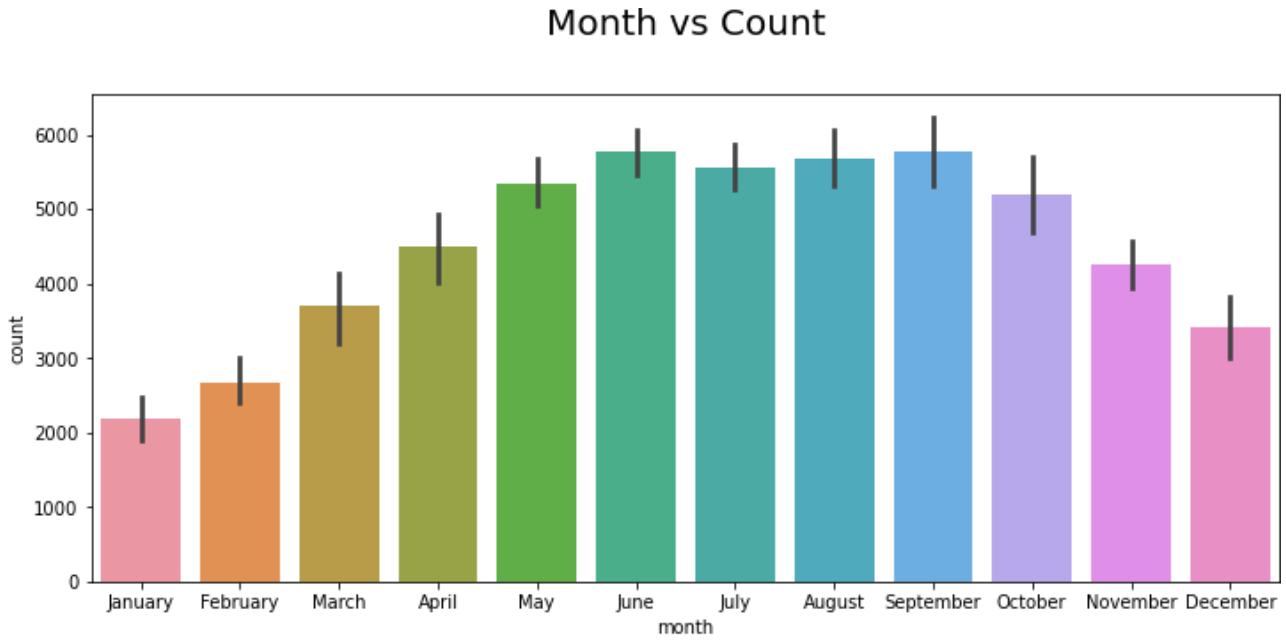
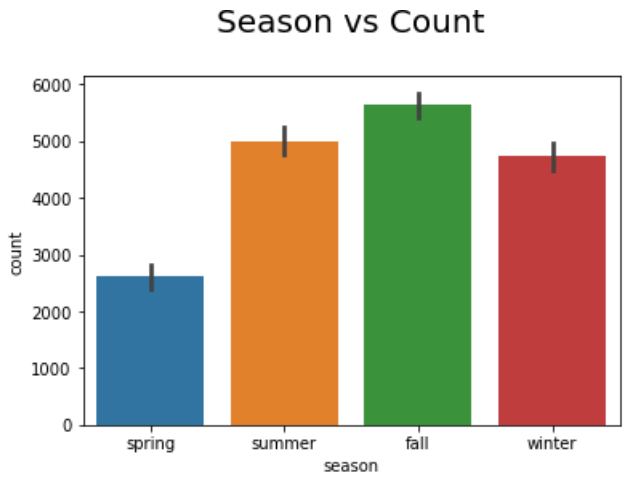
Assignment-based Subjective Questions

