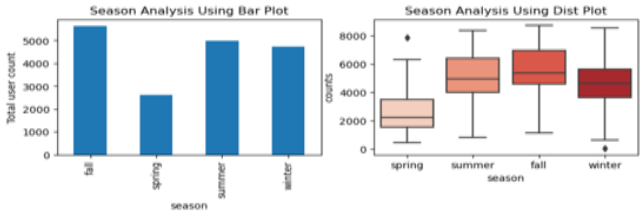
Assignment Questions

I. 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?

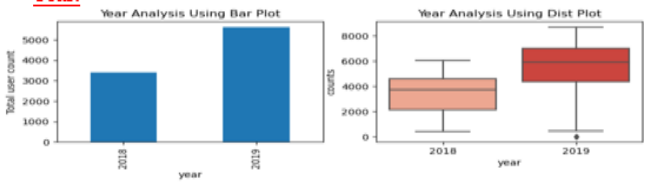
categorical_vars_list = ['season', 'year', 'month', 'holiday', 'weekday', 'weathersit'] All are the categorical variable list of our dataset. A **box plot** and **dist plot** is used to analyze the effect on the dependent variable.

Season:



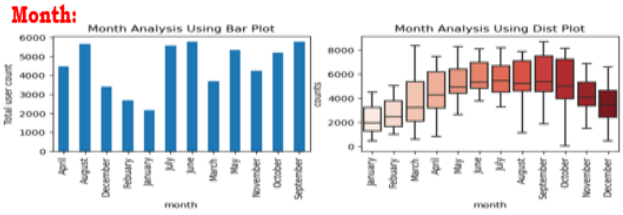
INSIGHTS --
Maximum people renting bikes in **Fall Season** fallowed by **summer** then **winter** and **spring**.

Year:



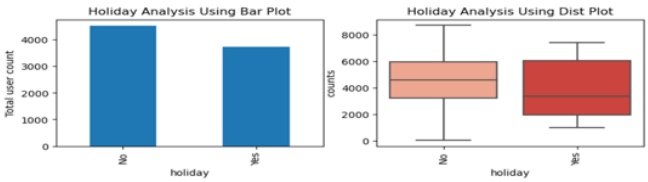
INSIGHTS --
Maximum people renting bikes in the **Year 2019** as compared to **2018**

Month:



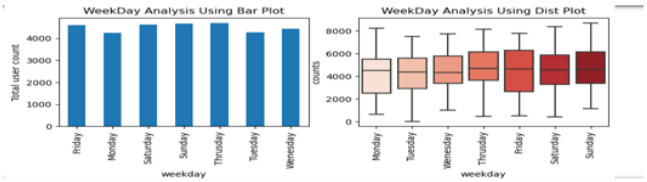
INSIGHTS --
Maximum people renting bikes in the month of **June ,July, August , September** . Very less people rents bikes in the month of **January , Febuary ,December**

Holiday:



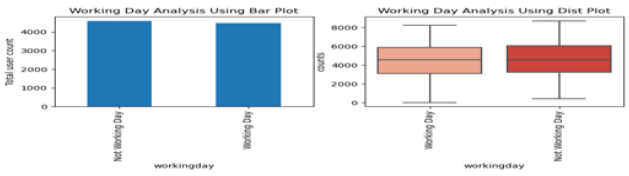
INSIGHTS --
Less people rents bikes on **Holidays**.

