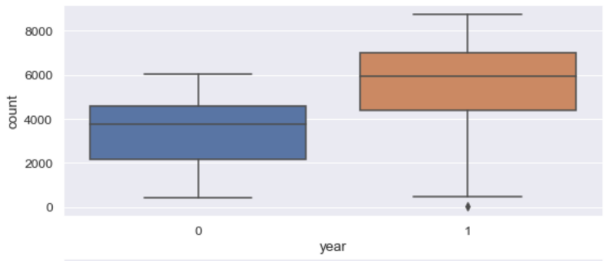
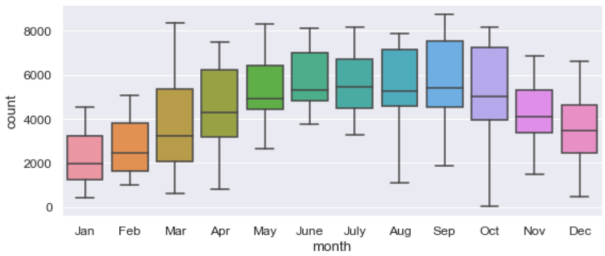
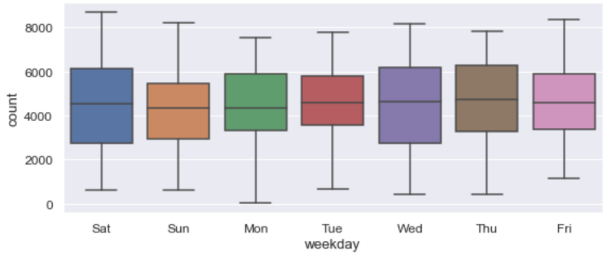
Assignment-based Subjective Questions

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

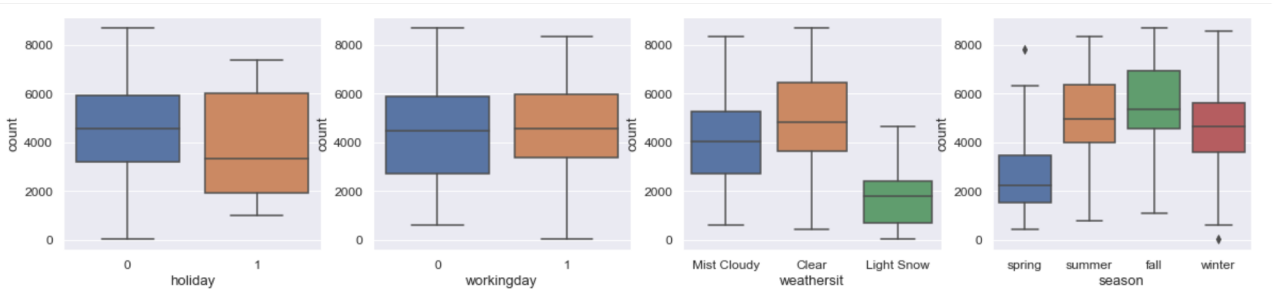
The categorical variable in the dataset were yr, mnth, weekday, holiday, workingday, weathersit and season . Categorical variables had the following effect on our target variable:







1. Year:
    - The number of rentals in 2019 was more than 2018
2. Month:
    - The count of bike sharing values increases in summer months
    - Almost 10% of the bike booking were done in the months 5,6,7,8 & 9
    - There is a median of over 4000 booking per month.
    - This indicates, month has some trend for bookings and can be a good predictor for count.
    - The month of September have highest no of rentals while December have least. This observation is on par with the observation made in weathersit. The weather situation in December is usually have heavy snow.
3. weekday:
    - All weekdays have between 13.5%-14.8% of total booking and thus shows very close trend.
    - Weekdays have medians between 4000 to 5000 bookings.
    - This variable can have some or no influence towards the predictor



