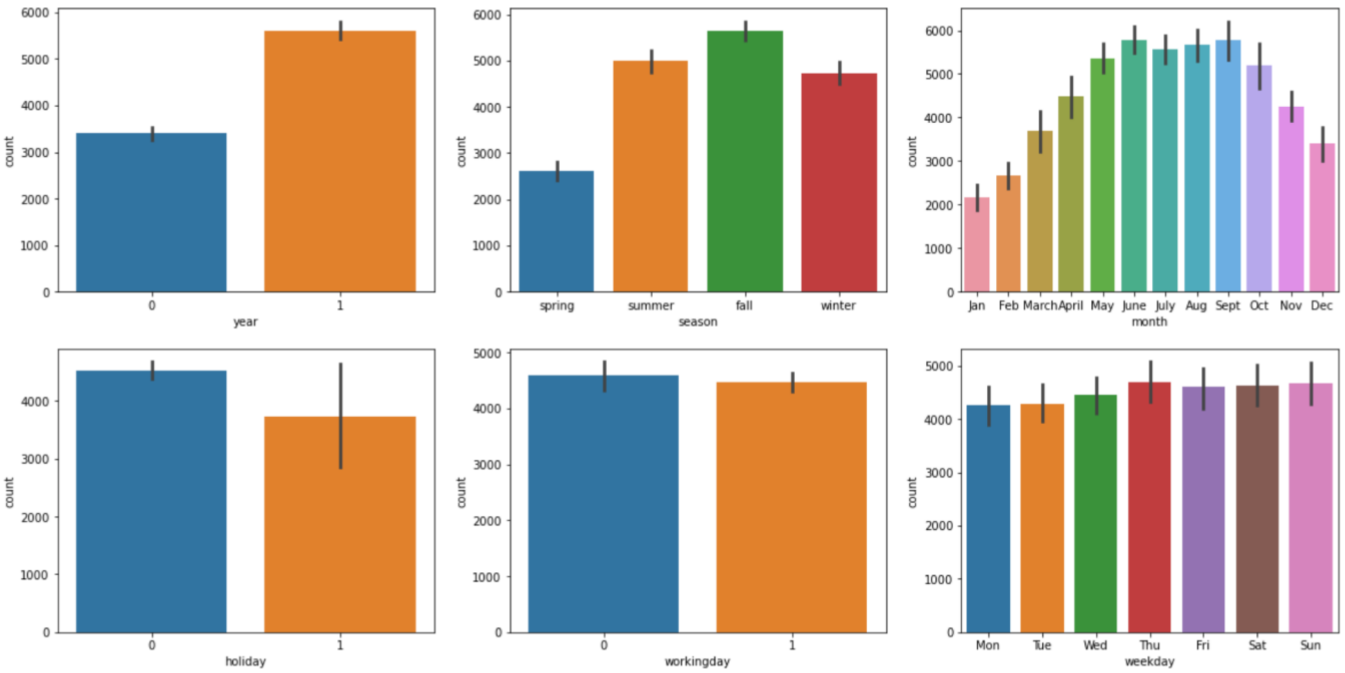
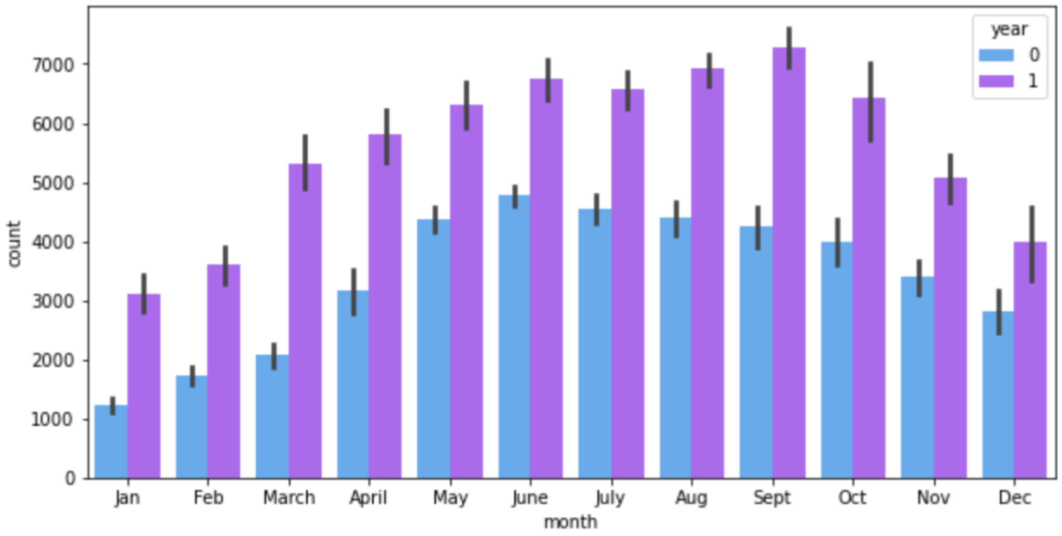
Assignment-based Subjective Questions