Weekday:



INSIGHTS --
Weekdays does not affect the renting bikes **Counts**.As all are somewhat equal in scale

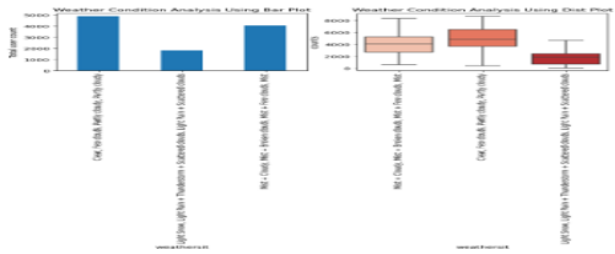
Working day:



INSIGHTS –

Working Day does not affect the renting bikes **Counts**. As all are somewhat equal in scale

Weather Conditions:



INSIGHTS –

People likes to rent bikes when the weather condition is **Clear, Few clouds, Partly cloudy.** Also when the weather is **Mist, Cloudy, Broken clouds, or Few clouds** . people likes to rent bikes.

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

Dummy Variable:

Some data science tools will only work when the input data are numeric. This particularly true of machine learning. Many **machine learning** algorithms – like **linear regression and logistic regression** – strictly require numeric input data. If you try to use them with string-based categorical data, they will throw an error.

So before you use such tools, you need to encode your categorical data as numeric dummy variables.

drop_first=True:

The drop_first parameter specifies whether or not you want to drop the first category of the categorical variable you're encoding.

By default, this is set to **drop_first = False**. This will cause get_dummies to create one dummy variable for every level of the input categorical variable.

If you set **drop_first = True**, then it will drop the first category. So if you have K categories, it will only produce K – 1 dummy variables.

Example:

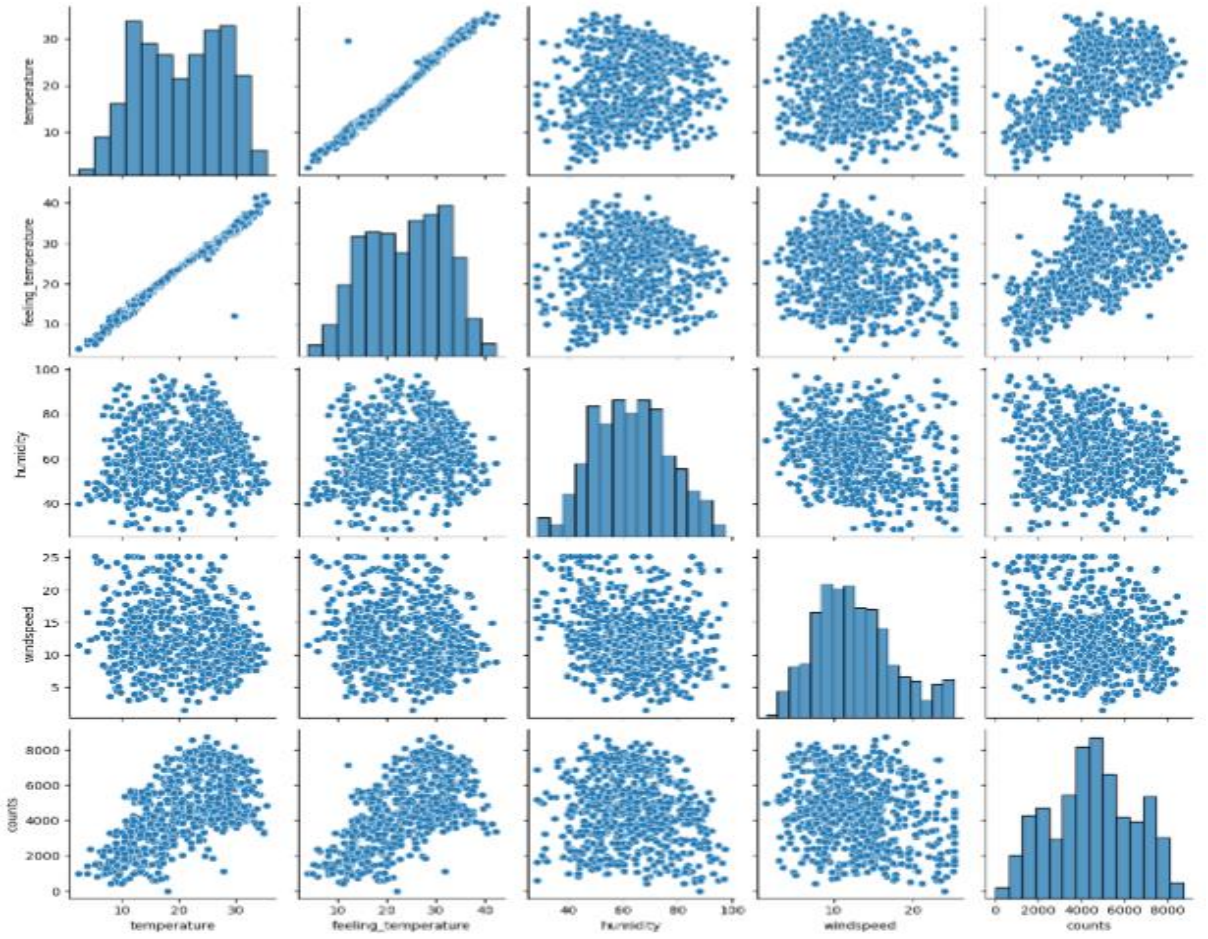
Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi furnished, then It is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished.