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

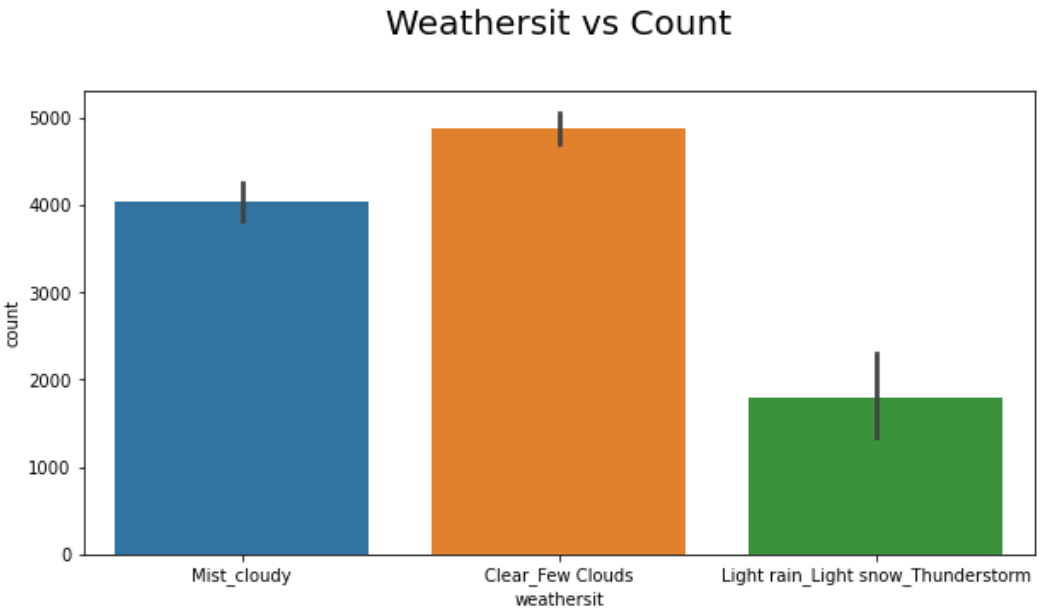
Ans:



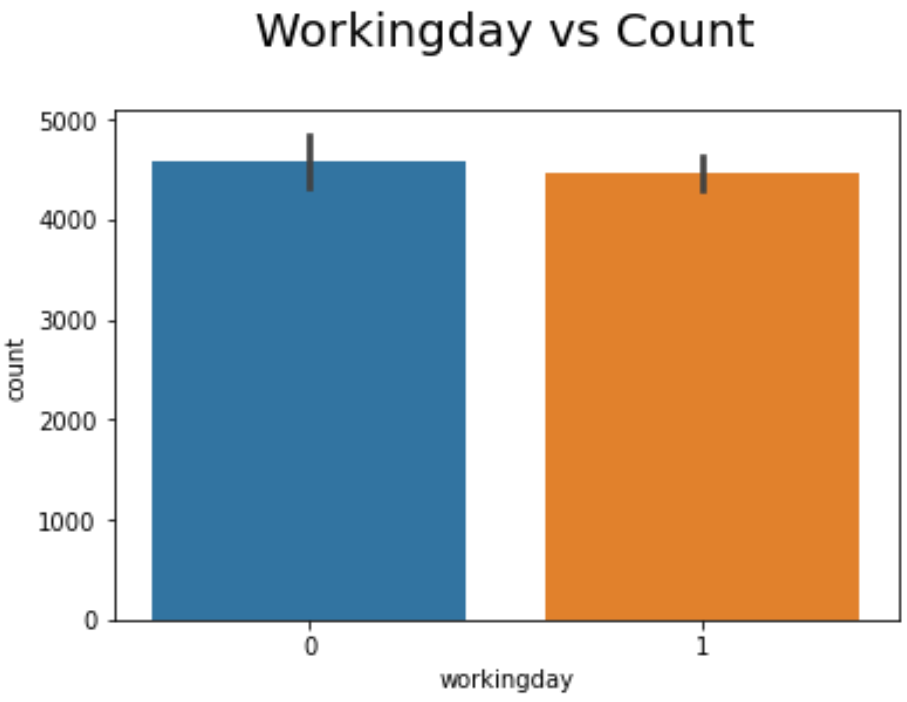
Months from **May** to **October** recorded highest bike sharings.



