Multiple Linear Regression Subjective Questions

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?

Ans:

The **Categorical Feature Variables** in the Case Study are –

'season' , 'workingday', 'weathersit', 'weekday', 'yr', 'mnth', 'holiday'

The effect of these features on the **dependent variable "cnt"** which denotes count of bike rentals-

- 'season' –
  - The data in the dataset provided is almost equally distributed amongst the 4 season category – Fall, Summer, Spring, Winter
  - Fall has the highest number of bike rentals while Spring Season has significant lowest bike rentals compared to other seasons.
  - Summer has the second highest bike rental numbers followed by Winter.
  - Fall and Summer are best season having more bike rentals and can be leveraged to improve the number of rentals even further.
- 'workingday' –
  - Working day represents if the day is weekday or weekend/holiday.
  - We can see that the rentals number is marginally higher for working day compared to weekend or holiday.
  - There is more variability in data when it is not a working day.
- 'weathersit' –
  - Most number of rentals can be seen when the weather is Clear and Few Clouds
  - There is a small number of rentals also on Light Rain/ Snow weather days indicating it is from the registered users.
  - There is no data for Heavy Rains/ Snow weather.
- "weekday" –
  - The median count of rentals on all the days of week is similar so no pattern observed.
  - When we observe the correlation with cnt variable , we can see that weekdays tend to have positive correlation indicating that rentals are mostly used for daily commute.
- "yr" –
  - 2 years data is available and the increase in the bikes has increased from 2018 to 2019.
- "holiday" –
  - In total Non Holidays have more bike rentals compared to holidays.
- "mnth" –
  - The Bike rentals is higher in the middle months of the year - From June to October the count of rentals is above 5000 per month.
  - Also from boxplot we can observe that the demand of rentals in these months are expanding since the difference between the median and 75th percent count has expanded.

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

Ans :

Inorder to avoid multicolliearity ,we use one hot encoding to encode the categorical variable values for the feature which cover the values and are mapped using 1 and 0 values. For the features with 2 values we do not need to create dummy data as they are already binary. This can be done by using pandas.get_dummies() which will return dummy-coded data.

drop_first = True parameter is used to drop the first dummy variable from the n discrete variables created for the n categorical levels of the feature, generating n-1 dummy variables for n levels.
This makes sense as the one value dropped can be explained by the other categories.

If this is not done, then we will observe multicollinearity since the generated dummy variables will themselves be co related to each other and lead to Dummy Variable Trap.

If we have a feature with 3 category and we generate 3 dummy variables- v1, v2 , v3 using one hot encoding - 100 , 010 , 001 then since these are categorical variables, v1 + v2 + v3 = 1

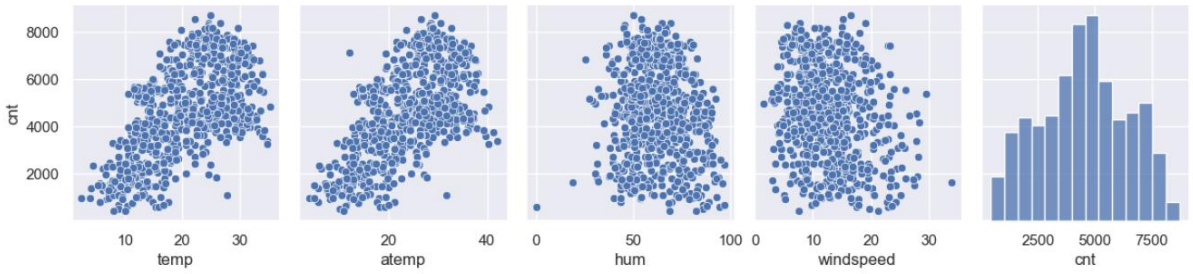
so v3 = 1 – (v1 + v2) when used in MLR equation it will lead to multicollinearity.

Conclusion - if there are n dummy variables, n-1 dummy
variables will be able to predict the value of the nth dummy variable, so one dummy variable should be dropped to avoid multicollinearity.

