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 variables?

Ans – Below are the categorical variables in the dataset and there effect on the dependant variable –

1. **Season** – Summer and winter season affect the demand of bike i.e the dependant variable 'cnt'

2. **WeatherSit** – weather sit condition of light snow & Rain and Mist and cloudy are inversely related to 'Cnt' variable ie during these weather condition demand of bike decreases

3. **Yr** – The demand of bike increased with passing year which is indicated by high positive coefficient of yr in regression coefficient

4. **Holiday** – bike rentals are reduced during holiday , this is indicated by a negative coefficient of holiday in the regression equation

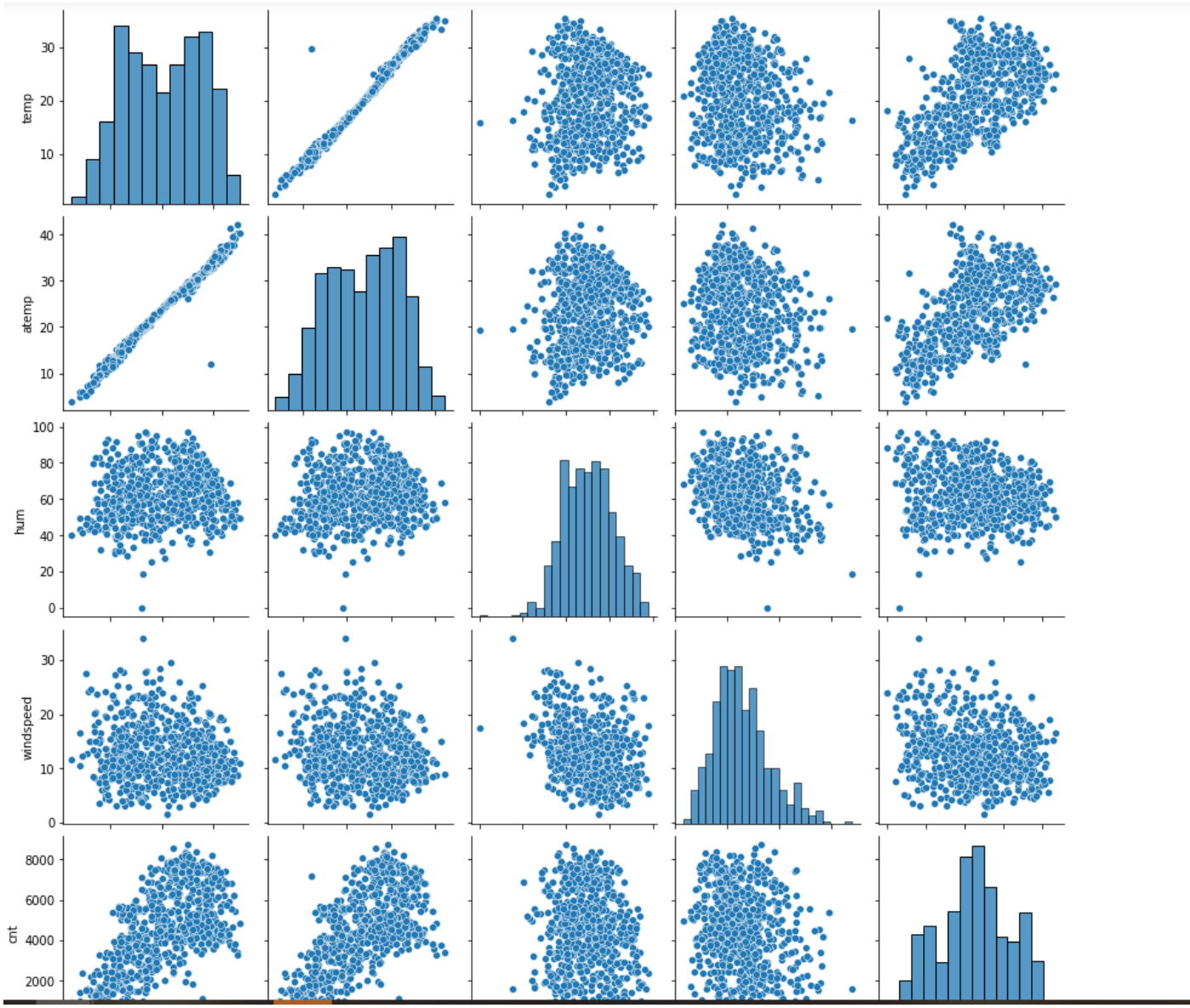
5. **Mnth** – bike rental for the month of aug and sept are positively correlated with the dependant variable

