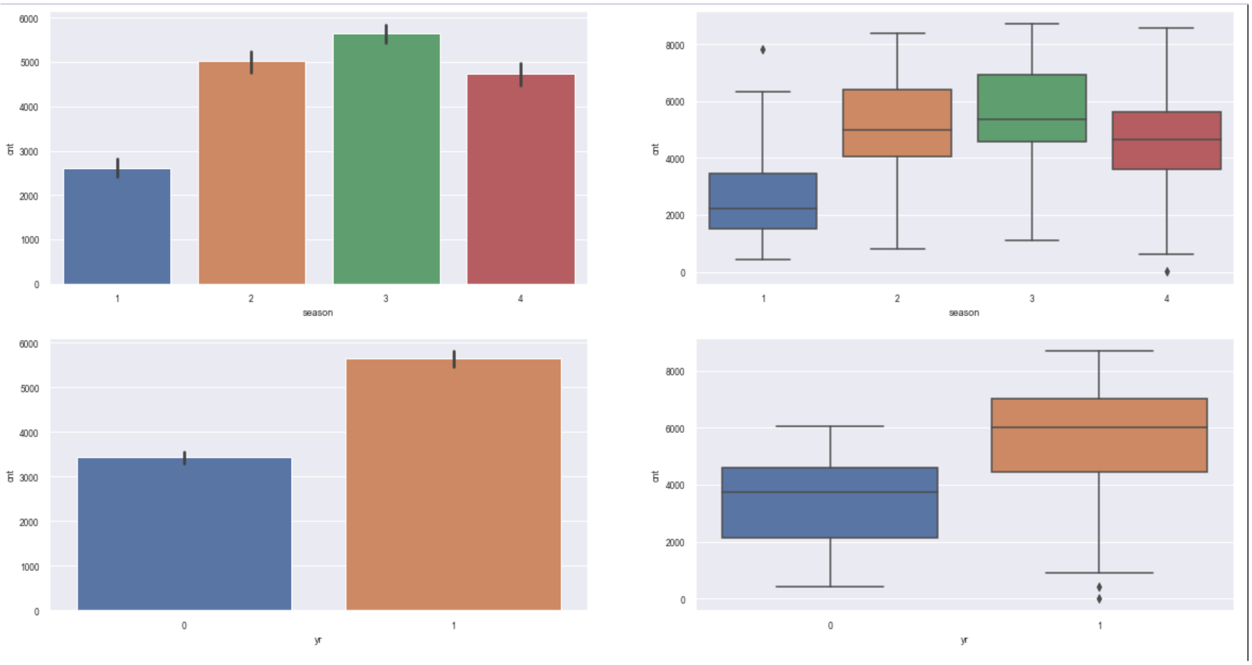
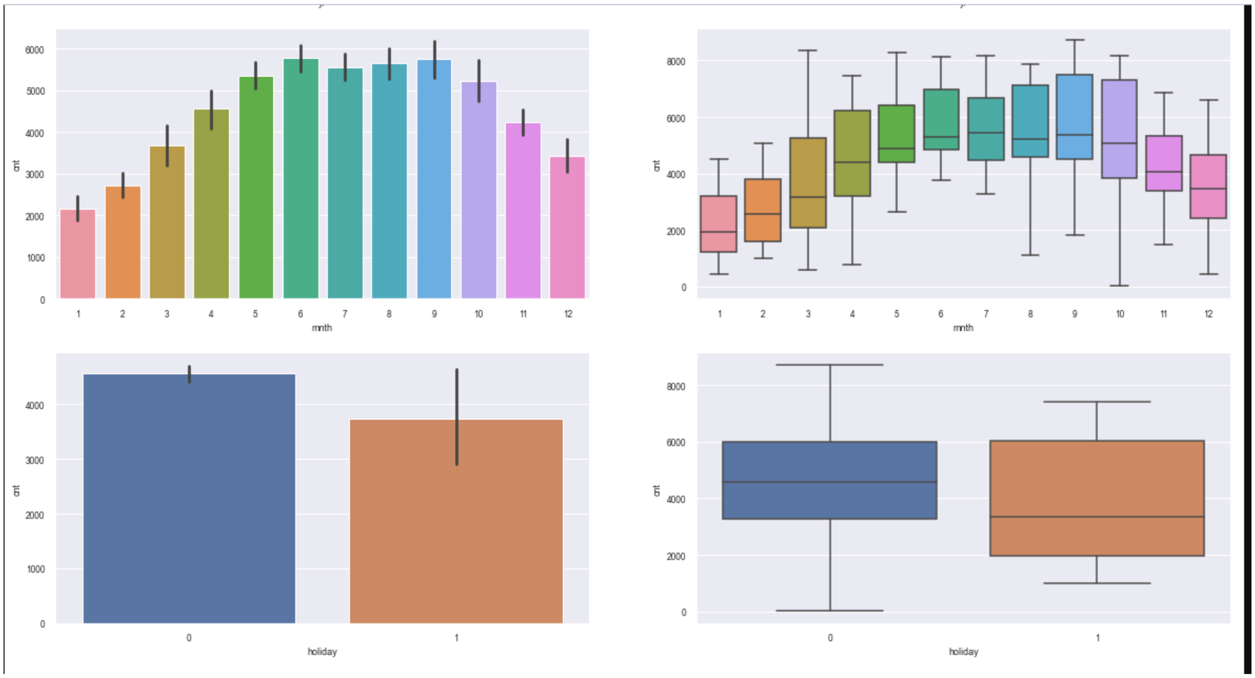
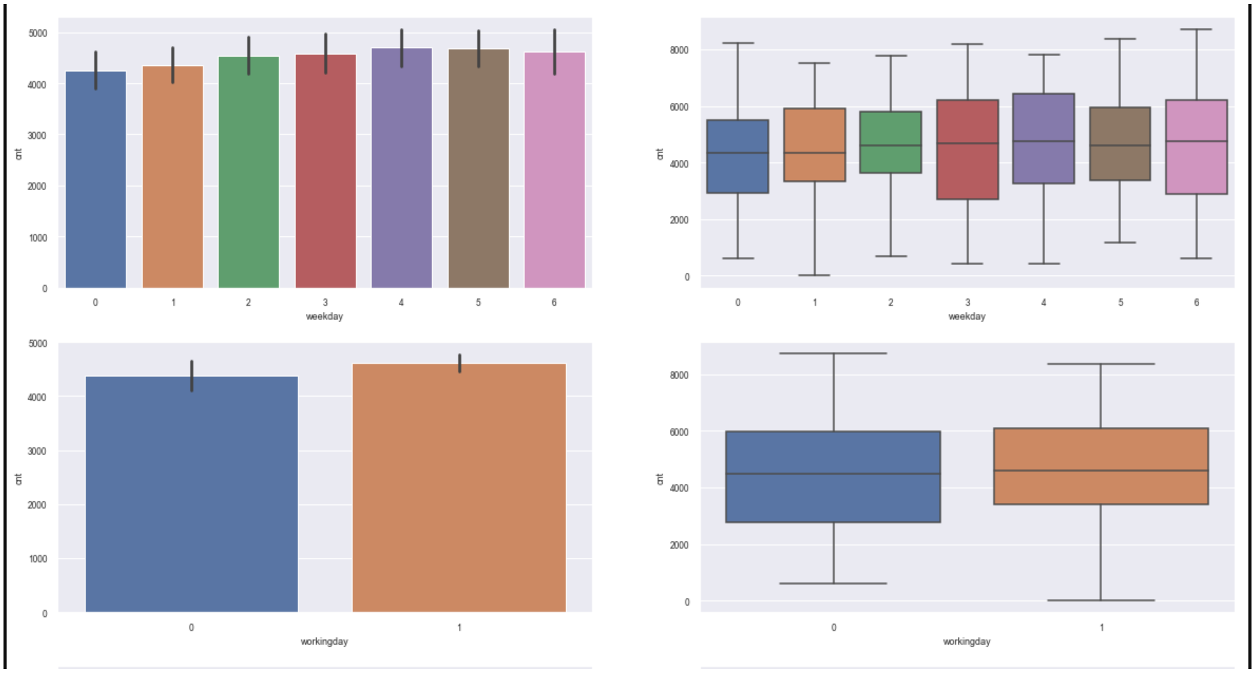
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

