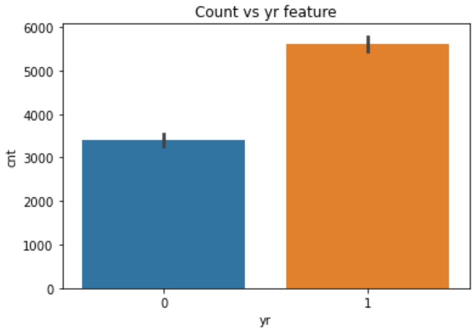
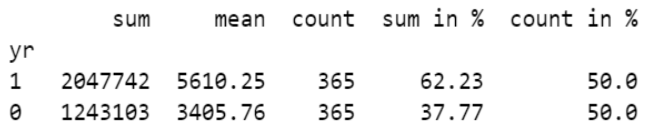
Assignment-based Subjective Questions

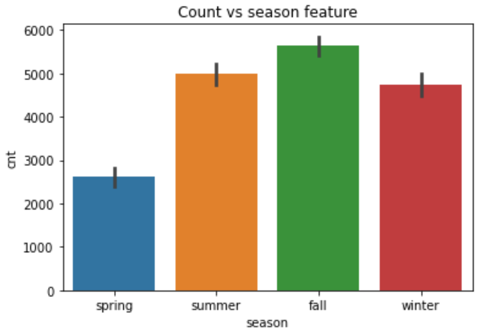
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

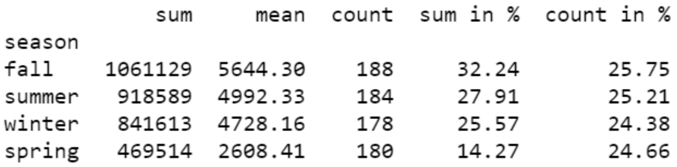




```
        sum     mean  count  sum in %  count in %
yr
1   2047742  5610.25    365     62.23        50.0
0   1243103  3405.76    365     37.77        50.0
```

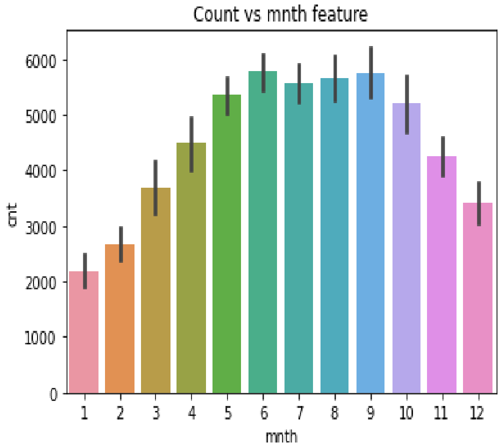
Since we know year 0,1 implies 2018 and 2019 respectively, it can be concluded that there is an increasing trend with increase in year. Hence, yr could a good predictor variable of total count.

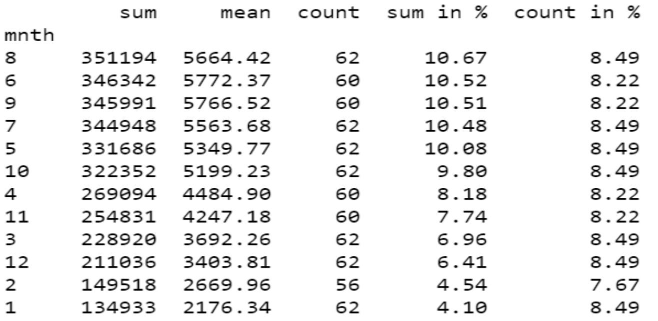




```
            sum     mean  count  sum in %  count in %
season
fall    1061129  5644.30    188     32.24       25.75
summer   918589  4992.33    184     27.91       25.21
winter   841613  4728.16    178     25.57       24.38
spring   469514  2608.41    180     14.27       24.66
```

