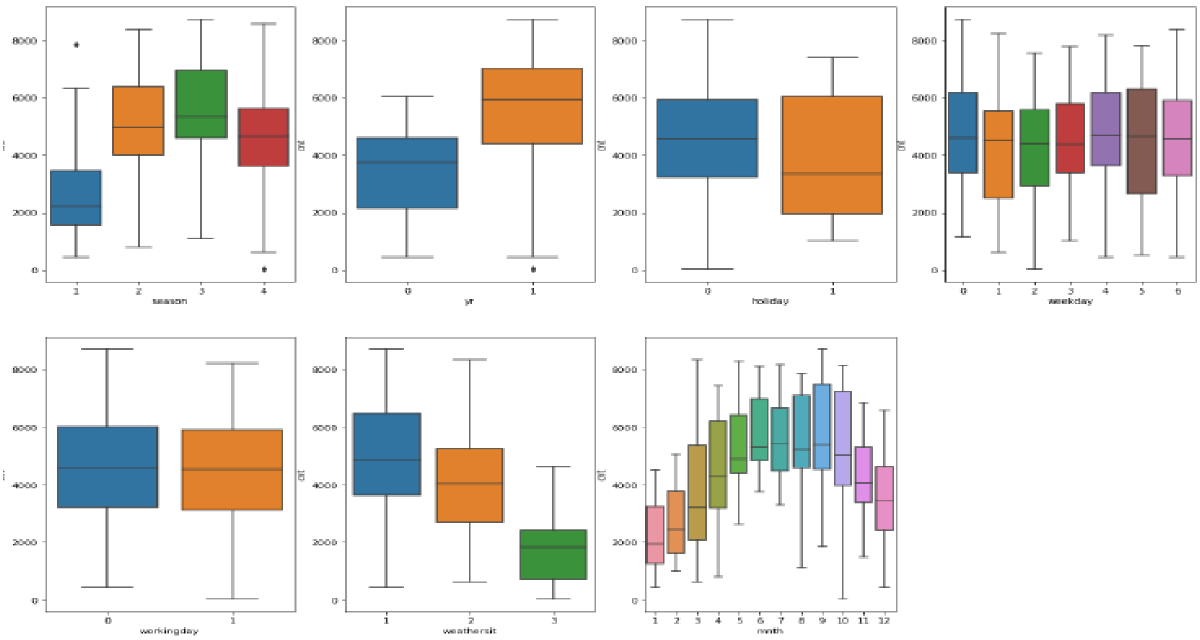
Assignment-based Subjective Questions :

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



The categorical variable in the dataset were season, yr , holiday, weekday ,workingday, and weathersit and mnth . These were visualized using a boxplot.

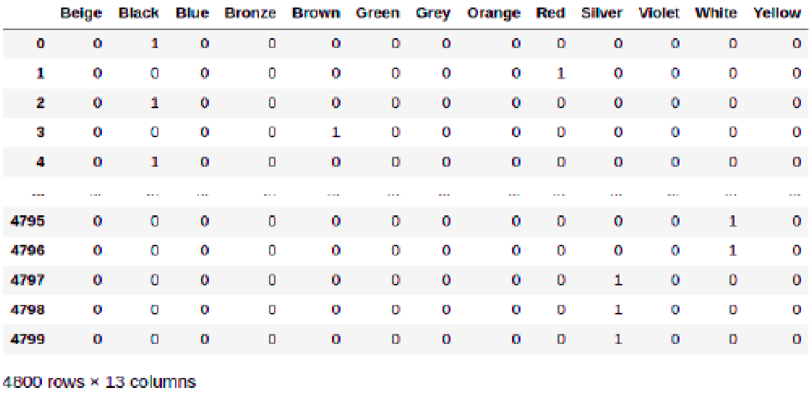
These variables had the following effect on our dependant variable: -

- Season - The boxplot showed that spring season had least value of cnt whereas fall had maximum value of cnt. Summer and winter had intermediate value of cnt.
- Weathersit - There are no users when there is heavy rain/ snow indicating that this weather is extremely unfavourable. Highest count was seen when the weathersit was’ Clear, Partly Cloudy’.
- Yr - The number of rentals in 2019 was more than 2018
- Holiday - rentals reduced during holiday.
- Mnth - September saw highest no of rentals while December saw least. This observation is in accordance with the observation made in weathersit. The weather situation in December is usually heavy snow due to which the rentals might have dropped.
- Weekday - The count of rentals is almost even throughout the week
- Workingday – The median count of users is constant almost throughout the week.

