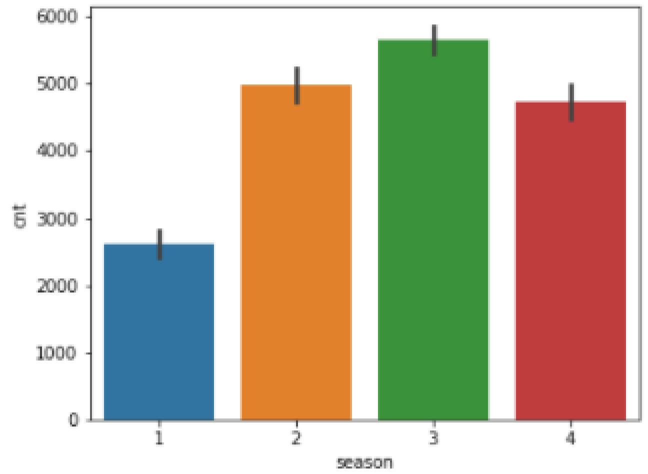
Assignment-based Subjective Questions

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

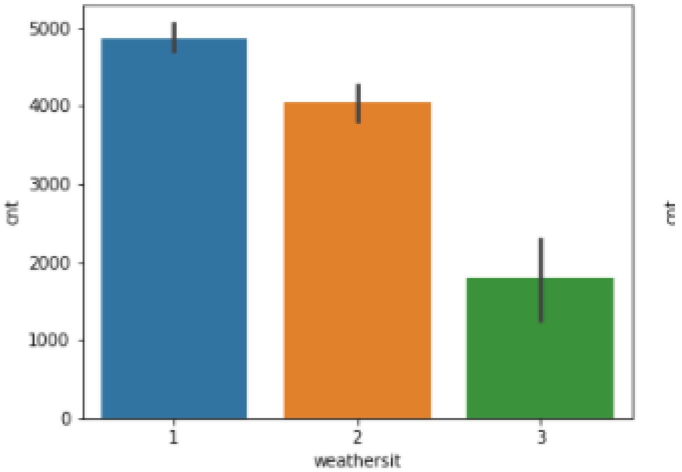
Answer: The categorical variable in given dataset were yr,mnth , holiday , season and weathersit . We can use box plot to see their effect on target variable –cnt as well.

Season: The barplot shows that spring season has lowest count and fall had maximum value of count. Summer and winter has moderate value of count.



```
1: spring, 2: summer, 3: fall, 4: winter
```

Weathersit: There are no users at the time heavy rain/snow .Highest count was time when weather is Clear, Partly Cloudy.



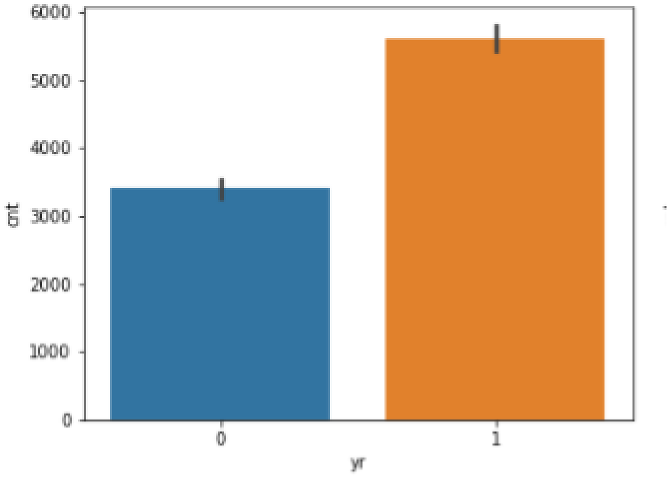
1: Clear, Few clouds, Partly cloudy, Partly cloudy

2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

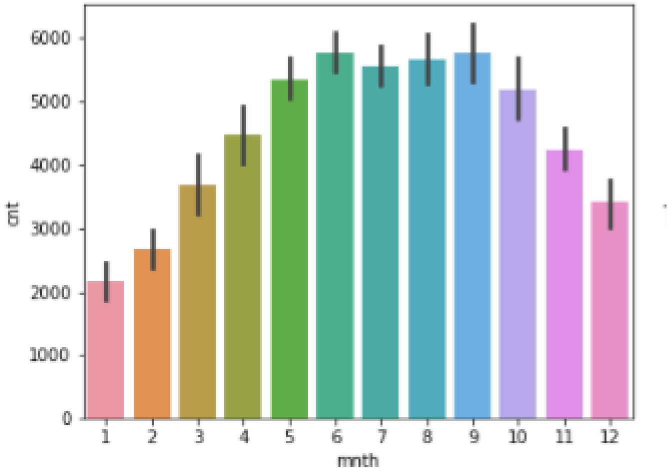
3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

