Assignment-based Subjective Questions (Solutions)

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

Solution:

After few many iterations, we found the suitable model with significant variables and that is Model 6 and the cost function for the same is:

Cnt = 0.264576 * const + 0.234151 * yr -0.109022 * holiday - 0.022600 * workingday + 0.436214 * temp -0.159031 * windspeed - 0.071883 * season_spring + 0.033886 * season_summer + 0.089846 * season_winter - 0.046208 * mnth_Dec -0.050298 * mnth_Jan - 0.050926 * mnth_Jul -0.041068 * mnth_Nov + 0.067735 * mnth_Sep - 0.293908 * weathersit_Light Snow -0.083300 * weathersit_Mist + Cloudy

All the variables that have (+ve) coefficient such as: "Year, Temperature, Season_Summer, Season Winter, Mnth_Sep. Since they have the positive co efficient, the cnt(Dependent Variable) will increase if these variables increase.

On the other hand, the variables that have negative coefficients such as: holiday, windspeed, season_spring, mnth_Dec, mnth_Jan, mnth_Jul, mnth_Nov, weathersit_light show and wathersit_mist+cloudy. Since they have the negative co efficient, the cnt(Dependent Variable) will decrease if these variables increase.

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

Solution:

The get_dummies is a function in pandas that creates the category wise dummy variables with respect to the columns. This function has features **drop_first**, which drops the first column of the output dummies variable columns. This **drop_first=True** is important while creating the dummy columns for below advantages:

1. Multicollinearities between the variables will be avoided by dropping one of the columns

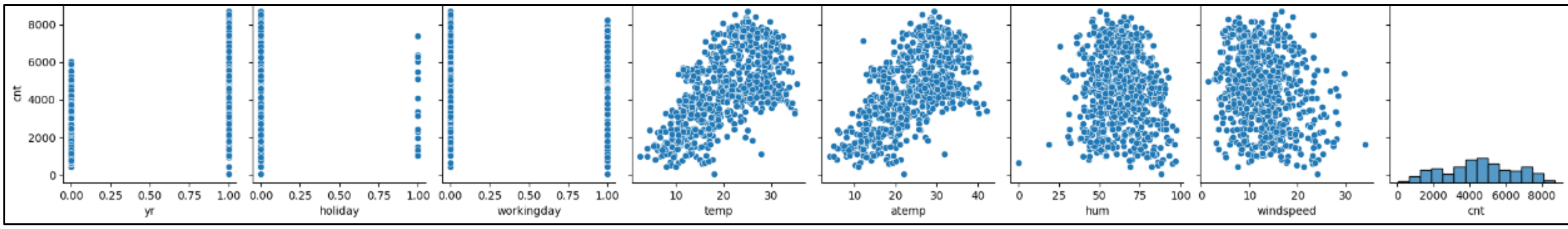
   a. Including all category levels of a categorical variable as dummy variables can introduce multicollinearity in the model. Multicollinearity occurs when there is a high correlation between predictor variables, which can cause issues in regression analysis. By dropping the first category level, you create a reference category, and the remaining dummy variables become independent of each other, reducing multicollinearity.

2. Uniqueness of coefficient
3. The efficiency will increase and the complexity of the model will be decreased
    a. **Interpretability:** Dropping the first category level makes it easier to interpret the coefficients of the remaining dummy variables. When all category levels are included, the interpretation of the coefficients becomes less straightforward. The dropped category becomes the reference category, and the coefficients of the remaining variables represent the difference between each category level and the reference category.

