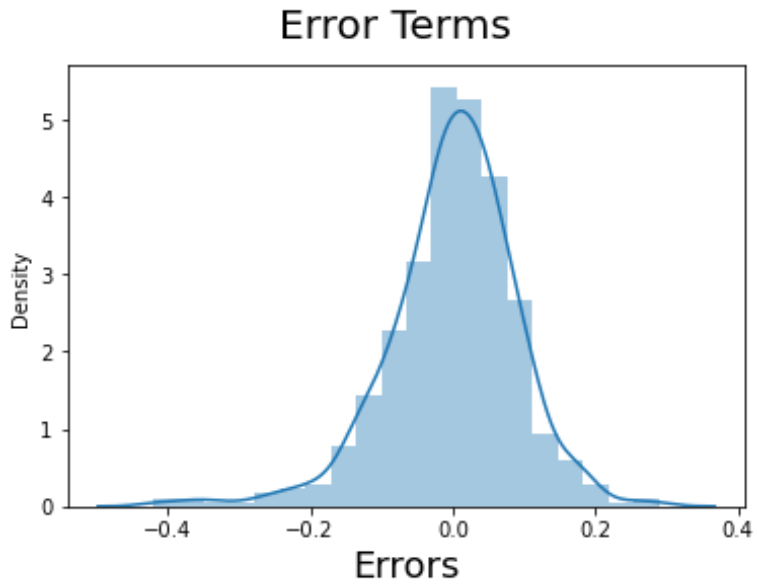
## Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**
   **Answer:** We observe -
   - The demand for bike increases from month of April to October. Month of January is less in demand.
   - The demand for bike throughout the weekdays is almost similar.
   - In Summer and Fall season, there is very high demand for bikes. While, it is very less in Spring season
   - Clear Weather have the high demand for bikes. While, Heavy/Rain have the very less demand.
   - The demand for bike is high for year 2019 as compared to year 2018.
2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**
   **Answer:** The main reason to use the **drop_first = True** is to remove the extra column created during creation of dummy variable. It helps in reducing the correlation between the dummy variables. It also reduces redundancy as the data can be represented by the remaining variables created by dummy variable method.

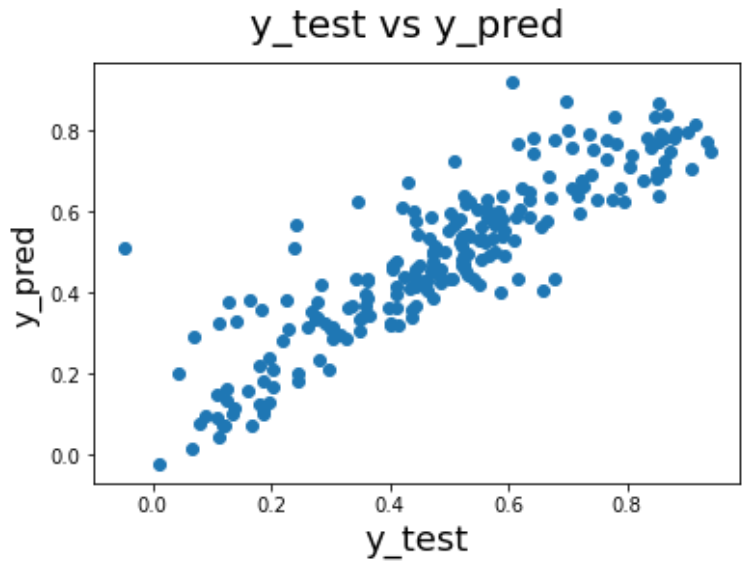
**3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