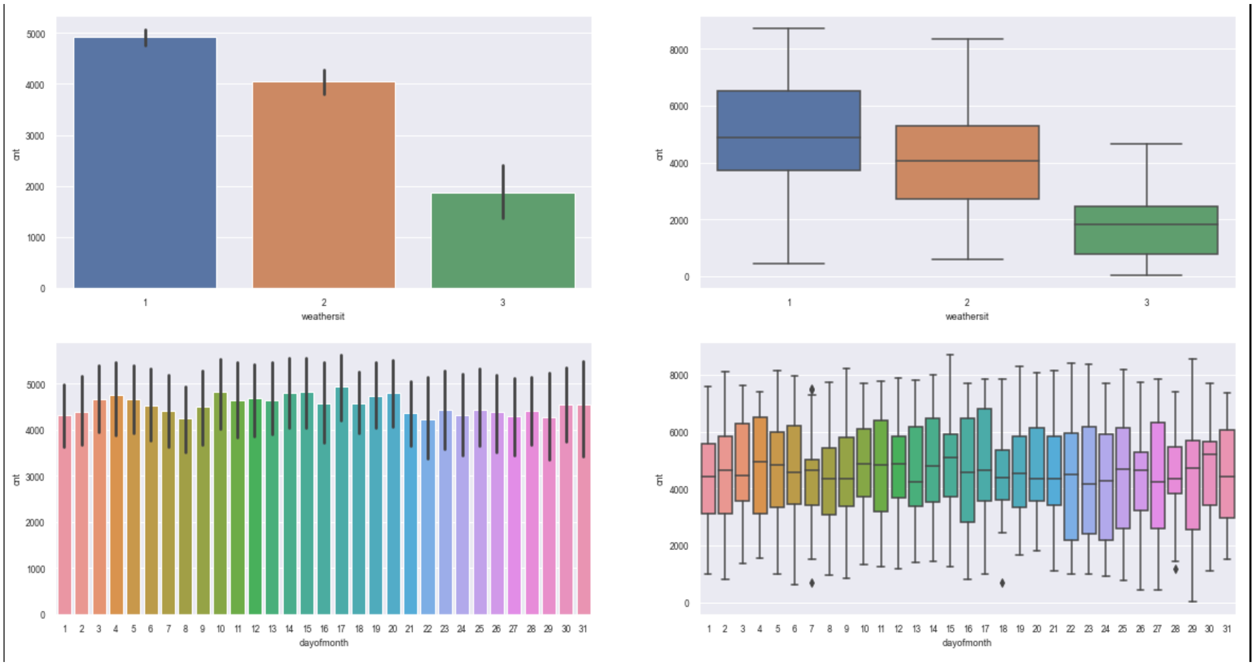
Ans: The important categorical variables in the dataset are 'season', 'yr', 'mnth', 'holiday', 'weekday', 'workingday', 'weathersit'. The observations from EDA are as mentioned below

1. Demand was least in the spring season
2. Demand was the highest in the fall season with median at around 5000
3. Demand in 2019 was higher than 2018. The median was around 6000 which was also the upper whisker of the box plot of 2018.
4. Demand in 2018 was approximately equal to quartile 1 (25%) of 2019 demand as per data points in dataset
5. Month 5 to 10 had good demand and co-relates with season trend observed. Median during these months were near 5000
6. Demand was low during holidays. Median of demand on holidays was equal to quartile 1 (25%) of demand on non-holidays as per data points in dataset
7. 2nd day of the week had Q1 higher than rest of the days. Medians seem to be almost the same throughout the week
8. There is slightly higher demand on working days.
9. Demand was high when the weather was clear with few clouds or partly cloudy
10. Demand was the least with weathersit 3, i.e., light rain or snow.
11. Demand was nil when weather was thunderstorm, heavy rain




cnt
0
1000
2000
3000
4000
5000
6000
1
2
3
4
5
6
7
8
9
10
11
12
mnth
cnt
0
2000
4000
6000
8000
1
2
3
4
5
6
7
8
9
10
11
12
mnth
cnt
0
1000
2000
3000
4000
0
1
holiday
cnt
0
2000
4000
6000
8000
0
1
holiday


cnt
0
1000
2000
3000
4000
5000
0
1
2
3
4
5
6
weekday
cnt
0
2000
4000
6000
8000
0
1
2
3
4
5
6
weekday
cnt
0
1000
2000
3000
4000
5000
0
1
workingday
cnt
0
2000
4000
6000
8000
0
1
workingday



The top three categorical predictor variables and its inference with target variable cnt are as mentioned below

A unit increase in temperature affects the demand by 0.5704 units.
A unit increase in year affects the demand by 0.2351 units.
A unit increase in weather situation affects the demand negatively by 0.2316 units.

Season and windspeed were other significant contributing features.

