**Data Preparation & Visualisation**

# Process of acquiring raw data

First of all I searched dataset related to agriculture. As my task is related to Ireland so I searched datasets on Ireland official website I found many datasets but I choose crop production.

I get two detests related to Ireland because both detests are same. But here problem is that these datasets cannot be used for country comparisons as it contains the data of only Ireland so I got another dataset which is crop production international dataset.

So there are three datasets which I got from internet sources

# Positive and negative aspects of the research and acquisition

Positive aspects:

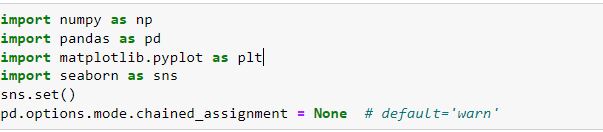
* We know about what is going what are trends in countries.
* We can analyse that a country on which position by exploring data.
* We can make good decisions

Negative aspects:

* It is very time consuming process to acquire datasets.
* data may be incomplete or inconsistent so we have to structure it .
* it can be used for wrong purposes.

# Importing libraries:

The first step is to import all required libraries.so here I am importing four Major Libraries which are here:



1)**pandas**:

pandas is used for data crunching.

2)**numpy** :

It is used for scientific computing it provides some mathematics and statistics like functions.

3)**Matplotlib:**

It is used for data Visualization For example to make Charts and graphs.

4)**Seaborn**:

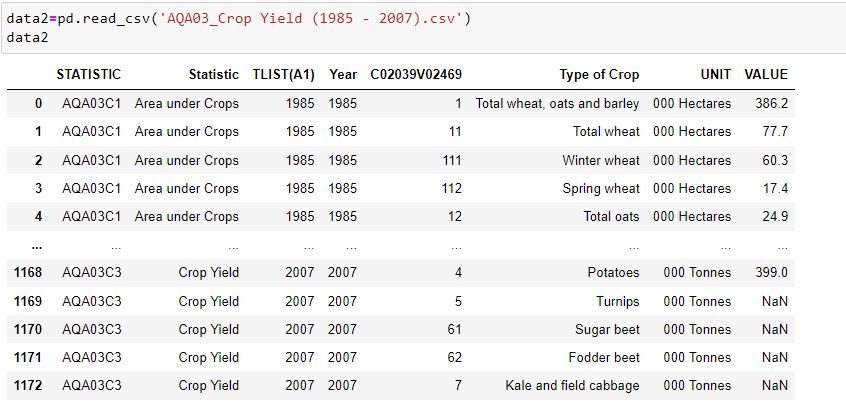
It is also used for data Visualization For example to make Charts and graphs.

***Importing first dataset***

Now here I am loading data. As data is in excel file so I am importing Excel File with .csv file format. ***t***

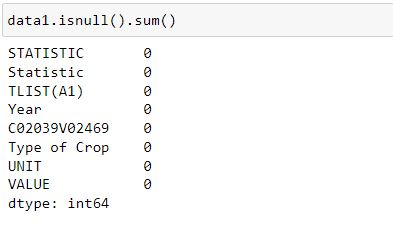


***Importing second dataset***

******

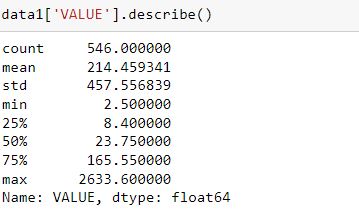
***EDA***

***Checking for any missing values***

******

As shown above there are no null or missing values in data.

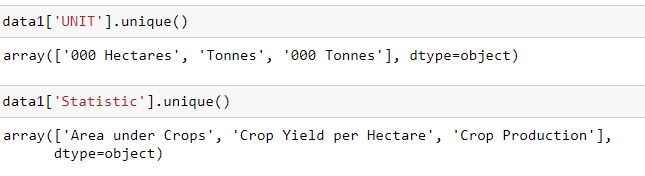
***Description of data1***

******

As by looking on mean and median it is clear that there is huge difference is observed so this show that data is inconsistent.

So it is needed to structure data.

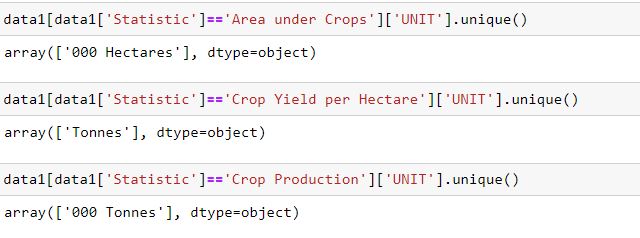
***Exploring unit and Statistic feature***

******

As analysing above there are three units and three types of measurements. It is must that unit must be same. Here above the huge difference in mean and median is observed due to this reason.

But here problem is that every unit belongs to different quantity not same quantity.

***Checking every unit based on every measure***

******

As looking above following is concluded:

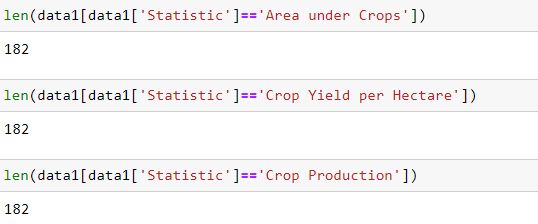
Area under crop has unit of hectares

Crop yield per Hectare has unit of Tonnes

Crop production also has unit of Tonnes but there is difference of some characters

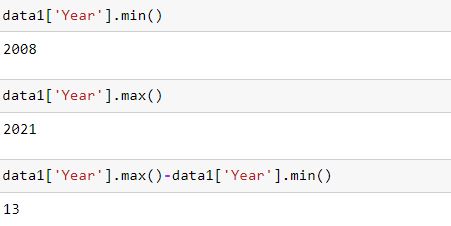
So three measures representing three units

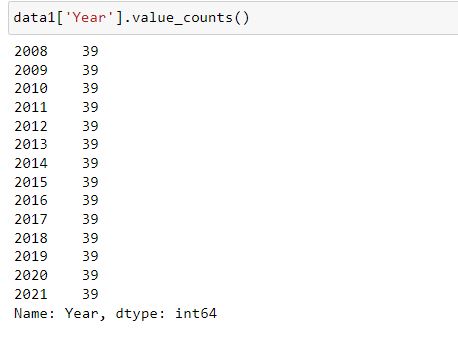
Checking Length of every measure

******

As observing above it is cleared that every measure have 182 records

***Exploring Year***

******

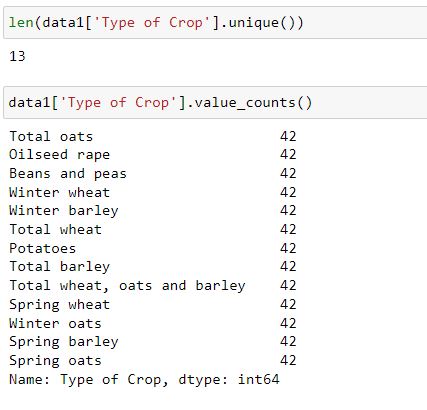
******

This is agriculture record from 2008 to 2021

So this is record contains 13 years

Every year having same number of records as 39

***Exploring type of crop feature***

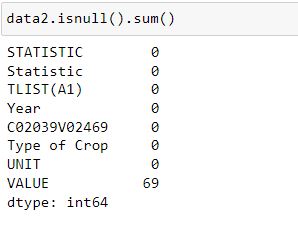
******

There are 13 types of crops are mentioned in dataset

Every crop having same number of records as 42

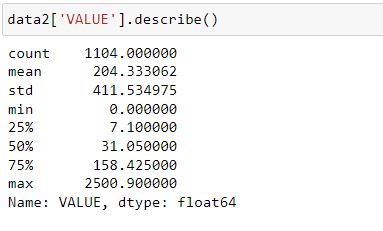
***Exploring 2nd dataset***

***Checking for null values***

******

There are 69 null values present in 2nd dataset

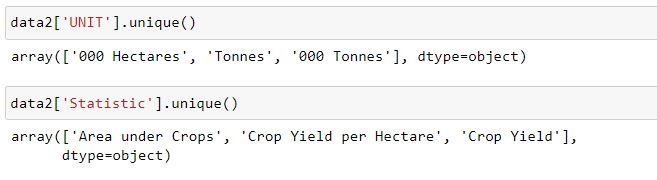
***Description of data2***

******

As by looking on mean and median it is clear that there is huge difference is observed so this show that data is inconsistent.