**Answer:** The variable having highest correlation with the target variables is 'temp'.

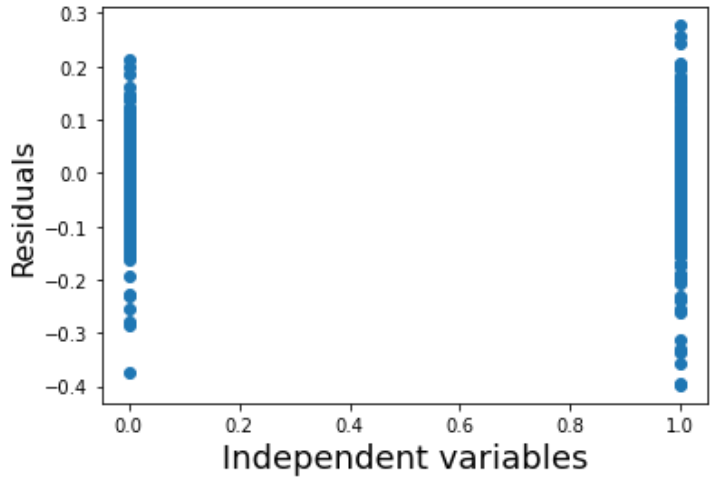
3. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**
   **Answer:** To validate the Assumptions of Linear Regression after building model, I check for the **Normal Distribution of Error Terms**. Below is the distribution plot, which shows the Normal Distribution with mean at zero.




The Error Term satisfies to have reasonably constant variance (homoscedasticity)




It shows that the line fits the model well.




Checked assumption of homoscedasticity and autocorrelation.

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

**Answer:** The top three feature contributing significantly towards explaining the demand of shared bikes are –

a. Temperature (0.5449)
b. Year (0.2403)
c. Month – September (0.1009)

## General Subjective Questions

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

**Answer:** Linear Regression is a technique which is used to model the linear relationship between the independent variable and the dependent variable (target variable). It is mandatory to keep the Predictor and target variable be Numerical values.

Linear Regression is of two types:

(a) Simple Linear Regression: It has only one independent variable and one dependent variable (target variable). The model is defined on these singular variables.

$$y = b_o + b_1 x$$

where $b_o$ is the intercept, $b_1$ is coefficient or slope, x is the independent variable and y is the dependent variable.

(b) Multi Linear Regression: It has more than one independent variable and one dependent variable (target variable). The model is created on basis of these variables.

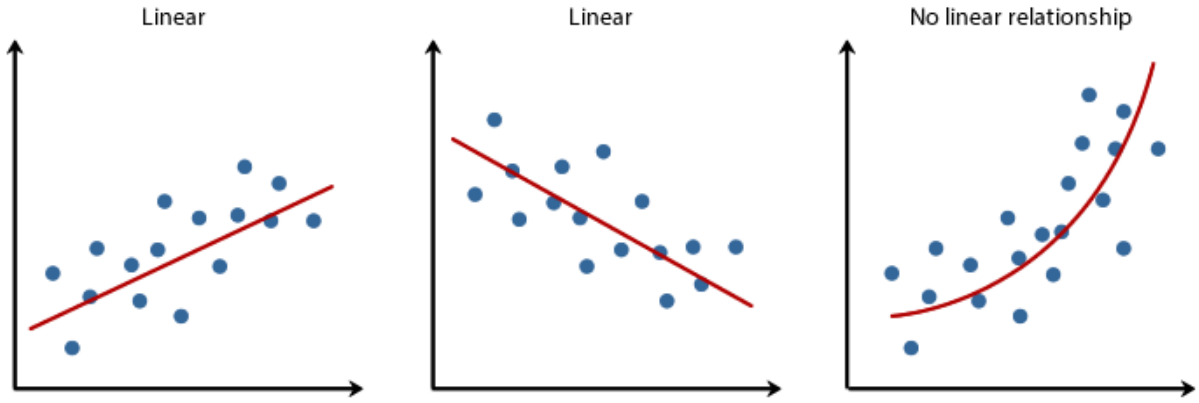
$$y = b_o + b_1 x_1 + b_2 x_2 + b_3 x_3 \ldots . + b_n x_n$$

where $b_o$ is the intercept, $b_1, b_2, b_3, b_4 \ldots, b_n$ are coefficients or slopes of the independent variables $x_1, x_2, x_3, x_4 \ldots, x_n$ and y is the dependent variable.

