Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer abouttheir effect on the dependent variable? (3 marks)**

Season, month, weathersit, weekday, holiday, year, working day were plotted on a boxplot to see their effects on the dependent variable cnt. Below are the insights:

1. Fall (Autumn) season has the highest demand for rental bikes. It could be due to the fact that fall is considered to be as transition from summer to winter during which temperature is neither too high nor too low. The weather is pleasant. So bike rentals are more in this season. We can say that weathersit can be a good predictor for the dependent variable.

2. The demand for rental bike has decreased on holidays. More bike rental on non-holidays.

3. Bike rental demand has gone up from 2019 to 2019.

4. The rentals on weekdays i.e. days of the week are almost same. We can't see any obvious trend. It can effect or not effect the dependent variable. We will se later at the model building stages.

5. When the weather is clear, the demand for the bike is highest which is obvious. In bad weather the demand is lowest. Weather sit can be a good predictor for the dependent variable.

6. Demand is continuously growing till June. September month has the highest demand for the bikes. After September, demand is decreasing.

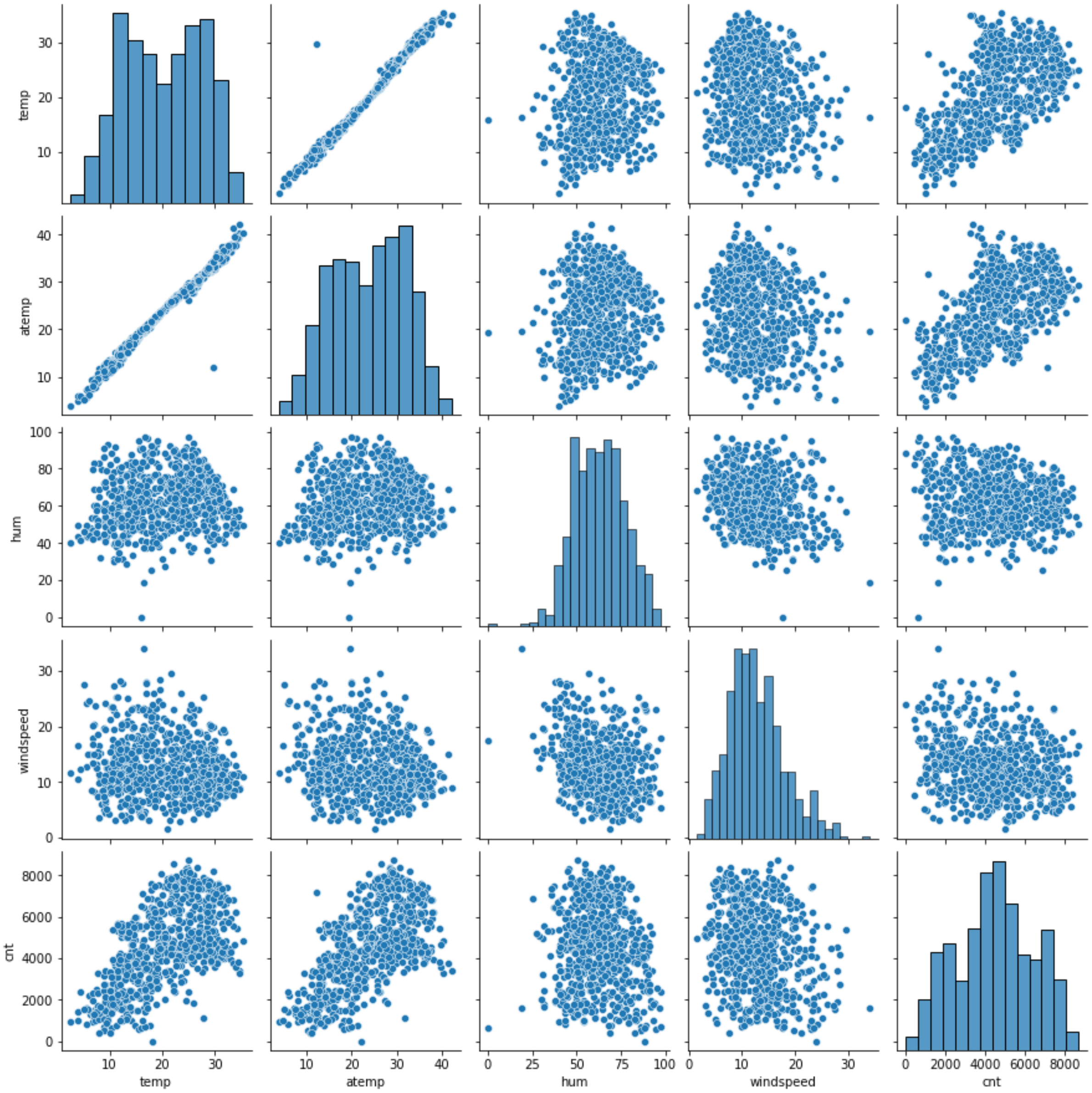
7. Bike sharing is more during September. It is less at beginning and at the end of the year. It could be due to cold/extreme weather conditions. So we can say that month is a good predictor of dependent variable.

