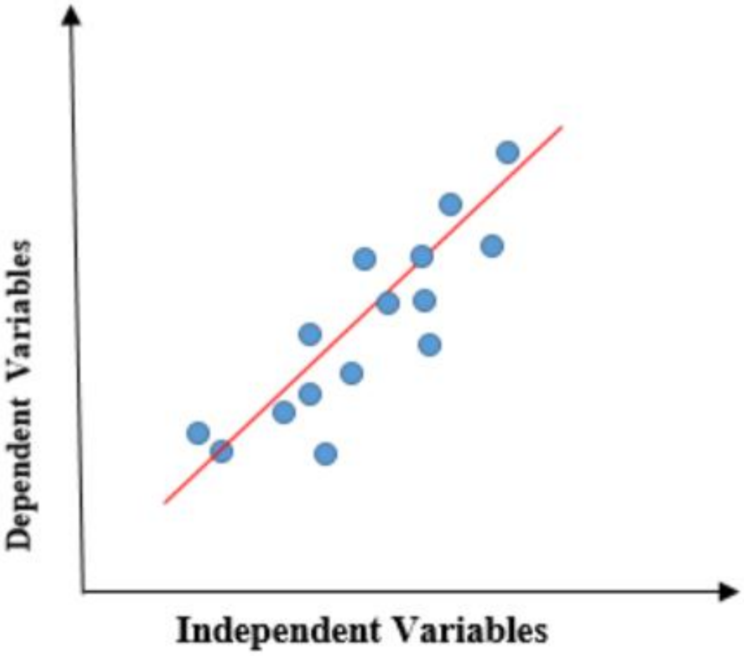
General Subjective Question

01. Explain the linear regression algorithm in detail?

Ans: Linear regression is one of the very basic forms of the machine learning where we train a model to predict the behaviour of our data based on some variables. In the case of linear regression as we can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

If there is a single input variable (x), such linear regression is called **simple linear regression**. And if there is more than one input variable, such linear regression is called **multiple linear regression**. The linear regression model gives a sloped straight line describing the relationship within the variables.



Mathematically, we can write a linear regression equation as:

y = mx + c

where c and m given by the formulas:

m(slobe)= (n∑xy – (∑x) (∑y))/(n∑x^2 – (∑x)^2)

c(intercept) = (n∑y- m (∑x))/n

Here, x and y are two variables on the regression line.

