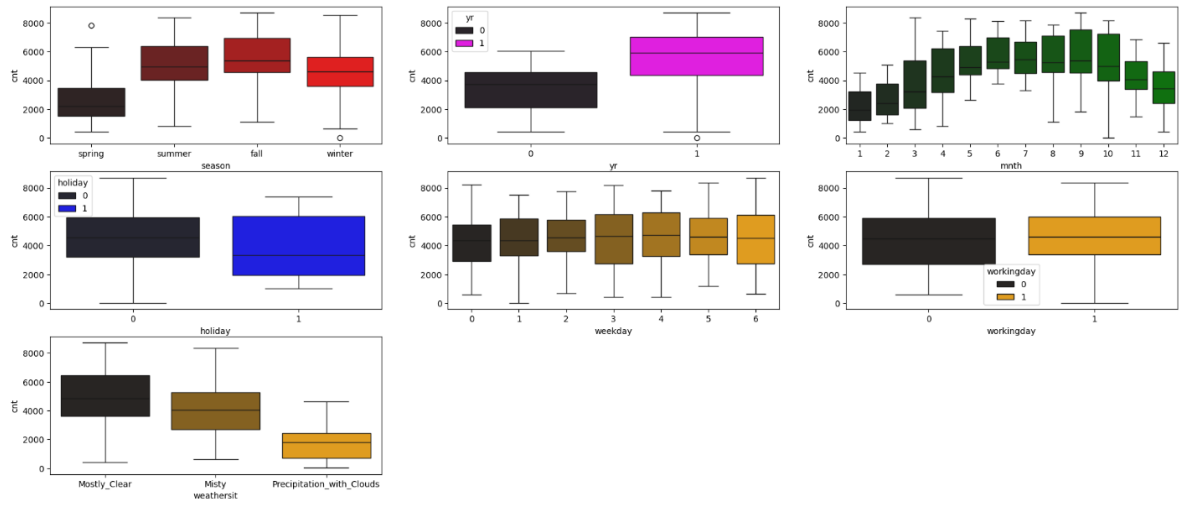
Assignment-based Subjective Questions

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

Visual analysis:



Insight:

a. season: there is higher demand of rental bikes in summer and fall compared to winter and spring
b. yr: there is an uptrend in demand from 2018 to 2019 which is generally a good macro trend, need to see further data post covid to generalize the trend
c. mnth: the general trend is higher demand in 2nd and 3rd quarter compared to 1st and 4th quarter
d. holiday: generally, the demand for rental bikes is higher for not a holiday
e. weathersit: demand for rental bikes is higher when it 'Mostly_Clear'