So it is needed to structure data.

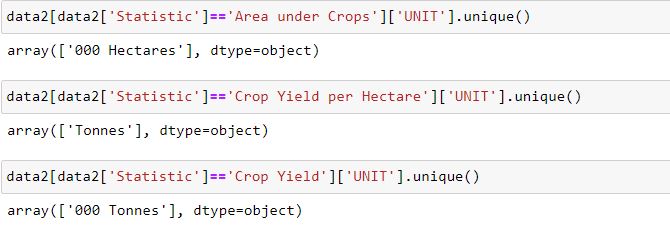
***Exploring unit and Statistic feature***

******

As analysing above there are three units and three types of measurements. It is must that unit must be same. Here above the huge difference in mean and median is observed due to this reason.

But here problem is that every unit belongs to different quantity not same quantity.

***Checking every unit based on every measure***

******

As looking above following is concluded:

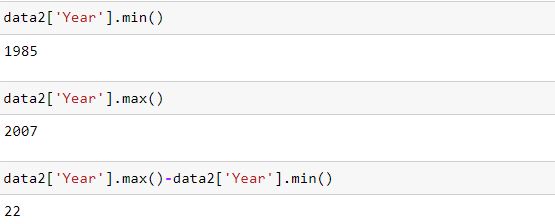
Area under crop has unit of hectares

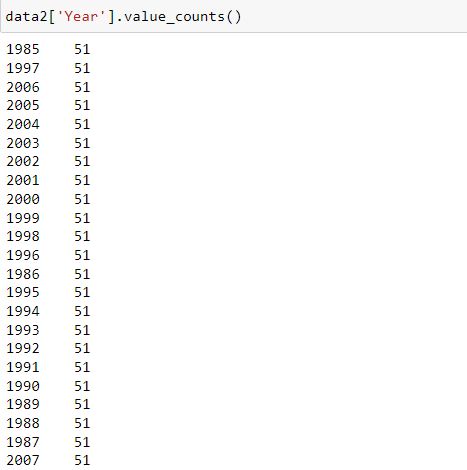
Crop yield per Hectare has unit of Tonnes

Crop Yield also has unit of Tonnes but there is difference of some characters

So three measures representing three units

***Exploring Year***

******

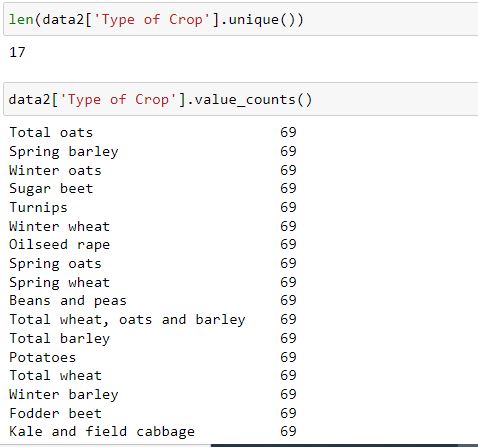
******

This is agriculture record from 1985 to 2007

So this is record contains 22 years

Every year has 51 records.

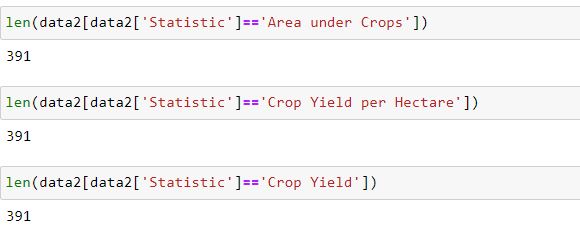
***Exploring type of crop feature***



There are 17 types of crops are mentioned in dataset

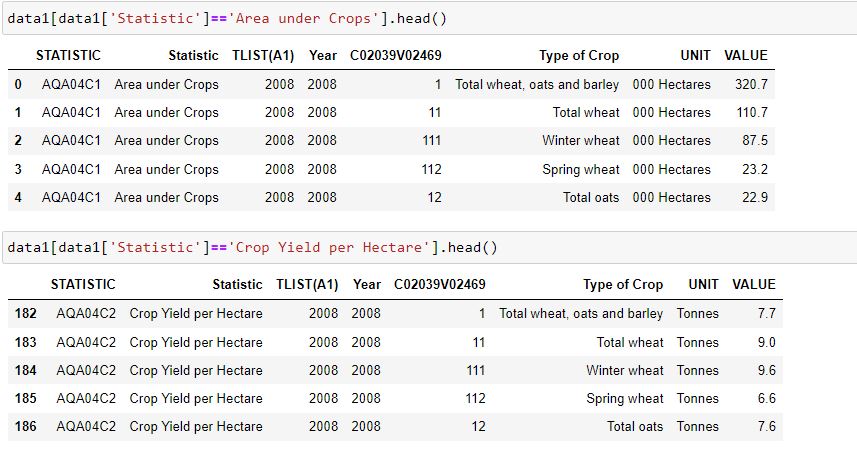
Every crop having same number of records as 69

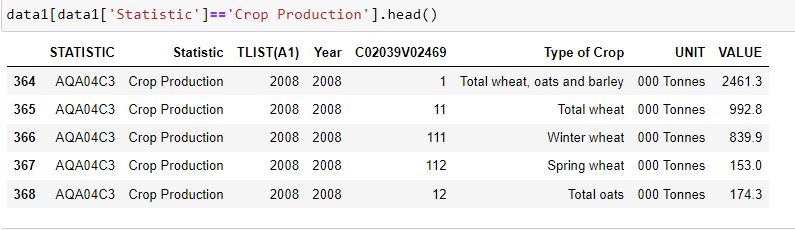
**Checking Length of every measure**



As observing above it is cleared that every measure have 391 records

**Comparing 1st five records first dataset of every measure**



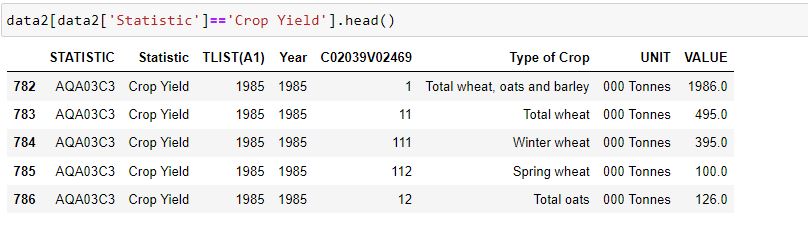
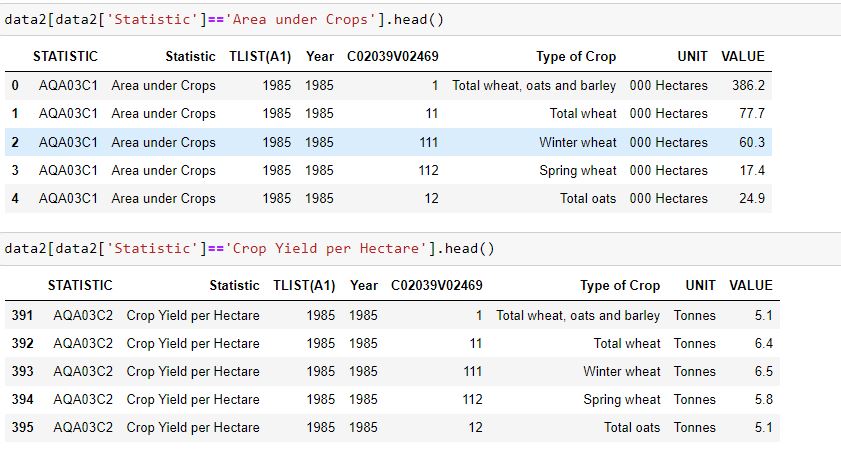


As by comparing first five records of three measures it concluded that:

Features of year, Co2039VO2469 and type of crop are same for all three measures.

There is difference of value and unit in every measure

**Comparing 1st five records first dataset of every measure**



As by comparing first five records of three measures it concluded that:

Features of year, Co2039VO2469 and type of crop are same for all three measures.