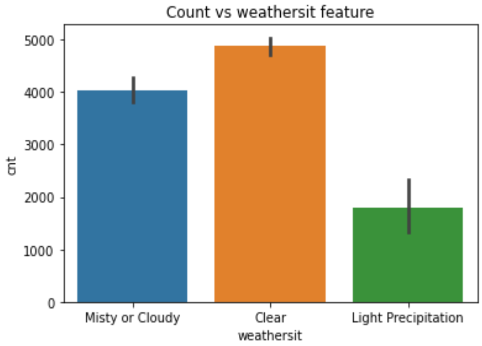
Analysing the four seasons, we can see that there is a significant increase in bike rentals in summer and fall seasons. Since there is a visible influence, season could be a potential predictor of the total count.

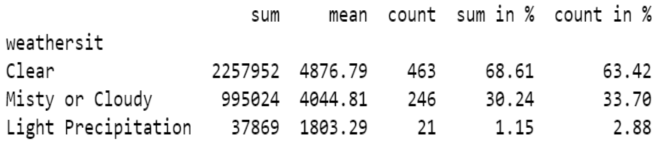




```
         sum     mean  count  sum in %  count in %
mnth
8     351194  5664.42     62     10.67        8.49
6     346342  5772.37     60     10.52        8.22
9     345991  5766.52     60     10.51        8.22
7     344948  5563.68     62     10.48        8.49
5     331686  5349.77     62     10.08        8.49
10    322352  5199.23     62      9.80        8.49
4     269094  4484.90     60      8.18        8.22
11    254831  4247.18     60      7.74        8.22
3     228920  3692.26     62      6.96        8.49
12    211036  3403.81     62      6.41        8.49
2     149518  2669.96     56      4.54        7.67
1     134933  2176.34     62      4.10        8.49
```

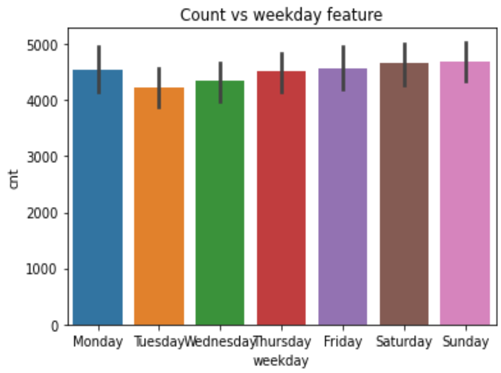
Analysing the 12 months, a steep increase in bike rentals can be observed from March to September. Hence, month might be a possible influencing variable of the total bike rentals.

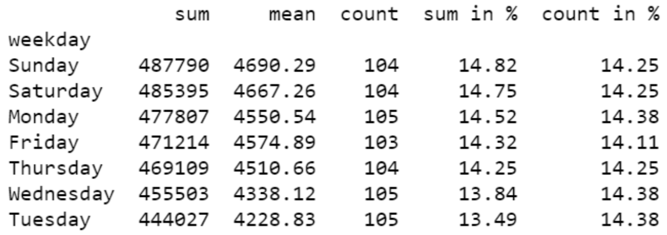




```
                         sum     mean  count  sum in %  count in %
weathersit
Clear                2257952  4876.79    463     68.61       63.42
Misty or Cloudy       995024  4044.81    246     30.24       33.70
Light Precipitation    37869  1803.29     21      1.15        2.88
```

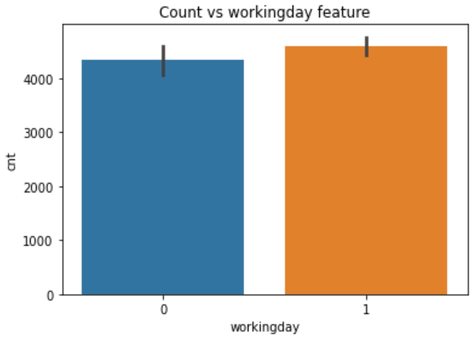
A clear increase in bike rentals can be seen in clear weather condition, indicating that weather might be a potential predictor of bike rental count.

