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? (3 marks)

Below are the insights from the below plots:

1. Fall season has more booking followed by summer season.
2. Most of the bookings are in the months from June to September.
3. During weekdays, most of the bookings are on thursday, friday and saturday.
4. Year (0: 2018, 1:2019) - Clear weather leads to an increased number of bookings. In 2019, the booking number increased as compared to the number in 2018.
5. Year 2019 had more bookings than 2018.
6. On holiday, the numbers of bookings are less.
7. There is not too much difference on the number of bookings if it's a working day or not. However, working day slightly has the large number of booking.

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

While using get_dummies, the drop_first parameter specifies whether or not we want to drop the first category of the categorical variable we're encoding.

If its parameters drop_first = False, the get_dummies will create one dummy variable for every level of the input categorical variable

However, we don't dummy variable for every level of the input categorical variable. If we set drop_first=True during dummy variable creation, we can avoid:

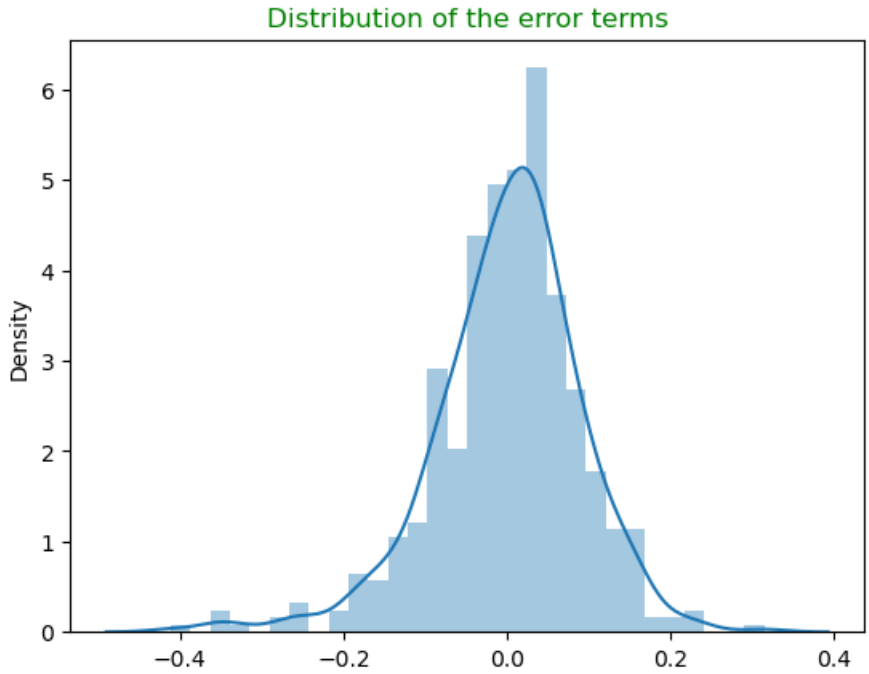
- Dummy variable trap issue
- Reduce Multicollinearity
- Memory efficiency: The reason we are trying to delete a column is because of memory efficiency. If the same information can be represented by two variables, then why use three variables.