There is difference of value and unit in every measure

**Final conclusion of EDA**

By exploring 1st and 2nd dataset it is concluded that:

Both datasets having same columns so we can combine both to make dataset efficient

Both datasets has three measures represented by three units respectively so this thing makes dataset inconsistent

Year and type of crop column is same for all measures

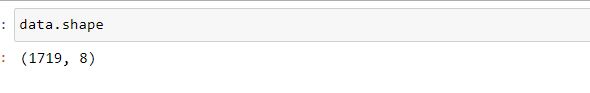
Every measure have same number of records

**Feature Engineering**

Combining two datasets by concatenation

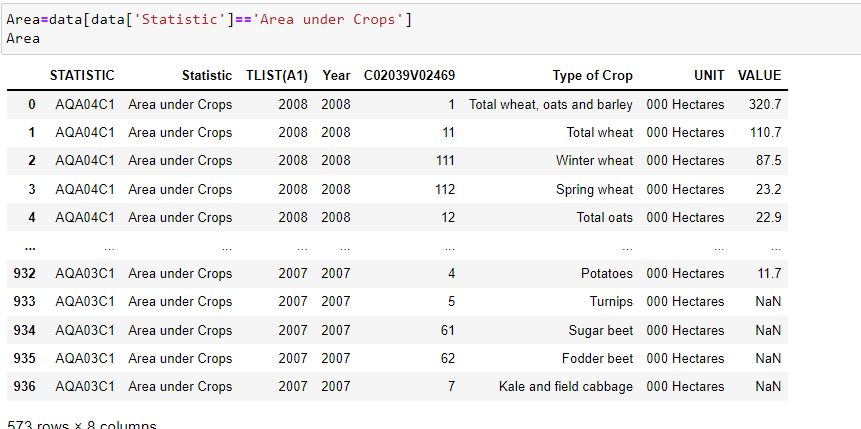
As two datasets has same features so two datasets can be combined in such a way that rows of 2nd dataset are added where rows of 1st dataset terminating



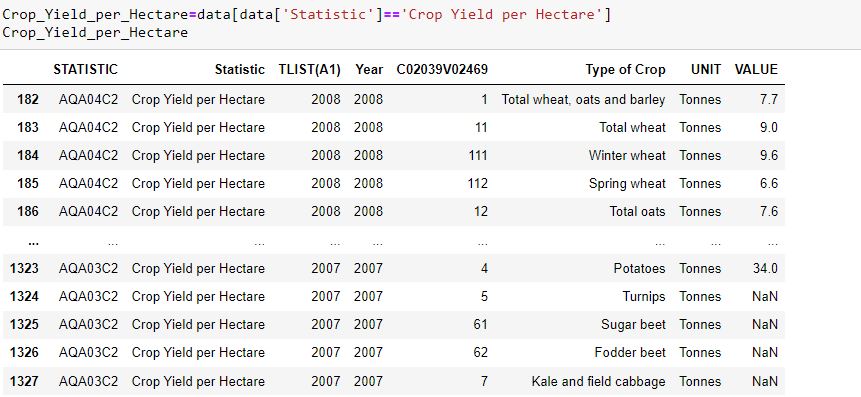


So after concatenation the total records in concatenated data is 1719

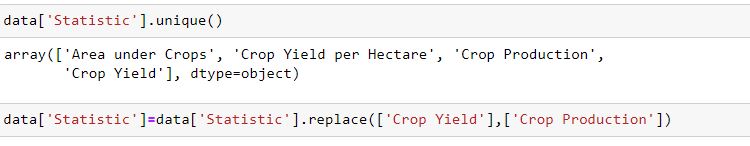
**Separating data related to measurement of area**



**Separating data related to measurement of crop yield per Hectare**



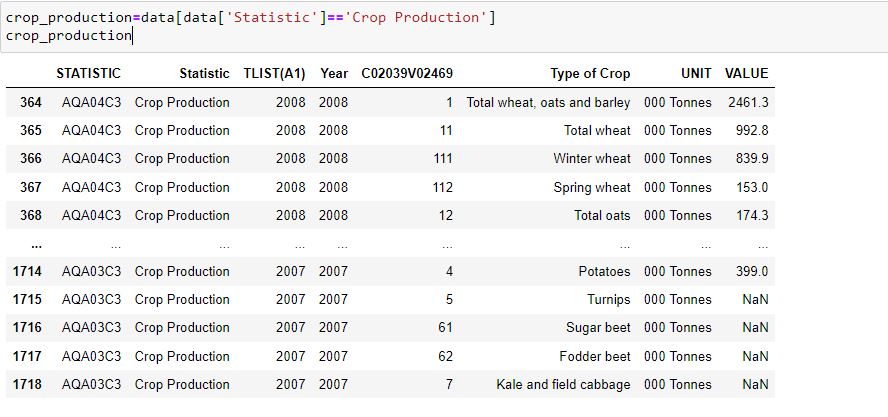
**Replacing crop yield with crop production**



As in resultant data there are four measurements but Crop production and Crop Yield is same.

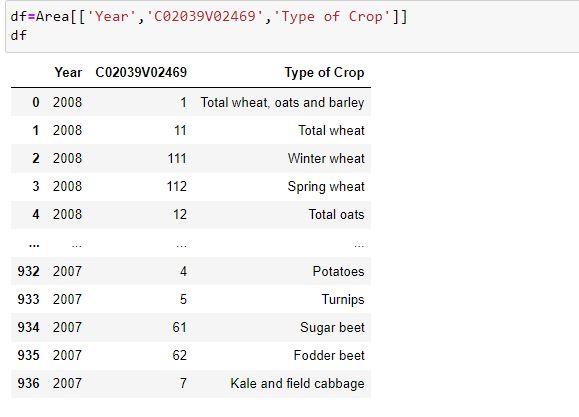
So crop yield is replaced by crop production

**Separating data related to measurement of production**

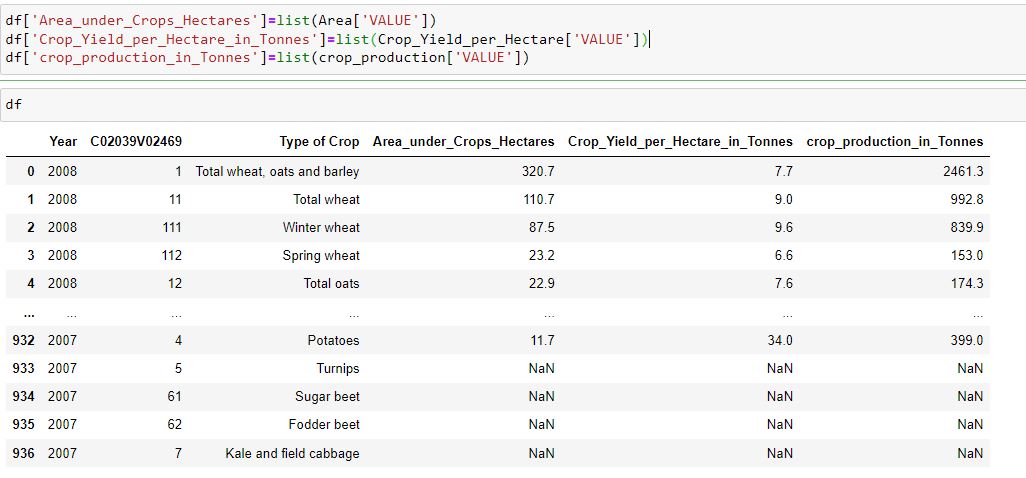


**Preparing data frame with year and type of crop feature**

As year and type of crop is same for all measurements and number of records in all measurements are same three columns are derived from area

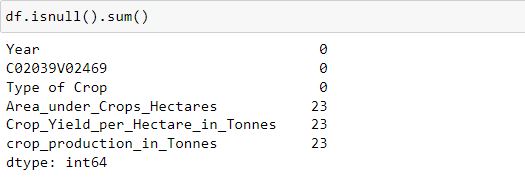


**Adding three new columns to data frame**



As three new columns are added which are derived from area under crop, crop yield per hectare and crop production

**Checking missing values**



In three newly added columns total 69 null values are found

**Removing null values**



As null values are removed so remaining rows are 550

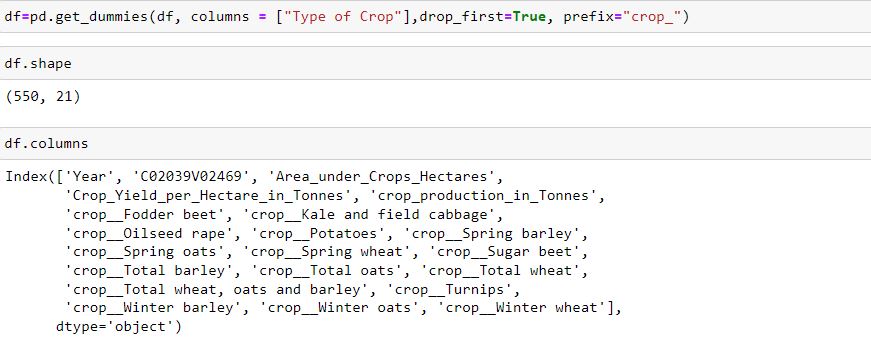
**Saving dataset**

As now dataset is structured and cleaned it can be used for analysis and machine learning



**One hot encoding**

One hot encoding is used to convent categorical feature into numerical feature.