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

Ans: Let us understand the concept of dummy variable creation using an example. Consider gender as a categorical variable. It can have 3 possible values/categories

Gender
Female
Male
Transgender

Dummy variable creation is the process of encoding the categorical data present in rows and converting it to fields/predictors. For our example, the data can be represented in the below encoded format

Gender	Female	Male	Transgender
Female	1	0	0
Male	0	1	0
Transgender	0	0	1

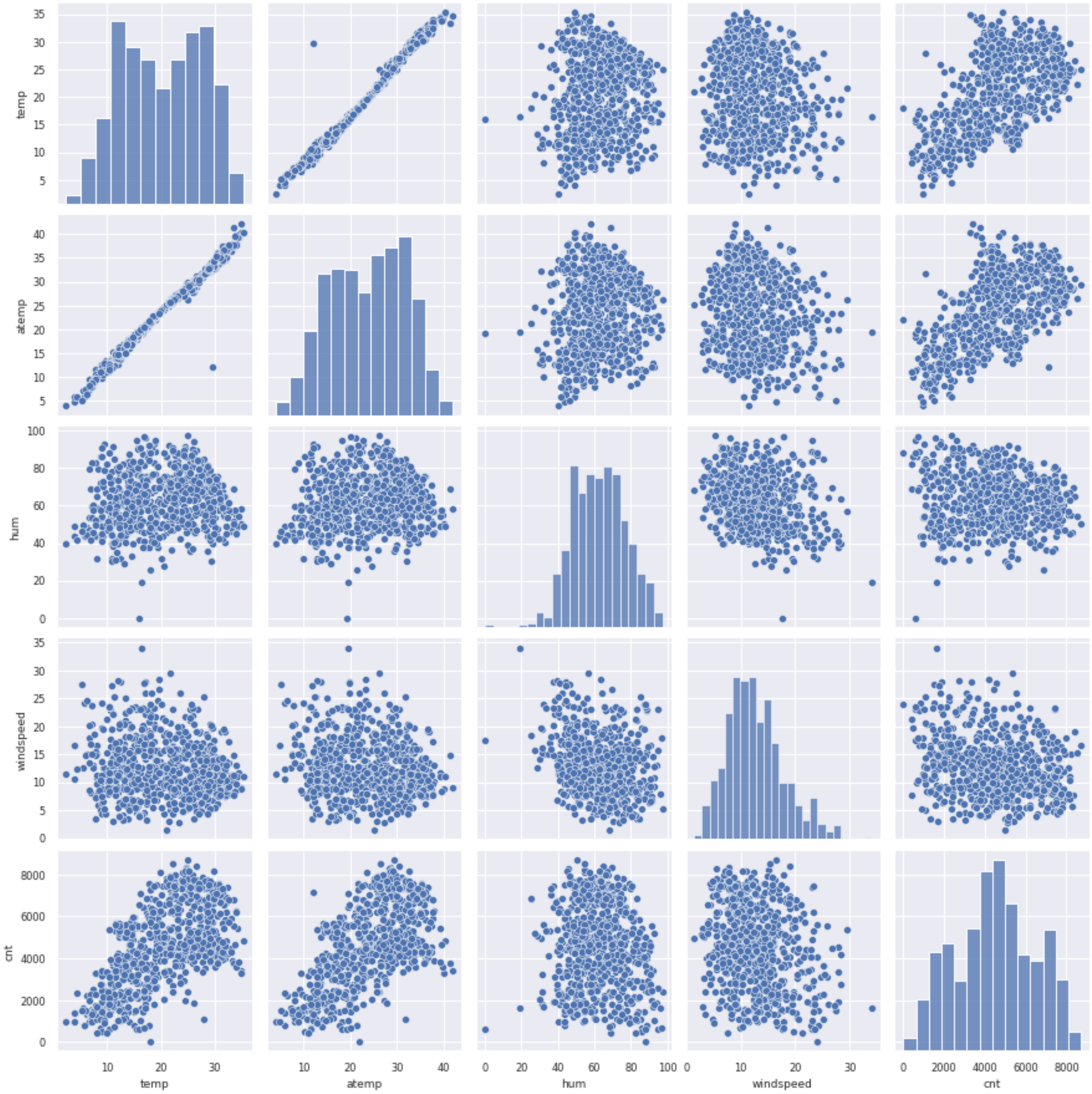
On observation, to represent 3 categories of a single field gender, we created 3 fields Female, Male and Transgender. We can optimize this using the below encoding

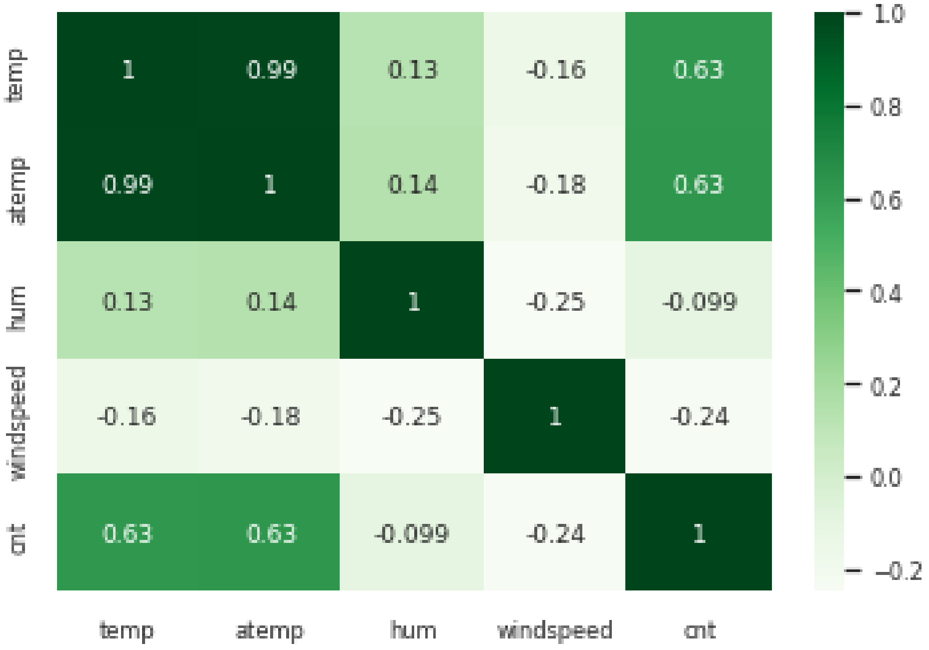
Gender	Male	Transgender
Female	0	0
Male	1	0
Transgender	0	1

Observe, we dropped the first category "**Female**" from its previous table. To infer the gender of a record is female, it is sufficient if we verify the condition "**Male**" is 0 and "**Transgender**" is 0 i.e., for M categories, M-1 fields/predictors are sufficient. The pandas get_dummies method has a parameter drop_first=True, which when set to true, generates M-1 predictors by dropping the first category as seen in the example. Dropping this field helps reduce correlations between predictors of dummy variables.

