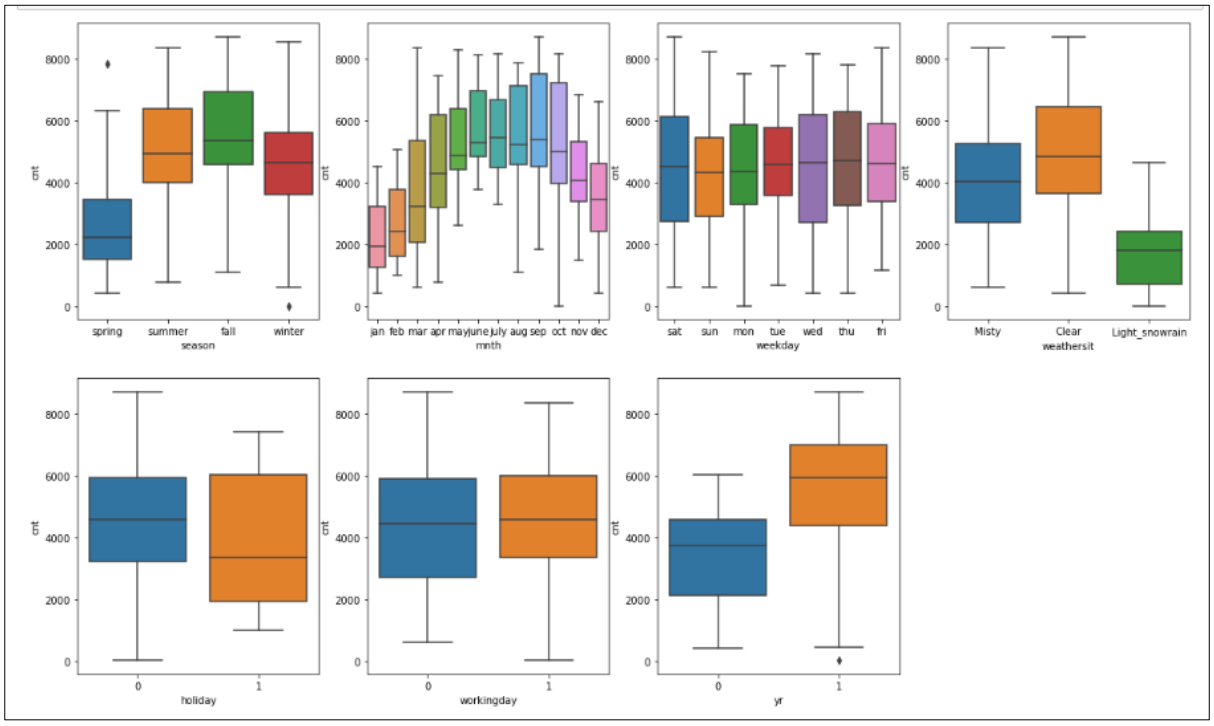
ASSIGNMENT-BASED SUBJECTIVE QUESTIONS

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

   Answer:

   - As per the analysis, following categorical variables were used to check the effect on the dependent variable "**cnt**".
     - Season
     - Mnth
     - Yr
     - weekday,
     - working day and
     - weathersit.
   - To check the effect boxplot and bar plot was used.



   - Following are the points we observe from the boxplot:
     - Fall season seems to have attracted more booking.
     - Sep, Oct month have attracted more booking.
     - There is no impact of working day on bike booking.
     - As we can see, for weathersit - when it is clear or misty booking is high else low.
     - Having holiday increases the bike booking.
     - there is drastic increase in bike booking in year 2019 in comparison to 2018.

