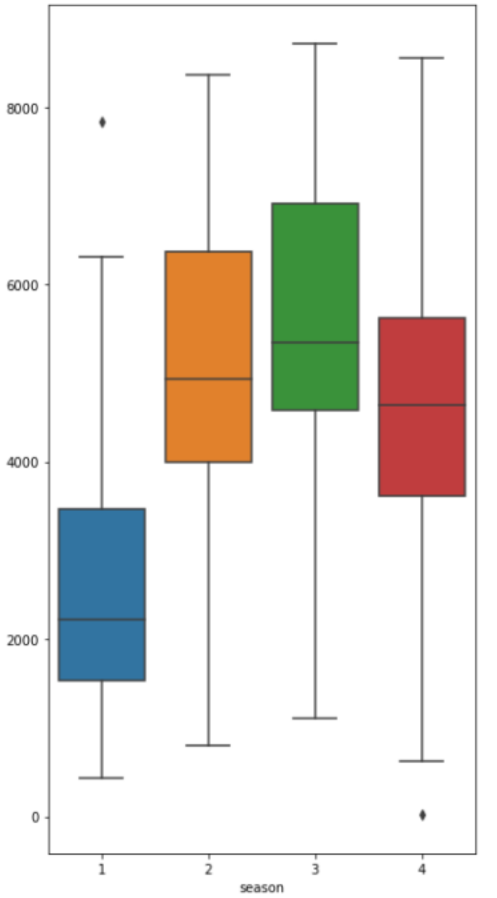
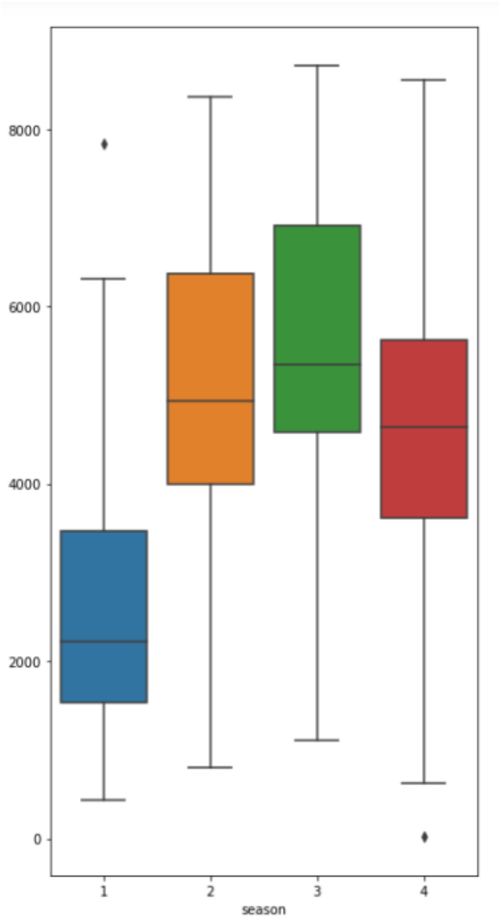
Assignment-based Subjective Questions

