Assignment-based Subjective Questions

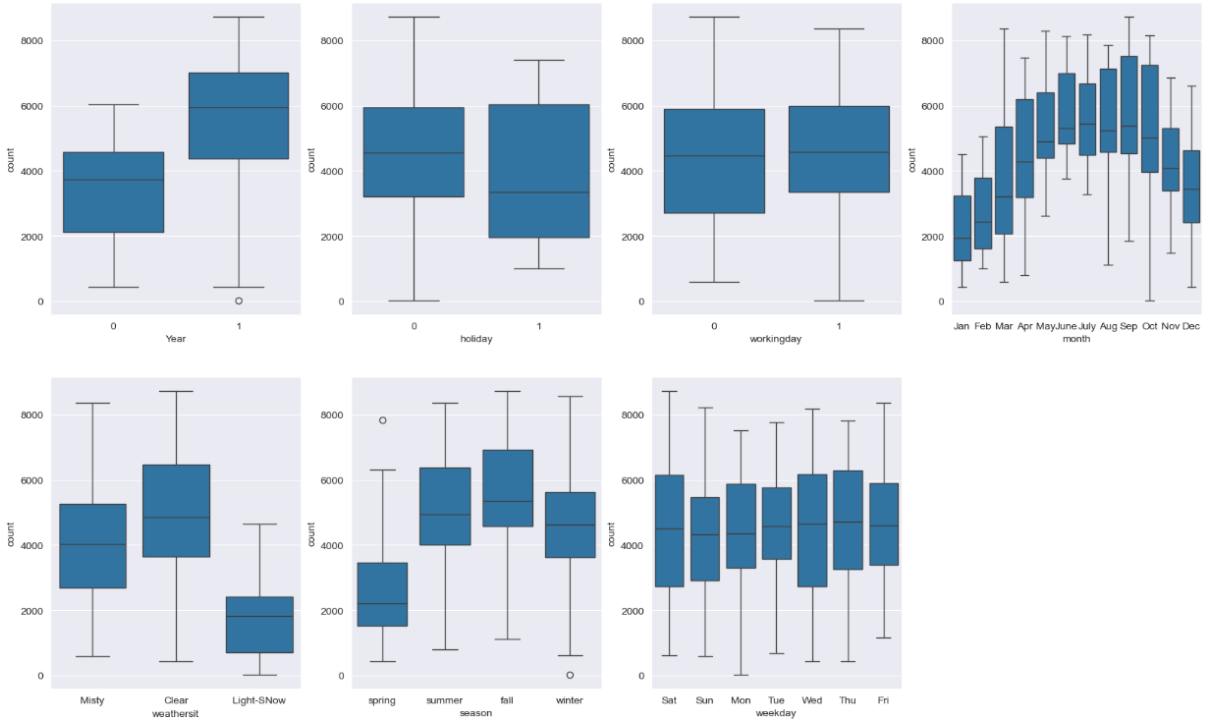
Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans:

The categorical variable used in the dataset,

1. Year
2. Holiday
3. Working day
4. Month
5. Weather-situation
6. Season
7. Weekday

Here is a visualisation of each variable against target variables,



Observations:

- Bike rentals have **increased** significantly in **2019** from 2018, indicating healthy business growth.
- Bike rentals are **less** on holidays.
- Bike Rentals **does not vary** much on working/non-working day
- Bike rentals **increases** during month of July- September, it reaches **peak** in **September** and **decline** in **winter** months, it is directly related to weather conditions.

- Weather has a significant effect on bike rentals, **clear weather** indicate increase in rentals.
- Season has a significant effect on bike rentals, bike rentals are more in **fall and summer**.