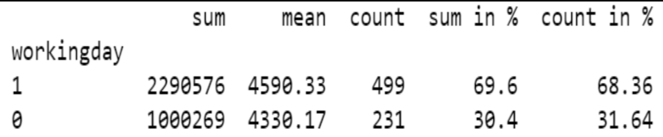




```
                sum      mean   count   sum in %   count in %
weekday
Sunday       487790   4690.29     104      14.82        14.25
Saturday     485395   4667.26     104      14.75        14.25
Monday       477807   4550.54     105      14.52        14.38
Friday       471214   4574.89     103      14.32        14.11
Thursday     469109   4510.66     104      14.25        14.25
Wednesday    455503   4338.12     105      13.84        14.38
Tuesday      444027   4228.83     105      13.49        14.38
```

The trend of bike rentals with weekdays is almost flat, and hence weekday might not be having any significant effect on the target variable.





```
                 sum      mean   count   sum in %   count in %
workingday
1            2290576   4590.33     499       69.6        68.36
0            1000269   4330.17     231       30.4        31.64
```

The trend of bike rentals with respect to working day is almost flat, and hence working day might not be having any significant effect on the target variable.

Hence, it can be concluded that, a variation in demand of bike rentals is visible with respect to change in **season, yr, month, and weathersit** while no significant trend has been observed with variation in the variables like **working day and weekday.**

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

When dummy variables are created for a categorical variable with n number of levels, without passing the **drop_first=True** argument, then n number of dummy variables will be created. But, actually to describe n levels of a categorical variable only n-1 dummy variables are required. Hence, we pass **drop_first=True,** the first column of the created dummy variables will be automatically dropped resulting in n-1 variables as required.

For example: We have season as a categorical variable with 4 levels. So, when we create dummy variables for it using the following command: pd.get_dummies(bike['season']), we will get all the 4 levels hot encoded as below.

Categorical Season Value	fall	spring	winter	summer
fall	1	0	0	0
spring	0	1	0	0
winter	0	0	1	0
summer	0	0	0	1

But we can explain this with only 3 dummy variables as below:

Categorical Season Value	spring	winter	summer
fall	0	0	0
spring	1	0	0
winter	0	1	0
summer	0	0	1

Because the fall can be denoted as 000, that is the season when it is neither spring nor winter nor summer. This is achieved by passing on **drop_first=True** argument into pd.get_dummies command.

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

The highest correlation with numeric variable is shown by temp and atemp. In turn, these two variables are highly correlated to the point that only one of them is necessary. Hence, atemp was later removed after VIF analysis. Hence, temperature would become the most correlated variable to the target count variable.

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

The assumptions of linear regression made to create the model are the following:

a. The predicted values have linear relationship with the actual values.

   To validate this assumption a scatter plot between actual values and predicted values in the train and test dataset were generated and best fit line were drawn to ensure linear relationship.

b. The error terms are normally distributed.

   A histogram of error terms was plotted to check whether it follows the normal distribution, and the result was further confirmed using normal Q-Q plot.

c. The training and testing accuracy are nearly equal hence there is no Overfit/Underfit situation.

   The R^2 value of models on train and test datasets were calculated and was found to be 82.99% and 80.44% which are very close to each other, ensuring that there is not over or under fitting of the data.

d. There is no Multicollinearity between two independent variables.

   A heatmap of all variables was created to show that none of the variables are highly correlated to each other. Further, VIF of all the variables were also calculated, which indicated that there is no multicollinearity between any variables since all VIF values are well within the acceptable range of less than 5.

e. Homoscedasticity of Residuals.

   A residual plot was drawn and from that, it was clear that there is no visible trend in the distribution of residuals. Hence, it was confirmed that the residuals are homoscedastic.

f. There are at least 20 records of all independent variables.