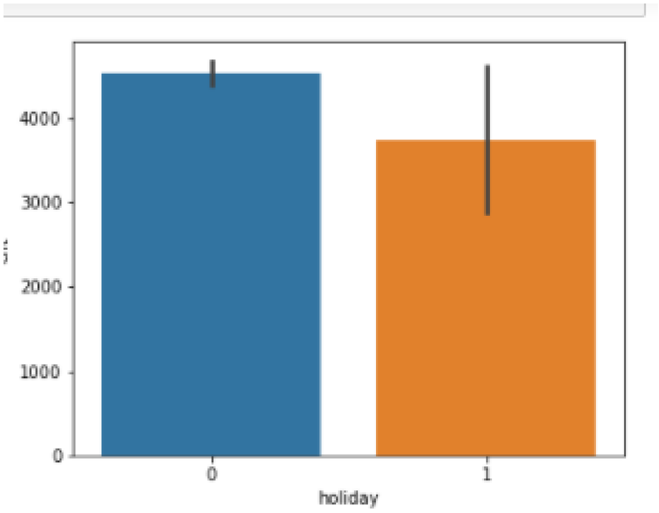
Yr: Count was higher in year 2019 than 2018.



Mnth: September saw highest count and in December/January count was lowest. This is matching with earlier findings at the time of snow –count is low.



Holiday: Count was low during holiday.



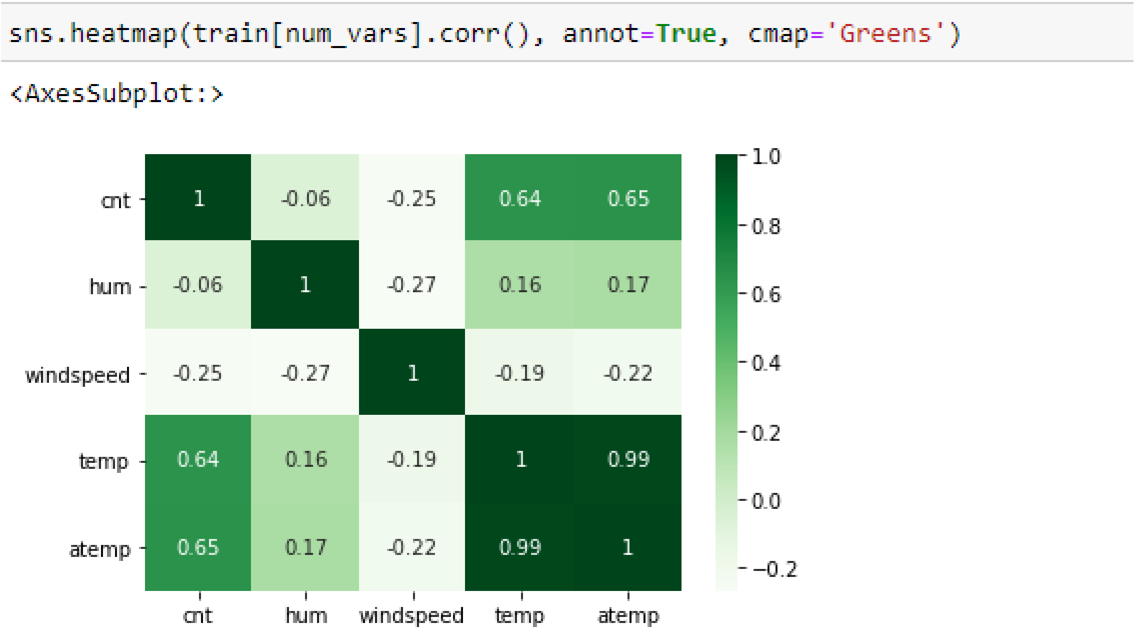
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Answer: To stop redundancy in dummy variables we need to use drop_first=True. drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

If we have categorical variables with n-levels , then we need to use n-1 columns to represent the dummy variables .

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

Answer : “temp” and “atemp” are two numerical columns which are highly correlated with target variable cnt .