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

   The dependent variable in our dataset was count. Fall weather came around to be highest as is it a very pleasant weather for a bike ride. The counts increased quite a number in just a year from 2018 to 2019, reason being much greater awareness among people and it is becoming a popular trend. People rented more on holidays than on non-holidays. Whereas working and non-working days had the same median.

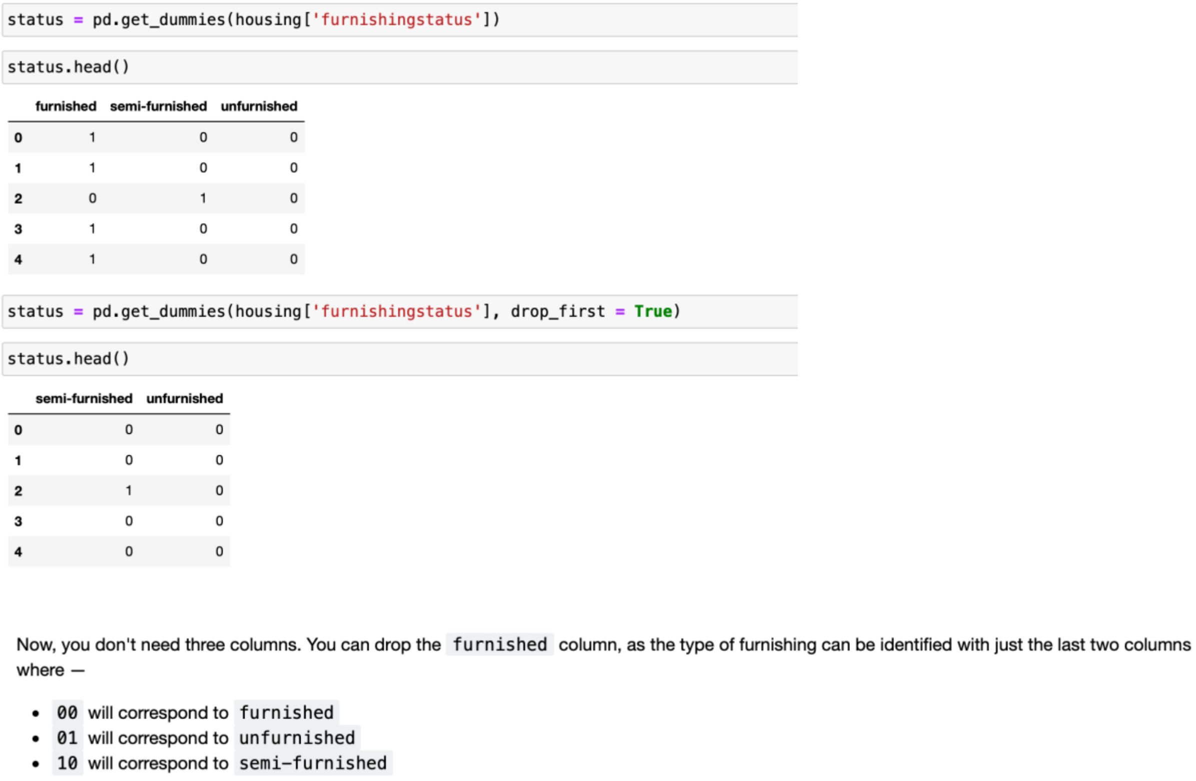
   Clear weather is more optimal for bike renting.





