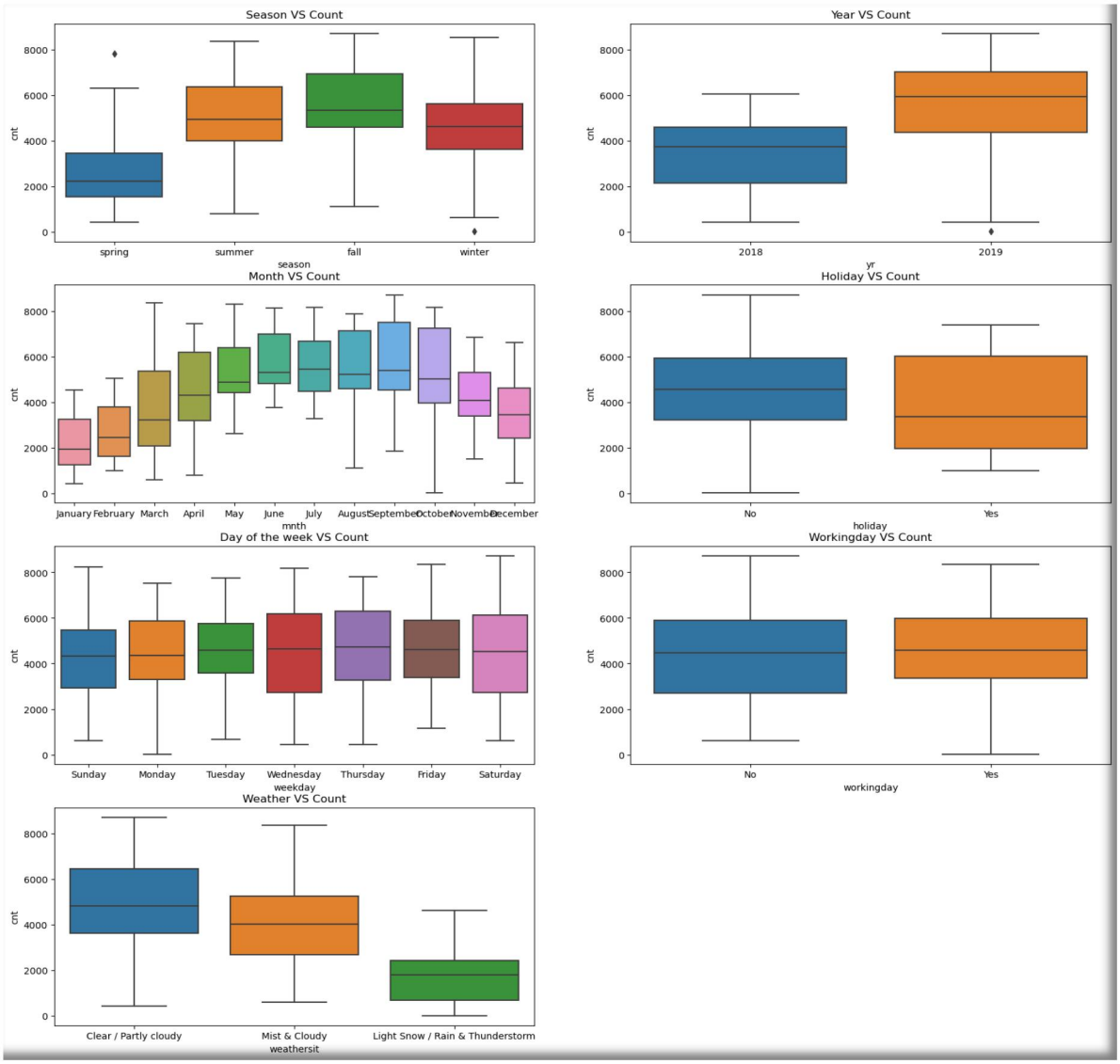
Assignment-based Subjective Questions

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



Categorical Variables:

Based on the above box-plots the observations are,

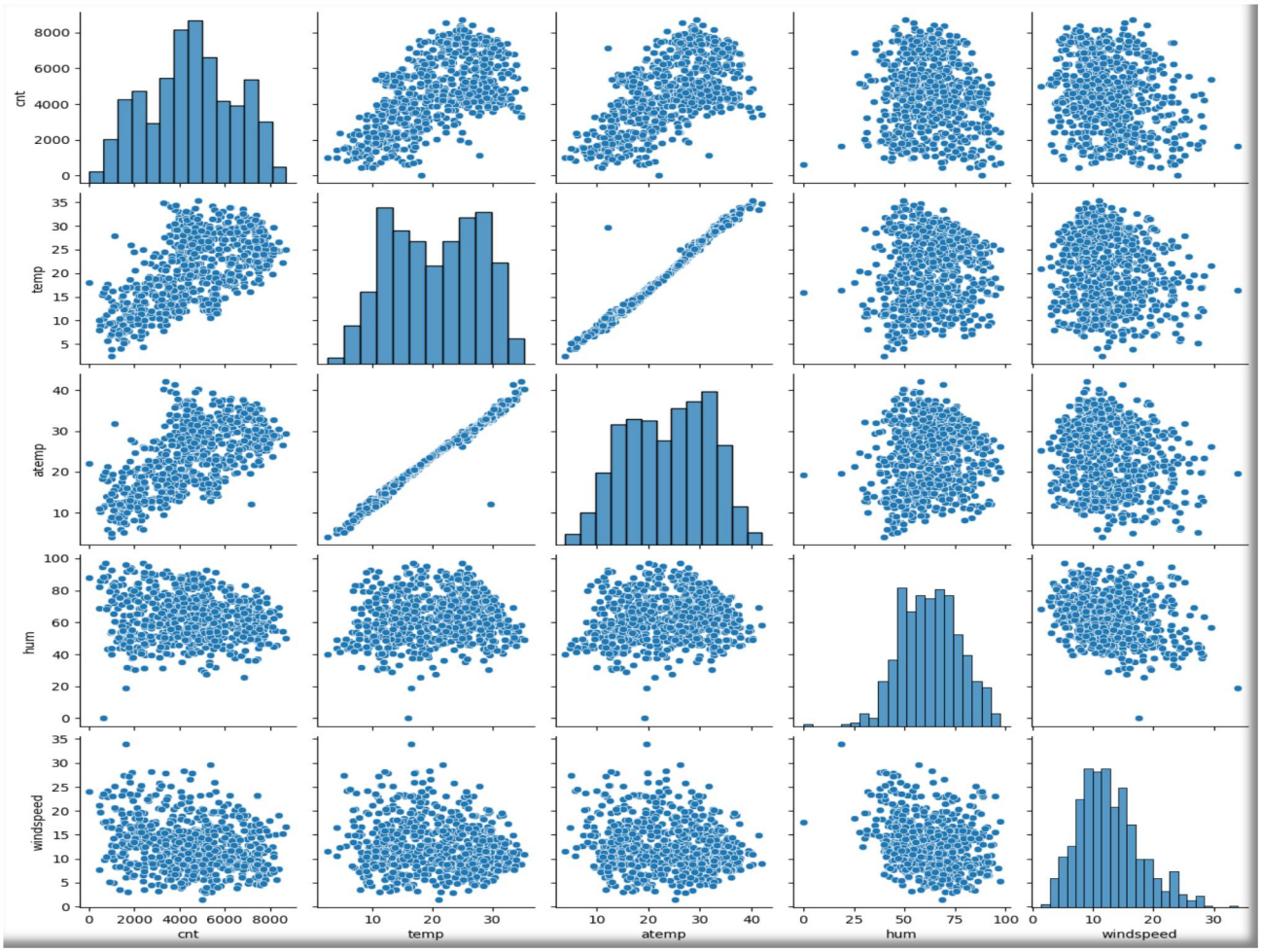
- The demand for rental bikes is high when the season is fall and during spring season the demand is very low compared to seasons summer and winter.
- The demand for rental bikes gradually increases from Jan to June hits peak during July and then gradually decreases from August to December.
- Whether the day is a holiday or not, weekend or weekday the demand doesn't get affected much by these factors.
- The demand for rental bikes is high when there is a clear weather and it is very low during bad weathers (snow or rain).

