Iris

2019年6月3日　　下午 04:31

1. 資料集介紹

IRIS資料集也稱作鳶尾花資料集，整個資料集共有150條資料，分為三類，每類50條資料，每一條資料都有四個屬性：花萼長度，花萼寬度，花瓣長度，花瓣寬度，標籤資料共有三種，分別是Setosa，Versicolour，Virginica。一般使用前面的四種屬性資料來預測樣本屬於哪種鳶尾花

來自 <https://www.itread01.com/content/1546315322.html>

2. 建置

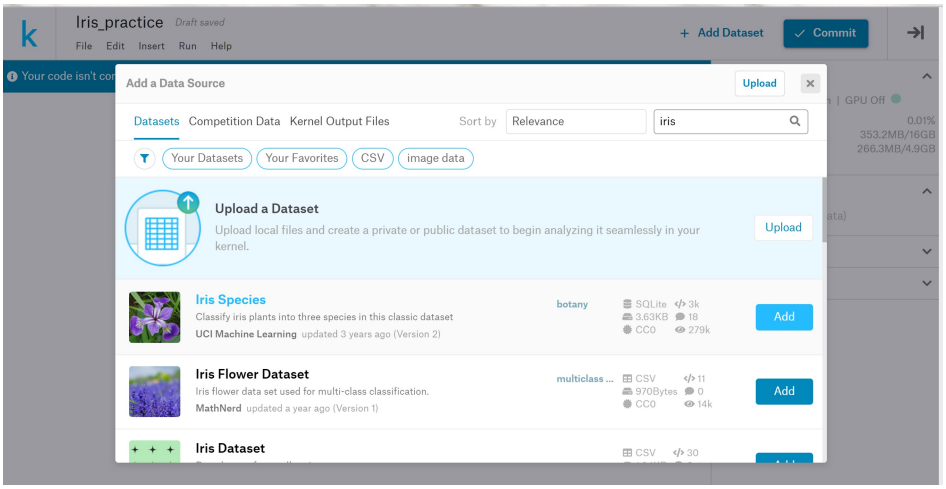
a. Kernel建置

i. Kaggle→Kernel→New Kernel→notebook

b. 加入資料集

i. Add Dataset→搜Irirs

ii.



3. Set up

a. 把預設程式碼刪掉後

b. 匯入所需的套件

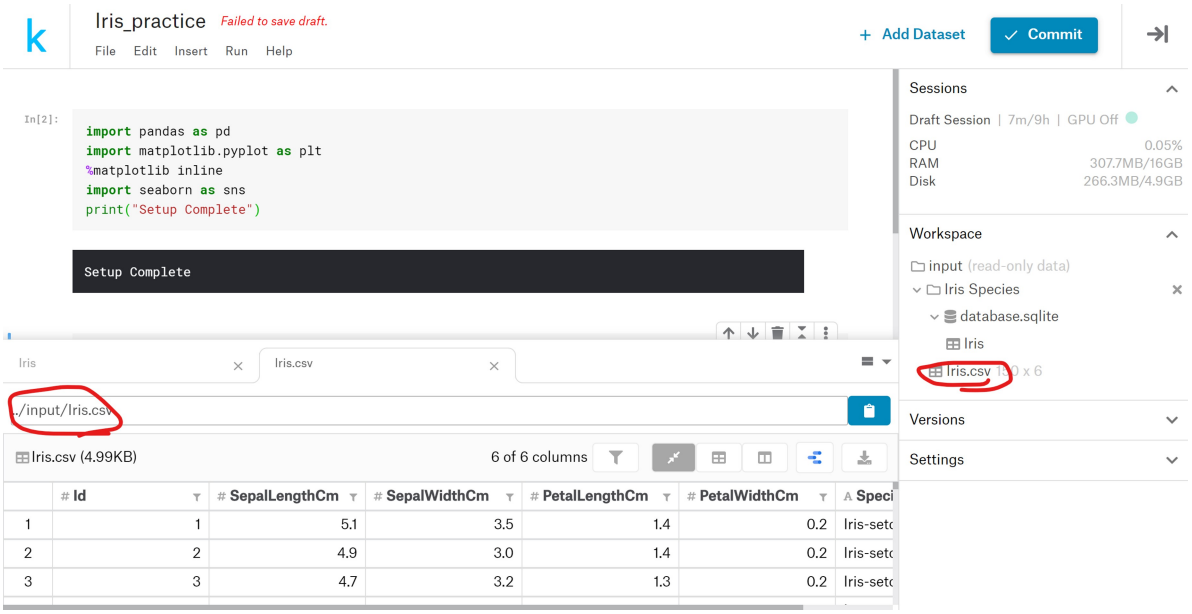
```
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
print("Setup Complete")
```

```
Setup Complete
```

4. Specify the filepath

a. 先找Iris資料集的檔案位置

b.



c. 將Iris.csv的路徑複製下來

5. Load the data

```
iris_filepath="../input/Iris.csv"
iris_data=pd.read_csv(iris_filepath)
iris_data.head()
```

a.