m = Slope of the line

c = y-intercept of the line

x = independent variable from dataset

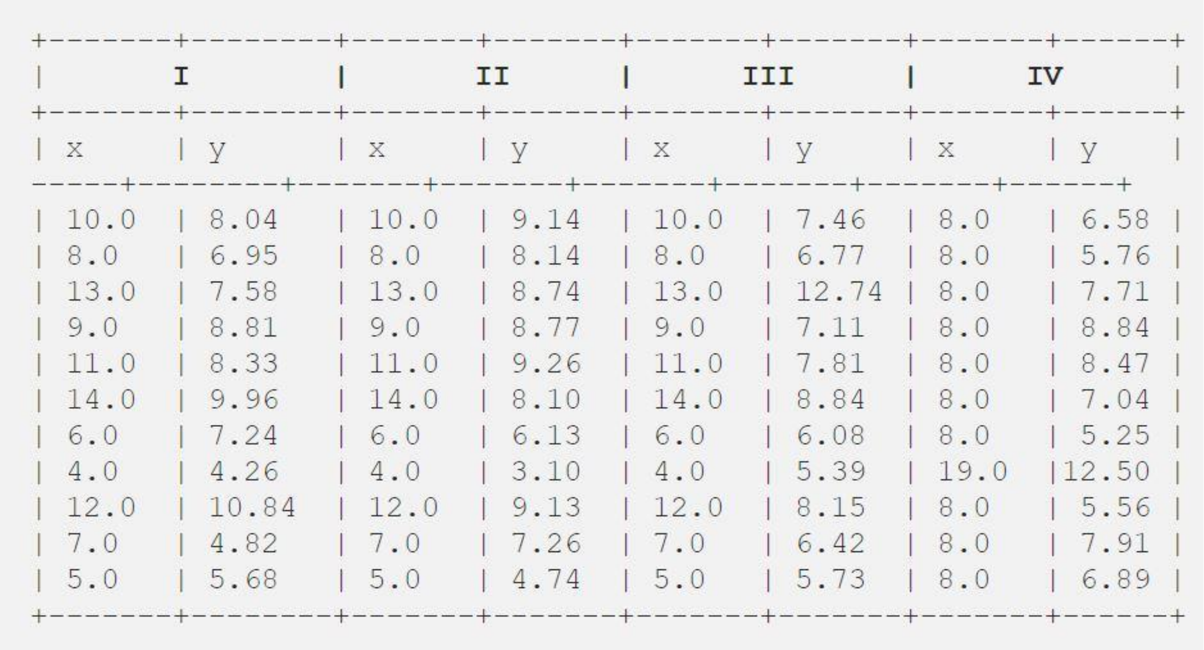
y = dependent variable from dataset

02. Explain the Anscombe’s quartet in detail?

Ans: Anscombe’s quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of the graphing data before analyzing it and the effect of outliers on statistical properties.

Simple Understanding:

Once Francis John “Frank” Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot these points. These 4 sets of 11 data-points are given below.



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

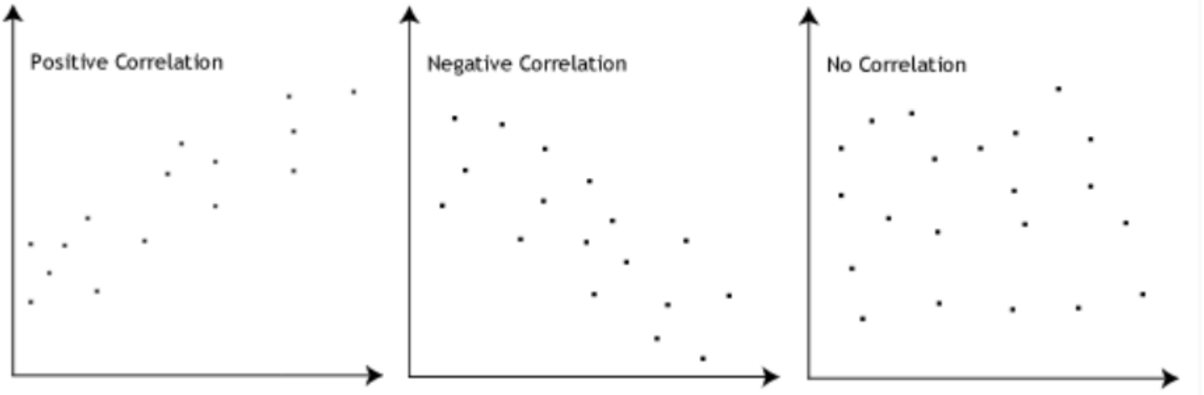
After that, the council analyzed them using only descriptive statistics and found the mean, standard deviation and correlation between x and y.

03. What is Pearson’s R?

Ans: In statistics, the Person correlation coefficient (PCC), also referred to as Person’s *r*, the Person product-moment correlation coefficient (PPMCC) or the bivariate correlation is a measure of linear correlation between two sets of the data. It’s the covariance of two variables, divided by the product of their standard deviation; thus it’s essentially a normalised measurement of the covariance such that the result always has a value between -1 and 1.

The Person’s correlation coefficient varies between -1 and +1 where:

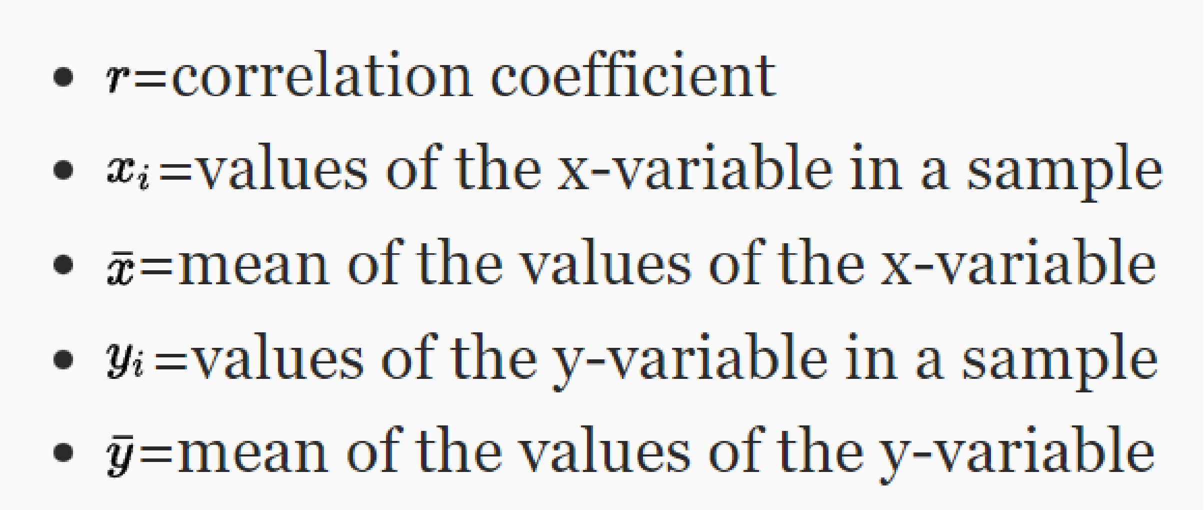
- r =1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e. both variables tend to change in different direction)
- r = 0 means there is a weak association
- r > 0 < 5 means there is a week association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association