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

If we don't drop the first column then the dummy variables will be correlated (redundant). This 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. In the figure below, check the number of columns, instead of 13 we got 12 columns. It removes the first column of the get_dummies dataframe. The first column for the "Body Color" column is Beige. If there is a beige car, all columns are 0. When all columns are 0, the model knows it's a beige car. More columns mean less performance and more training time. Imagine we have 20 columns that are not numerical. If we use 'drop_first', we get 20 columns less. So, it is useful to use the drop_first = True parameter for model performance.

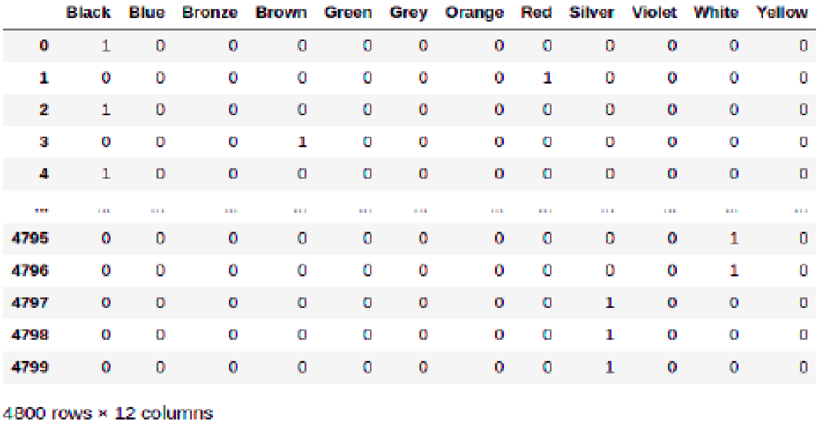
When we use, pd.get_dummies(df['Body Color'])



	Beige	Black	Blue	Bronze	Brown	Green	Grey	Orange	Red	Silver	Violet	White	Yellow
0	0	1	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	1	0	0	0	0
2	0	1	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	1	0	0	0	0	0	0	0	0
4	0	1	0	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
4795	0	0	0	0	0	0	0	0	0	0	0	1	0
4796	0	0	0	0	0	0	0	0	0	0	0	1	0
4797	0	0	0	0	0	0	0	0	0	1	0	0	0
4798	0	0	0	0	0	0	0	0	0	1	0	0	0
4799	0	0	0	0	0	0	0	0	0	1	0	0	0

4800 rows × 13 columns

When we use, pd.get_dummies(df['Body Color'], drop_first = True)