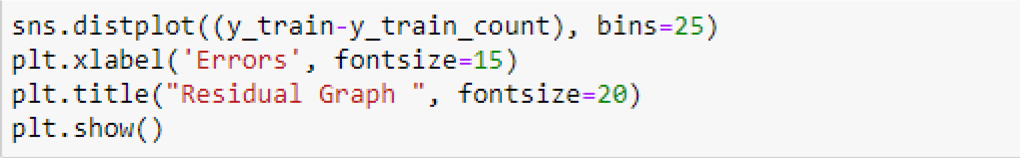
```
sns.heatmap(train[num_vars].corr(), annot=True, cmap='Greens')
```

```
<AxesSubplot:>
```

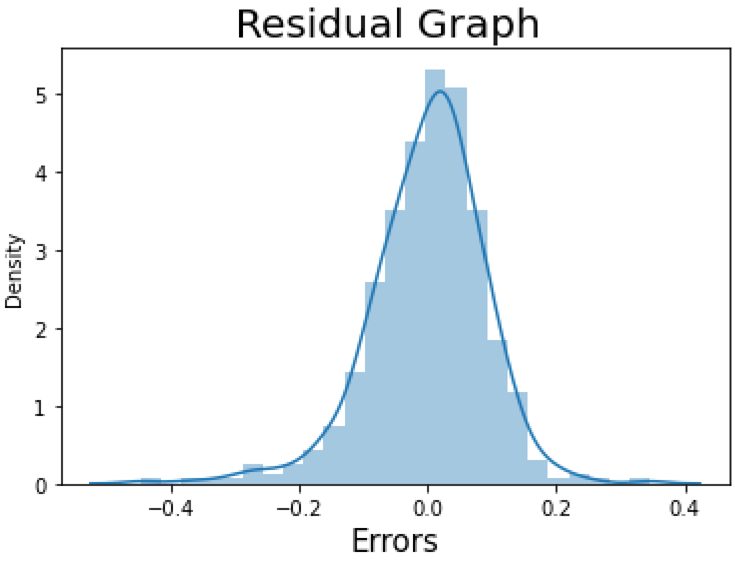
	cnt	hum	windspeed	temp	atemp
cnt	1	-0.06	-0.25	0.64	0.65
hum	-0.06	1	-0.27	0.16	0.17
windspeed	-0.25	-0.27	1	-0.19	-0.22
temp	0.64	0.16	-0.19	1	0.99
atemp	0.65	0.17	-0.22	0.99	1

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

Answer : Residual Distribution shows follow normal distributions and mean was centered around 0 . In python sheet, we checked this assumption by plotting residuals with help of distplot and found residual follows normal distribution.



```
sns.distplot((y_train-y_train_count), bins=25)
plt.xlabel('Errors', fontsize=15)
plt.title("Residual Graph ", fontsize=20)
plt.show()
```



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

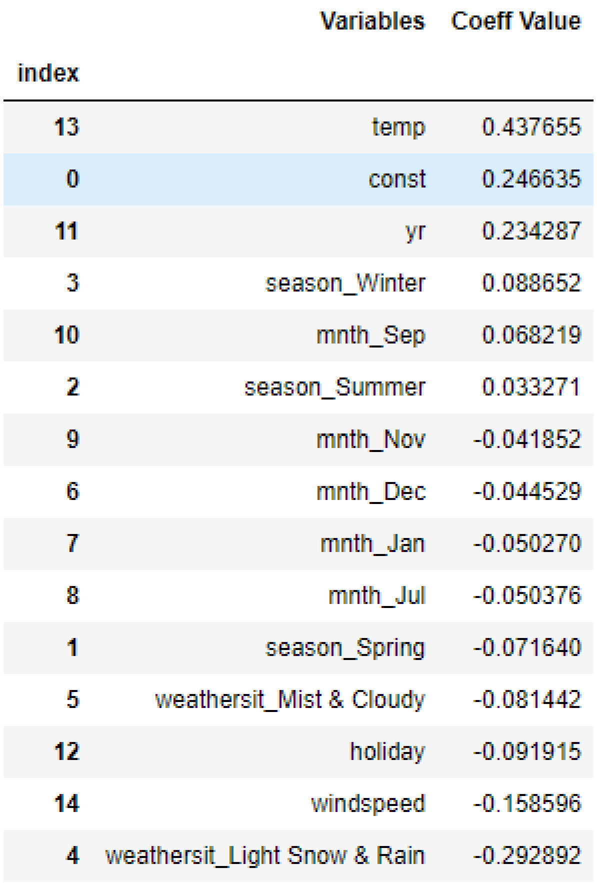
Answer:

Top 3 positive features are

temp – coefficient: .4376
yr –coefficients : .23
Season_Winter-.08

Top 3 , Negative features are

weather_Light Snow & Rain – .2928
windspeed -.15
holiday-.09



	Variables	Coeff Value
index		
13	temp	0.437655
0	const	0.246635
11	yr	0.234287
3	season_Winter	0.088652
10	mnth_Sep	0.068219
2	season_Summer	0.033271
9	mnth_Nov	-0.041852
6	mnth_Dec	-0.044529
7	mnth_Jan	-0.050270
8	mnth_Jul	-0.050376
1	season_Spring	-0.071640
5	weathersit_Mist & Cloudy	-0.081442
12	holiday	-0.091915
14	windspeed	-0.158596
4	weathersit_Light Snow & Rain	-0.292892

General Subjective Questions

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

Answer: Linear regression is used for finding linear relationship between target and one or more predictors/Numerical variables.

There are two types of linear regression- Simple and Multiple.

Simple Linear Regression: for example, if we have dataset to find relationship between ‘number of hours studied’ and ‘marks obtained’.

Goal is to design a model that can predict marks if given the number of hours studied. Using the training data, a regression line is obtained which will give minimum error. This linear equation is then used for any new data. That is, if we give number of hours studied by a student as an input, our model should predict their mark with minimum error.

y(pred)=b0+b1*x

Value of b0 and b1 will be such so that error will be minimized.

Observation on ‘b1’:

If b1 > 0, then x(predictor) and y(target) have a positive relationship. That is increase in x will increase y.

If b1 < 0, then x(predictor) and y(target) have a negative relationship. That is increase in x will decrease y.

b0 –Intercept/Constant.

Multiple Linear Regressions: it is used when dependent variable is predicted using multiple independent variables.

Equation of MLR will be

y(pred)=b0+b1*x+b2*x2+b3*x3

b1- coefficients of x1

b2- coefficients of x2

b0- Intercept (Constant Term)

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

Answer :- According to wiki , "*Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed*."

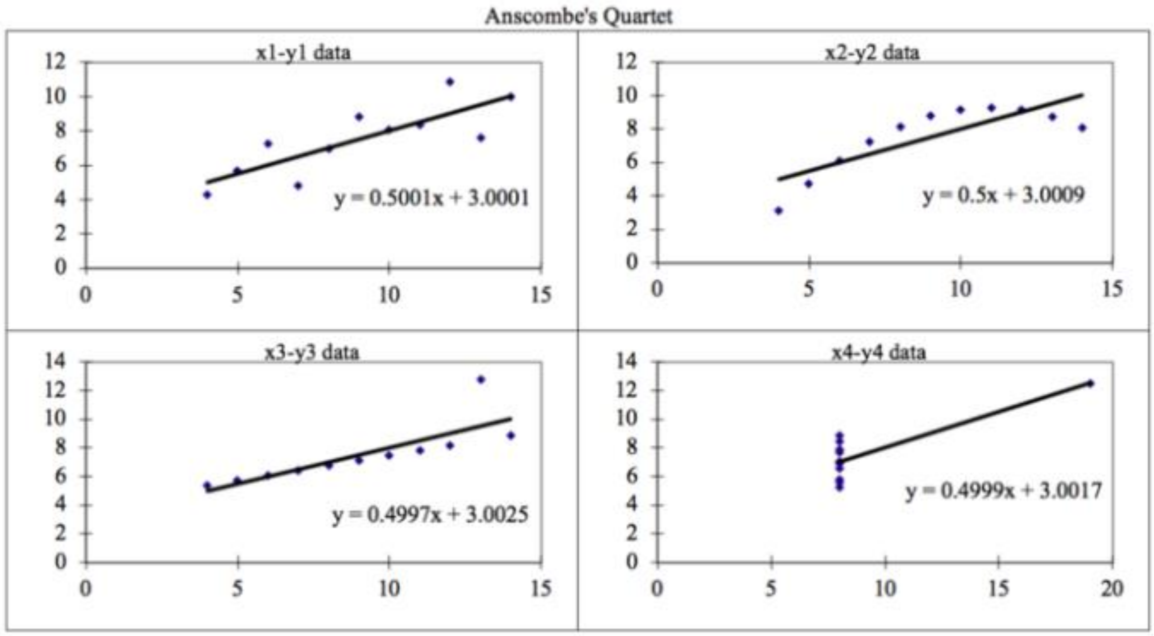


Image by Author

it was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

The four datasets can be described as:

Dataset 1: this fits the linear regression model pretty well.

Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.

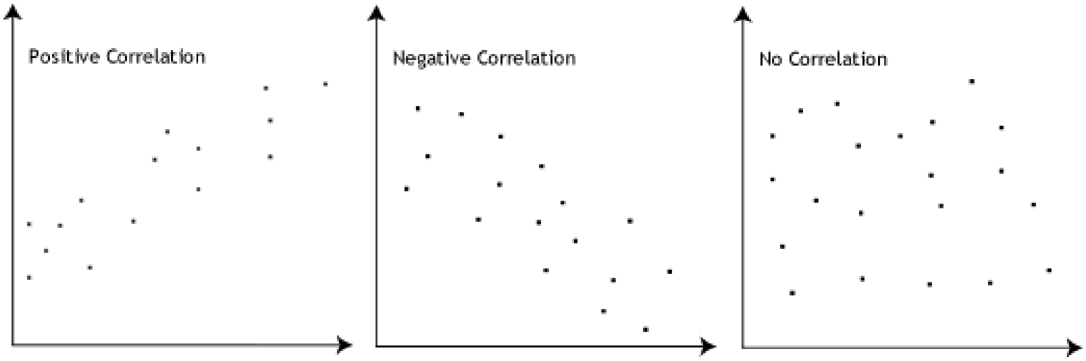
Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model

Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

In brief, *four datasets that were intentionally created to describe the importance of data visualization and how any regression algorithm can be fooled by the same. Hence, all the important features in the dataset must be visualized before implementing any machine learning algorithm on them which will help to make a good fit model*.

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

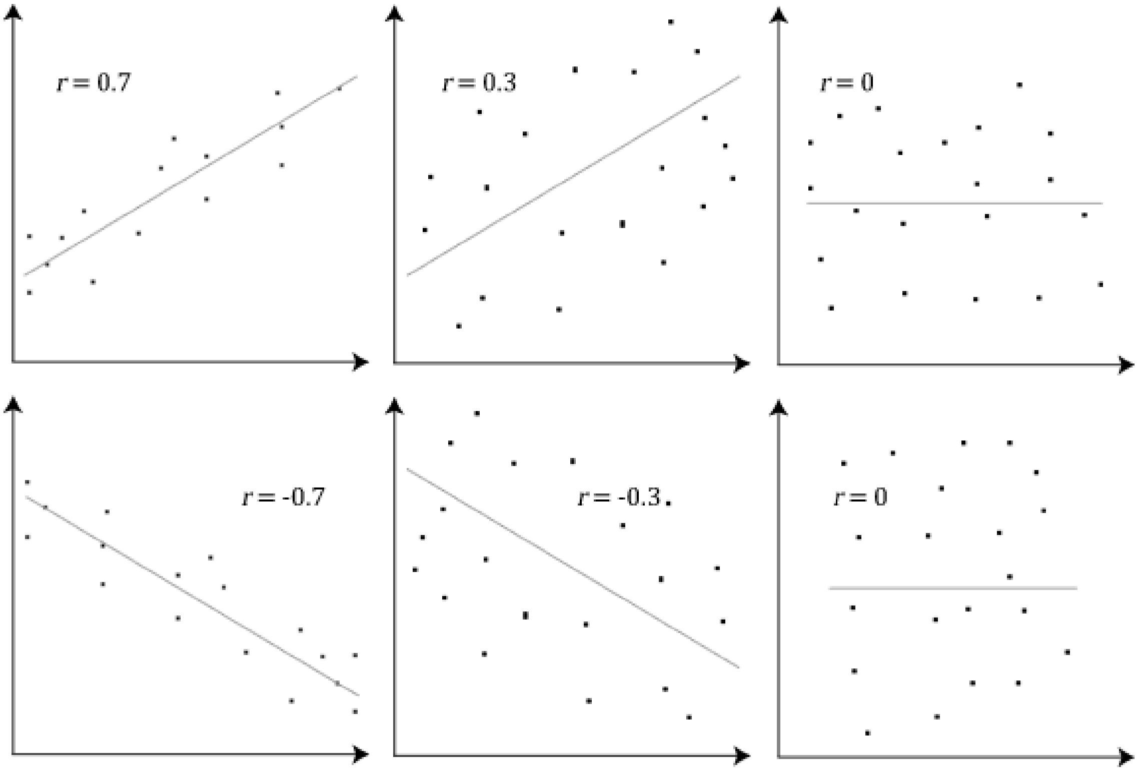
Answer :- The Pearson product-moment correlation coefficient (or Pearson correlation coefficient, for short) is a measure of the strength of a linear association between two variables and is denoted by *r*. Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, *r*, indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit).



r=1 means data is perfectly linear with positive slope.
r=-1 means data is perfectly linear with negative slope.
r=0 means there is no linear relationship.