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

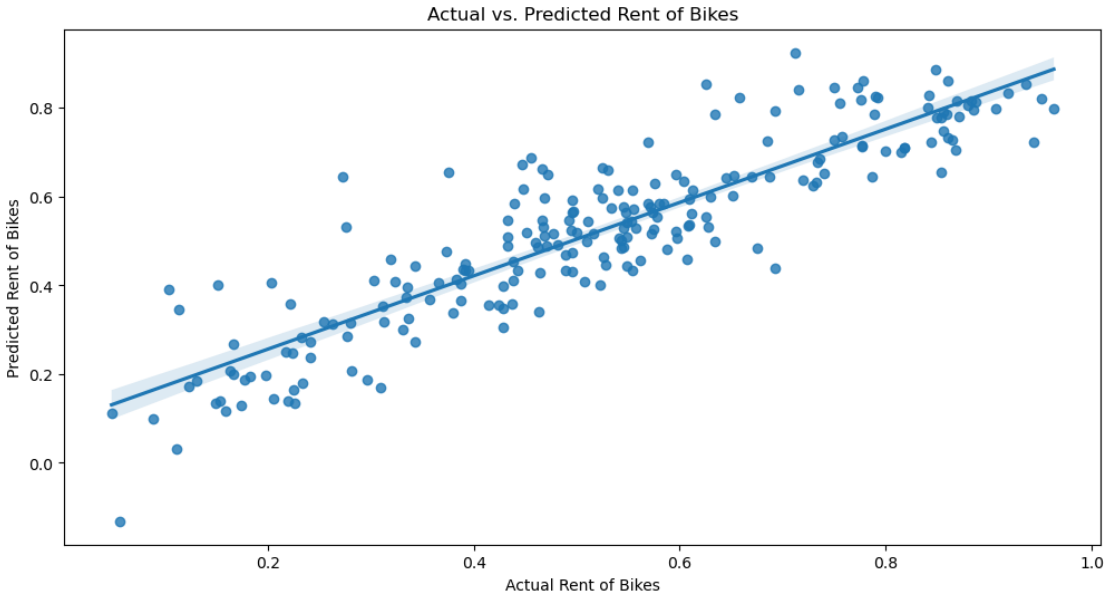
Temp, atemp with correlation of 65%

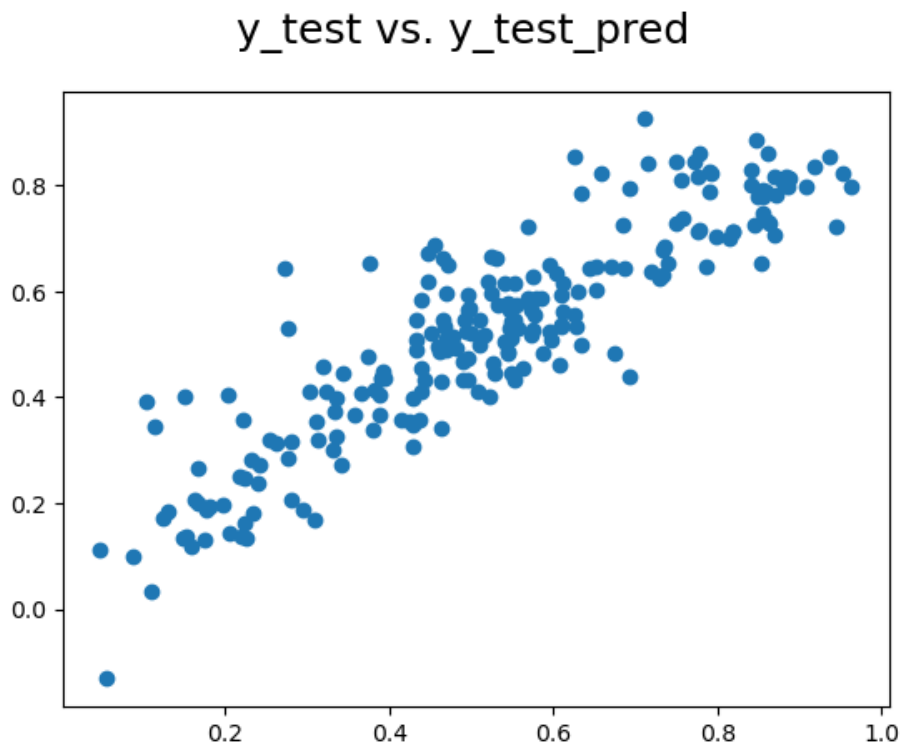
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