   The info() function was called on the source data to ensure there are more than 20 non-null values for all the variables under consideration, and it was found that the data contained 730 non-null values for all variables in the dataset.

g. All categorical are converted to numeric dummy variables.

   The info() function was called on the train and test datasets to verify all variables are of the numeric data type, hence ensuring all categorical variables were successfully converted to dummy numeric variables.

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

The final Linear Regression Equation is:

cnt = 0.2347 x yr + 0.4509 x temp - 0.1513 x windspeed -0.0935 x spring + 0.0759 x winter - 0.2854 x Light Precipitation - 0.0789 x Misty or Cloudy +0.0533 x 3 +0.0431 x 4 +0.0558 X 5 +0.0802 X 9

From the equation we can see that the top 3 features with highest influence on the count of rentals are temperature (temp), Light Precipitation (weathersit), year (yr) respectively.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Regression is a statistical method that is used to determine the connection between an outcome variable, and one or more independent factors. This approach is most commonly used for predicting and determining cause-and-effect correlations among variables. When the dependent variable is of continuous data type, regression can be used, irrespective of the data type of the predictors or independent variables. The regression approach aims to identify the best fit line that accurately depicts the connection between the dependent and independent variables.

The most fundamental type of regression analysis is linear regression, which is a supervised machine learning algorithm. The Linear Regression assumes that there is a linear relationship between the dependent variable and independent variables, the accuracy of train data and test data are almost equal so that there is not over/under fitting, and the error terms are normally distributed to ensure there are no non-linear relationships in the data.

When all these conditions are satisfied, LR can be applied to develop a model for accurate predictions of the target variable. It strives to find the best fit line in regression to explain the connection between the predictors and the predictive variable. In linear regression, the output variable is formulated as a function of the independent variables, their coefficients, and an error term of regression as given below:

$$\mathbf{y = \beta_0 + \beta_i X_{i.}}$$

where X_i is ith independent variable in the training input data, β_0 is the y-intercept, β_i is the co-efficient of ith independent variable in the input X data, and y is the predicted target variable. The model seeks to predict y value in such a way that the error difference between predicted and real value is as little as possible by reaching the best-fit regression line.

When there is only one predictor variable a Simple Linear Regression model is created, whereas when multiple predictor variables are present Multi Linear Regression is used. In MLR, there are a few more assumptions that needs to be satisfied in addition to the three specified in SLR. They are absence of Multicollinearity, Homoscedasticity of residuals, Sample size and Categorical Variable Conversion.

Multicollinearity condition requires no two independent variables to be highly correlated. Homoscedasticity of residuals means the variance of errors must be constant across all independent variables. To do MLR, there should be at least 20 records of all independent variables, and if categorical variables are present, they should be converted to numeric dummy variables.

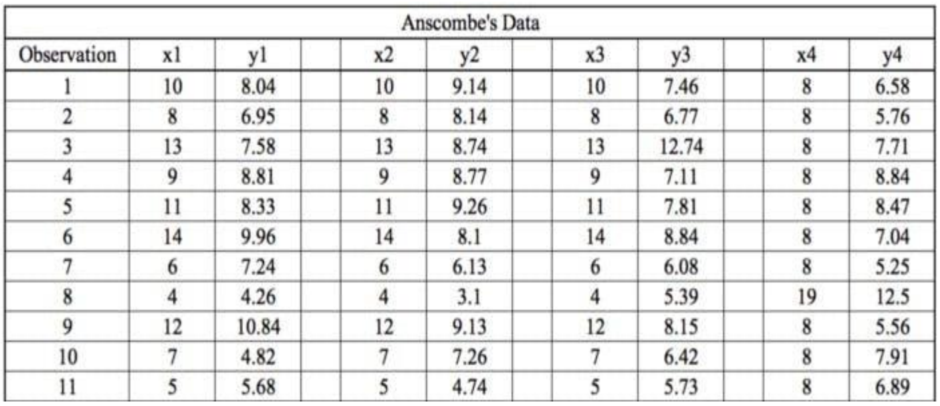
Some examples of use of linear regression in real life would be:

- For understanding the relationship between budget allocation and the return on investment from each of the invested sectors in an organization.
- To predict the impact of weather change on yield of crops in agricultural industry.
- To forecast revenue of the company in next month based on previous years performance and this year's data.
- To comprehend the relationship between drug usage and change in blood pressure in a control group of patient in the medical field.