As type of crop is a categorical feature so it is converted into numerical to apply machine learning on it.

After applying one hot encoding there are 21 features total.

**Logical justification based on the reasoning for the specific choice of machine learning approaches:**

*Why Supervised Machine Learning is used rather than Unsupervised Machine Learning?*

Here Supervised Machine Learning is used because with the help of supervised learning, the model can predict the output on the basis of training data.

Unsupervised Machine Learning is used to view patterns or clusters

As here the task is to prepare a Machine Learning Model which predicts the value of crop production so Supervised Machine Learning Techniques are solution here.

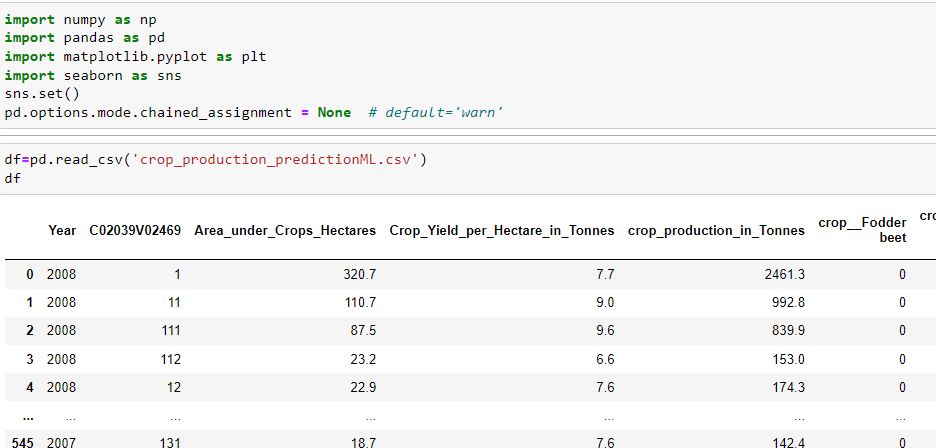
*Why Regression is used here rather than classification?*

Regression is used where dependant feature or target feature is continues and classification is used Where dependant feature or target feature is categorical

As here target variable crop production contains continues values so regression is used here.

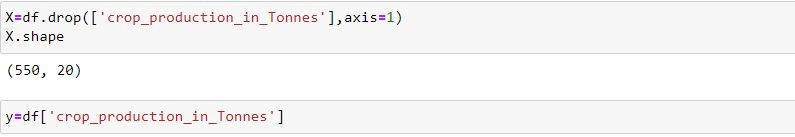
**Import dataset for Machine Learning**

Here below structured dataset is imported for building machine learning models.



**Splitting dataset into dependant and independent**

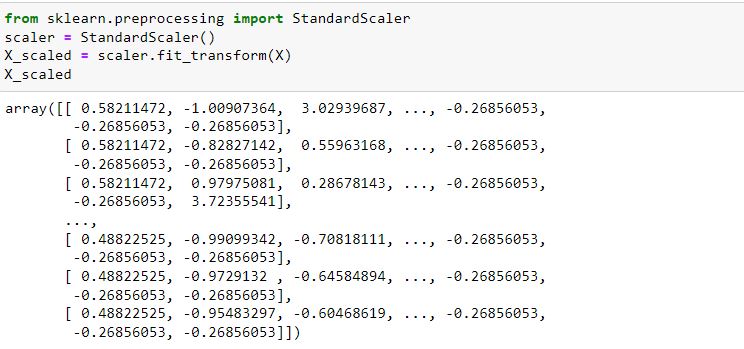
Here data is divided into dependent and Independent features. Dependant features are inputs for training process and Independent features is output for machine learning process.



As crop production is targeted variable so it is removed from inputs and added on outputs

**Feature Scaling**

Standardization is scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.



**Dimensional reduction:**

Dimensionality reduction refers to techniques for reducing the number of input variables in training data. When dealing with high dimensional data, it is often useful to reduce the dimensionality by reducing features.

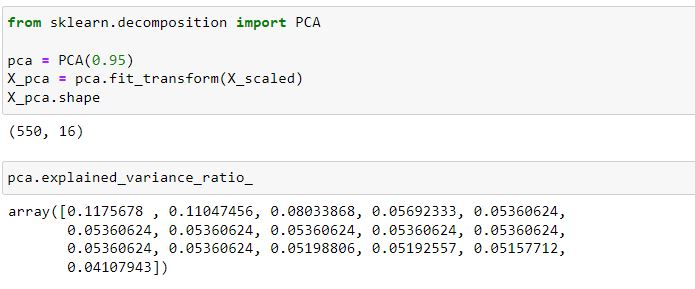
It has two main techniques:

● Principal Component Analysis (PCA)

● Linear Discriminant Analysis (LDA)

**Principal Component Analysis (PCA)**

Principal Component Analysis, or PCA, is a dimensionality-reduction method to find lower-dimensional space by preserving the variance as measured in the high dimensional input space. It is an unsupervised method for dimensionality reduction.



Here above number of features are reduced to 16 .which makes 95 % importance to Model.

#### Train test splitting

The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model.

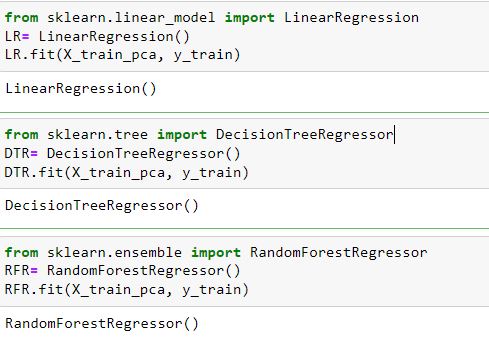


Here above size of test data is 20% so size of training dataset is 80%.

**Model building**

This is Machine Learning Model building stage in which threetypes of regression algorithims are used along with parameters.

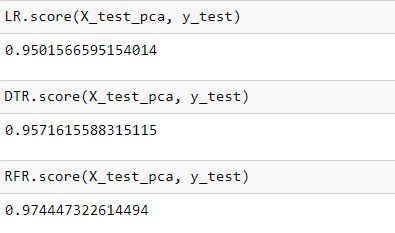
In this stage training data is given to model so in this stage classification models learns from training data.



**Comparison between the chosen modelling approaches:**

Here below performance of regression models are compared using Accuracy Score.

**Accuracy of classification model:**



As observed accuracy of linear regression model is 95%,

Accuracy of decision Tree Regression is 95% and accuracy of Random Forest Regression is 97%

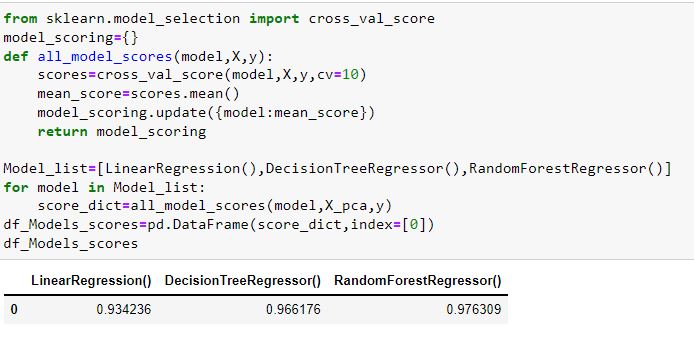
So Random Forest Regression is best performing model here

But it is not finalized that this is best performing model

So other Evaluation techniques are also used.

**K Fold Cross validation**

Here K Fold Cross validation is performed on Machine Learning Models to to find which model is performing best by giving highest accuracy.