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

drop_first = True : For 'n' categorical variables 'n-1' dummy variables are created by dropping the first column.

If we set drop_first = False for 'n' categorical variables 'n' dummy variables will be created, it will lead to increase in multicollinearity among the predictor variables as the predictor themselves are correlated which will lead to redundant or insignificant predictor variable and so we remove the first column and reduce the number of insignificant predictor variables.

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

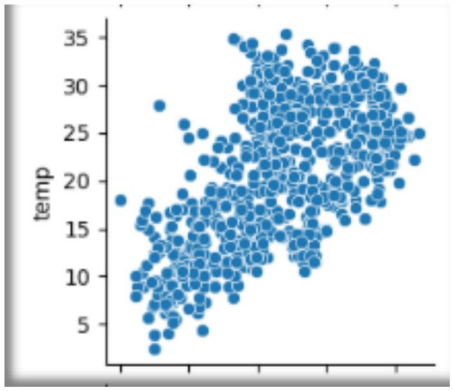


Numeric Variables:

Based on the above pair-plots the observations are,

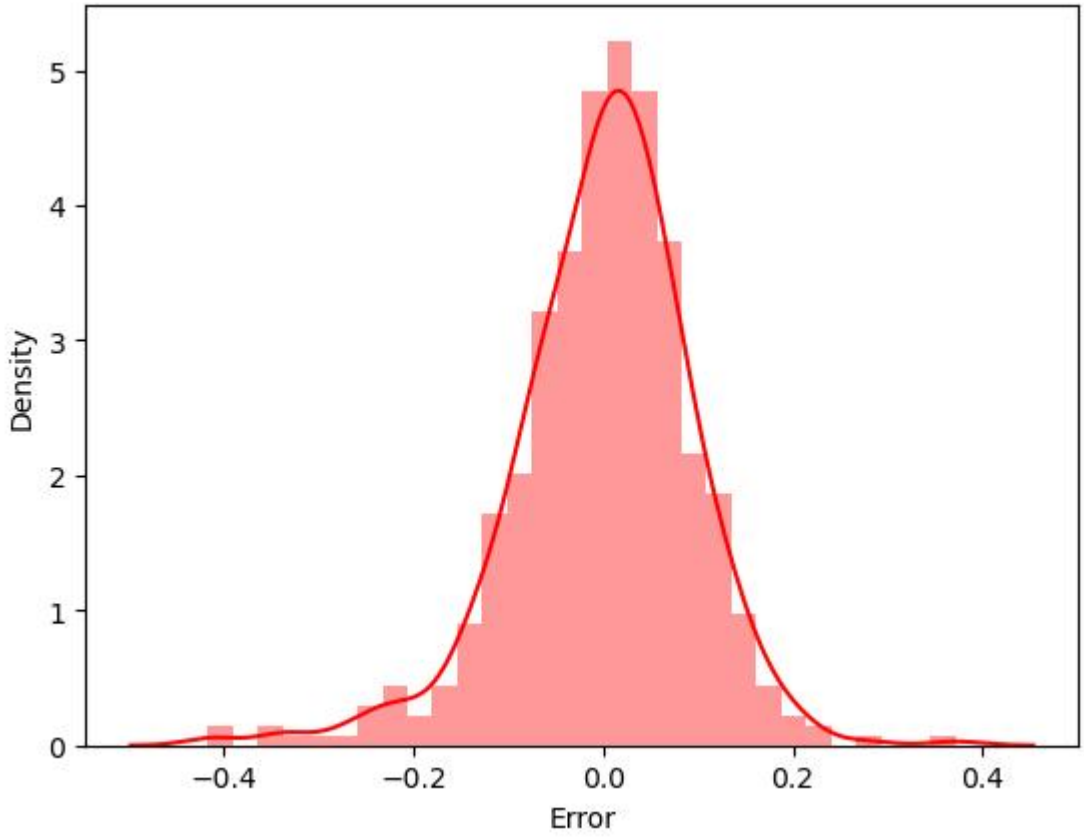
- There is a positive correlation between count of bikes and temperature.
- Temperature looks like a strong predictor of count of bikes.

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



- Temperature is a strong predictor variable to determine the demand of rental bikes.