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

Ans :

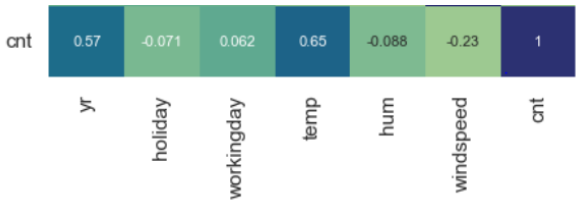
Below is the pair plot of all the numerical variables with the target variable cnt.



From the Above pair plot we can clearly see that **the temp has the highest correlation with cnt variable, and cnt increases with increase in temp**

atemp feature also shows similar treads but since adjusted temp (atemp) might be derived from temp variable we should ignore this variable drop from our dataset.

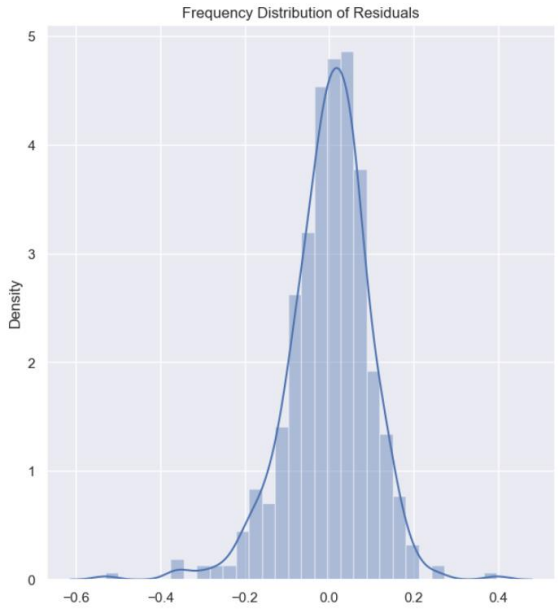
Also from correlation matrix we can see the **correlation of temp with cnt is 0.65**

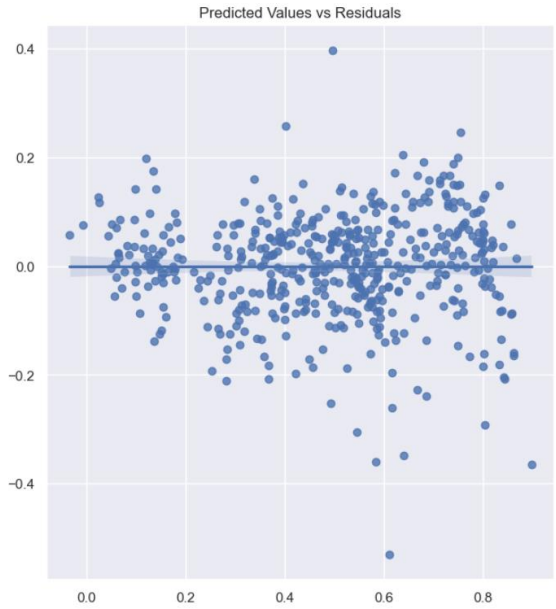


4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans:

1. **Residual Analysis – Residuals should be normally distributed**
    a. Histogram and distribution plot and Regplot helps to understand the normal distribution of error terms along with the mean of 0. The figure below clearly depicts the same.



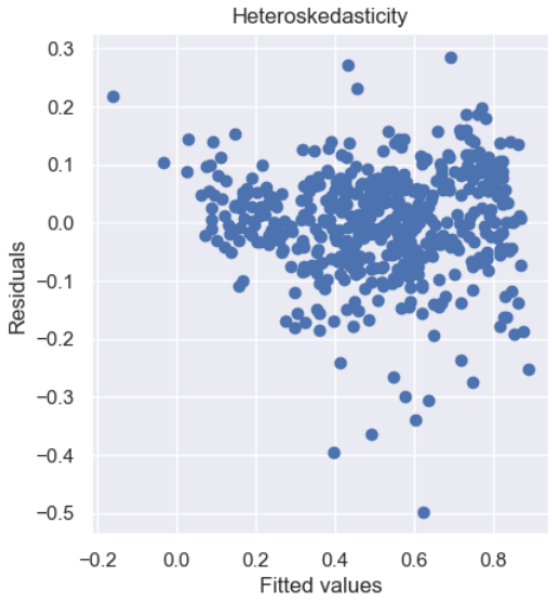


Residual errors follow a normal distribution with mean=0
Variance of Errors doesnot follow any trends
Residual errors are independent of each other since the Predicted values vs Residuals plot doesn't show any trend.