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

Ans- dummy variables will be correlated if we don't do drop_first = True. What drop_first =true does is it drops the first column of the categorical on hot encoding so that the encoded value are not redundant and prevents dummy variables from being correlated .

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

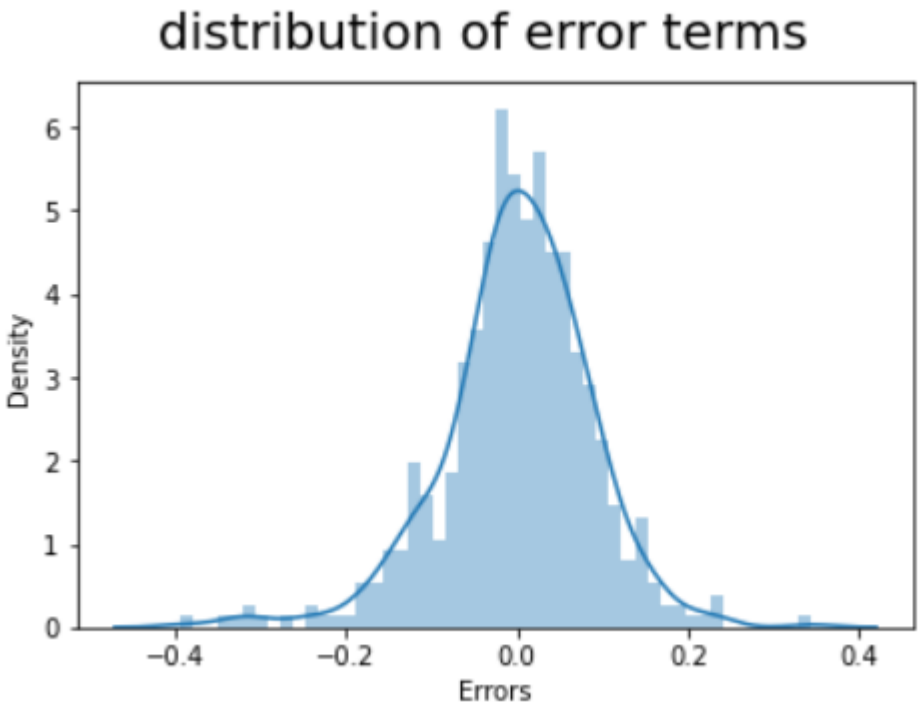
Ans – temp and a temp are highly correlated with the target variable both having correlation coefficient of 0.63, infact these two variables are highly correlated to each other as well. Below is the pair plot and heat map for reference



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

Ans – There are basically 4 assumptions of linear regression, they are explained and validated below