4. Holiday:
    - The count of bike sharing values ars less during holidays
    - This data is clearly biased as almost 97.6% of the bike booking were done during non-holiday.
    - Bike Booking counts reduced during holiday
    - Indicates holiday cannot be a good predictor for the count.
5. Working day
    - The median of non-working and working days are almost same
    - The median of close to 5000 booking for the period of 2 years.
    - Indicated working day can be a good predictor for the dependent variable.
6. weathersit:
    - - Almost 67% of the bike booking were happening during ‘weathersit1
    - - and has a median of close to 5000 booking for the period of 2 years.
    - - This was followed by weathersit2 with 30% of total booking.
    - - This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the count.
    - - The count of bike sharing has few zero values for weather situation - 'Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog'
    - There are no users when there is heavy rain or snow indicating that this weather is extremely unfavorable. Highest counts were seen when the weathersit was ‘Clear, Partly Cloudy’.
7. Season
    - Fall has a median of over 5000 booking, This is followed by summer & winter.
    - The count of bike sharing is least for spring
    - The boxplot indicates that spring season had least value of booking counts whereas fall had maximum values. Summer and winter had moderate value for bike booking counts.
    - This Indicates, season can be a good predictor for the dependent variable.

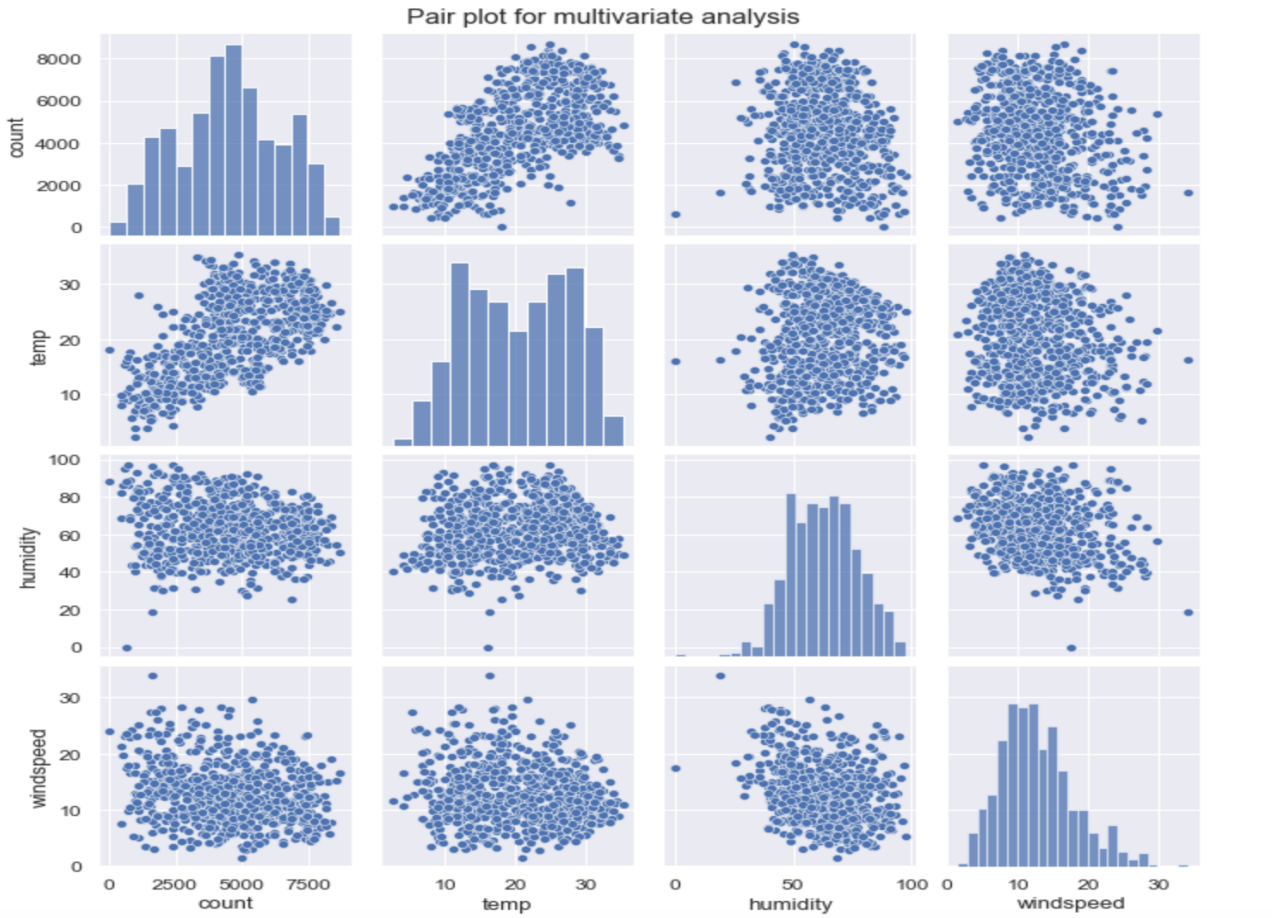
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