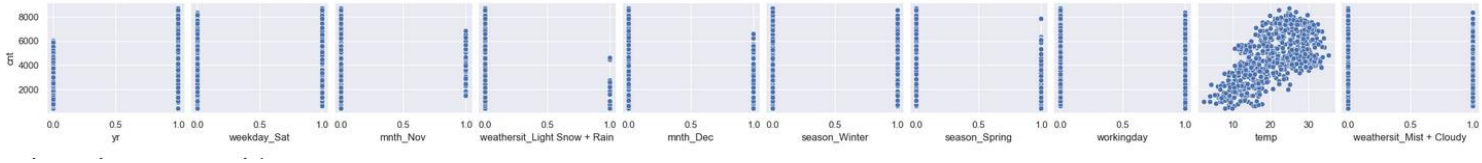
2. **No Multicollinearity** –
    a. As we can see **VIFs of all feature variables below 5**, so there is no multicollinearity.

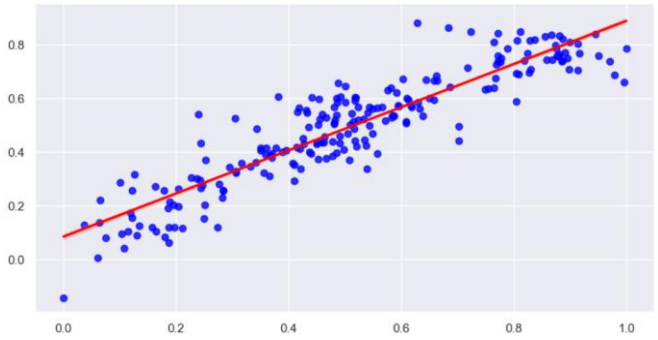
Features	VIF
temp	4.43
workingday	4.35
season_Winter	2.50
yr	2.08
mnth_Nov	1.75
weekday_Sat	1.63
weathersit_Mist + Cloudy	1.53
season_Spring	1.52
mnth_Dec	1.45
weathersit_Light Snow + Rain	1.07

3. **No Heteroskedasticity :**
    a. From the scatter plot, **we do not see a funnel like pattern and most of the points are centered around zero.** So we do not have any heteroskedasticity.

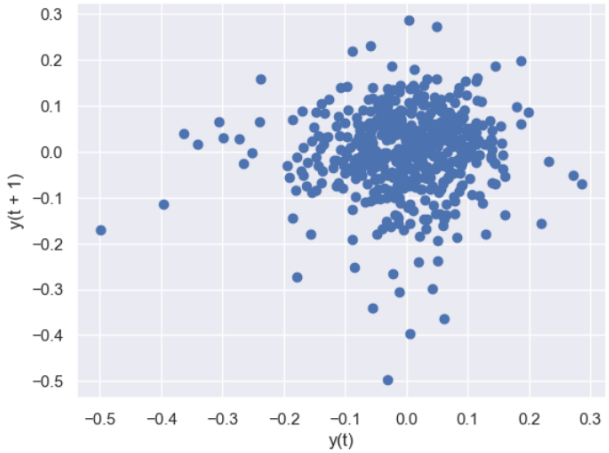


4. **Linear relationship between target and feature variables**
    a. Here we can see that for numerical feature cnt increases with increase in value





5. **Lagplot**
    a. Lagplot of residuals shows no trend. Hence the error terms have constant variance



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

Ans :

The Equation for best fitted line -

- cnt = **0.131488 + (**temp **× 0.453612) + (**yr **× 0.240374) - (**weekday_Sat **× 0.071721) - (**mnth_Nov **× 0.099894) - (**mnth_Dec **× 0.067850) + (**workingday **× 0.054807) + (**season_Winter **× 0.115089) - (**season_Spring **× 0.125942) - (**weathersit_Light Snow + Rain **× 0.280320) - (**weathersit_Mist + Cloudy **× 0.077951)**
- As per the above equation, the **top 3 predictor variables** that influences the bike booking are:
  - **Temperature** – Temperature is the most significant driver and influences the business positively
  - **Light Snow and Rain Weather – This has negative impact on the business so can introduce offers to boost the demand and number of bike rentals**
  - **Year** – The Growth in year on year seem organic and should be maintained using additional insights from other features