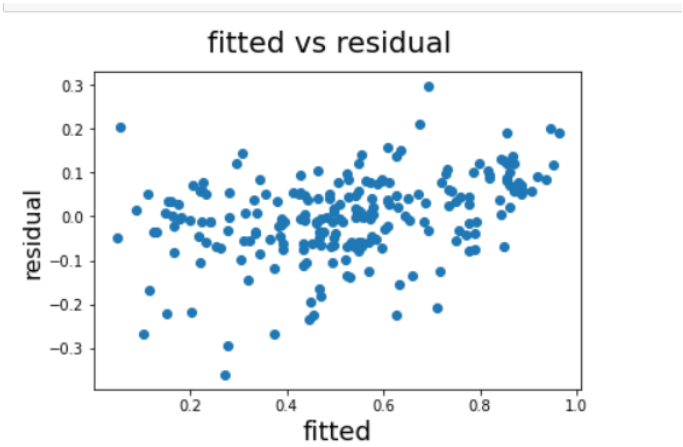
1. One of the major assumption of linear regression is error terms are distributed normally centred around 0 ie mean 0.This assumption is validated by performing Residual analysis and plotting a distplot of error terms –



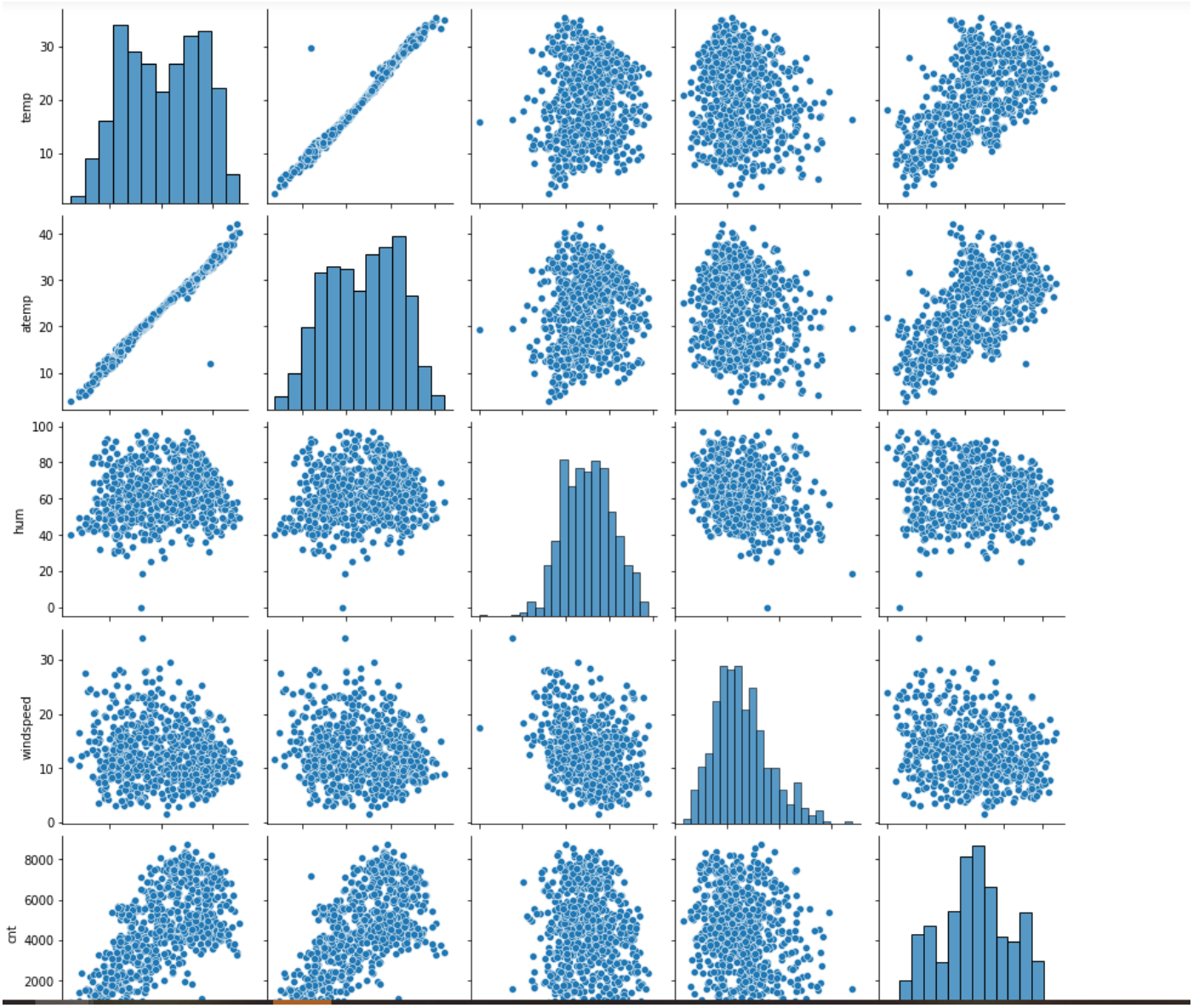
2 .3To check for multicollinearity we check the VIF values of the final independent variables , all VIF values are lower than expect the constant –

