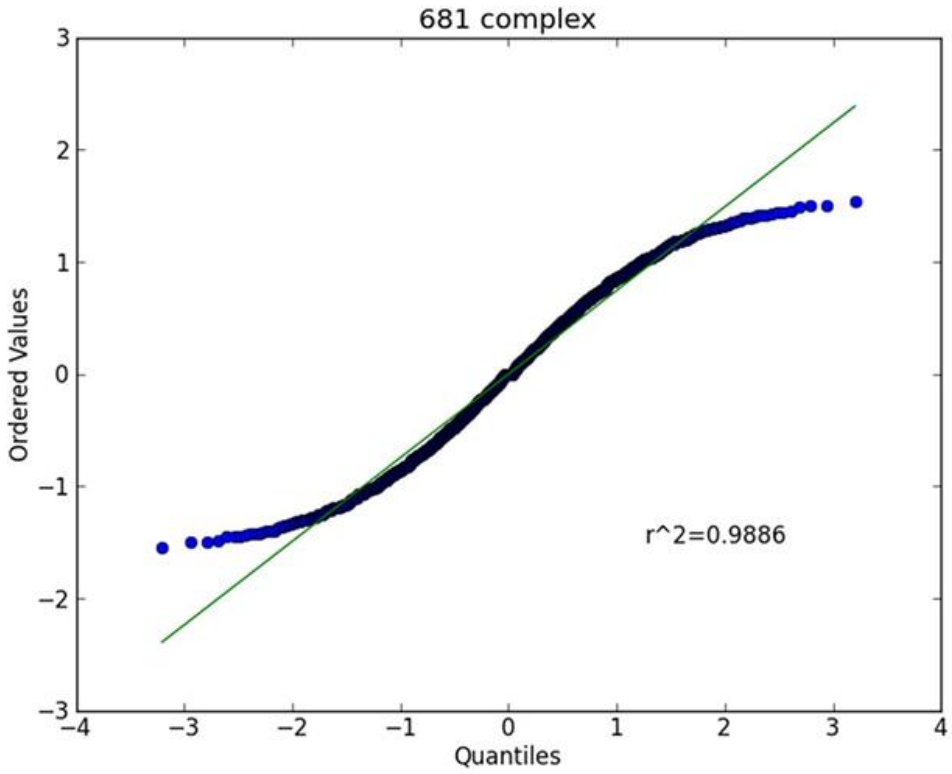
Addressing High VIF: Techniques to handle high VIF values include removing highly correlated predictors.

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

Answer:

Definition: Quantile-quantile plot is used to determine whether two samples of data belong to the same population. The quantiles of the first dataset is plotted against the quantiles of the second dataset, if the two samples belong to the same population then the points will lie along the same line.

For an instance,



Importance of Q-Q plot:

1. **Checking Normality of Residuals**: One of the key assumptions of linear regression is that the residuals should be normally distributed. A Q-Q plot helps visually assess this assumption.
2. **Detecting Outliers**: Q-Q plots can help identify outliers, as points that deviate significantly from the reference line indicate unusual data points.
3. **Model Validation**: Ensuring that the residuals are normally distributed helps validate the appropriateness of the linear model. Deviations from normality may suggest the need for a different model or transformations.
4. **Identifying Data Transformations**: If the residuals are not normally distributed, a Q-Q plot can indicate the type of transformation (e.g., log, square root) needed to meet the normality assumption.

Use of Q-Q plot:

- Determine the distribution of the sample (normal, uniform etc…)
- Identify whether two samples belong to the same population