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

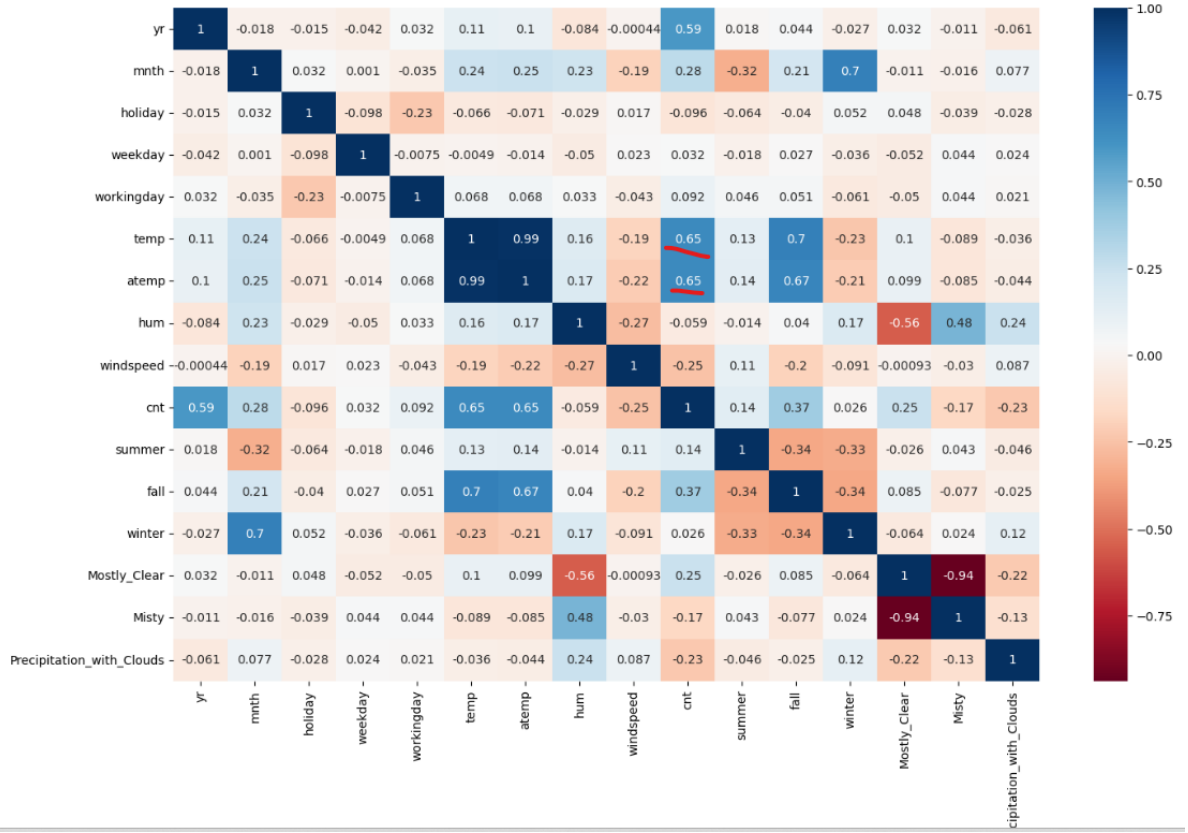
If there are 5 unique values for a categorical variables they can be represented by (5-1=4) dummy columns . When we use drop_first=True it drops the 1st column and we efficiently retain the information with minimum columns which helps in computation while modelling.

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

   Numerical variables

   a. temp
   b. atemp

N.B: Correlation value of cnt vs temp = 0.65 , cnt vs atemp = 0.65 , other variables with high correlation are



4. How did you validate the assumptions of Linear Regression after building the model on the training set?
   a. Used pairplot to validate if the relationship between some independent variables with target variable is linear
   b. Iterated the base model till the VIF score for all the remaining independent variables are < 5 to handle multicollinearity
   c. Calculated residuals by subtracting y_train_pred from y_train and visualized the residuals through a distribution plot. Qualitatively speaking the distribution plot should be centred around zero and approximately normal

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

The dummy variables created out of season are most important:

a. fall
b. summer
c. winter

The dummy variable 'Precipitation_with_Clouds' from weathersit reduces shared bike demand by 0.2947

OLS Regression Results

Dep. Variable:	cnt	**R-squared:**	0.763
Model:	OLS	**Adj. R-squared:**	0.758
Method:	Least Squares	**F-statistic:**	160.5
Date:	Wed, 28 Aug 2024	**Prob (F-statistic):**	7.12e-149
Time:	12:41:04	**Log-Likelihood:**	405.80
No. Observations:	510	**AIC:**	-789.6
Df Residuals:	499	**BIC:**	-743.0
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2438	0.019	12.559	0.000	0.206	0.282
yr	0.2478	0.010	25.222	0.000	0.229	0.267
holiday	-0.0682	0.032	-2.118	0.035	-0.131	-0.005
weekday	0.0084	0.002	3.426	0.001	0.004	0.013
workingday	0.0208	0.011	1.926	0.055	-0.000	0.042
windspeed	-0.1757	0.030	-5.844	0.000	-0.235	-0.117
summer	0.2563	0.014	18.256	0.000	0.229	0.284
fall	0.3130	0.014	22.011	0.000	0.285	0.341
winter	0.2280	0.014	15.967	0.000	0.200	0.256
Misty	-0.0882	0.010	-8.445	0.000	-0.109	-0.068
Precipitation_with_Clouds	-0.2947	0.030	-9.944	0.000	-0.353	-0.236

Omnibus:	28.748	**Durbin-Watson:**	2.011
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	59.210
Skew:	-0.324	**Prob(JB):**	1.39e-13
Kurtosis:	4.539	**Cond. No.**	26.8

General Subjective Questions

1. Explain the linear regression algorithm in detail.

a. Linear regression models can be classified into two types depending upon the number of independent variables:

- **Simple linear regression:** When the number of independent variables is 1
- **Multiple linear regression:** When the number of independent variables is more than 1

b. The equation of the best fit regression line $Y = \beta_0 + \beta_1 X$ can be found by minimising the cost function (RSS in this case, using the Ordinary Least Squares method) which is done using the following two methods:

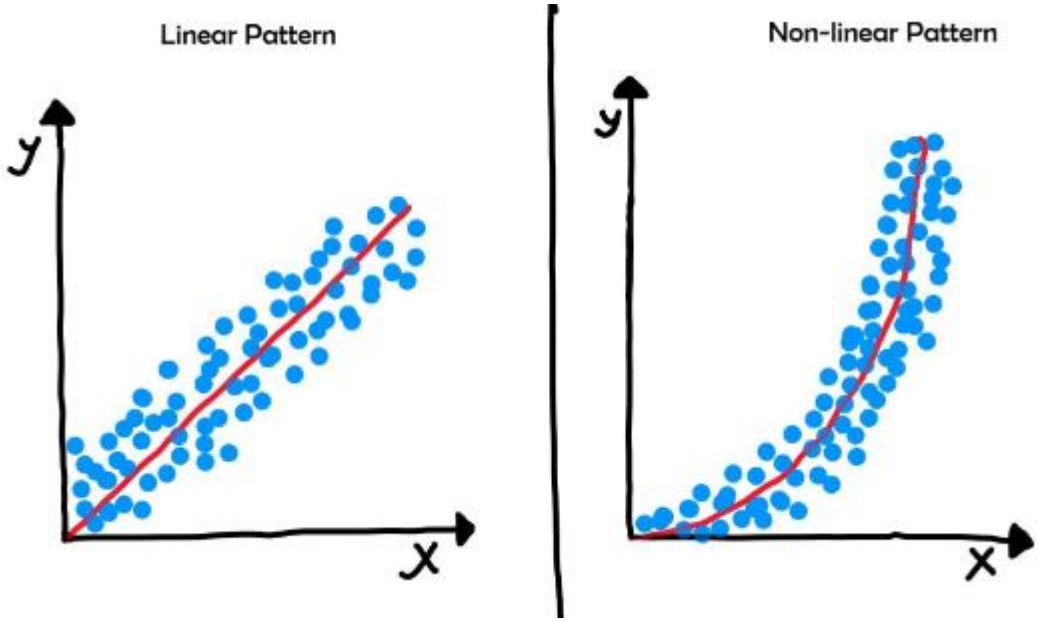
- **Differentiation**
- **Gradient descent method**

c. The strength of a linear regression model is mainly explained by R^2, where $R^2 = 1 - (RSS / TSS)$

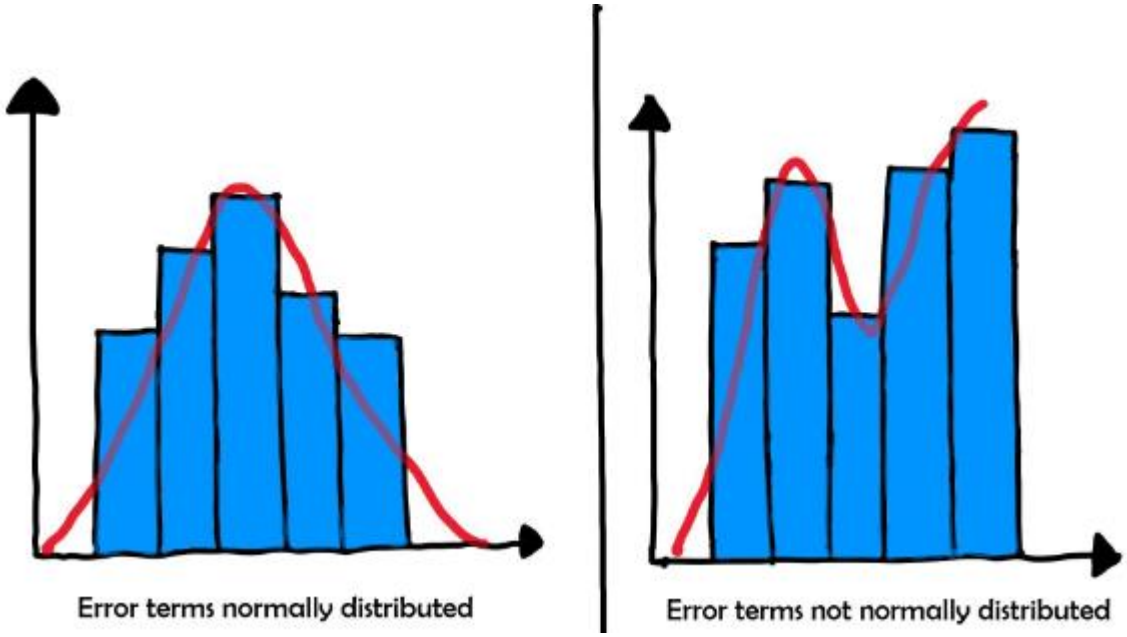
- **RSS:** Residual Sum of Squares
- **TSS:** Total Sum of Squares