1. I ensured whether the error terms are normally distributed i.e. the distribution is centered around 0.



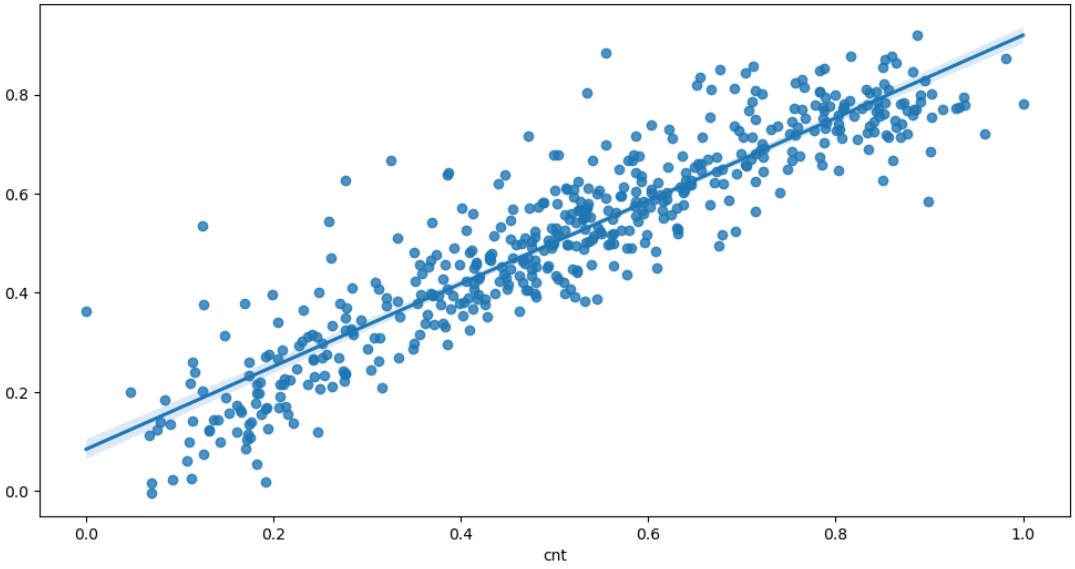
2. Linear relationship validation





3. I performed Multicollinearity check to ensure there is no high variable correlation

4. Ensured whether error terms have approximately a constant variance.



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

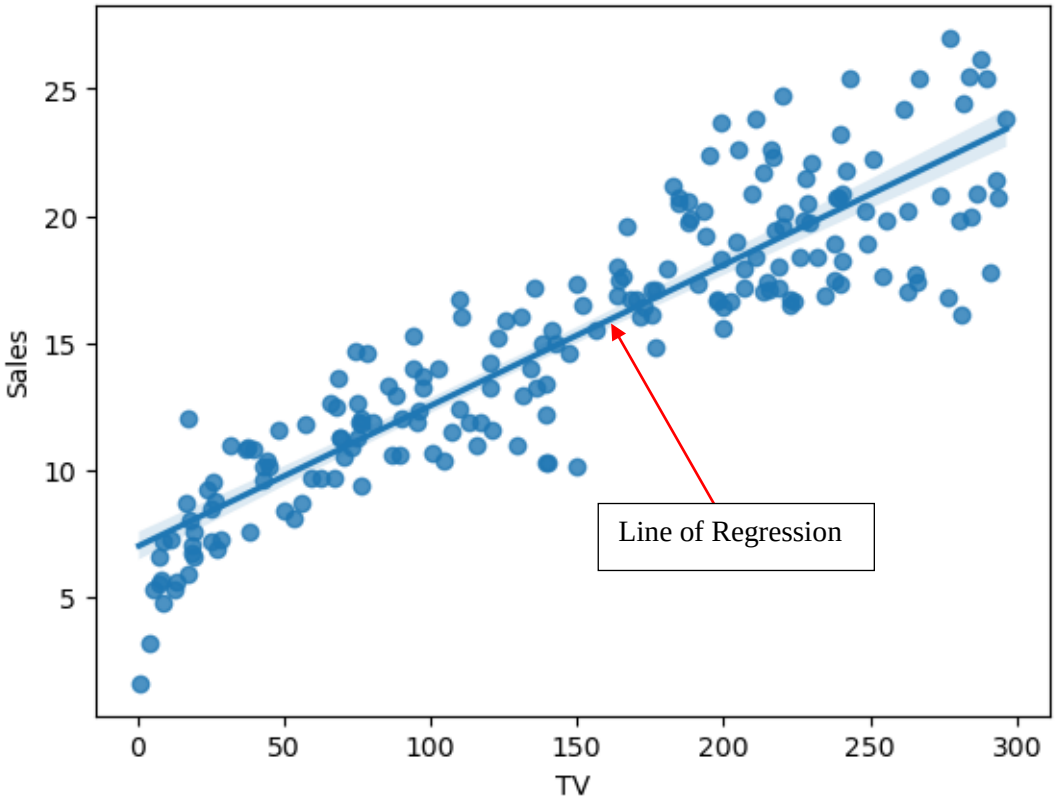