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

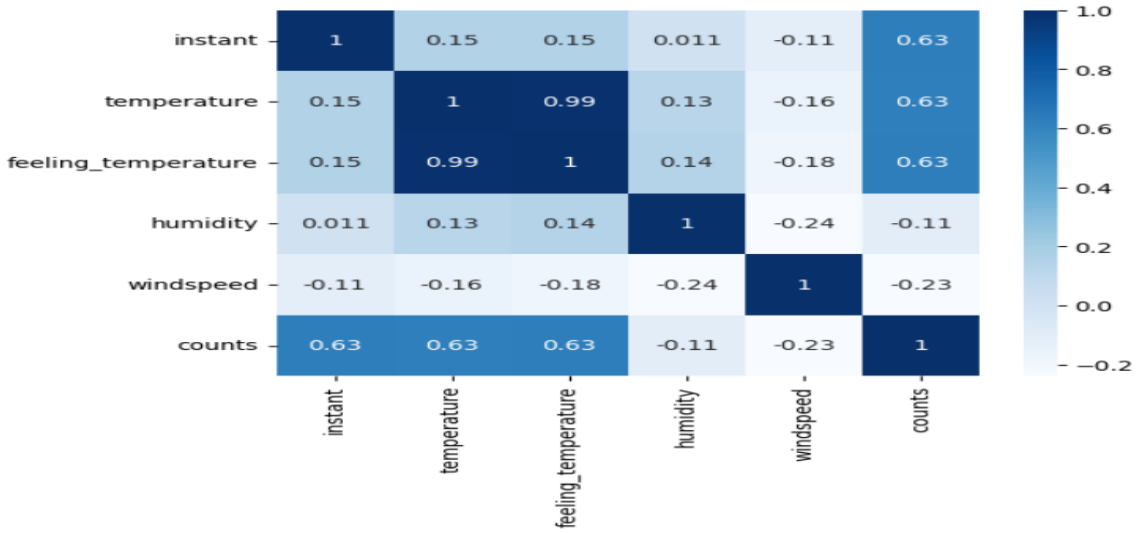
Pair-Plot for Numerical Variables:

```
In [61]: plt.figure(figsize=(20,8))
         sns.pairplot(daysData[numeric_vars_list])
         plt.show()
```

```
<Figure size 2000x800 with 0 Axes>
```



Heatmap corrs() for numerical variable:



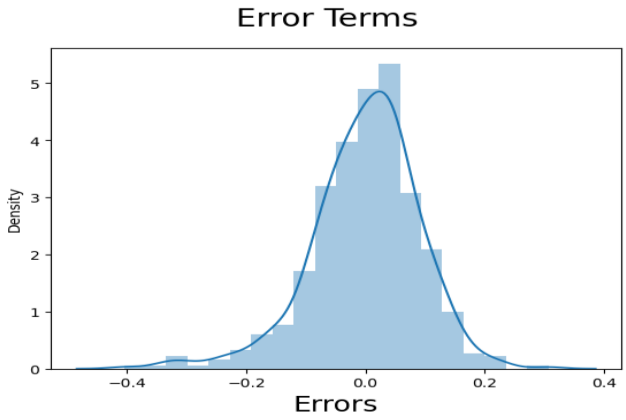
INSIGHTS –

We can see **temperature** (i,e temp) and **feeling temperature** (i,e atemp) are highly corelated with our Target variable '**counts** (i,e 'cnt').

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

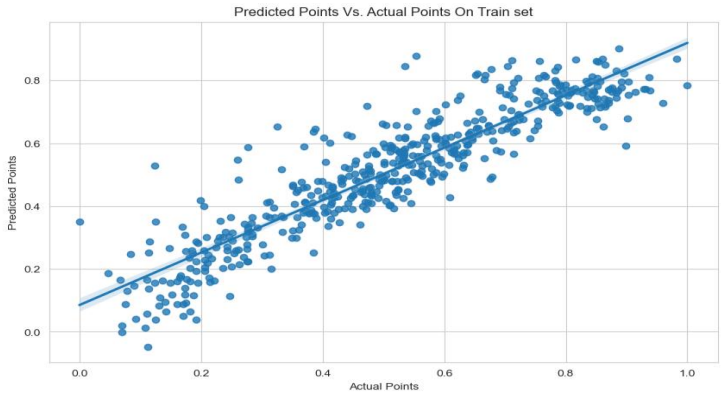
To validate our assumptions of Linear Regression , we do Residual Analysis of the train data and check various assumptions of Linear Regression.

Analysing for the Assumption: Normality



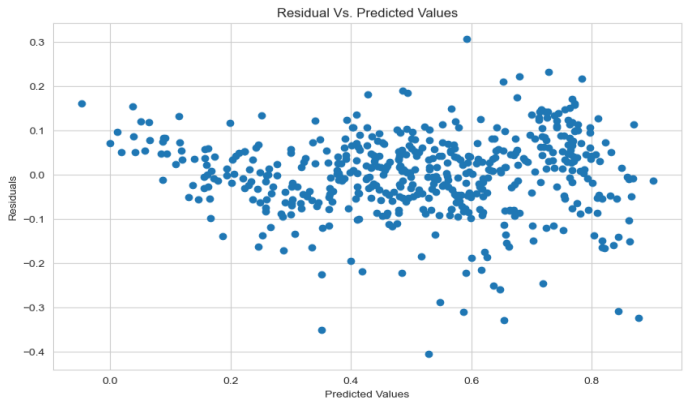
We can see there is no specific Pattern observed in the Error Terms with respect to Prediction, hence we can say **Error terms are independent of each other**.

Analyzing for Constant variance



We can see Error Terms have approximately a Constant Variance, Hence it follows the **Assumption of Homoscedasticity**

Analyzing patterns in residuals



We can see there is no specific Pattern observed in the Error Terms with respect to Prediction, hence we can say **Error terms are independent of each other.**

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 explaining the demand of the shared bikes are :

1.temperature : with the co-efficient of **'0.518857'** , a unit increases in the temperature variable **increase** the number of bike rentals by **0.518857** unit

2. year : with the co-efficient of **' 0.233548'** , a unit increase in the year variable **increases** the number of bike rentals by **0.233548** units per year.

3. summer season : with the co-efficient of **' 0.081277'** , a unit increase in particularly Summer season **increases** the number of bike rentals by **0.081277** units.

Some other features:

4. windspeed : with the co-efficient of **'-0.117426'** a unit increase in windspeed, **reduces** the bike rentals by **'+ 0.117426'** units

5. workingday: with the co-efficient of ' **-0.021774'** , a unit increase in workingday, **reduces** the bike rentals by **'+ 0.021774'** units.