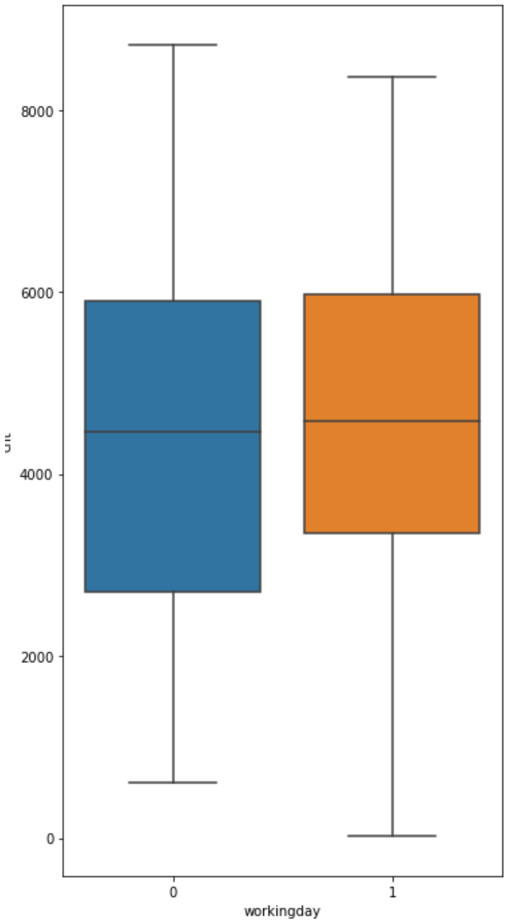
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**
    - We have a 3 categorical variables season, mnth, workingday
    - From **Season** We have a two outlier From Spring and Summer,
    - We Saw drastically change in cnt from spring to other season In spring median is around 2100 and summer, fall and winter median is around 4300 to 5000



- Month vs Count is a extended graph of season vs count From month 1 to 7 count is increasing and then decreasing.



- From Working day vs count we there is no drastic change in count



2. **Why is it important to use drop_first=True during dummy variable creation?**
   When you have a categorical variable with say 'n' levels, the idea of dummy variable creation is to build 'n-1' variables. From the Assignment Example From Season we have a ‘4’ levels spring, summer, fall, winter if we drop_first is spring than we conclude that if summer, fall and winter are Zero then that is spring. So We don’t need to carry one extra column for the model preparation.
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   - Atemp
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   - According To definition of Linear Regression linear regression is the simplest regression model which involves only one predictor and also which is not categorical one.There are only one predictor involve in the dataset. Our target variable is count(cnt). Which is not categorical one
5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes**
   Atemp,summer,winter

General Subjective Questions

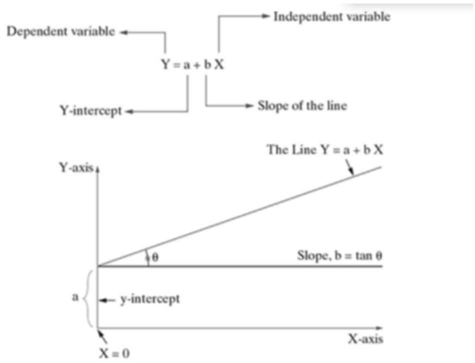
1. **Explain the linear regression algorithm in detail.**

There are Two Type of linear regression

- Simple Linear Regression
- Multiple Linear Regression

Simple Linear Regression

simple linear regression is the simplest regression model which involves only one predictor. This model assumes a linear relationship between the dependent variable and the predictor variable.



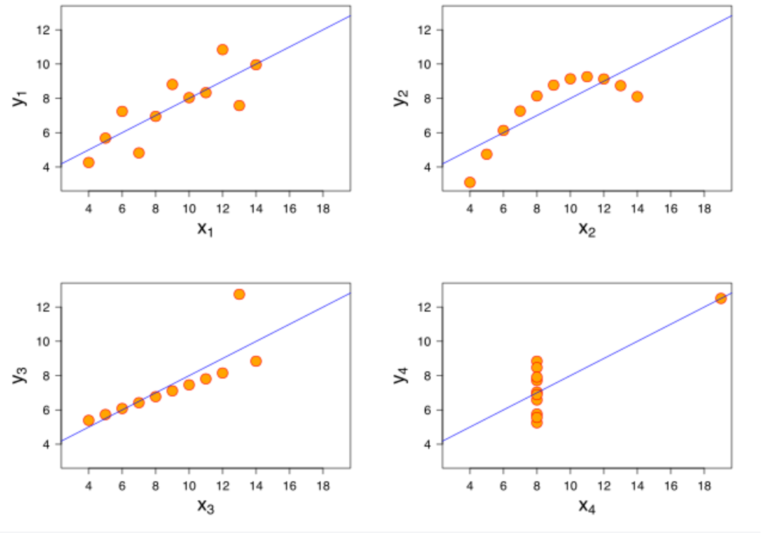
Here The Formula is y = mx + C form where C is a constant and y is a target variable of the Dataset. And x will be any independent variable which is correlate with dependent variable y And m is a slope is equal to tan A where A is a Angle. We can observe in the figure Y = a + bX and here b is a slop.