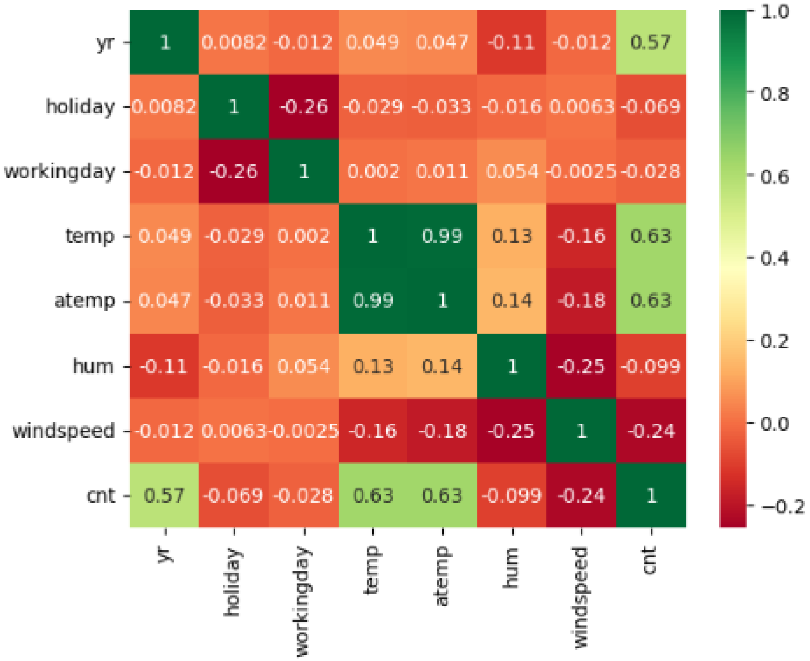
Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Solution:

The linear relationship Target Variable (cnt) and all independent variables can be seen in the below Scatter plot (pair-plot) image



The variable 'temp' and 'atemp' has the highest correlation of **0.63** with the Target variable **'cnt'.** The second highest is **0.57** from the Variable 'Year'.

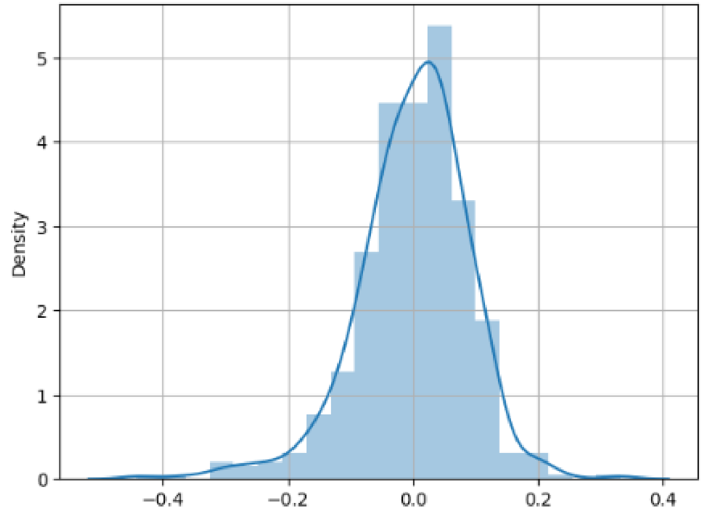


Note: this is the minimized version of the heatmap. The full version is available in the notebook.

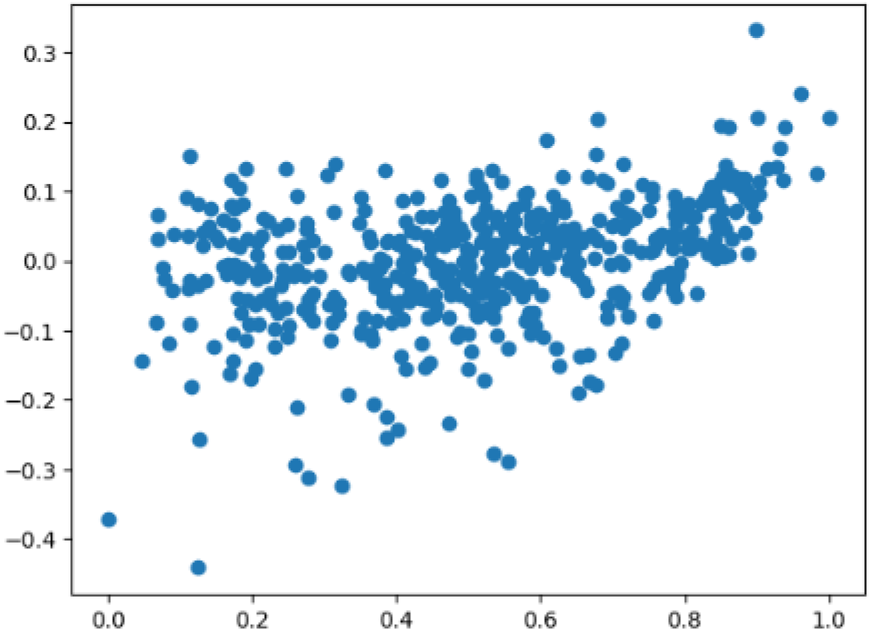
Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Solution:

