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?**

The categorical variable in the dataset were season, weathersit, holiday,mnth,yr and weekday. These were visualized using a boxplot .These variables had the following effect on our dependant variable:-

- **Season -** The boxplot showed that spring season had least value of cnt whereas fall had maximum value of cnt. Summer and winter had intermediate value of cnt.
- **Weathersit** - There are no users when there is heavy rain/ snow indicating that this weather is extremely unfavourable. Highest count was seen when the weathersit was' Clear, Partly Cloudy'.
- **Holiday -** rentals reduced during holiday.
- **Mnth -** September saw highest no of rentals while December saw least. This observation is on par with the observation made in weathersit.The weather situation in december is usually heavy snow.
- **Yr** - The number of rentals in 2019 was more than 2018

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

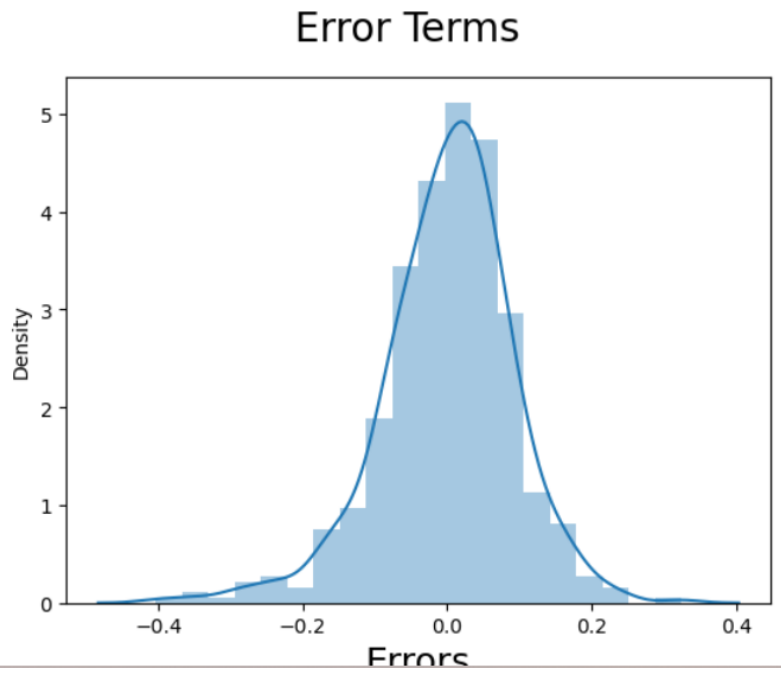
If you don't drop the first column then your dummy variables will be correlated (redundant). This may affect some models adversely and the effect is stronger when the cardinality is smaller. For example iterative models may have trouble converging and lists of variable importance's may be distorted. Another reason is, if we have all dummy variables it leads to Multicollinearity between the dummy variables. To keep this under control, we lose one column.

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

"**temp**" and "**atemp**" are the two numerical variables which are highly correlated with the target variable (cnt).

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

```
In [58]: # Plot the histogram of the error terms
fig = plt.figure()
sns.distplot((y_train - y_train_cnt), bins = 20)
fig.suptitle('Error Terms', fontsize = 20)                  # Plot heading
plt.xlabel('Errors', fontsize = 18)                         # X-label
plt.show()
```



Residuals distribution should follow normal distribution and centered around 0.(mean = 0). We validate this assumption about residuals by plotting a distplot of residuals and see if residuals are following normal distribution or no. The above diagram shows that the residuals are distributed about mean = 0.

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

The top features are:

- Temp- 0.569
- Yr – 0.232
- Weathersit - Light_snow and rain - -0.281

General Subjective Questions

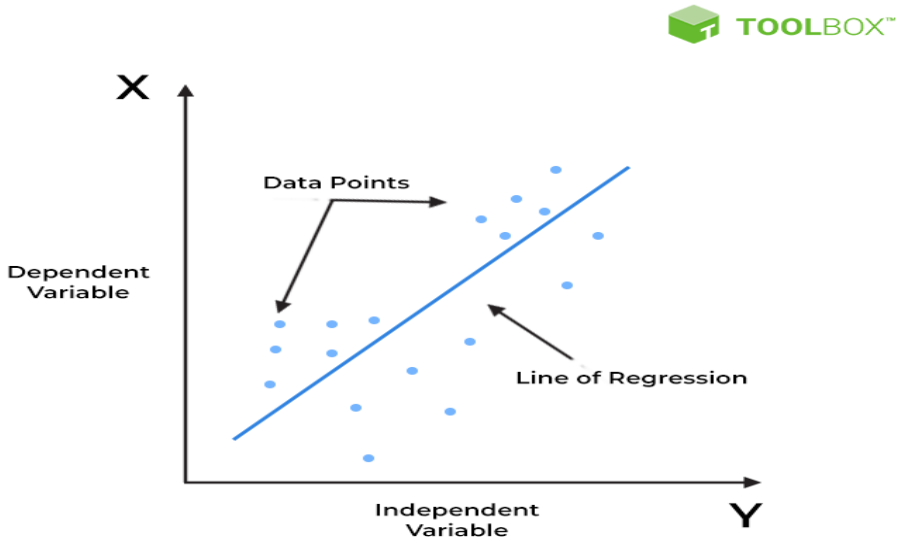
1. **Explain the linear Regression algorithm in detail.**

Linear Regression is a method of finding best straight line fitting to given data, i.e. finding the best linear relationship between the independent and dependent variables. In technical terms linear regression is a machine learning algorithm that finds best linear-fit relationship on any given data, between independent and dependent variables. It is mostly done by the sum of squared Residual method.

The independent variable is also called predictor or explanatory variable that remains unchanged due to the change in other variables. However, the dependent variable change with fluctuations in the independent variables. The regression model predicts the value of the dependent variable, which is the response or outcome variable being analyzed.

Thus, linear regression is a supervised learning algorithm that simulates a mathematical relationship between variables and makes predictions for continuous or numeric variables such as sales, salary, age, product price etc.

This analysis method is advantageous when atleast two variables are available in the data, as observed in stock market forecasting, portfolio management, scientific analysis, etc.



Here, a line is plotted for the given data points the suitably fitted all the issues. Hence, it is called a “best fit line”. The goal of the linear regression is to find this best fit line as seen in the above figure.

Linear Regression Equation:

$Y = m*X + b$

Where X = Independent variable

Y – Dependent variable (Target variable)

M = slope of the line

Types of Linear Regression:

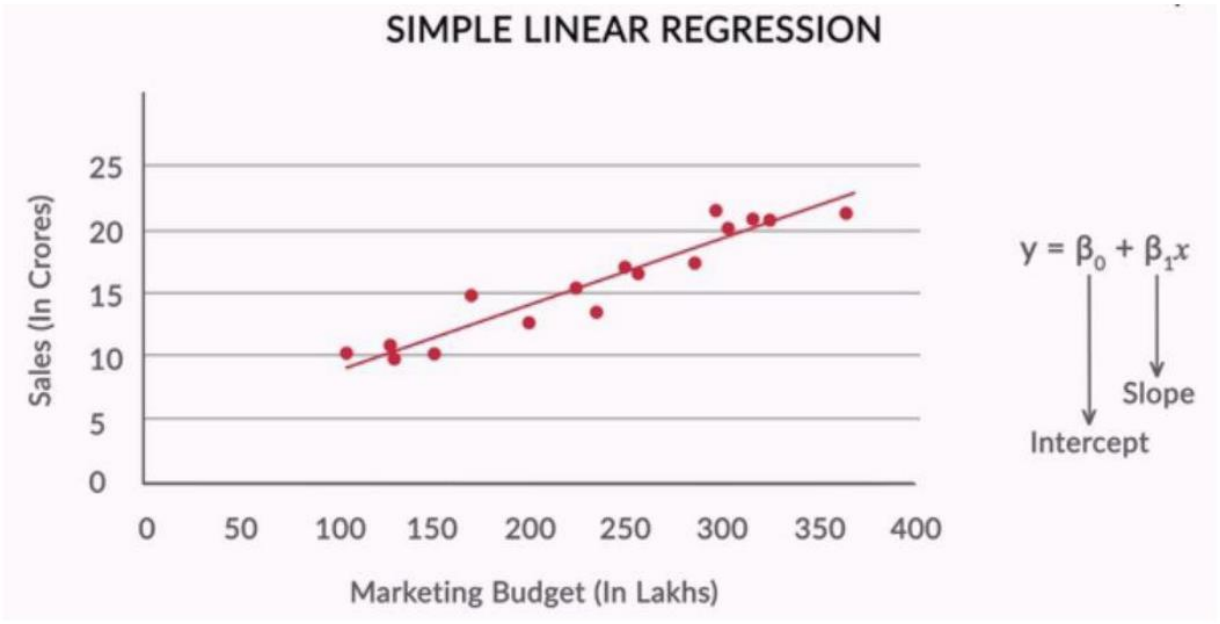
1. **Simple Linear Regression**
2. **Multiple Linear Regression**

Simple Linear Regression:

The most elementary type of regression model is the simple linear regression which explain the relationship between a dependent variable and one independent variable using a straight line.