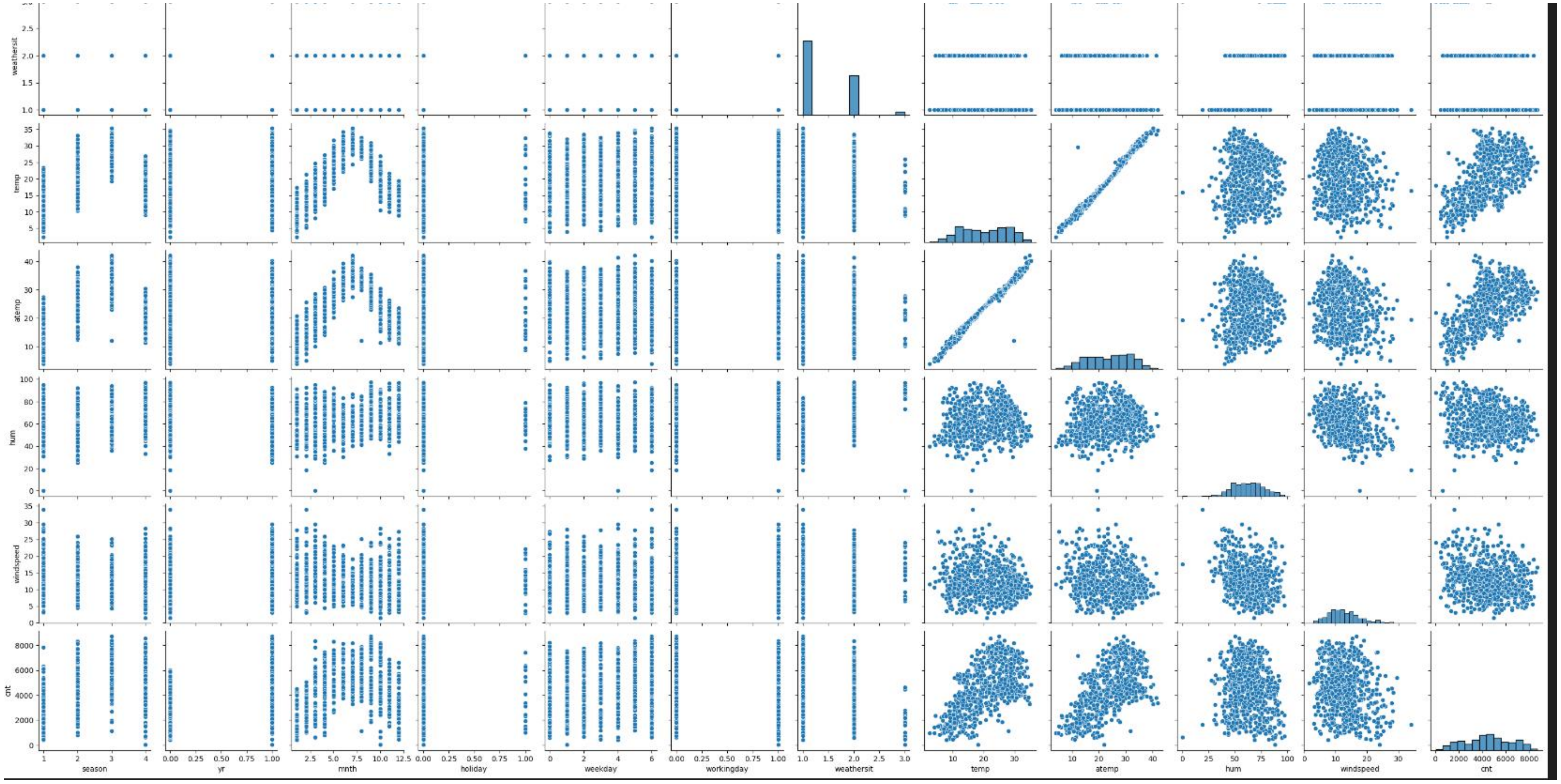


	Black	Blue	Bronze	Brown	Green	Grey	Orange	Red	Silver	Violet	White	Yellow
0	1	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	1	0	0	0	0
2	1	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	1	0	0	0	0	0	0	0	0
4	1	0	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...
4795	0	0	0	0	0	0	0	0	0	0	1	0
4796	0	0	0	0	0	0	0	0	0	0	1	0
4797	0	0	0	0	0	0	0	0	1	0	0	0
4798	0	0	0	0	0	0	0	0	1	0	0	0
4799	0	0	0	0	0	0	0	0	1	0	0	0