d. Assumptions of linear regression

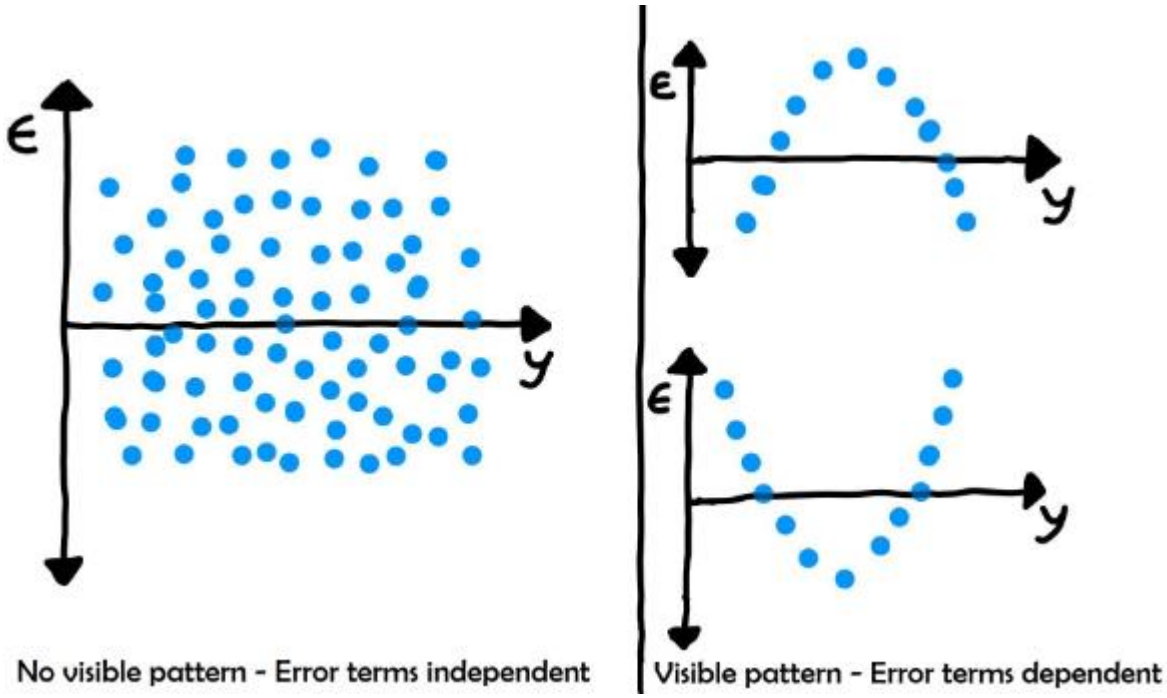
- There is a *linear relationship* between X and Y:



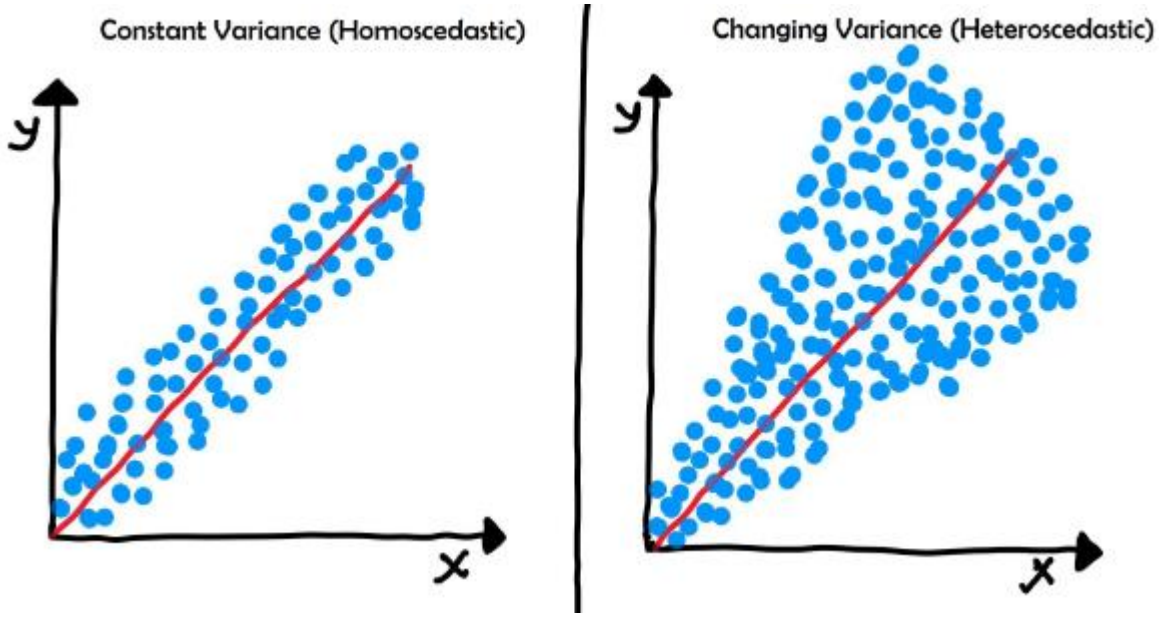
- Error terms are normally distributed with mean zero(not X, Y)



- Error terms are *independent* of each other



- Error terms have *constant variance* (homoscedasticity)



2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a group of datasets (x, y) that have the same mean, standard deviation, and regression line, but which are qualitatively different.

It is often used to illustrate the importance of looking at a set of data graphically and not only relying on basic statistic properties.

x = [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]

y1 = [8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68]

y2 = [9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74]

y3 = [7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73]

x4 = [8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8]

y4 = [6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89]

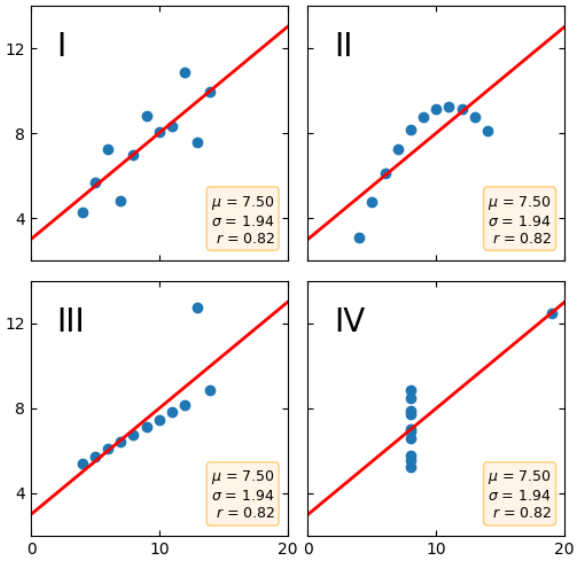
datasets

'I': (x, y1),

'II': (x, y2),

'III': (x, y3),

'IV': (x4, y4)



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.

Pearson correlation coefficient (*r*)	**Correlation type**	**Interpretation**	**Example**
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the **same direction**.	Baby length & weight: The longer the baby, the heavier their weight.
0	No correlation	There is **no relationship** between the variables.	Car price & width of windshield wipers: The price of a car is not related to the width of its windshield wipers.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the **opposite direction**.	Elevation & air pressure: The higher the elevation, the lower the air pressure.

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

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

In machine learning, scaling is a data preparation technique that involves transforming the values of a dataset's features to a specific range. This process brings data points that are far apart closer together, which can help algorithms run faster and more effectively. It also helps models learn and understand problems better.