Other Features like Winter season and Working Days are positive contributors as well and can be used further to increase the business

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans:

- It is a statistical, linear, predictive algorithm that uses regression to establish a linear relationship between the dependent and the independent variable. It comes up with a line of best fit, and the value of Y (variable) falling on this line for different values of X (variable) is considered the predicted values.
- 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
    - $Y = \beta_0 + \beta_1 X$ is the line equation used for SLR.
  - Multiple linear regression: When the number of independent variables is more than 1
    - $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \in$ is the line equation for MLR.

  Note: $\beta_1, \beta_2, \ldots, \beta_p = Slope\ or\ coefficients\ .\ \beta_0\ is\ the\ y\ intercept$

- Also in case of MLR , We need to check the Model for Overfitting, Multicollinearity.
- For Categorical Variables, dummy variables should be used by using one hot encoding.
- Feature Scaling must be done using below techniques to prevent influence of variable values -
  - Standardisation
  - MinMaxScaling
  - Categorical variables should be converted to numeric before scaling

- While finding the best fit line we encounter that there are errors while mapping the actual values to the line. These errors are nothing but the residuals. To minimize the error squares OLS (Ordinary least square) is used.
  - $e_i = y_i - y_{pred}$ is provides the error for each of the data point.
  - OLS is used to minimize the total e^2 which is called as Residual sum of squares.
  - $\text{RSS} = = \sum_{i=1}^{n} (y_i - y_{pred})^2$
- Cost Function helps to find the best possible values for the coefficients to better predict the target variables.

- Minimising the cost function (RSS using the Ordinary Least Squares method) which is done using the following two methods:
    - Differentiation
    - Gradient descent method
- 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
    - Adjusted R^2 - The **adjusted R-squared** value increases only if the new term improves the model more than would be expected by chance.
    - AIC, BIC - Various types of criteria used for automatic feature selection

2. Explain the Anscombe's quartet in detail.

Ans:

Statistical methods like variance and standard deviation are usually considered good enough parameters to understand the variation of some data without actually looking at every data point. The statistics are great to for describing the general trends and aspects of the data.

Anscombe's Quartet is a classic statistical demonstration devised by the statistician Francis Anscombe in 1973. It consists of four distinct datasets, each containing eleven (x, y) data points. What makes Anscombe's Quartet remarkable is that despite having very similar statistical properties, the datasets exhibit vastly different relationships when graphed, showcasing the importance of visualizing data alongside numerical analysis.

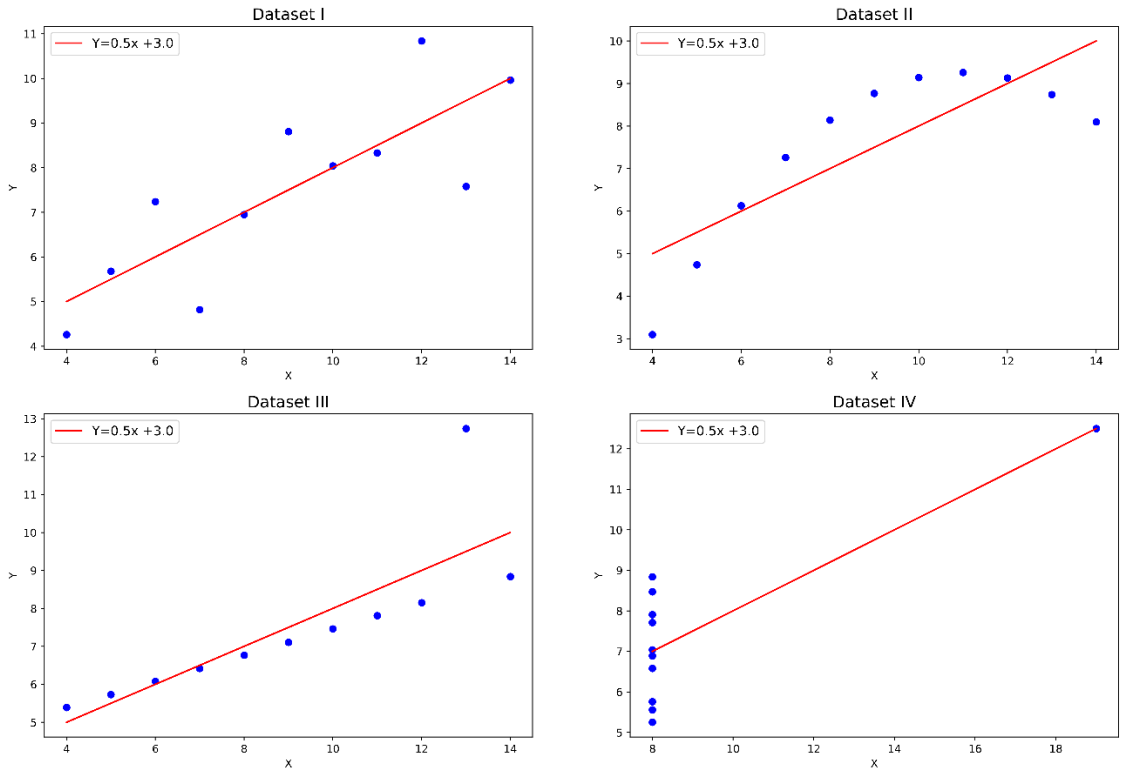
Lets understand this with an e.g. dataset –