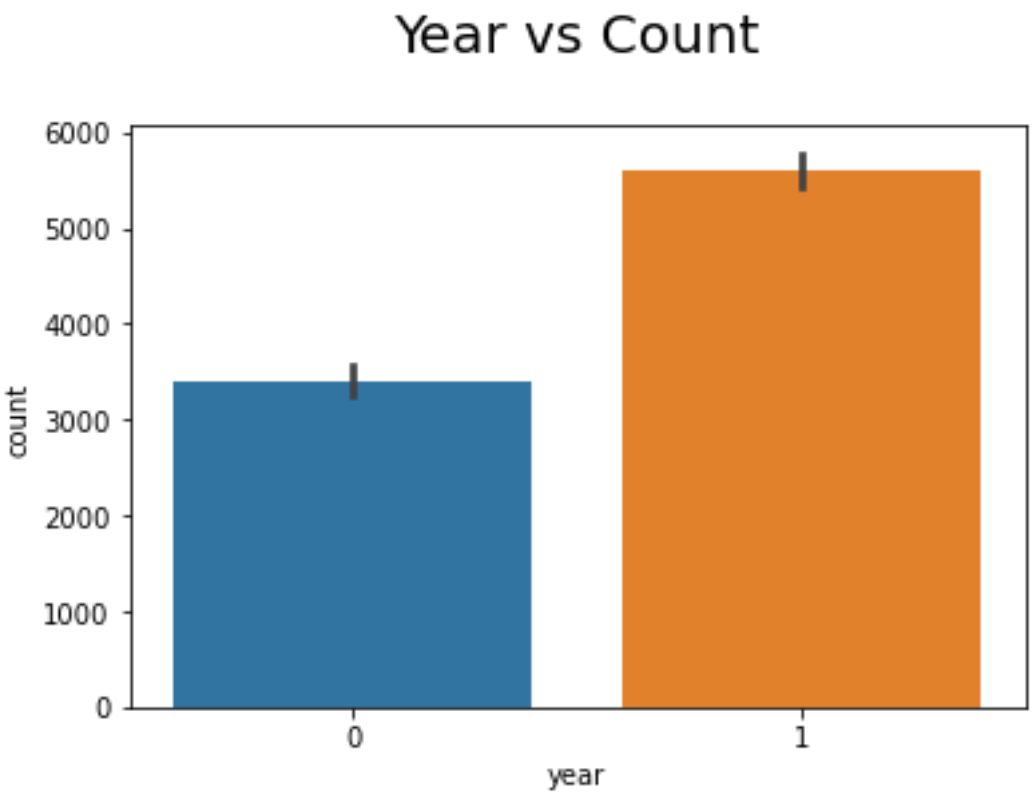
We can clearly observe that the bikes were shared mostly in the **fall** season followed by **summer**, **winter** and the least in **spring**.



Lowest bike sharing was recorded in the weather of Light rain_Light snow_Thunderstrorm, which is quite understandable.



There's no such significant difference between the number of bikes shared in workingday.



Year 2019 i.e.1 recorded the more numbers of bike sharing compared to that in year 2018 i.e.0.