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

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.



```
status = pd.get_dummies(housing['furnishingstatus'])
```

```
status.head()
```

	furnished	semi-furnished	unfurnished
0	1	0	0
1	1	0	0
2	0	1	0
3	1	0	0
4	1	0	0

```
status = pd.get_dummies(housing['furnishingstatus'], drop_first = True)
```

```
status.head()
```

	semi-furnished	unfurnished
0	0	0
1	0	0
2	1	0
3	0	0
4	0	0

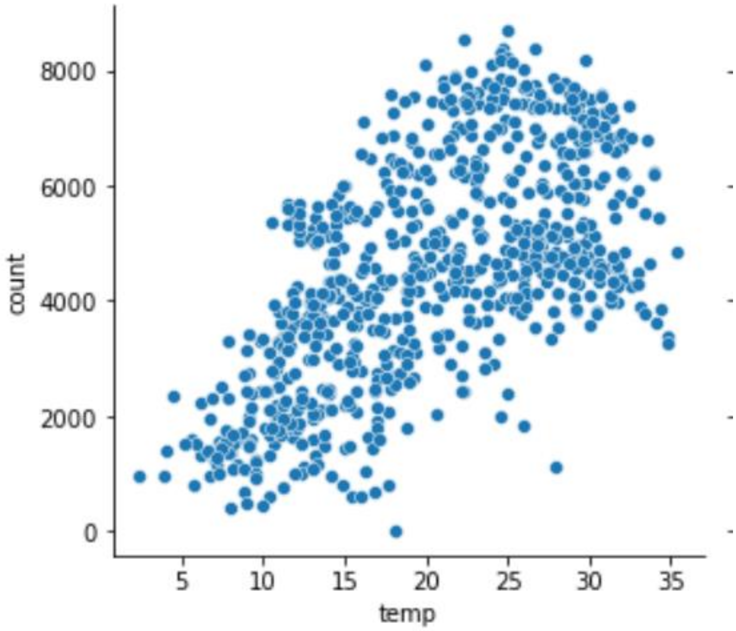
Now, you don't need three columns. You can drop the `furnished` column, as the type of furnishing can be identified with just the last two columns where —

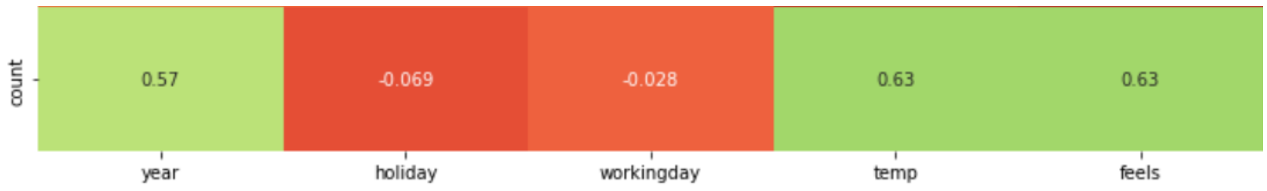
- `00` will correspond to `furnished`
- `01` will correspond to `unfurnished`
- `10` will correspond to `semi-furnished`

Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables. This also helps to avoid multicollinearity.

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

   By looking at the pair plot temp variable has the highest (0.63) correlation with target variable 'count'.





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

   We can check the multicollienearity using heat map for the final X_train and we can also check the y_train predicted y_trian using scatterplot.

   We can also validate by checking the distribution of the residuals, they should be normally distributed. If it is so then we can say that the model is working well.

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

   The Top 3 features contributing significantly towards the demands of share bikes are:

   Day_weather_Light_Snow(negative correlation)(0.2535).

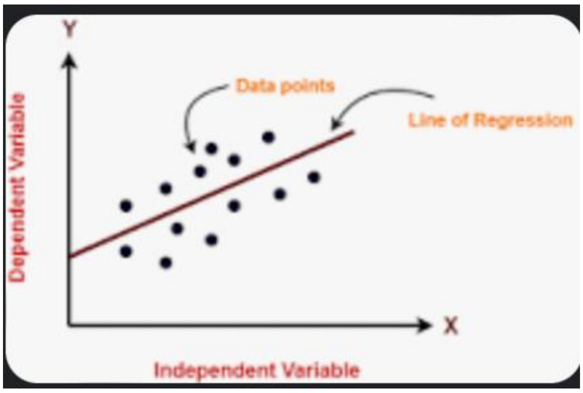
   year_2019(Positive correlation)(0.2334).

   temp(Positive correlation)(0.5682).

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Based on supervised learning, linear regression is a Machine Learning algorithm which performs the task of regression which helps to target prediction values based on independent variables.



Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of our data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

Mathematically, we can write a linear regression equation as:

y = a + bx

$$b(slobe) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

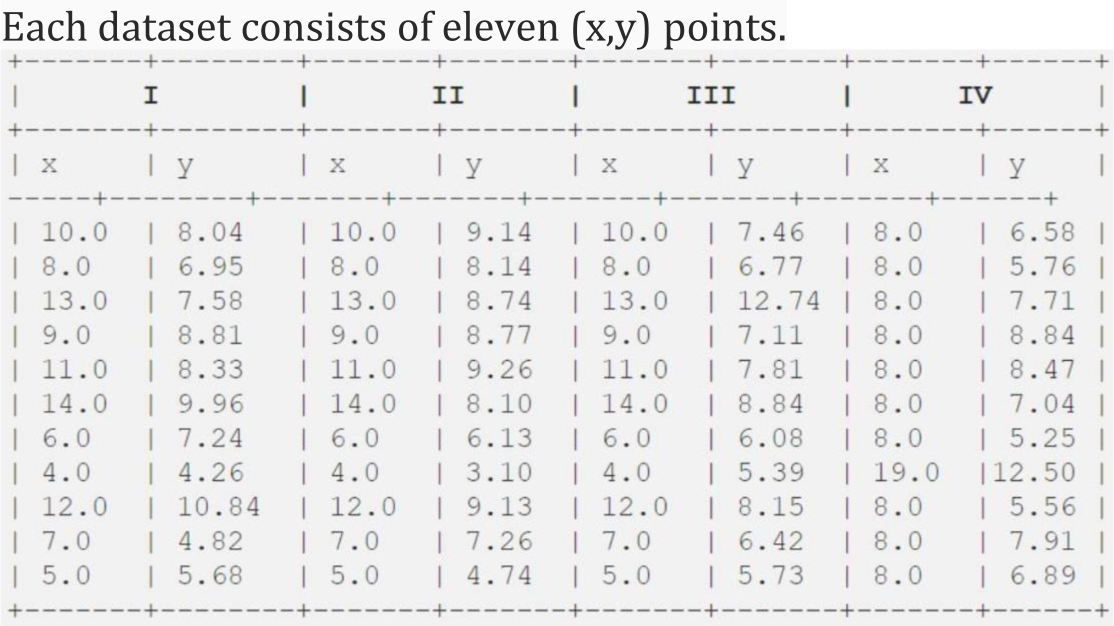
$$a(intercept) = \frac{n\sum y - b(\sum x)}{n}$$

2. Explain the Anscombe’s quartet in detail.

Constructed in 1973 by the statistician Francis Anscombe. It is a group of 4 datasets having nearly identical statistical properties, but somewhere during the analysis when the regression model is applied the analysis is led into false pretexts. However being identical datasets their scatterplot comes out to be different from one another.

It was used to show the importance of graphing the data before analysing it and it shed light on the on the effect of outliers on statistical properties.

⇒ Contents of Anscombe’s quartet :



Each dataset consists of eleven (x,y) points.

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

3. What is Pearson’s R?

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

⇒ r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)

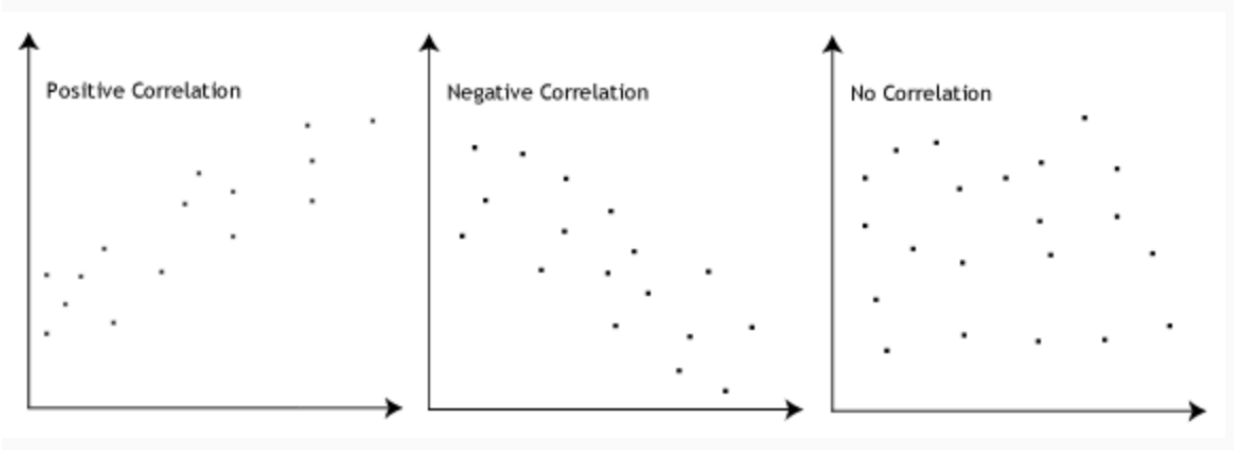
⇒ r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)

