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:

- Categorical Variable List =

['season','yr','mnth','holiday','weekday','workingday','weathersit','day_of_month']

Season:

Summer and Fall Season have observed high utilization of bikes, as compared to other seasons

Yr:

2019 has observed high utilization of bikes.

Mnth:

May, June and Jul have observed high utilization of bikes

Holiday:

Slightly high utilization is observed when there is no holiday

Weekday:

Seems to be same

Workingday:

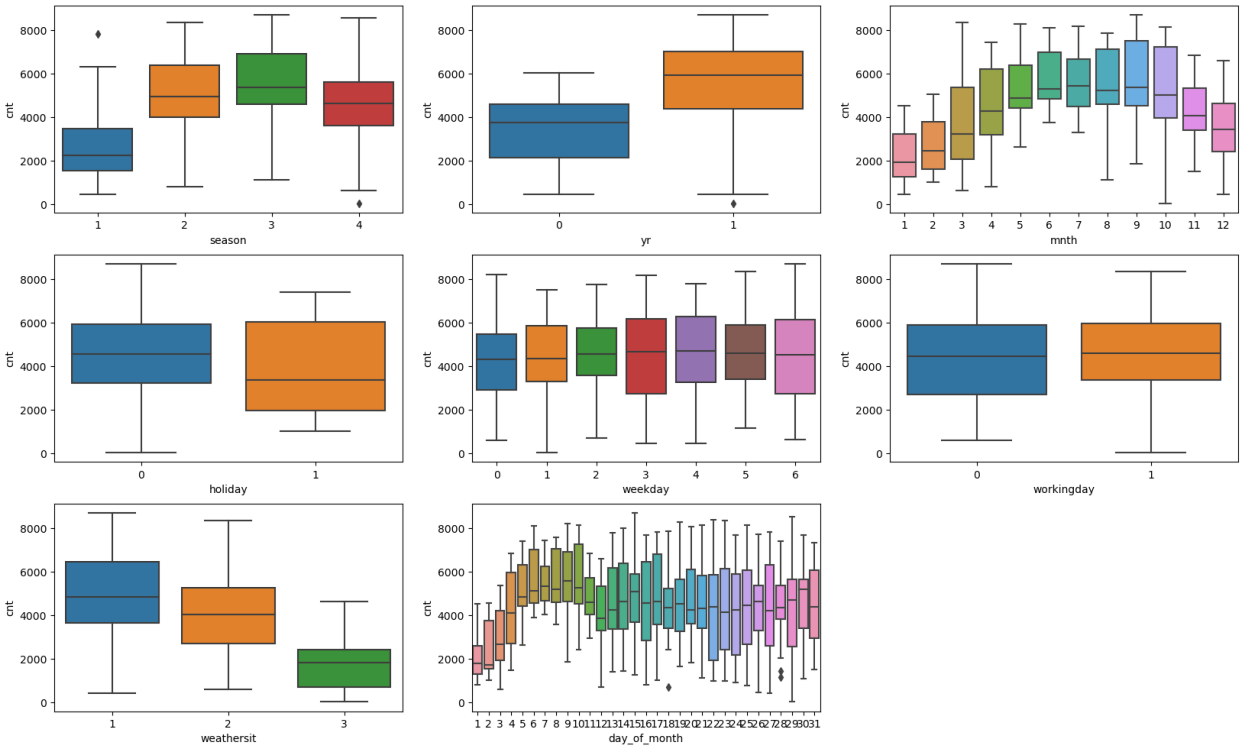
Seems to be same

Weathersit:

Clear, Few clouds, Partly cloudy, Partly cloudy has observed highest utilization of bikes , then comes Mist and cloudy

Day_of_Month: (Created Manually)

2nd Week of the month have observed comparatively high utilization of bikes, at majority of in between dates.



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

Answer:

- drop_first=True, drops the first column from the newly generated dummy variables dataframe.

- You can drop any column it is not mandatory to drop first one. The get_dummies function supports **drop_first=True,** hence we have used this.