1. temp
2. Yr.
3. Sep

General Subjective Questions

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

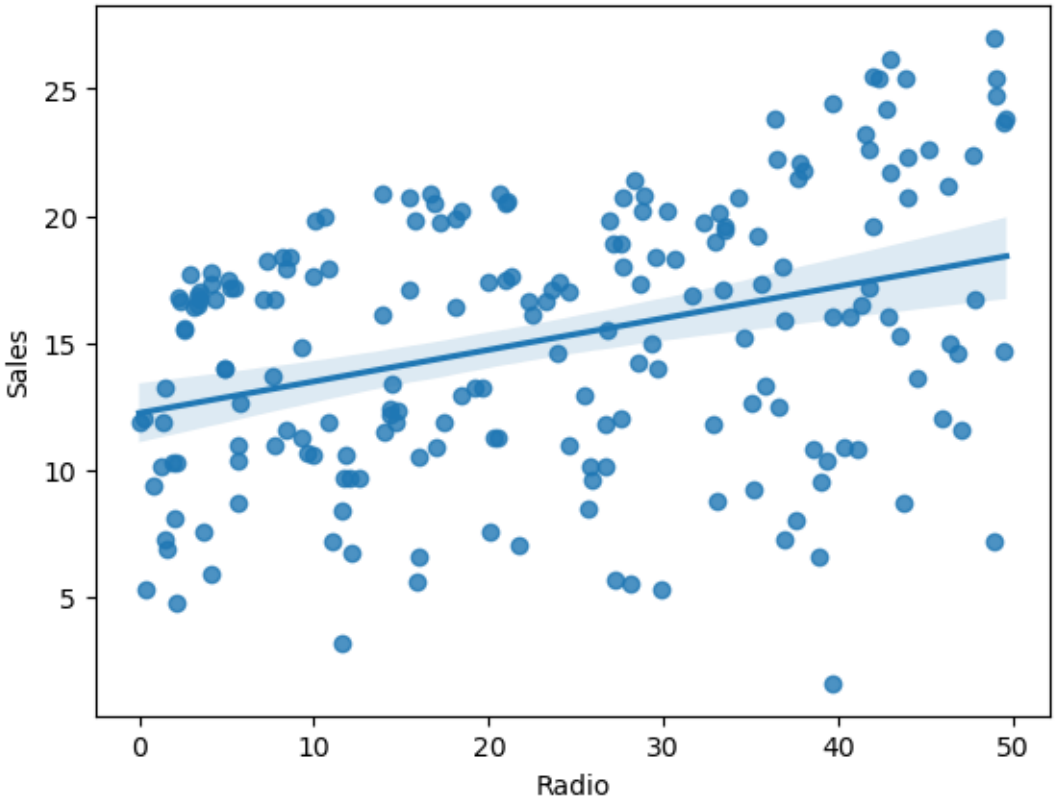
Solution: Linear regression is a supervised machine learning algorithmthat identifies a linear relationship between a target variable and a predictor variable. In other words, we can say that the value of an unknown data is predicted by using the known data value.

