1. Unsupervisedیک روش یادگیری ماشین است که به مدل اجازه می ده تا روی الگوها و اطلاعاتی که قبلاً پیدا نشده بودند کار کند، این کار با داده های بدون label است.

2. یادگیری supervised روش یادگیری ماشین است که با استفاده از داده های آموزشی دارای برچسب استنباط می کند که داده تست یا داده ی جدید شامل کدام مجموعه های آموزش است.

3. ترکیب دو تای بالایی است یعنی مقداری داده ی label خورده و مقدار بیشتری داده ی label نخورده. برای classify ام استفاده میشه

4. Outlier به داده هایی میگویند که خارج از انتظار ما از پراکندگی داده ها(بیشتر داده ها)، هست.

5. تعداد متغیرهای ورودی یا ویژگی های یک مجموعه داده به عنوان بعد dimension تعریف میشود.

6. Training Dataset: نمونه داده ها تا مدل با آن fit شود و مدل آموزش داده شود، مدل داده ها را میبیند و یاد میگیرد.

Validation Dataset: نمونه داده ای از آموزش نگه داشته میشود به طور جداگانه برای استفاده در ارزیابی از یک مدل که با مجموعه داده های آموزش fit شده است، همزمان با تنظیم hyperparameterهای مدل.

Test Dataset: نمونه ای از داده های خارجی برای ارائه ارزیابی از مدل نهایی fit شده مجموعه داده های آموزشی استفاده شده است.

7. Data warehousing فرآیند ساخت و داده با ادغام داده های چندین منبع تا توانایی گزارش تحلیلی ، query و تصمیم گیرید داشته باشد.

8. داده های دنیای واقعی اغلب مقدار زیادی داده ی از دست رفته دارند. علت از دست رفتن مقادیر می تواند خرابی داده ها یا ضبط نکردن داده ها باشد. مدیریت داده های از دست رفته در طی پیش پردازش مجموعه داده بسیار مهم است.

9. Independent Variable: متغیر مستقل به عنوان متغیری که در یک تست تغییر یا کنترل می شود، تعریف می شود. این نشان دهنده علت یا دلیل یک نتیجه است. متغیرهای مستقل متغیرهایی هستند که تست کننده برای آزمایش متغیر وابسته خود تغییر می دهد.

2) 1. Missing Values Ratio: بعید است ستون های داده با مقادیر زیادی از دست رفته اطلاعات مفیدی را حمل کنند. بنابراین ستون های داده با تعداد مقادیر از دست رفته بیشتر از یک آستانه داده شده را می توان حذف کرد. مثال:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
train=pd.read_csv("Train_UWu5bXk.csv")
# checking the percentage of missing values in each variable
train.isnull().sum()/len(train)*100
```

```
Item_Identifier              0.000000
Item_Weight                 17.165317
Item_Fat_Content             0.000000
Item_Visibility              0.000000
Item_Type                    0.000000
Item_MRP                     0.000000
Outlet_Identifier            0.000000
Outlet_Establishment_Year    0.000000
Outlet_Size                 28.276428
Outlet_Location_Type         0.000000
Outlet_Type                  0.000000
Item_Outlet_Sales            0.000000
dtype: float64
```

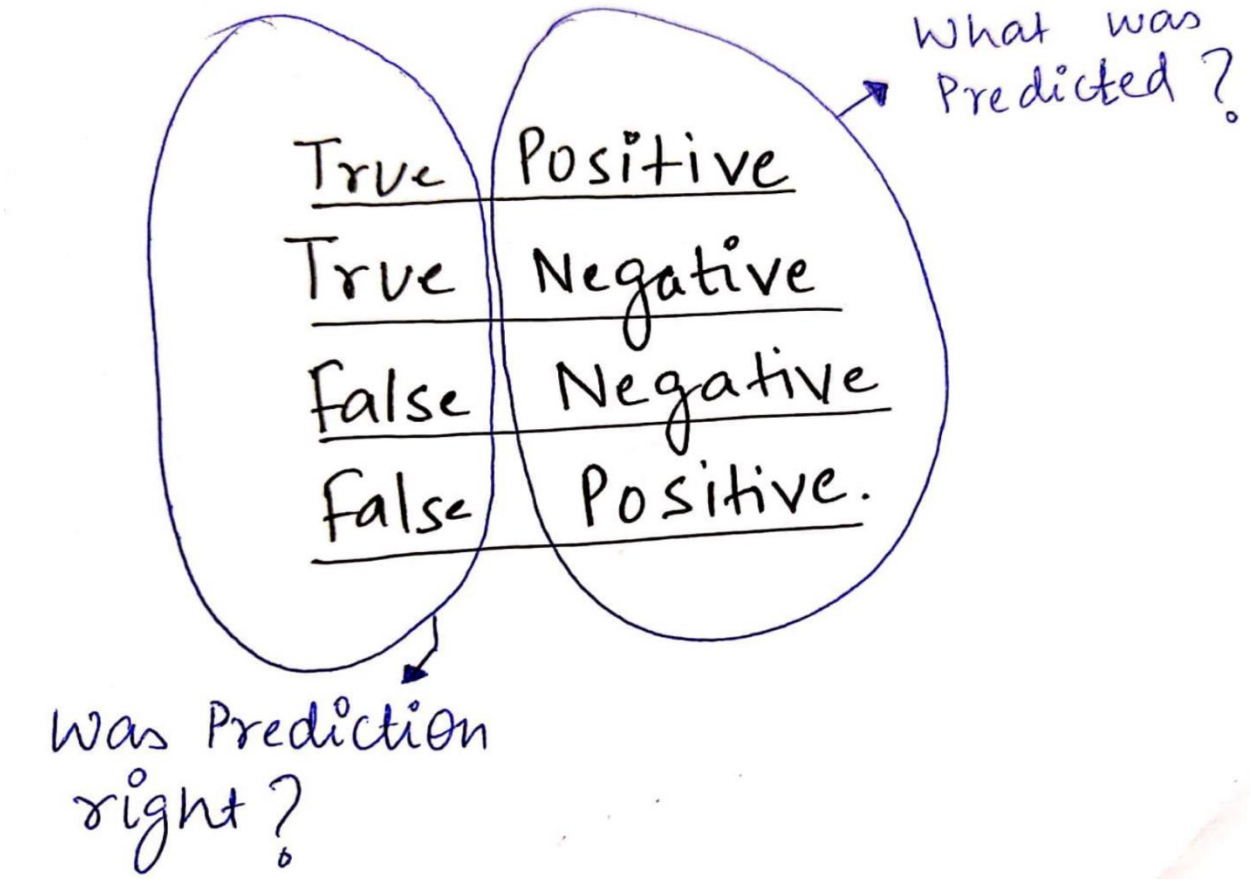
```
# saving missing values in a variable