4800 rows × 12 columns

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

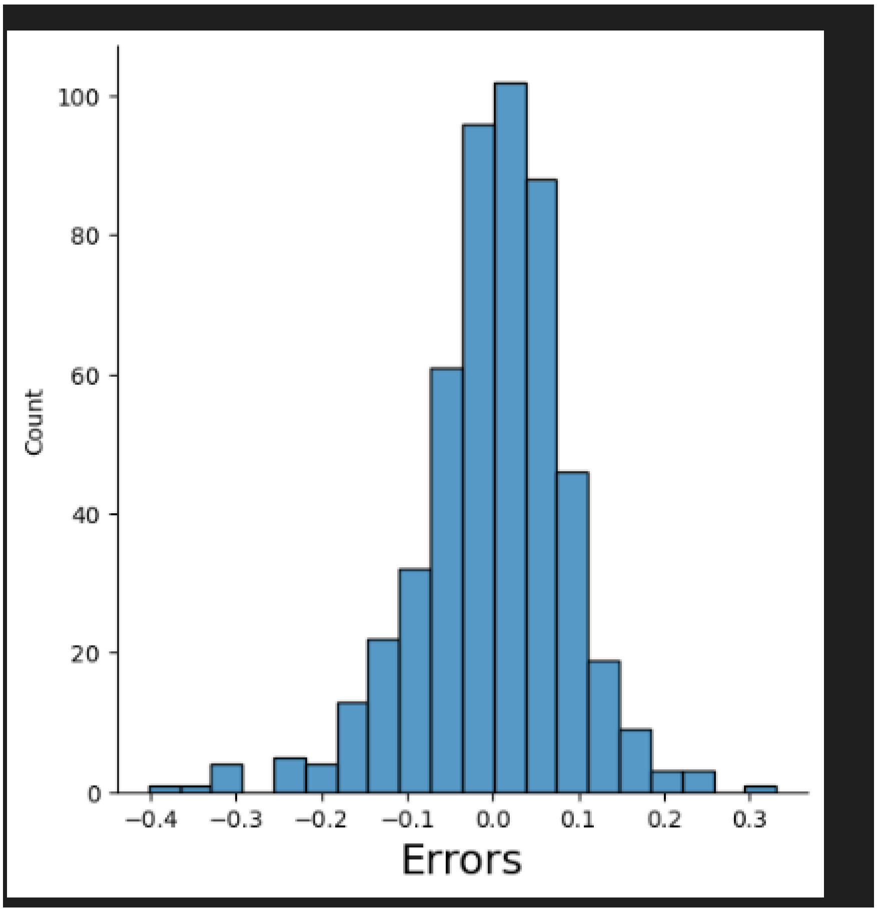
Using the below pairplot it can be seen that, "temp" and "atemp" are the two numerical variables which are highly correlated with the target variable (cnt)


weathersit
temp
atemp
hum
windspeed
cnt
season
yr
mnth
holiday
weekday
workingday
weathersit
temp
atemp
hum
windspeed
cnt
1 2 3 4
0.00 0.25 0.50 0.75 1.00
2.5 5.0 7.5 10.0 12.5
0 2 4 6
1.0 1.5 2.0 2.5 3.0
10 20 30
10 20 30 40
0 25 50 75 100
0 10 20 30
0 2000 4000 6000 8000

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

The following tests were done to validate the assumptions of linear regression:
1. First, linear regression needs the relationship between the independent and dependent variables to be linear. We visualised the numeric variables using a pairplot to see if the variables are linearly related or not. Refer to the notebook for more details.

2. Secondly, Residuals distribution should follow normal distribution and centred around 0 (mean = 0). We validated this assumption about residuals by plotting a distplot of residuals and saw if residuals are following normal distribution or not. The diagram below shows that the residuals are distributed about mean = 0.



3. Thirdly, linear regression assumes that there is little or no multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other. We calculated the VIF (Variance Inflation Factor) to get the quantitative idea about how much the feature variables are correlated with each other in the new model. Refer to the notebook for more details.

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

The top 3 features are:
1. temp - coefficient: 0.437655
2. yr - coefficient: 0.234287
3. weathersit_Light Snow & Rain - coefficient: -0.292892

General Subjective Questions

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

Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric values. Linear Regression is the most basic form of regression analysis. Regression is the most commonly used predictive analysis model.
Linear regression is based on the popular equation “y = mx + c”.
It assumes that there is a linear relationship between the dependent variable(y) and the predictor(s)/independent variable(x).
In regression, we calculate the best fit line which describes the relationship between the independent and dependent variable. Regression is performed when the dependent variable is of continuous data type and Predictors or independent variables could be of any data type like continuous, nominal/categorical etc. 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.
Regression is broadly divided into simple linear regression and multiple linear regression.