If we don't drop the first column then our dummy variables will be correlated (redundant). This may affect some models adversely and the effect is stronger when the cardinality is smaller. Below are few observation regarding drop_first=True with an example from our assignment:

1. Iterative models may have trouble converging and lists of variable importance may be distorted.
2. If we have all dummy variables it leads to Multicollinearity between the dummy variables.

weekday_Mon	weekday_Sat	weekday_Sun	weekday_Thu	weekday_Tue	weekday_Wed
0	1	0	0	0	0
0	0	1	0	0	0
1	0	0	0	0	0
0	0	0	0	1	0
0	0	0	0	0	1

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

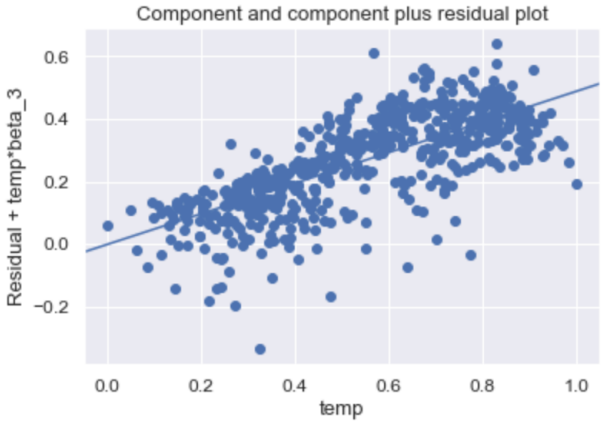


"temp" variable is highly correlated with the target variable count(cnt) .

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

There are five principal assumptions which justify the use of linear regression models for purposes of prediction:

1. linearity of the relationship between dependent and independent variables



   - Indicates temp variable have high corelation with target variable.

2. Little or no Multicollinearity between the feature

	Features	VIF
2	temp	2.66
0	year	2.04
5	month_July	1.33
4	season_winter	1.31
3	season_spring	1.22
6	month_Sep	1.18
7	weekday_Sun	1.16
1	holiday	1.04
8	weathersit_Light Snow	1.03

   - From the VIF calculation we could observe that there is no multicollinearity exists between the predictor variables as all the values are within permissible range of below 5.

3. Little or No autocorrelation in the residuals (no serial correlation)
   - Hypothesis testing states that:
     - $H0$:$B1$=$B2$=...=Bn=0
     - $H1$: at least one Bi!=0
   - From the model summary below, it is evident that all coefficients are not equal to zero which means we can Reject the null hypothesis
   - Durbin-Watson Statistic = 1.983, A value near 2 indicates non-autocorrelation

OLS Regression Results

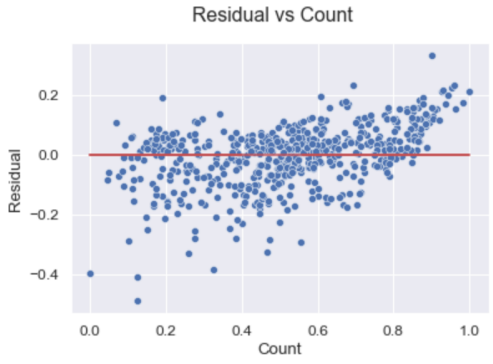
Dep. Variable:	count	R-squared:	0.793
Model:	OLS	Adj. R-squared:	0.790
Method:	Least Squares	F-statistic:	244.2
Date:	Wed, 09 Feb 2022	Prob (F-statistic):	8.39e-190
Time:	01:38:20	Log-Likelihood:	505.97
No. Observations:	584	AIC:	-991.9
Df Residuals:	574	BIC:	-948.2
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1595	0.022	7.182	0.000	0.116	0.203
year	0.2365	0.009	27.652	0.000	0.220	0.253
holiday	-0.0712	0.027	-2.636	0.009	-0.124	-0.018
temp	0.4880	0.032	15.488	0.000	0.426	0.550
season_spring	-0.1101	0.016	-6.923	0.000	-0.141	-0.079
season_winter	0.0607	0.013	4.697	0.000	0.035	0.086
month_July	-0.0355	0.018	-2.027	0.043	-0.070	-0.001
month_Sep	0.0546	0.016	3.451	0.001	0.024	0.086
weekday_Sun	-0.0413	0.012	-3.348	0.001	-0.066	-0.017
weathersit_Light Snow	-0.2899	0.028	-10.360	0.000	-0.345	-0.235