Multiple Linear Regression

Multiple Linear Regression is a basically extended form of Simple Linear Equation. Here Formula become y = $\sum mx + C$. There are multiple dependent variables which is correlate with the independent variable.

2. **Explain the Anscombe's quartet in detail.**

Anscombe's quartent is a way of to know the linear regression model is fit to linear regression or not. Acording to Wikipedia **Anscombe's quartet** comprises four data sets that have nearly identical simple descriptive statistics.



Here x1 is appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on *x*.

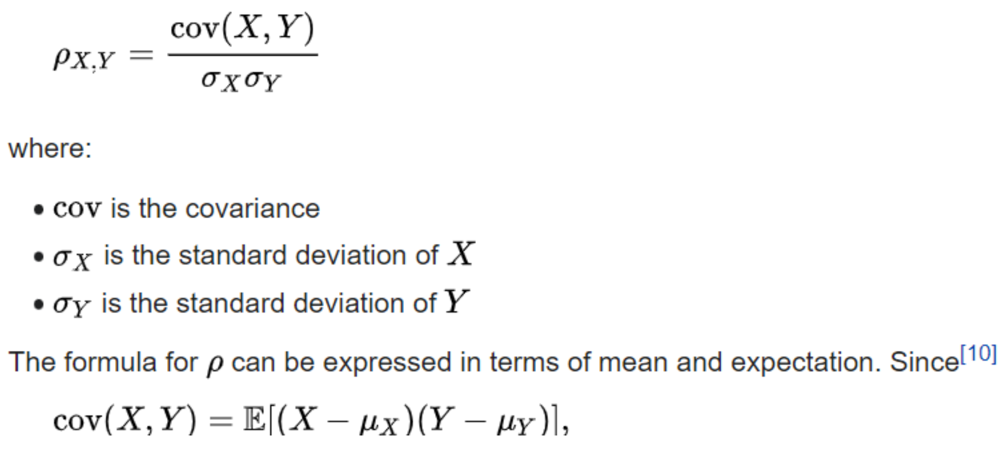
Here x2 is not a normally distributed. while a relationship between the two variables is obvious, it is not linear,

Here x3 and x4 has a outlier in linear relationship.

3. What is Pearson's R?

According to Wikipedia Defination of Pearson's R which is known as **Pearson correlation coefficient** (PCC) is

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier *product-moment* in the name.



$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where:

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

The formula for ρ can be expressed in terms of mean and expectation. Since[10]