Determining the strength of association based on the Pearson correlation coefficient-

The stronger the association of the two variables, the closer the Pearson correlation coefficient, *r*, will be to either +1 or -1 depending on whether the relationship is positive or negative, respectively. Achieving a value of +1 or -1 means that all your data points are included on the line of best fit – there are no data points that show any variation away from this line. Values for *r* between +1 and -1 (for example, *r* = 0.8 or -0.4) indicate that there is variation around the line of best fit. The closer the value of *r* to 0 the greater the variation around the line of best fit. Different relationships and their correlation coefficients are shown in the diagram below:



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

Answer: Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. 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, regardless of the unit of the values.

For example, If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

Normalization/Min-Max Scaling : it brings all of the data in range of 0 and 1 . sklearn.preprocessing.MinMaxScaler helps to implement normalization in python .

MinMaxScaling : x =x-min(x)/max(x)-min(x)

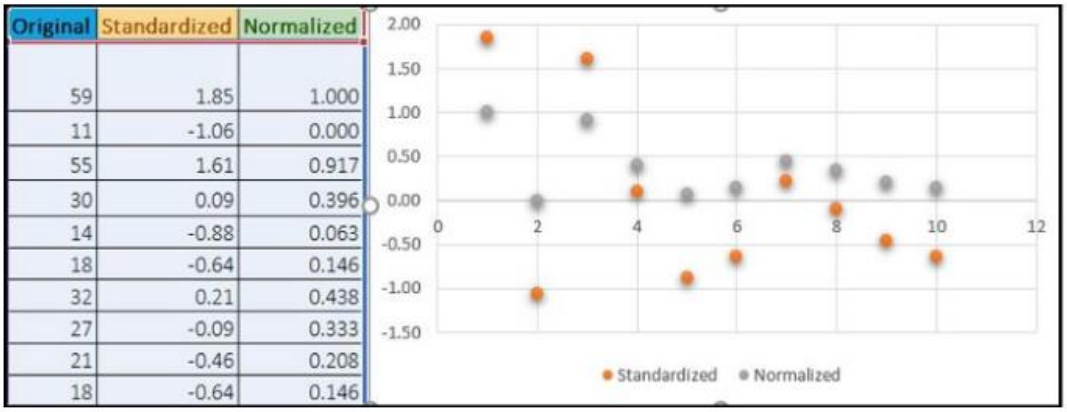
Standardization Scaling: Standardization replaces the values by their Z scores .it brings all of the data into a standard normal distribution which has mean zero and standard deviation .

Standardization =x-mean(x)/sd(x)

sklearn.preprocessing.scale helps to implement standardization in python .

One disadvantage of normalization over standardization is that is loses some information in data , especially outliers .

Below example shows 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

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

Answer: VIF (variance inflation factor) is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to co linearity. In order to determine VIF, we fit a regression model between the independent variables.

VIF infinity means there is perfect relationship between two independent variables .In case of perfect co-relation , we get R2=1 ,which leads to 1/(1-R2) infinity . To overcome this we need to drop one of variables from dataset which is causing perfect multi-co linearity.

In short, an infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

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 (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

The advantages of the q-q plot are:

1. The sample sizes do not need to be equal.
2. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.

The q-q plot is formed by:

Vertical axis: Estimated quantiles from data set 1

Horizontal axis: Estimated quantiles from data set 2

Both axes are in units of their respective data sets. That is, the actual quantile level is not plotted. For a given point on the q-q plot, we know that the quantile level is the same for both points, but not what that quantile level actually is.

If the data sets have the same size, the q-q plot is essentially a plot of sorted data set 1 against sorted data set 2. If the data sets are not of equal size, the quantiles are usually picked to correspond to the sorted values from the smaller data set and then the quantiles for the larger data set are interpolated.

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 behavior?