One good idea when we are doing linear regression or any sort of regression is to look at the scatter plot. Because if the scatter plot shows a positive trend between the predictor variable and the target variable, then, we will know whether linear regression or any sort of regression makes sense or not.



The above scatter plot describes the linear relationship between the target variable (‘Sales’) and the predictor variable (‘TV’). As we can see that there is a straight line totallyappropriate for the linear regression.

The below scatter plot describes the scenario, where the relationship is not as smooth between the target variable (‘Sales’) and the predictor variable(‘Radio’)than it was with TV. The data is more scattered in the y direction, and the slope is alsolesser.



Equation of linear regression is:

$y=c+m_1x_1+m_2x_2+.....+m_nx_n$

- y is the response
- c is the intercept
- m_1 is the coefficient for the first feature
- $m_n$is the coefficient for the nth feature

Building a model in Linear regression includes the below steps:

1. Create x (predictor) and y(target)
2. Create train and test sets (70-30, 80-20)
3. Train model on the training set (i.e., learn the coefficients)
4. Evaluate the model (training set, test set)

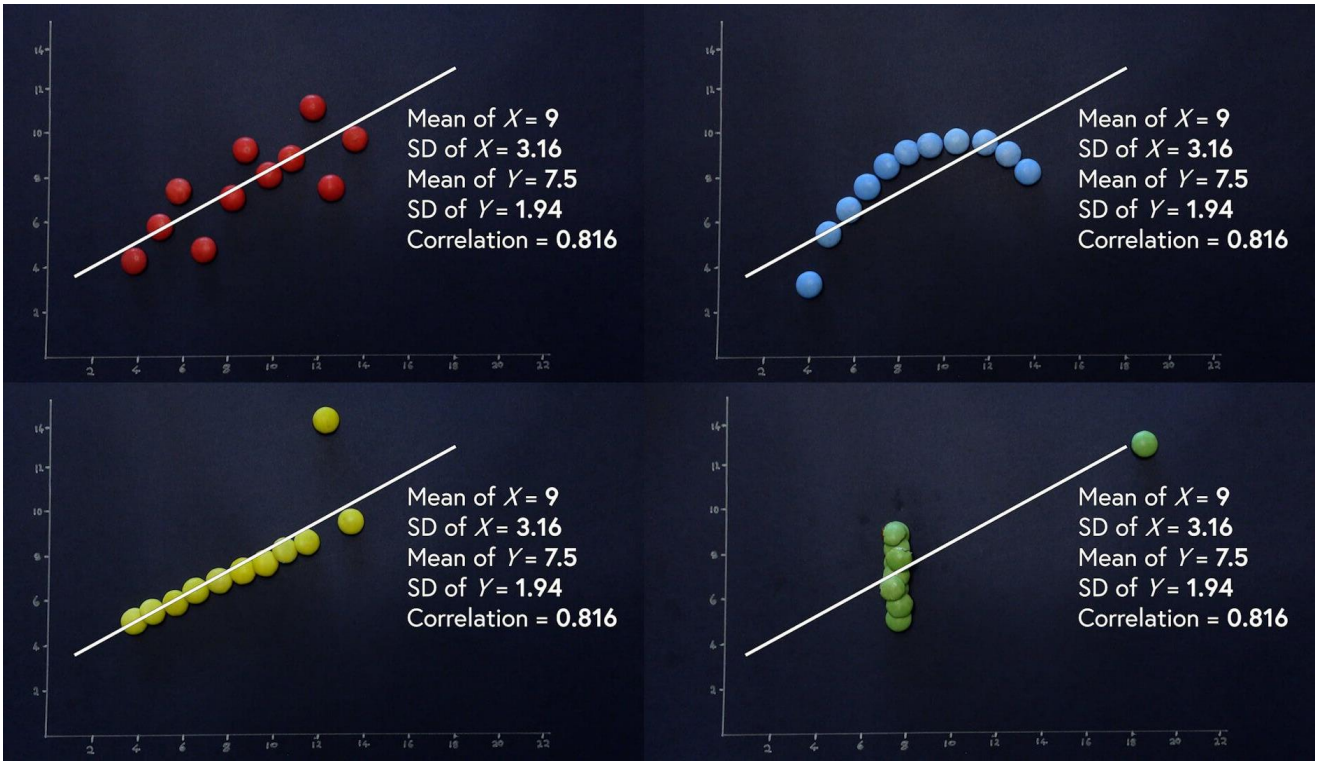
2. Explain the Anscombe's quartet in detail. (3 marks)