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa

6. 視覺化

用各種圖示和參數來找資料關聯性

a. 這是用lineplot去作呈現，參數選擇 species和sepalLengthCm，想看看兩者之間關係，發現種類是setosa的花萼長度最短

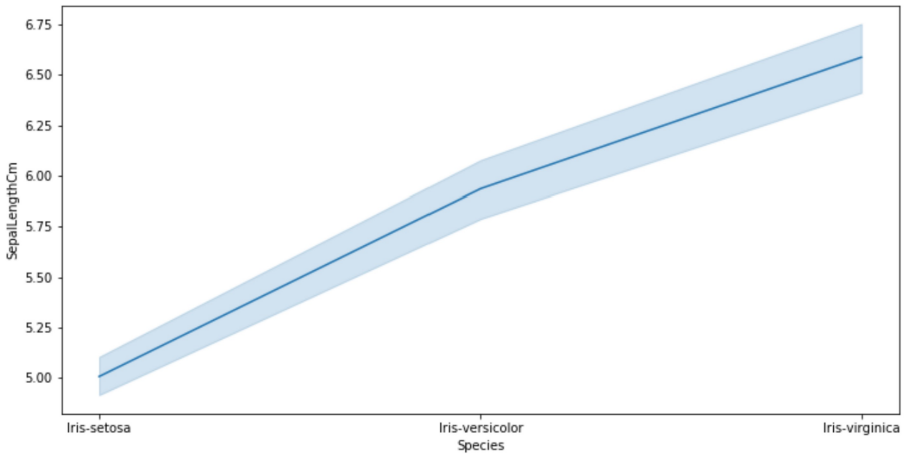
```
plt.figure(figsize=(12,6))
sns.lineplot(x='Species',y="SepalLengthCm",data=iris_data)
```

```
/opt/conda/lib/python3.6/site-packages/scipy/stats/stats.py:1713: FutureWarning: Using a non-tupl
q)]` instead of `arr[seq]`. In the future this will be interpreted as an array index, `arr[np.arr
  return np.add.reduce(sorted[indexer] * weights, axis=axis) / sumval
```

)ut[11]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fd192cbe630>
```

i.



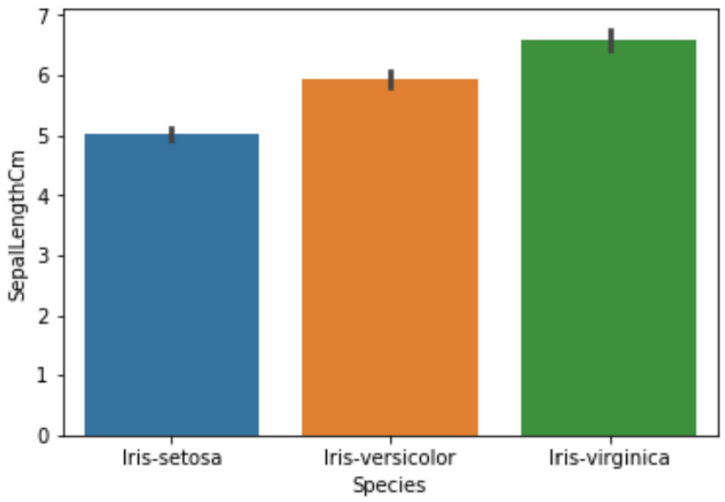
b. 這是用barplot看兩者關係，比起lineplot更明顯去辨別差異

```
sns.barplot(x=iris_data['Species'],y=iris_data['SepalLengthCm'])
```

```
/opt/conda/lib/python3.6/site-packages/scipy/stats/stats.py:1713: FutureWarn
dexing is deprecated; use `arr[tuple(seq)]` instead of `arr[seq]`. In the fu
p.array(seq)]`, which will result either in an error or a different result.
  return np.add.reduce(sorted[indexer] * weights, axis=axis) / sumval
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fd19298b080>
```

i.

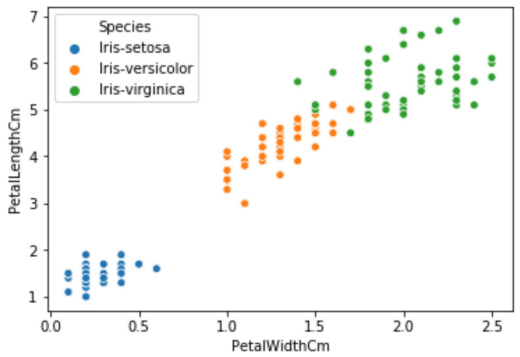


c. Scatter

```
sns.scatterplot(x=iris_data['PetalWidthCm'],y=iris_data['PetalLengthCm'],hue=iris_data['Species'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fd1927e58d0>
```

i.



ii. 這是觀察花瓣寬度和長度的相關性，發現呈正相關

iii. 用hue參數以Species為分類去上色

iv. 可以看到不管是哪種種類它的花瓣寬度長度都呈正相關，其中又以setosa的花瓣長寬特別短小

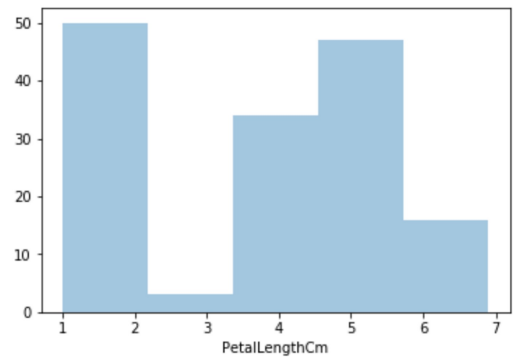
d. Distplot

```
sns.distplot(a=iris_data['PetalLengthCm'], kde=False)
```

```
/opt/conda/lib/python3.6/site-packages/scipy/stats/stats.py:1
dexing is deprecated; use `arr[tuple(seq)]` instead of `arr[s
p.array(seq)]`, which will result either in an error or a dif
  return np.add.reduce(sorted[indexer] * weights, axis=axis)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fd19275ca58>
```

i.



ii. 可發現所有的花瓣長度中，長度在1~2間的數量最多