Person r formula:

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

Here



- 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

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

Ans:

Scaling mean It is a step of data pre-processing which is applied to independent variables to normalize the data within a particular range. It also helps in speed up the calculations in an algorithm

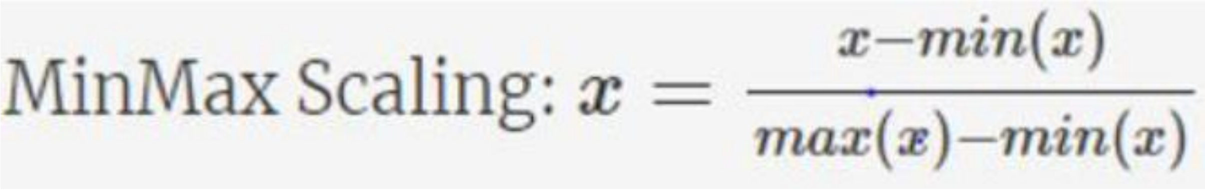
Reason to perform the scaling:

Most of the times, the 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. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that **scaling just affects the coefficients** and none of the other parameters like **t-statistic, F-statistic, p-values, R-squared**, etc.

Normalization/Min-Max Scaling:

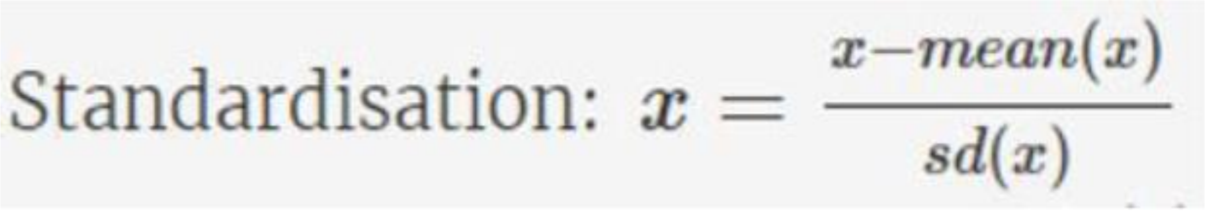
- ✓ It brings all of the data in the range of 0 and 1. **sklearn.preprocessing.MinMaxScaler** helps to implement normalization in python.



$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

- ✓ Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean **(μ)** zero and standard deviation one (**σ**).

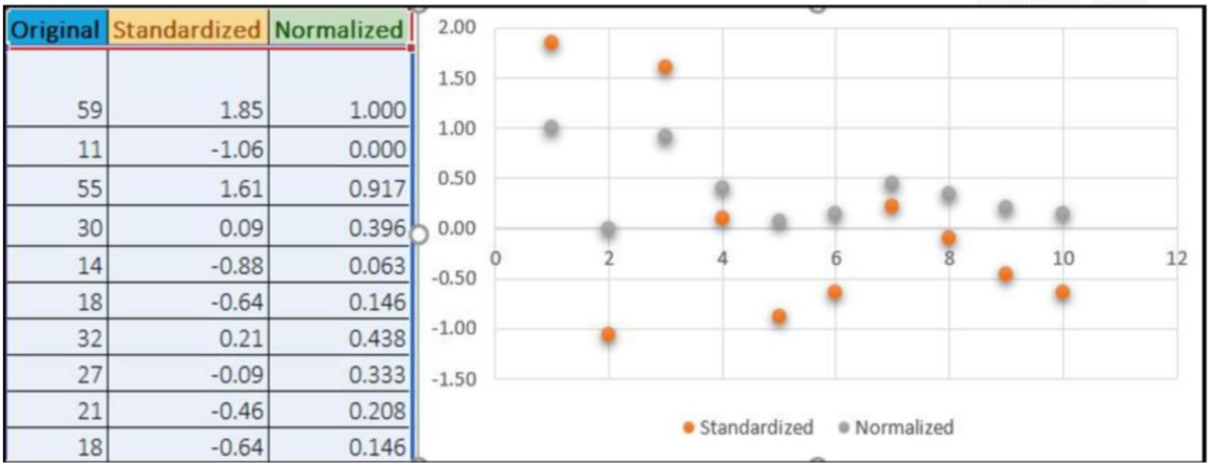


$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

- ✓ **sklearn.preprocessing.scale** helps to implement standardization in python.
- ✓ One disadvantage of normalization over standardization is that it **loses** some information in the data, especially about **outliers**.