Result:

Random Forest Regression is regression model here with highest score of accuracy so Random Forest Regression is best performing model

**Evaluation techniques:**

**R2 score**

The R2 score is a very important metric that is used to evaluate the performance of aregression-based machine learning model. It is pronounced as R squared and is also known as the coefficient of determination. It works by measuring the amount of variance in the predictions explained by the dataset

**Mean Absolute Error**

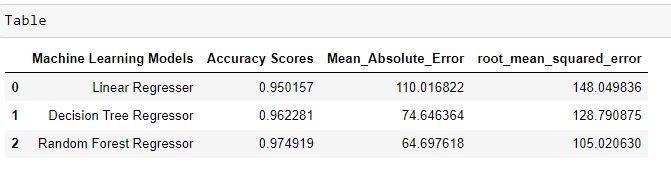
Mean absolute error refers to **the magnitude of difference between the prediction of an observation and the true value of that observation**. MAE takes the average of absolute errors for a group of predictions and observations as a measurement of the magnitude of errors for the entire group.

**Root Mean Square Error:**

**Root mean squared error** (**RMSE**) is the square root of the mean of the square of all of the **error**. The use of **RMSE** is very common

**Table of accuracy scores, Mean Absolute Error and Root Mean Square error:**



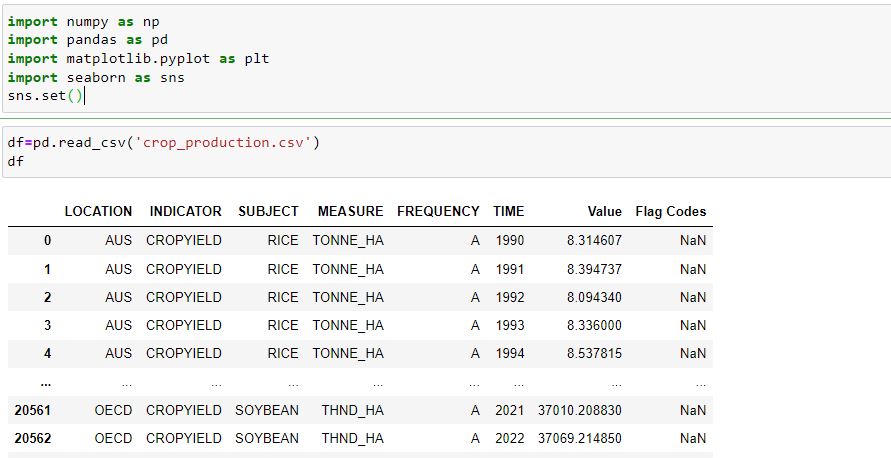


This table shows similarities and difference of Accuracy Scores, Mean Absolute Error and Root Mean squared error.

So random Forest Regression is best performing machine learning model as it gives largest accuracy scores but gives smallest Mean Absolute Error and Root Mean Square Error .

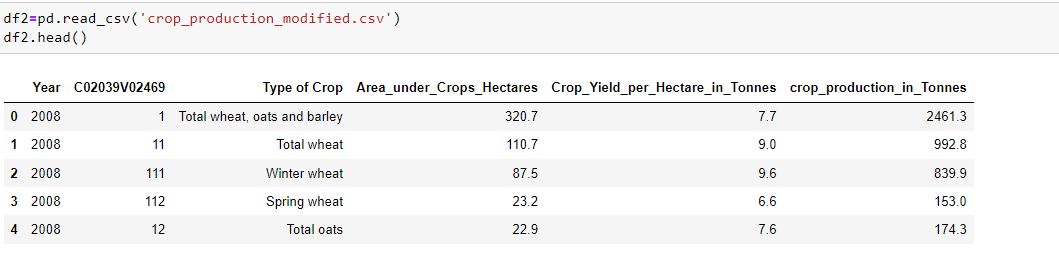
**Importing dataset for stats:**

Here a dataset is loaded for statistical analysis and comparisons. It is international data because it is used to compare the results of Ireland with rest of countries in world.



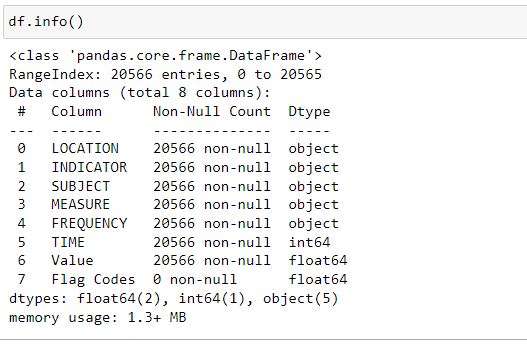
**Importing Ireland dataset**

This is structured dataset of Ireland so these two datasets are imported to do statistical analysis



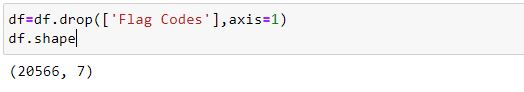
**Overview of features:**

This is mean, median and mode value of features



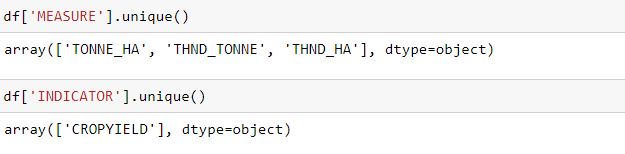
**Removing FLAG Codes feature**

Flag Codes feature is removed as it contains all null values



As there are 20566 records and 7 features

**Exploring Measure**



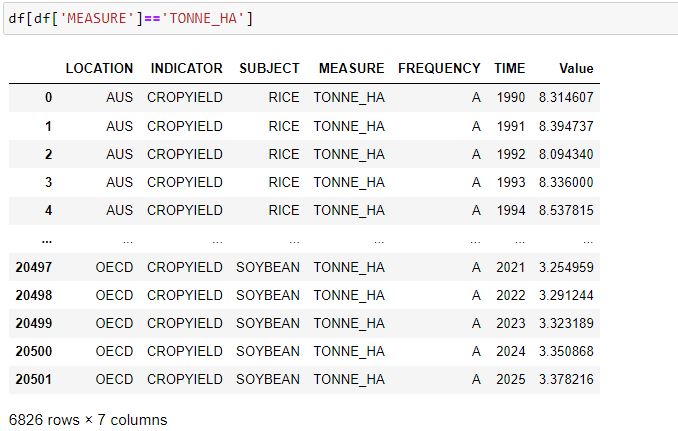
There are three types of measurements are here:

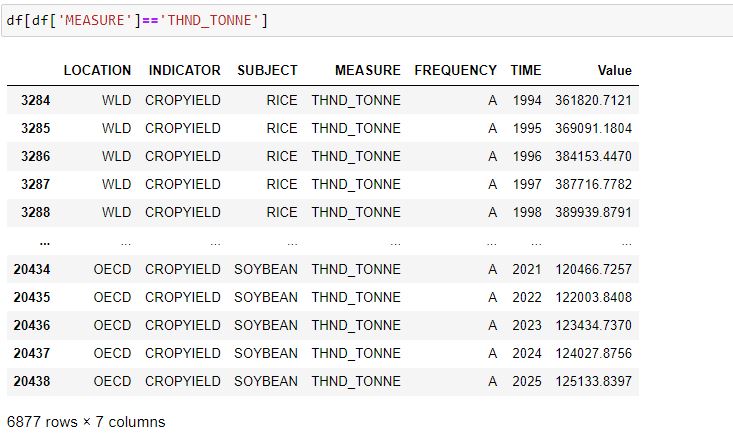
* Tonne per hectare :this is unit of crop production per area
* Thousand Tonne: this is unit of Crop production
* Thousand Hectare: This is unit of Land area

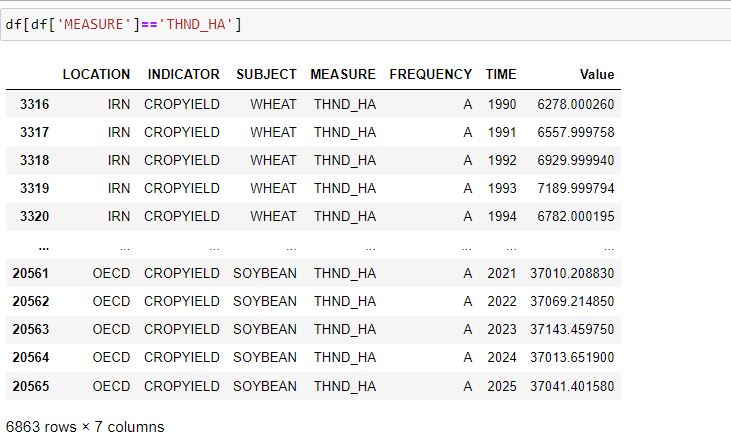
So this data representing three different types of quantities