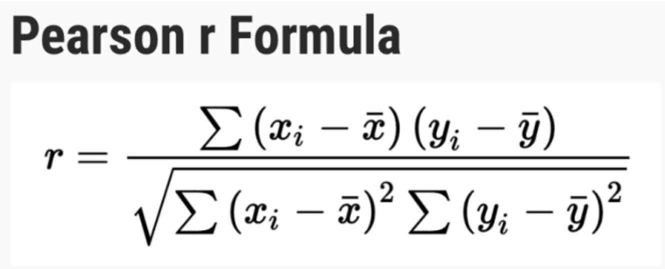
⇒ r = 0 means there is no linear association

⇒ r > 0 < 5 means there is a weak association

⇒ r > 5 < 8 means there is a moderate association

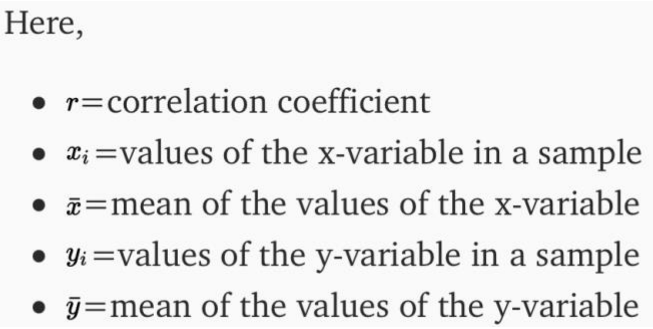
⇒ r > 8 means there is a strong association





Pearson r Formula

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



Here,

- 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

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

When we measure certain objects/values and assign them or categorise them according to certain specified rules, then that procedure is known as scaling in other words it is a procedure of locating the measured objects on the continuum.

WHY SCALING?

Most of the times, collected data set contains features highly varying in magnitudes, units and range. In simple words data with different measuring units cannot be grouped together. So to avoid such situation pre-processing of data is done to independent variables so that the data can be normalised within a particular range and side by side helping to ease out the speed as well as the calculation accuracy in an algorithm.

Difference between standardised and normalised scaling:

Normalisation	Standardisation
Minimum and maximum value of features are used for scaling.	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is really affected by outliers	Not so much
Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
It is useful when we don’t know about the distribution.	It is useful when the feature distribution is Normal or Gaussian.
It is often called as scaling normalisation.	It is often called as z-score normalisation.

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

This happens due to the presence of a perfect correlation between two different independent variables.

This is bound to happen when the value of r^2 = 1.

As we know the formula of V.I.F =

$$VIF_i = \frac{1}{1 - R_i^2}$$

So when r^2 gets to 1 the V.I.F stretches out to infinity.

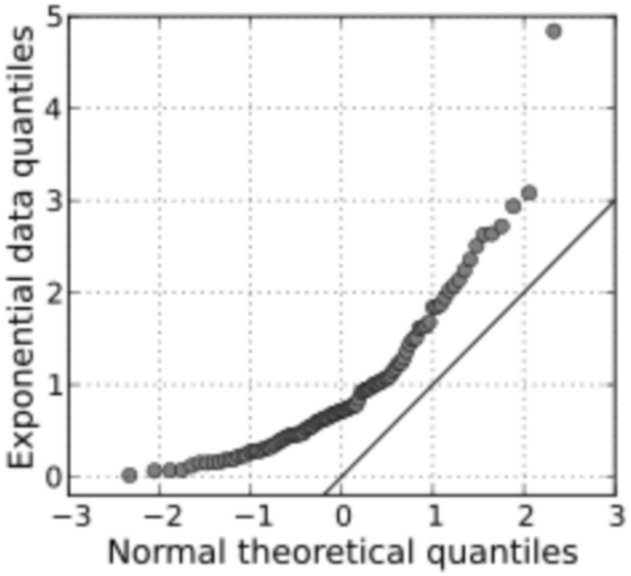
To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

A quantile to quantile or as we say Q-Q plot is a graphical tool used to represent the data's distribution that whether it is a Normal, exponential or Uniform distribution.

These are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it.



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.