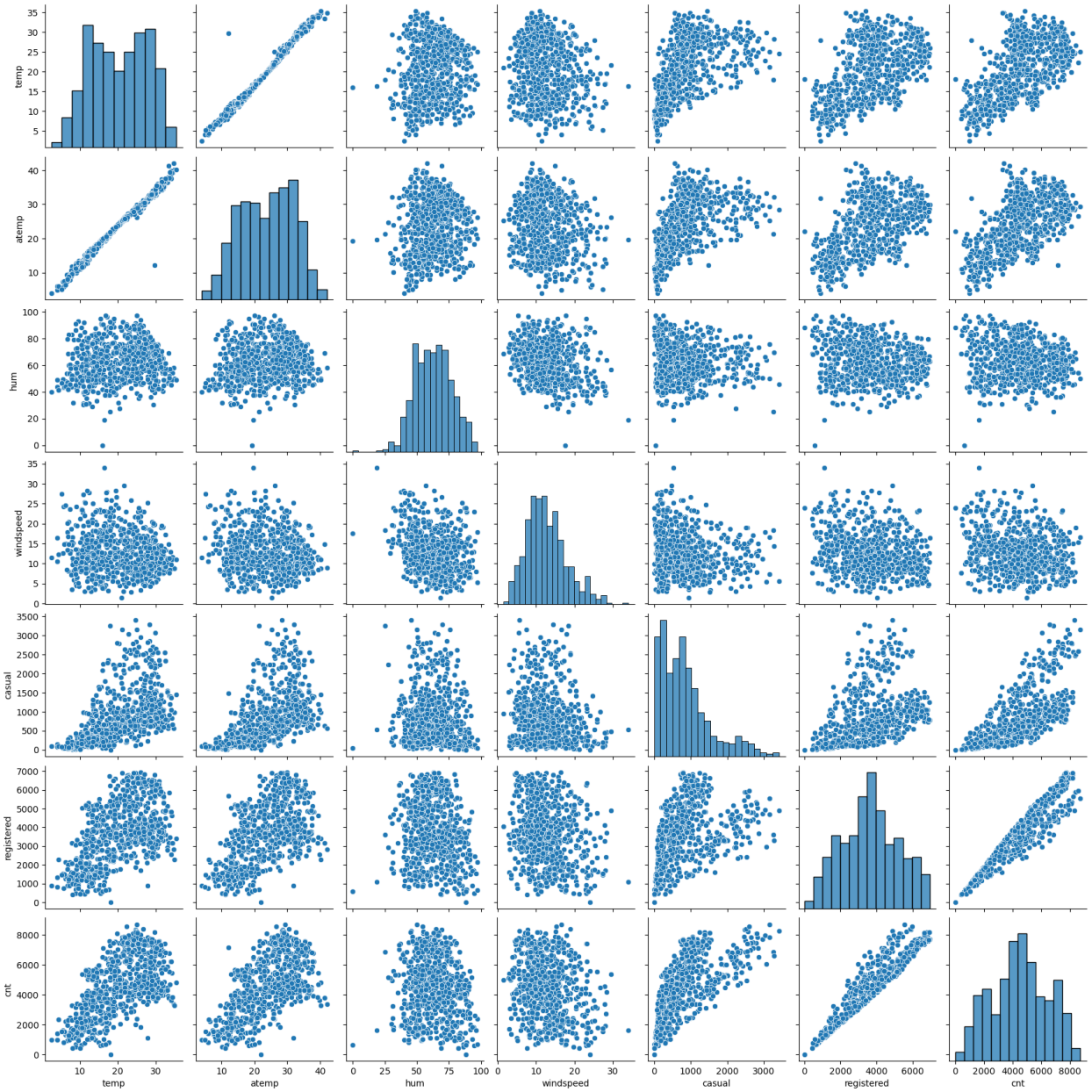
- This is to address the multicollinearity issues which may arise due to correlation between the newly introduced variables.