The main motive of Linear regression model is to fit the line and find the optimal values of coefficients and intercept in such a way that the error is minimized.

There are some assumptions used in Linear Regression Model:

1. Linear Relationship: It defines the relationship between the independent and target variable is linear.




2. Independence or No multicollinearity: It defines that the variables must be independent of each other or there should be no Multicollinearity.
3. Normal Distribution: The variables should define Normality.
4. Homoscedasticity: It define that the variance of the error term should be constant.

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

**Answer:** Anscombe's quartet was originally constructed by the statistician named Francis Anscombe in 1973. He constructed it to illustrate the importance of plotting the graphs before analysing and modelling.

He took four datasets which have same statistical observations (variance & mean) of all x,y values in all four datasets. And when he plot these datasets, those plots look very different from each other.

This experiment tells us the importance of visualising the datasets before using any algorithms on them and modelling. These plots tell us about the distribution of the samples which is going to help find anomalies in the datasets.

We can define it by example as follows:

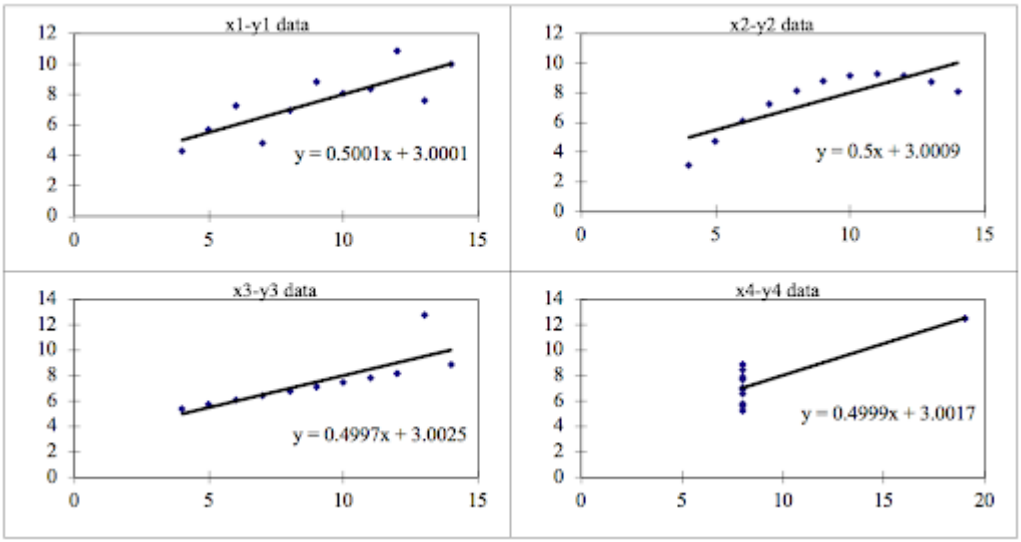
These are four datasets-

| Anscombe's Data | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Observation | x1 | y1 | x2 | y2 | x3 | y3 | x4 | y4 |
| 1 | 10 | 8.04 | 10 | 9.14 | 10 | 7.46 | 8 | 6.58 |
| 2 | 8 | 6.95 | 8 | 8.14 | 8 | 6.77 | 8 | 5.76 |
| 3 | 13 | 7.58 | 13 | 8.74 | 13 | 12.74 | 8 | 7.71 |
| 4 | 9 | 8.81 | 9 | 8.77 | 9 | 7.11 | 8 | 8.84 |
| 5 | 11 | 8.33 | 11 | 9.26 | 11 | 7.81 | 8 | 8.47 |
| 6 | 14 | 9.96 | 14 | 8.1 | 14 | 8.84 | 8 | 7.04 |
| 7 | 6 | 7.24 | 6 | 6.13 | 6 | 6.08 | 8 | 5.25 |
| 8 | 4 | 4.26 | 4 | 3.1 | 4 | 5.39 | 19 | 12.5 |
| 9 | 12 | 10.84 | 12 | 9.13 | 12 | 8.15 | 8 | 5.56 |
| 10 | 7 | 4.82 | 7 | 7.26 | 7 | 6.42 | 8 | 7.91 |
| 11 | 5 | 5.68 | 5 | 4.74 | 5 | 5.73 | 8 | 6.89 |