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

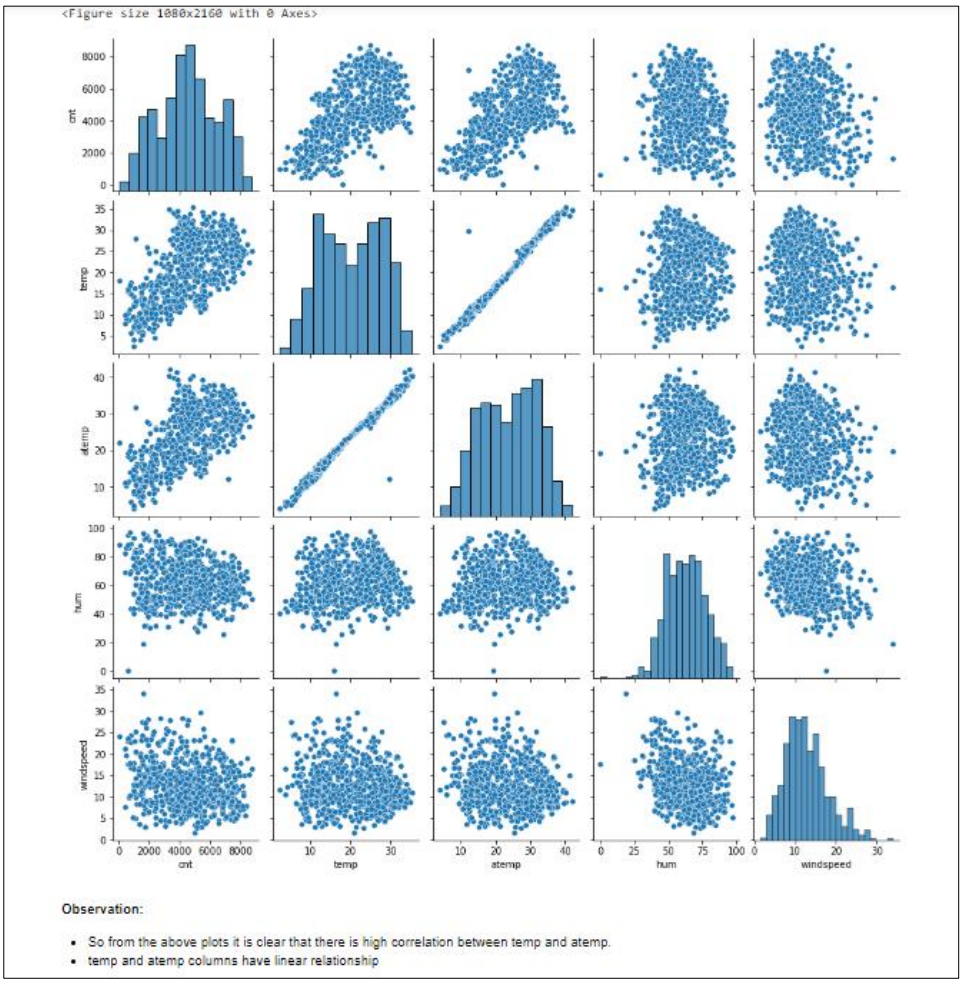
   Answer:

   - drop_first=True is important to use because it helps in reducing the extra column created during dummy variable creation. Which means, it reduces the correlations created among dummy variables.
   - For example, if 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.
   - drop_first=True helps to match the categorical variable with n-1 columns to represent the dummy variables.

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

   Answer:

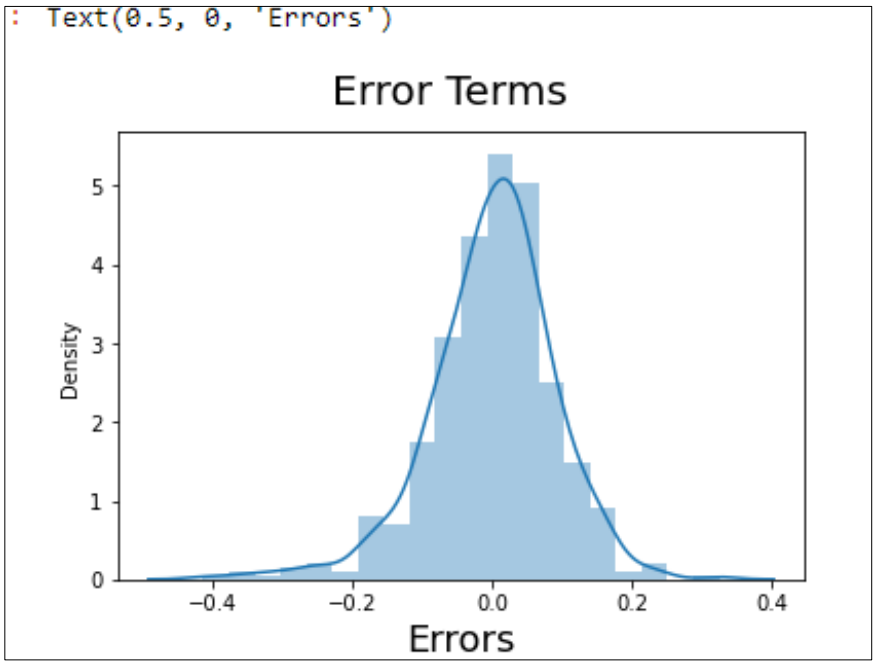
   - The ‘temp’ and ‘atemp’ variables have highest correlation when compared to the rest with target variable as ‘cnt’.