$$\operatorname{cov}(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)],$$

Source :Wikipedia

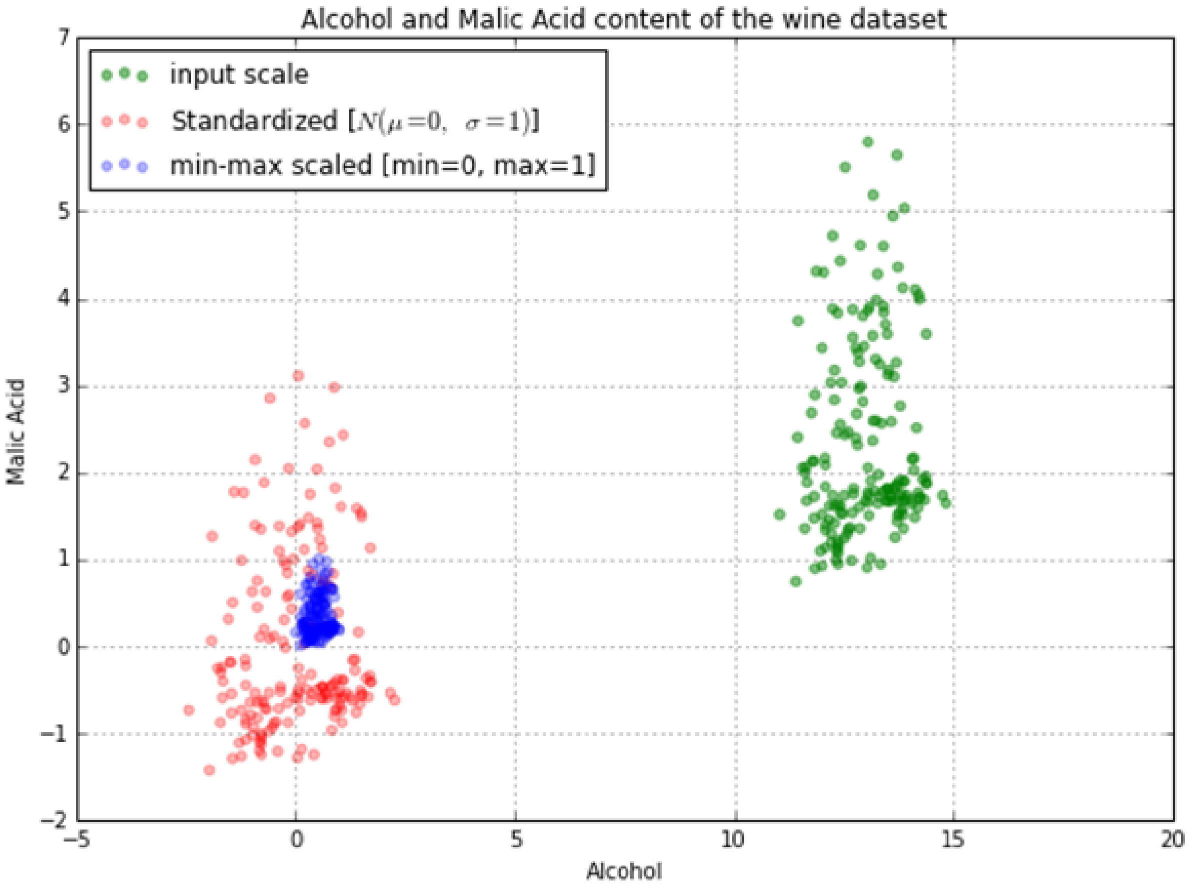
Importance in Machine Learning Model

PCC is help to find out relationship between two variable. Value of PCC is lies on close interval [-1,1]. Here 1 means that they are highly correlated and 0 means no correlation. -1 means that there is a negative correlation. Its helps us to decision to take dependent variable in model or not ?

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**
   - **Scaling :**

scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.

Refer From www.atoti.io



We observe that data is real data(input scale) lies between open interval (0,6). Some times it is difficult to handle data when the range is high. So we scale the data to minimize the range of the data.

Range = (max - min)

Normalisation :

Normalisation basically convert data in close interval [0,1]. Normalisation is Known as Min-Max Scaling or Min-Max Normalisation.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Standardization

Normalisation basically convert data in close interval [-1,1].

$$x' = \frac{x - \bar{x}}{\sigma}$$

There is no specific rule when use normalisation or standardization while make machine learning model. We can just hit and try the model show which model is more fit.

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

Formula of VIF is

$$VIF_i = \frac{1}{1 - R_i^2}$$

The Simple rule of mathematics is if denominator is zero than value become Infinite.

So that if R^2 is tends to 1 than 1 - R^2 become approximate 1 – 1 = zero

if $R^2 \rightarrow$ 0 than

$$\approx \mathbf{1 - 1 = 0}$$

$\boldsymbol{VIFi} = \frac{\mathbf{1}}{\mathbf{0}}$ = inf

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

Q-Q Plots is called as Quantile-Quantile plots. Q-Q plot helps to find the data is normally distributate or not ?