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

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.

For 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. Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables. Not dropping the column may affect some models adversely and the effect is stronger when the cardinality is smaller. For example iterative models may have trouble converging and lists of variable importance may be distorted. Another reason is, if we have all dummy variables it leads to Multicollinearity between the dummy variables. To keep this under control, we drop one column.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlationwith the target variable? (1 mark)**

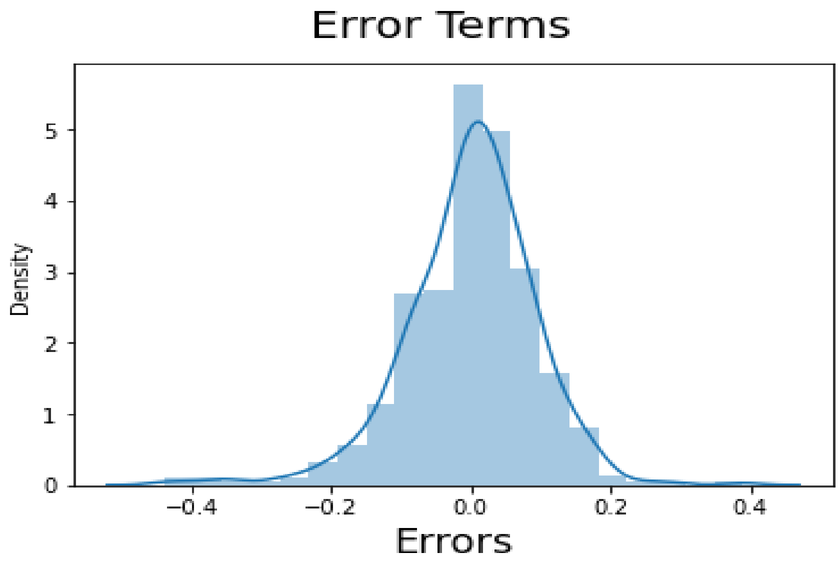


As we can see from the above pairplot, temp and atemp are the two numerical variables which have the highest correlation with the target variable cnt.

