**Week4\_課堂練習**

1. **Data preprocessing is an important issue for data mining. What are the tasks in data preprocessing?**

**Ans:**

**資料清理 (Data cleaning): Missing, Noisy, Inconsistent, Intentional, outliers, repetition.**

**資料整合 (Data integration): Schema integration, Entity identification problem, Different representations, different scales, Remove redundancies, Detect**

**inconsistencies (chi squared, covariance, correlation)**

**資料轉換 (Data transformation and data discretization): Normalization, Discretization**

**資料精簡 (Data reduction): Dimensionality reduction, Numerosity reduction, Data compression**

1. **Please use “Equal Width” and “Equal Depth” Binning Method to smooth the Dataset {17,42,35,29,12,21,6,29,22} into 3 boxes. Then, please use “Means”,” Median”, and “Boundaries” methods to eliminate the noisy data in final results, respectively.**

**Ans:**

**Sort {6, 12, 17, 21, 22, 29, 29, 35, 42}**

**Equal Width:**

**(42 - 6) / 3 = 12**

**(6-18): {6, 12, 17}**

**(19-30) : {21, 22, 29, 29}**

**(31-42) : {35, 42}**

**Means:**

**{11.67, 11.67, 11.67}**

**{25.25, 25.25, 25.25, 25.25}**

**{38.5, 38.5}**

**Median:**

**{12, 12, 12}**

**{25.5, 25.5, 25.5, 25.5, 25.5}**

**{38.5, 38.5}**

**Boundaries:**

**{6, 17, 17}**

**{21, 21, 29, 29}**

**{35, 42}**

**Equal Depth:**

**{6, 12, 17}**

**{21, 22, 29}**

**{29, 35, 42}**

**Means:**

**{11.67, 11.67, 11.67}**

**{24, 24, 24}**

**{35.33, 35.33, 35.33}**

**Median:**

**{12, 12, 12}**

**{22, 22, 22}**

**{35, 35, 35}**

**Boundaries:**

**{6, 17, 17}**

**{21, 21, 29}**

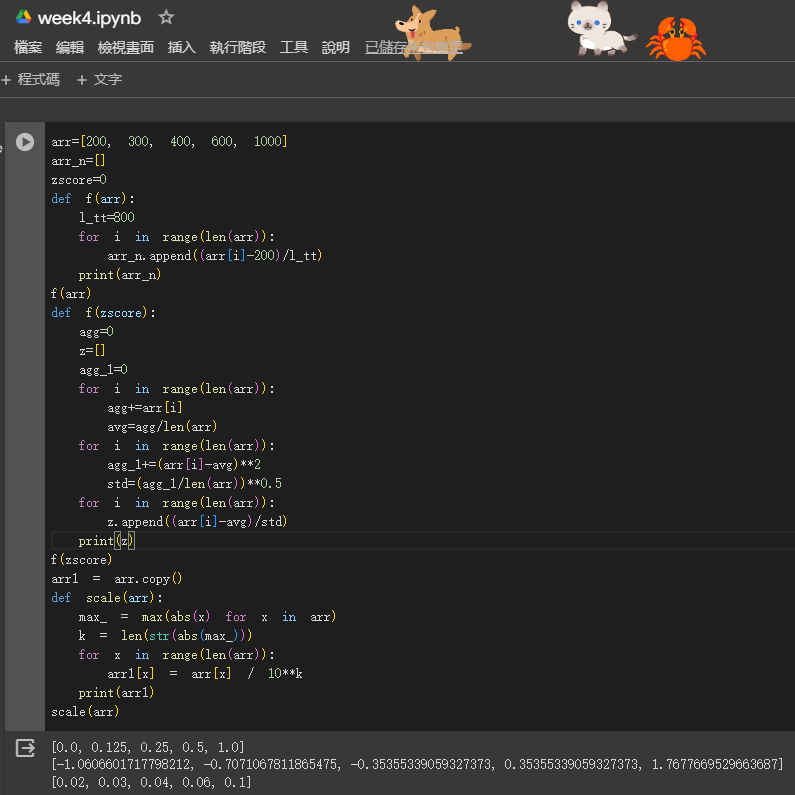
**{29, 29, 42}**

**--------------------------------------------------------------------------------------------------------------**

**Week4\_課堂練習**

**1. Use the two methods below to normalize the following group of data:**

**200, 300, 400, 600, 1000**

****

1. **min-max normalization by setting min = 0 and max = 1**

**Ans:**

**200 => (200-200) / (1000-200) = 0**

**300 => (300-200) / (1000-200) = 0.125**

**400 => (400-200) / (1000-200) = 0.25**

**600 => (600-200) / (1000-200) = 0.5**

**1000 => (1000-200) / (1000-200) = 1**

1. **z-score normalization**

**Ans:**

**Means: (200+300+400+600+1000) / 5 =500**

**SD=282.84271**

**200 => (200-500) / 282.84271 = -1.0607**

**300 => (300-500) / 282.84271 = -0.7071**

**400 => (400-500) / 282.84271 = -0.3536**

**600 => (600-500) / 282.84271 = 0.3536**

**1000 => (1000-500) / 282.84271 =1.7678**

1. **Decimal scaling**

**Ans:**

**Max = 1000**

**104=10000**

**200/10000=0.02**

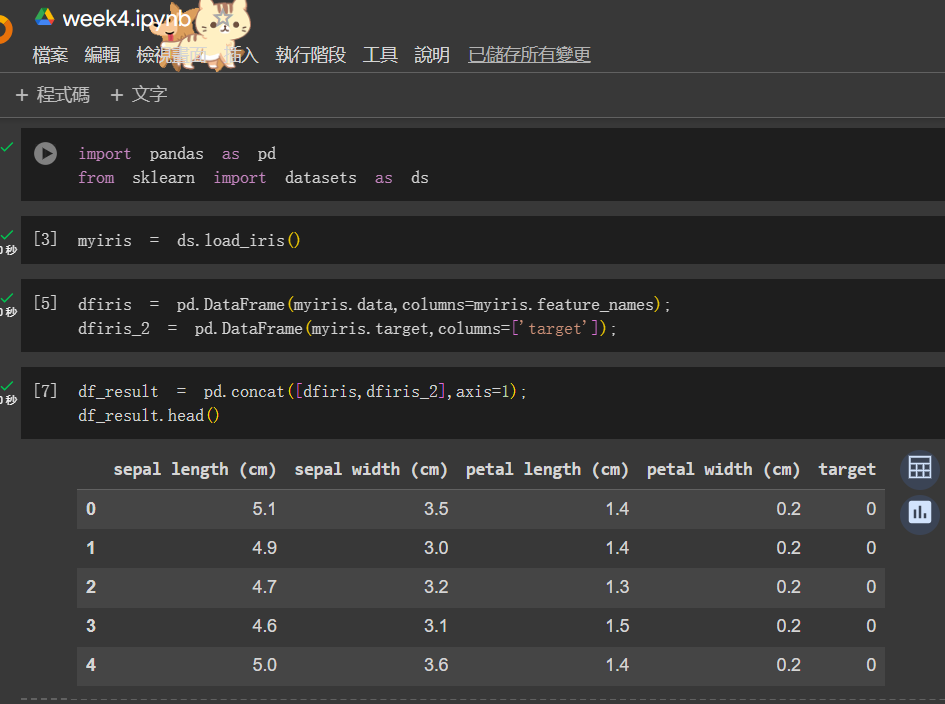
**300/10000=0.03**

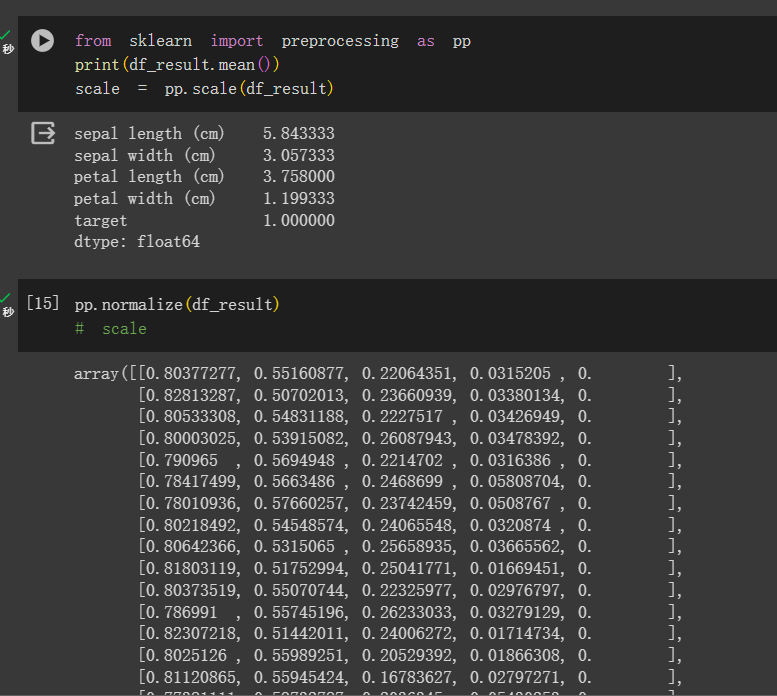
**400/10000=0.04**

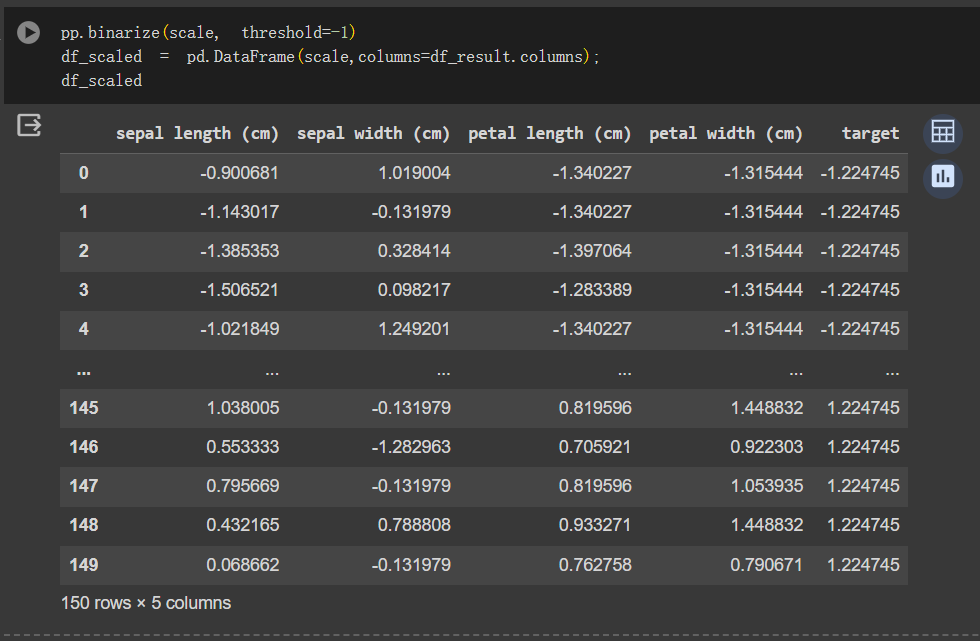
**600/10000=0.06**

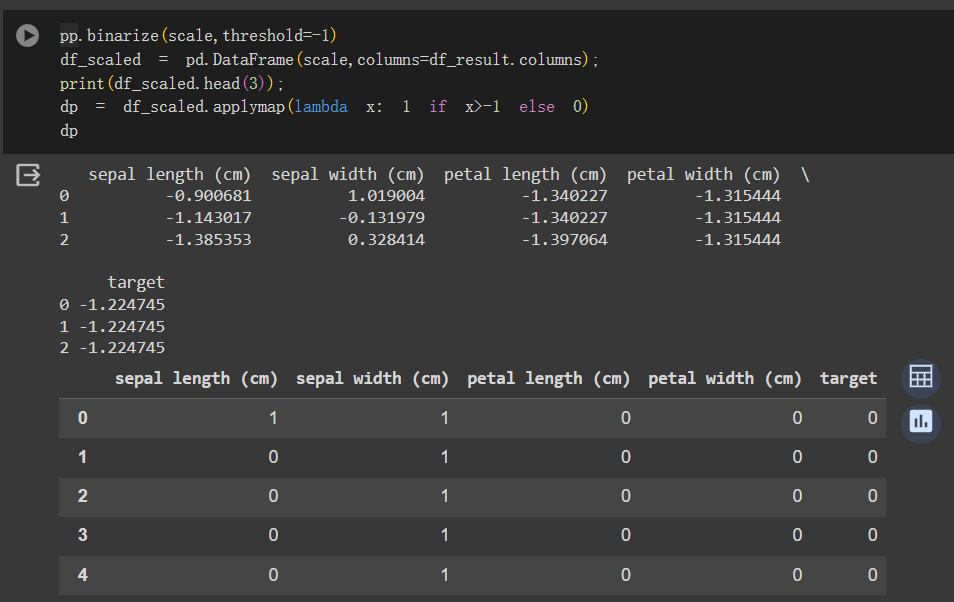
**1000/10000=0.1**

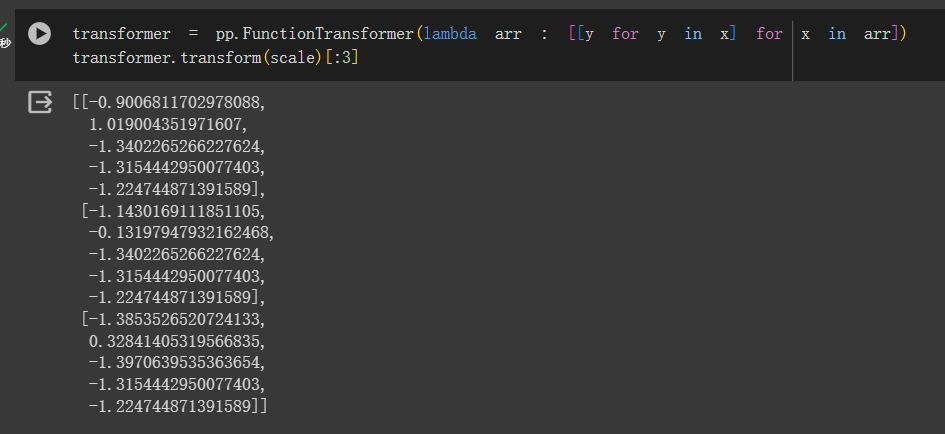
**-----------------------------------------------------------------------------------------------------------------week4 實作內容 影片中的程式碼**

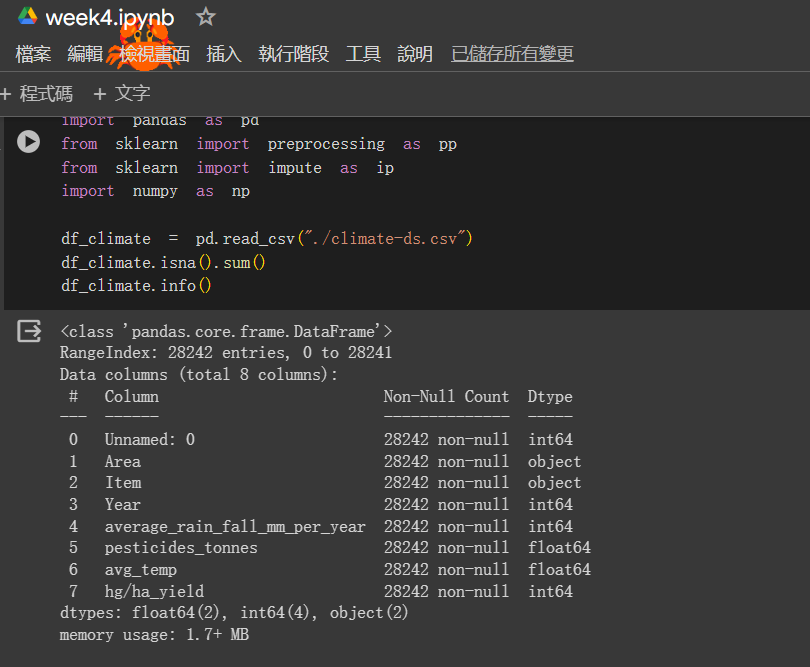
****

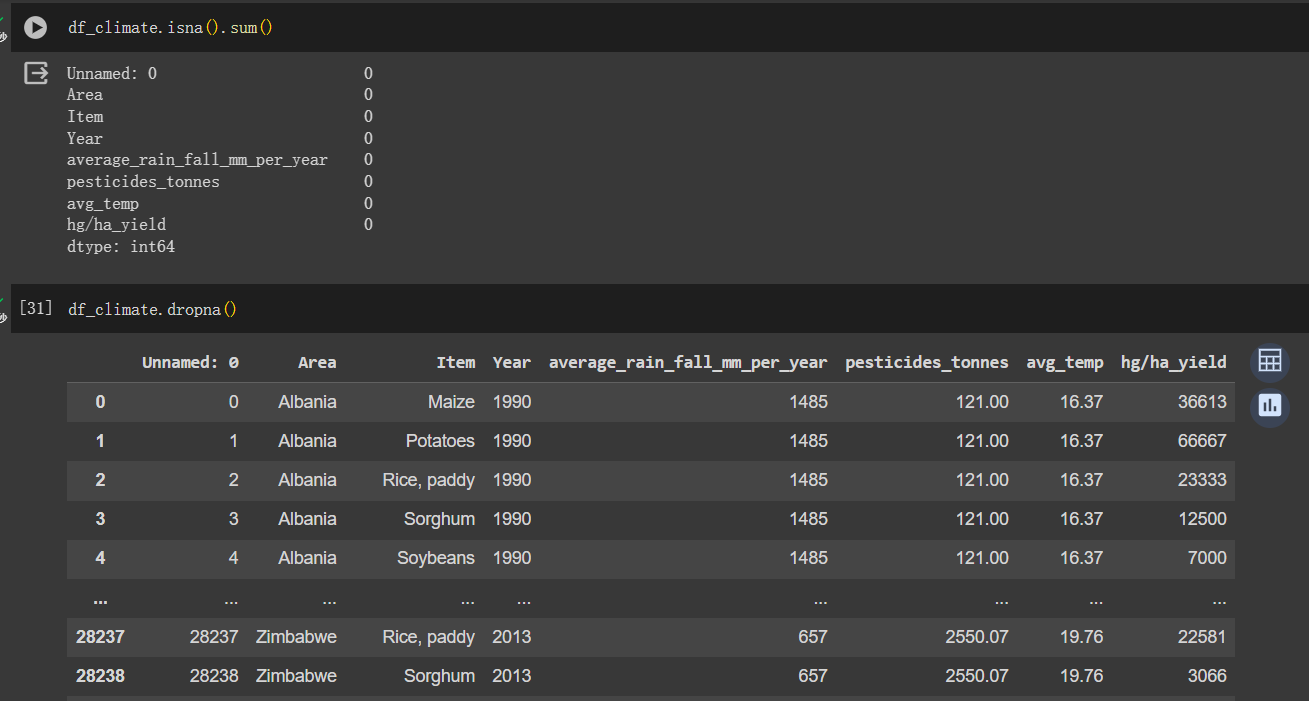
****

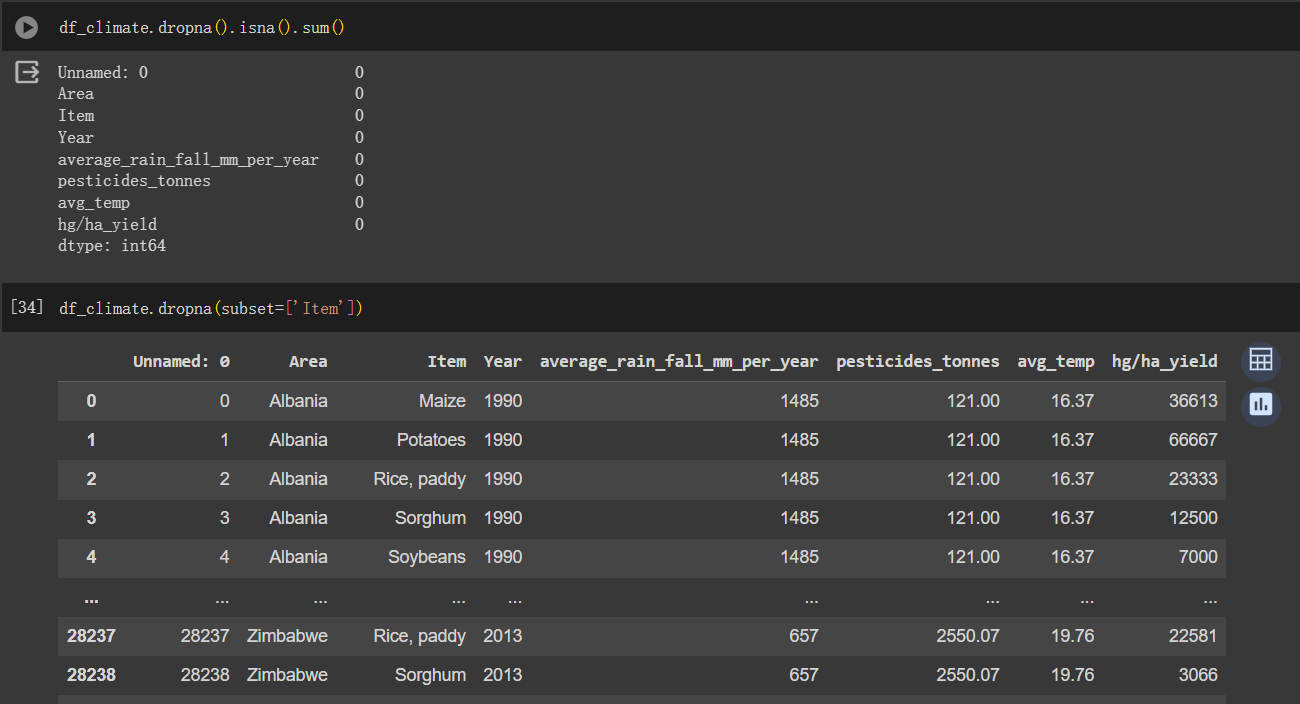
****

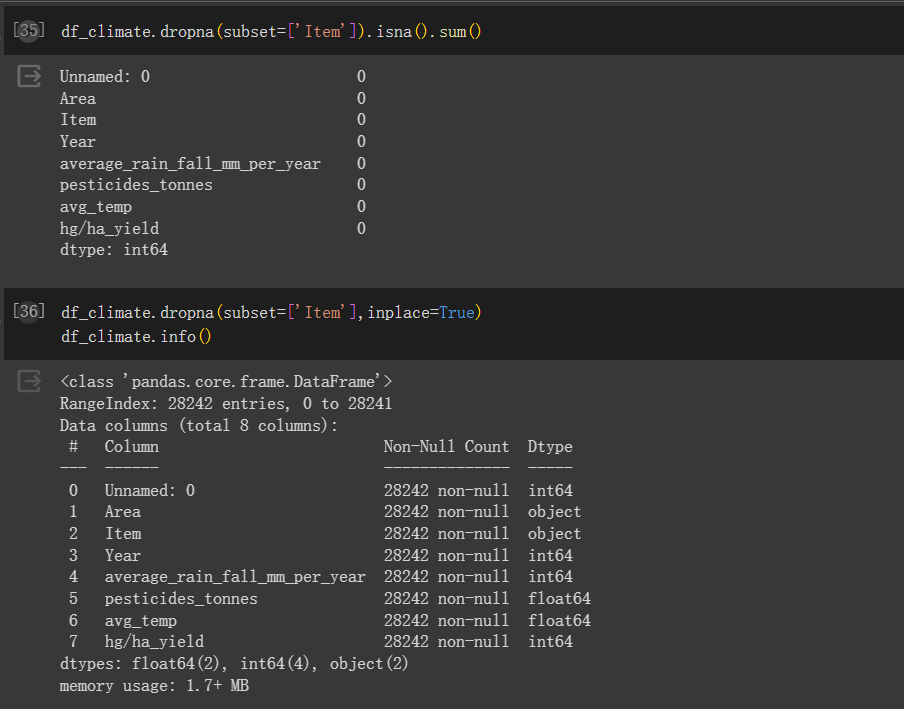
****

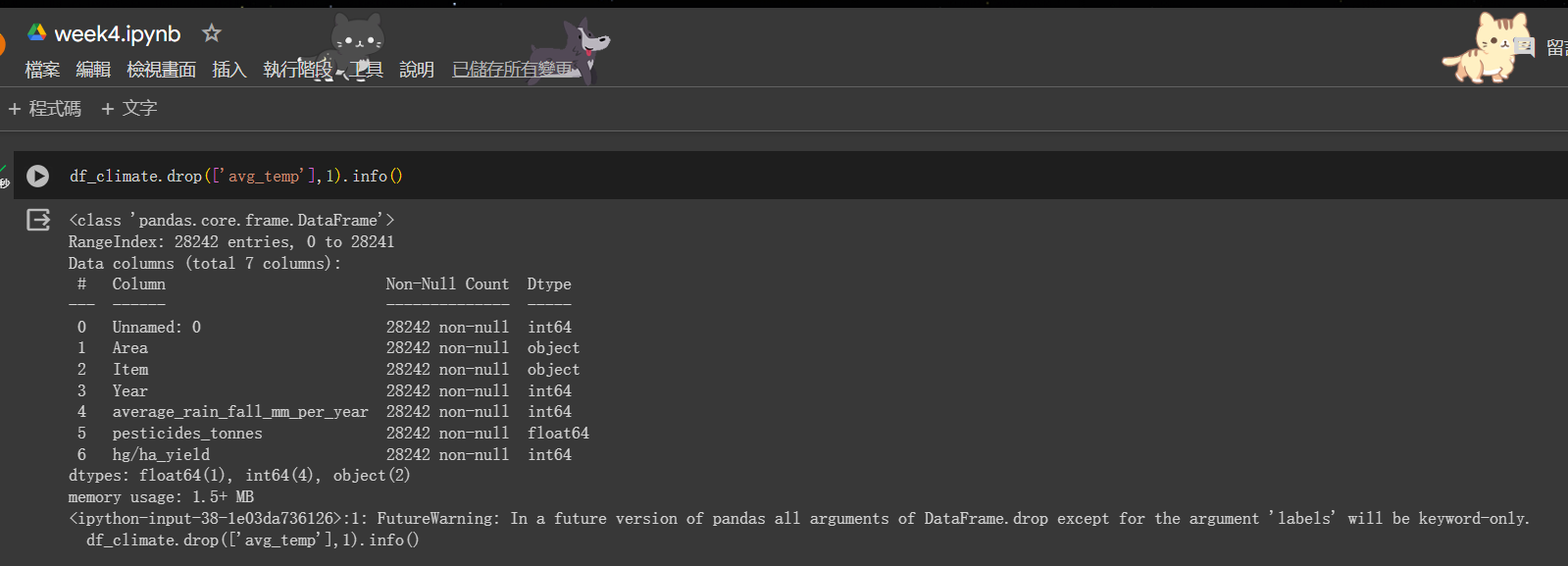
****

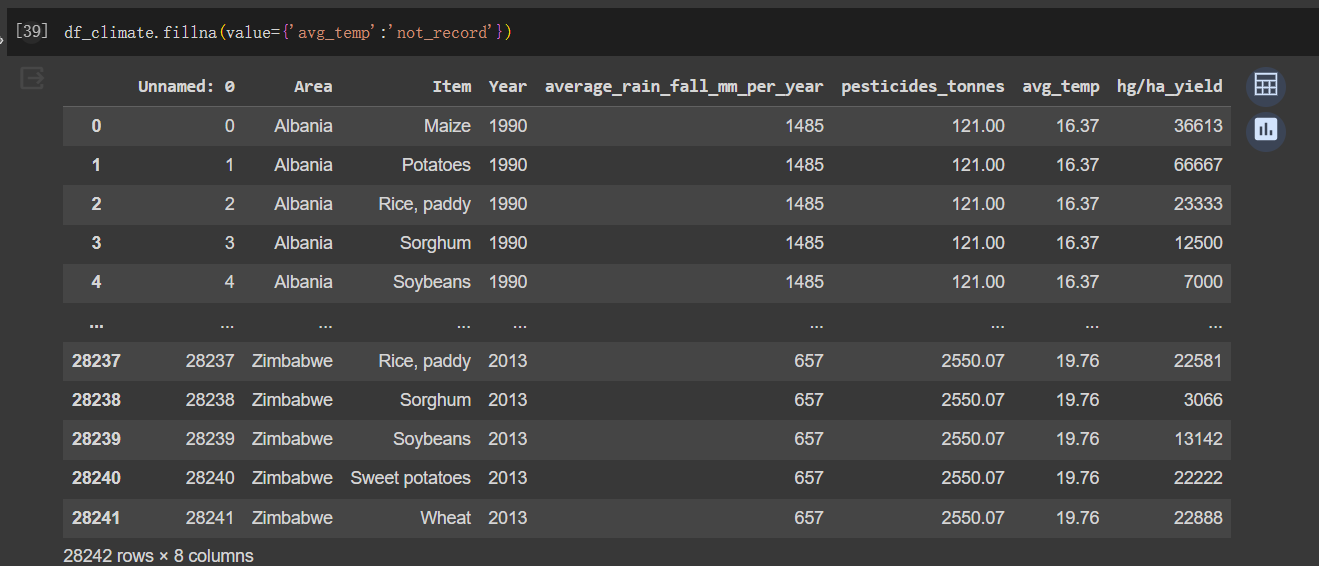
****

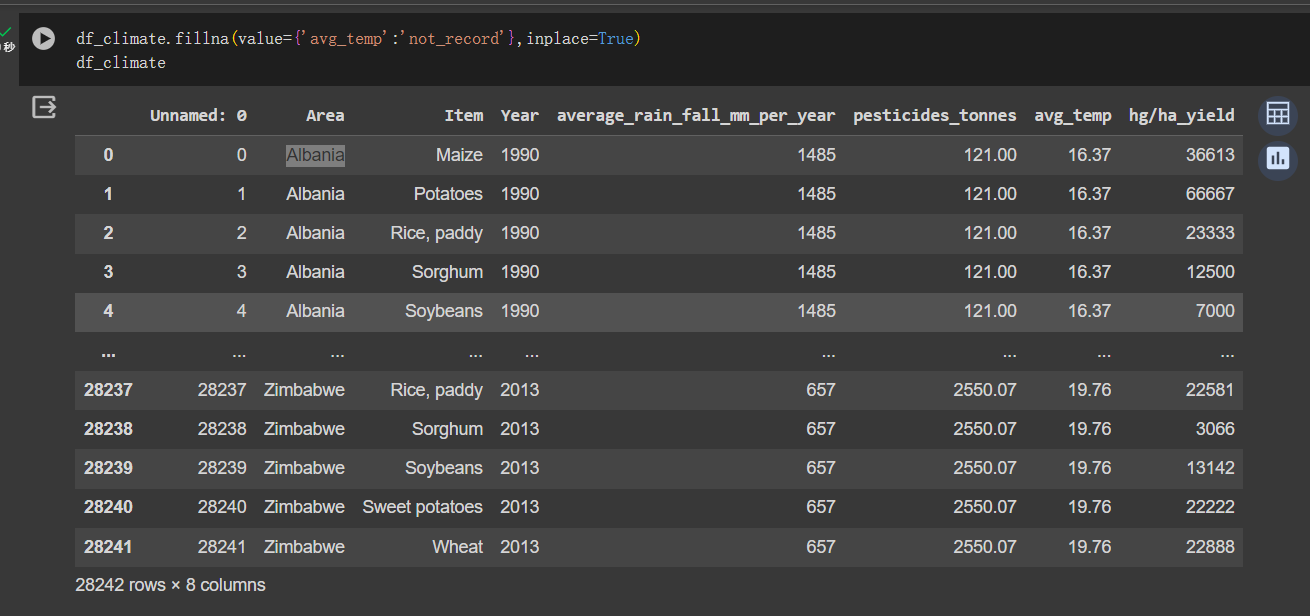
****

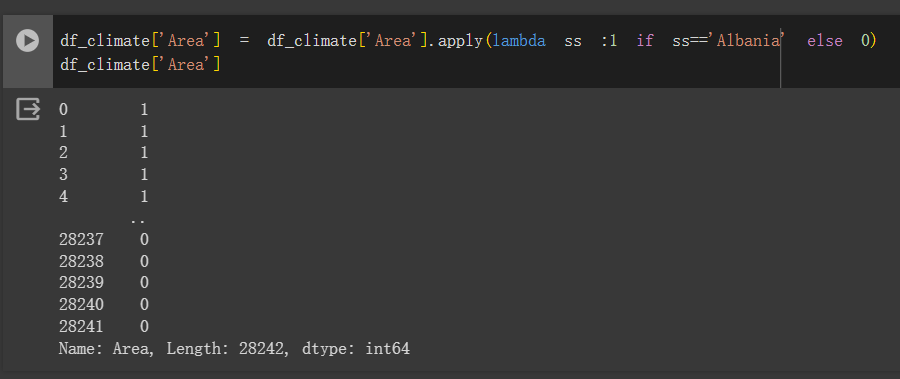
****

****

****

****

****

****