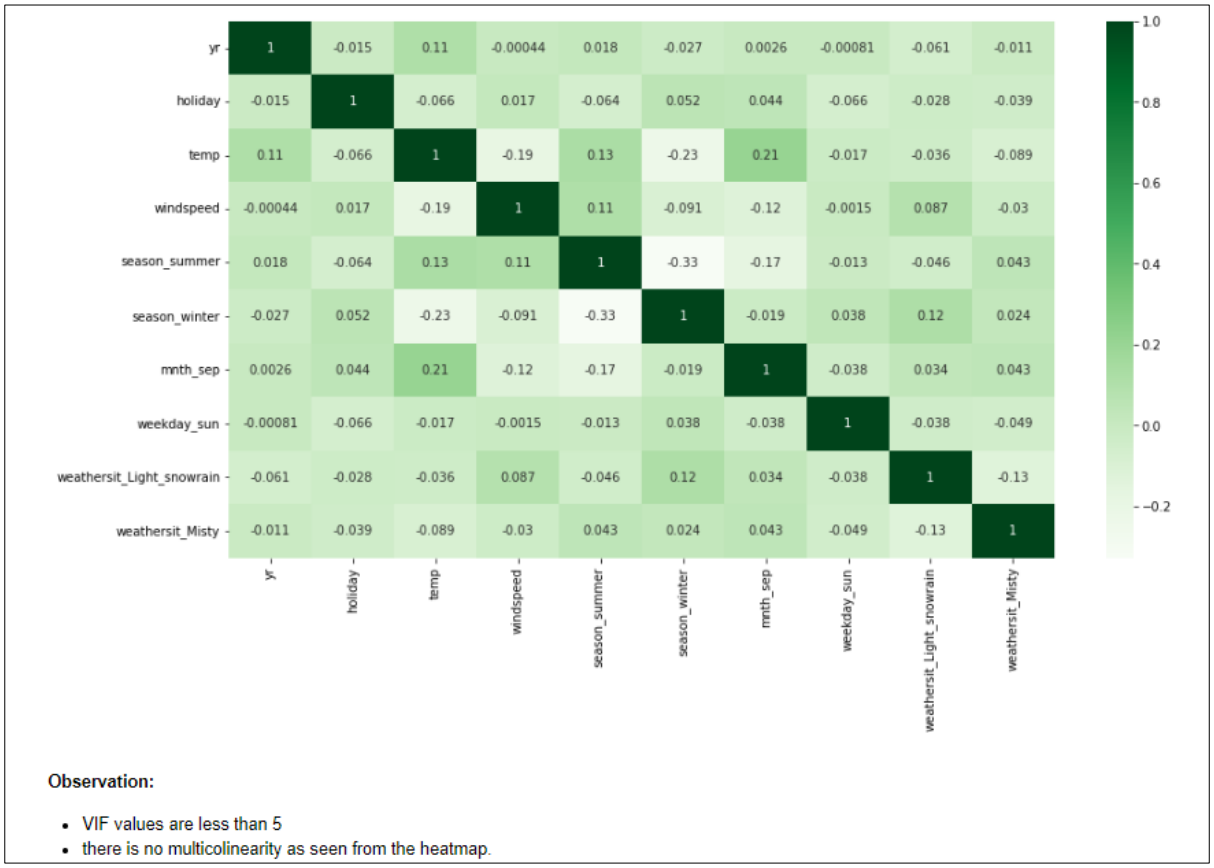
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

Answer:

- Following validation performed to validate the assumptions on the training data set:
  - Error terms should be normally distributed