	independent variables	VIF
0	const	45.193002
11	hum	1.869481
7	weathersit_Mist & Cloudy	1.566869
10	temp	1.539063
3	mnth_aug	1.443623
1	season_summer	1.373170
2	season_winter	1.318758
6	weathersit_Light Snow & Rain	1.240303
4	mnth_sept	1.213076
12	windspeed	1.181424
8	yr	1.027273
9	holiday	1.021329
5	weekday_tuesday	1.019193

3 .To check for homoscedasticity we plot a graph between fitted value and residual value. Absence of a pattern determines absence of homoscedasticity. Homoscedasticity basically means error terms have constant variance



4 . Last assumption is there should be linear relationship between X and Y. From the below plot we see a linear relationship exist between cnt and temp,atemp



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

Ans – the regression equation came out to be –

*cnt = 0.233*const +*

*0.102*season_summer +*

*0.149*season_winter +*

*0.053*mnth_aug +*

*0.119*mnth_sept -*

*0.045*weekday_tuesday -*

0.243 weathersit_Light Snow & Rain -*

*0.058*weathersit_Mist & Cloudy +*

*0.229*yr -*

*0.103*holiday +*

*0.539*temp -*

*0.166*hum -*

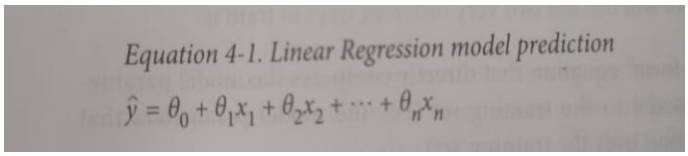
*0.182*windspeed '*

So the top 3 features are –

- **Temp** – positive coefficient of 0.539
- **weathersit_Light** Snow & Rain- negative coefficient of 0.243
- **yr** – positive coefficient of 0.229

General Subjective Questions

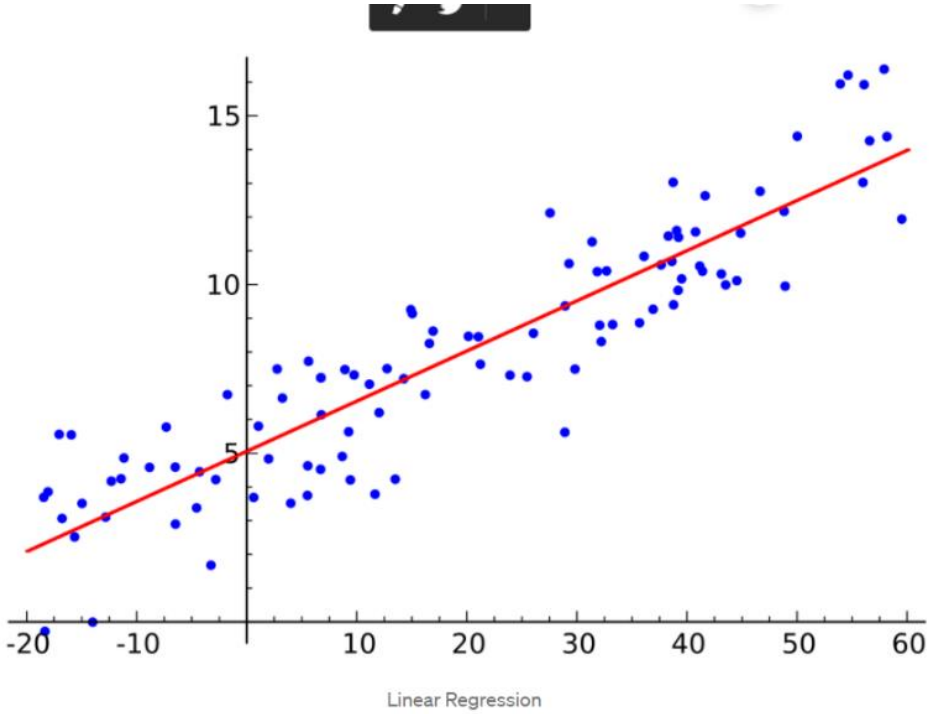
1. **Explain the linear regression algorithm in detail.**
   Ans – Linear regression is the most basic form of regression analysis**.** Regression is a method of modelling a target value on independent predictors. Linear regression model makes prediction by simply computing a weighted sum of the input features, plus a constant term often called bias term or intercept.
   Equation of linear regression is shown below-