The standard equation of the simple linear regression line is given by the following expression.

$Y = B_0 + B_1X$



Multiple Linear Regression:

Multiple linear regression is a statistical technique to understand the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X

The formulation for multiple linear regression is also similar to simple linear regression with the small change that instead of having beta for just one variable, you will now have betas for all the variables used. The formula now can be simply given as:

$Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + \ldots\ldots + B_pX_p$

Apart from the formula, a lot of other ideas in multiple linear regression are also similar to simple linear regression, such as:

1. Model now fits a 'hyperplane' instead of a line.
2. Coefficients still obtained by minimizing sum of squared error (Least squares criterion)
3. For inference, the assumptions from Simple Linear Regression still hold Zero means, independent, normally distributed error terms that have constant variance.

Assumption to build Linear Regression model:

1. Linear relationship between X and Y
2. Error terms are normally distributed (not X, Y
3. Error terms are independent of each other.
4. Error terms have constant variance (homoscedasticity)
5. Model may "Overfit" by becoming too complex.
6. Multicollinearity – Association between predictor variables.

2. **Explain the Anscombe's quartet in detail.**

Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

Purpose of Anscombe's Quartet:

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

Importance of Anscombe's Quartet:

Visualization: It highlights the crucial role of visualizing data. Graphs can reveal patterns, trends, and anomalies that descriptive statistics alone may not uncover.

Outliers: It underscores the influence of outliers on statistical measures and the importance of detecting them.

Model Validity: The datasets demonstrate how the same statistical summaries can arise from very different data distributions, emphasizing the need for careful analysis and model validation.

Anscombe's Quartet Dataset

The four datasets of Anscombe's Quartet.

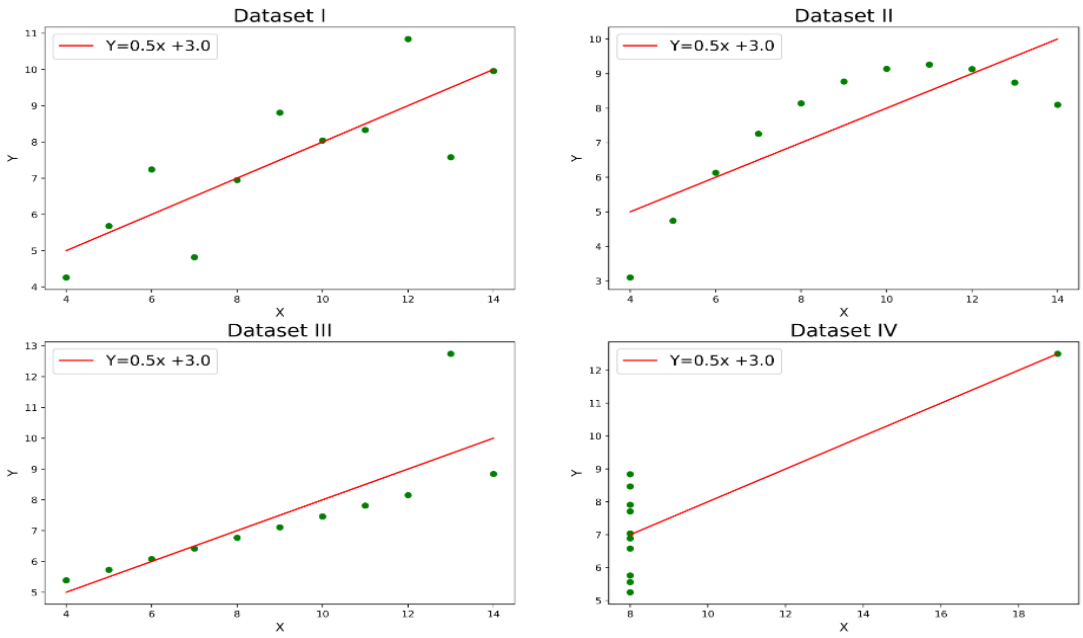
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Common Statistics:

All four datasets share the following statistics:

- **Mean of x**: 9
- **Mean of y**: 7.5
- **Variance of x**: 11
- **Variance of y**: Approximately 4.12
- **Correlation between x and y**: 0.816
- **Linear regression line**: y=3.00+0.50xy=3.00+0.50x

Despite these similarities, when the datasets are plotted, they reveal very different patterns:



Explanation of this output:

1. **Dataset 1:** A simple linear relationship with some random noise.
2. **Dataset 2:** A clear curve, indicating a non-linear relationship.
3. **Dataset 3:** A linear relationship but with an outlier that significantly affects the regression line.
4. **Dataset 4:** Most data points are the same with one extreme outlier, leading to a distorted regression line.

