**PROBLEM DEFINITION**

The data represents the details about the crops harvested by various farmers at the end of harvest seasons. It contains details like type of crop, season the crop was sown in, type of soil, type of pesticide used and the amount, etc. A good harvest depends on several factors like availability of water, soil fertility, protecting crops from rodents, timely use of pesticides & other useful chemicals and nature. While a lot of these factors are difficult to control for, the amount and frequency of pesticides used is something that the farmer can control.

**Inspiration:**

**The aim is to try and predict/determine the outcome of harvest season, i.e., whether the crop would be healthy, damaged by pesticides or damaged by other reasons.**

**DATA DESCRIPTION**

Fields or columns present in the dataset :

* ID - UniqueID - String
* Estimated\_Insects\_Count - Estimated insects count per square meter - Integer - Continuous Variable
* Crop\_Type - Category of Crop - (0, 1) - Categorical Variable
* Soil\_Type - Category of Soil - (0, 1) - Categorical Variable
* Pesticide\_Use\_Category - Type of pesticide uses - (1-Never, 2-Previously Used, 3-Currently Using) - Categorical Variable
* Number\_Doses\_Week - Integer - Continuous Variable
* Number\_Weeks\_Used - Integer - Continuous Variable
* Number\_Weeks\_Quit - Integer - Continuous Variable
* Season - Season Category - (1, 2, 3) - Categorical Variable
* Crop\_Damage - (0-Alive, 1-Damage due to other causes, 2-Damage due to pesticides) - Categorical Variable

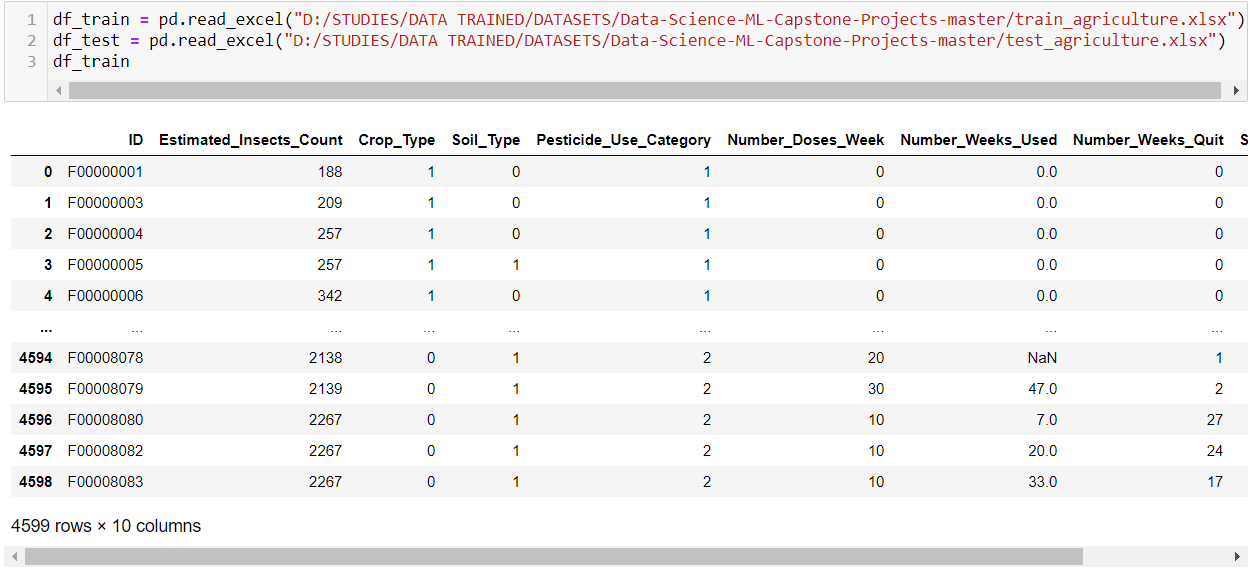
**DATA ANALYSIS & PEE-PROCESSING**

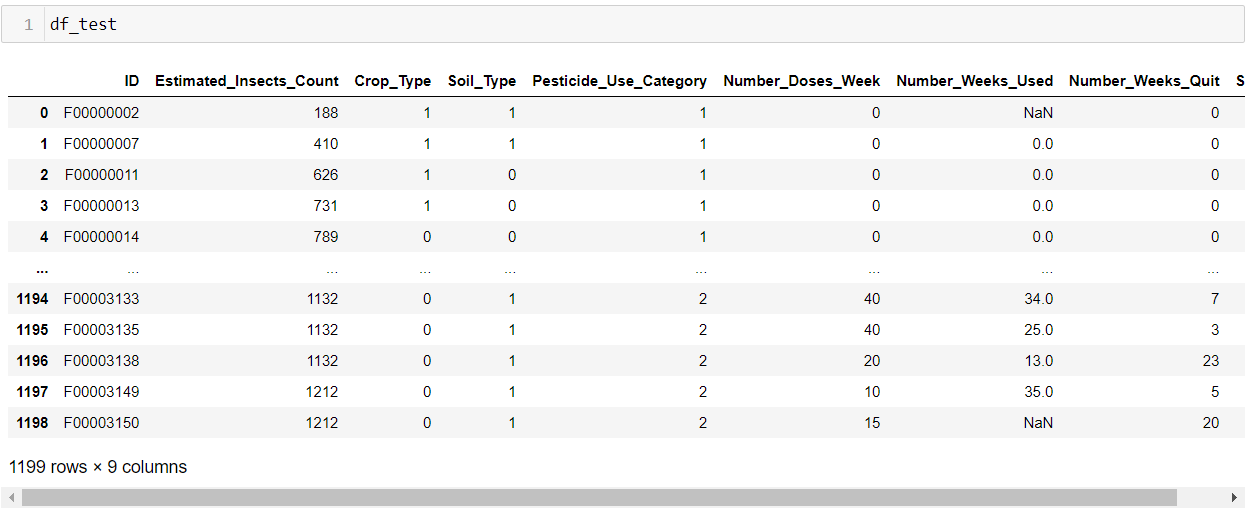
**Importing Libraries:**



These are all the libraries that are required at various stages - EDA, visualization, prediction and evaluation - for executing this project.

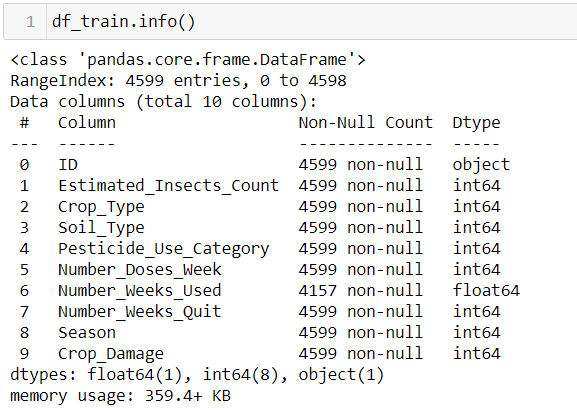
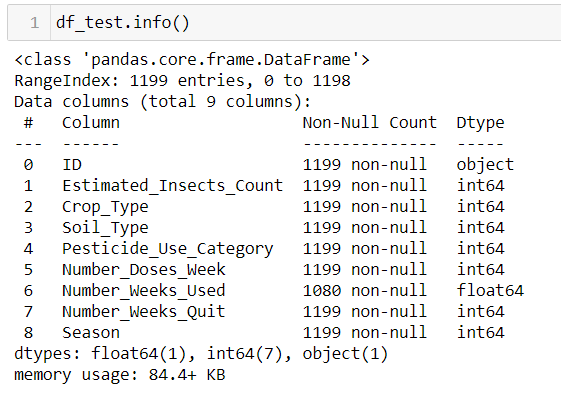
**Loading Dataset:**



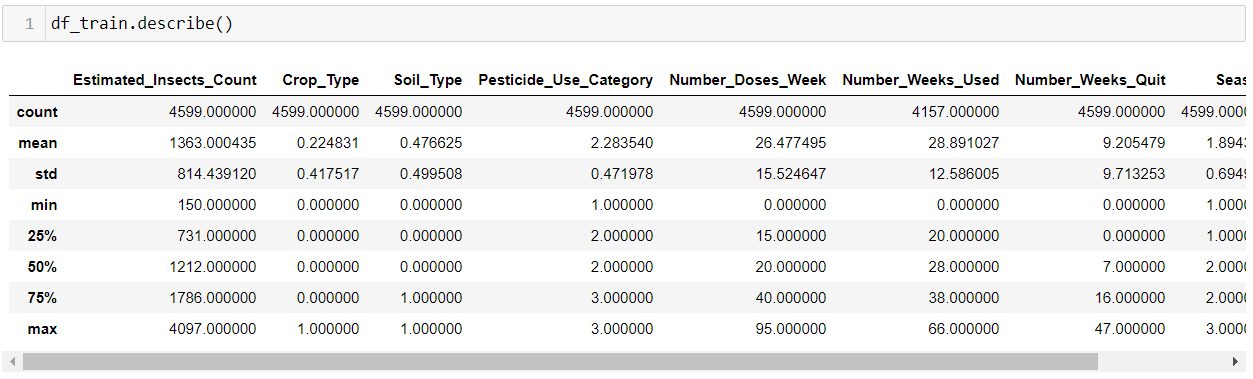


I have loaded the training and testing dataset in variables ”df\_train” and “df\_test” respectively. The training dataset contains 4599 rows and the tesind dataset contains 1199 rows.

**Statistical Data Analysis:**

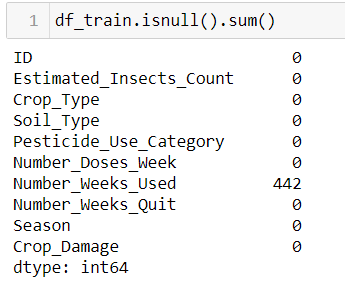
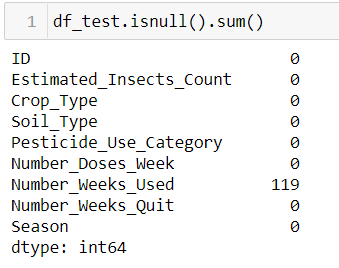
Apart from the ‘ID’ - *object* & ‘Number\_Weeks\_Used’ - *float* all the other variables are of data type *integer.*



* ‘Estimated\_Insects\_Count’ varies from 150 to 4097
* ‘Number\_Weeks\_Used’ varies from 0 to 66
* ‘Number\_Weeks\_Quit’ varies from 0 to 47
* ‘Number\_Doses\_Week’ varies from 0 to 95

All the above columns may contain outliers considering the significantly big difference between 3rd quartile and max.

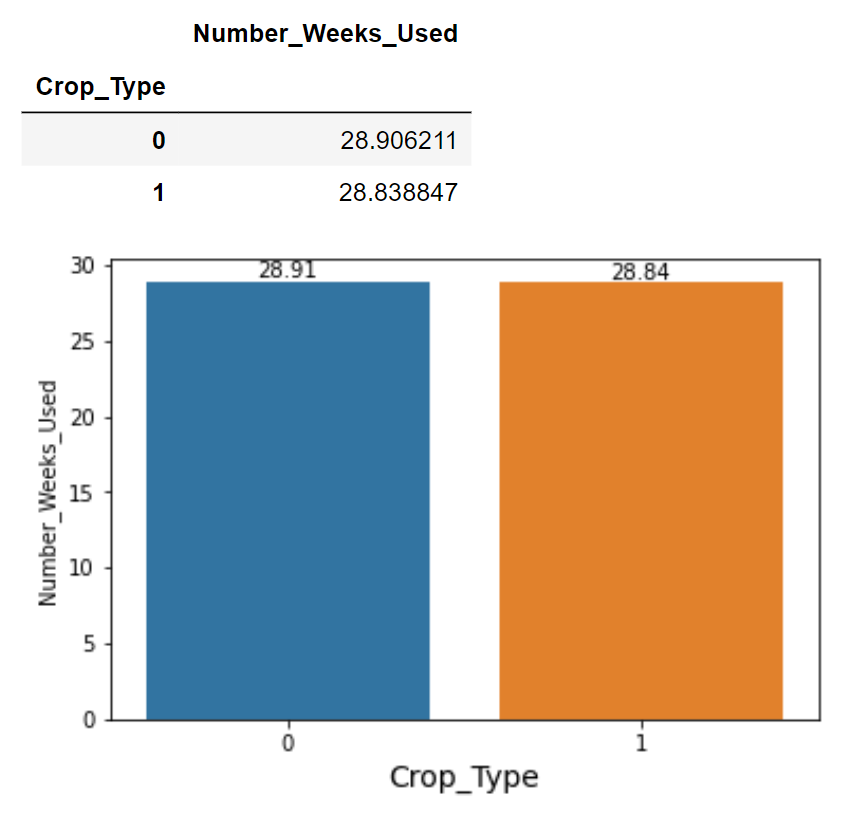
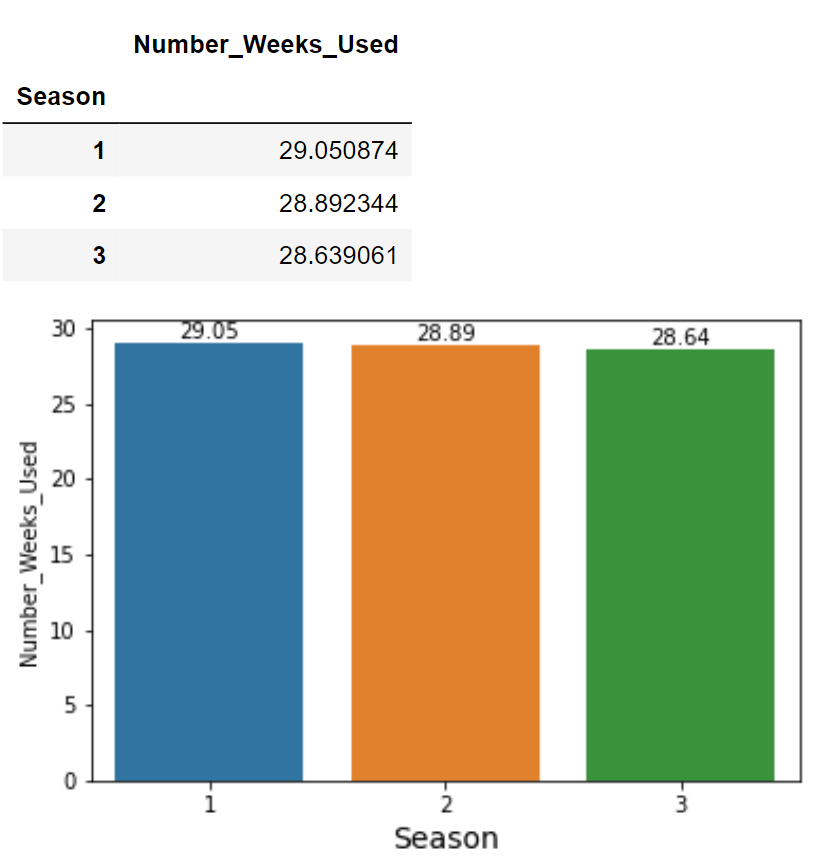
**Checking Null Values:**

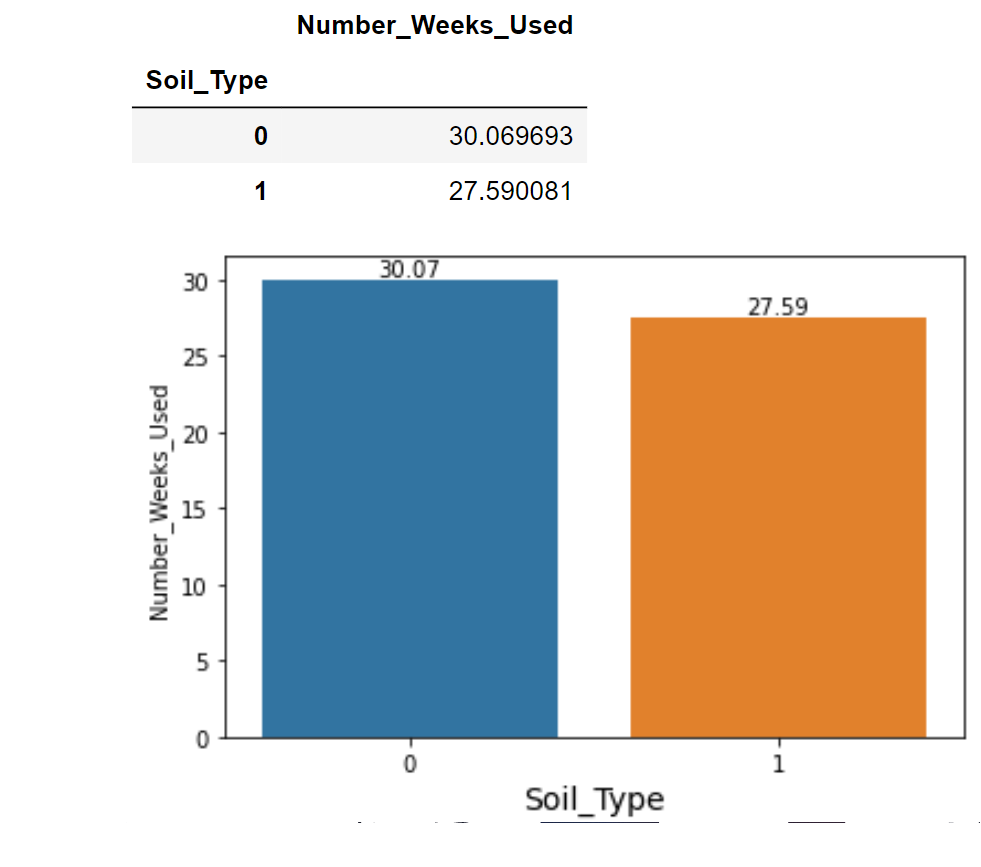
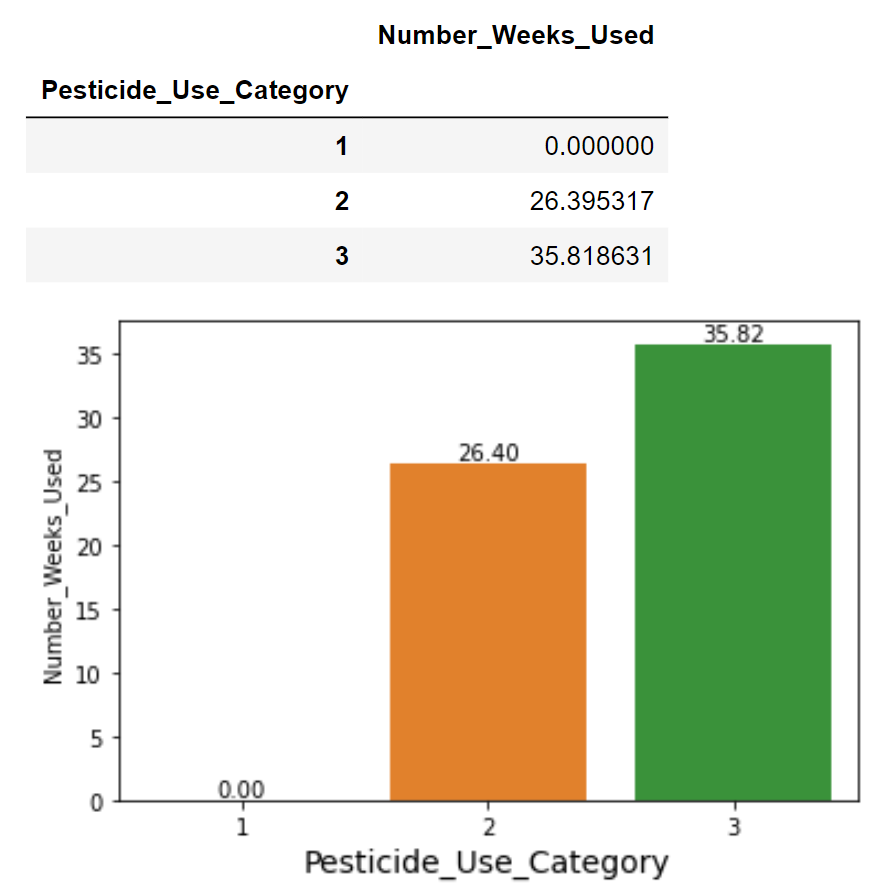
##### There are null values present in 'Number\_Weeks\_Used' field in both training and testing dataset.

**Imputing Null Values:**

The null values can be filled based based on the other categorical variables present in the dataset. Initially, the dataset is grouped by different categories and the variation of the field 'Number\_Weeks\_Used' with each category is analyzed.

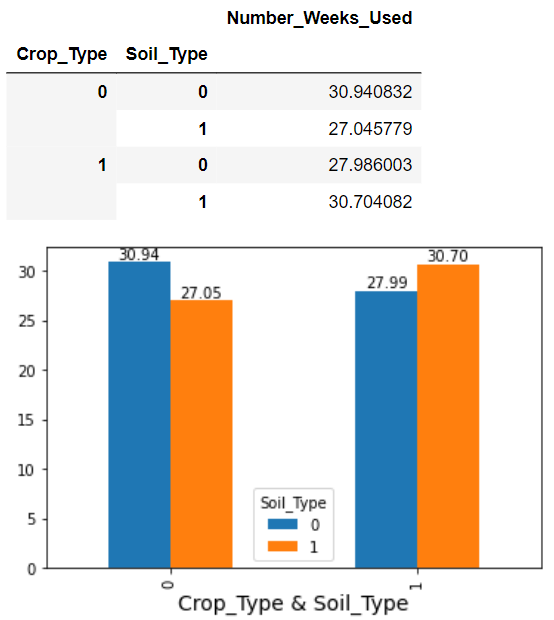
 

Initially, when the dataset is grouped by ‘Crop\_Type’ or ‘Season’ there is no significant difference in the mean values of 'Number\_Weeks\_Used' field for different categories.

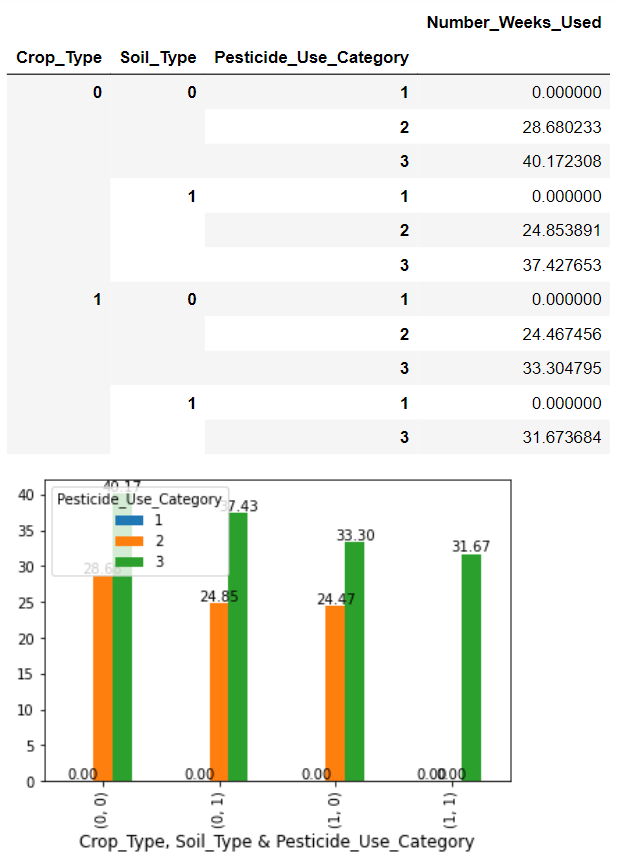
 

When the dataset is grouped by ‘Soil\_Type’ there is only a slight difference, whereas ‘Pesticide\_Use\_Category’ there is a significant difference in in the mean values of 'Number\_Weeks\_Used' field for different categories.

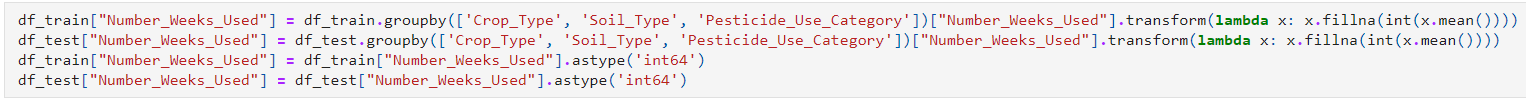
To make the imputation even more specific, the grouping can be done with more than one field.



When the dataset is grouped by ‘Crop\_Type & Soil\_Type’ there is only a slight difference. To make the imputation even more specific a group with ‘Crop\_Type, Soil\_Type & Pesticide\_Use\_Category’ is created and compared against ‘Number\_Weeks\_Used’.



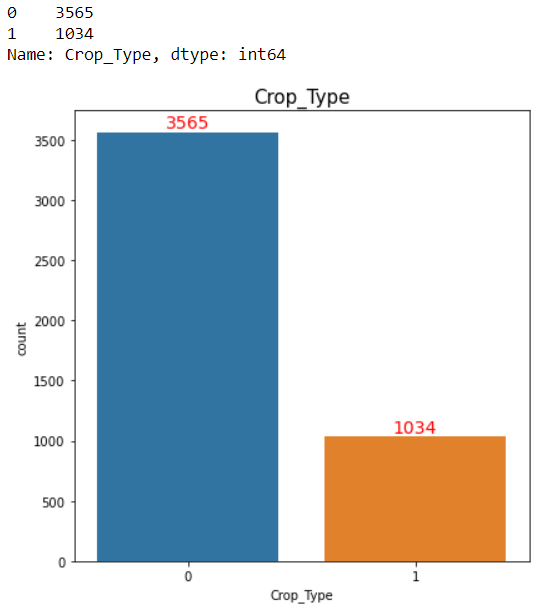
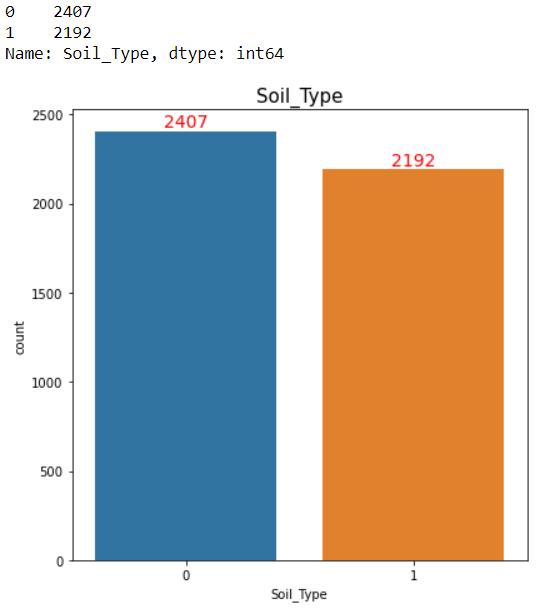
The field ‘Number\_Weeks\_Used’ varies significantly for the above group hence the missing values can be filled using it.

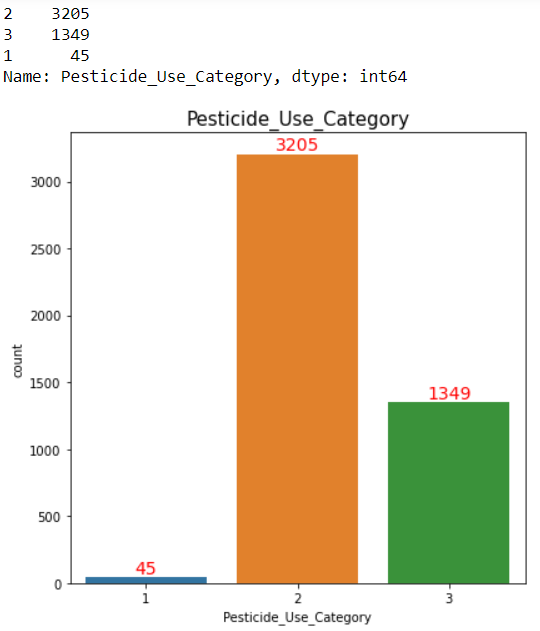
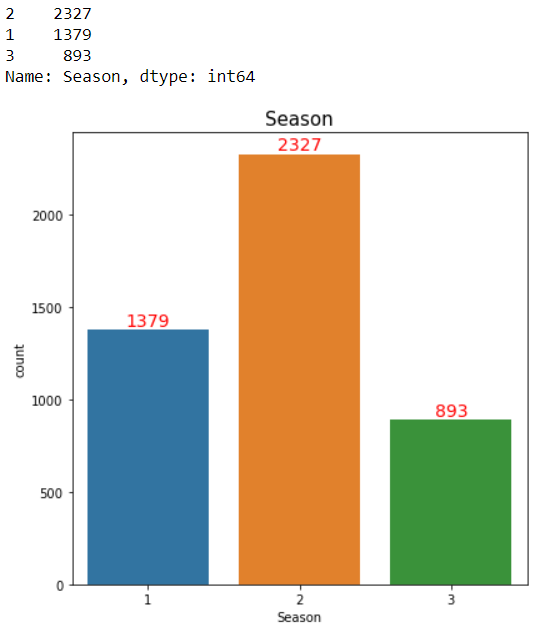


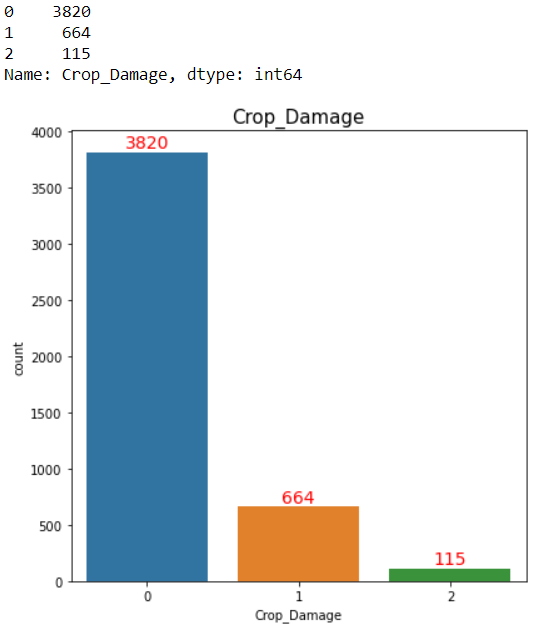
The null values in the field are filled after grouping using the transform function. Then the field is converted to integer data type.

**Uni-variate Analysis:**

First the count of different categories in the categorical variables are checked using countplots. A **countplot** is kind of like a histogram or a bar graph for some categorical area. It simply shows the number of occurrences of an item based on a certain type of category.

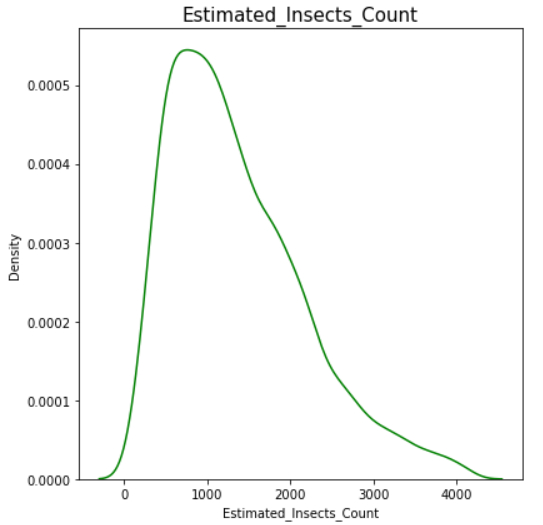
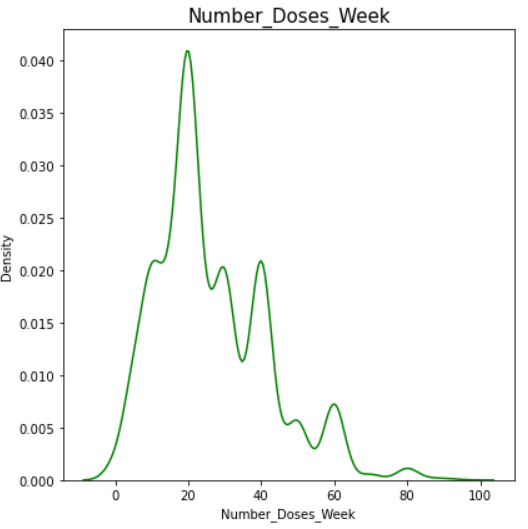
 

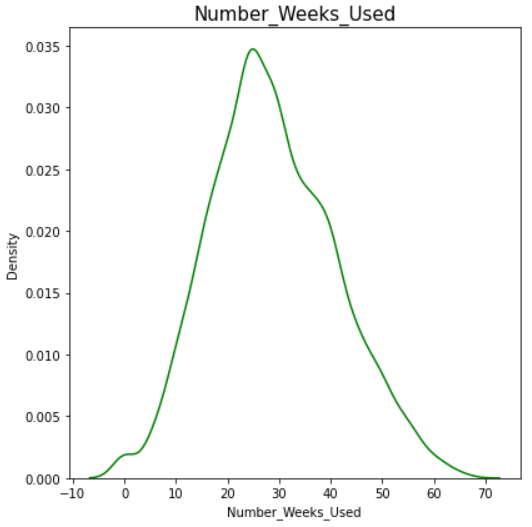
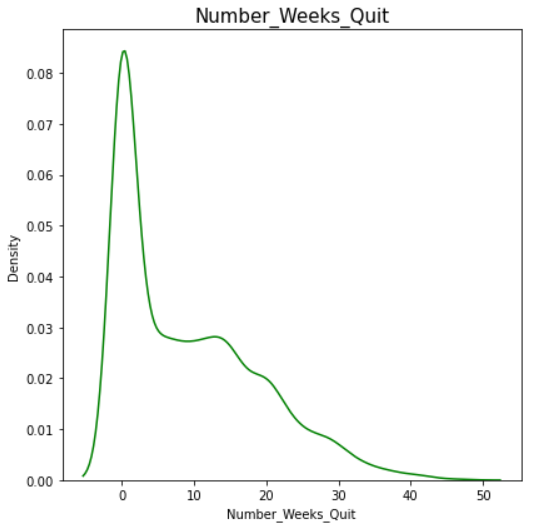


From the above count-plots the following insights can be drawn :

* Fewer crops were damaged by pesticides than by other causes.
* Only very few pesticides have never been used before and farmers prefer to use the pesticides that they have previously used.
* Crop\_Type 0 is preferred more than Crop\_Type 1.
* Soil\_Type 0 is preferred more than Soil\_Type 1.
* Farmers prefer to sow crops in season 2 than in any other season.

The distribution of continuous variables can be checked using distribution plots. A **distribution plot** displays a distribution and range of a set of numeric values plotted against a dimension. Distribution plot helps in identifying if the field is skewed or normally distributed.

From the above distribution plots the following insights can be drawn :

* The 'Estimated\_Insects\_Count' column is slightly right skewed.
* The 'Number\_Doses\_Week' column is multi modal and right skewed.
* The 'Number\_Weeks\_Used' column is normally distributed with low standard deviation.
* The 'Number\_Weeks\_Quit' column is multi modal and right skewed.

The **outliers** of the continuous variables can be checked using boxplots. **Boxplots** are a standardized way of displaying the distribution of data based on a five number summary (“minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”).

First quartile (Q1/25th Percentile): The middle value between the smallest number (not the “minimum”) and the median of the dataset.

Median (Q2/50th Percentile): The middle value of the dataset.

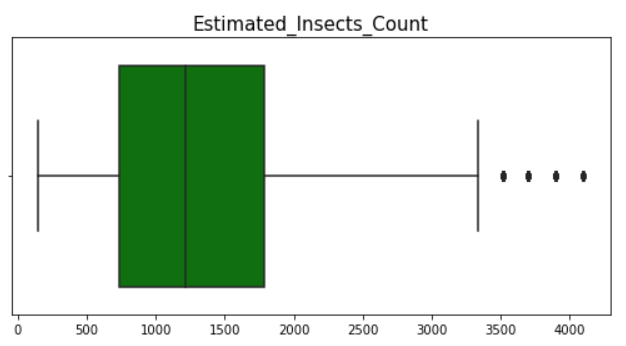
Third quartile (Q3/75th Percentile): The middle value between the median and the highest value (not the “maximum”) of the dataset.

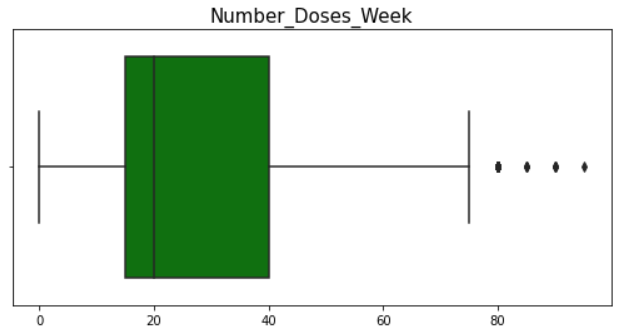
Inter-quartile range (IQR): Range of values between the 25th to the 75th percentile.

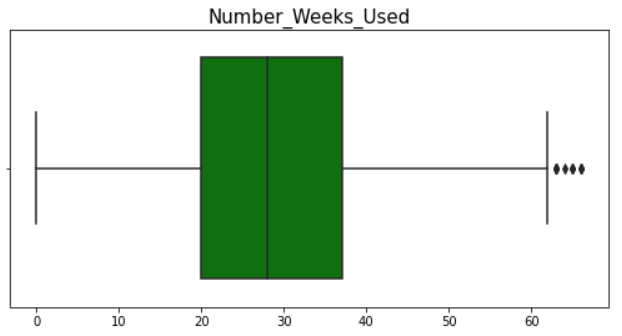
Maximum: Q3 + 1.5 \* IQR

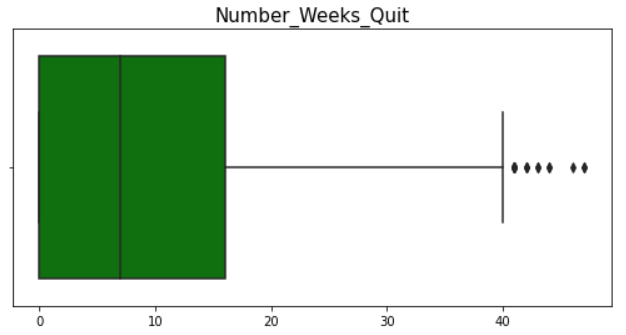
Minimum: Q1 - 1.5 \* IQR

Any value greater than maximum and less than minimum is considered as an outlier.









From the above box-plots the following insights can be drawn :

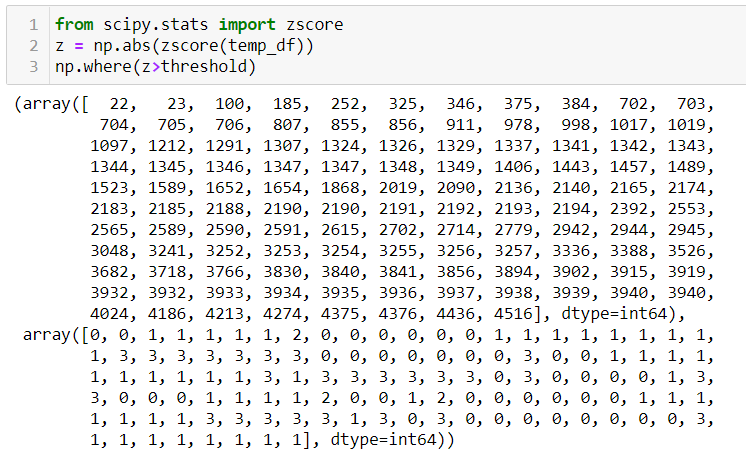
* All of the above four continuous columns have higher outliers and no lower outliers.
* 'Estimated\_Insects\_Count' and 'Number\_Doses\_Week' columns are slightly right skewed.
* ‘Number\_Weeks\_Quit’ column has high right skewness.

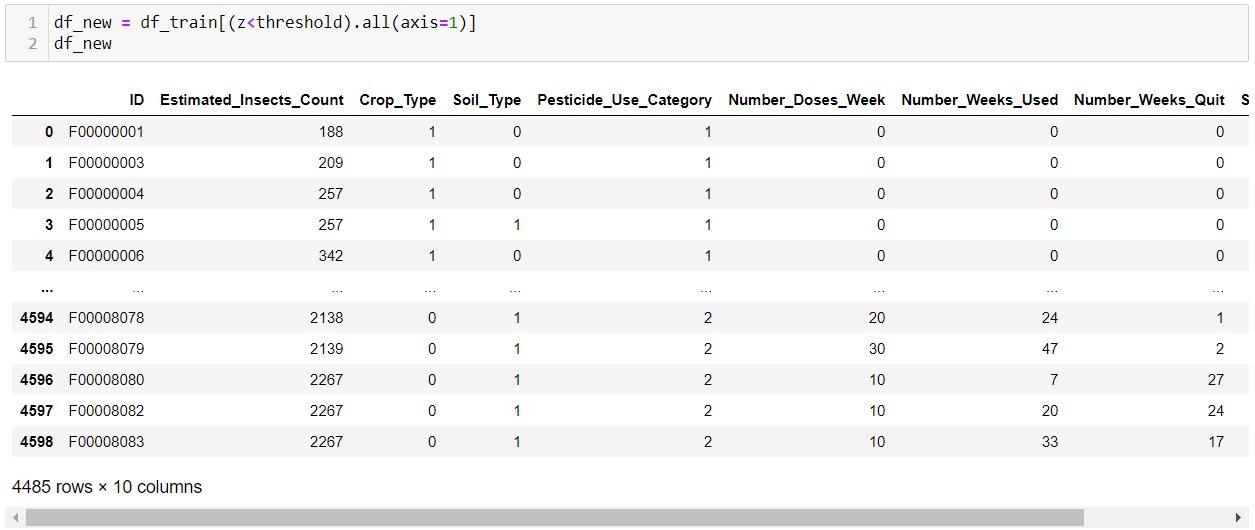
**Outlier Removal:**

Data outliers can spoil and mislead the training process resulting in longer training times, less accurate models, and, ultimately, more mediocre results. So, we must remove the outliers present in the dataset. But, we must be careful not to delete large amounts of data from the dataset as it may cause over fitting of the models. Up to 5% loss in data is considered acceptable. We can remove the outliers by two methods :

1. Z- Score Method - **Z-score** is a numerical measurement of how many [standard deviations](https://www.statisticshowto.com/probability-and-statistics/standard-deviation/) below or above the [population mean](https://www.statisticshowto.com/population-mean/)a [raw score](https://www.statisticshowto.com/raw-score/) or value is. If a Z-score is 0, it indicates that the data point's score is identical to the mean value. A Z-score of 1.0 would indicate a value that is one standard deviation away from the mean. Z-scores may be positive or negative, with a positive value indicating the score is above the mean and a negative score indicating it is below the mean. . Z-scores range from -3 standard deviations (which would fall to the far left of the normal distribution curve) up to +3 standard deviations (which would fall to the far right of the normal distribution curve). Therefore, we set threshold as 3 and delete all the values that are greater than the threshold.
2. Inter-Quartile Range(IQR) Method - In this method, IQR, “maximum” and “minimum” are calculated. Values less than minimum and greater than the maximum are deleted.

First, we try with the Z-score method, if the data loss is more than 5%, then we can try with IQR method.





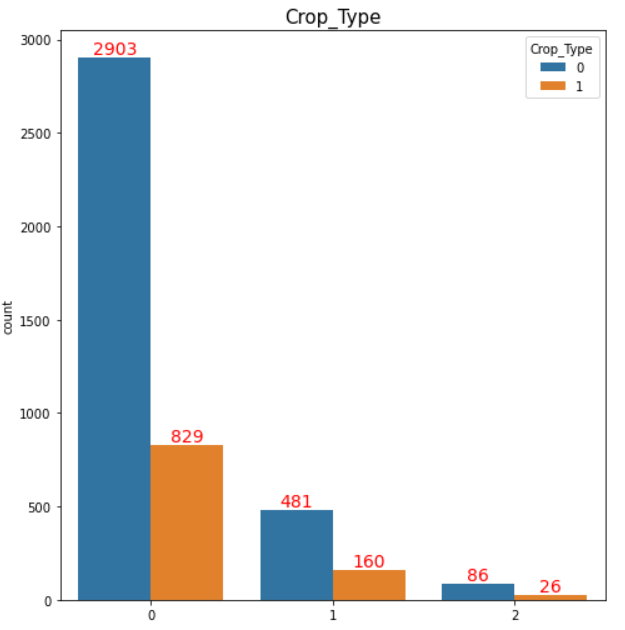
The new dataset has 4485 rows and 10 columns. The data loss percentage is 2.47%. Since the data lost in Z-score method is less than 5% we need not proceed with the IQR method.

**Bi-variate Analysis:**

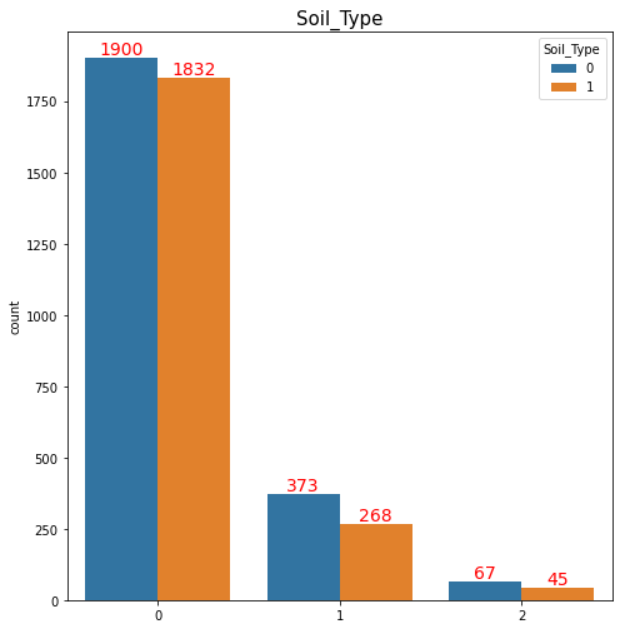
**Bi-variate analysis** is stated to be an analysis of any concurrent relation between two variables or attributes. This study explores the relationship of two variables as well as the depth of this relationship to figure out if there are any discrepancies between two variables and any causes of this difference.

Since our target variable is categorical we analyse how the other variables vary for each class in our target variable.

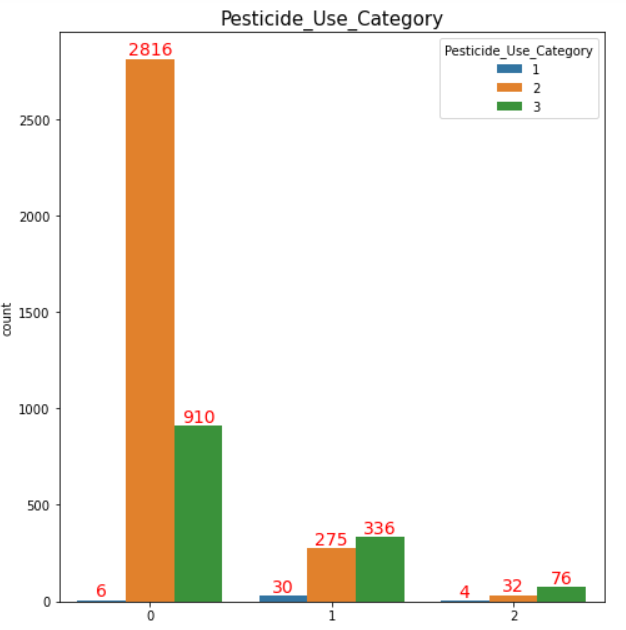
First, we analyze how the categorical variables vary with the target using count-plots.



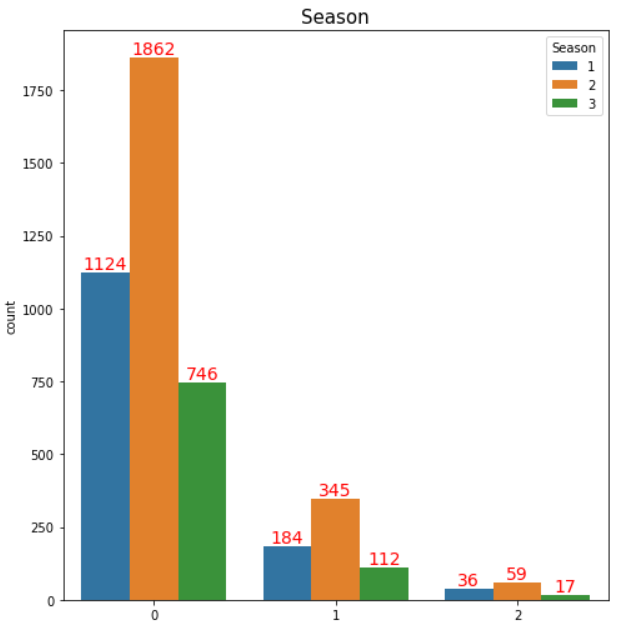
The ‘Crop\_Type’ field is distributed in similar ratios across all the different classes/categories of the target. So we can’t draw great insights from this comparison.



The ‘Crop\_Type’ field is distributed in similar ratios across classes 1 (damage due to other causes) and 2 (damage due to pesticides). In class 0 (Healthy crops) there is a higher ratio between soil type1 and 2 due to increase in count of Soil\_Type 2 category. We can say that higher percentage of crops sowed on Soil\_Type 2 tend to stay healthy.

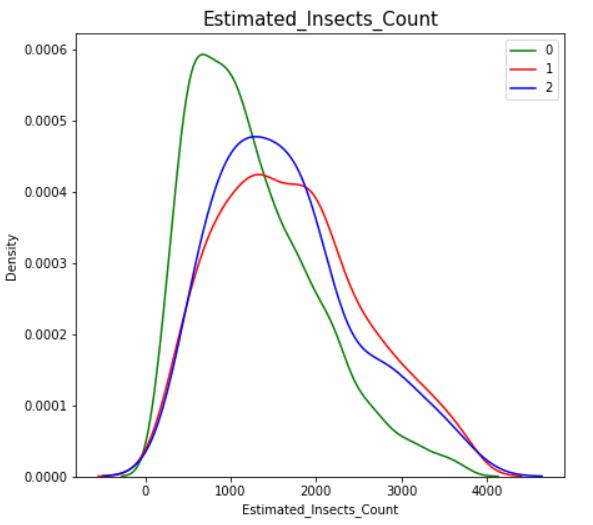


Crops that used type 3 pesticides seemed to be getting damaged due to the pesticide and other reasons. Most crops that are healthy had used type 2 pesticide. Type 1 pesticide was rarely used by the farmers.

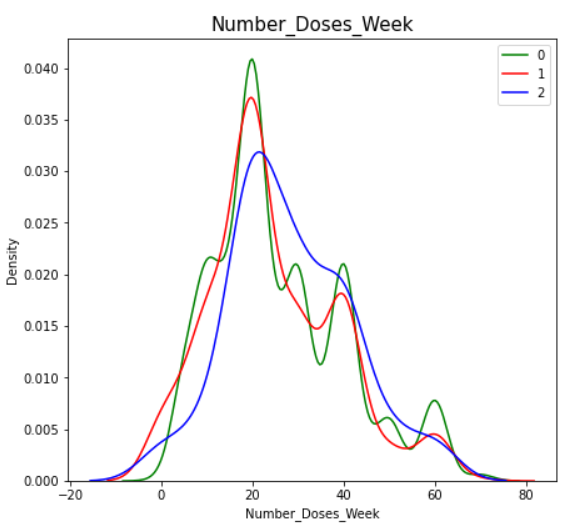


The ‘Season’ field is distributed in similar ratios across all the different classes/categories of the target. So we can’t draw great insights from this comparison.

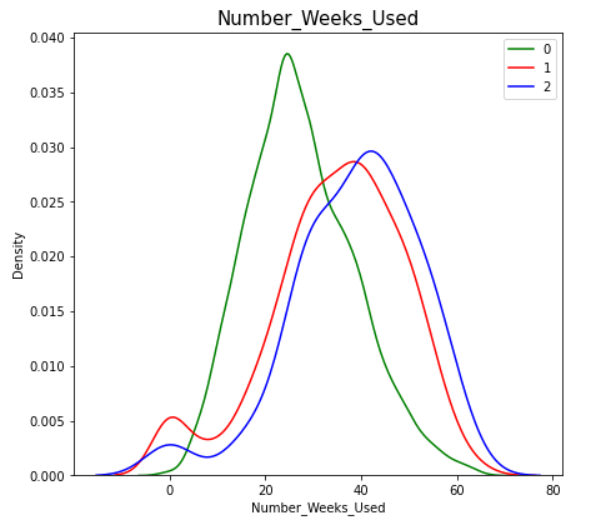
We can now analyze how the continuous variables vary with the target variable using distribution plots.



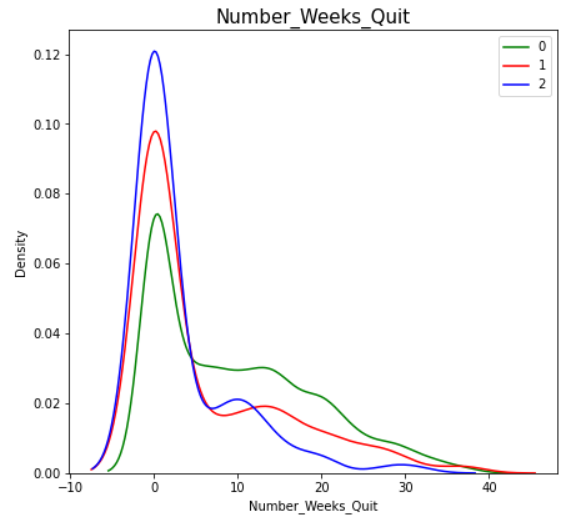
'Estimated\_Insects\_Count' column has a low value when crop is healthy/alive.



The , 'Number\_Doses\_Weeks' column is slightly right skewed for class 2 - Crops that are damaged by pesticide, meaning that higher number of doses per week can result in damage to the crops.

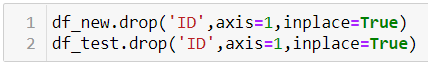


'Number\_Weeks\_Used' column has a slightly lower value for class 0- that is when the crop is healthy.



We can’t draw much of an information from the above graph.

We can drop the ID Columns from the test and training dataset.



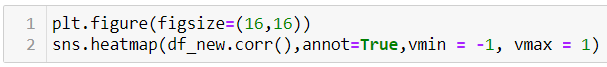
**Correlation:**

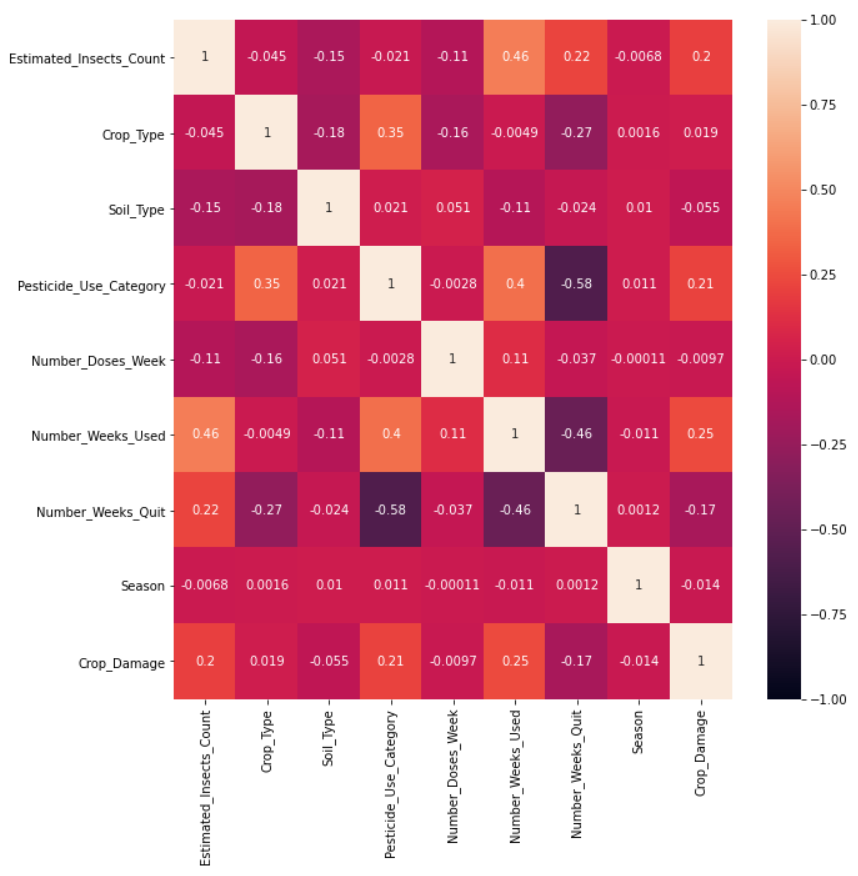
Correlation Matrix is basically a covariance matrix. A summary measure called the correlation describes the strength of the linear association. Correlation summarizes the strength and direction of the linear (straight-line) association between two quantitative variables. Denoted by r, it takes values between -1 and +1.

Correlation value of each column is categorized into mainly 2 parts that are:

* Positive correlated value means that when one variable decreases as the other variable decreases, or one variable increases while the other increases.
* Negative correlated value is the vice versa of positive correlated value.

Correlation matrix is plotted using the seaborn heatmap.





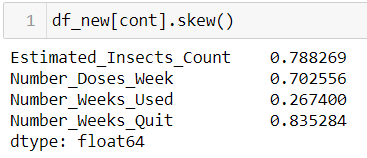
'Estimated\_Insects\_Count', 'Pesticide\_Use\_Category' and 'Number\_Weeks\_Used' have a positive correlation with Crop\_Damage variable.

**EDA Conclusion:**

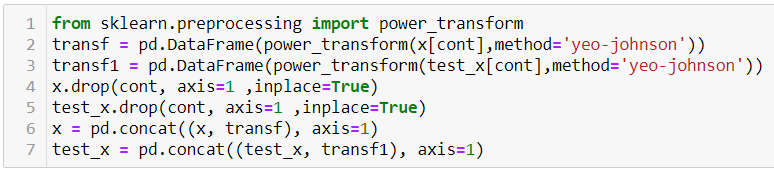
* High percentage of crops sowed on Soil\_Type 2 tend to stay healthy.
* Farmers prefer to sow the crops in season 2.
* Healthy crops have lower insects and lower value for number of weeks that pesticide is used.
* More number of pesticide doses per week tend to damage the crops.
* Type 2 pesticide is the safest compared to others.
* Type 1 pesticide has rarely been used by the farmers.

**Removing Skewness:**

We first check for skewness in the continuous columns.



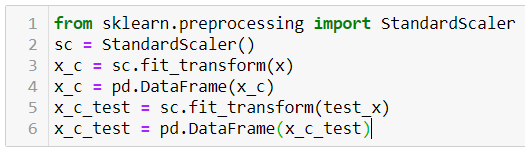
We can remove skewness in many ways, I am currently using Yeo-Johnson transformation - this is one of the older transformation techniques which is very similar to Box-cox transformation but does not require the values to be strictly positive. This transformation also has the ability to make the distribution more symmetric.



**Feature Scaling:**

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. 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.

I have used the StandardScaler (It transforms the data in such a manner that it has mean as 0 and standard deviation as 1) from the sklearn library to scale the features.

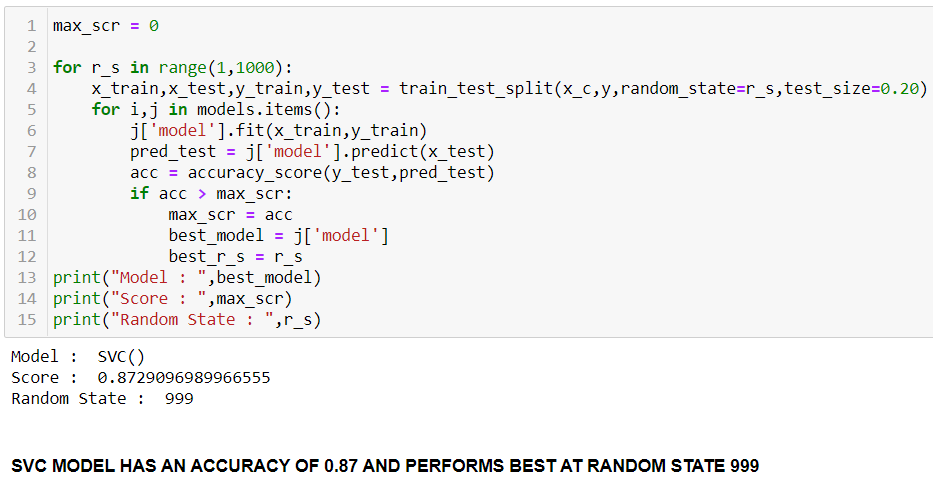


**MODEL BUILDING**

The given problem is a classification, so initially we can define a dictionary with some of best the classifier models and their parameters.

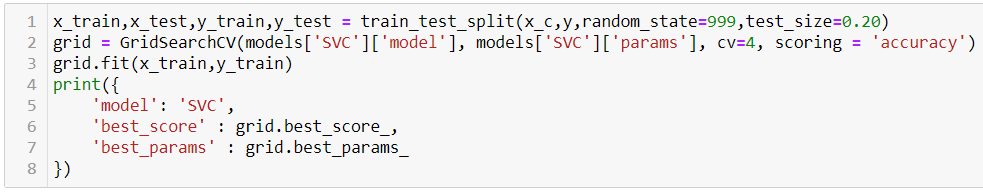


I am using the for loop which help me to provide the accuracy score at each random state and for the best state where accuracy score is maximum is displayed come as output value.

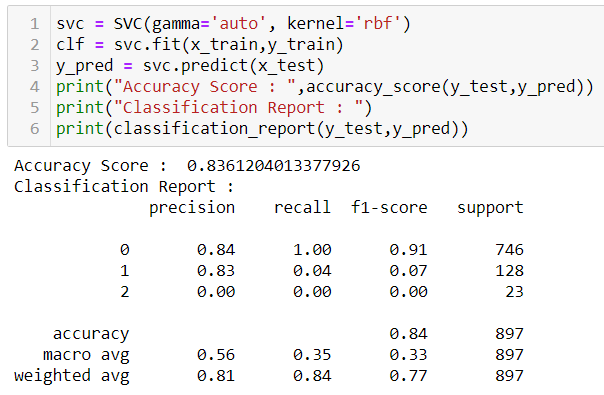


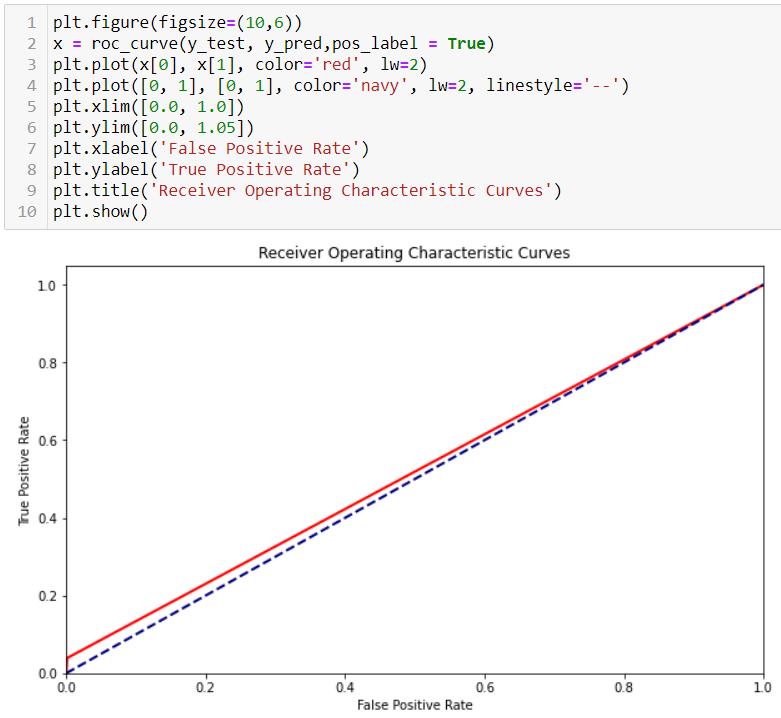
SVC MODEL HAS THE HIGHEST ACCURACY OF 0.87 AND PERFORMS BEST AT RANDOM STATE 999.

After obtaining the best model we find the best parameters using GridSearchCV. GridSearchCV uses cross validation to evaluate the model’s performance and this helps in preventing the over-fitting of the model.



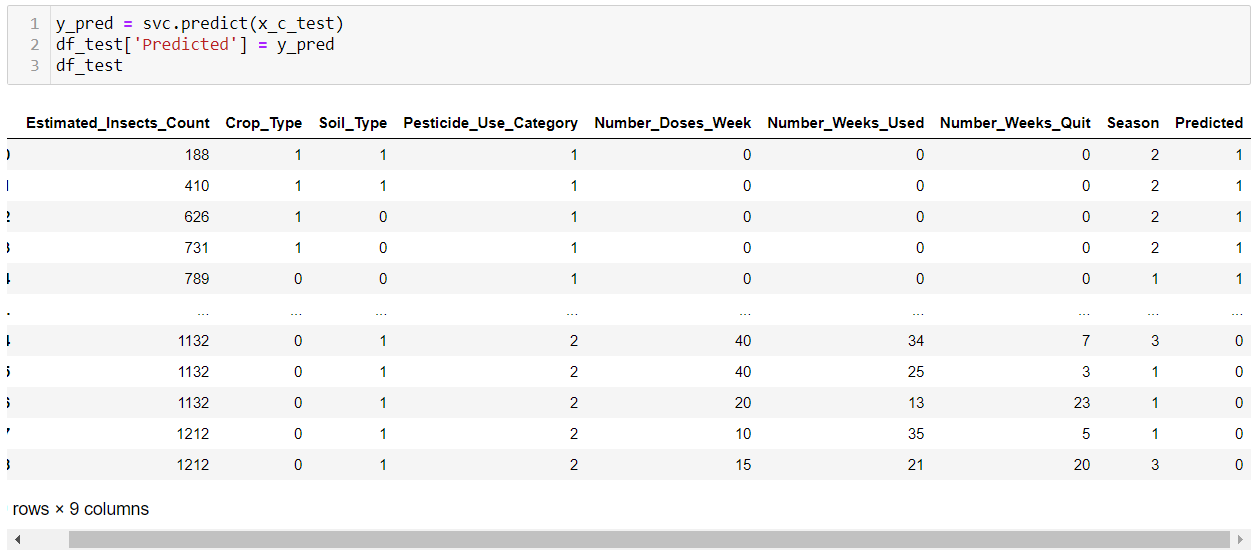
We can now evaluate the model’s performance.





As we can see from above that the model’s performance is quite good.

We can now predict the values for the test dataset.





**CONCLUDING REMARKS**

* SVC was found to be the best classifier for the dataset with an accuracy score of 0.87.
* The SVC model performed best for the following parameters : *gamma='auto', kernel='rbf'.*
* The model predicted 1177 crops to be alive out of the 1199 entries in the dataset, 22 to be damaged by other causes and 0 crops to be damaged due to pesticides.