Solution: Anscombe's quartet describes the value of plotting data before analyzing it with the statistical properties (such as mean, correlation, variance etc.) and building the

model. Anscombe's quartet consists of four data-set and each data-set further consists of eleven (x,y) points, such as x1 & y1, x2 & y2, x3 & y3, x4 & y4.

The point of interest is that all of the datasets have the same statistics information (such as mean, variance, standard deviation etc) but have different graphical representation. As an example, there are four datasets Red, Blue, Yellow and Green.

Red		Blue		Yellow		Green	
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



As we can see in the above images, all datasets have the same summary statistics (mean, standard deviations (SD), correlation coefficient, and linear regression line) but different graphical representation.

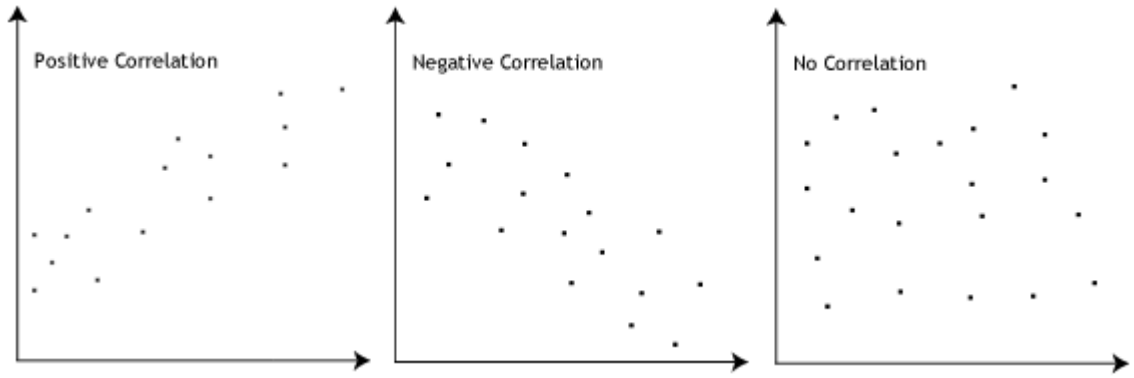
The first dataset i.e. Red describes the linear relationship. The second dataset i.e. Yellow describes a linear trend, but the regression line is affected by a single outlier. The third dataset i.e. Blue describes an unexpected turn.

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

Pearson's *R*, also called Pearson's correlation, is a correlation coefficient commonly used in linear regression to measure the statistical relationship, or association, between two continuous variables. It is denoted by r. Pearson's correlation attempts to draw a line of best fit through the data of two variables, In other words, *r* indicates how far away all the data points are to the line of best fit.