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

Ans:

Dummy variable creation is a technique used in statistical modelling and machine learning to represent categorical variables with binary values (0 or 1). It involves creating new binary (dummy) variables for each category of the original categorical variable. These dummy variables serve as indicators for the presence or absence of a specific category.

drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

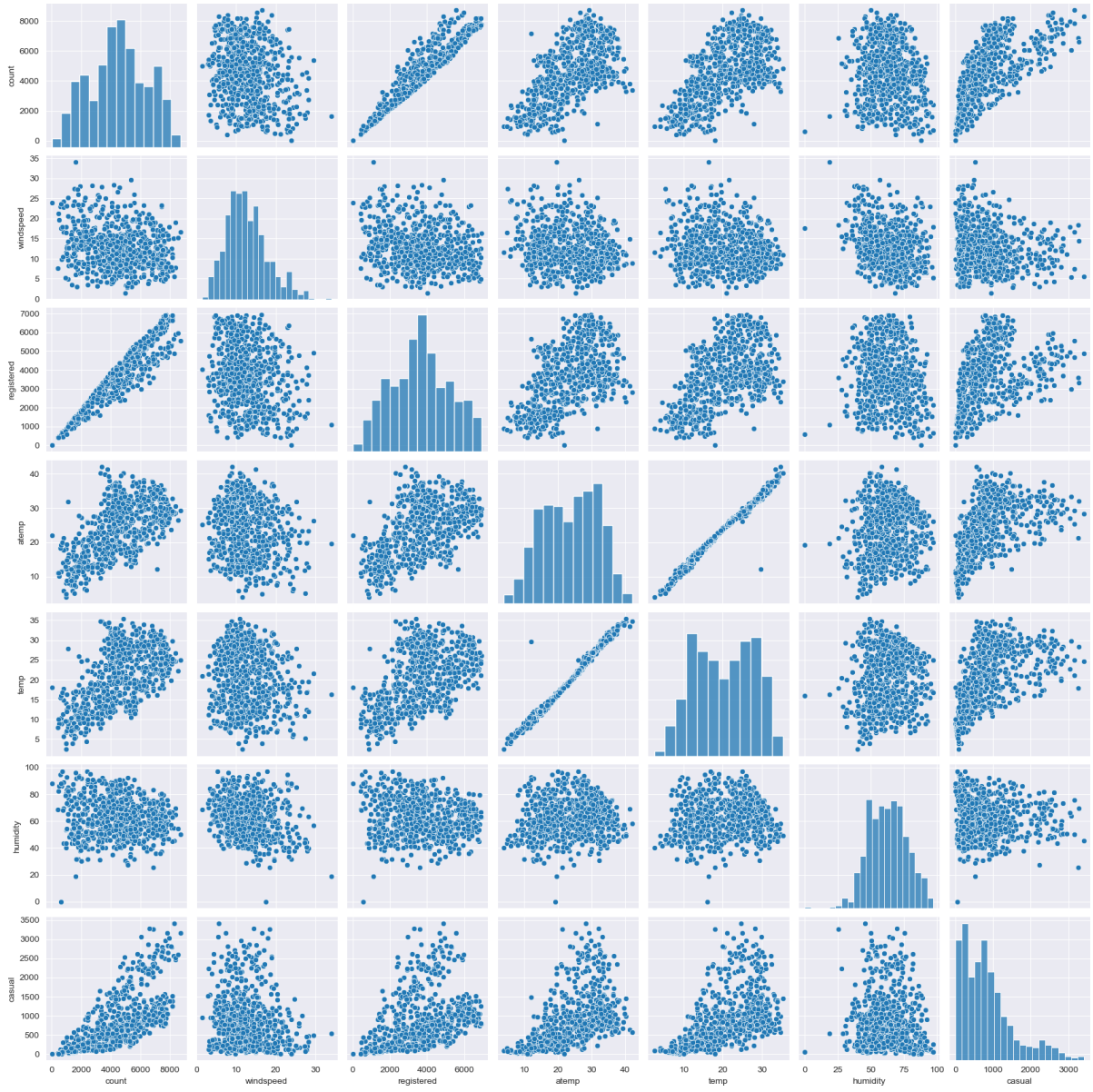
Ex. There are 4 seasons in assignment, we can represent them using 3 dummy variables like,

- 000 will corresponds to fall.
- 100 will corresponds to spring.
- 010 will corresponds to summer.
- 001 will corresponds to winter.