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

Ans: Highest correlation was observed between the independent variables temp and atemp with the target variable cnt.
The correlation is 0.63 and positive.


temp
atemp
hum
windspeed
cnt
35
30
25
20
15
10
5
40
100
80
60
40
20
0
8000
6000
4000
2000
0
10
20
30
40
0
20
40
60
80
100
0
10
20
30
0
2000
4000
6000
8000



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

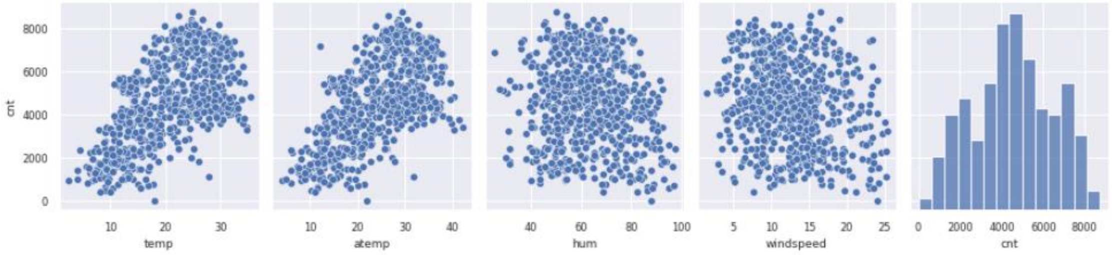
Ans: The assumption for linear regression is as listed below

1. X and Y have a linear relationship
2. Normally distributed error terms with mean zero
3. Error terms should be independent of each other

Additionally, the below concerns are to be taken for consideration when multiple predictors are used.

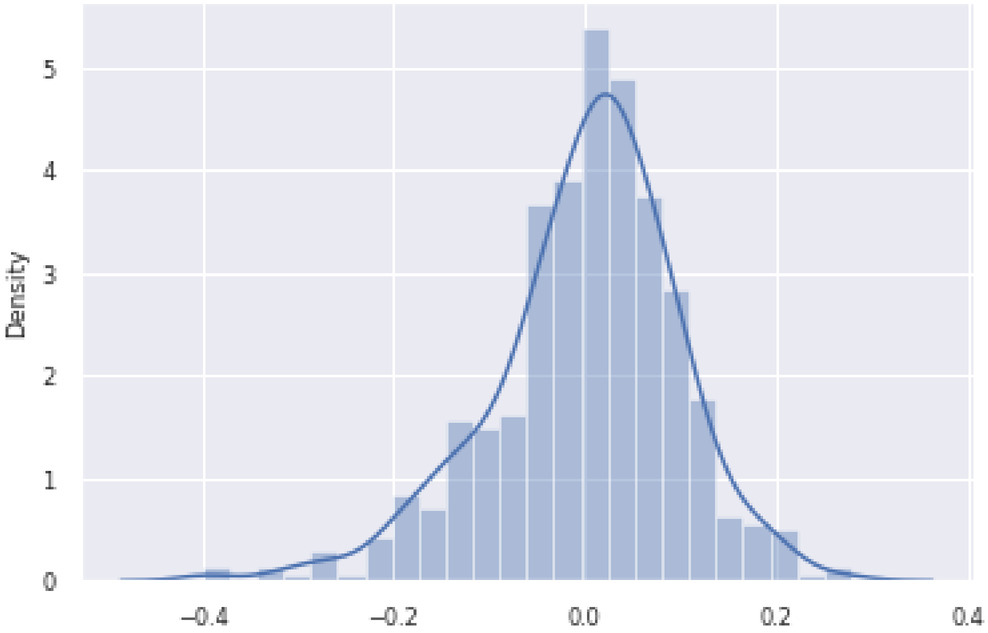
1. Overfitting
2. Multicollinearity
3. Homoscedasticity

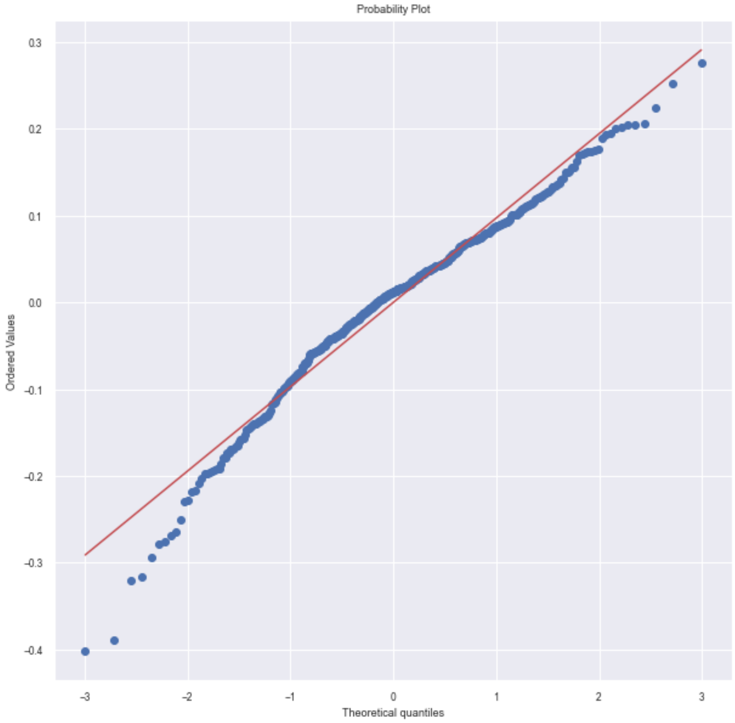
Validated that a linear relationship exists between X(predictors) and Y by drawing a pair plot. If the below scatter plot is observed, there exists some sort of linear relationship between cnt and temp fields.



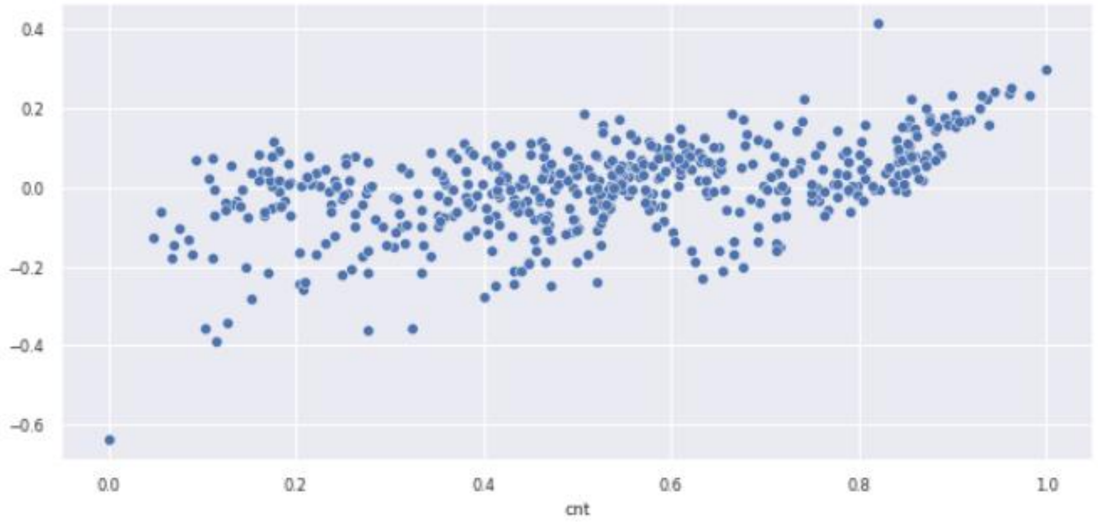
Next, validated that the error terms are normally distributed with mean at zero by plotting a dist and Q-Q plot on the residual.