Pearson correlation coefficient, *r*, can take a range of values from +1 to -1.where:

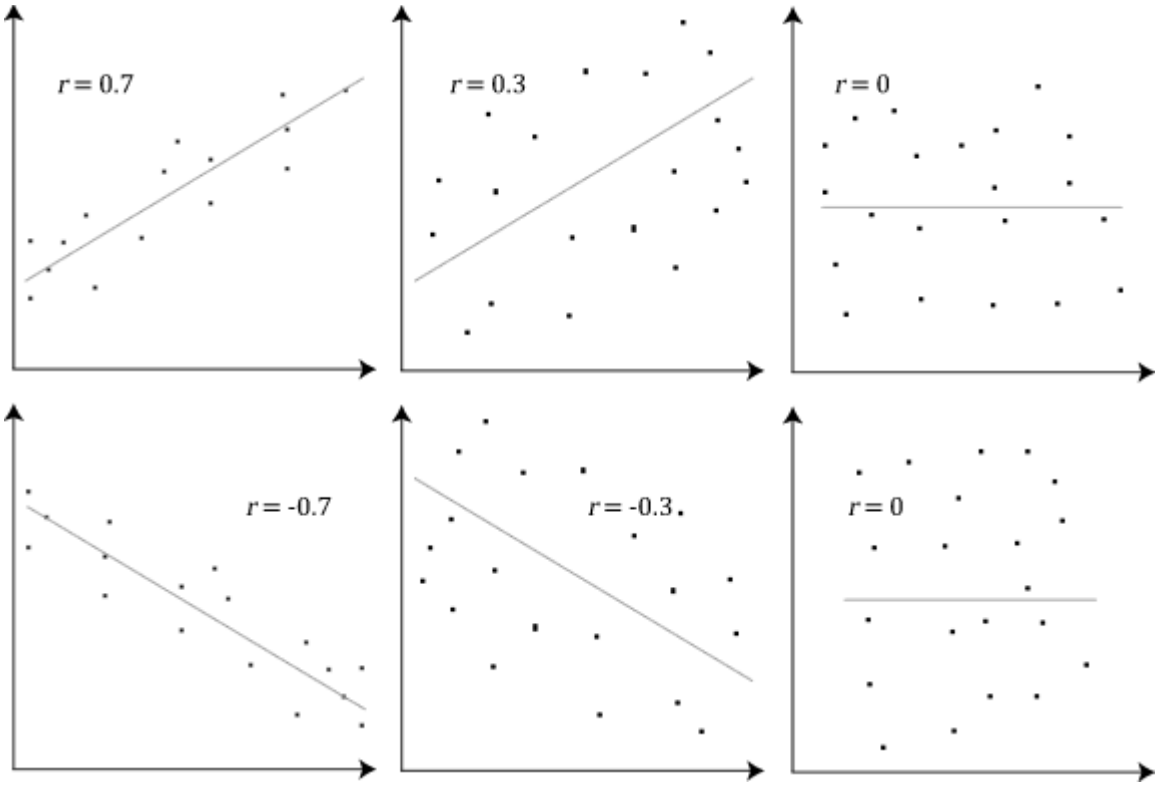
- 0 indicates that there is no association between the two variables.
- A value greater than 0 indicates a positive association i.e. if the value of one variable increases, the value of the other variable too increases.
- A value less than 0 indicates a negative association; i.e. if the value of one variable increases, the value of the other variable decreases.



Pearson correlation coefficient *r* closer to either +1 or -1 represents the stronger association between two variables.

- A value of +1 or -1 means that all the data points are included on the line of best fit i.e. there are no data points that show any variation away from this line.
- Value of *r* between +1 and -1 indicates 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:



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

Solution: Before model building, we first need to perform the test-train split and scale the features.

In a dataframe, some variables are on a different scale with respect to all other numerical variables taking very small values. Hence, it is extremely important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. This might become very annoying at the time of model evaluation. So it is advised to use standardization or normalization so that the units of the coefficients obtained are all on the same scale. There are two common ways of rescaling:

1. Min-Max scaling
2. Standardisation (mean-0, sigma-1)

Min-Max scaling (Normalization) compresses all the data between 0 and 1. The max value of the data is 1 and min value is 0. For Normalized scaling, we use MinMaxScaler() function.