2. Explain the Anscombe's quartet in detail.

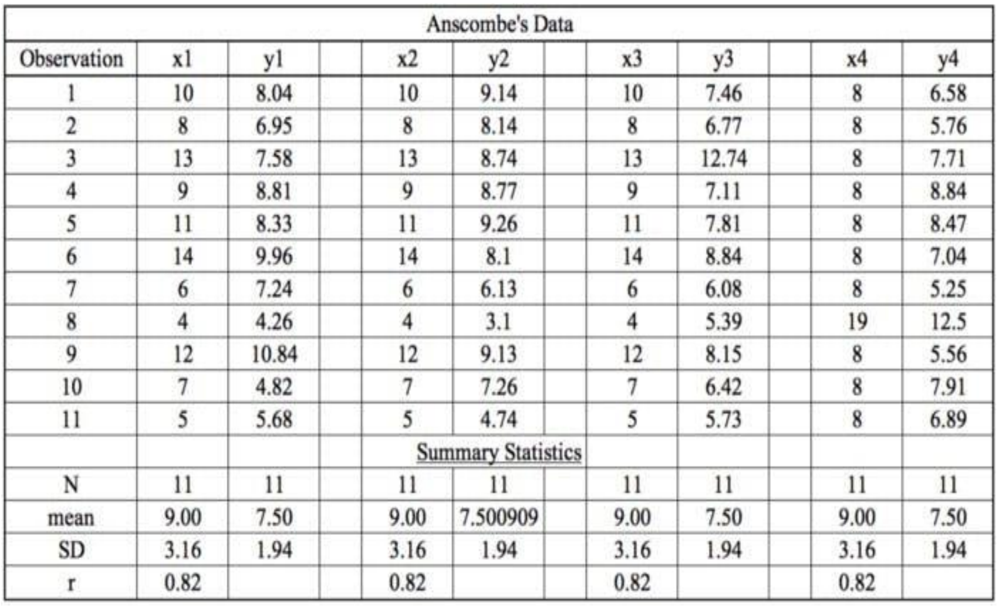
Francis Anscombe created Anscombe's Quartet in 1973 to demonstrate the significance of plotting graphs before analysing and modelling, as well as the impact of additional observations on statistical features. It is described as a set of four data sets that are almost equal in terms of simple descriptive statistics but have extremely distinct distributions and look differently on scatter plots. The necessity of visualising data before using various algorithms to develop models is illustrated by Anscombe's Quartet, since data feature plots may help discover abnormalities in the data such as outliers, diversity, and linear separability.

For Example: The four plots of Anscombe's quartet can be defined as below:



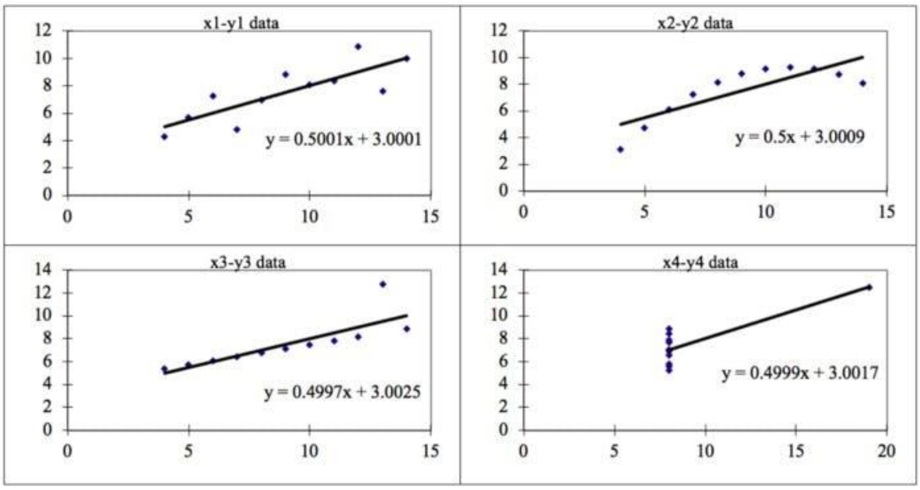
Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		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