Due to which this data is inconsistent so we have to make it structured

**Exploring Measure Deeply**

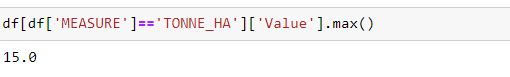






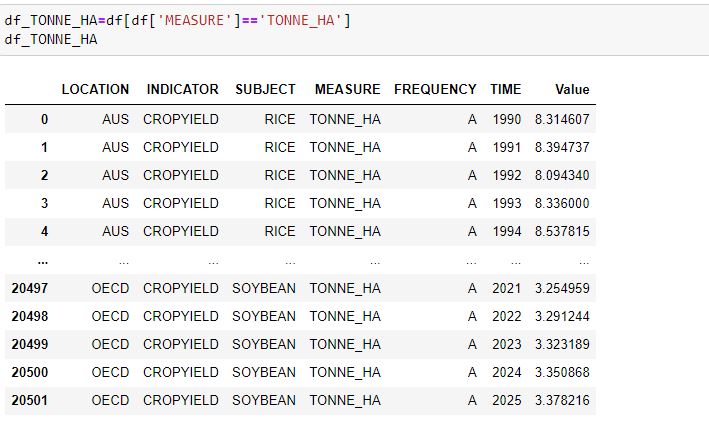
As by looking above three quantities it is clear that there is huge difference between Values of three quantities

**Finding maximum of Tonne per hectare**

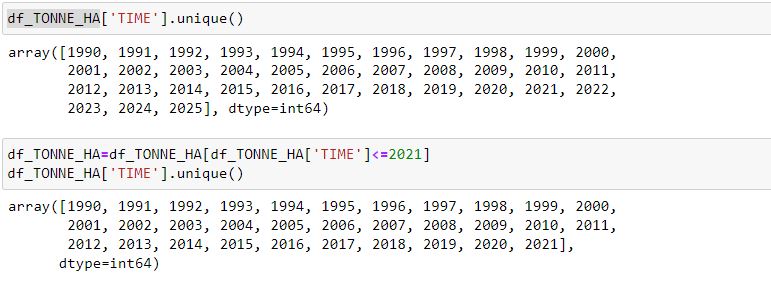


**Making a new data frame**

Here a new dataset is developed by using measurement of Tonne per hectare so it can be compared with crop production per hectare feature of Ireland dataset

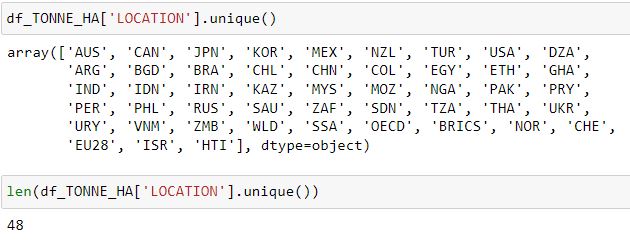


**Exploring Time feature**



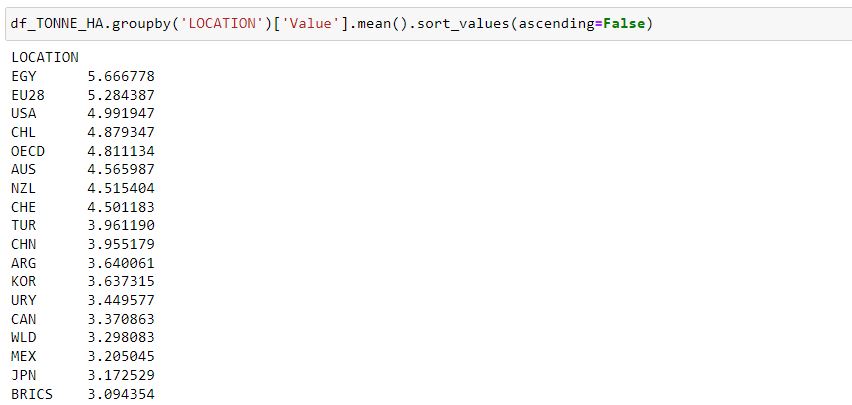
As here a problem is detected that year data shows years of 2025, 2024, 2023. These are feature years so these years are eliminated from original dataset.

**Exploring Location**



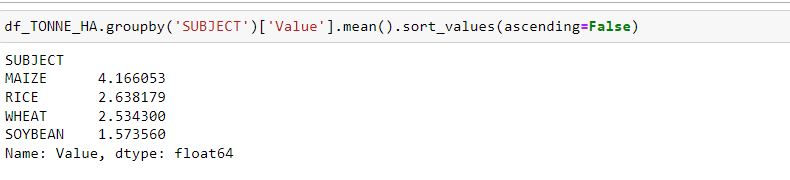
Here location shows names of countries in world so there are 48 countries In data

**Showing Mean values of crop production of every country**

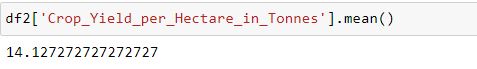


Here above are mean values of crop production of countries

**Mean Value on the basis of types of crops:**



**Mean of crop production value in Ireland :**

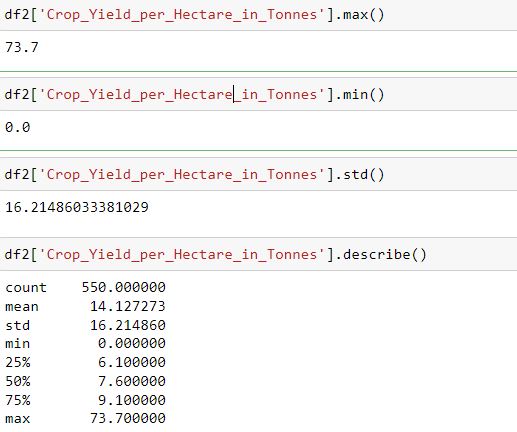


Here 14.12 Tonnes per hectare is Average Value of crop production per hectare

**Comparison of Ireland with other:**

As looking above Ireland average crop production is 14 tonnes per hectare it is highest value as comparing with other countries has less mean value of crop production

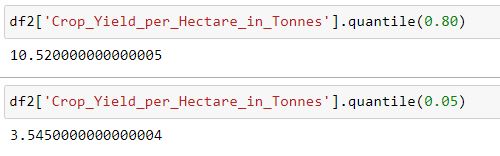
**Descriptive stats**



Here crop yield of Ireland is explored as this data has huge variance is observed between mean and median

This show that this data is very skewed and not normally distributed

An huge evidence of greater variance is that seventy five percentile is 9 and maximum (100 percentile) is 73 so this shows that 73 is outlier



80 percentile is 10

**Inferential statistics**

Inferential statistics use measurements from the sample of subjects in the experiment to compare the treatment groups and make generalizations about the larger population of subjects.

**Why inferential statistics is used?**

While descriptive statistics summarize the characteristics of a data set, inferential statistics help our come to conclusions and make predictions based on your data. When we have collected data from a sample, we can use inferential statistics to understand the larger population from which the sample is taken.

**Hypothesis testing:**

Hypothesis testing is a form of statistical inference that uses data from a sample to draw conclusions about a population parameter or a population probability distribution.

**Chi-square test**

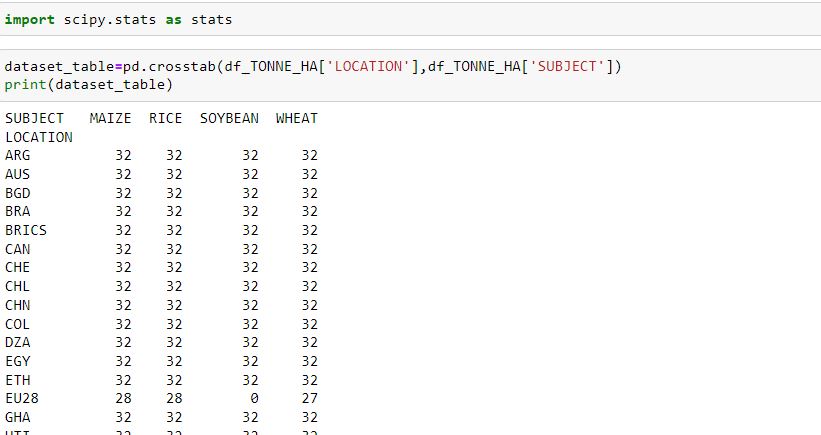
A chi-square test is a statistical test used to compare observed results with expected results. The purpose of this test is to determine if a difference between observed data and expected data is due to chance, or if it is due to a relationship between the variables you are studying.