Normalization = $(x - x_{min})/(x_{max} - x_{min})$

When $x=x_{max}$, we will get 1, if $x_{max}=x_{min}$, we will get 0

So, the entire data is being compressed between 0 and 1.

Standardisation: Standardization converts our data so that it has mean of 0. So, it centres the data around mean 0, and the standard deviation is 1.

The standardized value of [(x - mu)/sigma] will have a mean of 0 and standard deviation will be 1.

It is advisable to use min-max scaling because it takes care of outliers. So, if there is an outlier in the original data points, they should have been mapped to 1 and other data points should have been between 0 and 1.

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

Checking VIF

Variance Inflation Factor or VIF, gives a basic quantitative idea about how much the feature variables are correlated with each other. It is an extremely important parameter to test our linear model. The higher a variable is related to all the other variables, the higher the VIF of that variable.

The common heuristic we follow for the VIF values is:

> 10: Definitely high VIF value and the variable should be eliminated.

> 5: Can be okay, but it is worth inspecting.

< 5: Good VIF value. No need to eliminate this variable.

The infinite values of the Variance Inflation Factor (VIF) indicate perfect multicollinearity within the dataset.

The formula for calculating VIF is:

$VIF_i = 1/ (1-R_i^2)$

When the correlation between predictor variables is perfect, i.e. $R_i^2 = 1$ the denominator in the VIF formula becomes zero, leading to a division by zero and resulting in an infinite VIF value.

One of the possible reasons for perfect multicollinearity includes the linear relationships among predictor variables. To overcome multicollinearity, one of the alternatives is to remove the correlated variables from the model.

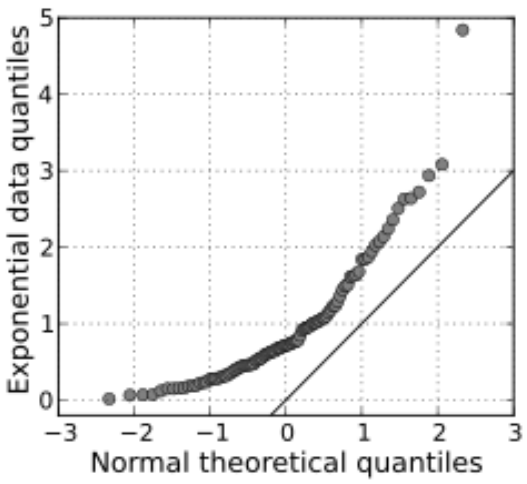
5: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

A Q-Q plot, short for quantile-quantile plot, is a scatterplot that compares the quantiles of two distributions. One distribution is usually the observed data, and the other is a theoretical or reference distribution, such as the normal distribution. The idea is to see how well the data fit the expected distribution by checking if the points lie on or near a straight line. Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or if it follows some other known distribution.

A Q-Q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. If the two data sets have come from populations with different distributions then the data points will be far from the reference line.

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.



To draw a Quantile-Quantile (Q-Q) plot, following steps are used:

1. **Collect the Data**: Gather the dataset for which we want to create the Q-Q plot..
2. **Sort the Data**: Arrange the data in either ascending or descending order.
3. **Choose a Theoretical Distribution**: Determine the theoretical distribution against which we want to compare the dataset.
4. **Calculate Theoretical Quantiles**: Compute the quantiles for the chosen theoretical distribution. For example, if we are comparing against a normal distribution, we would use the inverse cumulative distribution function (CDF) of the normal distribution to find the expected quantiles.
5. **Plotting**:
   - Plot the sorted dataset values on the x-axis.
   - Plot the corresponding theoretical quantiles on the y-axis.
   - Each data point (x, y) represents a pair of observed and expected values.
   - Connect the data points to visually inspect the relationship between the dataset and the theoretical distribution.

The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.