Prediction Based on above data:

1.**Temperature** is the Most Significant Feature which affects the Business positively.
2.Whereas the other Environmental condition such as **mist**, **Windspeed** and **light_scattered_clouds** affects the Business negatively.
3.Company need to think how to tackle with bad weather condition and come up with sort of solution which protect users from Raining.
4.The Demand of Bikes is more in the **Winter** and **Summer** season
5.Demand of Bike Rent has been significantly increased in the **2019** than **2018** so it is clear sign that Boom Bikes is doing a Business.

II. General Subjective Questions

1. Explain the linear regression algorithm in detail.

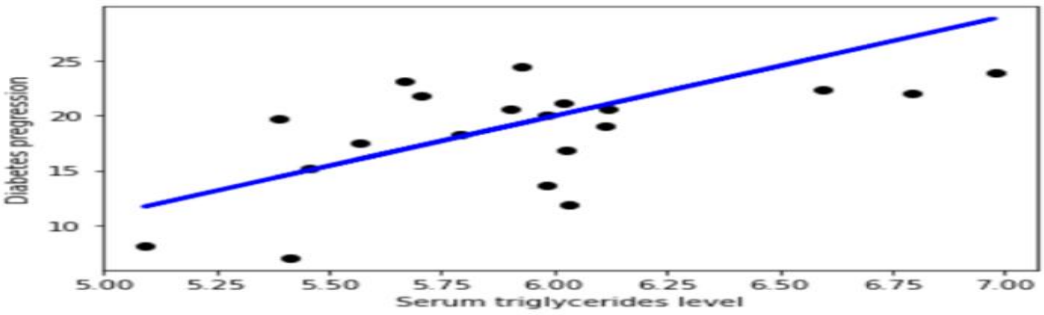
Linear regression:

Linear regression is a supervised machine learning method that is used by the **Train Using AutoML** tool and finds a linear equation that best describes the correlation of the explanatory variables with the dependent variable. This is achieved by fitting a line to the data using least squares. The line tries to minimize the sum of the squares of the residuals. The residual is the distance between the line and the actual value of the explanatory variable. Finding the line of best fit is an iterative process.

The following is an example of a resulting linear regression equation:

$$y = b_0 + b_1x_1 + b_2x_2 + ...$$

Here, y is the dependent variable, and x_1, x_2, and so on, are the explanatory variables. The coefficients (b_1, b_2, and so on) explain the correlation of the explanatory variables with the dependent variable. The sign of the coefficients (+/-) designates whether the variable is positively or negatively correlated. b_0 is the intercept that indicates the value of the dependent variable assuming all explanatory variables are 0. In the following image, a linear regression model is described by the regression line y = 153.21 + 900.39x. The model describes the relationship between the dependent variable, Diabetes progression, and the explanatory variable, Serum triglycerides level. A positive correlation is shown. This example demonstrates a linear regression model with two variables. Although it is not possible to visualize models with more than three variables, practically, a model can have any number of variables.



A linear regression model helps in predicting the value of a dependent variable, and it can also help explain how accurate the prediction is. This is denoted by the R-squared and p-value values. The R-squared value indicates how much of the variation in the dependent variable can be explained by the explanatory variable and the p-value explains how reliable that explanation is. The R-squared values range between 0 and 1. A value of 0.8 means that the explanatory variable can explain 80 percent of the variation in the observed values of the dependent variable. A value of 1 means that a perfect prediction can be made, which is rare in practice. A value of 0 means the explanatory variable doesn't help at all in predicting the dependent variable. Using a p-value, you can test whether the explanatory variable's effect on the dependent variable is significantly different from 0.

Assumption for Linear Regression Model

Linear regression is a powerful tool for understanding and predicting the behavior of a variable, however, it needs to meet a few conditions in order to be accurate and dependable solutions.