3. What is Pearson's R?

The **Pearson correlation coefficient (*r*)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

- 1 indicates a perfect positive linear correlation.
- -1 indicates a perfect negative linear correlation.
- 0 indicates no linear correlation.

The Pearson correlation coefficient (*r*) is the most widely used correlation coefficient and is known by many names:

- Pearson's *r*
- Bivariate correlation
- Pearson product-moment correlation coefficient (PPMCC)

- The correlation coefficient

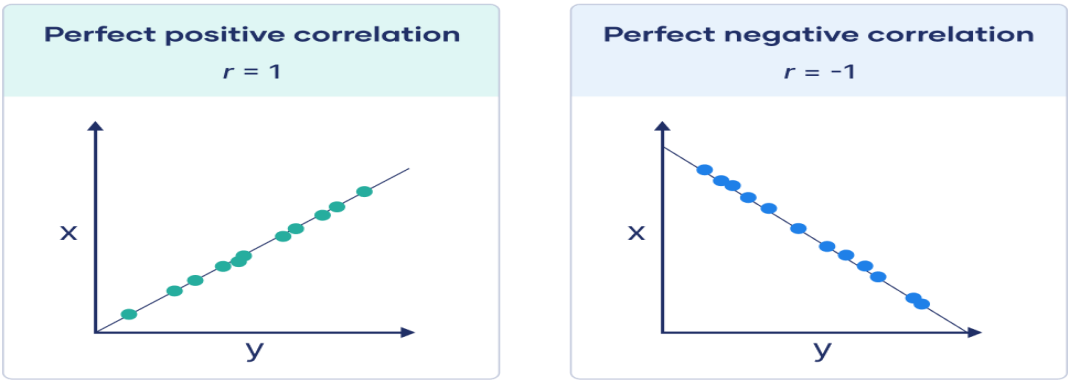
The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

The Pearson correlation coefficient is also an inferential statistic, meaning that it can be used to test statistical hypotheses. Specifically, we can test whether there is a significant relationship between two variables.

Visualizing the Pearson correlation coefficient

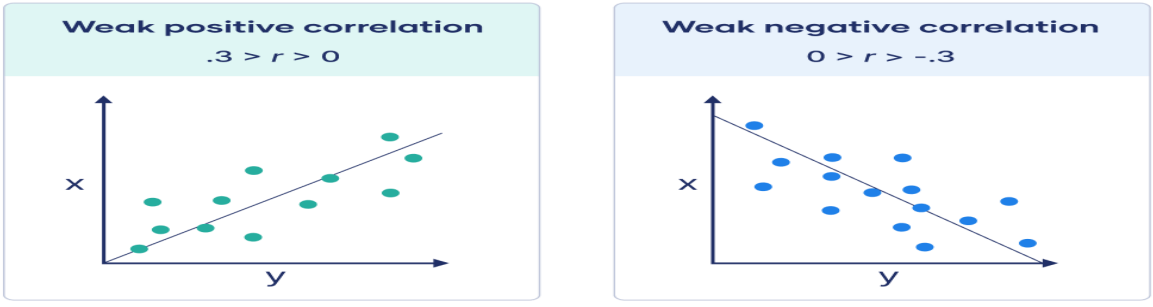
Another way to think of the Pearson correlation coefficient (r) is as a measure of how close the observations are to a line of best fit. The Pearson correlation coefficient also tells you whether the slope of the line of best fit is negative or positive. When the slope is negative, r is negative. When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