1. Simple Linear Regression: SLR is used when the dependent variable is predicted using only one independent variable.
   Y(i) = B (0) + B(1)X(i) + E(i)
   Where Y(i) is a dependent variable, B (0) is the Population Y intercept, B(1) is the population sloe coefficient, X(i) is the independent variable and E(i) is the error term

The equation for SLR will be:
2. Multiple Linear Regression: MLR is used when the dependent variable is predicted using multiple independent variables.
The equation for MLR will be:

$$\textit{observed data} \rightarrow \quad y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p + \varepsilon$$

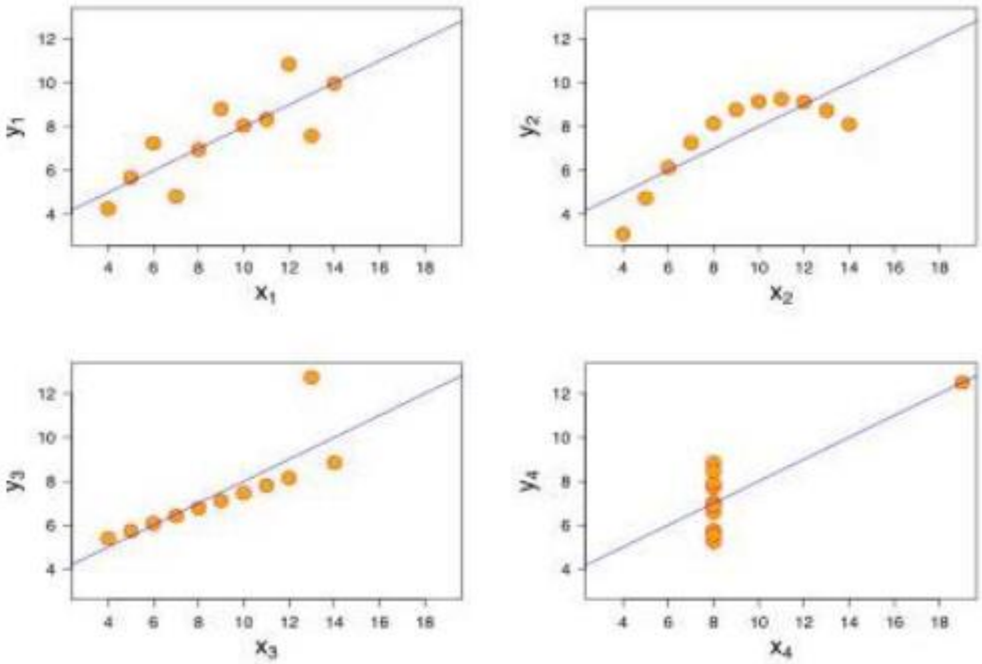
$$\textit{predicted data} \rightarrow \quad y' = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

$$\textit{error} \rightarrow \quad \varepsilon = y - y'$$

B1 = coefficient for X1 variable
B2 = coefficient for X2 variable
B3 = coefficient for X3 variable and so on…
B0 is the intercept (constant term)

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

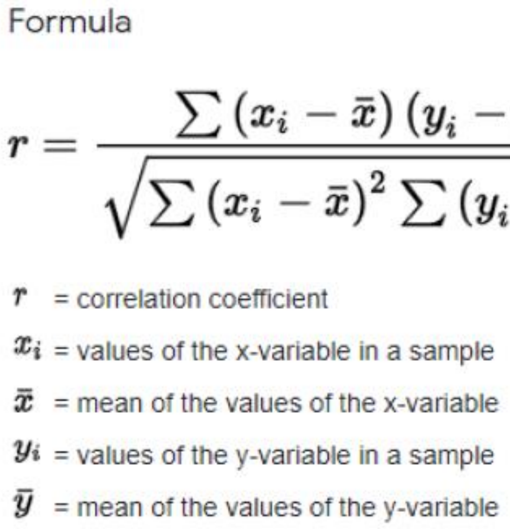
Anscombe’s Quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph. It was developed to emphasize both the importance of graphing data before analyzing 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? (3 marks)

Pearson's r is a numerical summary of the strength of the linear association between the variables.It value ranges between -1 to +1. It shows the linear relationship between two sets of data. In simple terms, it tells us “can we draw a line graph to represent the data? “



Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

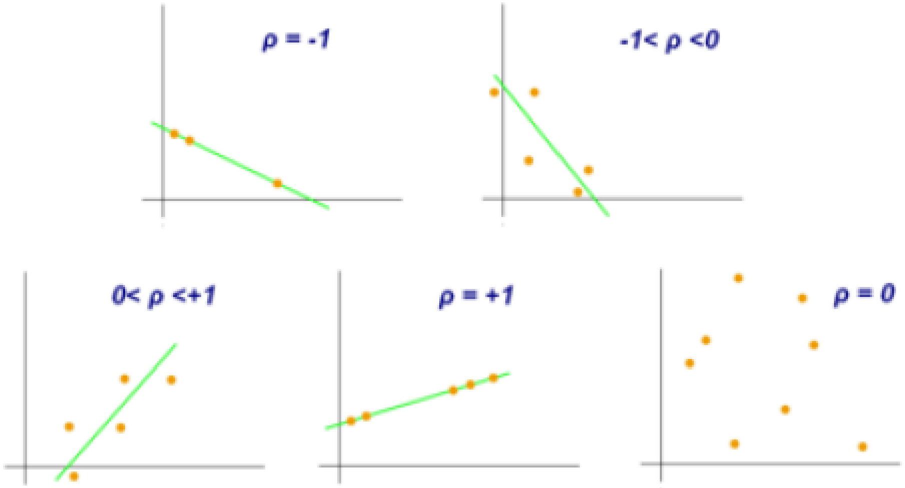
x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

As can be seen from the graph below,
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? (3 marks)

Feature scaling is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data preprocessing stage to deal with varying values in the dataset. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values.

- Normalization is generally used 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)

$$VIF = \frac{1}{1 - R^2}$$

VIF - Variance Inflation Factor
The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity. If there is perfect correlation, then VIF = infinity. It 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.
Where R-1 is the R-square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables. If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and its R-squared value will be equal to 1. So, VIF = 1/(1-1) which gives VIF = 1/0 which results in "infinity"
The numerical value for VIF tells you (in decimal form) what percentage the variance (i.e. the standard error squared) is inflated for each coefficient. For example, a VIF of 1.9 tells you that the variance of a particular coefficient is 90% bigger than what you would expect if there was no multicollinearity — if there was no correlation with other predictors.
A **rule of thumb** for interpreting the variance inflation factor:

- 1 = not correlated.
- Between 1 and 5 = moderately correlated.
- Greater than 5 = highly correlated.

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

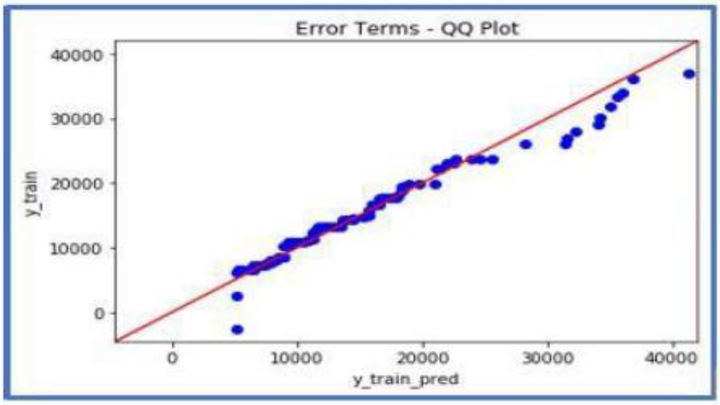
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. It is used to compare the shapes of distributions. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. 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 behaviour?

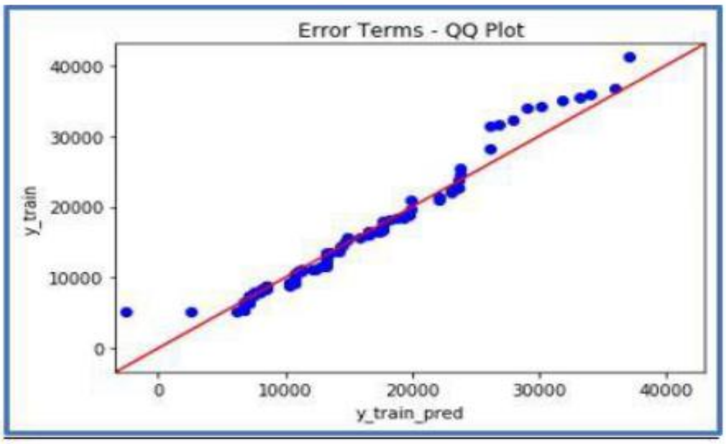
Below are the possible interpretations for two data sets using a Q-Q plot:

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.


Error Terms - QQ Plot
y_train
y_train_pred