The quartet consists of four different datasets, each containing 11 points, with two variables: x and y; such as x1 & y1, x2 & y2, x3 & y3, x4 & y4.

	x1	x2	x3	x4	y1	y2	y3	y4
0	10	10	10	8	8.04	9.14	7.46	6.58
1	8	8	8	8	6.95	8.14	6.77	5.76
2	13	13	13	8	7.58	8.74	12.74	7.71
3	9	9	9	8	8.81	8.77	7.11	8.84
4	11	11	11	8	8.33	9.26	7.81	8.47
5	14	14	14	8	9.96	8.10	8.84	7.04
6	6	6	6	8	7.24	6.13	6.08	5.25
7	4	4	4	19	4.26	3.10	5.39	12.50
8	12	12	12	8	10.84	9.13	8.15	5.56
9	7	7	7	8	4.82	7.26	6.42	7.91
10	5	5	5	8	5.68	4.74	5.73	6.89

	I	II	III	IV
Mean_x	9.000000	9.000000	9.000000	9.000000
Variance_x	11.000000	11.000000	11.000000	11.000000
Mean_y	7.500909	7.500909	7.500000	7.500909
Variance_y	4.127269	4.127629	4.122620	4.123249
Correlation	0.816421	0.816237	0.816287	0.816521
Linear Regression slope	0.500091	0.500000	0.499727	0.499909
Linear Regression intercept	3.000091	3.000909	3.002455	3.001727

Despite the variations in each dataset, they have the same summary statistics such as same mean, same standard deviations (SD), correlational coefficient, and linear regression line.



- **Data-set I** — the first dataset appears to be a simple linear relationship, where y increases as x increases.
- **Data-set II** — this follows a perfectly quadratic relationship, with a clear curve. This highlights the fact that data can exhibit nonlinear patterns, and relying solely on linear regression can lead to incorrect conclusions.
- **Data-set III** — the dataset, shows a linear trend, a single outlier affects the regression line, creating a misleading representation of the data.
- **Data-set IV** — the fourth dataset adds a new layer of complexity to the situation. There is one data point that is outlier from the others and entirely contradicts the pattern, which causes the linear regression line to shift in a significant way.

Anscombe's Quartet Significance:

- **Diverse Relationships**: Despite having identical means, variances, correlation coefficients, and linear regression lines, the datasets portray drastically different relationships when graphed. This illustrates the importance of visualizing data to understand its underlying structure fully.
- **Cautionary Tale**: Anscombe's Quartet serves as a cautionary tale in statistical analysis, reminding researchers not to rely solely on summary statistics. Even when statistical properties seem similar, the actual data may be fundamentally different.
- **Exploratory Data Analysis (EDA):** Anscombe's Quartet supports the necessity of exploratory data analysis (EDA) before drawing conclusions from statistical analyses. Visualization techniques can reveal nuances and patterns that summary statistics alone might miss.
- **Outliers need to be treated:** Based on the last 2 graphs we can see how outliers impact the overall model.