Example:

Below shows example of Standardized and Normalized scaling on original values:



Original	Standardized	Normalized
59	1.85	1.000
11	-1.06	0.000
55	1.61	0.917
30	0.09	0.396
14	-0.88	0.063
18	-0.64	0.146
32	0.21	0.438
27	-0.09	0.333
21	-0.46	0.208
18	-0.64	0.146

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

Ans

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.

To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

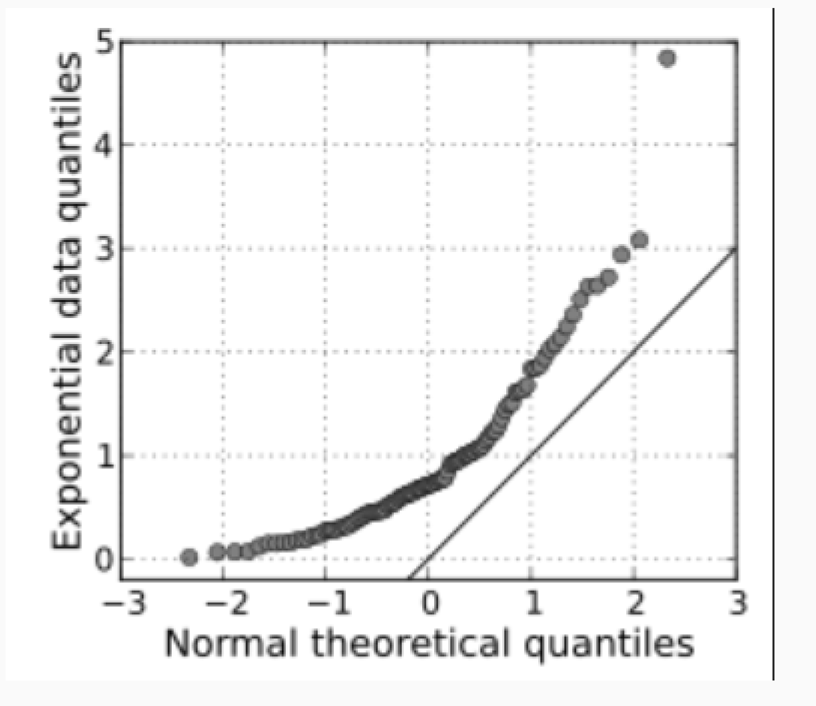
An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Ans:

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. 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.

A Q Q plot showing the 45 degree reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

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.

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

   Ans: Temp and month, seasons are more dependant variables as per data observation and final model results

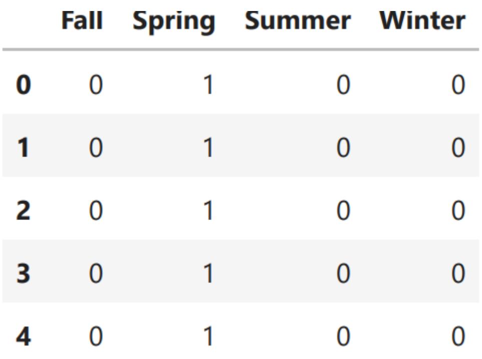
2. Why is it important to use drop_first=True during dummy variable creation?

   Ans

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

   # creating dummy variables for season

   season = pd.get_dummies(bikedata['season']) season.head()



	Fall	Spring	Summer	Winter
0	0	1	0	0
1	0	1	0	0
2	0	1	0	0
3	0	1	0	0
4	0	1	0	0

Now, we don't need four columns. You can drop the Fall column, as the type of season can be identified with just the last three columns where –

- `100` will correspond to `Spring`
- `010` will correspond to `Summer`
- `001` will correspond to `Winter`