Omnibus:	82.200	Durbin-Watson:	1.983
Prob(Omnibus):	0.000	Jarque-Bera (JB):	161.951
Skew:	-0.812	Prob(JB):	6.81e-36
Kurtosis:	5.004	Cond. No.	12.9

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

4. Homoscedasticity (constant variance) of the errors



- Residuals are equal across the regression line

5. Normality of the error distribution.

- Residuals distribution should follow normal distribution and centred around 0 (mean = 0).We can validate this assumption about residuals by plotting a distplot of residuals and see if residuals are following normal distribution or not. The below diagram shows that the residuals are normally distributed and centred around mean=0.

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

As per our final linear Model, the top 3 predictors that influences the bike booking are:

1. Temperature (temp) - A coefficient value of '0.49' tells that a unit increase in temp increases the bike counts by 0.49 units.
2. Year (yr)- A coefficient value of '0.24' tells that a unit increase in year increases the bike counts by 0.24 units.
3. Weather Situation (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) - A coefficient value of '-0.29' tells that a unit increase in weathersit_Light Snow decreases the bike counts by 0.29 units

General Subjective Questions

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

Regression is a method of modelling a target value based on independent predictors. It is a statistical tool which is used to find out the relationship between the outcome variable also known as the dependent variable, and one or more variable often called as independent variables. This method is mostly used for forecasting and finding out cause-and-effect relationships between variables. Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric values. It is the most basic form of regression analysis.

The equation for the best-fit line:

$$y = mx + c.$$

The equation given above assumes that there is a linear relationship between the dependent variable(y) and the predictor(s)/independent variable(x). Through the best fit line, we can describe the impact of change in independent variables on the dependent variable.

Regression is performed when the dependent variable is of continuous data type and Predictors or independent variables could be of any data type like continuous, nominal/categorical etc. Regression method tries to find the best fit line which shows the relationship between the dependent variable and predictors with least error.

Regression is broadly divided into Simple Linear Regression and Multiple Linear Regression:
Simple Linear Regression: SLR is used when the dependent variable is predicted using only one independent variable.
Multiple Linear Regression: MLR is used when the dependent variable is predicted using multiple independent variables.

The equation for MLR will be:

$$yi = \beta 0 + \beta 1xi1 + \beta 2xi2 + ... + \beta pxip + \epsilon$$

where, for $i=n$ observations:
yi=dependent variable
xi=explanatory variables
$\beta 0$=y-intercept (constant term)
βp=slope coefficients for each explanatory variable
ϵ=the model's error term (also known as the residuals)

There are broadly four assumptions associated with a linear regression model:

1. Linearity: The relationship between independent variables and the mean of the dependent variable is linear.
2. Multicollinearity: Observations are independent of each other. Multicollinearity can be checked using correlation matrix, Tolerance and Variance Influencing Factor (VIF).
3. Homoscedasticity: The variance of residuals should be equal. If Variance of errors are constant across independent variables, then it is called Homoscedasticity. The residuals should be homoscedastic. Q-Q plots are also used to check homoscedasticity.
4. Normality: For any fixed value of an independent variable, the dependent variable is normally distributed. Residuals should be normally distributed

Applications of Linear Regression:

1. Effect of independent variable on dependent variable can be calculated.
2. Used to predict trends.
3. Used to find how much change can be expected in a dependent variable with change in an independent variable.