   Here In this equation –

   - Y^ is the predicted variable
   - n is the number of features
   - xi is the ith feeture value

Graphically linear regression can be visualized by below figure ,it is a simple linear regression between a dependant an independent variable–



There are four assumptions associated with a linear regression model:

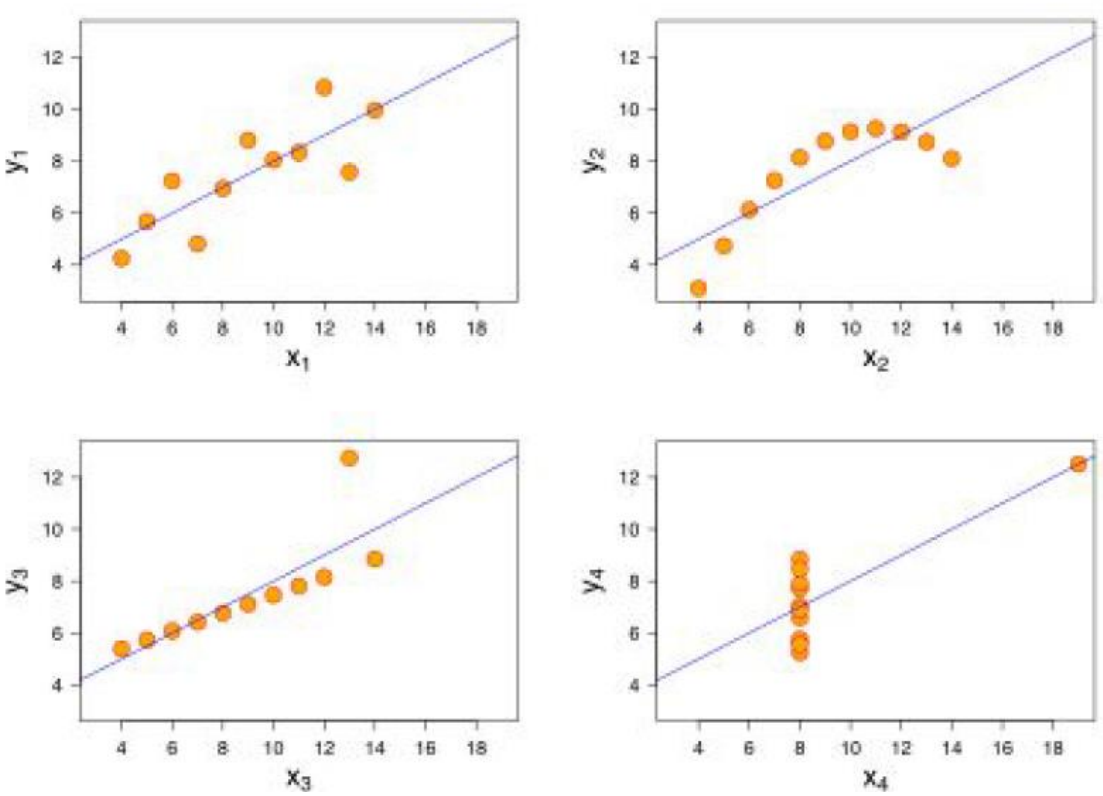
1. **Linearity**: The relationship between X and the mean of Y is linear.
2. **Homoscedasticity**: The variance of residual is the same for any value of X.
3. **Independence**: Observations are independent of each other.
4. **Normality**: For any fixed value of X, Y is normally distributed.