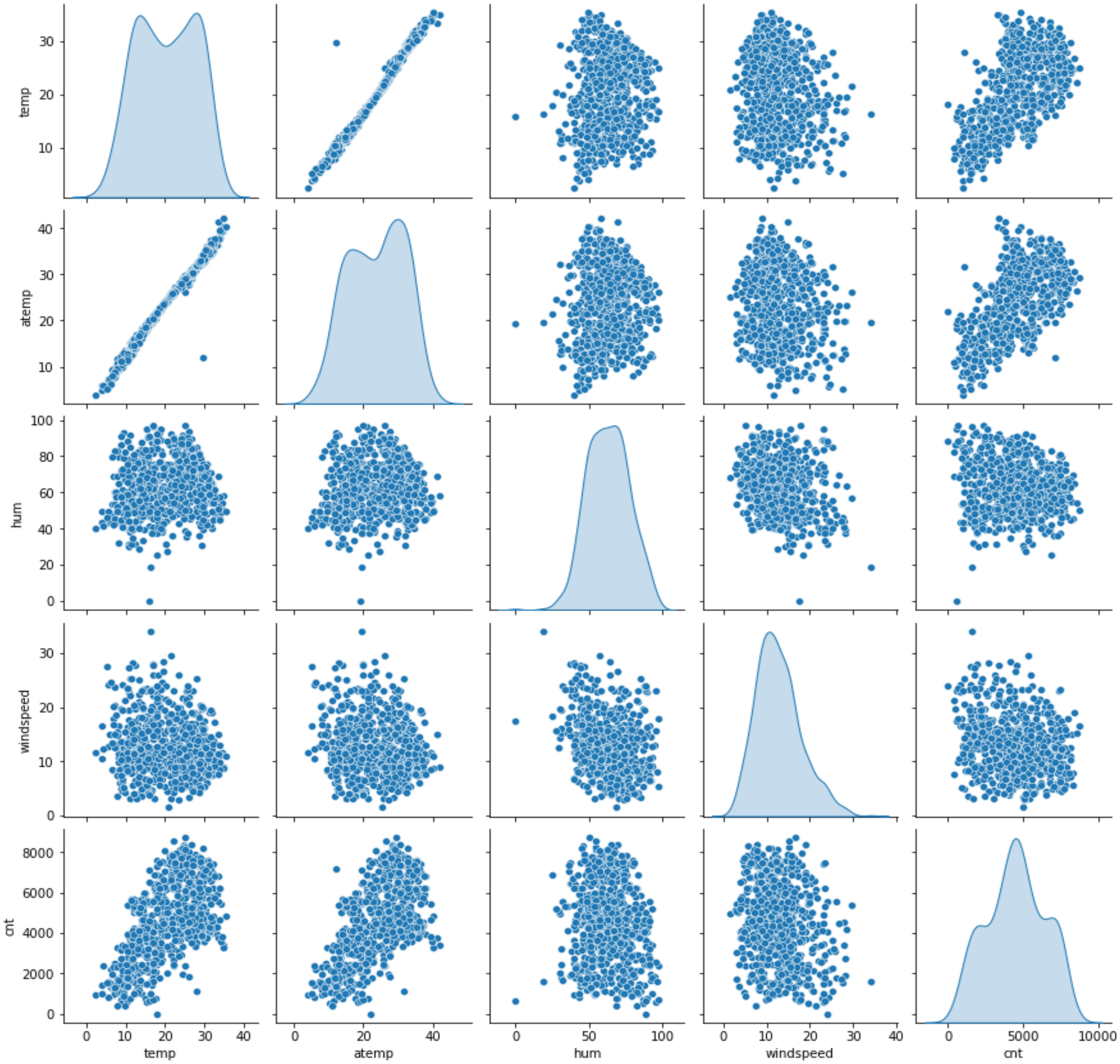
4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

   1. *Error terms are normally distributed with mean zero (not X, Y)*

      We validate this assumption about residuals by plotting a distplot of residuals and see if residuals are following normal distribution or not. Below graph shows that the residuals are distributed about mean 0.

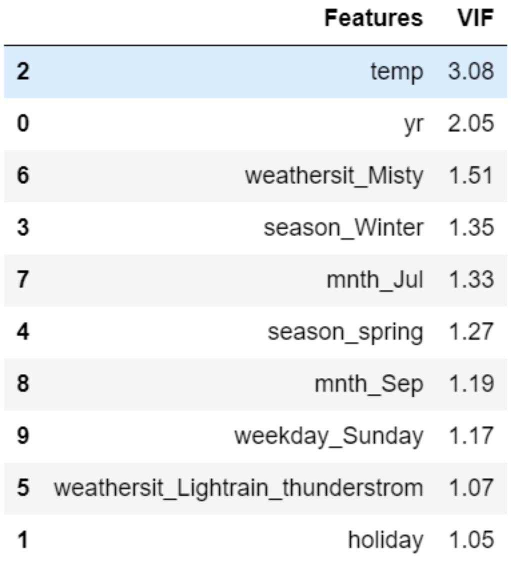


2. *linear relationship between X and Y*



We validate this assumption by plotting this pair plot. we can say that there is a linear relation between temp and atemp variable with the predictor variable ‘cnt’.