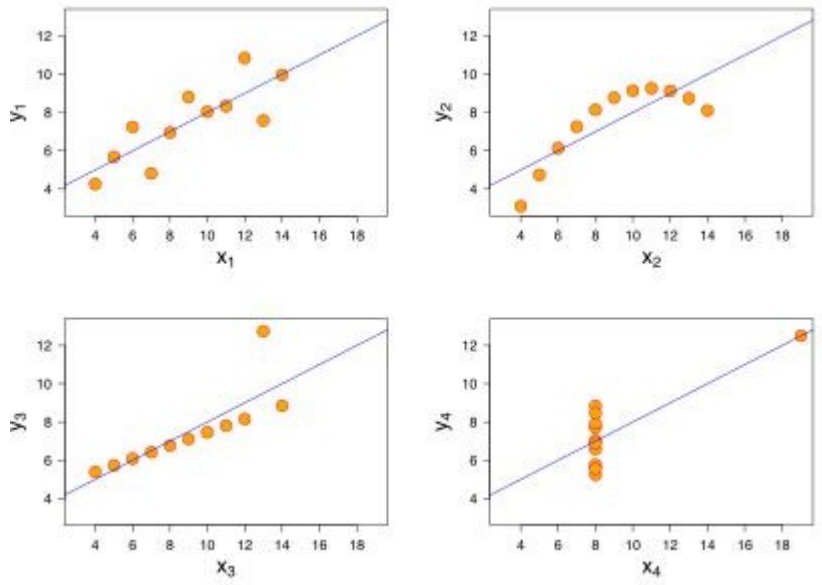
2. Explain Anscombe's quartet in detail. (3 marks)

Anscombe's Quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph.
The quartet is often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, the effect of outliers and the inadequacy of basic statistic properties for describing realistic datasets.

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 those points. Those 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.
The scatter graphs for above four data set are given below:



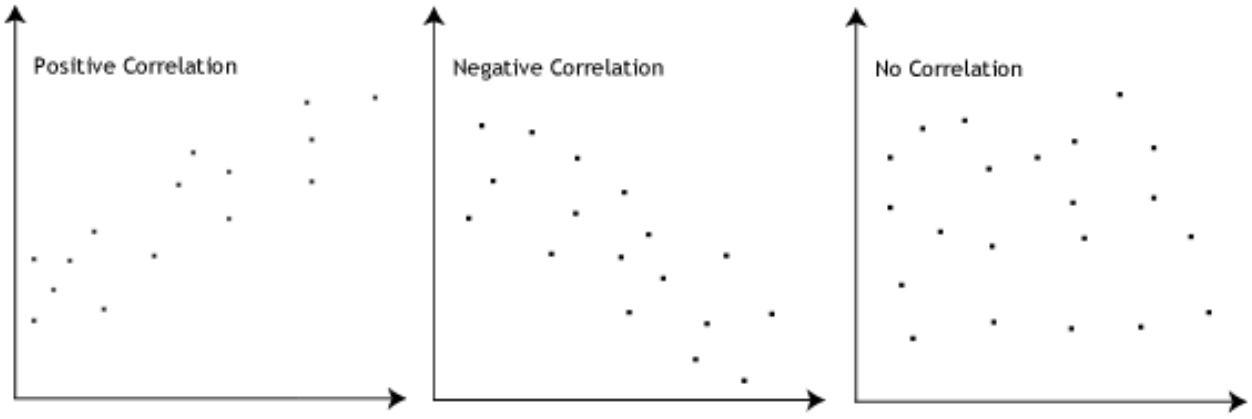
Inferences of the graphs:

- The first scatter plot appears to be a simple linear relationship.
- The second graph on the top right is not linear and not even distributed normally; while there is a relation between them.

- In the third graph at the bottom left , the distribution is linear, but should have a different regression line The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph at bottom right, shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

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

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's R. Pearson's R is a numerical summary of the strength of the linear association between the variables. It is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.