Residual= ytrain – ytrainpred





Verified the Homoscedasticity by plotting the fitted vs residual scatter plot in which we should not see any pattern.



Additional verification for multicollinearity by using the VIF values (variance inflation factor) was performed. Normally the VIF of all predictors should be less than 5.

Finally, checked the model is overfit by using the test data and calculating the R-squared and adjusted R-squared scores for it post prediction. The scores should be near to that calculated on training set. If its drastically low for the test set, that means the model is overfit. we have considered up to 5% difference as normal.

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

Ans: The demand is explained by these three features

1. temp
2. weather_sit
3. yr

A unit increase in temperature affects the demand by 0.5704 units.
A unit increase in year affects the demand by 0.2351 units.
A unit increase in weather situation affects the demand negatively by 0.2316 units.

Season and windspeed were other significant contributing features.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: The relationship between an independent and dependent variable if explained using a straight line and its equation is called Linear regression.

In machine learning, linear regression is categorized under the supervised learning.

Let us try and understand what the equation of straight line is

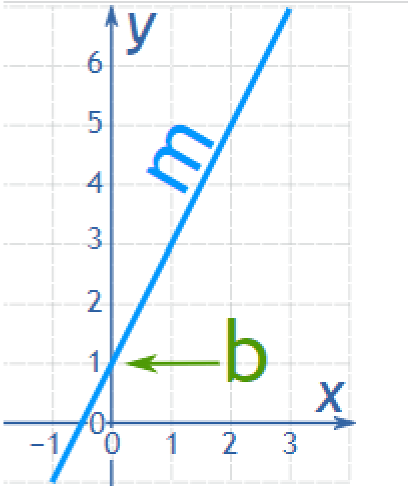
Y=mX + b

where, Y is the target variable also known as dependent variable

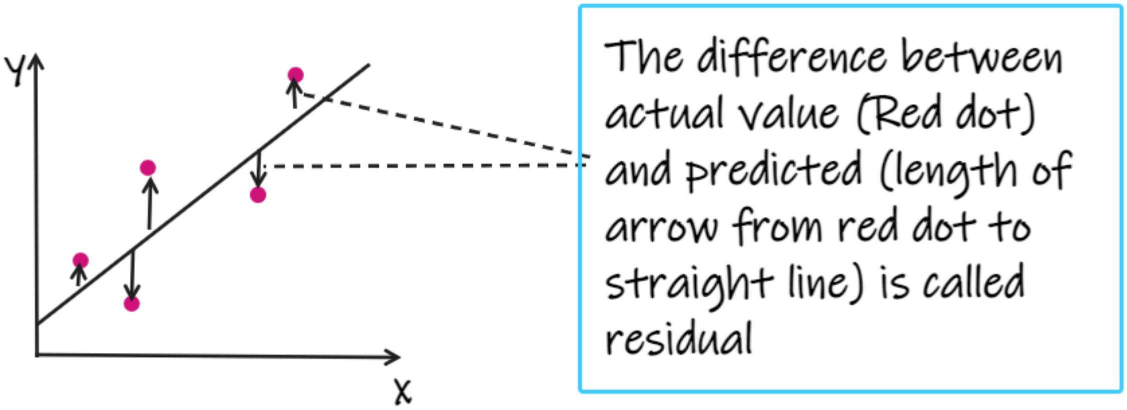
M is the slope (change in Y) / (change in X)

X is the independent variable

b is known as intercept. point common on y axis and the linear line.



Best fit line: There can be multiple straight lines that can be drawn between points on the graph. The algorithm must choose the best fit line. Cost function is used to select the best fit line. Linear regression model uses the least sum of squares of errors as a cost function.


Y
X
The difference between actual value (Red dot) and predicted (length of arrow from red dot to straight line) is called residual

Residual error

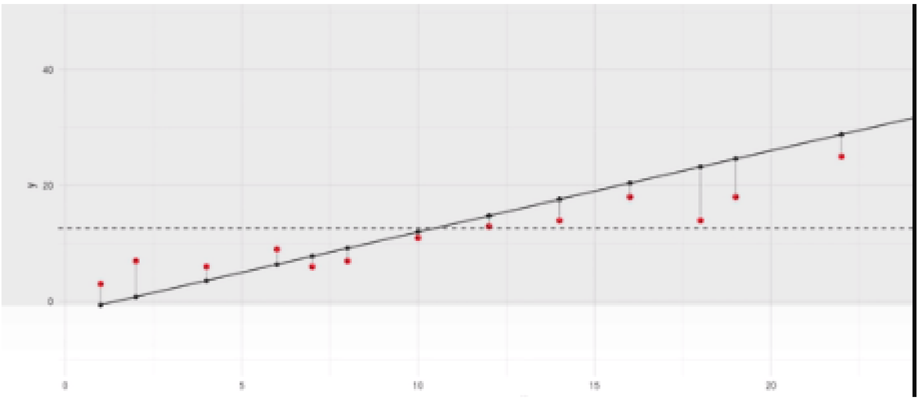
$$e_i = y_i - y_{pred}$$

Residual sum of squares (RSS). RSS is a cost function which is absolute in nature. It is dependent on the unit of both the dependent and independent variable.

$$RSS = \sum_{i=1}^{n} (e_i)^2$$

Total sum of squares (TSS). TSS is a cost function which is relative in nature. It uses the difference between actual and mean value of the dependent variable instead of the predicted value.

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$



where n is the sample size
The quality of the model is measured using R-squared score.
R-squared= 1- (RSS/TSS)
A good model is represented by a R-squared of 1.