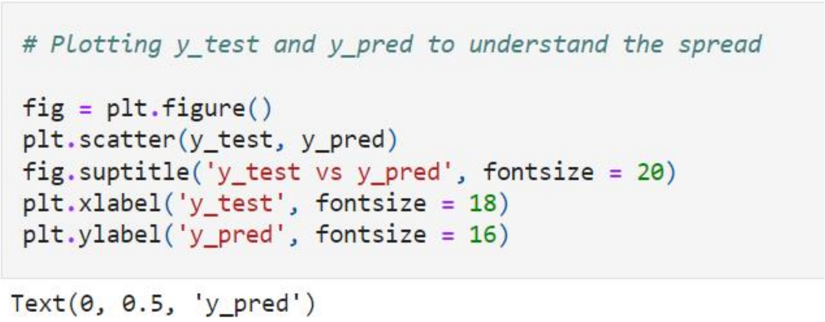
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans: As per Pairplot graph we can observe the “temp” is highest correlation with CNT dependant variables

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

Ans: if we observe the ml notebook, we show in the scatter plot y prediction value against the y test values

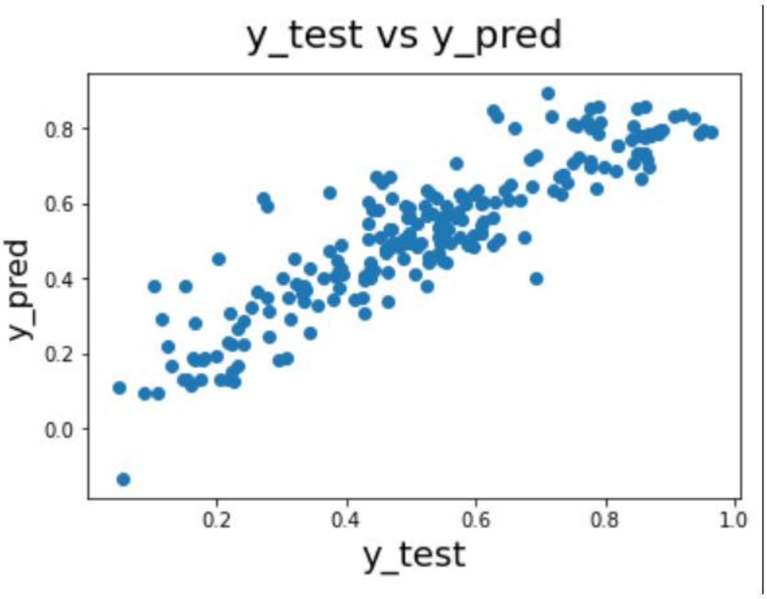
```
fig = plt.figure()
plt.scatter(y_test, y_pred)
fig.suptitle('y_test vs y_pred', fontsize = 20)
plt.xlabel('y_test', fontsize = 18)
plt.ylabel('y_pred', fontsize = 16)
```



```
# Plotting y_test and y_pred to understand the spread

fig = plt.figure()
plt.scatter(y_test, y_pred)
fig.suptitle('y_test vs y_pred', fontsize = 20)
plt.xlabel('y_test', fontsize = 18)
plt.ylabel('y_pred', fontsize = 16)
```

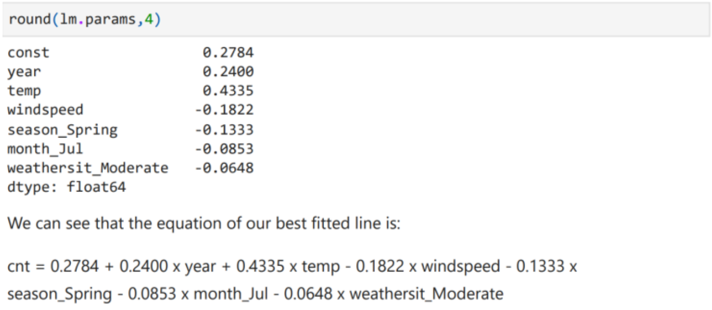
```
Text(0, 0.5, 'y_pred')
```



It’s represent the linear line as per data point

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

Ans: Year, Temp and windspeed more dominating features for contributing significatly towards the Bike share



```
round(lm.params,4)
```

```
const                  0.2784
year                   0.2400
temp                   0.4335
windspeed             -0.1822
season_Spring         -0.1333
month_Jul             -0.0853
weathersit_Moderate   -0.0648
dtype: float64
```

We can see that the equation of our best fitted line is:

cnt = 0.2784 + 0.2400 x year + 0.4335 x temp - 0.1822 x windspeed - 0.1333 x season_Spring - 0.0853 x month_Jul - 0.0648 x weathersit_Moderate

Year, Temp and windspeed more dominating features for contributing significatly towards the Bike share