1. **Normal distribution of error:** Calculated the residuals from the training module and plotted the distribution plot/histogram and verify the mean of this normal distribution of error/residuals near zero. Refer the below image for reference:



2. **Homosc** and the difference between original value and predicated values and verify if the values follow any type of pattern format. If any specific pattern on the scatter plot may indicate homoscedasticity. If No patters then the values are fit for the further predictions. Refer the below image for reference



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

Solutions:

The Cost function of our model is

> **Cnt** = 0.264576 * const + 0.234151 * yr -0.109022 * holiday - 0.022600 * workingday + 0.436214 * temp -0.159031 * windspeed - 0.071883 * season_spring + 0.033886 * season_summer + 0.089846 * season_winter - 0.046208 * mnth_Dec -0.050298 * mnth_Jan - 0.050926 * mnth_Jul -0.041068 * mnth_Nov + 0.067735 * mnth_Sep - 0.293908 * weathersit_Light Snow -0.083300 * weathersit_Mist + Cloudy

Based on the Cost function, the top 3 features contributing significantly towards explaining the demand of the shared bikes are:

1. Temp (Temperature)
2. Yr (Year)
3. season_Winter

General Subjective Questions

Q1. Explain the linear regression algorithm in detail**.**

Solution:

Linear Regression is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable i.e it finds the linear relationship between the dependent(y) and independent variable(x).

Linear Regression is of two types: **Simple** and **Multiple**.

Simple Linear Regression is where only one independent variable is present and the model has to find the linear relationship of it with the dependent variable
Equation of Simple Linear Regression, where b_o is the intercept, b_1 is coefficient or slope, x is the independent variable and y is the dependent variable.

$$y= b_o + b_1 x$$

In **Multiple Linear Regression** there are more than one independent variables for the model to find the relationship.
Equation of Multiple Linear Regression, where bo is the intercept, b_1,b_2,b_3,b_4…,b_n are coefficients or slopes of the independent variables x_1,x_2,x_3,x_4…,x_n and y is the dependent variable.

$$y = b_o + b_1x_1 + b_2x_2 + b_3x_3 \ldots. + b_nx_n$$

A Linear Regression model’s main aim is to find the best fit linear line and the optimal values of intercept and coefficients such that the error is minimized.

The line that fits the data best will be the one for which the n prediction errors (one for each observed data point) are as small as possible in some overall sense. One way to achieve this is to use the least squares criterion, which minimizes the sum of all the squared prediction errors. The cost function

$$\frac{Min}{J\left(\beta_0, \beta_1\right)} \quad where,\ J\left(\beta_0, \beta_1\right) = \sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n}\left(y_i - y_{Predicted}\right)^2 = \sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)^2$$

where β0 is intercept & β1, β2, β3,…βn are coefficient of independent variable

The coefficients (β values) are estimated using a method called Ordinary Least Squares (OLS). OLS aims to find the values of β that minimize the MSE. This is done by taking the derivative of the loss function with respect to each coefficient and setting it to zero. The resulting equations can be solved to obtain the optimal coefficient values.

Once the coefficients are estimated, the model is considered "fit" to the data. It can be used to make predictions on new data points by plugging in the values of the independent variables into the linear equation.

The coefficients obtained from linear regression represent the relationship between each independent variable and the dependent variable. Positive coefficients indicate a positive relationship, while negative coefficients indicate a negative relationship. The magnitude of the coefficient indicates the strength of the relationship, and the intercept term represents the predicted value when all independent variables are zero.

There are two approaches to minimizing/optimizing the cost function:
1) Standard mathematical approach
2) Iterative methods
    a. First Order differentiation (Gradient Decent)
    b. Second Order differentiation (Newton’s method)

The linear regression algorithm output model strength is verified by R-Squared and Root Mean Squared Error (RMSE) metrics.