3. What is Pearson's R?

Ans:

Pearson's correlation coefficient, often denoted by r, is a measure of the linear relationship between two continuous variables. It quantifies the strength and direction of the association between the variables. Pearson's r ranges from -1 to 1, where:

- $r = 1$*: Perfect positive correlation. As one variable increases, the other variable increases proportionally.*
- $r = -1$*: Perfect negative correlation. As one variable increases, the other variable decreases proportionally.*
- $r = 0$*: No linear correlation. There is no systematic relationship between the variables.*

Pearson's correlation coefficient is calculated using the following formula:

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

Where:

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

Pearson's r measures the degree to which the points in a scatterplot cluster around a straight line. However, it is important to note that Pearson's correlation coefficient only measures linear relationships and may not capture non-linear associations between variables. Additionally, correlation does not imply causation, meaning that a high correlation between two variables does not necessarily mean that one variable causes the other to change.

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

Ans:

Scaling refers to the process of transforming the features of a dataset so that they fall within a similar numerical range. This transformation does not change the shape of the distribution of the data but merely rescales it. It is the data preparation step for regression model.

Scaling only affects the coefficients. The prediction and precision of prediction stays unaffected after scaling.

Scaling is performed for several significant reasons:

- **Improving Model Performance**: Many machine learning algorithms are sensitive to the scale of the input features. Features with larger scales may dominate those with smaller scales, leading to biased model training. Scaling ensures that all features contribute equally to the model's learning process, preventing dominance by any particular feature.

- **Faster Convergence**: Scaling can help algorithms converge more quickly during optimization, especially for algorithms that use gradient descent-based optimization techniques. Normalizing the scale of features can result in a more efficient optimization process.

- **Facilitating Interpretation**: Scaling can make it easier to interpret the coefficients or weights of features in linear models. When features are on the same scale, the coefficients represent the relative importance of each feature more accurately.

- **Regularization**: Some regularization techniques, such as Ridge Regression and Lasso Regression, are sensitive to the scale of features. Scaling ensures that regularization penalties are applied uniformly across all features.

There are **two common scaling techniques:**

- **Normalized Scaling (also known as Min-Max scaling):**
  - Normalized scaling transforms the features so that they fall within a specified range, typically between 0 and 1. It also **helps in taking care of outliers.**
  - The formula for normalized scaling is:

$$MinMaxScaling: x = \frac{x - \min(x)}{\max(x) - \min(x)}$$

  - Normalized scaling **preserves the relative relationships between data points** and is suitable for algorithms that require input features to be within a bounded range, such as neural networks with activation functions that have bounded output ranges.

- **Standardized Scaling (also known as Z-score scaling):**
  - Standardized scaling transforms the features **so that they have a mean of 0 and a standard deviation of 1.**
  - The formula for standardized scaling is:

$$Standardization: x = \frac{x - mean(x)}{sd(x)}$$

  - Standardized scaling centers the data around zero and ensures that it has a consistent scale, regardless of the original distribution of the data.
  - Standardized scaling is suitable for algorithms that assume normally distributed data or rely on distance-based calculations, such as K-nearest neighbors (KNN) and support vector machines (SVM).

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

Ans:

This typically occurs when one or more features in the dataset are perfectly collinear, meaning that they can be expressed as exact linear combinations of each other. In such situations, the VIF calculation involves dividing by zero, resulting in an infinite value.

Perfect Collinearity: Perfect collinearity occurs when two or more independent variables in the dataset are perfectly correlated with each other. Mathematically, if one variable can be expressed as a linear combination of others, it indicates perfect collinearity.

VIF Calculation: The VIF for each feature is calculated by regressing that feature against all other features in the dataset. Mathematically, the VIF for a feature is equal to $\boldsymbol{VIF = \frac{1}{1-R^2}}$**, where $\boldsymbol{R^2}$ is the coefficient of determination from the linear regression** of that feature against all other features.