Now we define the Statistical behaviour of the datasets-

| Anscombe's Data | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Observation | x1 | y1 | x2 | y2 | x3 | y3 | x4 | y4 |
| 1 | 10 | 8.04 | 10 | 9.14 | 10 | 7.46 | 8 | 6.58 |
| 2 | 8 | 6.95 | 8 | 8.14 | 8 | 6.77 | 8 | 5.76 |
| 3 | 13 | 7.58 | 13 | 8.74 | 13 | 12.74 | 8 | 7.71 |
| 4 | 9 | 8.81 | 9 | 8.77 | 9 | 7.11 | 8 | 8.84 |
| 5 | 11 | 8.33 | 11 | 9.26 | 11 | 7.81 | 8 | 8.47 |
| 6 | 14 | 9.96 | 14 | 8.1 | 14 | 8.84 | 8 | 7.04 |
| 7 | 6 | 7.24 | 6 | 6.13 | 6 | 6.08 | 8 | 5.25 |
| 8 | 4 | 4.26 | 4 | 3.1 | 4 | 5.39 | 19 | 12.5 |
| 9 | 12 | 10.84 | 12 | 9.13 | 12 | 8.15 | 8 | 5.56 |
| 10 | 7 | 4.82 | 7 | 7.26 | 7 | 6.42 | 8 | 7.91 |
| 11 | 5 | 5.68 | 5 | 4.74 | 5 | 5.73 | 8 | 6.89 |
| | | | Summary Statistics | | | | | |
| N | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 |
| mean | 9.00 | 7.50 | 9.00 | 7.500909 | 9.00 | 7.50 | 9.00 | 7.50 |
| SD | 3.16 | 1.94 | 3.16 | 1.94 | 3.16 | 1.94 | 3.16 | 1.94 |
| r | 0.82 | | 0.82 | | 0.82 | | 0.82 | |

The Statistical observations seems pretty similar for four datasets.



Here, you can see that the plot for every dataset shows different plotting. This is not interpretable by the algorithm.

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

**Answer:** Pearson's R, which is also known as Pearson product moment correlation coefficient or bivariate correlation. It is a statistical method to measure the linear correlation between two variables.

There is an issue in Pearson R, that it cannot state the difference between the dependent variables and the target variables and also does not give any details about the slope of line. It only states about the relationship between the variables.

It is also defined as the ration between the covariance of two variables and product of their standard deviations.

Formula for Population,

$$\rho_{X,Y} = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]}{\sqrt{\mathbb{E}[X^2] - (\mathbb{E}[X])^2}\,\sqrt{\mathbb{E}[Y^2] - (\mathbb{E}[Y])^2}}.$$

Formula for Sample,

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\,\sqrt{n\sum y_i^2 - (\sum y_i)^2}}.$$

Where:

- $n$ is sample size
- $x_i, y_i$ are the individual sample points indexed with *i*
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$

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

**Answer:** Scaling is a technique, which is used to standardize the independent variable or features. The basic idea is to keep the value in range for all the variables.

It is basically used during data pre-processing to maintain the ranges of the variables(independent).

Scaling is performed to standardized/normalize the data points or values in the variables to make it easy for model to learn and illustrate the good results. It also helps in fasten the calculations in the regression algorithm.

| Normalization | Standardization |
|---|---|
| It is also known as the Minimum and Maximum value Method | It is also known as Z-Score Normalization. |
| It is used when features are of different ranges/scales. | It ensures zero mean and unit standard deviation. |
| It ranges/scales the value between [0,1] or [-1,1]. | There is no bounded range for it. |
| Scikit-Learn Library provides a transformer known as MinMaxScaler for Normalization. | Scikit-Learn provide transformer known as StandardScaler for standardization. |

| | |
|---|---|
| In case of unknown distribution, it is very useful. | In case of Normal or Gaussian distribution, it is very helpful. |