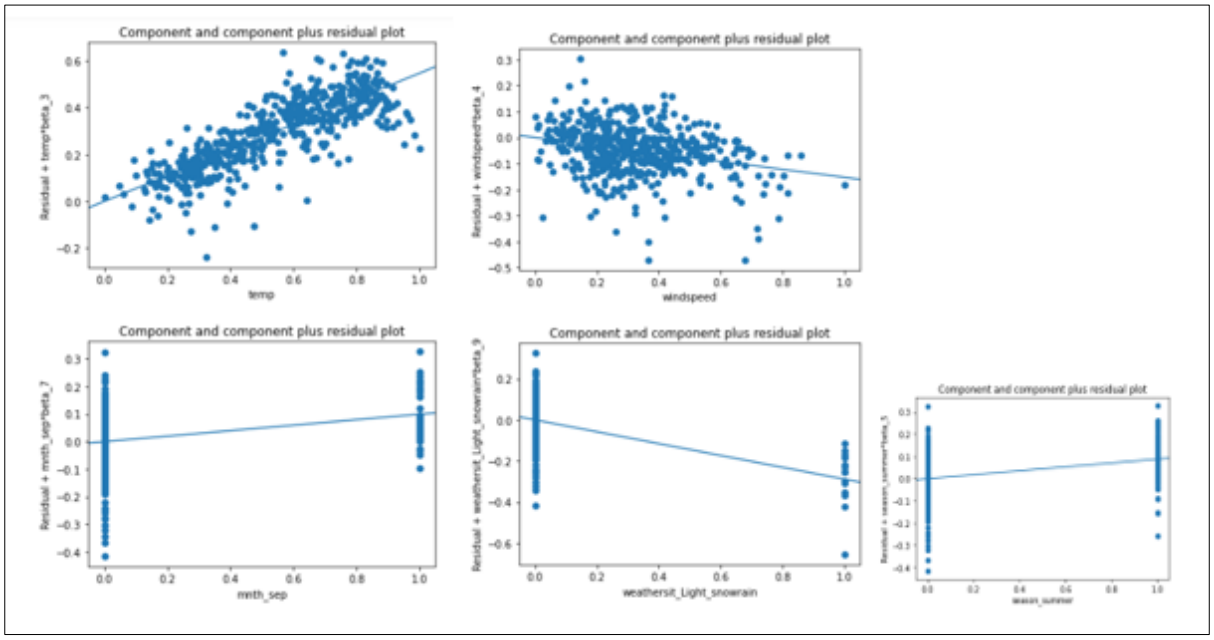
  - Multicollinearity Check



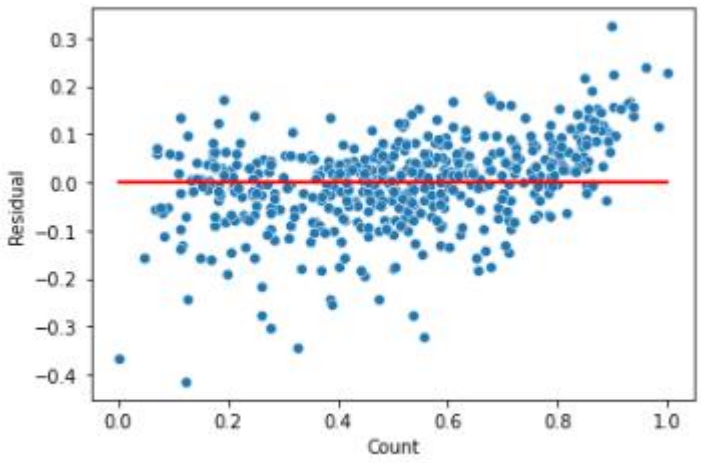
Observation:

- VIF values are less than 5
- there is no multicolinearity as seen from the heatmap.

- Linearity should be visible among variables



- Homoscedasticity - There should be no visible pattern in residual values.



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

   Answer:

   - Temp, windspeed, season

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

   Answer:

   - **Linear Regression** is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.
   - Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).
   - Linear regression is used to predict a quantitative response Y from the predictor variable X.
   - Mathematically, we can write a linear regression equation as:

     $$Y = mX + c$$

     Where Y is the dependent variable we are trying to predict.

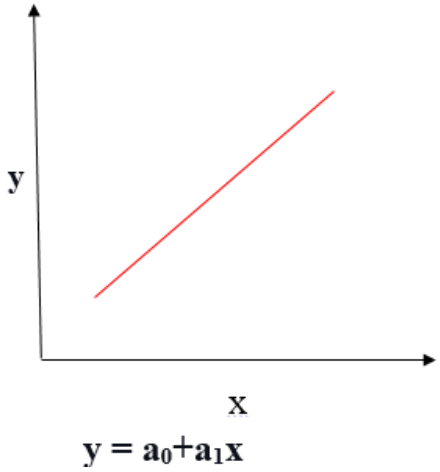
     X is the independent variable we are using to make predictions.

     m is the slope of the regression line which represents the effect X has on Y.