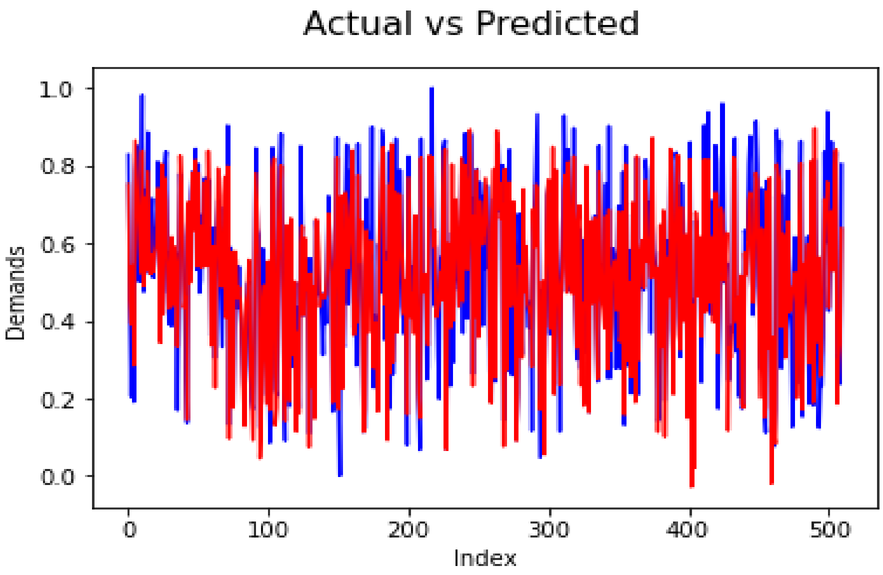
3. *No Multicollinearity between the predictor variables*

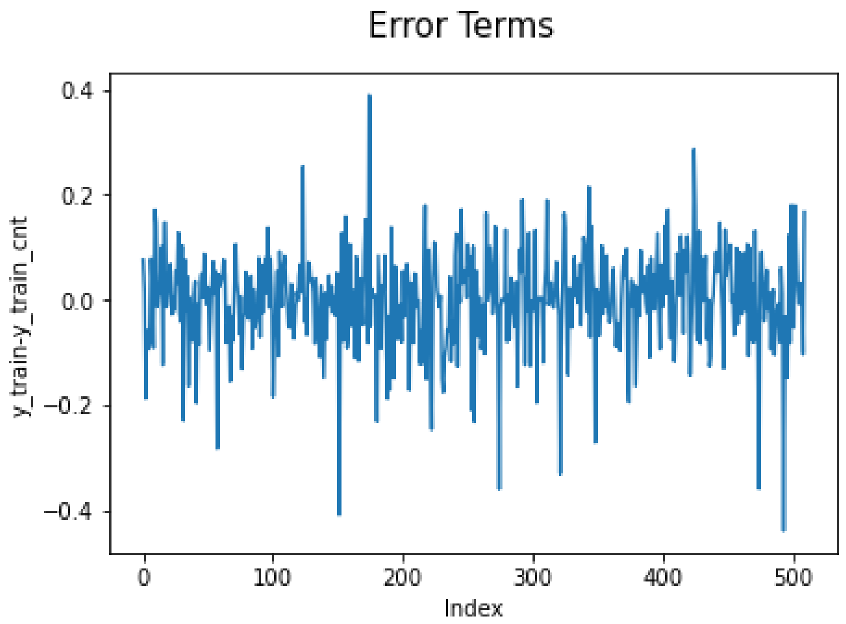


	Features	VIF
2	temp	3.08
0	yr	2.05
6	weathersit_Misty	1.51
3	season_Winter	1.35
7	mnth_Jul	1.33
4	season_spring	1.27
8	mnth_Sep	1.19
9	weekday_Sunday	1.17
5	weathersit_Lightrain_thunderstrom	1.07
1	holiday	1.05

From the VIF calculation we can say that there is no multicollinearity between the predictor variables, as all the values are within the allowed range of <5.