1. **Linearity:** The independent and dependent variables have a linear relationship with one another. This implies that changes in the dependent variable follow those in the independent variable(s) in a linear fashion.
2. **Independence**: The observations in the dataset are independent of each other. This means that the value of the dependent variable for one observation does not depend on the value of the dependent variable for another observation.
3. **Homoscedasticity**: Across all levels of the independent variable(s), the variance of the errors is constant. This indicates that the amount of the independent variable(s) has no impact on the variance of the errors.

4. **Normality:** The errors in the model are normally distributed.
5. **No multicollinearity**: There is no high correlation between the independent variables. This indicates that there is little or no correlation between the independent variables.

Types of Linear Regression Models:

There are two kinds of Linear Regression Model:-

- **Simple Linear Regression:** A linear regression model with one independent and one dependent variable. The formula for Simple linear regression is:

  $y = \theta x + b$

  where,

  θ – It is the model weights or parameters

  b – It is known as the bias.

- **Multiple Linear Regression:** A linear regression model with more than one independent variable and one dependent variable. The formula for Multiple linear regression is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon$$

where, for i=n observations:

yi=dependent variable

i=explanatory variables

β0=y intercept (constant term)

βp=slope coefficients for each explanatory variable

ϵ=the model's error term (also known as the residuals)

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet:

Is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

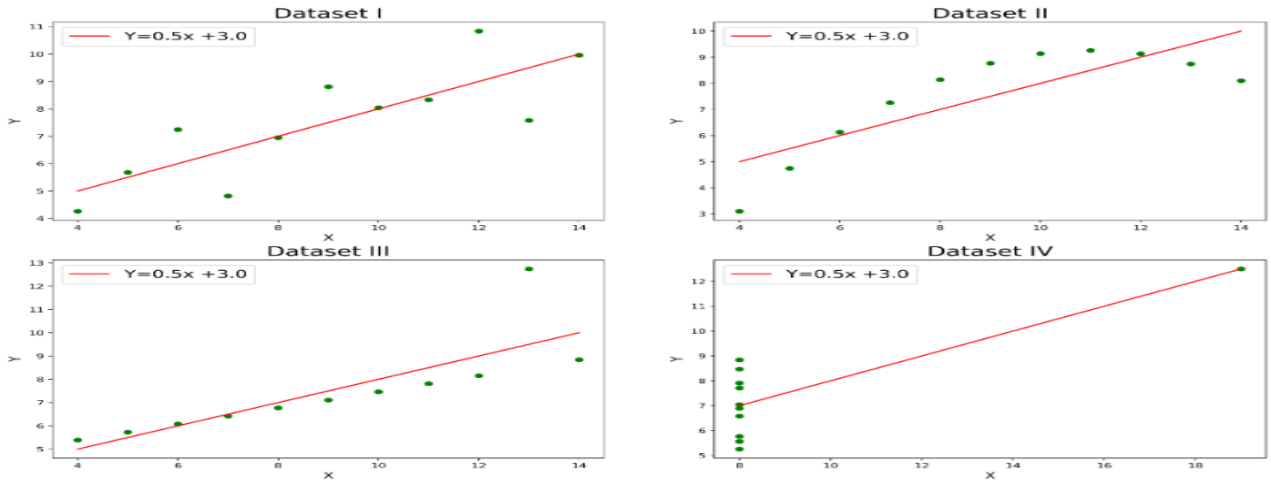
The four datasets of Anscombe's quartet.

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

Application:

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

Plotting The four datasets of Anscombe's quartet.



Anscombe's quartet Plot

Explanation of this plot:

- In the **first one(top left)** if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the **second one(top right)** if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the **third one(bottom left)** you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, **the fourth one(bottom right)** shows an example when one high-leverage point is enough to produce a high correlation coefficient.