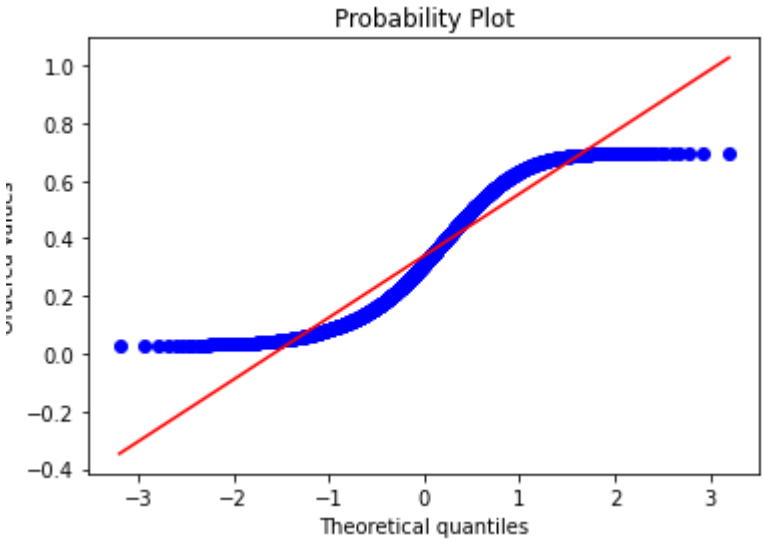
5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**

**Answer:** In case there is VIF = infinity. Then, it states that there is a strong correlation/Multicollinearity between the variables.
When there is a strong relationship then the value of R2 is 1, which makes 1/(1-R2) to infinity. The approach to solve this problem is to drop the features having perfect multicollinearity.
Another approach is to check for the independent variables and look for the duplicate rows. Then drop those rows.

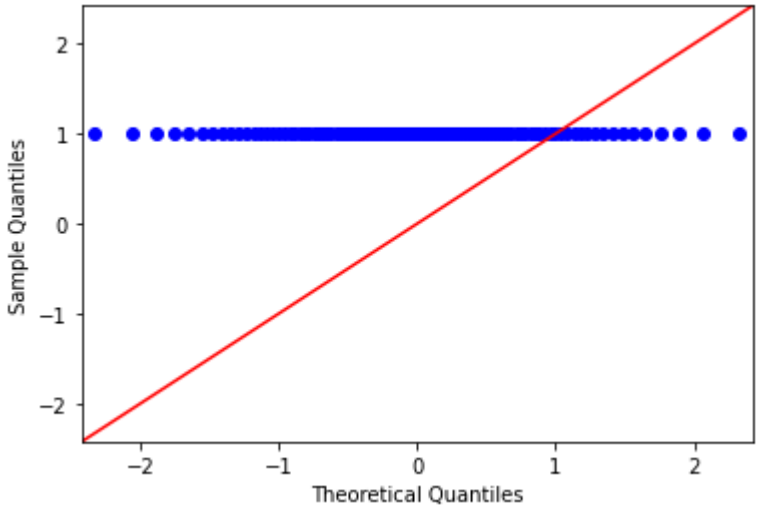
6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**

**Answer:** The Quantile-Quantile plot is graphical method, which is used for determining if two samples belong to same population or not. It is a plot of the quantiles one dataset to quantiles of another dataset.
There are two types of Q-Q plots:

1. For Left-tailed distribution:




2. For Uniform Distribution:




It is used for following reasons:

1. To check whether two samples are from same population or not.
2. To check if two samples have the same tail.
3. To check if two samples have same distribution shape.
4. To check if two samples have common location behaviour.

The importance of Q-Q plot in linear regression:

1. Q-Q plots are very useful in determining if two population are of same distribution.
2. Q-Q plots also helps in finding skewness of distribution.
3. It can also detect distributional aspects like Scale, change in symmetry, the outliers & shift in location.