4. *Homoscedasticity*





Actual and Predicted result following almost the same pattern so this model seems ok. Also we see the error terms are independent of each other. So the Homoscedasticity assumption holds true.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

Tempearture (Temp), weathersit_Lightrain_thunderstrom (Season 3), year (yr) are top 3 predictor of the dependent variable cnt. These variables influence the bike rentals the most.

1. Temperature with coefficient 0. 468240 indicates that temperature has positive effect on bike rental.
2. weathersit_Lightrain_thunderstrom with coefficient -0. 304531 indicates that the light snow and rain has negative impact on people from renting out the bikes.
3. Year with coefficient 0. 233100 indicates that bike rental will go up/increase in the subsequent year.

General Subjective Questions

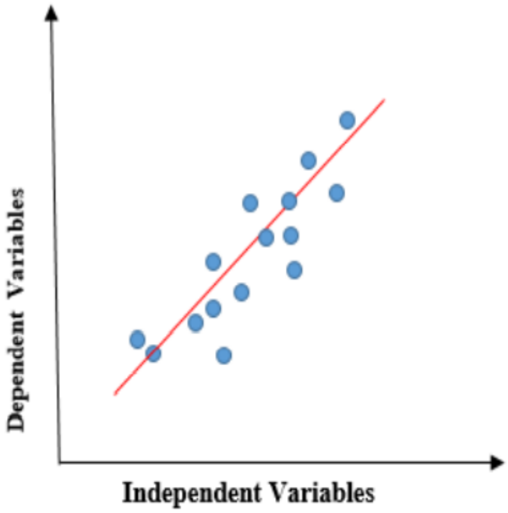
1. **Explain the linear regression algorithm in detail.** **(4 marks)**

Linear regression is a type of supervised machine learning algorithm used for predictive analysis and shows the relationship between the continuous variables. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis).

*If there is a single input variable (x), such linear regression is called **simple linear regression**.*

*And if there is more than one input variable, such linear regression is called **multiple linear regression**.*

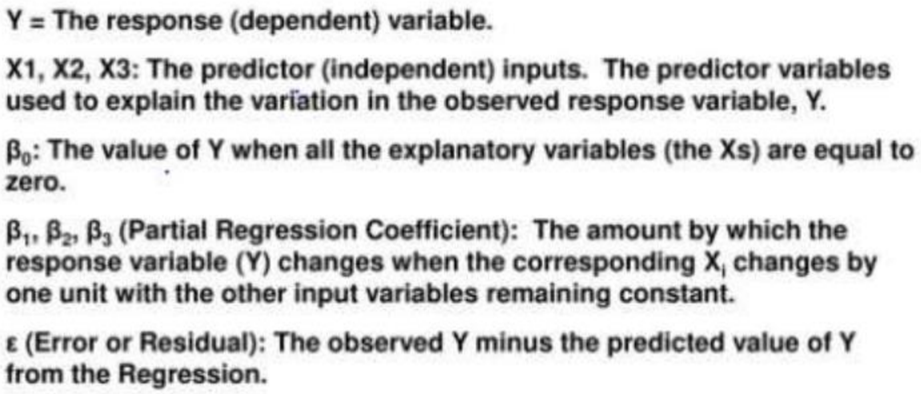
The linear regression model gives a sloped straight line describing the relationship within the variables.



The above graph presents the linear relationship between the dependent variable and independent variables. When the value of x (**independent variable**) increases, the value of y (**dependent variable**) is likewise increasing. The red line is referred to as the best fit straight line. *To calculate best-fit line linear regression uses a traditional slope-intercept form.*

Regression method tries to find the best fit line which shows the relationship between the dependent variable and predictors with least error. In regression, the output/dependent variable is the function of an independent variable and the coefficient and the error term. A general MLR equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$



Y = The response (dependent) variable.

X1, X2, X3: The predictor (independent) inputs. The predictor variables used to explain the variation in the observed response variable, Y.

β_0: The value of Y when all the explanatory variables (the Xs) are equal to zero.