- Let's, say we have a Categorical Column, which holds three unique data – 0, 1 and 2
    - Post creating the dummy variables, if we hold all the columns then we will land in a situation where the information about one category can be perfectly predicted from the other.
    - By removing one variable using drop_first=True, we are eliminating the perfect correlation between the dummy variables.
    - In summary, using *drop_first=True* helps improve the performance and interpretability of the linear regression model when dealing with categorical variables by avoiding multicollinearity issues associated with the dummy variable trap.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer:

Registered Column



4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer:

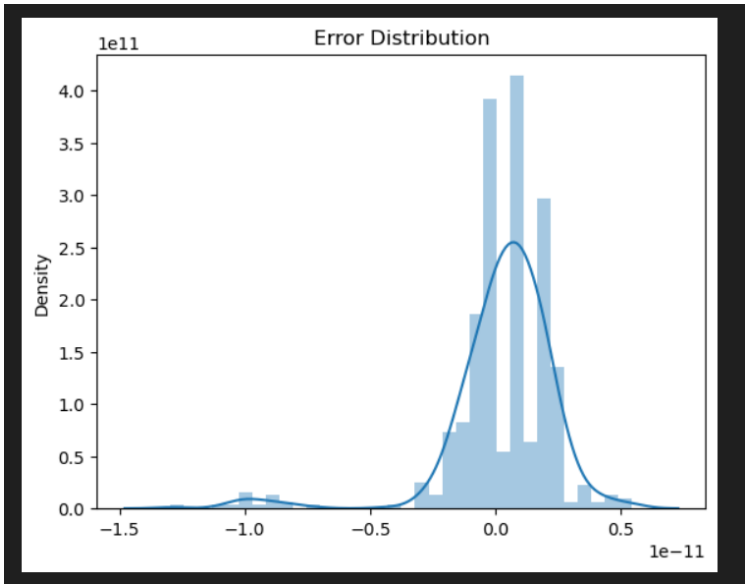
- Assumptions:

- Linear relationship:
    - Analysis of pair-plot suggest multiple columns do vary linearly with target variable, i.e., their plot shows a straight line. Variance in target variable can be explained by predictor variables.

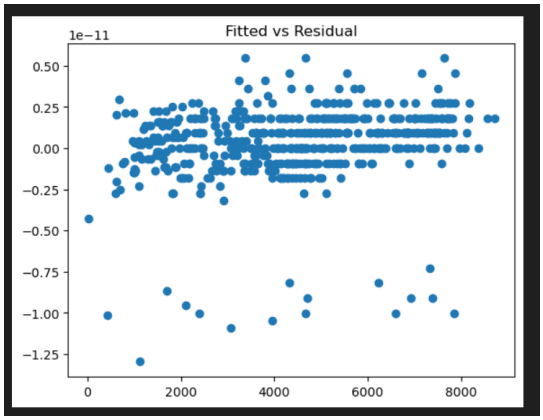
- No Multicollinearity:
    - We can confirm the independence of predictor variables by checking the statistics of VIF (Variance Inflation Factor).
    - The predicated model has predictor variables with VIF < 5.

	Feature	VIF
0	const	1.000000
1	atemp	3.158784
2	hum	1.142441
3	casual	1.561806
4	registered	1.557444
5	Fall	1.913074
6	Oct	1.109570
7	18	1.005799
8	27	1.001955

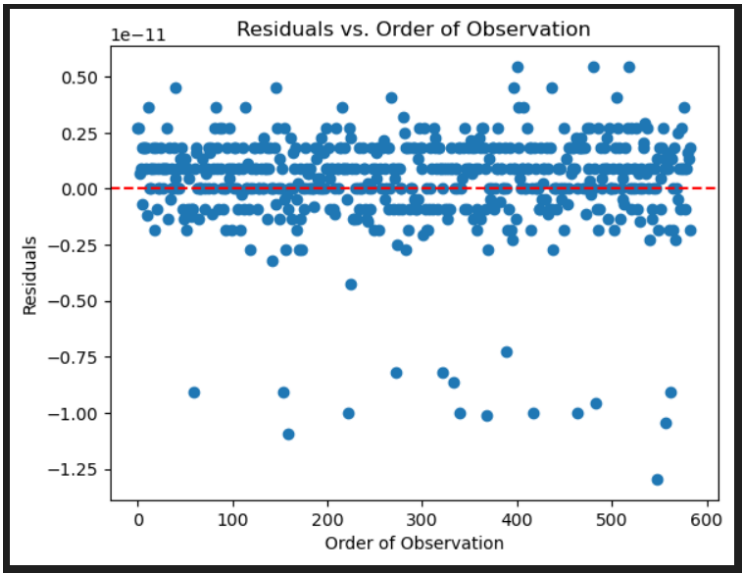
- Normality:
    - The errors follow a normal distribution. We can validate this in the Residual Analysis.