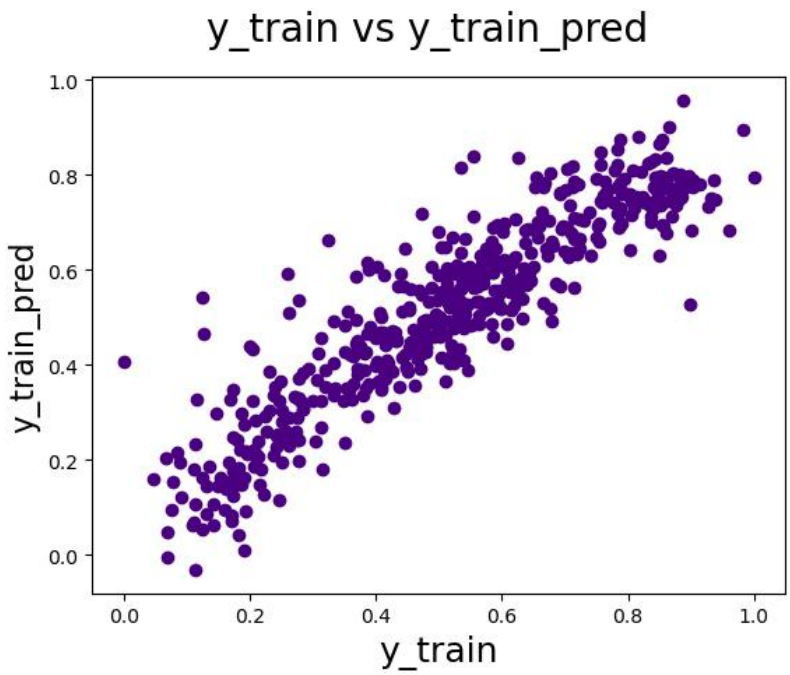
Residual Analysis on Training data-set:



The above distribution plot indicates that

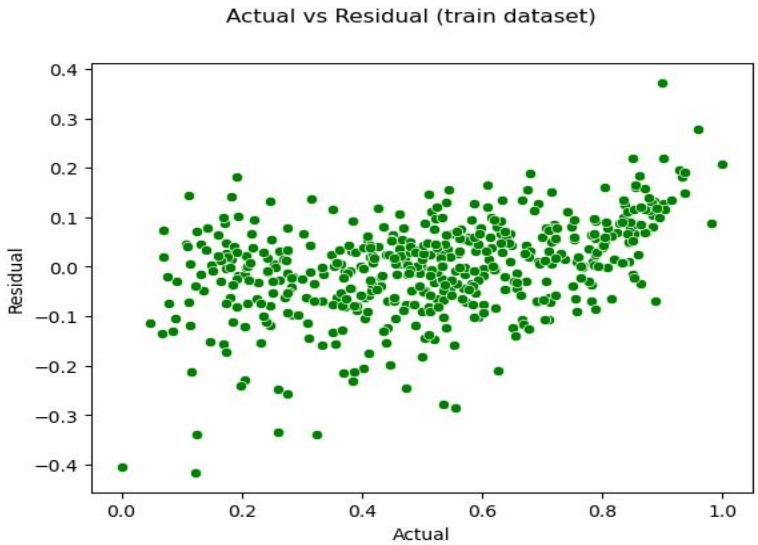
- The error terms are normally distributed.
- The mean of the error term is centered at 0.

Homoscedastic Model:



- The variance of the residual/error term, in this model is constant. That is, the error term does not vary much as the value of the predictor variable changes.
- Predicted and actual values in training set follows linear regression.
- There is not much deviation with respect to the error terms.

Pattern of error terms in training data set:



- The above scatter plot indicates that error terms doesn't follow any pattern (independent of each other) and it is randomly distributed which means the model is best fit.

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

```
lr_model_10.params
```

```
const                                 0.069762
yr                                    0.231336
holiday                              -0.099650
temp                                  0.539011
summer                                0.096809
winter                                0.144914
August                                0.058430
September                             0.123721
Mist & Cloudy                        -0.079552
 Light Snow / Rain & Thunderstorm    -0.296388
dtype: float64
```

Based on the co-efficient of the predictor variables from the best fit model, top 3 features contributing significantly towards explaining the demand of shared bikes are

- **Temperature:** The demand for Rental bikes increases when there is an increase in temperature.
- **Year :** Year on Year the demand for rental bikes increases ie. 2018 < 2019.
- **Light snow / Rain & Thunderstorm:** The demand for Rental bikes is negatively impacted when the weather is bad ie rainy or snowy condition.