If the concept is extended to have multiple independent variables or predictors, it is called multiple linear regression. In the case of multiple linear regression adjusted R-squared is used which punishes for including predictors which do not improve the model than by chance.
Let us consider the bike demand assignment as an example to understand multiple linear regression and assume the model equations
demand = 0.0766 +(0.5704) * temp -(0.0899) * windspeed +(0.0733) * season_2 +(0.1272) * season_4 +(0.2351) * yr_1 +(0.0901) * mnth_9 -(0.082) * holiday_1 -(0.2316) * weathersit_3

This equation is of the form, Y= b + m1X1 -m2X2 + ….
Here m1 and m2 are coefficients and X1, X2 are the independent variables.
The coefficients describe the relationship between each independent variable and dependent variable mathematically

The signs positive and negative in the equation represent the correlation is positive or negative. In the example, A unit increase in temperature increases the demand by 0.5704 times. Similarly, a unit increase in windspeed decreases the demand by 0.0899 units. Using this equation, coefficients, and predictors the model predicts/easily calculates the demand.
Assumptions of linear regression:

1. X and Y should have a linear relationship.
2. Normally distributed error terms with mean zero
3. Error terms should be independent of each other

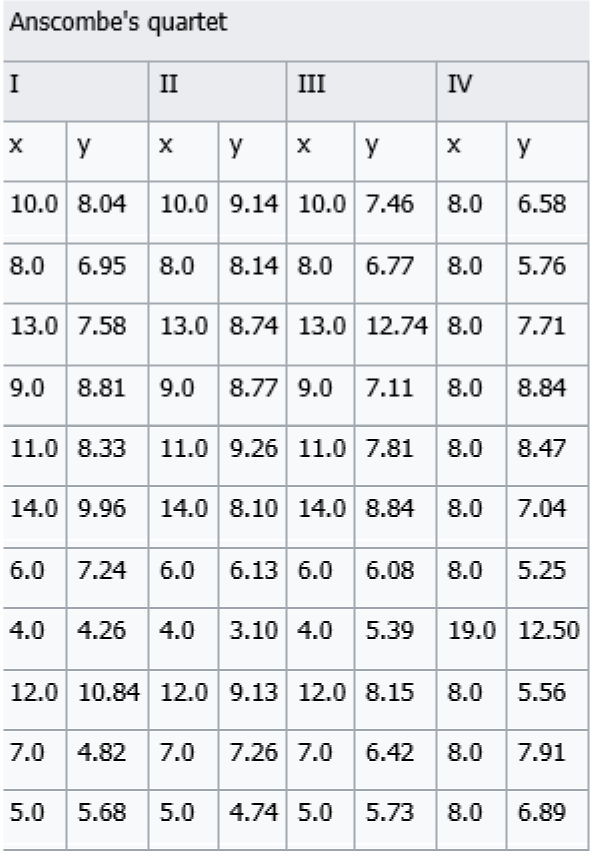
Additionally, the below concerns are to be taken for consideration when multiple predictors are used.

1. Overfitting
2. Multicollinearity
3. Homoscedasticity

2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's Quartet comprises of four datasets. There is one dependent(Y) and one independent variable(X) in each of these datasets. The actual data in these datasets are different but when descriptive statistical data summary is generated, it is identical.

Consider the quartet as shown in the table below



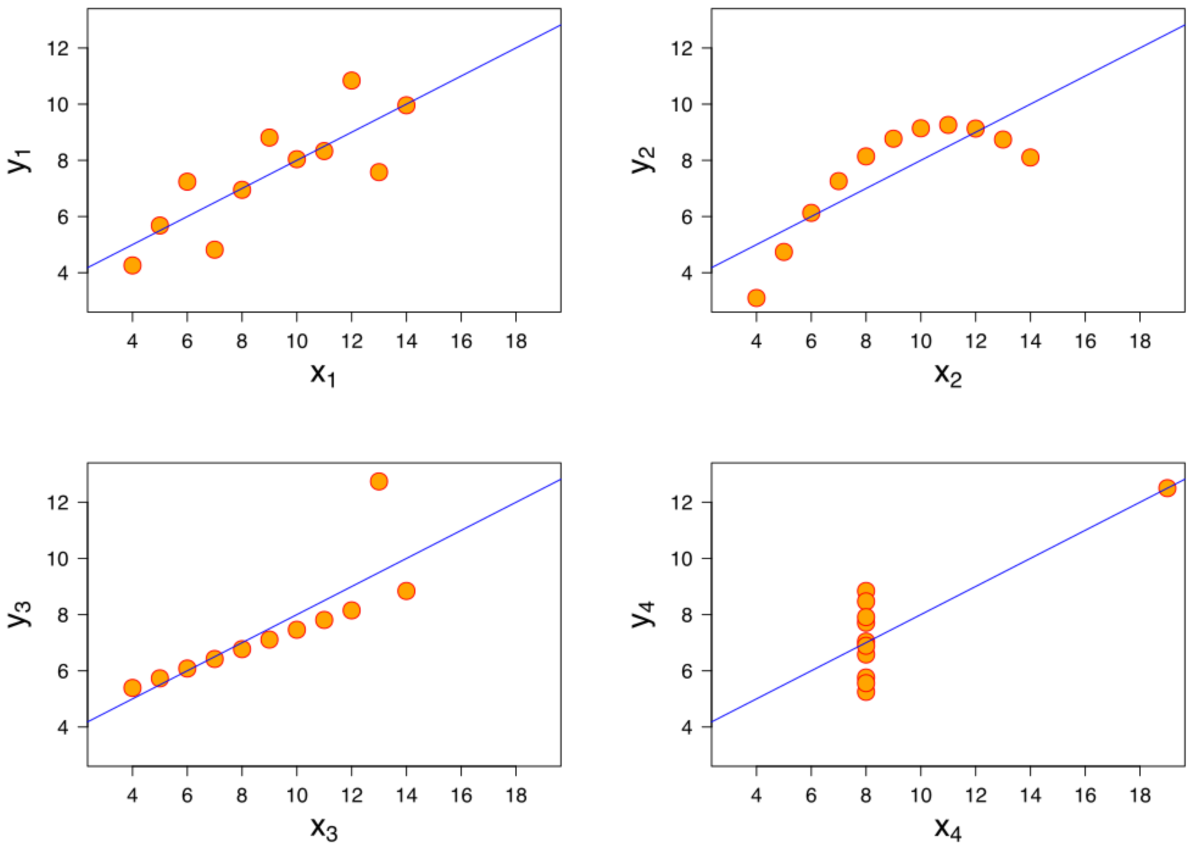
Anscombe's quartet

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

When stats are generated, for all the four datasets individually, we find that they are same as described below.

- Mean of x is 9
- Mean of y is 7.50
- Sample variance of x is 11
- Sample variance of y is 4.13
- Correlation between x and y is 0.816. Positive correlation.
- $y = 3.00 + 0.500x$ is the equation of straight line. 3 is the constant and a unit increase in x increase y by 0.500 units

However, when we use data visualization to plot graphs, these statistically identical datasets look completely different.



The first plot from the left depicts a simple linear relation between X1 and Y1.
The second plot on the top right depicts some relationship between X2 and Y2 but it is not linear. It can be observed that up to certain value of X2, the Y2 increases but at X2=8, this trend starts to reverse.
The third plot on the bottom left corner, depicts a linear relation but should have a different regression line (not the best fit). The calculated regression line is offset by an outlier which has enough influence to reduce the correlation between X3 and Y3 from 1 to 0.816.
The fourth plot on the bottom right corner, depicts that one high leverage point or outlier was enough to change the correlation to high although other data points in the dataset do not indicate any relationship between X4 and Y4.