     c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

   - A regression line can be a Positive Linear Relationship or a Negative Linear Relationship. The goal of the linear regression algorithm is to get the best values for a0 and a1 to find the best fit line and the best fit line should have the least error.
     - Positive Linear Relationship

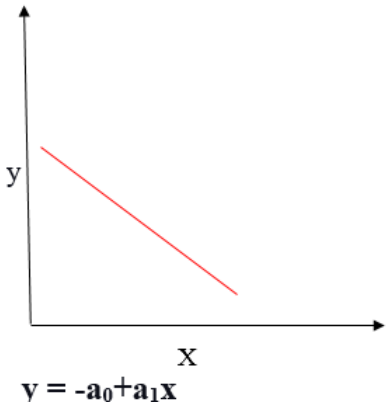
       If the dependent variable expands on the Y-axis and the independent variable progress on X-axis, then such a relationship is termed a Positive linear relationship.

       $y = a_0 + a_1 x$

- Negative Linear Relationship

  If the dependent variable decreases on the Y-axis and the independent variable increases on the X-axis, such a relationship is called a negative linear relationship.



$y = -a_0 + a_1 x$

- The goal of the linear regression algorithm is to get the best values for a0 and a1 to find the best fit line. The best fit line should have the least error means the error between predicted values and actual values should be minimized.
- In Linear Regression, RFE or Mean Squared Error (MSE) or cost function is used, which helps to figure out the best possible values for a0 and a1, which provides the best fit line for the data points.

2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's Quartet was devised by the statistician Francis Anscombe to illustrate how important it was to not just rely on statistical measures when analyzing data.

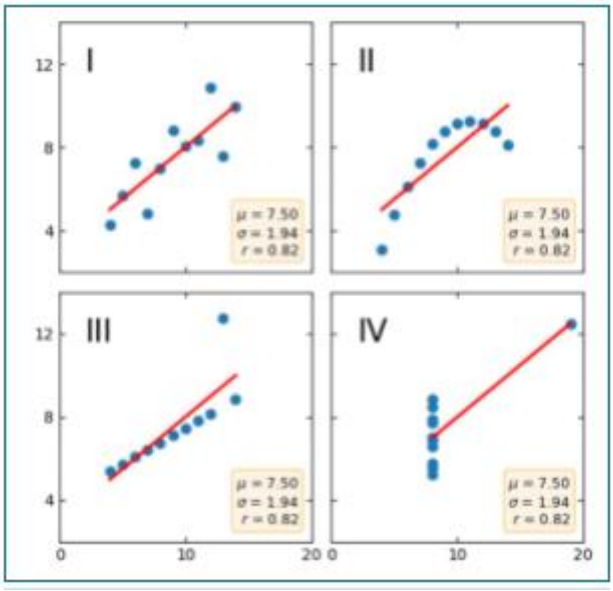
It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics

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

The summary statistics show that the means and the variances were identical for x and y across the groups:

- Mean of x values in each data set = 9.00
- Standard deviation of x values in each data set = 3.32
- Mean of y values in each data set = 7.50
- Standard deviation of x values in each data set = 2.03
- Pearson's Correlation coefficient for each paired data set = 0.82
- Linear regression line for each paired data set: y = 0.500x + 3.00

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story



- Data Set A does indeed fit a linear regression – and so this would be appropriate to use the line of best fit for predictive purposes.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

At last, regression algorithms can be fooled so, it's important to data visualization before build machine learning model.

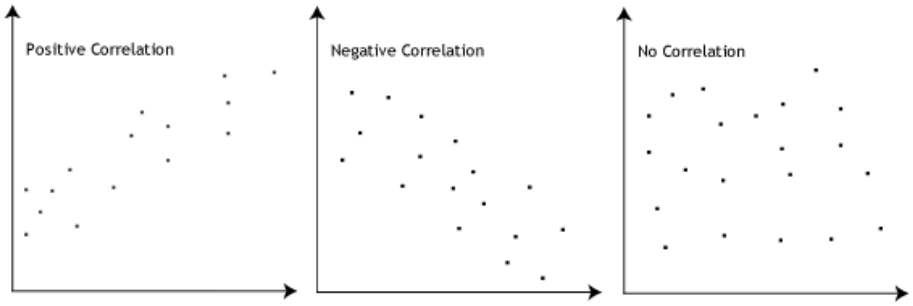
3. What is Pearson's R?

Answer:

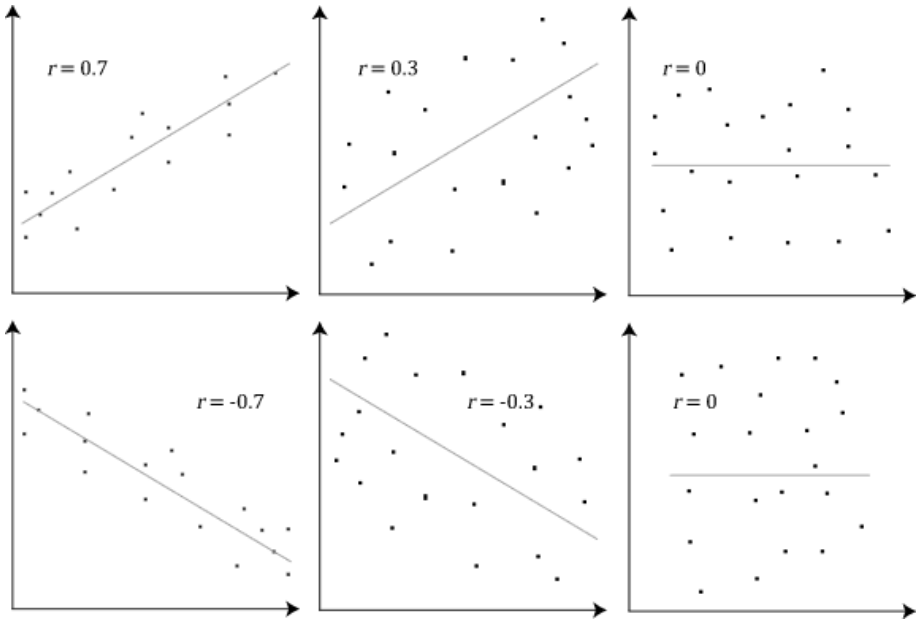
- The Pearson product-moment correlation coefficient (or Pearson correlation coefficient, for short) is a measure of the strength of a linear association between two variables and is denoted by *r*. Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson

correlation coefficient, *r*, indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit).

- The Pearson correlation coefficient, *r*, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



- The stronger the association of the two variables, the closer the Pearson correlation coefficient, *r*, will be to either +1 or -1 depending on whether the relationship is positive or negative, respectively. Achieving a value of +1 or -1 means that all your data points are included on the line of best fit – there are no data points that show any variation away from this line. Values for *r* between +1 and -1 (for example, *r* = 0.8 or -0.4) indicate that there is variation around the line of best fit. The closer the value of *r* to 0 the greater the variation around the line of best fit. Different relationships and their correlation coefficients are shown in the diagram below:



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

Answer:

- It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a range. It also helps in speeding up the calculations in an algorithm.
- It happens that 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.
- Now, scaling helps us to bring all the variables to the same level of magnitude.
- Again, Scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc

Why Scaling is Performed?

- if there is a vast difference in the range say few ranging in thousands and few ranging in the tens, and it makes the underlying assumption that higher ranging numbers have superiority of some sort. So these more significant number starts playing a more decisive role while training the model.
- Suppose we have two features of weight and price. The “Weight” cannot have a meaningful comparison with the “Price.” So the assumption algorithm makes that since “Weight” > “Price,” thus “Weight,” is more important than “Price.”
- feature scaling is needed to bring every feature in the same footing without any upfront importance.

Difference between normalized scaling and standardized scaling

S.NO.	Normalisation	Standardisation
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.

S.NO.	Normalisation	Standardisation
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.

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

   Answer:

   - When the value of VIF is infinite it shows a perfect correlation between two independent variables.
   - In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity.
   - To solve this, 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.**

   Answer:

   - Q-Q plots are also known as Quantile-Quantile plots.
   - They plot the quantiles of a sample distribution against quantiles of a theoretical distribution. This helps to determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.
   - QQ plot can also be used to determine whether two distributions are similar or not. If they are quite similar you can expect the QQ plot to be more linear.
   - The linearity assumption can best be tested with scatter plots. Secondly, the linear regression analysis requires all variables to be multivariate normal. This assumption can best be checked with a histogram or a Q-Q-Plot.
   - The advantages of the q-q plot are:
     - The sample sizes do not need to be equal.
     - Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.
     - For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.

- In Linear Regression when we have a train and test dataset then we can create Q-Q plot by which we can confirm that both the data train and test data set are from the population with the same distribution or not.
- QQ plots is very useful to determine
    - If two populations are of the same distribution
    - If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it’s met using this.
    - Skewness of distribution