General Subjective Questions:

1. Explain the linear regression algorithm in detail.

- Linear regression is a statistical approach in modeling the relation between a target
- variable (dependent variable) and predictors(independent variable).
- The goal of the algorithm is to find the best linear equation that can predict the value of the dependent variable based on the independent variables.
- The equation provides a straight line or plane that represents the relationship between the target and predictor variables.

There are two types of linear regression ,

✧ **Simple Linear Regression:**

In this we will be predicting the target variable with only one predictor variable. The equation for simple linear regression is:

$y=\beta_0+\beta_1 X$

where:
y - Target variable
X - Predictor variable
β0 - Intercept
β1 - Slope

✧ **Multiple Linear Regression:**

In this we will be predicting the target variable with multiple predictor variable (more than 1). The equation for multiple linear regression is:

$$y=\beta_0+\beta_1X+\beta_2X+\ldots\ldots\ldots\beta_nX$$

where:

Y - Target variable
X1, X2, …, Xp - Predictor variables
β0 - Intercept
β1, β2, …, βn- Slope

Assumptions in linear regression model:

✧ **Linear Relationship:** Assumption that there is linear relationship between the target and predictor variables.

✧ **Error terms normality:** The error terms (residual analysis) should be normally distributed with mean centered at zero.

✧ **Error terms are independent:** Error terms should not follow any pattern (independent of each other) and it should be randomly distributed so that the model will be a best fit model.

✧ **Homoscedasticity**: The variance of the residual/error term, in this model is constant. That is, the error term does not vary much as the value of the predictor variable changes.

2. Explain the Anscombe's quartet in detail?

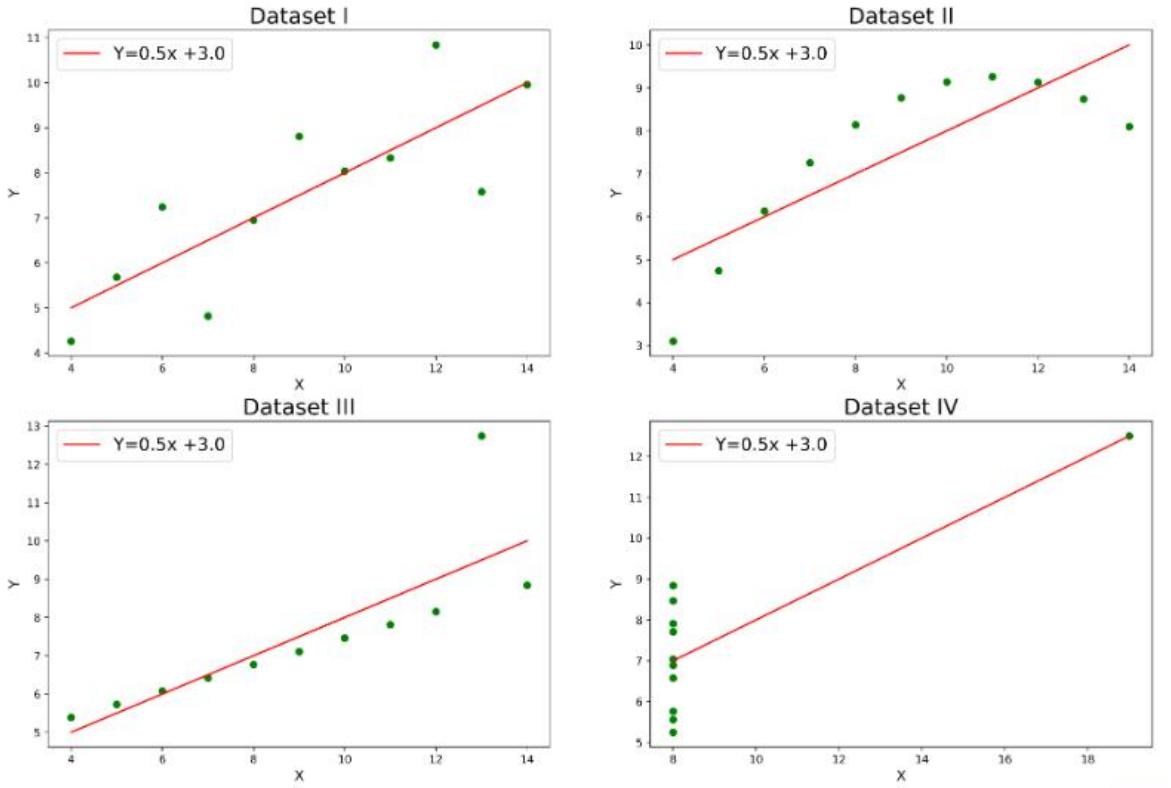
➢ Anscombe's quartet is used to demonstrate the importance of visualizing data-set.
➢ It comprises of a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