When the statistical summary of the above four datasets is generated, it would be as following:



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	

All four datasets have similar descriptive statistical features like mean, standard deviation, and variance, along with very similar x and y values. However, when the data is plotted using a scatter plot, we get a completely different picture.

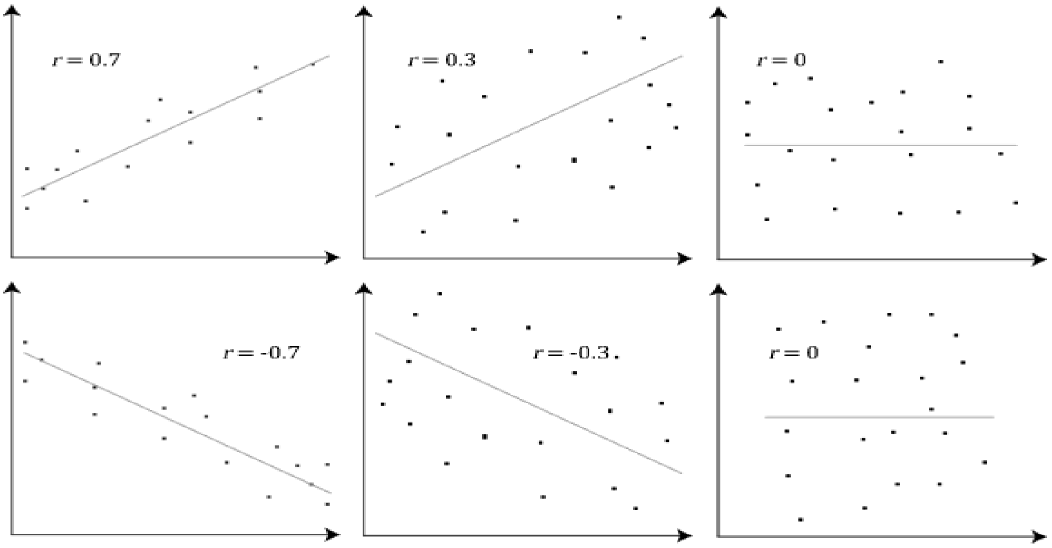


If we try to explain the data based on the above plots, only the first dataset can be fitter with a linear regression model satisfactorily. Second dataset has non-linear relationship, third and fourth datasets has outliers which can't be handled by the linear regression algorithm.

3. What is Pearson’s R?

Pearson's R, also known as the correlation coefficient, is a numerical representation of the strength of a linear relationship between two variables. The correlation coefficient will be positive if the variables tend to go up and down together. The correlation coefficient will be negative if the variables tend to go up and down in opposite directions, with low values of one variable correlated with high values of the other. That is the value of Pearson’s R vary between -1 and +1, where -1 indicates that the variables have a perfectly linear relationship but with a negative slope meaning the direction of change in variables are opposite; and a +1 indicates that the variables have a perfectly linear relationship with positive slope indicating, both variables change in the same direction. When R=0, it means there is no linear relationship between the variables under consideration. The points get increasingly closely clustered around a straight line across the data as the correlation coefficient increases in magnitude.