Suppose we have 1000 data of variable X that is X1,X2, …. X1000.then we follow several steps

Step 1 : sort the data.

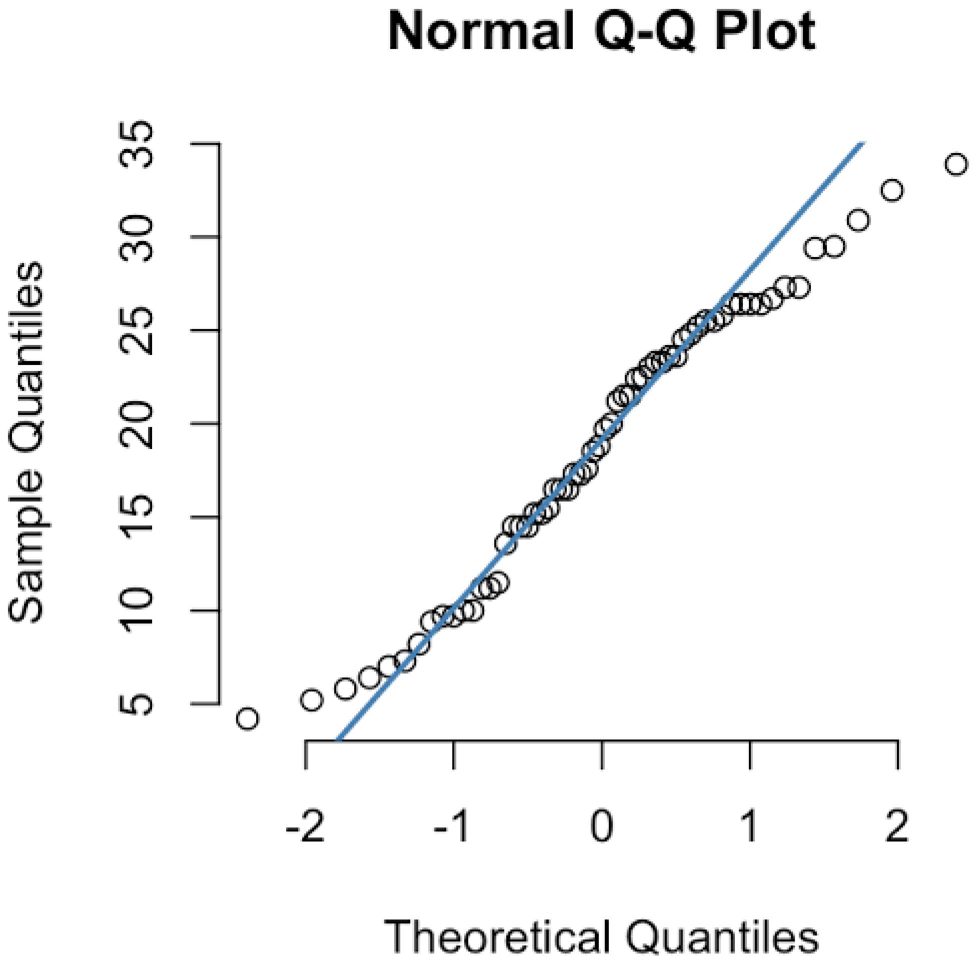
Step 2 : Find 50th percentile (median)

Step 3 : Create random variable y which is normally distributed(mean =0 and standard deviation = 1)

Step 4 : take 1000 observations from y and sort them and calculate 50th percentile (median)

Step 5 : Plot X and y as Q-Q Plot

If X vs y tends to be a linear hence X is homomorphic to y such that X dataset is normally distributed.



Reference : https://medium.com/analytics-vidhya/q-q-plots-scatter-plots-pair-plots-where-to-use-how-to-use-d16c682e33a2