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

From the result of statistic perspective, the usual conception would be that the data distribution is almost similar for all the 4 data sets as the statistical summary is similar.
However, when the data is actually plotted with scatter plot, each dataset generates different kind of plot.

Output:

	I	II	III	IV
Mean_x	9.000000	9.000000	9.000000	9.000000
Variance_x	11.000000	11.000000	11.000000	11.000000
Mean_y	7.500909	7.500909	7.500000	7.500909
Variance_y	4.127269	4.127629	4.122620	4.123249
Correlation	0.816421	0.816237	0.816287	0.816521
Linear Regression slope	0.500091	0.500000	0.499727	0.499909
Linear Regression intercept	3.000091	3.000909	3.002455	3.001727



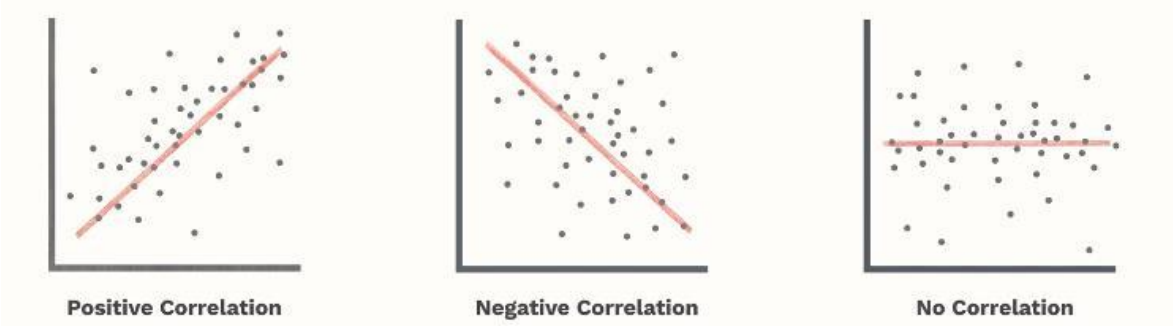
From the plot, we can say:
Dataset -I fits the linear regression
Dateset -II shows the non-linear pattern.
Dataset -III shows linear relationship for all data points except one which is an outlier in the data
Dataset -IV has one outlier with x=8 which disturbs the model
So from the four datasets, we can only use linear regression on Dataset-I.

3. What is Pearson's R?

It is the measure of linear correlation between the variables, used to measure how strong a relationship is between variables.

The value ranges from -1 to 1

Pearson Correlation Coefficient (r) Range	Type of Correlation	Description of Relationship
$0 < r \leq 1$	Positive	An increase in one variable associates with an increase in the other.
$r = 0$	None	No discernible relationship between the changes in both variables.
$-1 \leq r < 0$	Negative	An increase in one variable associates with a decrease in the other.