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



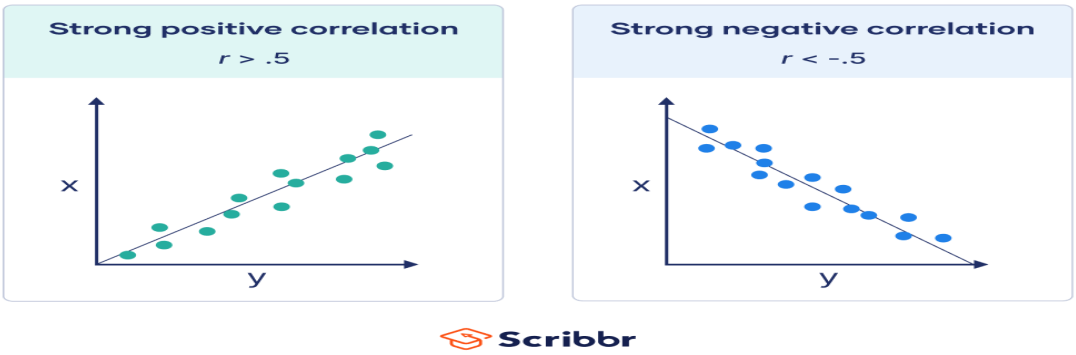
Scribbr

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

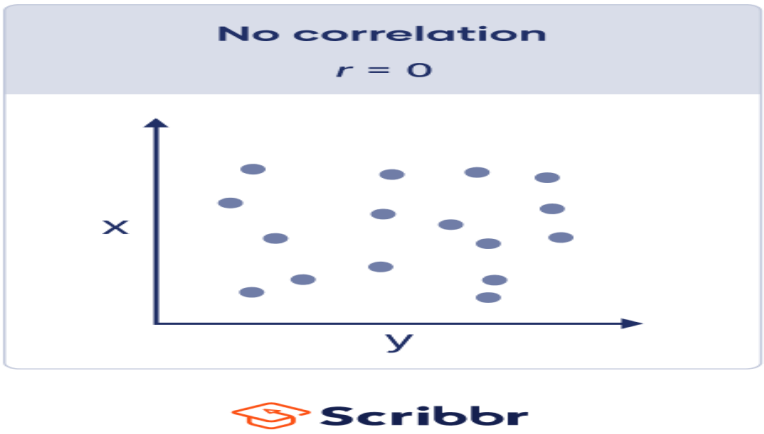


Scribbr

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



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



Calculation:

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

where:

- x_i and y_i are the individual data points.
- $\overline{x}$ and $\overline{y}$ are the means of the x and y variables, respectively.

Properties:

1. **Symmetry:** Pearson's R is symmetric, meaning the correlation between x and y is the same as the correlation between y and x.
2. **Unitless**: It is a dimensionless index, meaning it does not depend on the units of measurement of the variables.
3. **Sensitivity to Outliers**: Pearson's R is sensitive to outliers, which can disproportionately affect the correlation coefficient.

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

Scaling is a data preprocessing technique used to adjust the values of numeric features so that they are on a similar scale. This is essential in many machine learning algorithms that rely on the distances between data points (e.g., k-nearest neighbors, support vector machines, and gradient descent-based algorithms). Without scaling, features with larger ranges may dominate the computation, leading to biased models.

Scaling is performed due to,

1. **Improves Algorithm Performance**: Many machine learning algorithms perform better or converge faster when the features are on a similar scale. Algorithms that use gradient descent, for instance, converge more quickly when the feature values are standardized.
2. **Prevents Dominance of High-Range Features**: Features with larger ranges can disproportionately affect the results of models that depend on distance measurements, like clustering algorithms and k-nearest neighbors.
3. **Enhances Model Interpretability**: Consistent feature scaling allows for more meaningful comparison and interpretation of model coefficients, especially in linear models.
4. **Prepares Data for Models Sensitive to Feature Magnitudes**: Some algorithms are sensitive to the magnitude of feature values. Scaling ensures that all features contribute equally to the result.

Difference Between Normalized Scaling and Standardized Scaling

1. **Normalized Scaling:**

Definition: Normalization typically rescales the data to a fixed range, usually between 0 and 1 or -1 and 1.

Formula:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Purpose: It adjusts the values of features so that they are proportional to each other within a specific range. It is particularly useful when the distribution of the data does not follow a Gaussian distribution or when we want to retain the relationships in the data.

Use Case: Useful in algorithms where absolute values are important, like image processing.

2. Standardized Scaling:

Definition: Standardization transforms the data to have a mean of 0 and a standard deviation of 1.

Formula:

$$x' = \frac{x - \mu}{\sigma}$$

Where μ: μ is the mean of the feature and

σ: σ is the standard deviation.

Purpose: It adjusts the data to follow a standard normal distribution (bell curve). This is useful for algorithms that assume the data follows a Gaussian distribution.