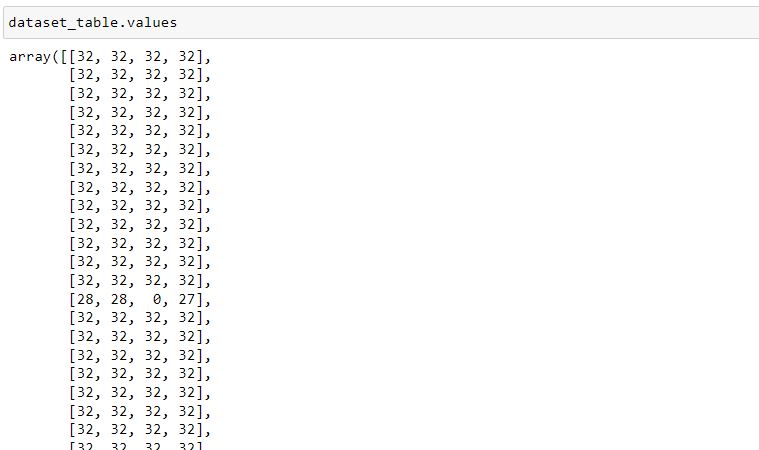
**We use the chi-square test to compare categorical variables**

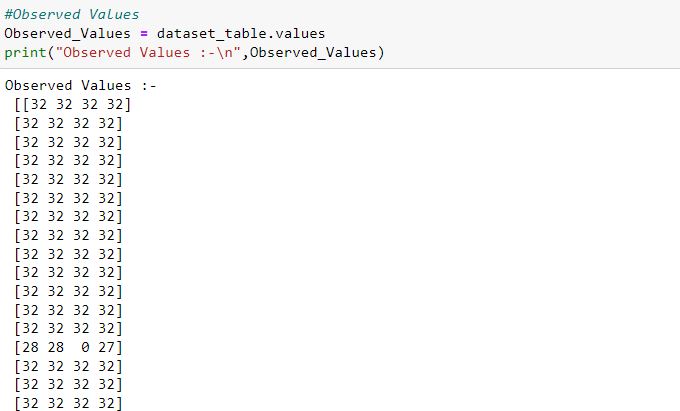
Here chi-square is used to show that is there is a relationship between two categorical features

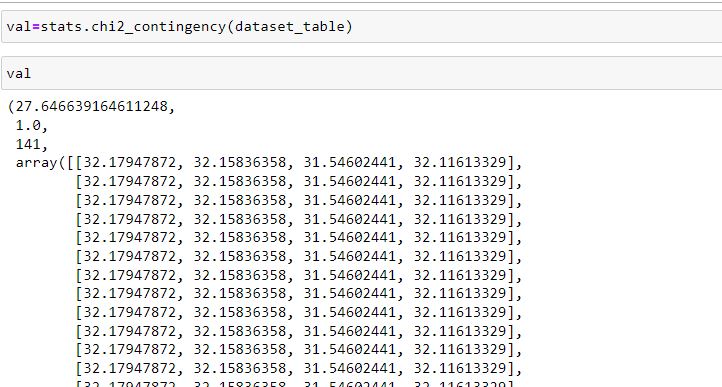
H0: relation exist between two categorical features Location and Subjuct

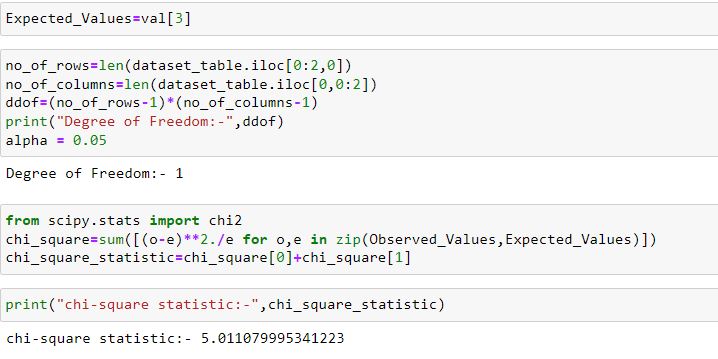
H1:there is no relationship between Location and Type of crop

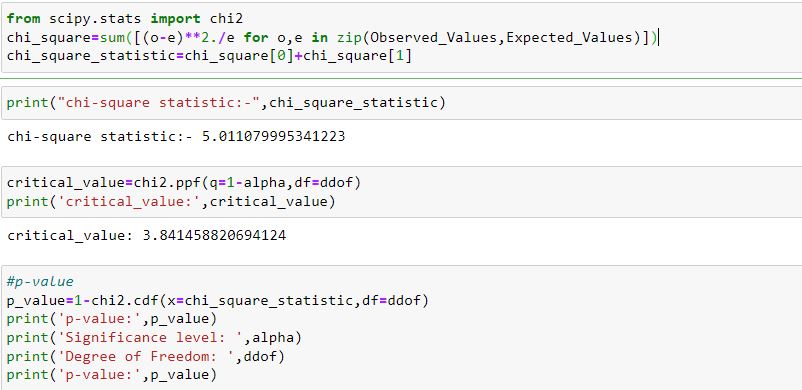


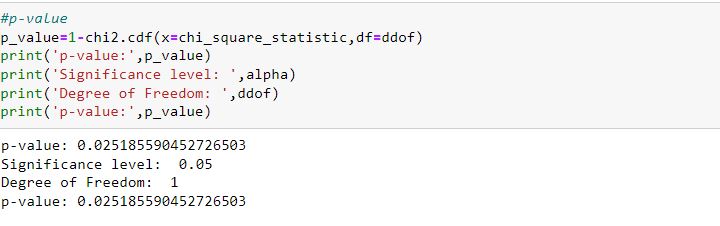


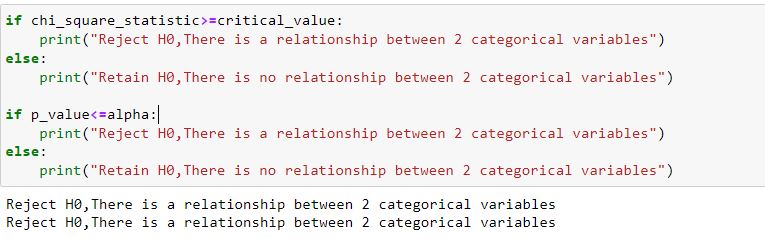








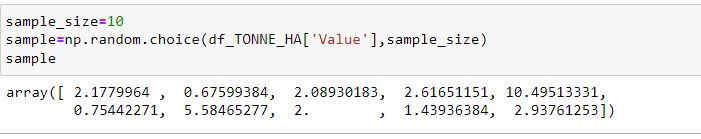




As there is relationship between two categorical features so null hypothesis is rejected

Because value of p value is less than 0.025

**Taking sample of 10 randomly chosen data points**

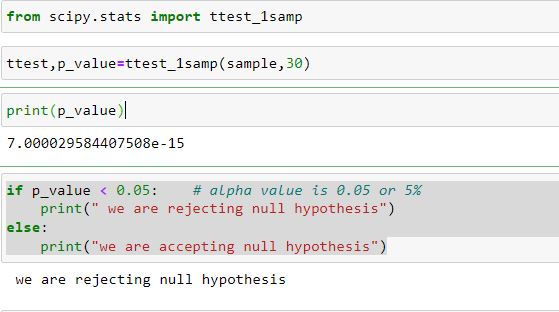


**T\_test:**

A t-test is a statistical test that is used to **compare the means of two groups**. It is often used in hypothesis testing to determine whether a process or treatment actually has an effect on the population of interest, or whether two groups are different from one another.

The **t test** tells us how significant the differences between group means are. It lets our know if those differences in means could have happened by chance.