Assumptions of Linear Regression algorithm:

1. Linearity of residual
2. Independence of residuals
3. Normal distribution of residual
4. Equal variance of residuals
5. No multicollinearity

Linear regression is a widely used and interpretable algorithm for predicting continuous outcomes based on the relationship between variables.

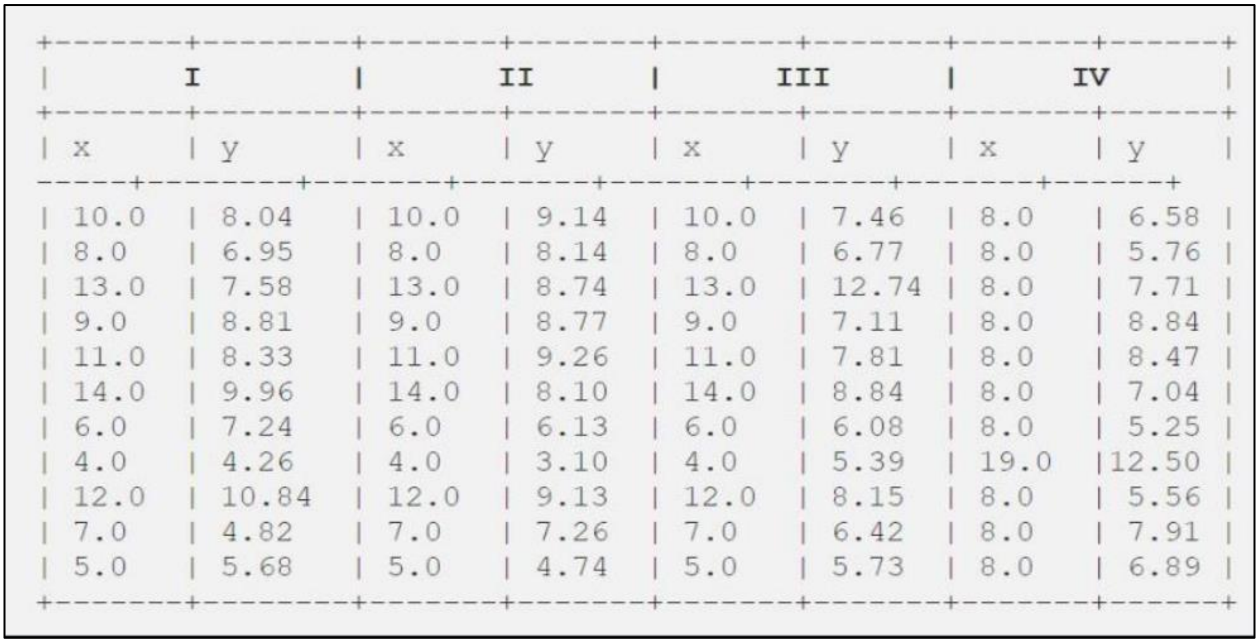
Q2. Explain the Anscombe's quartet in detail.

Solution:

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.