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

Ans:

Based on the pair-plot among numerical variables below,



And the final equation derived from model training,

count = 0.0902 + 0.4914×temp + 0.0916×Sep + 0.0645×Sat + 0.0527×summer + 0.0970×winter + 0.2334×Year + 0.0566×workingday − 0.03041×light-snow − 0.0786×misty − 0.065×spring

We can conclude that “**temp**” and “**atemp**” are the two numerical variables which are highly correlated with the target variable (**cnt**).

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

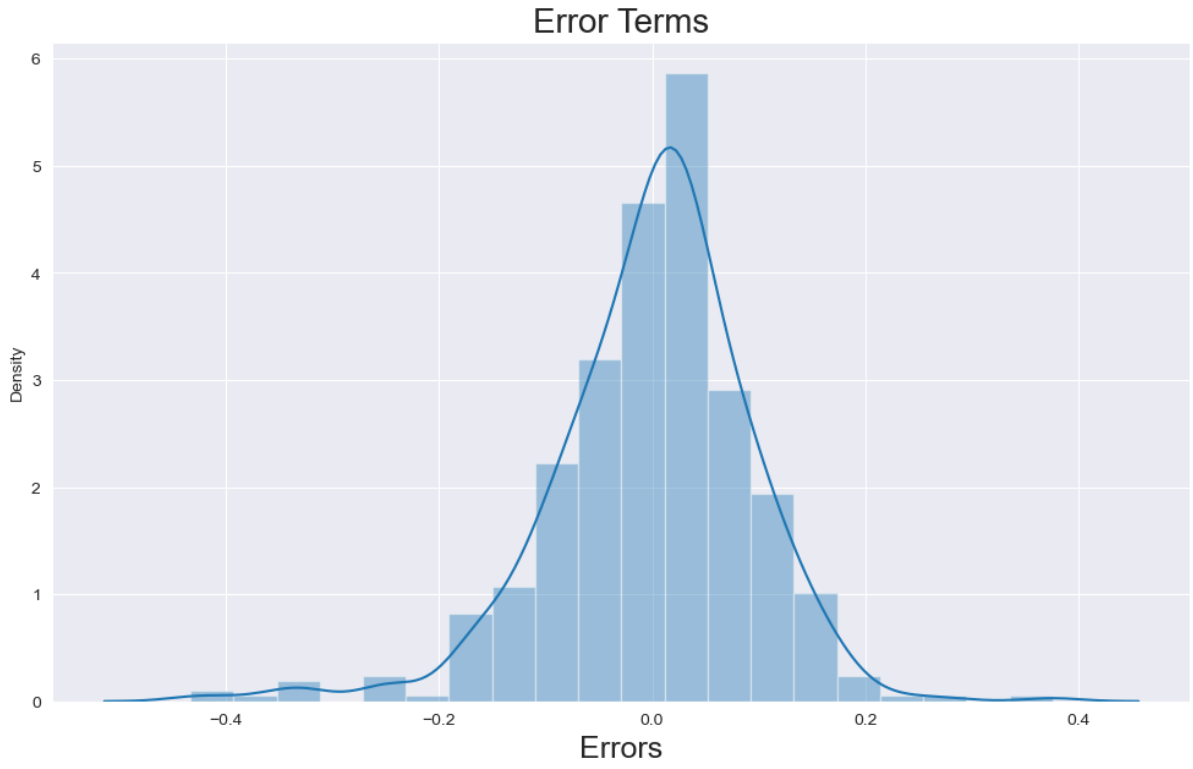
Ans:

Validating the assumptions of linear regression is a crucial step to ensure the reliability of the model. After building the model on the training set, here are the steps I followed to validate the assumptions:

1. Residual Analysis:

- Process: Examine the residuals (the differences between observed and predicted values).