We represent Pearson's R by below Formula

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

Here,

- r=correlation coefficient
- xi=values of the x-variable in a sample
- $\bar{x}$ =mean of the values of the x-variable
- yi=values of the y-variable in a sample
- $\bar{y}$ =mean of the values of the y-variable

The Pearson'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 directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association

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

To normalize the range of predictor features or independent variables of the data set, we use Feature scaling**.** If we ignore feature scaling mechanism then the machine learning algorithm often make wrong interpretations when we have different units associated to the variables. If scaling is not done then algorithm only takes values in account and not units associated to the features hence results in incorrect modelling.
We can consider an example to elaborate this further: If we are not using feature scaling method then it can consider the amount value 5000 rupees to be greater than an amount value 500 dollars but when we look closely to the amounts units we found that's actually not true and in this case, the algorithm will going to predict wrong. Therefore, we use Feature Scaling to make all values to the same magnitudes and thus, overcome this issue.

We generally performed feature scaling technique during the data preprocessing stage to deal with above mentioned scenarios in the dataset.

Normalization is generally used when we know that the distribution of your data does not follow a Gaussian distribution. We can observe that normalization have a bounding range. This technique re-scales a feature value between 0 and 1. Normalization can also be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbours and Neural Networks.
We use ***sklearn.preprocessing.MinMaxScaler*** *to implement normalization in python*

$$X\ new = (X - X\ min)/ (X\ max - X\ min)$$

Standardization on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. This technique re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1. Therefore, we can

observe unlike normalization, standardization does not have a bounding range. Even if we have outliers in our data, they will not be affected by standardization.

We use ***sklearn.preprocessing.scale*** *to implement standardization in python.*

$$x' = (x - \mu) / \sigma$$

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

VIF (Variance Inflation Factor) – is an index that provide a measure how much variance of the coefficient estimate is being inflated by collinearity. An infinite VIF value indicates that the corresponding Independent variable may be expressed exactly by a linear combination of other variables. In order to determine VIF, we fit a regression model between the independent variables.

- If all the independent variables are orthogonal to each other, then VIF = 1.0.
- If there is perfect correlation, then VIF = infinity.
- A large value of VIF indicates that there is a correlation between the variables.
- If the VIF is 5, this means that the variance of the model coefficient is inflated by a factor of 5 due to the presence of multicollinearity.

VIF is expressed by given formula:

$$(VIF) = 1/(1-R\_1^2).$$

Where R-1, is the R-square value of that independent variable which we want to verify how well this independent variable is explained by other independent variables. If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it's R-squared value will be equal to 1.

Therefore, VIF = 1/(1-1) which gives us VIF = 1/0 which would results in Infinity.

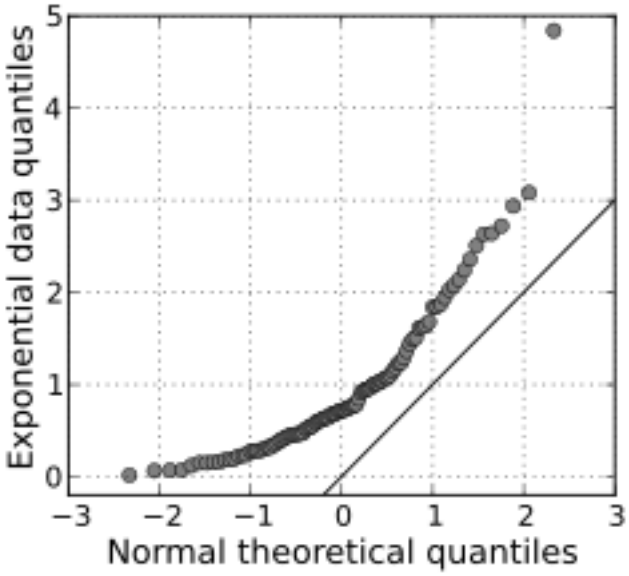
To solve this problem we need to drop one of the variables from the dataset which is causing multicollinearity.

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

As the name suggest Q-Q Plots are the Quantile-Quantile plots which plots two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. If we observe closely a boxplot, the median is a quantile where 50% of the data fall below that point and 50% lie above if.

Now, the purpose of Q-Q plots is to find out if these two sets(one above the median and another one below the median) of data come from the same distribution. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

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.



Observations of Q-Q plots:

- The points in the Q–Q plot will approximately lie on the line y = x, If the two distributions being compared are similar.
- The points in the Q–Q plot will approximately lie on a line(but not necessarily on the line y = x). If the distributions are linearly related.

Usage of Q-Q plots:

- Q–Q plots can be used to check Homoscedasticity.
- 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.