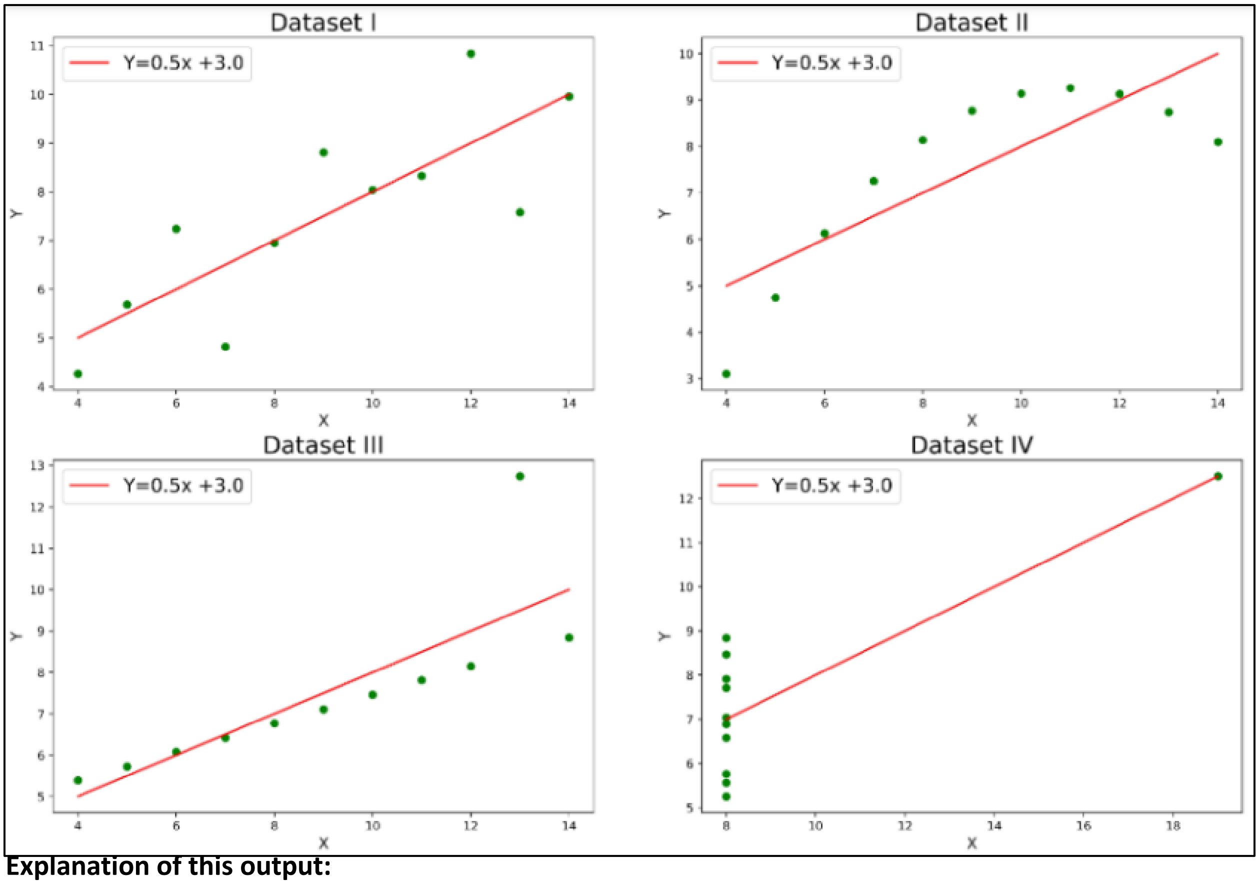
The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

Anscombe's quartet 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.



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

If we plot the scatter plot and linear regression line for each dataset



Explanation of this output:

1. In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
2. In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
3. In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated to be far away from that line.
4. Finally, the fourth one(bottom right) shows an example of when one high-leverage point is enough to produce a high correlation coefficient.

Q3. What is Pearson's R?

Solution:

Correlation coefficients are used to measure how strong a relationship is between two variables. There are different types of formulas to get a correlation coefficient, one of the most popular is Pearson's correlation (also known as Pearson's R) which is commonly used for linear regression. Pearson's correlation coefficient is denoted with the symbol "R". The correlation coefficient formula returns a value between 1 and -1.

Here,

-1 indicates a strong negative relationship
1 indicates strong positive relationships
And a result of zero indicates no relationship at all

Pearson’s correlation coefficient formula is the most commonly used and the most popular formula to get the correlation coefficient. It is denoted with the capital “R”. The formula for Pearson’s correlation coefficient is shown below,

$$R= n(\sum xy) - (\sum x)(\sum y) / \sqrt{[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2}$$

The interpretation of Pearson’s correlation coefficient is as follows:-

1. A correlation coefficient of 1 means there is a positive increase of a fixed proportion of others, for every positive increase in one variable. Like, the size of the shoe goes up in perfect correlation with foot length.

2. If the correlation coefficient is 0, it indicates that there is no relationship between the variables.

3. A correlation coefficient of -1 means there is a negative decrease of a fixed proportion, for every positive increase in one variable. Like, the amount of water in a tank will decrease in perfect correlation with the flow of a water tap.

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

Solution:

Scaling is the process of transforming numerical variables in a dataset to a specific range or distribution. It is performed to bring variables onto a similar scale, which can help in the analysis and modeling of the data. Scaling is particularly useful when variables have different units of measurement or varying ranges.