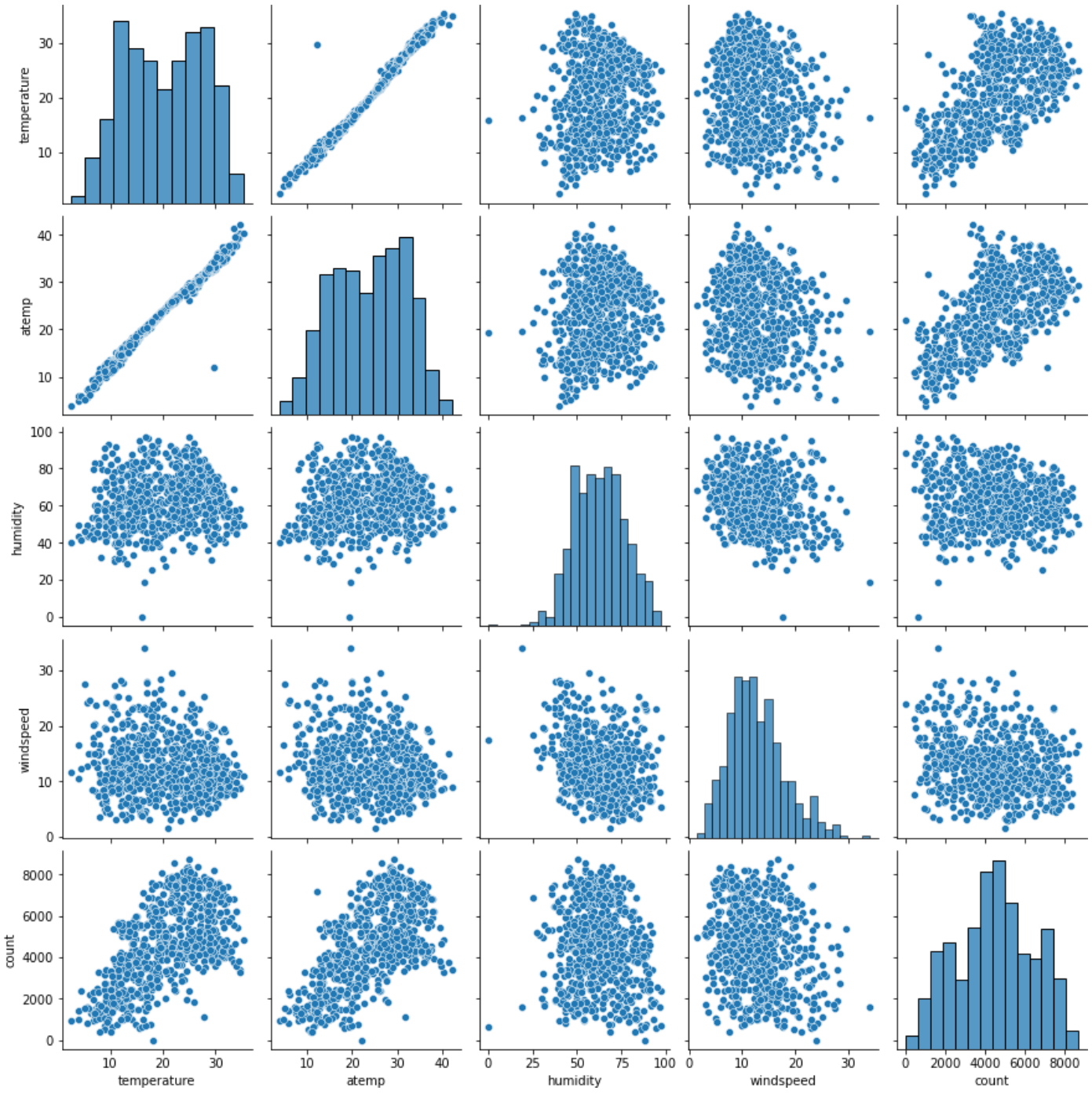
Q2. Why is it important to use drop_first=True during dummy variable creation?

Ans: drop_first=True is used because it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Also, we can say that drop_first allows us to drop our first variable and identify it through all other columns being 0.