- Homoscedasticity:
    - Homoscedasticity refers to the assumption in regression analysis that the variance of the residuals (the differences between observed and predicted values) is constant across all levels of the independent variables. In other words, it implies that the spread or dispersion of the residuals remains the same regardless of the values of the predictors.



- Independence of Residuals:
    - It implies that the residuals are not correlated with each other.
    - We can verify this by plotting a scatter plot between Residuals and there order of observation



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer:

- Registered, Casual and 27th Day of Month. They have the highest coefficient.

OLS Regression Results

Dep. Variable:	cnt	R-squared:	1.000
Model:	OLS	Adj. R-squared:	1.000
Method:	Least Squares	F-statistic:	5.633e+31
Date:	Fri, 17 Nov 2023	Prob (F-statistic):	0.00
Time:	10:18:54	Log-Likelihood:	14845.
No. Observations:	584	AIC:	-2.967e+04
Df Residuals:	575	BIC:	-2.963e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4505.2671	9.22e-14	4.89e+16	0.000	4505.267	4505.267
atemp	1.08e-12	1.64e-13	6.586	0.000	7.58e-13	1.4e-12
hum	6.537e-13	9.86e-14	6.629	0.000	4.6e-13	8.47e-13
casual	674.4972	1.15e-13	5.85e+15	0.000	674.497	674.497
registered	1595.3723	1.15e-13	1.39e+16	0.000	1595.372	1595.372
Fall	-4.547e-13	1.28e-13	-3.563	0.000	-7.05e-13	-2.04e-13
Oct	3.268e-13	9.72e-14	3.363	0.001	1.36e-13	5.18e-13
18	-6.253e-13	9.25e-14	-6.757	0.000	-8.07e-13	-4.44e-13
27	1.698e-12	9.24e-14	18.388	0.000	1.52e-12	1.88e-12

Omnibus:	345.488	Durbin-Watson:	1.960
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3269.612
Skew:	-2.496	Prob(JB):	0.00
Kurtosis:	13.462	Cond. No.	3.40

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Answer:

- Linear regression is a supervised machine learning algorithm used for predicting a continuous outcome variable (also called the dependent variable) based on one or more predictor variables (independent variables). The relationship between the variables is assumed to be linear, meaning that changes in the predictor variables are associated with a constant change in the outcome variable.
- The basic form of a linear regression model for a single predictor variable can be expressed as:

  $Y=b0+b1 \cdot X+\varepsilon$

  Y is the dependent variable (the variable you are trying to predict).

  X is the independent variable (the variable used for prediction).

  b0 is the y-intercept (the value of YY when XX is 0).

  b1 is the slope (the change in YY for a one-unit change in XX).

  ε is the error term, representing the difference between the observed and predicted values.

- In the case of multiple predictor variables, the formula is extended to:

  $Y=b0+b1 \cdot X1+b2 \cdot X2+\ldots+bn \cdot Xn+\varepsilon$

Here:

b0 is the y-intercept.

b1,b2,…,bn are the coefficients for the respective predictor variables X1,X2,…,Xn

ε is the error term.