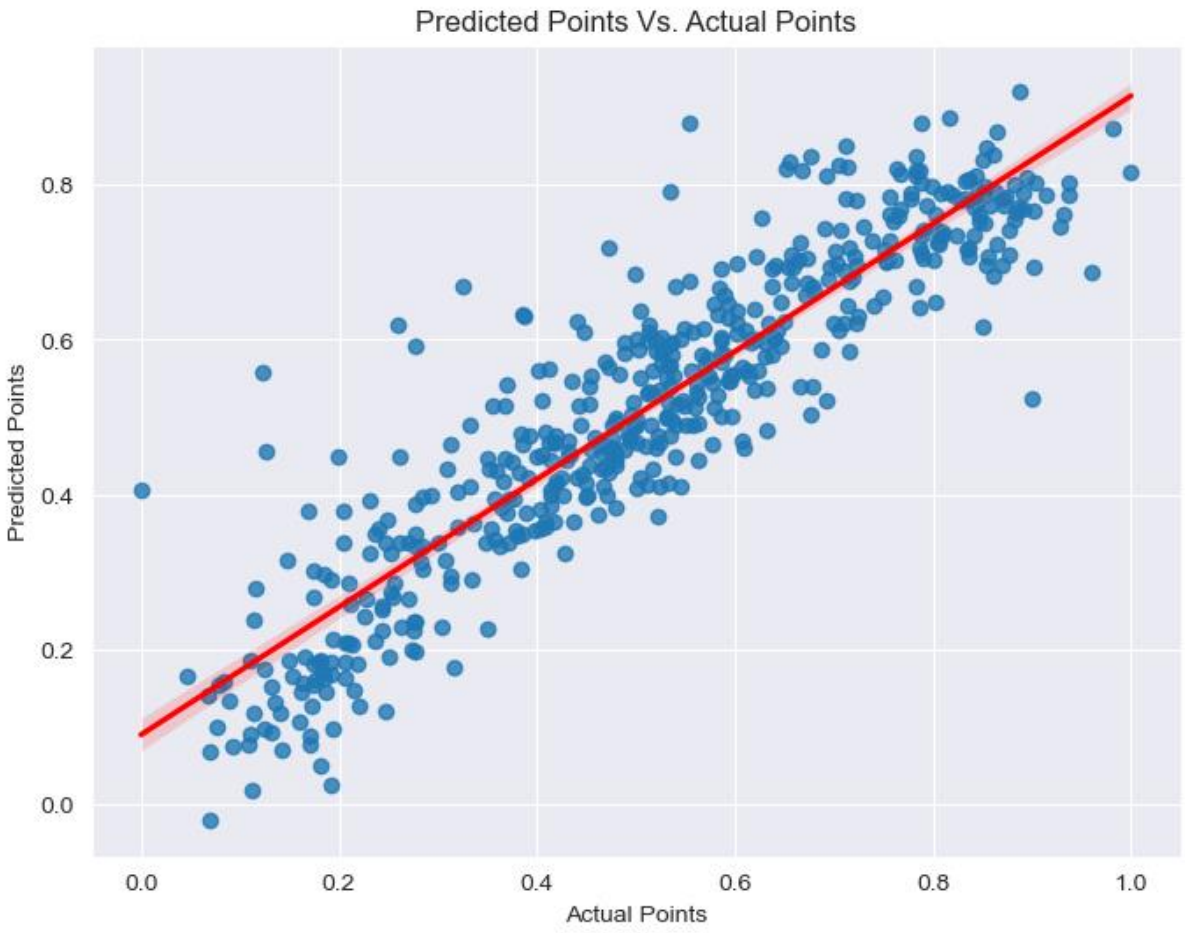
- Check: Residuals should be approximately normally distributed, and there should be no visible patterns in the residual plot.



2. Homoscedasticity (Constant Variance) and Linearity:

- Process: Plot residuals against predicted values.

- Check: The spread of residuals should be roughly constant across all levels of the predicted values.



3. Multicollinearity:

- Process: Calculate Variance Inflation Factors (VIF) for predictor variables.

- Check: VIF values should be below a certain threshold (commonly 5 or 10) to ensure no problematic collinearity.