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

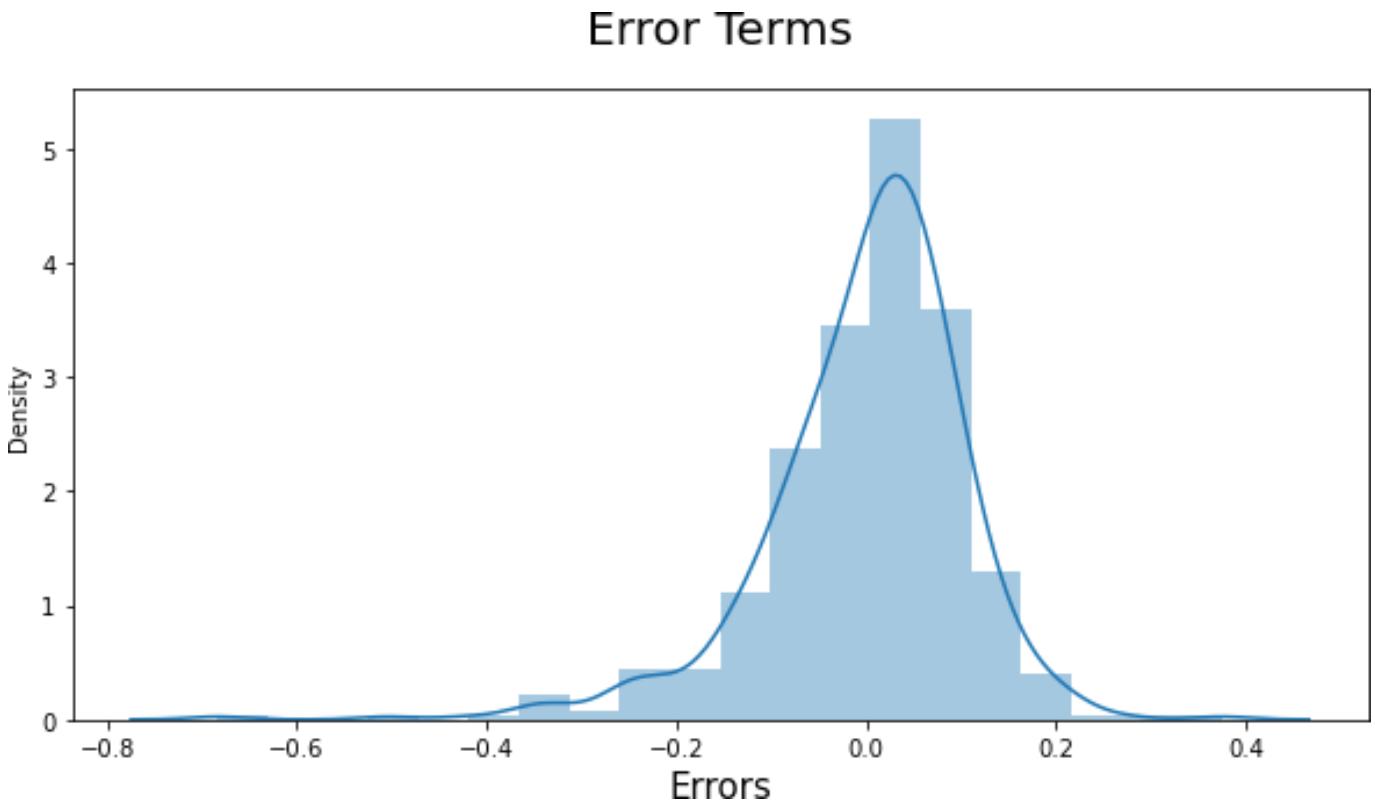
Ans:



Clearly temperature and atemp are highly correlated. Other than that, they both are correlated slightly well with the count variabl

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

Ans:



The above diagram shows that the residuals are normally distributed and follow a mean at zero.

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

Ans: The top three features for my model were:

1) Temperature: coefficient = 0.5184

2) Year: coefficient = 0.2382

3) Winter season(season_winter): coefficient = 0.0726

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Ans: When there's a relationship established between an independent and dependent variable using a straight line, this linear estimation is called as Linear regression. It's a basic form of regression analysis used in Machine learning for predicting models built of various functions.

In linear regression, we try to calculate the best fit straight line which describes the relationship between the predictors and predictive/dependent variable.

Mathematically, we can write a linear regression equation as:

$$y = mx+c$$

Were,

m = slope of the line

c = y-intercept of the line

x = independent variable in the dataset

y = dependent variable in the dataset

Generally, linear regression is divided into two types:

1. Multiple linear regression: As the word suggests, in this type of linear regression we try to discover the relationship between two or more independent variables or inputs and the corresponding dependent variable or output and the independent variables can be either continuous or categorical.

2. Simple linear regression: In a simple linear regression, we aim to reveal the relationship between a single independent variable or you can say input, and a corresponding dependent variable or output.