H0



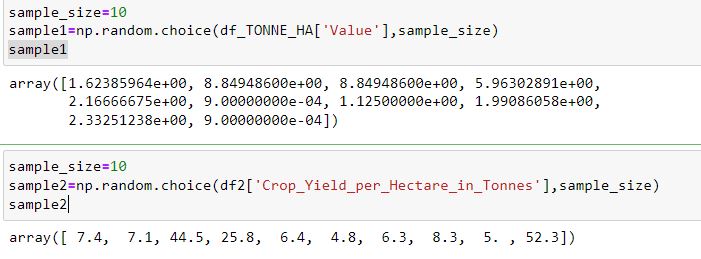
As p\_value is less than 0.05 so null hypothesis is rejected so this show that mean of sample and mean of overall population is not same

**T Test for two samples:**

H0: Mean is same as sample as population

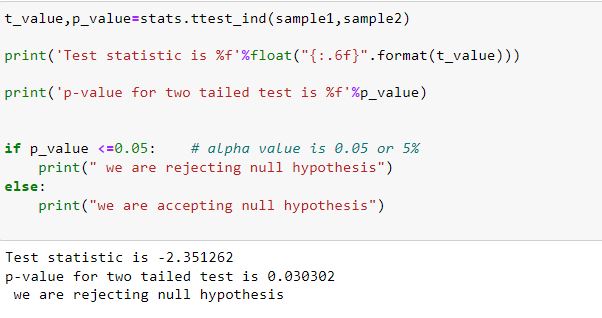
H1: mean is different in sample and their parent population

**Taking two samples**



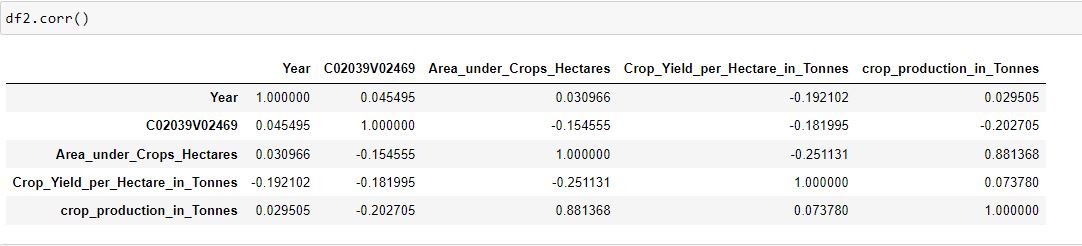
Here two samples are randomly chosen for T test to make comparisons

First sample from value of Tonnes from first dataset and 2nd sample is taken from 2nd dataset so this is used to make comparison



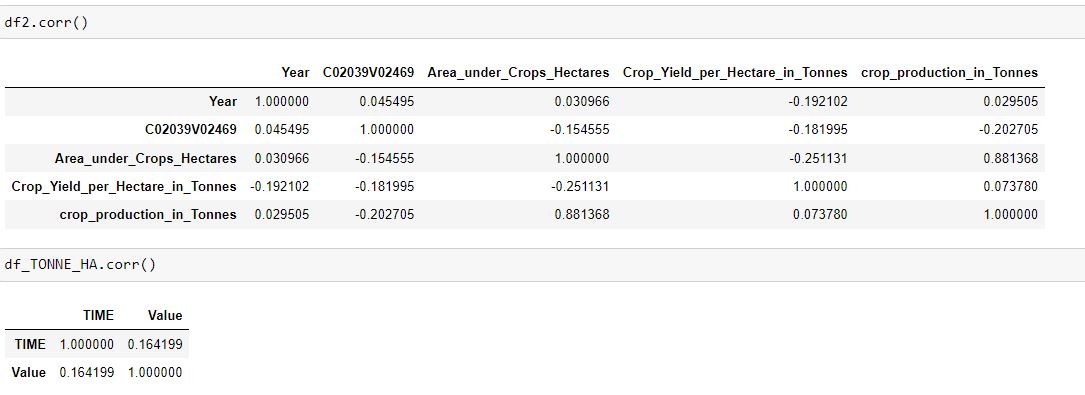
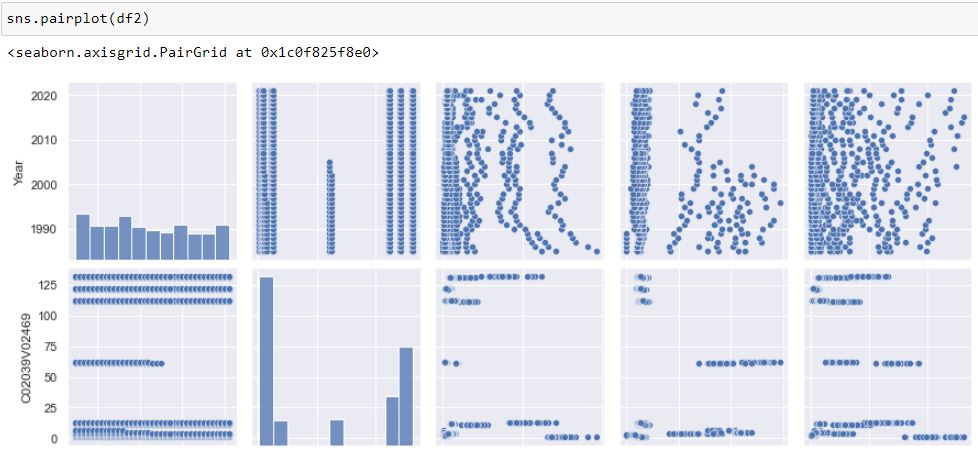
Here null hypothesis is rejected because p value is less than 0.05 so mean of two quantities are very different

**Correlation**



The correlation shows positive or negative relationship

Correlation is **an indication about the changes between two variables**.



It shown that area, year and crop production has direct relation as as area increased crop production also increases

**Z test**

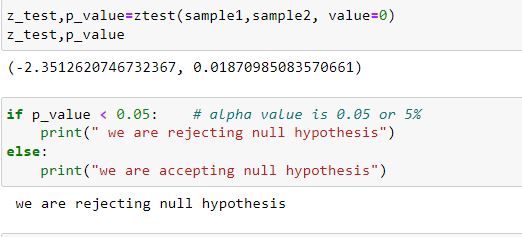
A z-test is a statistical test to determine whether two population means are different when the variances are known and the sample size is large. A z-test is a hypothesis test in which the z-statistic follows a normal distribution. A z-statistic, or z-score, is a number representing the result from the z-test.

Here H1 is null hypothesis and H1 is alternative hypothesis.

H0: NO difference of mean in two samples

H1: Difference of Mean in two samples





Are null hypothesis is rejected as there are difference between means .

**Outcome of my analysis**

Here two detests are used Ireland Crop production data and international crop production data of 48 countries as Ireland is not mentioned here so it is needed to import data of Ireland.

By comparing mean value with other countries it is cleared that Ireland crop production capability is more than other countries as Ireland has highest average production per hectare

As z test is used for comparison of two samples which shows mean value is different.

Here chie square test shows that there is relationship between categorical variables of subject and location.

**Challenges:**

There are many problems which I faced which are following:

* to find a good dataset.
* To arrange dataset.
* To find pattern of dataset
* Measurement problems

So these are problems which I faced

Datasets Finding:

* Ireland Crop production 1985-2007

This data is can be found on official website of Ireland datasets IRELAND'S OPEN DATA PORTAL. This is data from 1985 to 2007.

Link:

<https://data.gov.ie/dataset/aqa03-crop-yield-1985-2007?package_type=dataset>

* Ireland crop production 2008-2021

This data is found on official website of Ireland datasets IRELAND'S OPEN DATA PORTAL .this data contains record from 2008 to 2021

Link:

<https://data.gov.ie/dataset/aqa04-crop-yield-and-production?package_type=dataset>

* International crop production data:

This data is found on data. World. It contains data of 48 countries.

Link:

<https://data.world/oecd/crop-production>

**Reference:**

**Hypothesis testing in Machine learning using Python** posted by [Yogesh Agrawal](https://medium.com/@yugagrawal95?source=post_page-----a0dc89e169ce--------------------------------) on Towards data science portal

Link:

<https://towardsdatascience.com/hypothesis-testing-in-machine-learning-using-python-a0dc89e169ce>

Statistical Inference Using Python **posted by** [**ELLURU PAVAN KUMAR REDDY**](https://www.analyticsvidhya.com/blog/author/elluru_pavan_kumar/) **on analyticsvidhya.com**

Link: <https://www.analyticsvidhya.com/blog/2022/02/statistical-inference-using-python/>

**https://data.gov.ie/dataset/aqa03-crop-yield-1985-2007?package\_type=dataset**

**ELAND'S OPEN DATA PORTALIRELAND'S OPEN DATA PORTAL**