3. What is Pearson's R?

Pearson Coefficient

The Pearson coefficient is a type of **correlation coefficient** that represents the relationship between two variables that are measured on the same interval or ratio scale. The Pearson coefficient is a measure of the strength of the association between two continuous variables.

Understanding the Pearson Coefficient

To find the Pearson coefficient, also referred to as the Pearson correlation coefficient or the Pearson product-moment correlation coefficient, the two variables are placed on a scatter plot. The variables are denoted as X and Y. There must be some linearity for the coefficient to be calculated; a scatter plot not depicting any resemblance to a linear relationship will be useless. The closer the resemblance to a straight line of the scatter plot, the higher the strength of association. Numerically, the Pearson coefficient is represented the same way as a correlation coefficient that is used in linear regression, ranging from -1 to +1. A value of +1 is the result of a perfect positive relationship between two or more variables.

- **Positive correlations** indicate that both variables move in the same direction. Conversely, a value of -1 represents a perfect negative relationship.
- **Negative correlations** indicate that as one variable increases, the other decreases; they are inversely related.
- **A zero** indicates no correlation.

Benefits of the Pearson Coefficient It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

Limitations of the Pearson Coefficient A key limitation of Pearson's r is that it cannot distinguish between independent and dependent variables. Therefore, also if a relationship between two variables is found, Pearson's r does not indicate which variable was 'the cause' and which was 'the effect'.

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

Scaling:

Feature scaling is one of the most important data preprocessing step in machine learning.

Why is scaling performed?

Algorithms that compute the distance between the features are biased towards numerically larger values if the data is not scaled. Hence we have to do scaling .

Tree-based algorithms are fairly insensitive to the scale of the features. Also, feature scaling helps machine learning, and deep learning algorithms train and converge faster. There are some feature scaling techniques such as Normalization and Standardization that are the most popular and at the same time, the most confusing ones.

a) Normalization or Min-Max Scaling is used to transform features to be on a similar scale. The new point is calculated as:

$$X\_new = (X - X\_min)/(X\_max - X\_min)$$

This scales the range to [0, 1] or sometimes [-1, 1].

b) Standardization or Z-Score Normalization is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score.

$$X\_new = (X - mean)/Std$$

Difference Between Normalization and Standardization:

S.NO.	Normalization	Standardization
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called `MinMaxScaler` for Normalization.	Scikit-Learn provides a transformer called `StandardScaler` for standardization.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

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

VIF :

VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables. For example, we would fit the following models to estimate the coefficient of determination R1 and use this value to estimate the

$$\text{VIF:} \quad X\_1=C+ \alpha\_2 X\_2+\alpha\_3 X\_3+\cdots$$

$$〖\text{VIF}〗\_1=1/(1-R\_1^2)$$

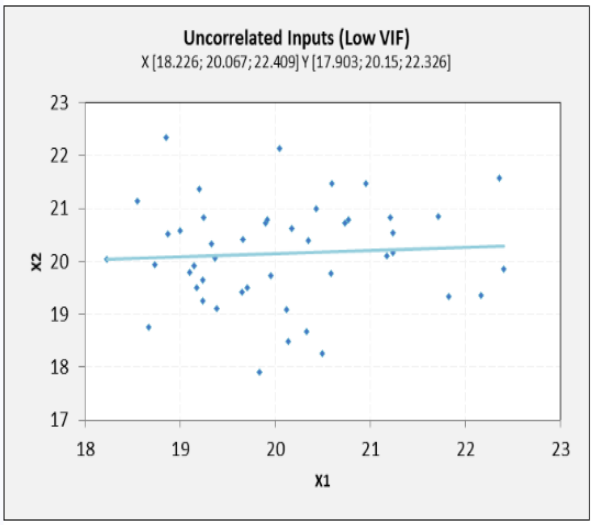
Next, we fit the model between X2 and the other independent variables to estimate the coefficient of determination R2:

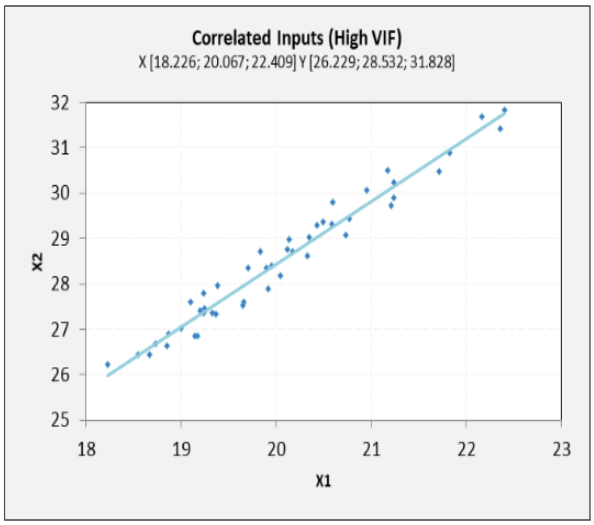
$$X\_2=C+ \alpha\_1 X\_1+\alpha\_3 X\_3+\cdots$$

$$〖\text{VIF}〗\_2=1/(1-R\_2^2)$$

If all the independent variables are orthogonal to each other, then VIF = 1.0. **If there is perfect correlation, then VIF = infinity**. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. This would mean that that standard error of this coefficient is inflated by a factor of 2 (square root of variance is the standard deviation). The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity. A general rule of thumb is that if **VIF > 10 then there is multicollinearity.** Note that this is a rough rule of thumb, In some cases we might choose to live with high VIF values if it does not affect our model results such as when we are fitting a quadratic or cubic model or depending on the sample size a large value of VIF may not necessarily indicate a poor model.

VIF	Conclusion
1	No multicollinearity
4 - 5	Moderate
10 or greater	Severe





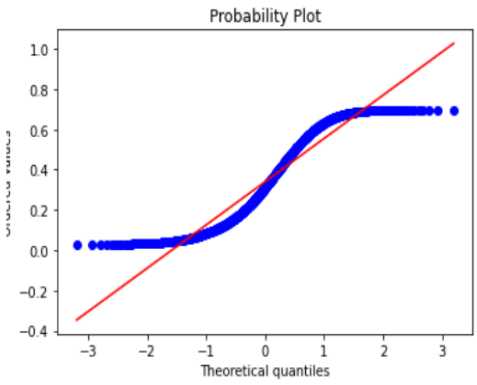
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Q-Q plot:

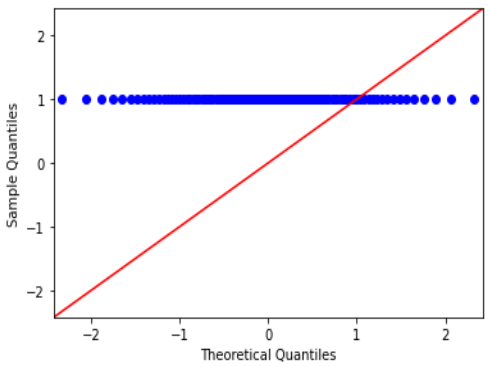
The quantile-quantile plot is a graphical method for determining whether two samples of data came from the same population or not. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

Types of Q-Q plots:

1. For Left-tailed distribution:



2. For Uniform distribution:



Usage of Q-Q plot:

- Determine whether two samples are from the same population.
- Whether two samples have the same tail
- Whether two samples have the same distribution shape.

- Whether two samples have common location behavior.

Importance of Q-Q plot:

- Since Q-Q plot is like probability plot. So, while comparing two datasets the sample size need not to be equal.
- Since we need to normalize the dataset, so we don't need to care about the dimensions of values.
- It is a tool using which we analyze the different shapes of distribution. Generally we use scatter –plot to study the Q-Q plot distribution.