Use Case: Suitable for many machine learning algorithms, especially those that assume the data is normally distributed, such as linear regression, logistic regression, and other algorithms that use distance calculations.

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

Variance Inflation Factor (VIF) is a measure used to detect the severity of multicollinearity in regression analysis. Multicollinearity occurs when one predictor variable in a model can be linearly predicted from the others with a substantial degree of accuracy.

How VIF is calculated:

$VIF(X_i) = 1/1\text{-}R_i^2$

Where R_i^2 is the co-efficient of determination obtained by regression X_i against all the other predictor variables in the model.

Why VIF Can Be Infinite

The VIF value can become infinite due to the following reasons:

Perfect Multicollinearity:

This occurs when a predictor variable is perfectly linearly related to one or more other predictor variables. In mathematical terms, this means that $R_i^2 = 1$ when X_i is regressed on the other predictors.

When $R_i^2 = 1$, the formula for VIF become,

$VIF(X_i) = 1/1-R_i^2 = 1/1-1 = 1/0$

Which results in an undefined or infinite value.

This indicates that the predictor variable X_i can be perfectly explained by other predictors, leading to an exact linear dependency.

Implications of Infinite VIF

Unreliable Estimates: The presence of perfect multicollinearity means that the regression coefficients are not uniquely estimable. This leads to unreliable estimates and highly inflated standard errors.

Model Instability: Infinite VIF suggests that the model is unstable, and the results are not interpretable.

Inability to Determine Individual Predictor Effects: With perfect multicollinearity, it is impossible to isolate the effect of each predictor variable on the dependent variable because the predictors are indistinguishable from each other in terms of their contribution to the model.

How to Address Infinite VIF

Remove Perfectly Collinear Predictors: Identify and remove one or more of the perfectly collinear predictors from the model.

Combine Predictors: Combine collinear variables into a single predictor through methods like Principal Component Analysis (PCA).

Regularization Techniques: Use regularization methods like Ridge Regression or Lasso, which can handle multicollinearity by adding a penalty to the regression coefficients.

Feature Engineering: Create new features or transform existing ones to reduce collinearity.

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

The quantile-quantile (q-q plot) plot is a graphical method for determining if a dataset follows a certain probability distribution or whether two samples of data came from the same population

or not. Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or if it follows some other known distribution. They are commonly used in statistics, data analysis, and quality control to check assumptions and identify departures from expected distributions.

How to Construct a Q-Q Plot

Order the Data: Sort the sample data in ascending order.

Calculate Quantiles: Compute the quantiles of the sample data and the corresponding quantiles of the theoretical distribution.

Plot the Points: Plot the sample quantiles on the y-axis against the theoretical quantiles on the x-axis.

Draw Reference Line: Draw a 45-degree reference line (y = x) which represents a perfect match between the sample and theoretical distributions.

Use and Importance of a Q-Q Plot in Linear Regression

In the context of linear regression, Q-Q plots are particularly useful for assessing the assumption of normality of the residuals (errors). Here's why this is important and how Q-Q plots are used:

1. **Assessing Normality of Residuals:**

- **Assumption**: One of the key assumptions in linear regression is that the residuals (errors) are normally distributed. This assumption underpins the validity of hypothesis tests and confidence intervals for the regression coefficients.
- **Usage**: By plotting the residuals against a normal distribution in a Q-Q plot, we can visually inspect whether the residuals follow a normal distribution. If the points lie along the reference line, the residuals are approximately normally distributed.

2. **Detecting Deviations from Normality:**

- **Heavy Tails**: If the residuals have heavier tails than the normal distribution, the points will deviate from the reference line, curving upwards or downwards at the ends.
- **Skewness**: If the residuals are skewed, the points will form a curve that deviates from the reference line, indicating either right or left skewness.

3. **Model Diagnostics and Improvement:**

- **Identify Problems**: A Q-Q plot can help identify issues such as outliers, skewness, and heavy tails in the residuals, which might suggest the need for transforming the data or using a different modeling approach.
- **Model Refinement:** If the Q-Q plot indicates non-normality, remedial measures such as transforming the dependent variable (e.g., using a logarithmic or square root transformation) or applying robust regression techniques can be considered.

Interpreting Q-Q Plots

Straight Line: If the points on the Q-Q plot follow the 45-degree reference line closely, it indicates that the residuals are normally distributed.

Systematic Deviations: Patterns such as S-shaped curves or deviations at the ends of the plot indicate departures from normality, such as skewness or heavy tails.

Outliers: Points that are far away from the reference line can indicate outliers in the residuals.