Pearson's R equation is

$$r = n(\Sigma xy) - (\Sigma x)(\Sigma y) / \sqrt{[n\Sigma x^2-(\Sigma x)^2][n\Sigma y^2-(\Sigma y)^2}$$

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

- Scaling is a process which standardize all the independent values in order to bring all the data in a same scale. It helps in handling different units and higher numerical values and convert into lower range.
- It helps to do calculations in algorithms very quickly and also it takes less time to train the model.
- If scaling is not done, then the algorithm will weigh greater values as high and smaller values as low since the algorithm works on numbers and not on units.
- Scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc
- There are two types of scaling,

Normalization	Standardization
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is really affected by outliers.	It is much less affected by outliers.
Scikit-Learn provides a transformer called `MinMaxScaler` for Normalization.	Scikit-Learn provides a transformer called `StandardScaler` for standardization.
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.
It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

Normalization or Min-Max Scaling :

```
X_new = (X - X_min)/(X_max - X_min)
```

Standardization:

```
X_new = (X - mean)/Std
```

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

VIF - Variance Inflation Factor is a measure of the amount of multicollinearity in regression analysis

$$VIF_i = \frac{1}{1 - R_i^2}$$

- Infinite VIF value means that a variable is perfectly expressed by linear combination of other variables. ie. Perfect correlation between two independent variables
- When R2 score is 1, denominator in VIF becomes 0 and value of VIF will become infinite.
- To solve this problem we need to drop one of the variables from the data-set which is causing this perfect multicollinearity.

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

- Q-Q plots are also known as Quantile-Quantile plots is a probability plot for comparing two probability distributions by plotting their quantiles against each other

- QQ plots is very useful to determine

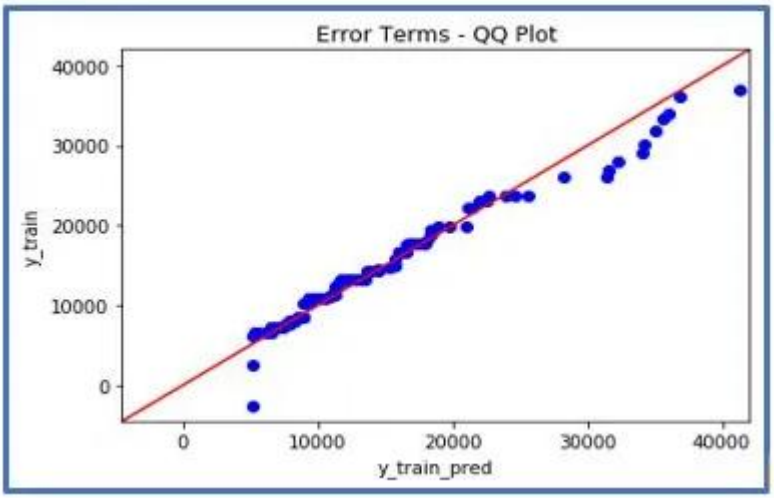
i. If two populations are of the same distribution
ii. If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it's met using this.

Interpretation:

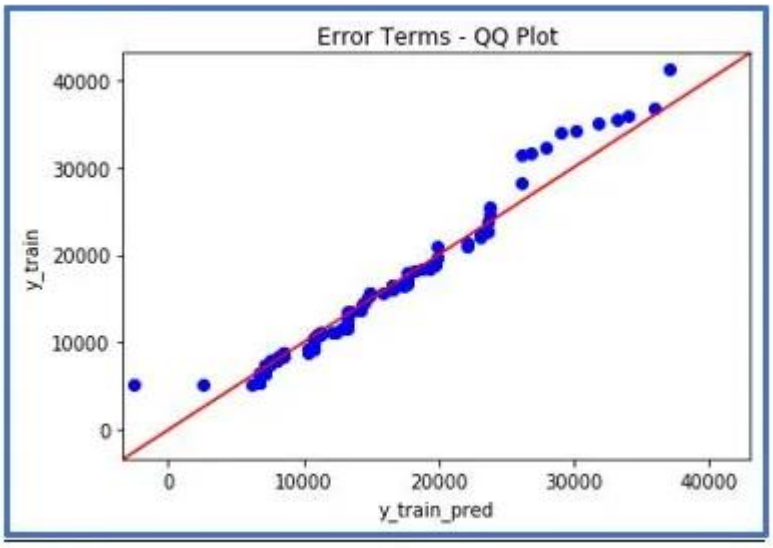
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. Below are the possible interpretations for two data sets.

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.



d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis