PROJECT

**Wine data set**

**Problem Definition:**

**the given wine dataset is related to red, white and many other types of wines.** the wine data set can be viewed as regression or classification tasks.

The given data has the following columns

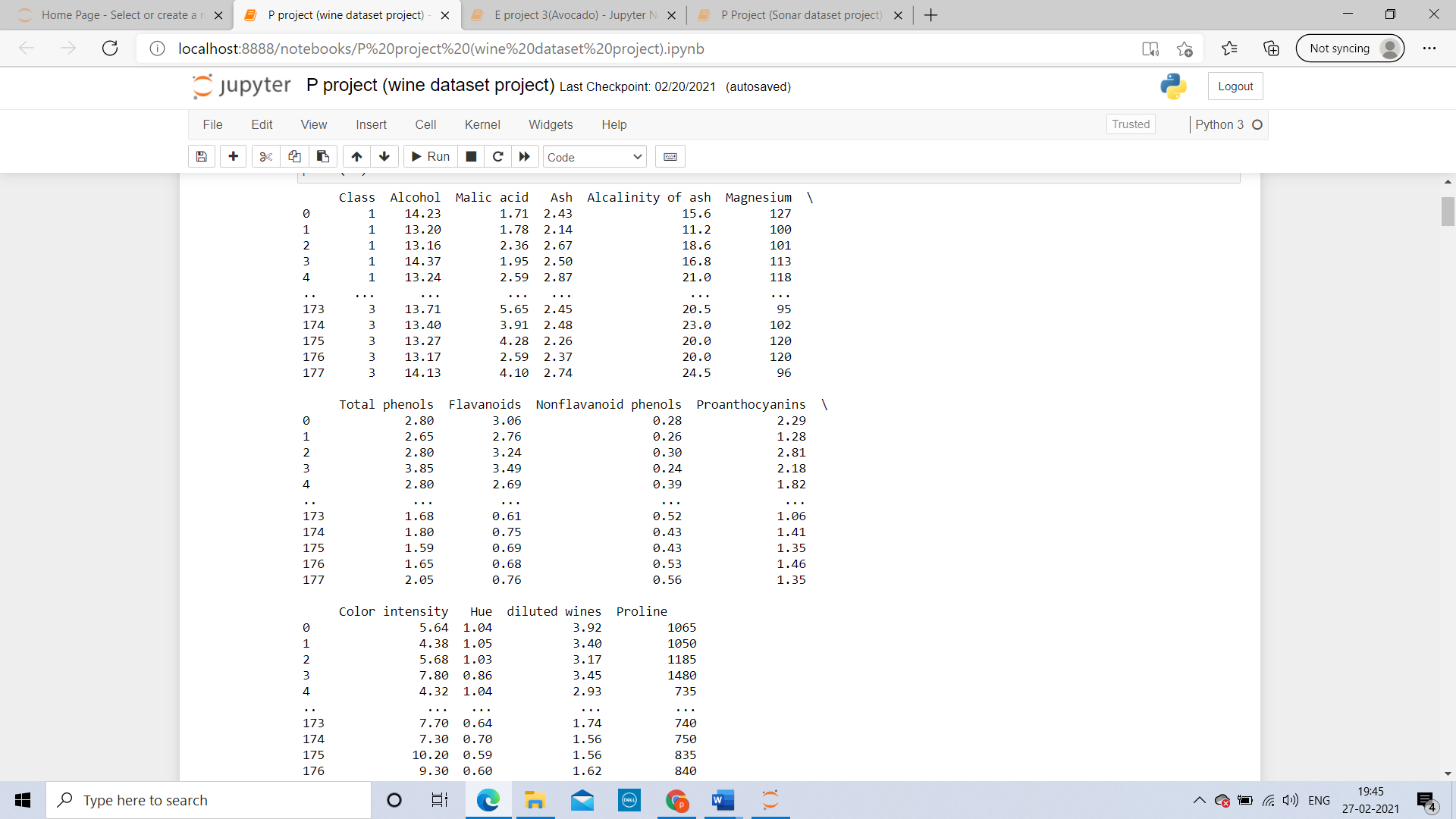
1) fixed acidity  
2) volatile acidity  
3) citric acid  
4) residual sugar  
5) chlorides  
6) free sulfur dioxide  
7) total sulfur dioxide  
8) density  
9) pH  
10) sulphates  
11) alcohol  
12) quality

The above columns when mixed in different ratio and proportion gives different type and taste of wine. In these problems, we are going to divide the wine basically into three classes

Let us start with importing the necessary libraries .



**Data Analysis**



The given data has 178 rows and 14 columns

**The columns are:**

1. Class
2. Alcohol
3. Malic acid
4. Ash
5. Alcalinity of ash
6. Magnesium
7. Total phenols
8. Flavanoids
9. Nonflavanoid phenols
10. Proanthocyanins
11. Color intensity
12. Hue
13. diluted wines
14. Proline

**Datatypes are:**

**Integer datatype:**

1)Class

2)Magnesium

3) Proline

**float datatype:**

1)Alcohol

2)Malic acid

3)Ash Alcalinity of ash

4)Total phenols

5)Flavanoids

6)Nonflavanoid phenols

7)Proanthocyanins

8)Color intensity

9) Hue diluted

**Null values:** there are no null values in the given data

**Missing values:** there are no missing or Nan values in the given datatype

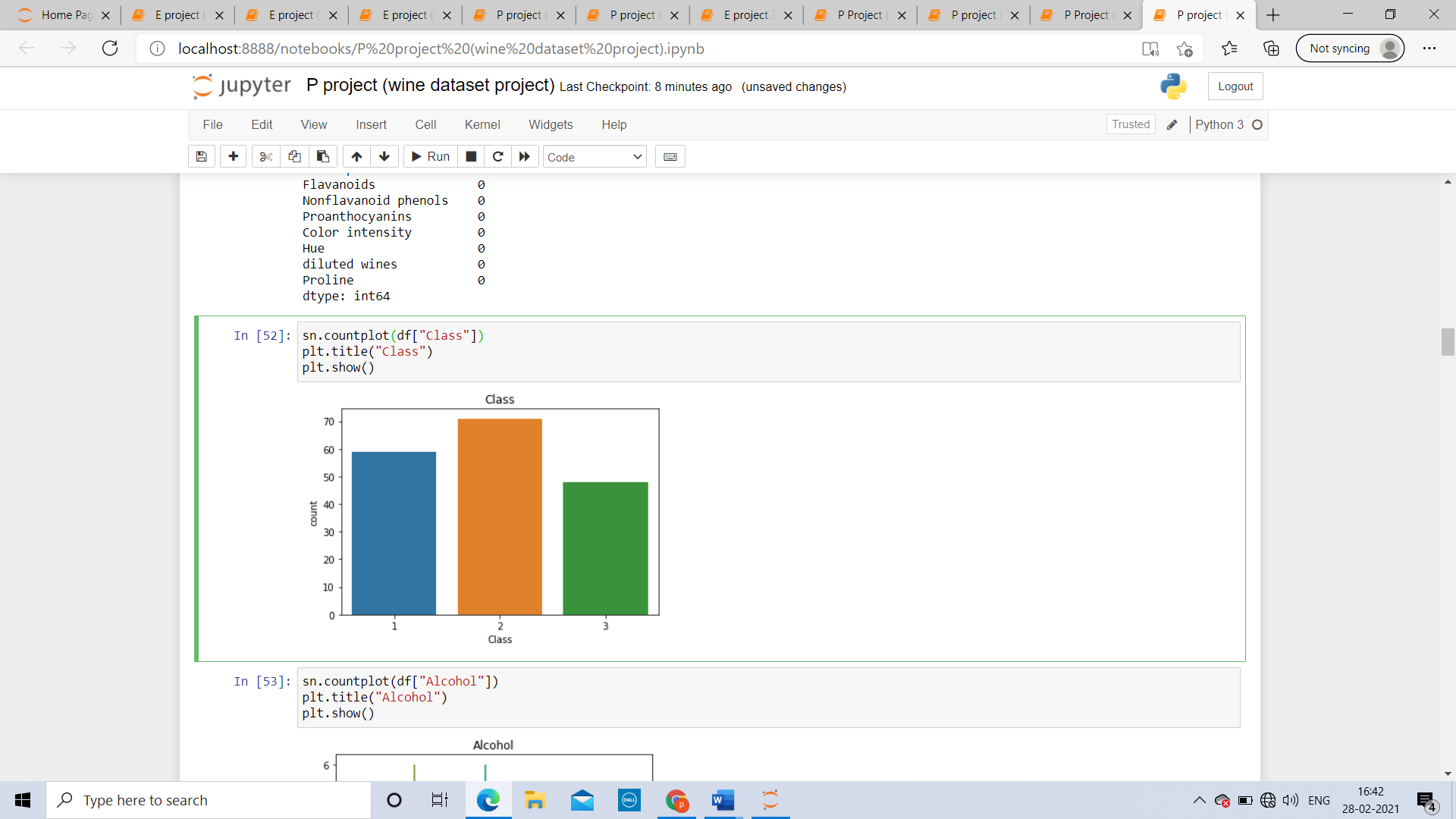
**Target variable**: our target variable in the given data will be class

**EDA:**

**UNIVARIANT ANALYSIS:**

CLASS: there are 3 types of class

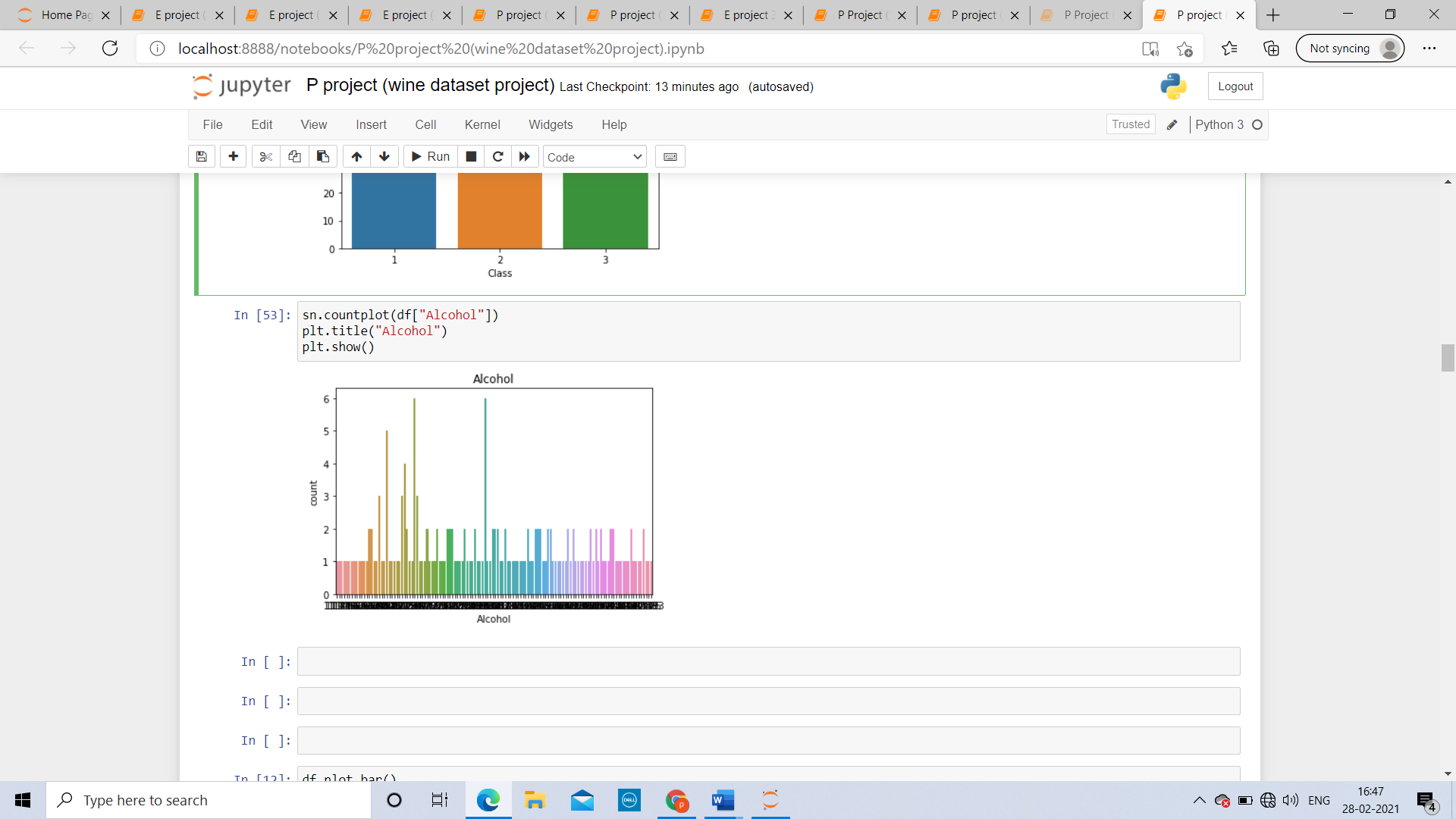
Class1, class2 and class3



Class 2 have the highest count around 70,

class 3 have the least count around 50 and class 1 have medium count around 60

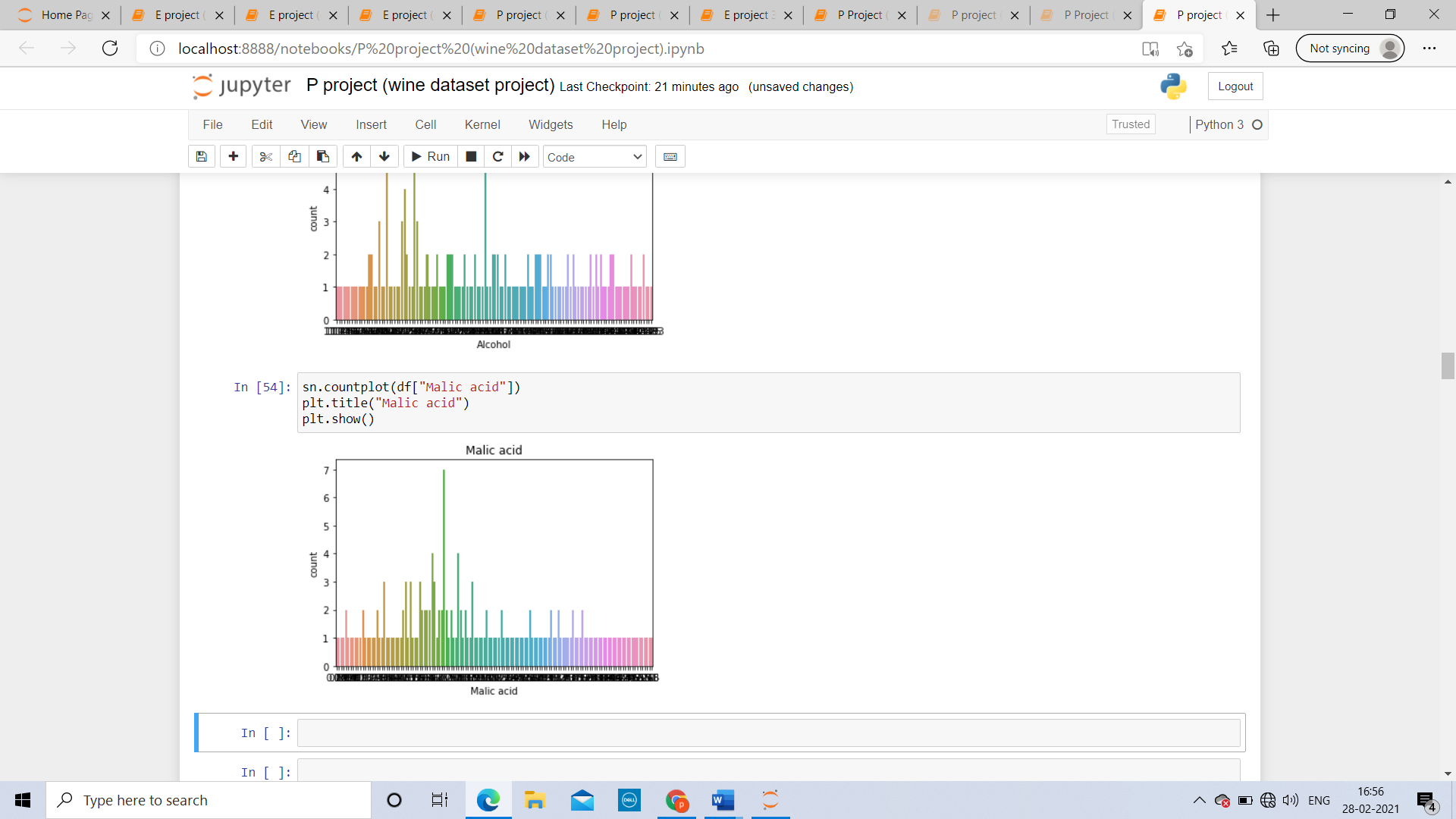
**Alcohol:**



The highest count of alcohol is 6 and the least count of alcohol is 1.

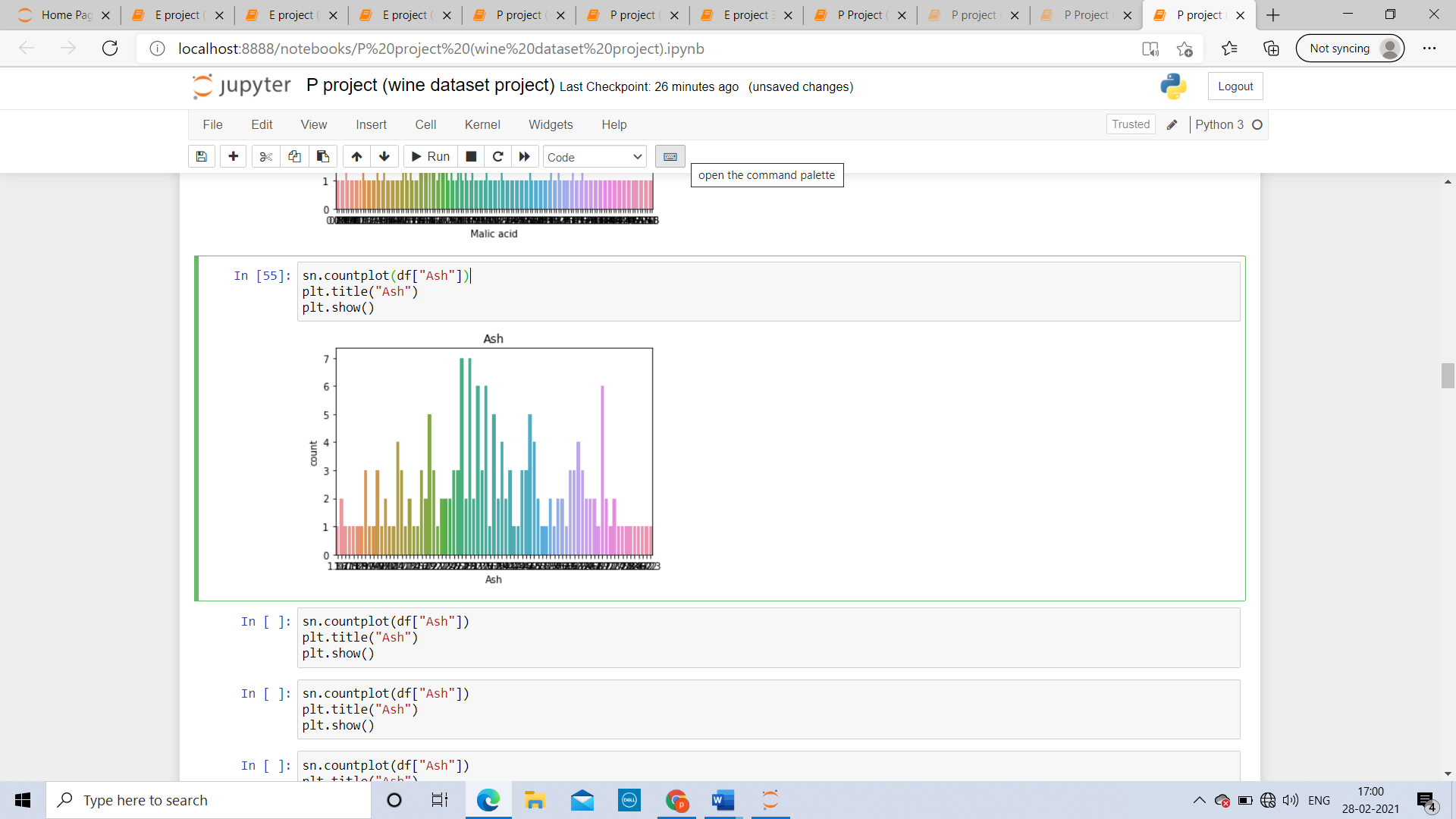
Average count is alcohol is 1.

**Malic acid:**



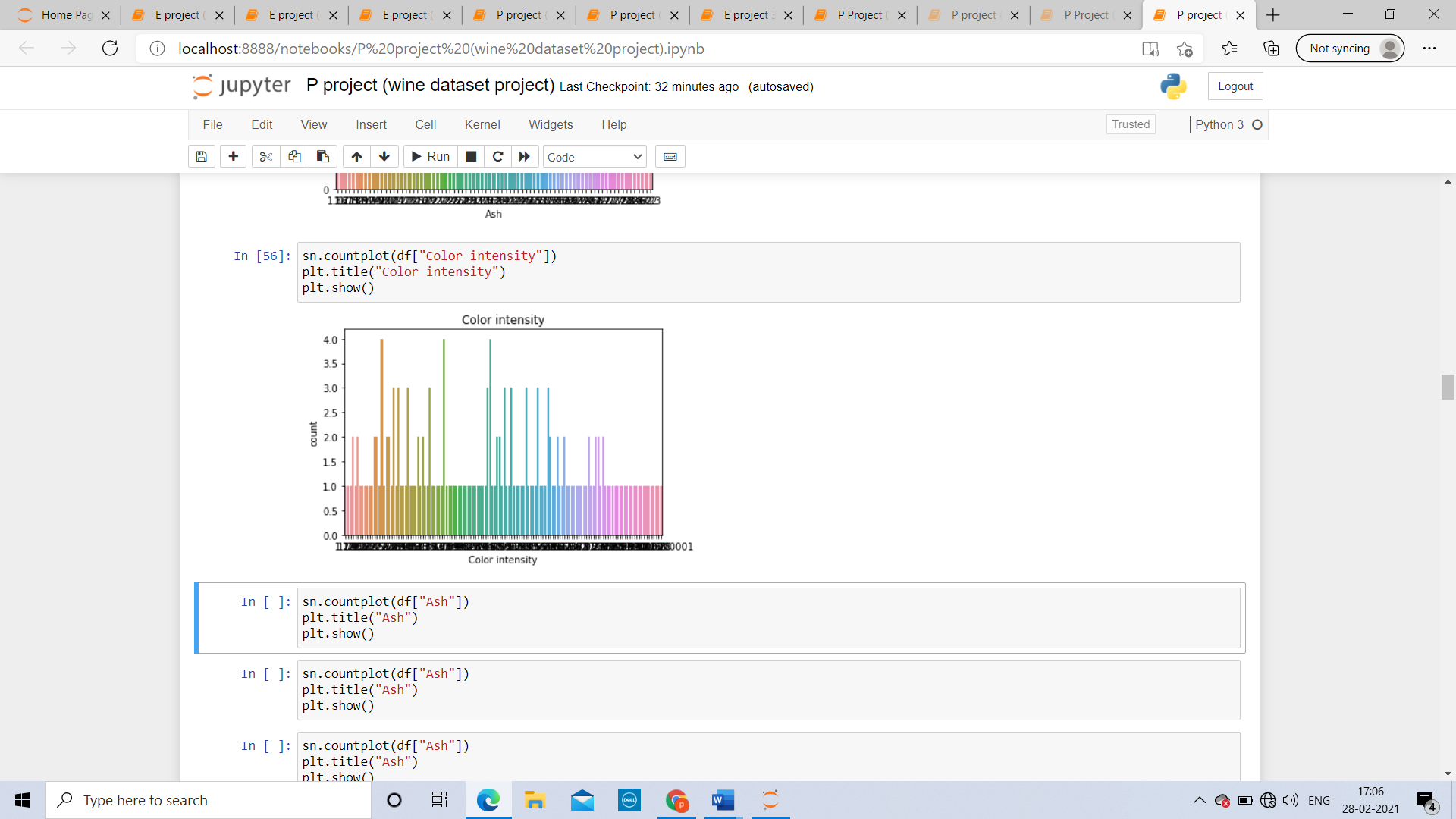
The highest count of Malic acid is 7 and the least count is 1. the average count is also 1

**Ash:**



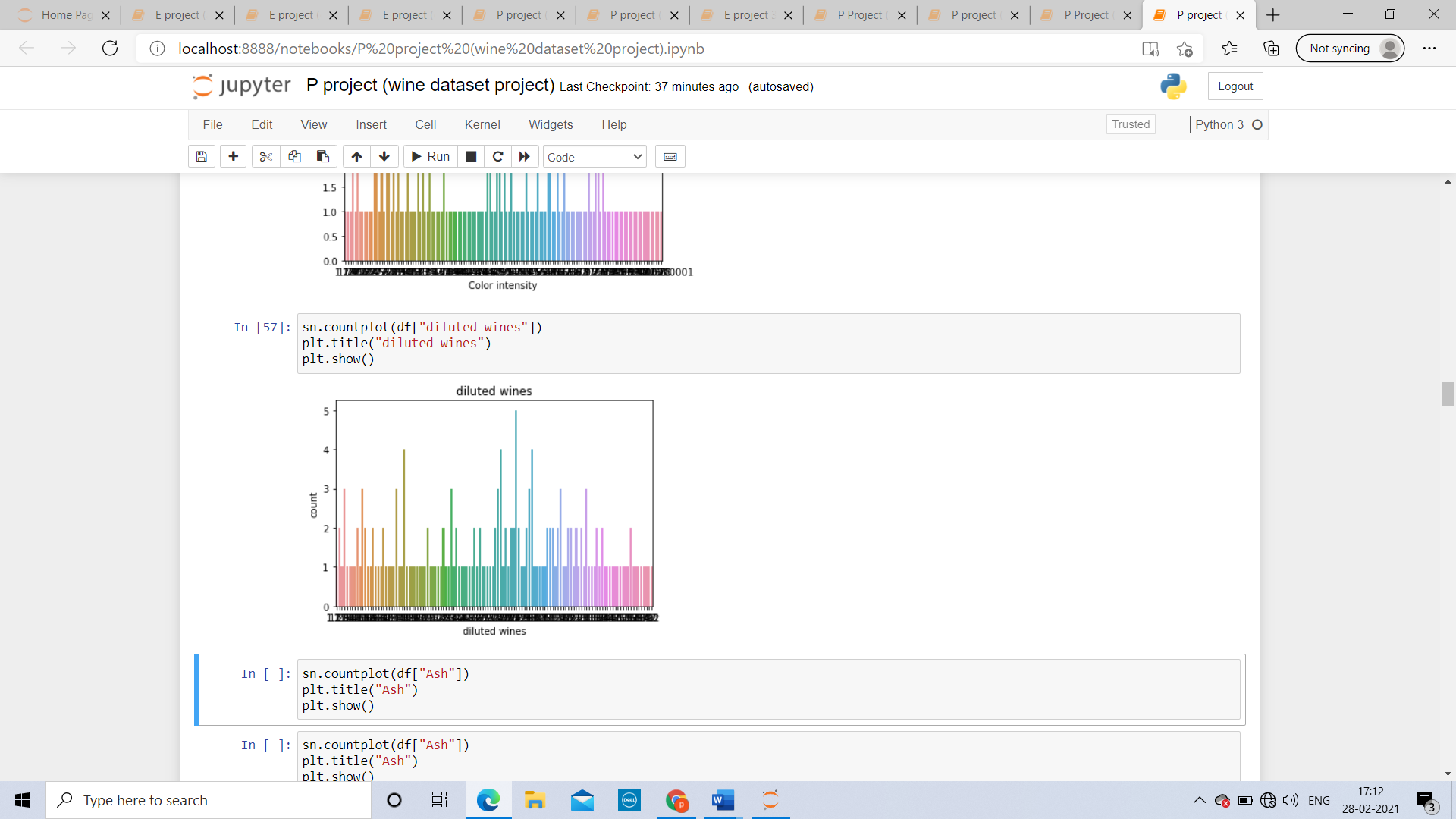
The highest count of the Ash is 7 and the least count is 1. the average count is also 1

**COLOR INTENSITY**:



The highest count of the Ash is 4 and the least count is 1. the average count is also

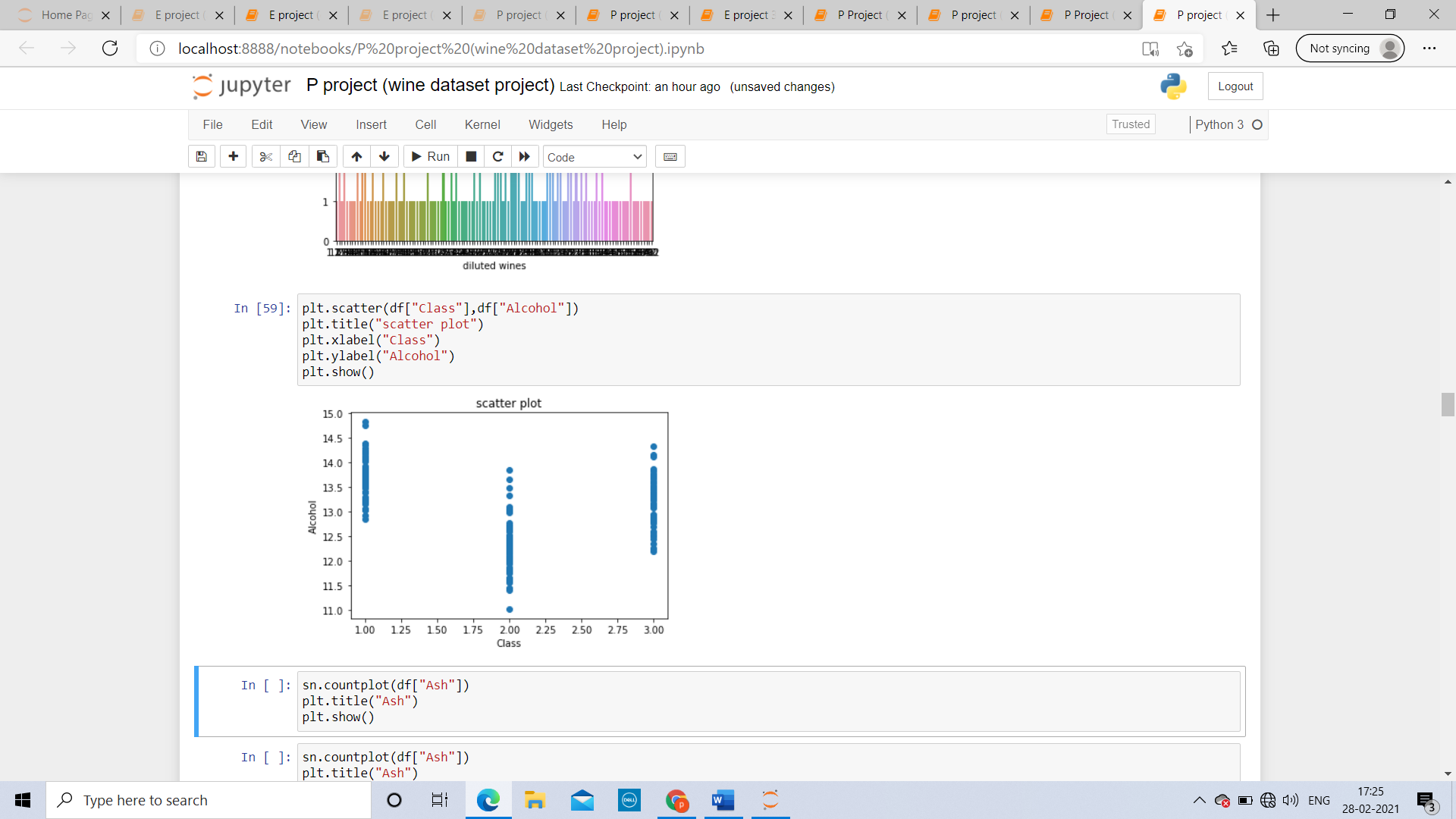
**diluted wines:**



The highest count of the Ash is 5 and the least count is 1. the average count is also 1

**BIVARIANT ANALYSIS:**

Between alcohol and class



The range of alcohol in class 1 is from 12.5 – 15

The range of alcohol in class 2 is from 11- 14

The range of alcohol in class 3 is from 12- 14.5

**Between color intensity and class:**

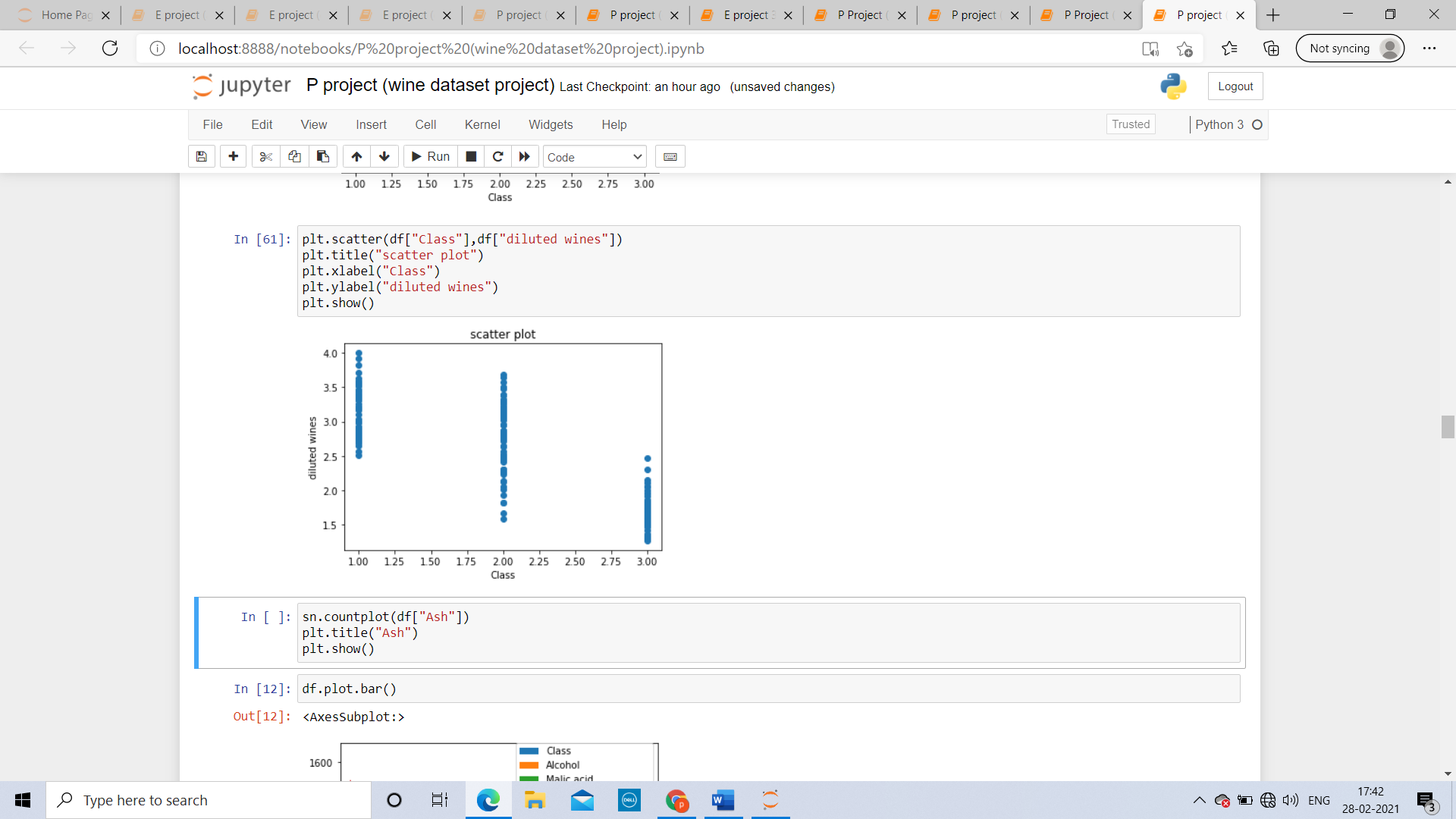


Color intensity in class 1 is range between 5.5 - 9

Color intensity is most in the class 3 wines range between 4 -13

Color intensity is least in class 2 wines range from around 0.5 - 5

**Between diluted wines and class:**



We can see that most of the diluted wines belongs to class 2

And least of the diluted wines belongs to the class 3

**MULTIVARIANT ANALYSIS:**



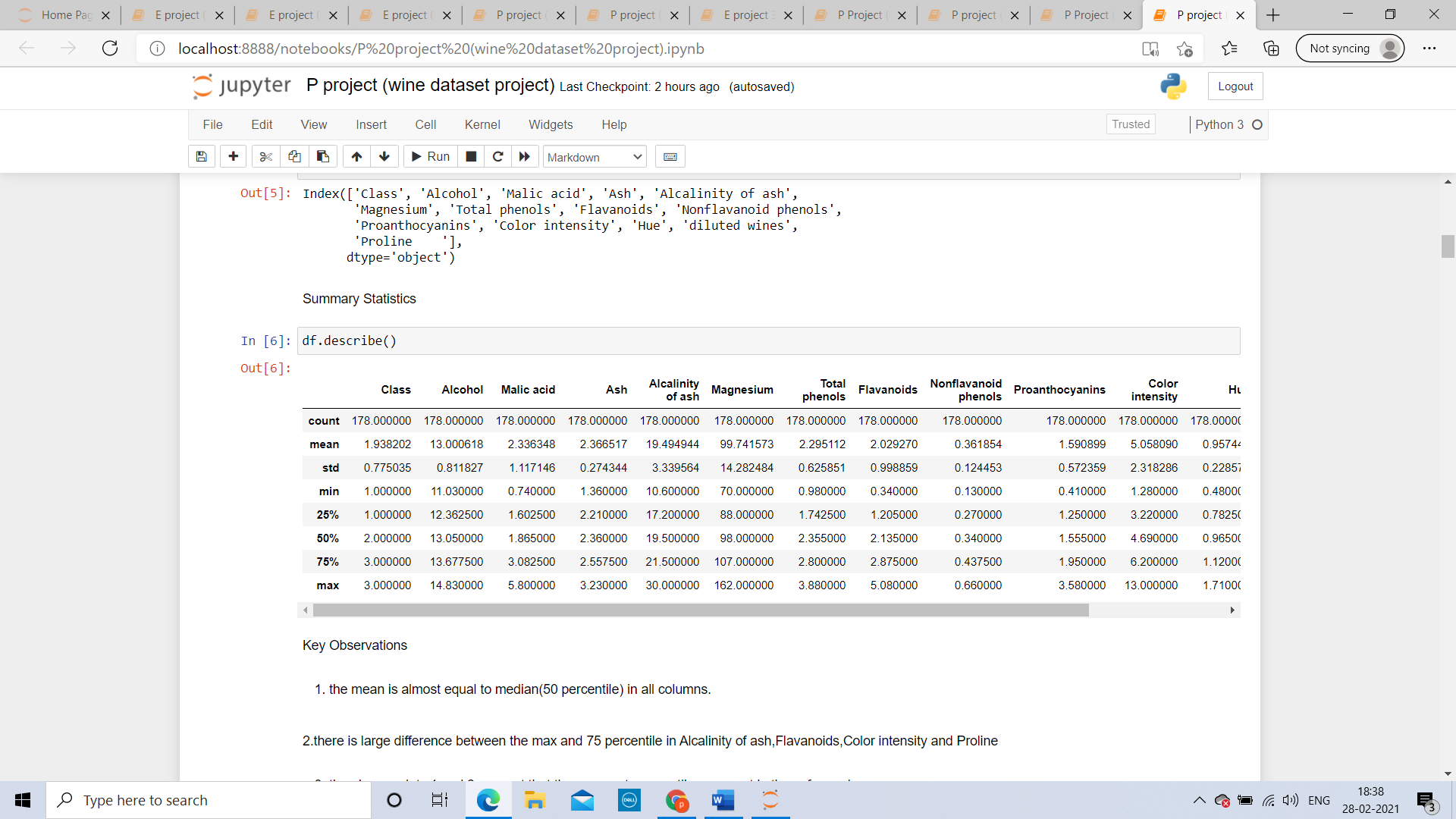
**Negative correlation**: Class has negative correlation with alcohol, ash, magnesium, total phenols, flavonoids, proanthocyanins, hue, diluted wines and proline

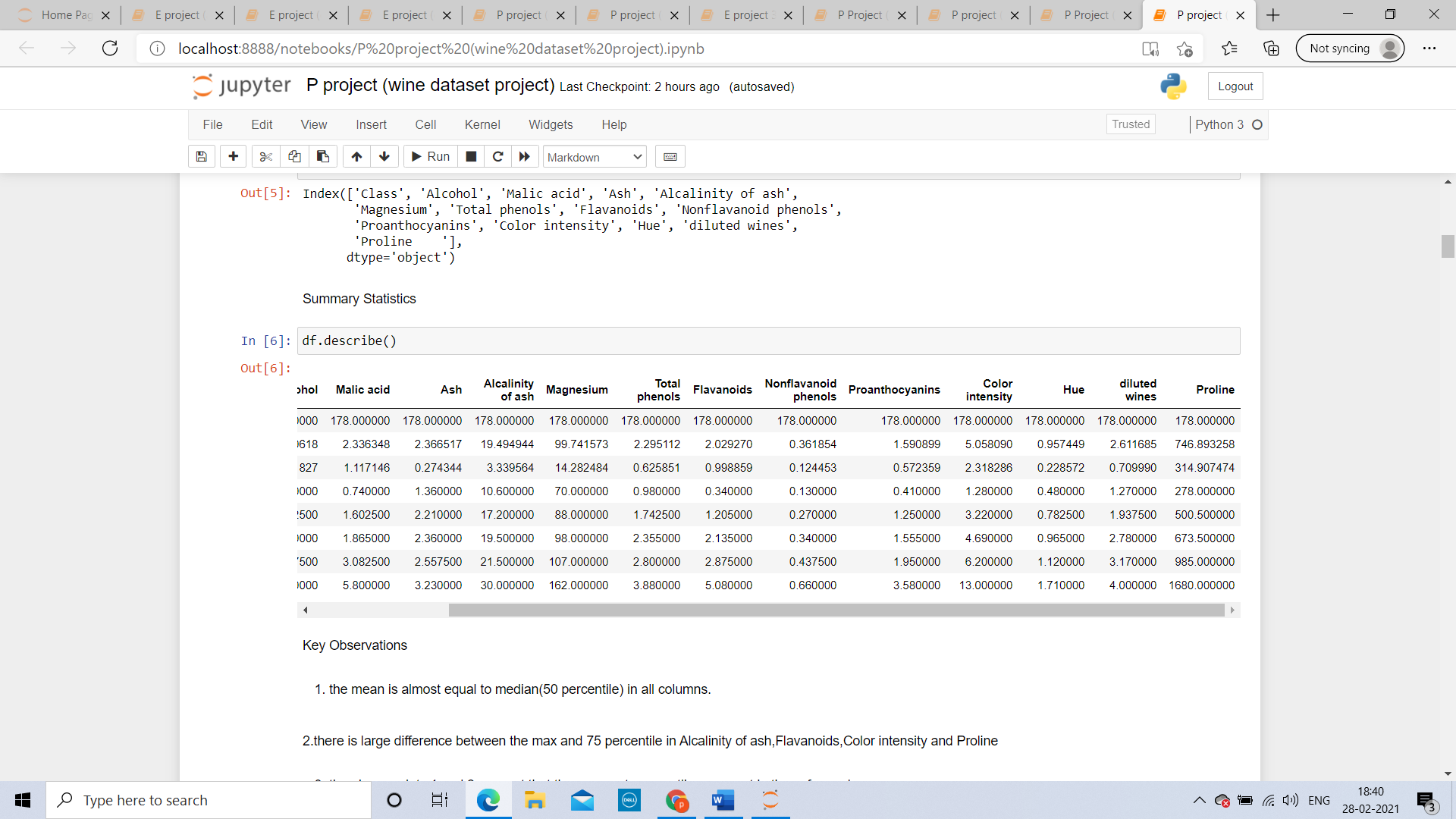
**Positive correlation**: Class has positive correlation with malic acid, alkalinity of ash, nonflavanoid phenols and color intensity

**Good correlation**: class has good correlation with alcohol, magnesium, total phenols, flavonoids, hue, diluted wines and proline, malic acid, alkalinity of ash, nonflavanoid phenols and color intensity

**No so good correlation:** class has not so good correlation withash andproanthocyanins

**PRE-PROSSECING PIPELINE:**





**Key Observations**

1. the mean is almost equal to median (50 percentile) in all columns.
2. there is large difference between the max and 75 percentiles in Alcalinity of ash, Flavanoids, Color intensity and Proline

3) the above points 1 and 2 suggest that there are extreme outliers present in these four columns.

**Removing outliers:** as analysis above there are outliers present in the given data, we need to remove the outliers using zscore method.

**Converting object values to int**: There are no object values to convert to integers.

Removing skewness: it is very important to remove skewness to avoid any kind of over fitting or under fitting, to remove the skewness we will use StandardScaler and fit method.

Now we will split the data into two parts

1. train data
2. test data

the split of our train data and test data will be 25% and 75%

**NOW FINDING THE BEST MODEL:**

**Linear Regression:**

we found Accuracy=100.0, cross validation score=100.0 & difference =0.0

**Random Forest Regressor:**

Accuracy=99.99453551912568, cross validation score =99.97087372216639 & difference =0.023661796959288495

**Ada Boost Regressor**:

Accuracy=100.0, cross validation score=97.87336443586445 & difference =2.1266355641355545

**SGD Regressor:**

Accuracy=94.11001904585422, cross validation score =94.95559601455462 & difference =-0.8455769687003993

**RESULT**:

Linear Regression and Ada Boost Regressor are performing the best with same accuracy and cross validation score. I will choose Linear Regression.

**Agriculture dataset**

**PROBLE DEFINITION:**

Int the given agriculture dataset types of crops are given , we need to analyze the crop damage depending on different attributes

The columns are

1) ID

2) Estimated\_Insects\_Count

3) Crop\_Type

4) Soil\_Type

5) Pesticide\_Use\_Category

6) Number\_Doses\_Week

7) Number\_Weeks\_Used

8) Number\_Weeks\_Quit

9) Season

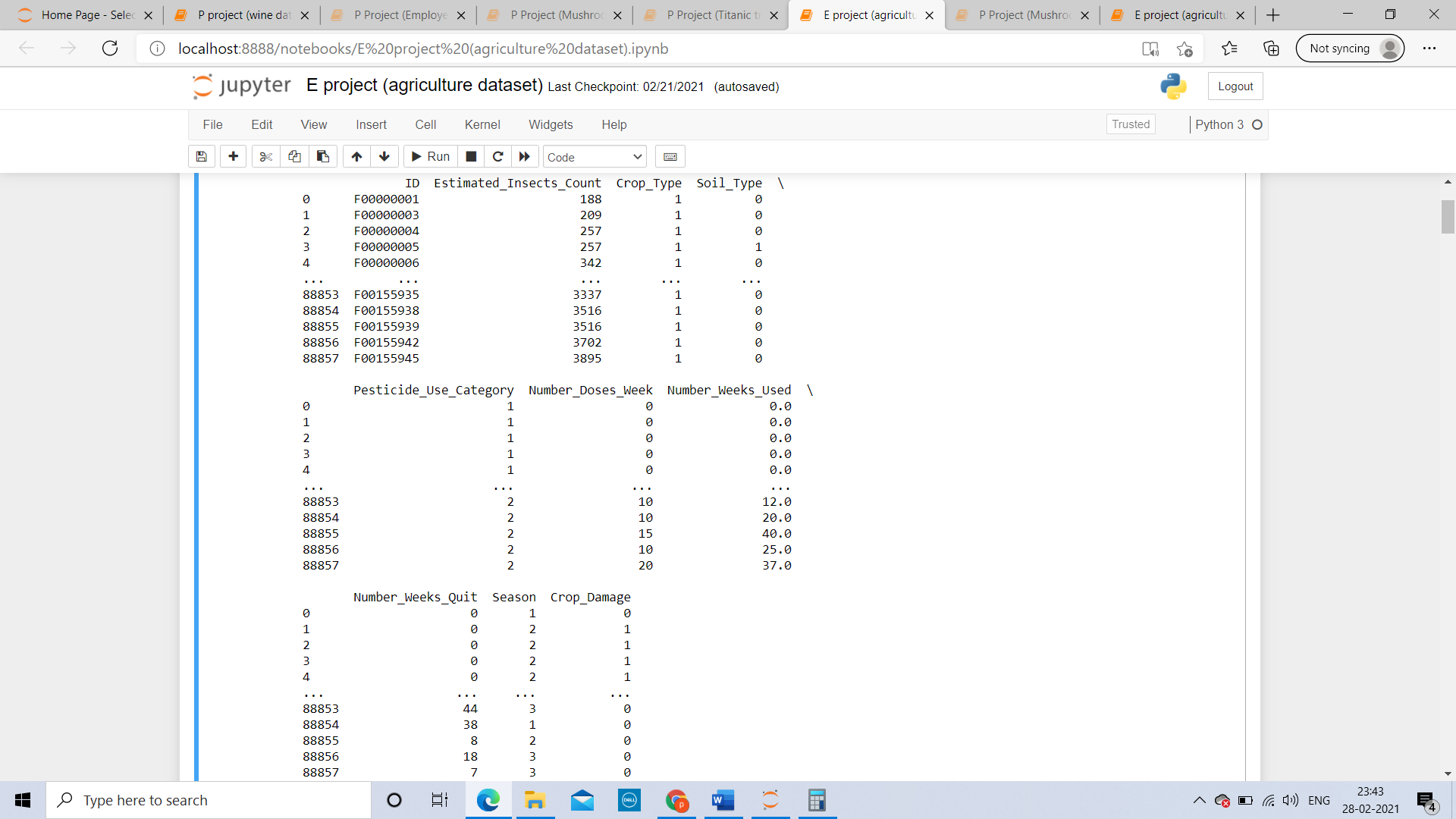
10)Crop\_Damage

Let us start with importing the necessary libraries.



**Data Analysis**

**Problem Definition:**



The given data has 88858 rows and 10 columns

**The columns** **are** :

1) ID

2) Estimated\_Insects\_Count

3) Crop\_Type

4) Soil\_Type

5) Pesticide\_Use\_Category

6) Number\_Doses\_Week

7) Number\_Weeks\_Used

8) Number\_Weeks\_Quit

9) Season

10)Crop\_Damage

**Datatypes**:

**Integer datatype**:

1)Estimated\_Insects\_Count

2)Crop\_Type

3)Soil\_Type

4)Pesticide\_Use\_Category

5)Number\_Doses\_Week

6)Number\_Weeks\_Quit

7)Season

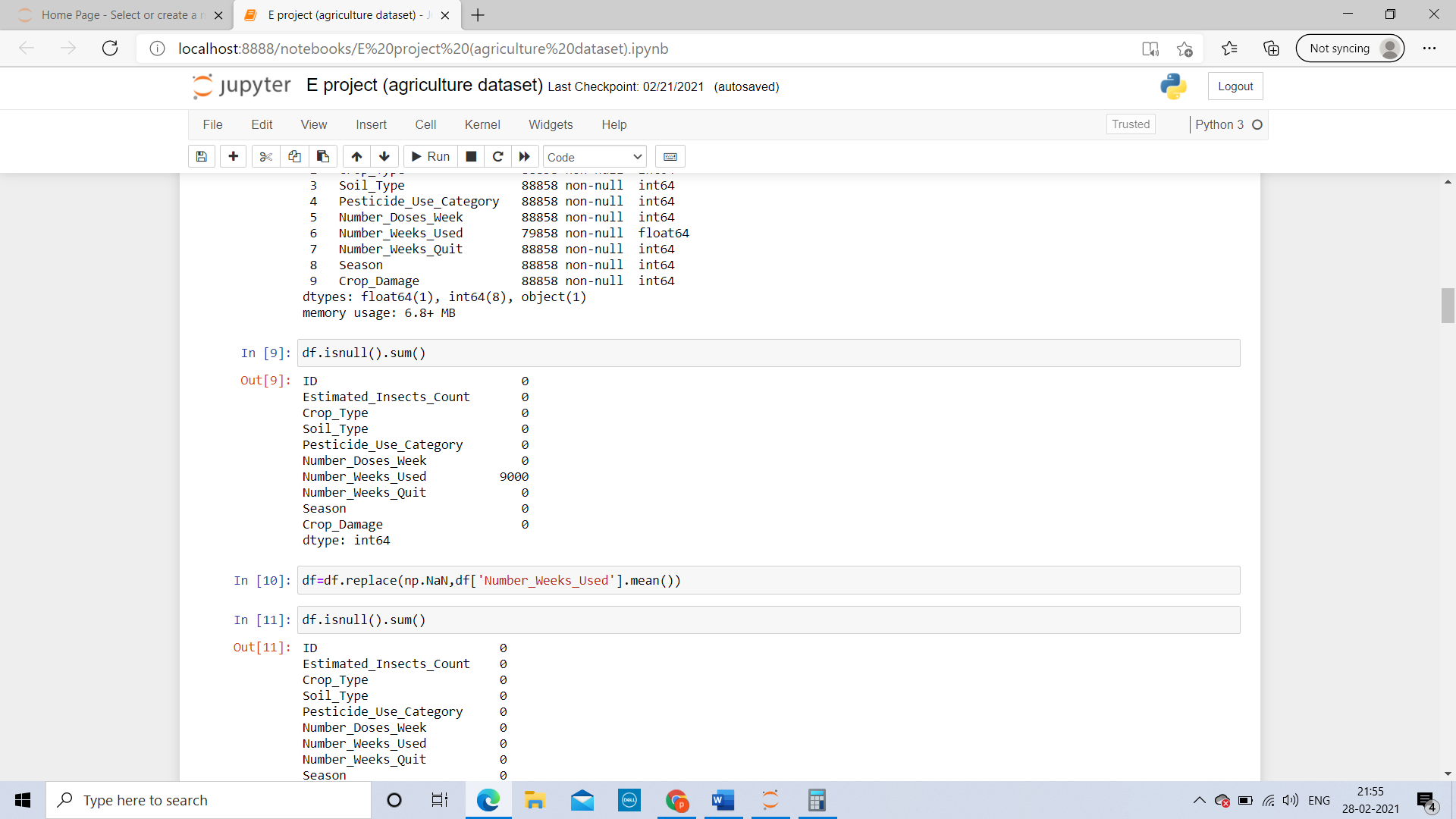
8)Crop\_Damage

**float datatype**: ID

**object datatype**: Number\_Weeks\_Used

**Null values:** there are no null values in the given data

**Missing values:** there are missing values in the given datatype



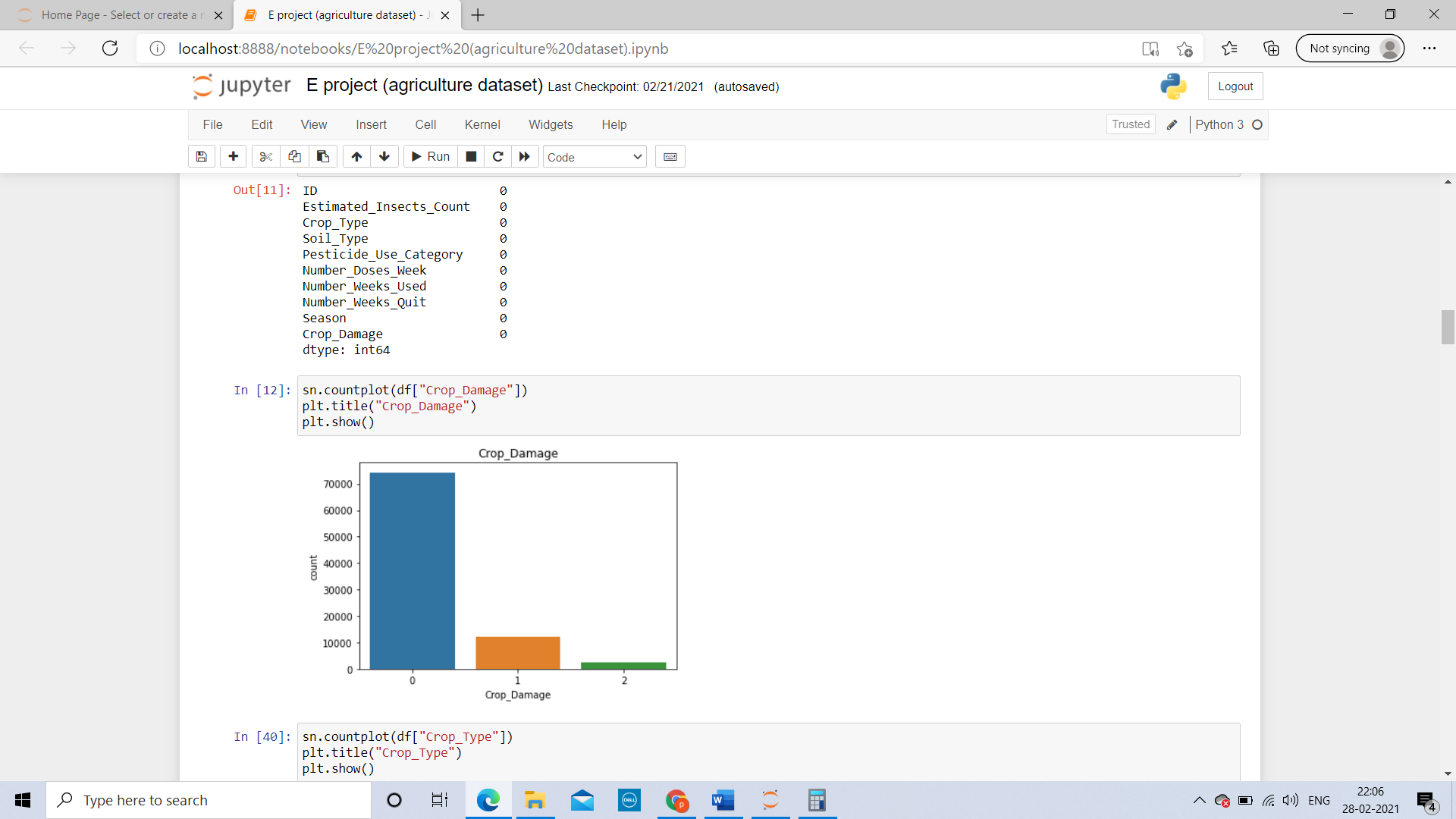
Using replace method we can fill the Nan values by mean

**Target variable**: our target variable in the given data will be Average Price

**EDA:**

**UNIVARIANT ANALYSIS:**

**Crop damage:**

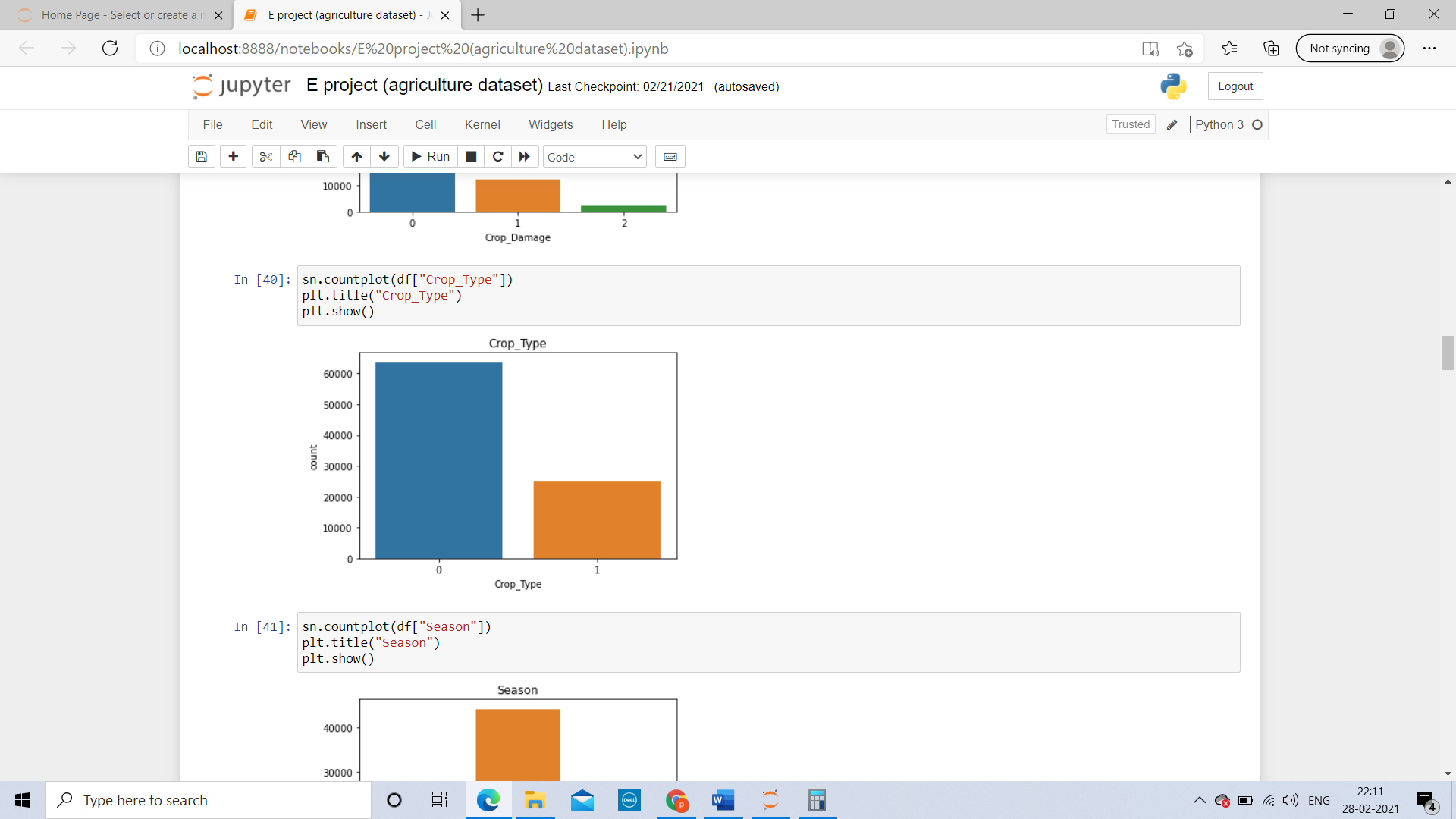


Crop type 0 has highest count of crop damage

Crop type 2 has least count of crop damage

Crop type 1 has crop damage around 10000.

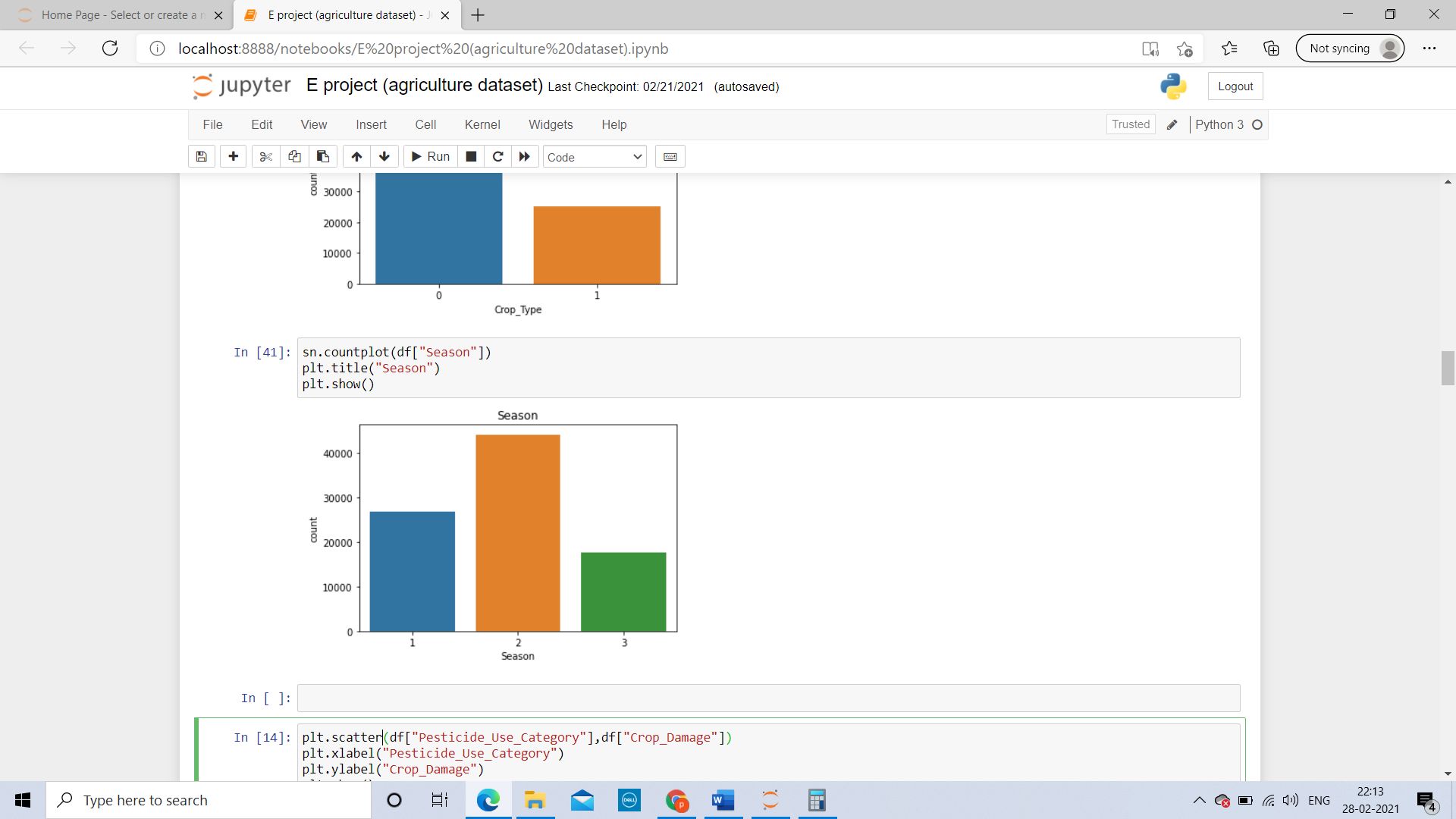
**CROP TYPE:**



Crop type 0 has the count of 60000

Crop type 1 has the count 25000

**SEASON:**



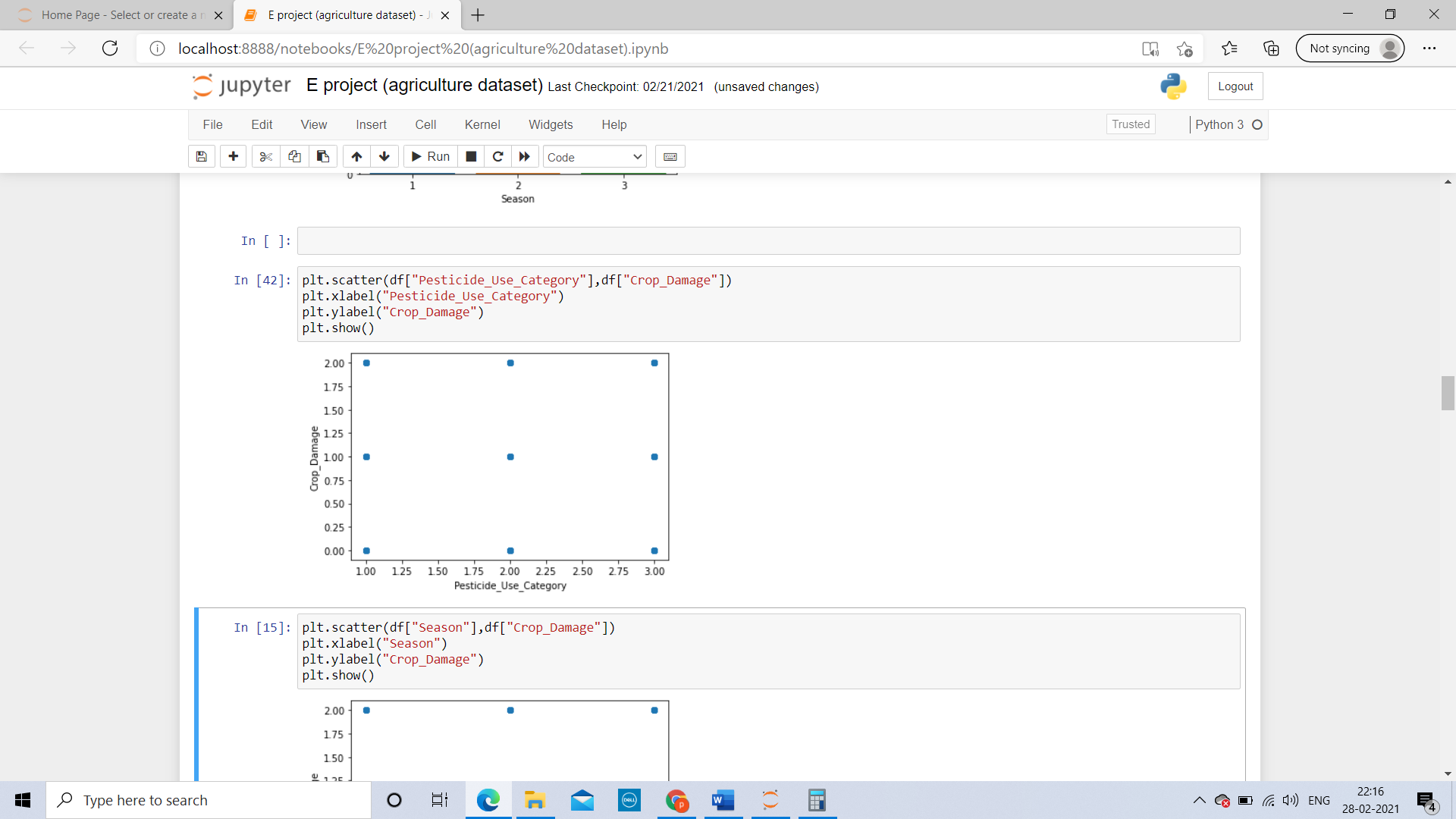
Season 1 has count of around 28000

Season 2 has count of around 45000

Season 3 has count of around 18000

**BIVARIANT ANALYSIS:**

**Between crop damage and pesticide use category**

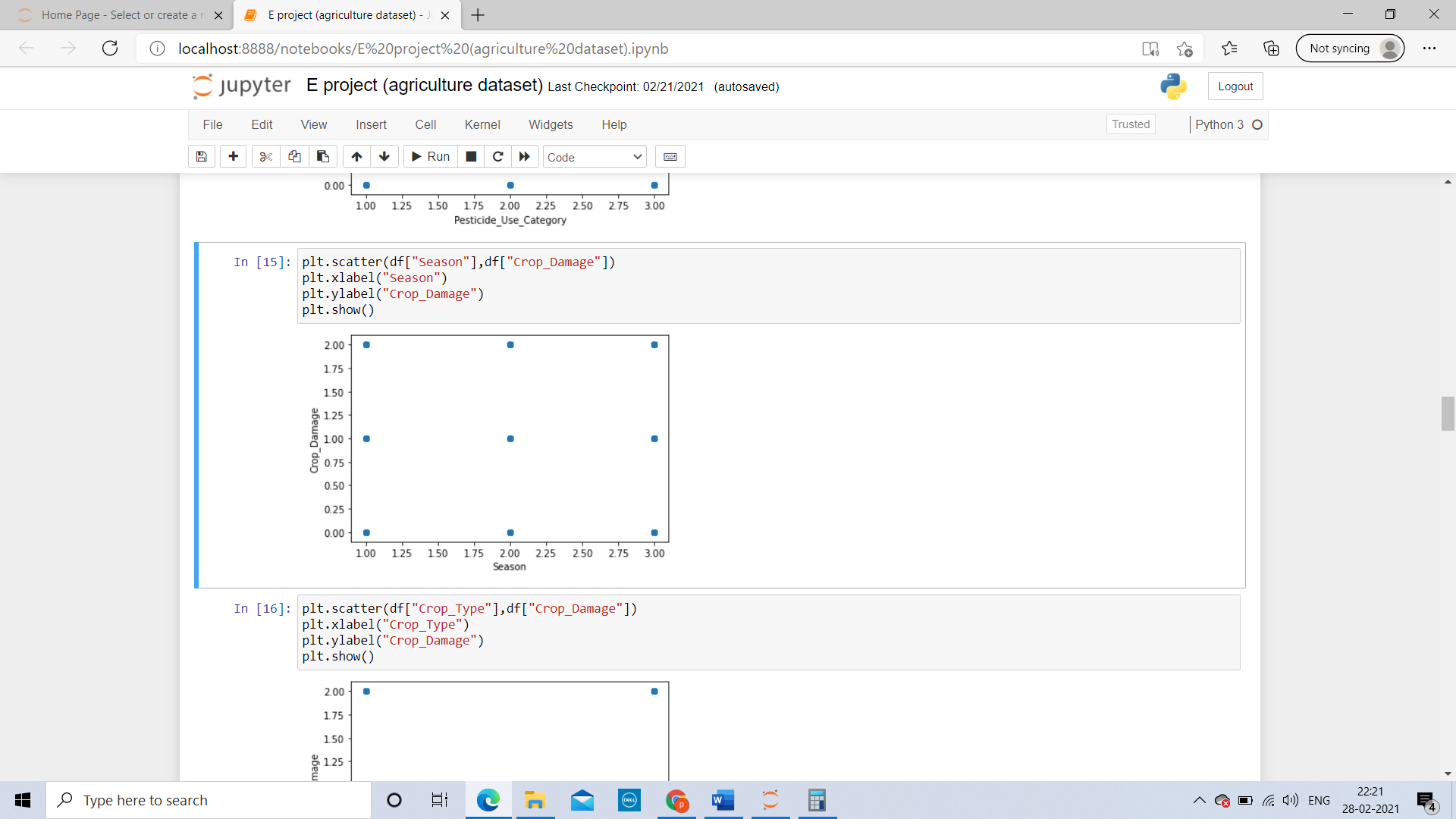


Pesticide category 1 has caused all three type of crop damage

Pesticide category 2 has caused all three type of crop damage

Pesticide category 3 has caused all three type of crop damage

**Between crop damage and season**

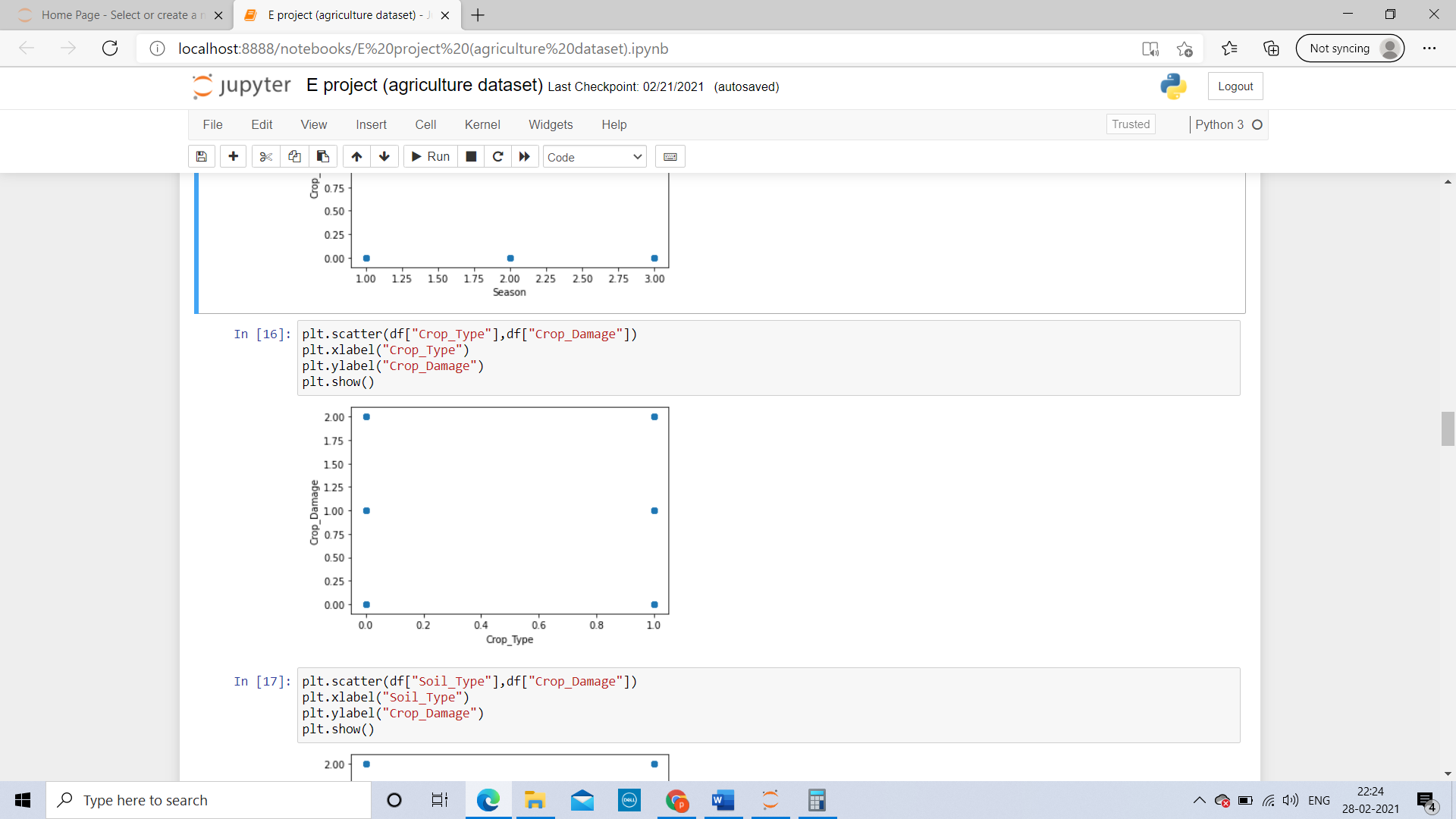


Season 1 has caused all three type of crop damage

Season 2 has caused all three type of crop damage

Season 3 has caused all three type of crop damage

**Between crop damage crop type**

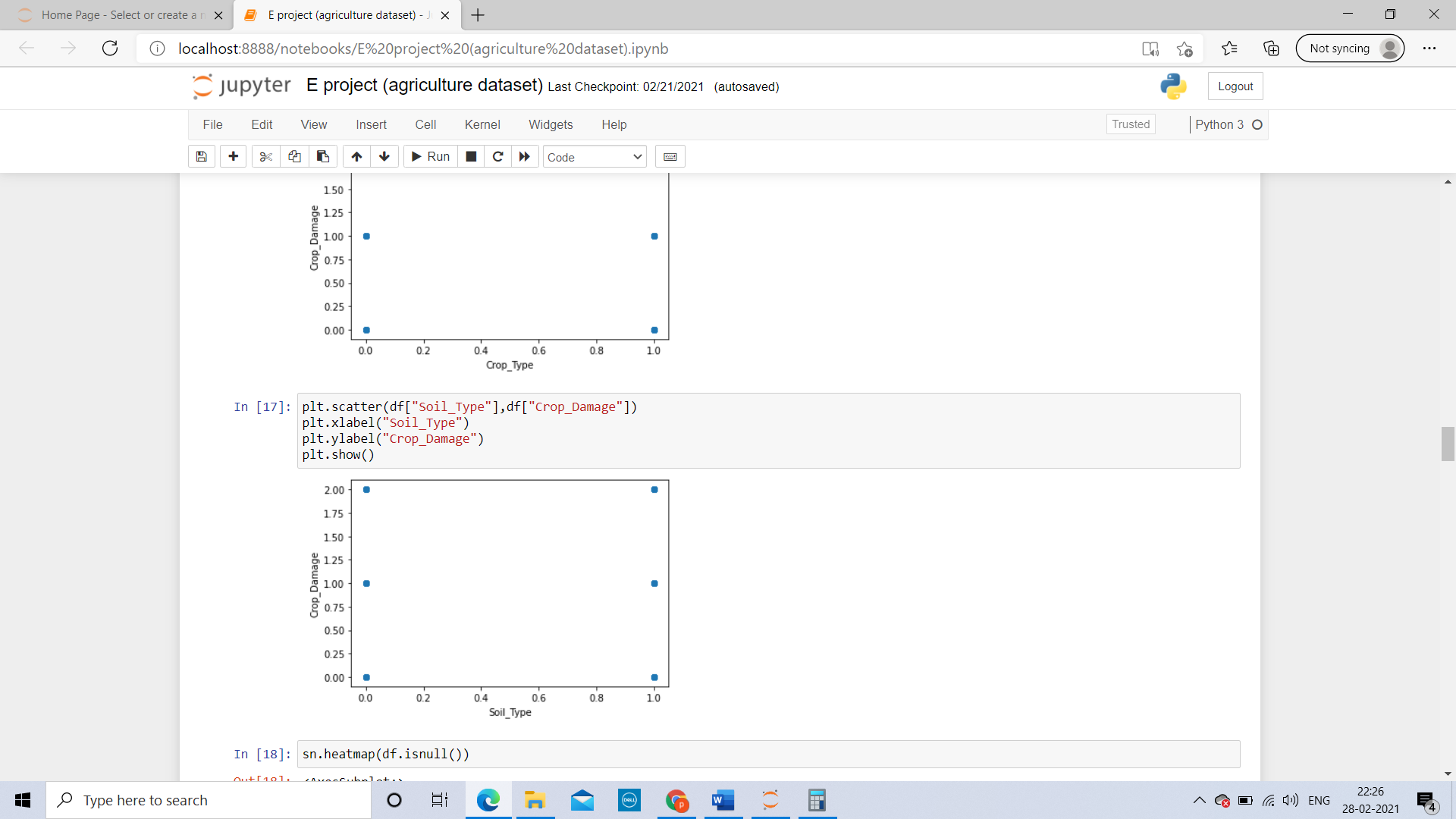


Crop type 1 has caused all three type of crop damage

Crop type 2 has caused all three type of crop damage

Crop type 3 has caused all three type of crop damage

**Between crop damage and soil type**

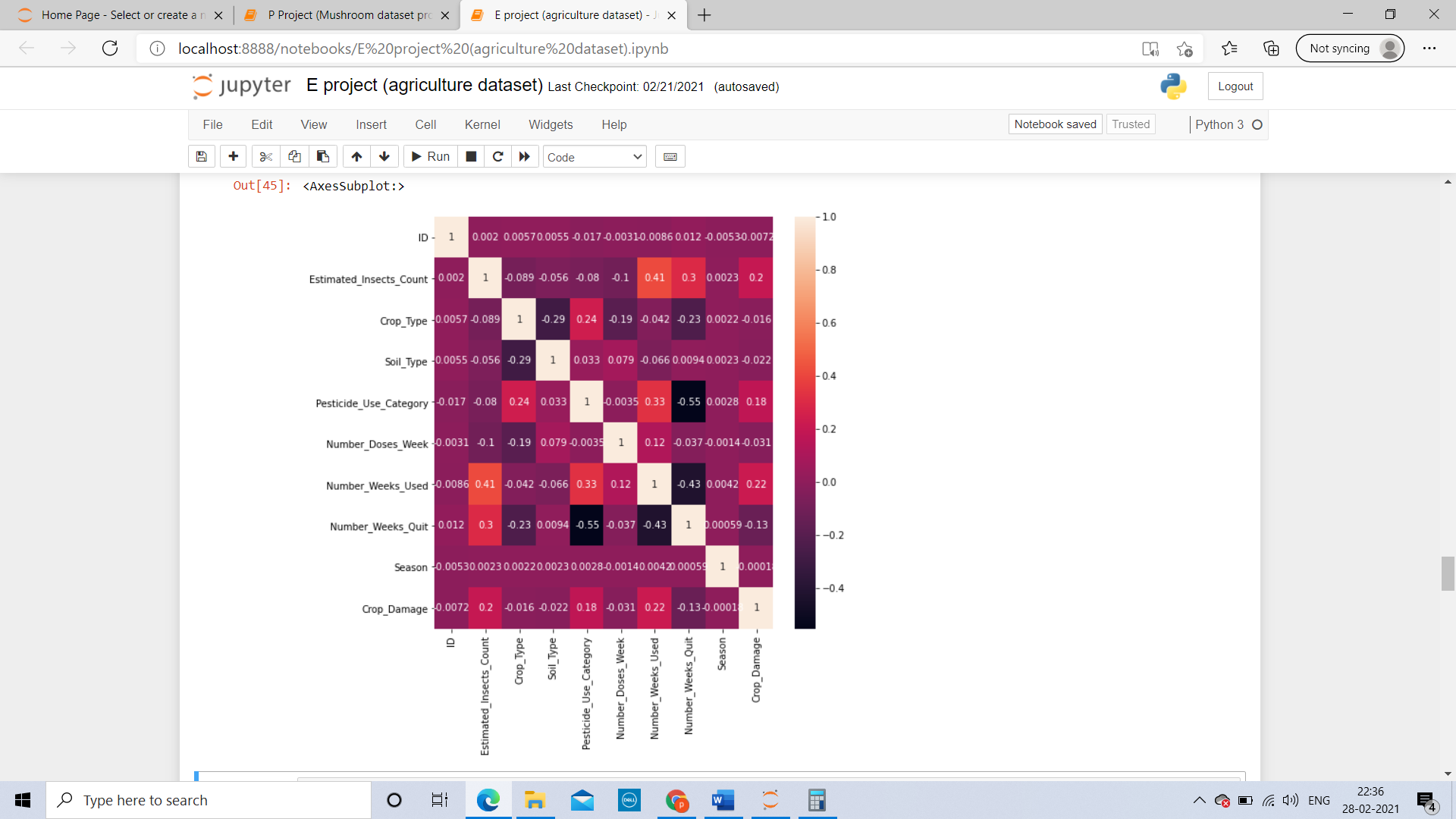


soil type 1 has caused all three type of crop damage

soil type 2 has caused all three type of crop damage

soil type 3 has caused all three type of crop damage

**MULTIVARIANT ANALYSIS:**



1) ID

2) Estimated\_Insects\_Count

3) Crop\_Type

4) Soil\_Type

5) Pesticide\_Use\_Category

6) Number\_Doses\_Week

7) Number\_Weeks\_Used

8) Number\_Weeks\_Quit

9) Season

10)Crop\_Damage

**Crop damage has positive correlation with with** : ID , Estimated\_Insects\_Count, Pesticide\_Use\_Category, Number\_Weeks\_Used and season

**Crop damage has negative correlation with**: Crop\_Type, Soil\_Type, Number\_Doses\_Week and Number\_Weeks\_Quit

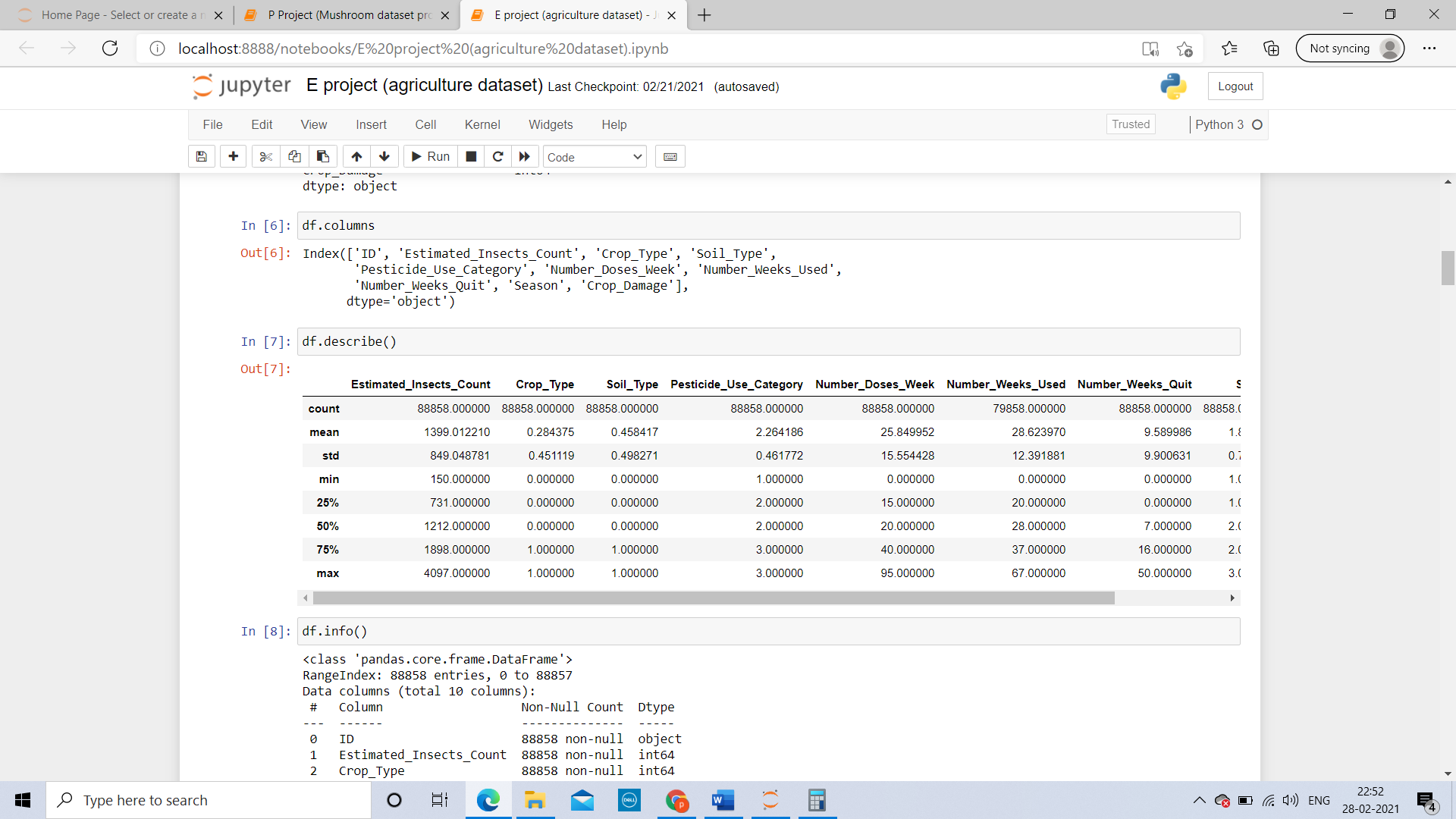
**Crop damage has good correlation with**: ID, Estimated\_Insects\_Count,

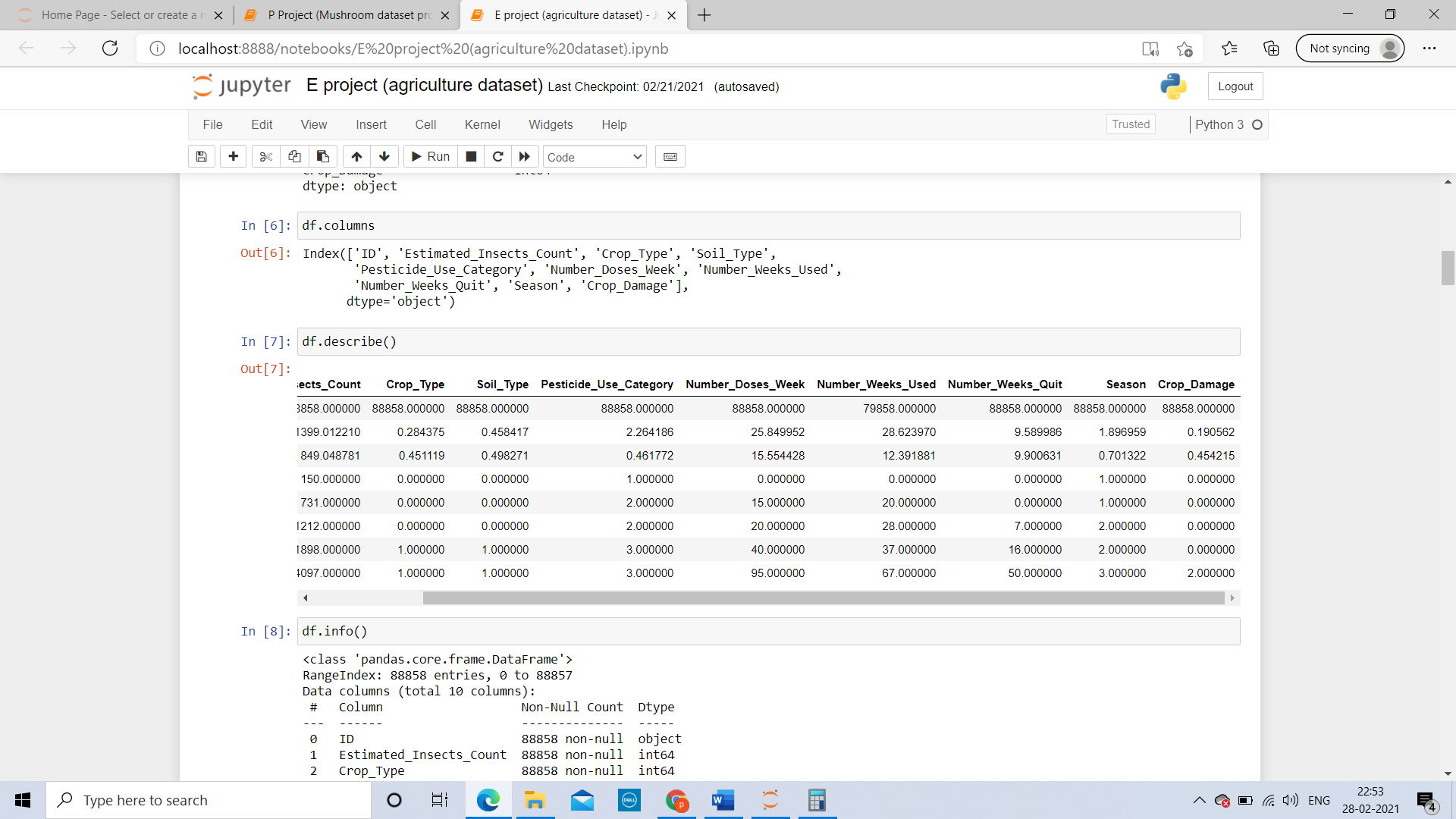
Pesticide\_Use\_Category, Number\_Weeks\_Used and Number\_Weeks\_Quit

**Crop damage has not so good correlation with**: Crop\_Type, Soil\_Type, Number\_Doses\_Week and season

**PRE-PROSSECING PIPELINE :**

**Removing outliers :**





**Key Observations**

1. the mean is greater than standard deviation in all columns.

2)there is large difference between the max and 75 percentiles in Estimated\_Insects\_Count, Number\_Doses\_Week, Number\_Weeks\_Used and Number\_Weeks\_Quit

3) the above points 1 and 2 suggest that there are extreme outliers present in these four columns.

**Removing outliers**: as analyzed above there are outliers present in the given data, we need to remove the outliers using Zscore method.

**Converting object values to int**: There are object values present which need to be converted into integers.



Using replace method we have convert object values into integer values .

**Removing skewness**: it is very important to remove skewness to avoid any kind of over fitting or under fitting, to remove the skewness we will use StandardScaler and fit method.

Now we will split the data into two parts

1. train data
2. test data

the split of our train data and test data will be 55% and 45%

**Target variable:** our target variable will be crop damage

**NOW FINDING THE BEST MODEL:**

**Linear Regression:**

we found Accuracy=100.0, cross validation score=100.0 & difference =0.0

**Random Forest Regressor:**

Accuracy=100.0, cross validation score =100.0 & difference =0.0

**Ada Boost Regressor:**

Accuracy=100.0, cross validation score =100.0 & difference =0.0

**SGD Regressor:**

Accuracy=99.99999890544973, cross validation score =99.99999885227527 & difference =5.317446039043716e-08

**RESULT:**

Linear Regression, Random Forest Regressor and Ada Boost Regressor are performing the best with same accuracy and cross validation score. I will choose Linear Regression.