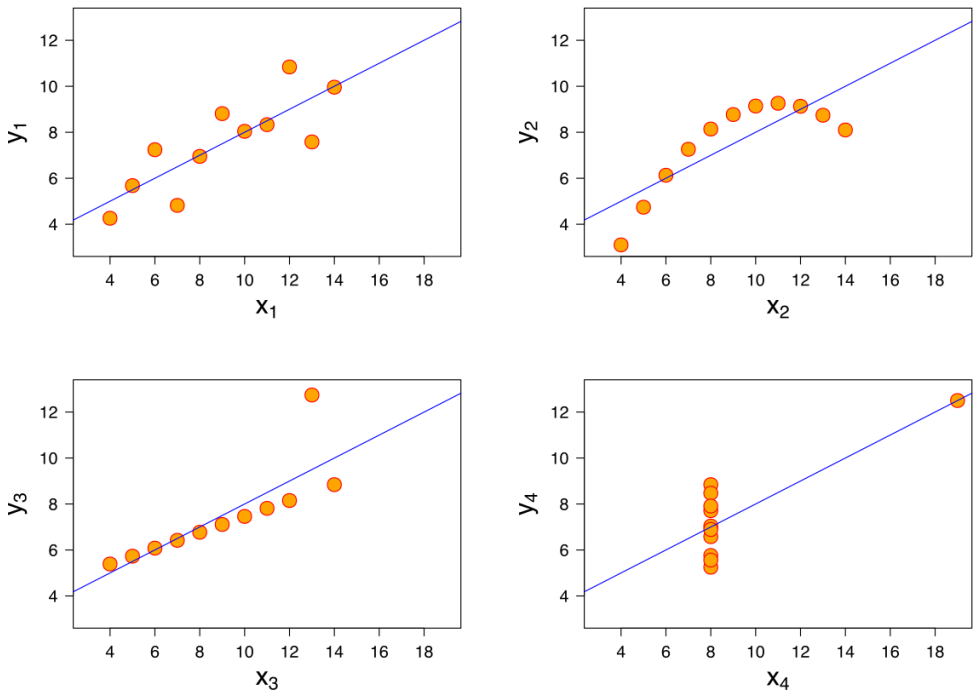
Q2. Explain the Anscombe’s quartet in detail.

Ans. Basically, an Anscombe's quartet comprises of four datasets that are nearly identical in terms of simple descriptive statistics and yet have very different distributions and appear very different when they are graphed.

It was constructed by statistician Francis Anscombe in the year 1973 to illustrate the importance of plotting the graphs before analyzing and building the model, and the effect of other observations on statistical properties.

This tells us a lot about the importance of visualizing the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc.

Example:



Q3. What is Pearson's R?

Ans: Pearson's r is the Pearson's coefficient which measures the correlation between two datasets. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

It's basically a covariance of the two variables divided by the product of their standard deviations.

It's is named after Karl Pearson who formulated this correlation coefficient from a related idea by Francis Galton in the 1880s.
According to him, using this coefficient we can calculate a linear relationship between the two given variables. But, Under certain criteria's, such as:

1. Scale of measurement should be interval or ratio
2. Variables should be approximately normally distributed
3. The association should be linear
4. There should be no outliers in the data

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

Ans: Scaling simply means that you're transforming your data so that it fits within a specific desired scale. Usually, the scaling is done in the range of 0 to 1 in industries. It is a step of data Pre-Processing which is usually applied to the independent variables to normalize the data within a particular range which helps in speeding up the calculations in an algorithm.

Usually, many a times the collected dataset contains features which are highly varying in terms of magnitudes, units and range. If scaling is not done then the algorithm only takes the magnitude in account and not units hence resulting in an incorrect modelling. Hence, to solve this issue, we have to do the scaling to bring all the variables on a same level of magnitude.

Also, scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

<u>Normalized Scaling:</u>

It brings all of the data in the range of 0 and 1.

sklearn.preprocessing.MinMaxScaler is the library that helps to implement normalization in python.

The formula for this is: $$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Here, Xmax and Xmin are the maximum and the minimum values of the feature respectively.

5. When the value of X is the minimum value in the column, the numerator will be 0, and hence X' is 0
6. On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X' is 1
7. If the value of X is between the minimum and the maximum value, then the value of X' is between 0 and 1

Standardized Scaling

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

The formula for standardized scaling is: $$X' = \frac{X - \mu}{\sigma}$$

Differences:

1. Minimum and maximum value of features are used for scaling in Normalized whereas mean and standard deviation is used for scaling in standardized scaling.
2. Scales values lies between [0, 1] or [-1, 1] in Normalized Scaling but there is no certain range in terms of Standardized Scaling.
3. Normalized Scaling is really affected by outliers but this isn't a case with the Standardized Scaling.

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

Ans: An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

The value of VIF is calculated by the following formula: $$VIF_i = \frac{1}{1-R_i^2}$$

Where, i refers to the ith variable.

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

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

Ans: A Q-Q plot or the quantile-quantile plot is a graphical technique which is used for determining that, if two data sets come from the population with a common distribution.

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.

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.

The q-q plot is used to know:

1. Whether data sets come from populations with a common distribution.
2. To know that if data sets have common location and scale.
3. To know that if two data sets have similar distributional shapes.
4. To know that if two data sets have similar tail behavior or not.