β_1, β_2, β_3 (Partial Regression Coefficient): The amount by which the response variable (Y) changes when the corresponding X_i changes by one unit with the other input variables remaining constant.

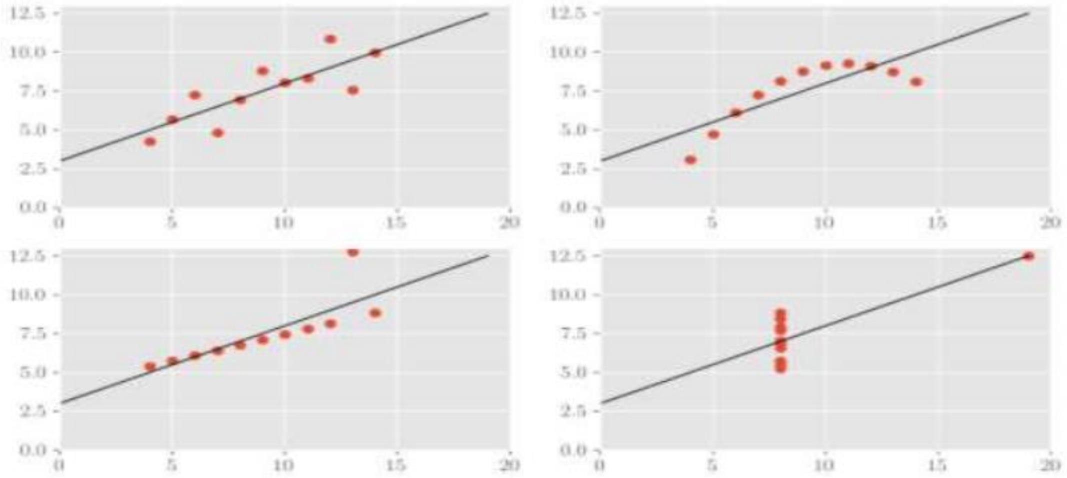
ε (Error or Residual): The observed Y minus the predicted value of Y from the Regression.

2. **Explain the Anscombe's quartet in detail. (3 marks)**

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

It comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.

It reminds us that graphing data prior to analysis is good practice, outliers should be removed when analyzing data, and statistics about a data set do not fully depict the data set in its entirety.



The Anscombe's quartet

All four of these data sets have the same variance in x, variance in y, mean of x, mean of y, and linear regression. While all four data sets have the same linear regression (the best fit line is same for all), it is obvious that the 2nd graph really shouldn't be analyzed with a linear regression at all because it's a curve. The 1st graph probably should be analyzed with a linear regression because it's a scatter plot that moves in a roughly linear manner. Dataset III looks like a tight linear relationship between x and y, except for one large outlier. Dataset IV looks like x remains constant, except for one outlier as well. That one outlier is fitting on the linear regression line but other points don't seem to have any relationship between the variables.

3. **What is Pearson's R?** **(3 marks)**

Pearson correlation (also called Pearson's R) is a correlation coefficient commonly used in linear regression. It measures the strength of association between two variables and the direction of the relationship. Its value ranges between -1 and 1. Simply, it tells us if we can draw a line graph to represent the data.

r=1 indicates a strong positive relationship. Data is perfectly linear with positive slope. It means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other.
r= -1 indicates a strong negative relationship. Data is perfectly linear with negative slope. It means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other.
r=0 indicates no relationship at all. Zero means that for every increase, there isn't a positive or negative increase. The two variables just aren't related.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scalingand standardized scaling?** **(3 marks)**

Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm. Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

Normalization/Min-Max Scaling:
It brings all of the data in the range of 0 and 1.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:
Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ**) zero and standard deviation one (**σ**).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

One disadvantage of normalization over standardization is that it **loses** some information in the data, especially about **outliers**.

- Normalization is good to use 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 Neighbors 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

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**

If there is perfect correlation, then VIF value is infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) = 1/1-1 =1/0 as 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. (3 marks)**

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. In other words it is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

This can be used to check following scenarios:
If two data sets —

1. come from populations with a common distribution
2. have common location and scale
3. have similar distributional shapes
4. have similar tail behavior.