The main reasons for scaling are:

1. **Comparability:** Scaling allows variables with different scales and units to be compared directly. It ensures that no single variable dominates the analysis or model simply because of its larger magnitude.

2. **Model Performance:** Many machine learning algorithms and statistical techniques are sensitive to the scale of variables. Scaling can improve the performance and stability of models by preventing large-scale variables from dominating the optimization process.

3. **Interpretability:** Scaling can make the coefficients or weights in a model more interpretable. When variables are on the same scale, their coefficients represent the relative importance and impact on the outcome more accurately.

There are two common approaches to scaling: normalized scaling and standardized scaling.

Normalized Scaling (Min-Max Scaling):

1. Normalized scaling transforms variables to a specified range, typically between 0 and 1.
2. It involves subtracting the minimum value of the variable and dividing by the range (maximum value minus minimum value).
3. Normalized scaling preserves the shape of the original distribution and ensures all variables have the same minimum and maximum values.
4. The formula for normalized scaling is: X_scaled = (X - X_min) / (X_max - X_min) Standardized

Scaling (Z-score Scaling):

1. Standardized scaling transforms variables to have a mean of 0 and a standard deviation of 1.
2. It involves subtracting the mean of the variable and dividing by the standard deviation.
3. Standardized scaling centers the data around zero, making the mean of the variable 0 and the standard deviation 1.
4. The formula for standardized scaling is: X_scaled = (X - X_mean) / X_std

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

Solution:

In some cases, the VIF can be calculated as infinite (∞). This occurs when one or more independent variables in the regression model are perfectly linearly dependent on a combination of other independent variables. Perfect multicollinearity means that one variable can be expressed as an exact linear combination of other variables in the model, leading to an infinite VIF.

Here are a few scenarios that can result in infinite VIF values:

1. Exact Linear Relationships: When one or more independent variables in the model are identical or perfectly predictable based on other variables, it leads to perfect multicollinearity. This can happen, for example, when two variables are calculated using the same formula or when there is a duplicate variable in the dataset.

2. Singular Matrix: In some cases, the matrix used to calculate the VIF can become singular, which means it is not invertible. This typically occurs when there are too many independent variables compared to the number of observations in the dataset.

3. Near Perfect Collinearity: While not resulting in infinite VIF, very high VIF values close to infinity (e.g., extremely large values) can occur when there is a very high degree of collinearity among the independent variables. This indicates a strong linear relationship among the variables, potentially causing numerical instability or inflated standard errors in the regression model.

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

Solution:

A Quantile-Quantile plot or Q-Q plot is a scatter plot created by plotting 2 different quintiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against.

For example, if you are testing if the distribution of age of employees in your team is normally distributed, you are comparing the quantile of your team members' age vs the quantile from a normally distributed curve. If two quantiles are sampled from the same distribution, they should roughly fall in a straight line.
Since this is a visual tool for comparison, results can also be quite subjective nonetheless useful in the understanding underlying distribution of a variable(s)

The use and importance of a Q-Q plot in linear regression are as follows:

1. **Checking Normality Assumption:** In linear regression, one of the key assumptions is that the residuals (the differences between the observed and predicted values) should follow a normal distribution. A Q-Q plot can help assess the normality assumption by visually examining whether the residuals align with the straight line representing the normal distribution. If the points deviate significantly from the line, it suggests departures from normality.

2. **Detecting Skewness and Outliers:** A Q-Q plot can reveal deviations from normality due to skewness or the presence of outliers. Skewed data may cause the points in the plot to deviate from the straight line, indicating a departure from normal distribution. Outliers can be identified as data points that significantly deviate from the expected quantiles. These departures can guide further investigation and potential data transformations or outlier handling.

3. **Assessing Model Fit:** A Q-Q plot can help evaluate the overall fit of the linear regression model. If the residuals closely follow the expected quantiles, it suggests that the model captures the underlying relationship between the variables well. Conversely, if the residuals exhibit a systematic departure from the expected quantiles, it indicates potential model inadequacies or misspecifications.

4. **Comparing Alternative Distributions:** Besides assessing normality, a Q-Q plot can also be used to compare the observed data against other theoretical distributions. This can be helpful in determining the best-fitting distribution for the data or evaluating the appropriateness of assumptions beyond normality.