Types of linear Regression

1. **Multiple Linear Regression**
   This is basically Linear Regression with Multiple Variables. In this we always try to find the linear relationship between two or more independent variables or inputs and the corresponding dependent variable or output in this sort of linear regression, and the independent variables can be either continuous or categorical.
2. **Simple linear Regression** – We want to know the relationship between a single independent variable, or input, and a matching dependent variable, or output, in simple linear regression. We can express this as y = 0 + 1x+ in a simple line.

2. Explain Anscombe's quartet in detail.

Ans- 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 analysing 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, it's 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.

3.What is Pearson's R?

Ans - Correlation measures the strength of association between two variables as well as the direction. There are mainly **three types** of correlation that are measured. One significant type is Pearson's correlation coefficient. This type of correlation is used to measure the relationship between two continuous variables. In Statistics, the Pearson's Correlation Coefficient is also referred to as **Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation**. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0. Whenever we discuss correlation in statistics, it is generally Pearson's correlation coefficient. However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables. **Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations**. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name. Pearson's Correlation Coefficient is named after Karl Pearson. He formulated the correlation coefficient from a related idea by Francis Galton in the 1880s.

In simple terms, Pearson's R tells us can we draw a line graph to represent the data?

- 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

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

Ans - Feature scaling refers to normalizing or standardizing the range of independent variables or features of dataset to bring them down to common scale. It is performed during the data preprocessing stage to deal with varying values in the dataset.

Scaling is performed as machine learning algorithms do not converge easily if the independent variables are on different scale or range

Two methods for feature scaling are –Normalization and Standardization

Normalization

- Also known as Min_max_scaling**.**
- Values are shifted and rescaled so that they end up ranging from 0 to 1
- Done by subtracting the min value and dividing by the max minus the min
- More affected by outliers

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Standardization

- Values are not bound between 0 to 1
- Done by subtracting the mean value then dividing by standard deviation.
- Much less affected by outliers

$$Z = \frac{x - \mu}{\sigma}$$

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

Ans-

VIF - **the variance inflation factor** -The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity.

(VIF) =1/(1-R_1^2).

If there is perfect correlation, then VIF = infinity.

Where R_1 is the R-square value of that variable which we want to check how well this independent variable is explained by other independent variables.

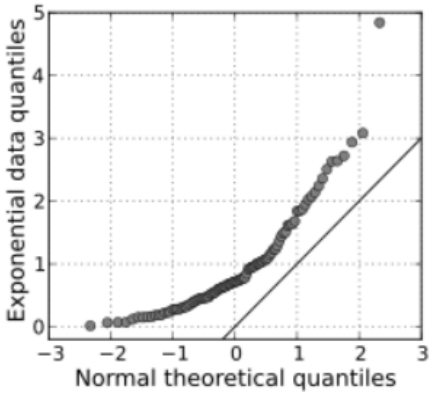
If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it's R-squared value will be equal to 1.

So, VIF = 1/(1-1) which gives VIF = 1/0 which results in infinity

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

Ans - Q-Q Plots (Quantile-Quantile plots) 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. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q Q plot showing the 45 degree reference line:



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 behavior?*