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

Ans:

From the equation of the best fit line:

count = 0.0902 + 0.4914×temp + 0.0916×Sep + 0.0645×Sat + 0.0527×summer + 0.0970×winter + 0.2334×Year + 0.0566×workingday − 0.03041×light-snow − 0.0786×misty − 0.065×spring

The following three features significantly contribute to explaining the demand for shared bikes:

- Temperature (temp)

- Winter season (winter)
- Calendar year (year)

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.
Linear regression is a machine learning algorithm based on supervised learning. It performs a regression task, which means it predicts a continuous output variable (y) based on one or more input variables (x). It is mostly used for finding out the linear relationship between variables and forecasting.
The basic idea of linear regression is to find a line that best fits the data points, such that the distance between the line and the data points is minimized. The line can be represented by an equation of the form:
$y = \theta 0 + \theta 1x$
where θ0 is the intercept (the value of y when x is zero) and θ1 is the slope (the change in y for a unit change in x). These are called the parameters or coefficients of the linear model.

To find the best values of θ0 and θ1, we need to define a cost function that measures how well the line fits the data. A common choice is the mean squared error (MSE), which is the average of the squared differences between the actual y values and the predicted y values:

$MSE = (1/n) * \sum(y - y')^2$

where n is the number of data points, y is the actual value, and y' is the predicted value.

The goal is to minimize the MSE by adjusting θ0 and θ1. There are different methods to do this, such as gradient descent, normal equation, or using libraries like scikit-learn.

Linear regression can also be extended to multiple input variables (x1, x2, …, xn), in which case the equation becomes:

$y = \theta 0 + \theta 1x1 + \theta 2x2 + \ldots + \theta nxn$

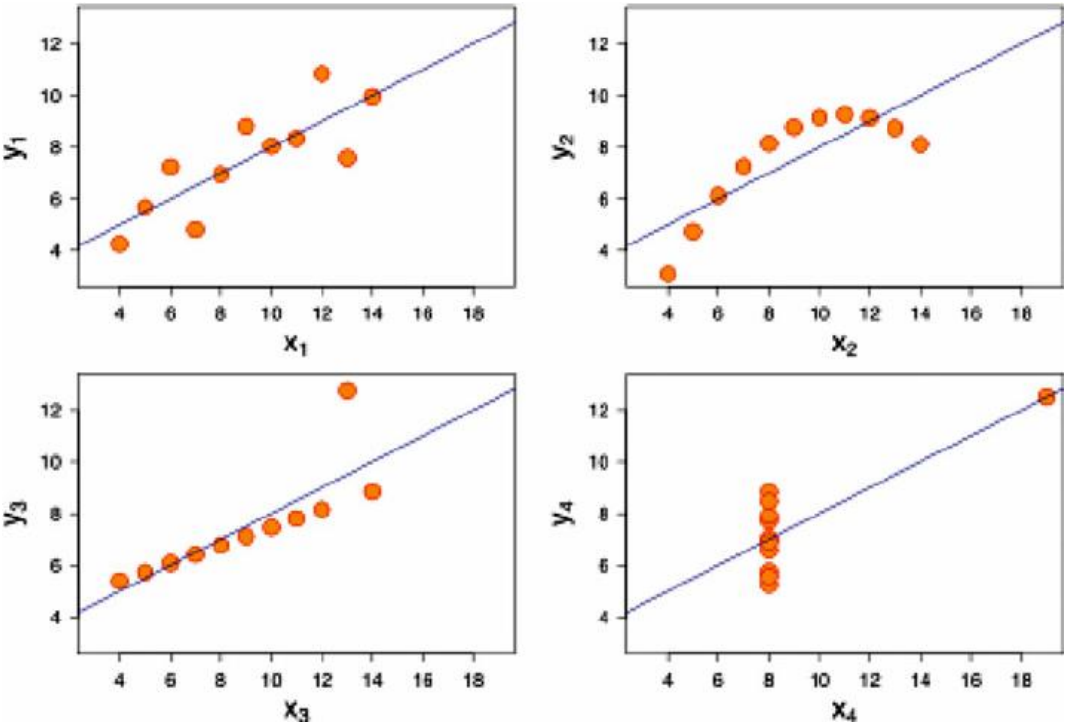
Limitations are it assumes a linear relationship between the input variables and the output variable, which may not always be the case. Another limitation is that it may be sensitive to outliers or multicollinearity.