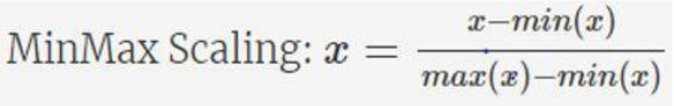
The approximate R value or at least the strength and direction of linear relationship could be seen from scatter plots itself. The following scatter plots illustrates this concept.



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

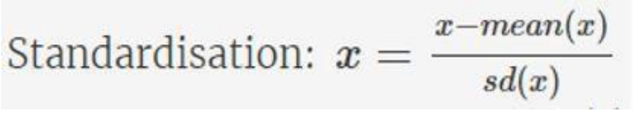
Scaling is a data pre-processing technique that normalises data within a certain range by applying it to independent variables. It also aids in the acceleration of algorithmic calculations. The majority of the time, the acquired data set comprises features with a wide variety of magnitudes, units, and ranges. If scaling is not done, the algorithm will only consider magnitude rather than units, resulting in erroneous modelling. To address this problem, we must scale all of the variables to the same magnitude level. Scaling only impacts the coefficients and has no effect on other factors like as the t-statistic, F-statistic, p-values, R-squared, and so on.

Normalization Scaling is done to bring all data into a range of 0 and 1 so that all variables under consideration are in the same range. In python, sklearn.preprocessing.MinMaxScaler is used to implement normalization. Min-Max scaling is advantageous when the data has categorical variables that need to be converted to numeric dummy variables, because normalization can handle dummy variables very well. The n-dimensional data is squished into an n-dimensional unit hypercube using this technique. It's also handy when we don't know how the data is distributed.



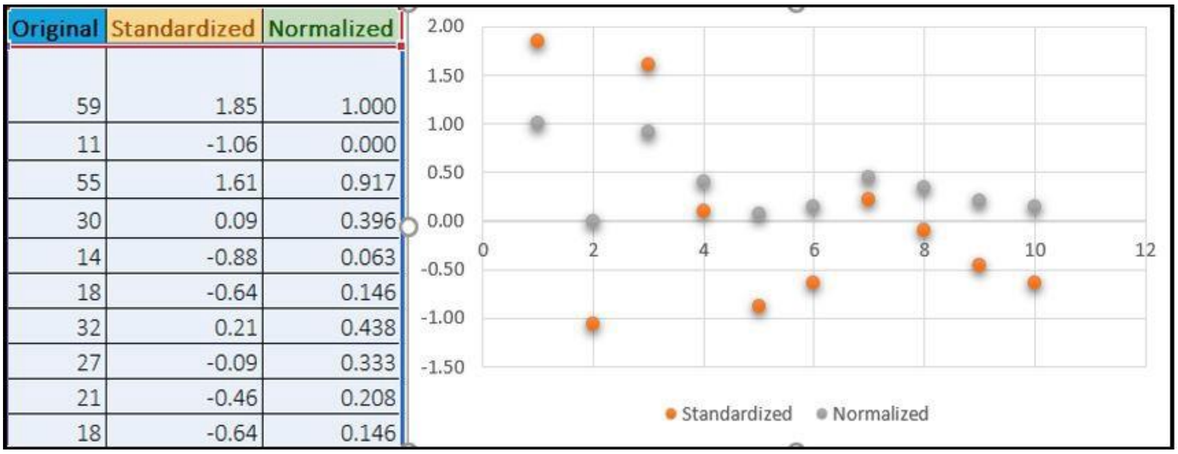
$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

In Standardized scaling, the data is converted into a standard normal distribution which has zero mean value and 1standard deviation, that is all data values are replaced with their Z scores. In python, sklearn.preprocessing.scale is used to implement standardization. Standardization is advantageous when the data is skewed with many outliers because normalization can’t cope with outliers, but standardization can preserve that information. It squishes or expands the data by translating it to the mean vector of the original data.