The quartet demonstrates the inadequacy of basic statistics for describing datasets. It stresses on importance of usage of graphical analysis to determine type of relationships in the dataset. Statistician Francis Anscombe constructed the quartet in 1973.

3. What is Pearson’s R?

Ans: It was developed by Karl Pearson in 1880’s and is also known as the bivariate correlation or the correlation coefficient in statistics. It is a measure of the linear correlation between two data sets. It is the ratio between the covariance of the variables X and Y to the product of their standard deviation.

For a population

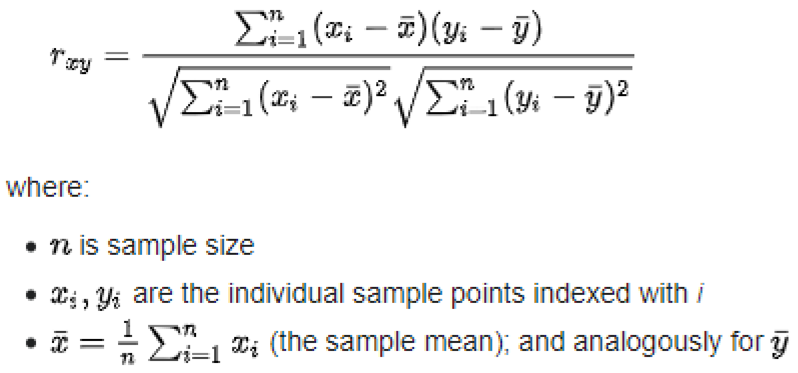
When applied to a population it is referred to as the population correlation coefficient. Given a pair of random variables X,y the Pearson’s population correlation coefficient is defined by the formula

$$\rho\ (X,y) = COV(X,y)/(\sigma X * \sigma y)$$

Where, COV is the covariance, .σX and σy are the standard deviations of X and y respectively

For a sample

When applied to a sample, is represented by r(X,y) and is referred to as the sample correlation coefficient.



$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

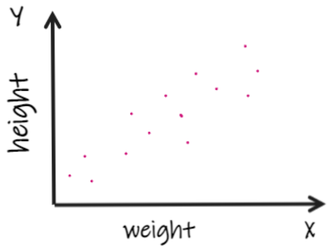
where:

- 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 $\bar{y}$

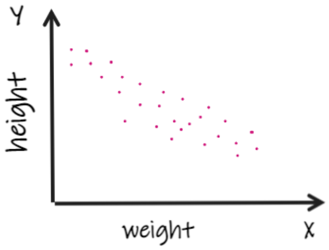
The coefficients values lie on or between −1 and 1. The sign + means the datasets are positively correlated and - means they are negatively correlated.
For sample correlations equal to +1 or −1 correspond to data points lying exactly on the line and in case of a population correlation, bivariate distribution entirely supported on a line.
If the correlation is 0, it signifies that the datasets are not correlated.

The Pearson correlation coefficient is symmetric i.e., correlation of (X, Y) = correlation of (Y, X).

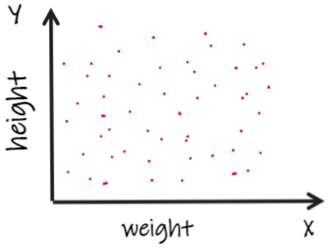
Example: Consider the datasets depicting height and weight of people. The scatter plots drawn depicts the various types of correlation between these datasets.



Positive



Negative



No Correlation

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

Ans: Scaling is a data pre-processing step. It is optional yet recommended for interpretation and faster convergence in the case of Gradient descent methods.

Example: Consider two predictors in the dataset represent time in seconds and minutes. On interpretation, it feels that 5 minutes is less than 2000 seconds which is false. To avoid confusion and for better interpretation feature scaling is recommended.

There are two methods to scaling

Normalization

Normalization is the process of compressing the data in the range of 0 to 1. Min-Max scaling is used for normalization. It uses the below formula for scaling

$$X_{\text{new}} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

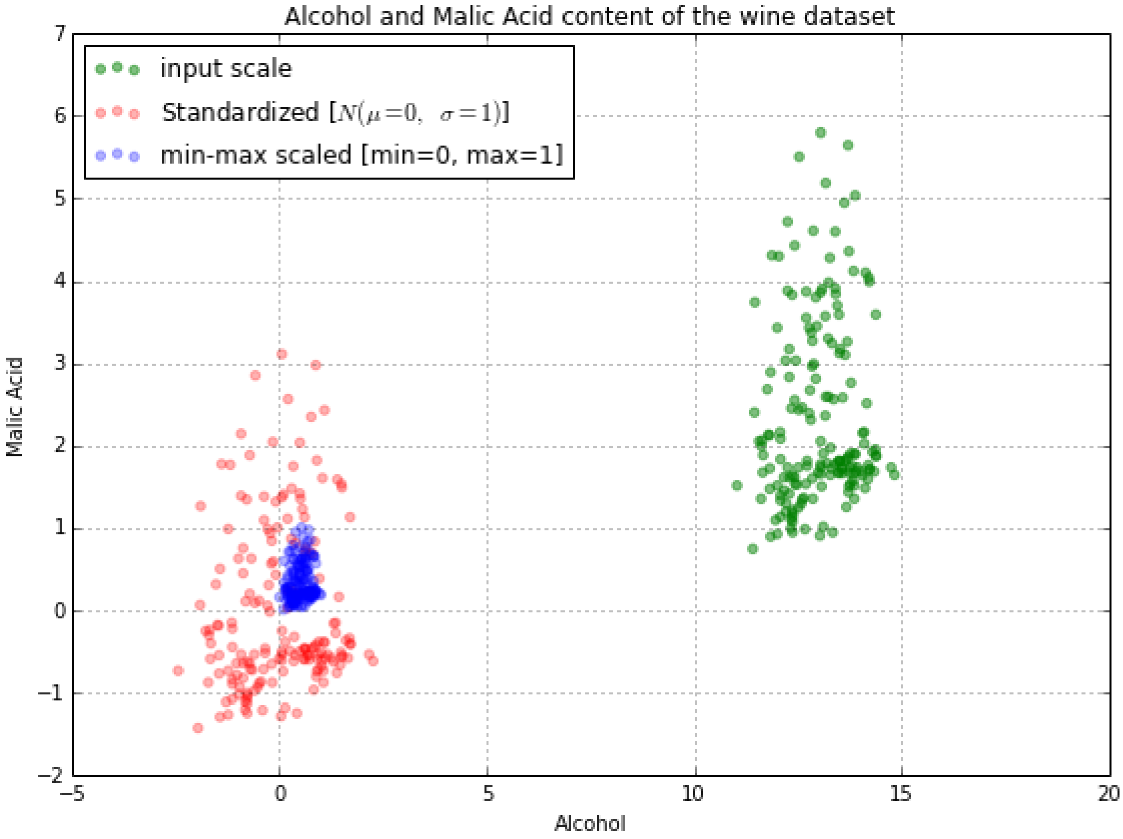
Standardization

Rescales the data into a standard normal distribution with mean at zero and standard deviation or variance equal to one.

$$X_{\text{new}} = \frac{X_i - X_{\text{mean}}}{\text{Standard Deviation}}$$

Advantage of standardization over normalization is the prevention of data loss especially the outliers.