Infinite VIF: **When perfect collinearity exists, the coefficient of determination ($\boldsymbol{R^2}$) in the linear regression becomes 1**, indicating that all variation in the dependent variable can be explained by the independent variable. **As a result, the denominator of the VIF formula becomes zero, leading to an infinite VIF value.**

An infinite VIF indicates severe multicollinearity issues in the dataset, making it challenging to estimate the coefficients of the regression model accurately and **impact the stability and reliability of model.**

Addressing Perfect Collinearity: To **address perfect collinearity, one of the correlated features can be removed from the model**. Alternatively, feature transformation techniques such as **principal component analysis (PCA) can be applied to reduce the dimensionality of the dataset** while retaining most of the information.

In summary, **infinite VIF values occur when there is perfect collinearity between features** in the dataset. Detecting and addressing multicollinearity issues is crucial for ensuring the stability and accuracy of regression models.

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

Ans:

Q-Q plots are the quantile-quantile plots. It is a graphical tool to assess if data sets are from common distribution. The theoretical distributions could be of type normal, exponential or uniform. The Q-Q plots are useful in the linear regression to identify the train data set and test data set are from the populations with same distributions.

Working of Q-Q plots –

Quantiles: Quantiles divide a dataset into intervals of equal probability. For example, the median is the 50th percentile, dividing the dataset into two equal parts.

Comparison: In a Q-Q plot, the quantiles of the sample dataset are plotted on the horizontal axis, while the quantiles of the theoretical distribution (e.g., normal distribution) are plotted on the vertical axis.

Linearity: If the sample dataset follows the theoretical distribution closely, the points on the Q-Q plot will fall close to a diagonal line (the line of equality). Deviations from the diagonal line indicate departures from the theoretical distribution.

Types of Q-Q plots

1. **Normal Distribution**: A symmetric distribution where the Q-Q plot would show points approximately along a diagonal line if the data adheres to a normal distribution.
2. **Right-skewed Distribution**: A distribution where the Q-Q plot would display a pattern where the observed quantiles deviate from the straight line towards the upper end, indicating a longer tail on the right side.
3. **Left-skewed Distribution**: A distribution where the Q-Q plot would exhibit a pattern where the observed quantiles deviate from the straight line towards the lower end, indicating a longer tail on the left side.
4. **Under-dispersed Distribution**: A distribution where the Q-Q plot would show observed quantiles clustered more tightly around the diagonal line compared to the theoretical quantiles, suggesting lower variance.
5. **Over-dispersed Distribution**: A distribution where the Q-Q plot would display observed quantiles more spread out or deviating from the diagonal line, indicating higher variance or dispersion compared to the theoretical distribution.

Advantages of Q-Q plot -

- Distribution aspects like loc, scale shifts, symmetry changes and the outliers all can be daintified from the single plot.
- Easily detects departures from assumed distributions, aiding in identifying data discrepancies.
- The plot has a provision to mention the sample size as well can compare datasets of different sizes without requiring equal sample sizes.

The use and importance of a Q-Q plot in linear regression include:

1. **Normality Assumption**: In linear regression, it is often assumed that the residuals (the differences between the observed and predicted values) follow a normal distribution. A Q-Q plot of the residuals helps assess whether this assumption holds. If the points on the Q-Q plot form a roughly straight line, it suggests that the residuals are normally distributed. Deviations from the line indicate departures from normality.
2. **Model Assessment**: Q-Q plots can be used to assess the goodness-of-fit of the regression model. If the residuals follow a normal distribution, it suggests that the model adequately captures the variation in the data. However, if the residuals deviate from normality, it indicates that the model may not be appropriate for the data.
3. **Identifying Outliers**: Q-Q plots can help identify outliers or data points that do not conform to the expected distribution. Outliers may appear as points that deviate significantly from the diagonal line on the Q-Q plot, suggesting that they may have a different distribution than the rest of the data.
4. **Model Validation**: Q-Q plots are a useful tool for validating the assumptions of linear regression models. By visually inspecting the Q-Q plot, analysts can assess whether the normality assumption holds and whether the model adequately fits the data.

Overall, Q-Q plots provide valuable insights into the distributional properties of the data and help ensure the validity and reliability of linear regression models. They are an essential tool in the diagnostic process of linear regression analysis.