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

The following example illustrates the two type of scaling on a sample data:



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

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

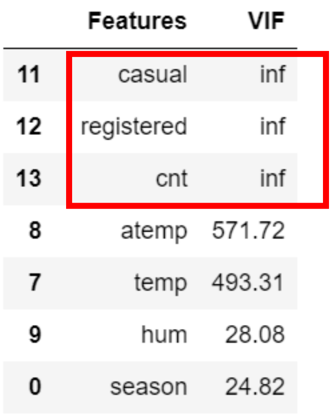
When VIF is infinite, it means that there is perfect correlation between the two independent variables under consideration. That’s the R^2 will be one with these two variables in the model.

The equation of VIF is: **VIF = 1/(1-R^2)**

So, when R^2 =1; the equation becomes **VIF = 1/(1-1) = 1/0 = Inf**

In a data when two independent variables are highly correlated, it’s said to have multicollinearity . Thus, to overcome multicollinearity, one of the variables under consideration that producing infinite VIF must be removed, then the other variable will have a finite VIF. Sometimes, the multicollinearity might be caused by more than two variables, that is a predictor variable in the data is perfectly represented by a linear combination more than one of other independent variables, resulting in infinite VIF again. In this case, one-by-one variables need to be removed and model need to be assessed to eliminate all multicollinearity present.

Example: In the bike dataset provided infinite VIF is obtained for 3 variables: cnt, casual and registered ass below. This because count (cnt) can be expressed as a linear relationship of casual and registered variables, since cnt = registered + casual. So here the infinite VIF is due to the high correlation and linear relationship between the 3 varaibles.

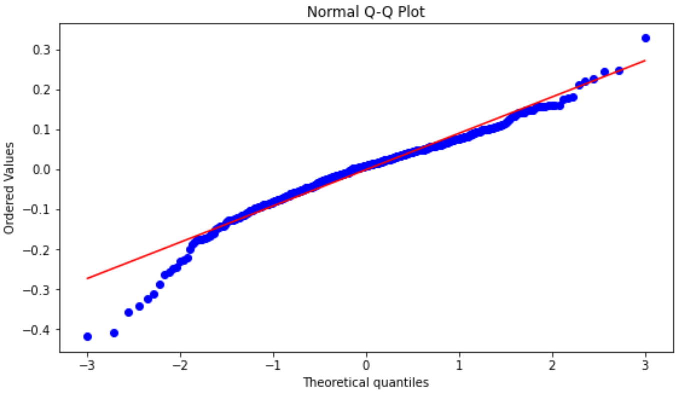


	Features	VIF
11	casual	inf
12	registered	inf
13	cnt	inf
8	atemp	571.72
7	temp	493.31
9	hum	28.08
0	season	24.82

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

The quantile-quantile (Q-Q) plot is a graphical tool for detecting if two data sets are from the same population. A Q-Q plot is a comparison of two dataset's quantiles. Quantiles are cut points that divide the range of a probability distribution into equal-probability continuous intervals or the observations in a sample in the same way. Q-quantiles are values that divide a finite collection of values into almost equal-sized Q subgroups. There exist one Q-quantile for each integer k that is satisfied by 0 < k < Q, resulting in Q-1 total quantiles. Continuous distributions also can use quantile to apply rank statistics to the continuous data.

These plots are useful in a linear regression situation where the training and test data sets are obtained separately, and the Q-Q plot is used to validate that both data sets are from populations with similar distributions. In a Q-Q plot, 45-degree reference line is plotted to determine the normality of populations. The points fall roughly along this reference line if the two sets originate from the same population with the same distribution. But if the two data sets are from populations with distinct distributions, then more is the deviation from this reference line.



Q-Q plots are also useful to confirm the normality of residuals in a linear regression model, which is one of the core assumptions in linear regression. If the predicted y terms plotted against theoretical quantile fall along the 45-degree reference line of a Q-Q plot, we can confirm that the residuals have a normal distribution.