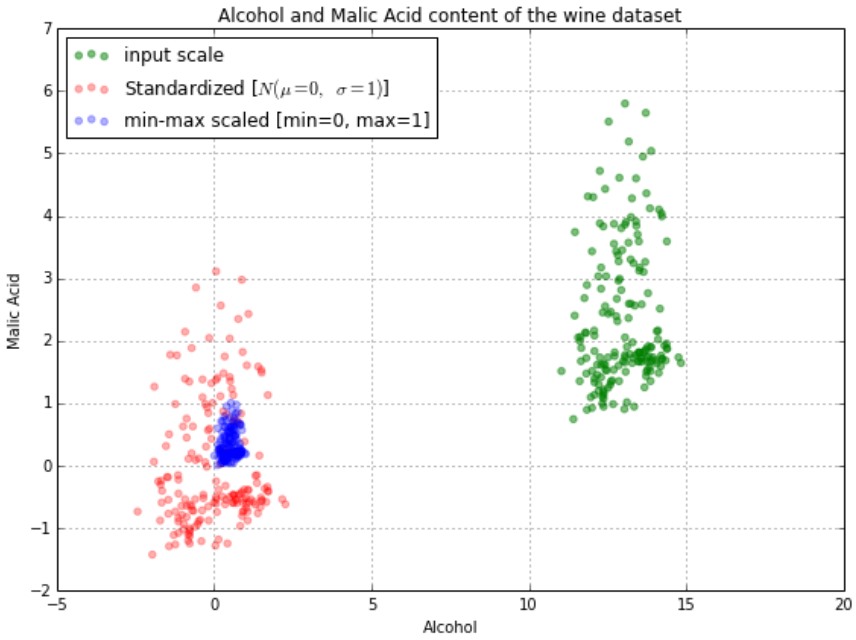
Scaling is a crucial step in preprocessing data before creating a machine learning model. It ensures that numerical features have a similar scale, which can help many algorithms perform better or converge faster. Scaling also prevents certain features from dominating due to their scale, which can help improve algorithm performance.

Normalized scaling (Min – Max scaling) compresses the data between 0 and 1.

Standardized scaling compresses with mean of zero and standard deviation of 1 .

Standardization ensures compression alongside the spread of original distribution

Visual representation of normalization and standardization on the same data below



Src: https://sebastianraschka.com/Articles/2014_about_feature_scaling.html

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

A VIF (Variance Inflation Factor) of infinity typically indicates a perfect multicollinearity problem. This means that one of your independent variables can be perfectly predicted by a linear combination of the others.

Here are some common reasons for infinite VIF:

- **Exact Linear Dependence:** Two or more independent variables are exactly linearly related. For example, if you have two variables "Age" and "Years of Experience" for employees, and these variables are always perfectly correlated (e.g., "Age" is always equal to "Years of Experience" + a constant), you'll get infinite VIF.
- **Near-Perfect Linear Dependence:** The variables might not be perfectly correlated, but they're very highly correlated. This can still lead to extremely high VIF values, making it difficult to interpret the model's coefficients.

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

The Quantile-Quantile plot is used for the following purpose:

- **Assessing Distributional Assumptions**: Q-Q plots are frequently used to visually inspect whether a dataset follows a specific probability distribution, such as the normal distribution. By comparing the quantiles of the observed data to the quantiles of the assumed distribution, deviations from the assumed distribution can be detected. This is crucial in many statistical analyses, where the validity of distributional assumptions impacts the accuracy of statistical inferences.
- **Detecting Outliers**: Outliers are data points that deviate significantly from the rest of the dataset. Q-Q plots can help identify outliers by revealing data points that fall far from the expected pattern of the distribution. Outliers may appear as points that deviate from the expected straight line in the plot.
- **Comparing Distributions**: Q-Q plots can be used to compare two datasets to see if they come from the same distribution. This is achieved by plotting the quantiles of one dataset against the quantiles of another dataset. If the points fall approximately along a straight line, it suggests that the two datasets are drawn from the same distribution.
- **Assessing Normality**: Q-Q plots are particularly useful for assessing the normality of a dataset. If the data points in the plot closely follow a straight line, it indicates that the dataset is approximately normally distributed. Deviations from the line suggest departures from normality, which may require further investigation or non-parametric statistical techniques.
- **Model Validation**: In fields like econometrics and machine learning, Q-Q plots are used to validate predictive models. By comparing the quantiles of observed responses with the quantiles predicted by a model, one can assess how well the model fits the data. Deviations from the expected pattern may indicate areas where the model needs improvement.
- **Quality Control**: Q-Q plots are employed in quality control processes to monitor the distribution of measured or observed values over time or across different batches. Departures from expected patterns in the plot may signal changes in the underlying processes, prompting further investigation.