Q2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph. It was developed to

emphasize both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties.

- The first scatter plot (top left) appears to be a simple linear relationship.
- The second graph (top right) is not distributed normally; while there is a relation between them is not linear.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line the calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.



Q3. What is Pearson's R?

Pearson's r is a numerical summary of the strength of the linear association between the variables. Its values range between -1 to +1. It shows the linear relationship between two sets of data. In simple terms, it tells us "Can we draw a line graph to represent the data?"

Formula

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

r = correlation coefficient

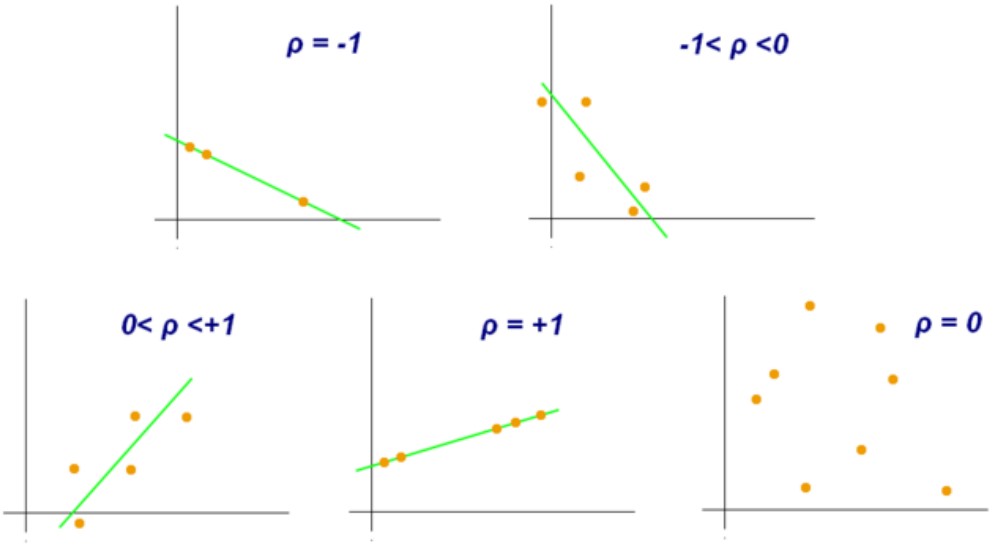
x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

As can be seen from the graph below, r = 1 means the data is perfectly linear with a positive slope r = -1 means the data is perfectly linear with a negative slope r = 0 means there is no linear association



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

Feature scaling is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data pre-processing stage to deal with varying values in the dataset. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values.

- Normalization is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbours and Neural Networks.
- Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

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

The VIF (Variance Inflation Factor) gives how much the variance of the coefficient estimate is being inflated by collinearity. If there is perfect correlation, then VIF = infinity. It gives a basic quantitative idea about how much the feature variables are correlated with each other. It is an extremely important parameter to test our linear model.

$$VIF = \frac{1}{1 - R^2}$$

Where R-1 is the R-square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables. If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and its R-squared value will be equal to 1. So, VIF = 1/(1-1) which gives VIF = 1/0 which results in “infinity” The numerical value for VIF tells you (in decimal form) what percentage the variance (i.e. the standard error squared) is inflated for each coefficient. For example, a VIF of 1.9 tells you that the variance of a particular coefficient is 90% bigger than what you would expect if there was no multicollinearity — if there was no correlation with other predictors.
A rule of thumb for interpreting the variance inflation factor:

- 1 = not correlated.
- Between 1 and 5 = moderately correlated.
- Greater than 5 = highly correlated.

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

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. It is used to compare the shapes of distributions. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that’s roughly straight. The q-q plot is used to answer the following questions:

- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?

- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behaviour?