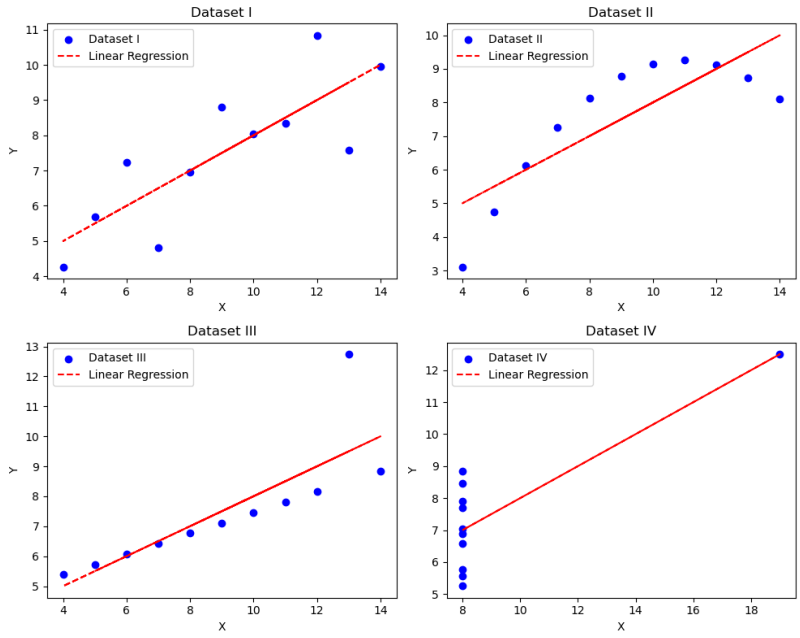
- The goal of linear regression is to find the values of b0,b1,…,bn that minimize the sum of squared differences between the observed and predicted values of the dependent variable.
- The model is trained using a dataset where both the predictor variables and the corresponding outcome variable are known. The training process involves adjusting the coefficients to minimize the error between the predicted and actual values. This is typically done using optimization techniques such as gradient descent.
- Once trained, the linear regression model can be used to make predictions on new, unseen data.
- Linear regression assumes that there is a linear relationship between the predictor variables and the dependent variable, and it also assumes that the errors (ε) are normally distributed and have constant variance (homoscedasticity). Additionally, linear regression is sensitive to outliers, and the presence of multicollinearity (high correlation between predictor variables) can affect the model's performance.

2. Explain the Anscombe's quartet in detail. (3 marks)

Answer:

- Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics but differ significantly when graphed.
- The four datasets in Anscombe's quartet have the following characteristics:
  - Dataset I: Linear Relationship
  - Dataset II: Non-linear Relationship
  - Dataset III: Outlier
  - Dataset IV: Influential Point
- Example:

```
# Anscombe's quartet data
data = {
    'I': {'x': [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5], 'y': [8.04, 6.95,
7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68]},
    'II': {'x': [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5], 'y': [9.14, 8.14,
8.74, 8.77, 9.26, 8.1, 6.13, 3.1, 9.13, 7.26, 4.74]},
    'III': {'x': [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5], 'y': [7.46, 6.77,
12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73]},
    'IV': {'x': [8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8], 'y': [6.58, 5.76,
7.71, 8.84, 8.47, 7.04, 5.25, 12.5, 5.56, 7.91, 6.89]}
}
```



3. What is Pearson's R? (3 marks)

Answer:

- Pearson's correlation coefficient, often denoted as r, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It ranges from -1 to 1, where:
  - *r=1: Perfect positive correlation (as one variable increases, the other increases proportionally).*
  - *r=−1: Perfect negative correlation (as one variable increases, the other decreases proportionally).*
  - *r=0: No linear correlation.*

- The formula for Pearson's correlation coefficient between two variables X and Y with n data points is given by:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

Where:

Xi and Yi are the individual data points.

$\bar{X}$ and $\bar{Y}$ are the means of X and Y respectively.

- Pearson's correlation coefficient measures the linear relationship between two variables. However, it assumes that the relationship is linear and may not capture non-linear relationships. Additionally, it is sensitive to outliers, meaning that a few extreme data points can disproportionately influence the correlation.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer:

Meaning & Why->

- Scaling is a preprocessing step in data analysis and machine learning that involves transforming the values of variables to a specific range. The primary goals of scaling are:
    - Normalization of Data:
        - Scaling ensures that all variables have the same scale or range. This is important for algorithms that are sensitive to the scale of variables, such as distance-based algorithms (e.g., k-nearest neighbors) and optimization algorithms (e.g., gradient descent).
        - Normalization helps prevent features with larger scales from dominating those with smaller scales during the learning process.
    - Improving Convergence:
        - Many optimization algorithms, particularly those used in machine learning models like neural networks and support vector machines, converge faster when the input features are on a similar scale. Scaling helps these algorithms reach convergence more quickly.
    - Enhancing Interpretability:
        - Scaling makes it easier to interpret the coefficients or weights of a model. If the variables are on different scales, the magnitude of the coefficients may not accurately represent their importance in the model.