The normalized values are shown in dark blue and are clustered tightly compared to standardized distribution shown in pink. Standardization helps in the case of algorithms such as gradient descent to converge faster to same solution compared to normalization.



Feature scaling affects only the coefficients and does not change parameters such as P-value, t-statistics, F-statistics, R-squared. It does not have an impact on the prediction of target variable and only affects the interpretation.

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

Ans: Variance inflation factor measures the extent of multicollinearity. It explains to what extent a predictor is explained by other predictors in the model. It plays a role only when model inference/interpretation is desired in the form of coefficients and p-values and does not affect the prediction, best fit or r-squared score of the model. Addition of more predictors can improve a model never decrease its R-squared. If we already have capability to predict something with certain accuracy, we can only improve and not degrade it by introducing additional predictors. However, introducing unnecessary predictors which does not add any value to the model should be avoided to save on compute, memory cost associated with handling that extra predictor during learning and prediction.

The formula for VIF is 1/(1-R^2)

For an infinite VIF, we need zero in the denominator which is only possible if R-squared is 1. When R-squared is 1, it signifies a strong correlation between the predictor and other predictors in the training dataset.

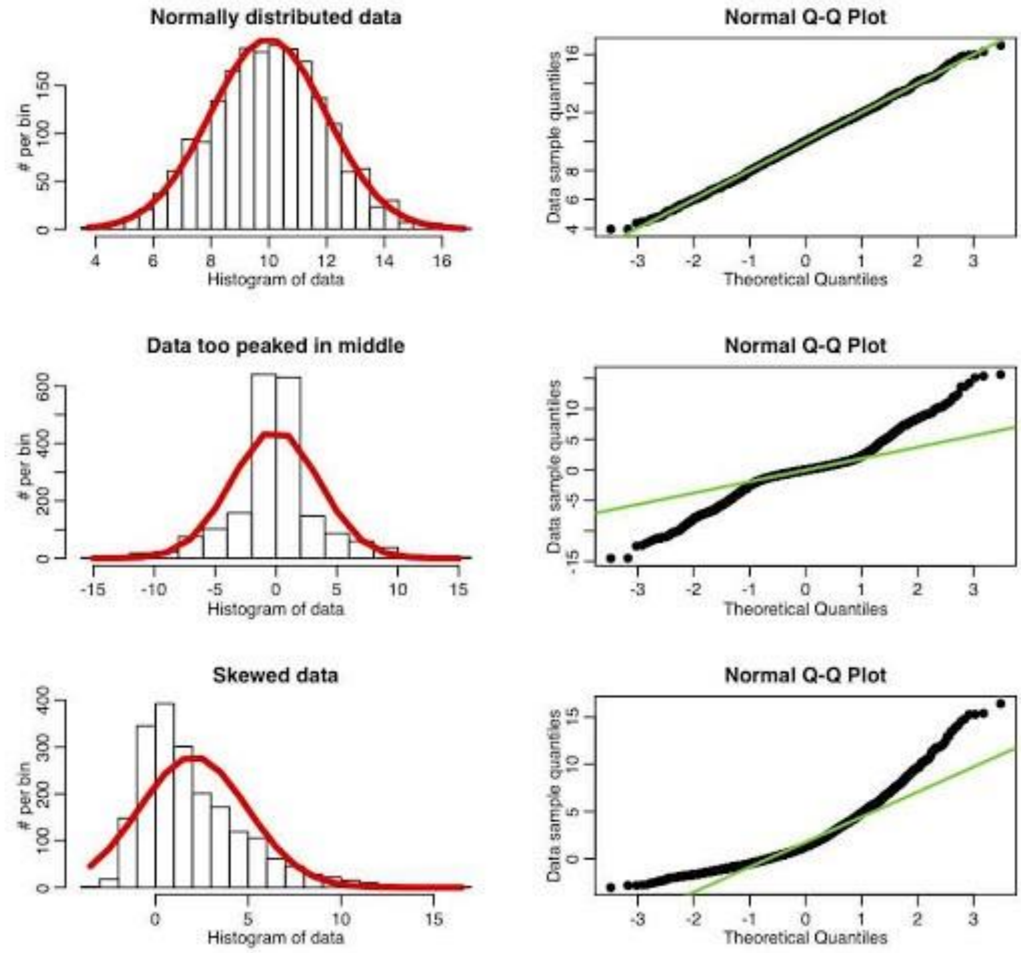
For example, if we consider two predictors temp and ftemp where temp is the current temperature in Celsius and ftemp is the current feels like temperature. These two variables have a near perfect correlation of 0.99 or 1. If we include both variables, R-squared will be near or equal to 1 and VIF will become infinity.

As a thumb rule, VIF above 10 is definitely very high. VIF above 5 and below 10 should be investigated in detail for its value to the overall model and prediction, coefficients and p values may become unreliable for inference. VIF below or equal to 5 is generally considered good.

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

Q-Q(quantile-quantile) plots is used for graphical analysis and to compare two probability distributions. It is created by plotting their quantiles against each other. They are used to detect the type of distribution for a random variable.

In a Q-Q plot, the data points will lie perfectly on the straight line if the distributions are equal. i.e., y=x
Visualization of the first distribution plot reveals a normal distribution. The data points are on the 45-degree straight line in the corresponding Q-Q plot.
The second distribution plot feels distributed normally. On closer inspection, the data has peaked in the middle. The start and end portion are scattered away from the straight line in the corresponding Q-Q plot. This trend signifies that the distribution is not normal, and we cannot say clearly that the X and Y variables are correlated or have a relationship.
Skewed data is visible in the third distribution plot. The data points are scattered away from straight line in the corresponding Q-Q plot. This trend signifies that the distribution is not normal.
Based on whether the bottom or the upper end has deviated, we can conclude if the distribution is left or right skewed.



In linear regression, we assume that the residual error terms (yactual - ypredicted) are a normal distributed with mean at 0 with constant variance.
Y=c + mx + error
Where Error= y-y predicted

If the above statement is false, the variance becomes a function of independent variable x. This condition is known as heteroskedasticity
The fit, stability and reliability are questionable. The calculated probabilities are valid due to the assumption and impacts the inferences or interpretations.