- Two common types of scaling are normalized scaling and standardized scaling:
    - Normalized Scaling (Min-Max Scaling):
        - Normalization, or Min-Max scaling, transforms the data to a specific range, typically between 0 and 1.
        - The formula for Min-Max scaling is given by:

$$X_{\text{norm}} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

            - *X is the original variable,*
            - *Xnorm is the normalized variable,*
            - *min(X) is the minimum value of X,*
            - *max(X) is the maximum value of X.*
    - Standardized Scaling (Z-score Standardization):
        - Standardization transforms the data to have a mean of 0 and a standard deviation of 1.
        - The formula for standardization is given by:

$$X_{\text{std}} = \frac{X - \text{mean}(X)}{\text{std}(X)}$$

            - *X is the original variable,*
            - *Xstd is the standardized variable,*

- *mean(X) is the mean of X, and std(X) is the standard deviation of X.*

Key Differences->

- Normalized scaling brings the values of a variable within a specific range (e.g., 0 to 1), while standardized scaling transforms the values to have a mean of 0 and a standard deviation of 1.
- Normalized scaling is sensitive to outliers, as the minimum and maximum values are directly influenced by extreme values. Standardized scaling is less affected by outliers because it relies on the mean and standard deviation.
- Normalized scaling is often preferred when the distribution of the data is not normal or when the algorithm used assumes normalized data. Standardized scaling is commonly used when the data is approximately normally distributed.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer:

- The Variance Inflation Factor (VIF) is a measure used to assess the extent of multicollinearity in a multiple regression analysis. VIF quantifies how much the variance of an estimated regression coefficient increases if the predictors are correlated.
- The formula for the VIF of a predictor variable Xi is:

$$\mathrm{VIF}(X_i) = \frac{1}{1-R_i^2}$$

  - Ri**2 is the coefficient of determination of the regression of Xi on all other predictor variables.
- Now, VIF can be infinite if the denominator becomes zero and then can become if we have Ri**2 has the value 1, which means perfect correlation.
- Perfect multicollinearity occurs when one or more independent variables in a multiple regression model can be exactly predicted from the others. This leads to an exact linear relationship among the predictors, making it impossible for the regression algorithm to estimate the coefficients accurately.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer:

- A Quantile-Quantile (Q-Q) plot is a graphical tool used to assess whether a given sample follows a particular theoretical distribution, such as the normal distribution. In the context of linear regression, Q-Q plots are often used to check the assumption of normality of residuals.

- Theoretical Quantiles:
    - Theoretical quantiles are the expected values from a specified theoretical distribution (e.g., normal distribution) based on the sample size.
- Sample Quantiles:
    - Sample quantiles are the observed values from the data.
- Plotting:
    - For a Q-Q plot, the theoretical quantiles are plotted on the x-axis, and the sample quantiles are plotted on the y-axis.
- Diagonal Line:
    - If the sample quantiles closely follow a straight line, it suggests that the data follows the theoretical distribution.
- Departures from Linearity:
    - Departures from the straight line indicate deviations from the assumed distribution. For example, if the points deviate from the line in the tails, it suggests that the data has heavier or lighter tails than the theoretical distribution.
- In the context of linear regression, the Q-Q plot is particularly useful for checking the normality assumption of residuals. The residuals are the differences between the observed values and the values predicted by the regression model. The normality assumption is important because many statistical inference procedures and hypothesis tests rely on the assumption that the residuals are normally distributed.
- Here's how the Q-Q plot is used in linear regression:
    - Residuals Calculation:
        - Calculate the residuals by subtracting the predicted values from the observed values in your linear regression model.
    - Q-Q Plot:
        - Create a Q-Q plot using the residuals.
    - Assessment:
        - Examine how closely the points on the Q-Q plot follow the diagonal line.
        - If the points deviate from the line, especially in the tails, it suggests that the residuals may not be normally distributed.
    - Normality Assessment:
        - The Q-Q plot is a visual tool to assess whether the residuals exhibit a normal distribution. Deviations from normality might indicate issues that need attention, such as outliers or non-linear relationships that are not captured by the model.
    - Remedial Actions:
        - If the Q-Q plot indicates significant departures from normality, it might be necessary to investigate and address the underlying issues. This could involve transforming the response variable, identifying influential points, or considering alternative modeling approaches.

- In summary, the Q-Q plot is a valuable diagnostic tool in linear regression for checking the normality assumption of residuals, providing insights into the appropriateness of the model and potential areas for improvement.