a = train.isnull().sum()/len(train)*100

# saving column names in a variable

variables = train.columns

variable = [ ]

for i in range(0,12):

    if a[i]<=20:   #setting the threshold as 20%

            variable.append(variables[i])
```

2. Low Variance Filter: به طور مشابه روش قبلی ، ستون های داده با کمی تغییر در داده ها ، اطلاعات با ارزش کمتری را حمل می کنند. بنابراین تمام ستون های داده با واریانس پایین تر از یک آستانه داده شده حذف می شوند. Variance به دامنه بستگی دارد. بنابراین قبل از استفاده از این روش normalization لازم است.

Feature Selection: فرایندی است که در آن به طور خودکار یا دستی آن ویژگی هایی را انتخاب می شود که بیشترین تأثیر را در متغیر پیش بینی یا خروجی مورد نظر دارند. داشتن ویژگی های بی ربط در داده های می تواند از دقت مدل ها بکاهد و مدل را بر اساس ویژگی های بی ربط بیاموزد.

Feature extraction: استخراج ویژگی فرآیندی در کاهش ابعاد است که طی آن یک مجموعه اولیه از داده های خام به گروه های قابل کنترل بیشتری برای پردازش کاهش می یابد. ویژگی این مجموعه داده های بزرگ تعداد زیادی متغیر است که برای پردازش به منابع محاسباتی زیادی نیاز دارند.

.3



دقت: نسبت پیش بینی های مثبت صحیح به کل موارد مثبت پیش بینی شده.

$$P = \frac{TP}{TP + FP}$$

فراخوانی: نسبت پیش بینی های مثبت صحیح به کل نمونه های مثبت.

$$R = \frac{TP}{TP + FN}$$

F-score: دقت و فراخوانی را می توان در یک معیار واحد به نام f-score ترکیب کرد. اگر تأکیدی زیادی بر روی اهمیت بیشتر دقت یا فراخوانی نباشد ،ترکیب مناسبی ست. فرمول زیر ترکیبی از هر دو معیار است.

$$\mathrm{F1} = 2 \cdot \frac{\mathrm{Precision} \cdot Recall}{\mathrm{Precision} + \mathrm{Recall}}$$

4. همبستگی صفر نشان می دهد که دو متغیر رابطه ی خطی ندارند و معنای عدم وجود رابطه نیست. همبستگی صفر اغلب با استفاده از مخفف r = 0 نشان داده می شود.

درباره ی مستقل بودن آن ها نمیتوان به قطعیت گفت چون مستقل بودن یعنی هیچگونه رابطه ای با هم ندارند.

.5

پاکسازی داده ها: پر کردن مقادیر از دست رفته ، مرتب کردن داده های noisy ، شناسایی یا حذف کردن داده های noisy و outlier.

ادغام داده ها: ادغام چندین پایگاه داده یا file تا به کاربر ها یک view از تمام این داده ها نشان بدهد.

تبدیل داده ها: در data warehouse ، ما از تبدیل داده برای تبدیل داده ها از قالب داده منبع به داده های مقصد استفاده می کنیم مانند:

Data mapping: map کردن داده های منبع به داده های دیگر با تابع.

Smoothing: حذف داده هایnoisy

Normalization: scale کردن داده ها به یک range خاص.

Aggregation: ادغام داده ها.

Generalization: داده های سطح پایین مثل سن با مفاهیم سطح بالاتر مثل جوانی یا تجربه جایگزین می شوند.

.6

Confidence = 60% , Support_Count = 6 × 33% = 2

نان	٤
الویه	۲
شیر	۲
کره	۳
مربا	٤

Sup_Count

$k = 1$

Sup_Count ≥ 2 ←

نان، الویه	۲
نان، شیر	۱
نان، کره	۲
نان، مربا	۲
الویه، شیر	۱
الویه، کره	۰
الویه، مربا	۰
شیر، کره	۰
شیر، مربا	۱
کره، مربا	۳

Sup_Count

نان، الویه	۲
نان، کره	۲
نان، مربا	۲
کره، مربا	۳

$k = 3$

Subsets

not frequent

~~نان، الویه، کره~~ ~~نان، الویه، مربا~~ نان، کره، مربا

$k = 4$

نان، کره، مربا	2

PaPCG

تبدیل اعضای

[نان، کره] → [مربا] = 2/2 = 100% ①

[نان، مربا] → [کره] = 2/2 = 100% ①

[کره، مربا] → [نان] = 2/3 = 66% ②

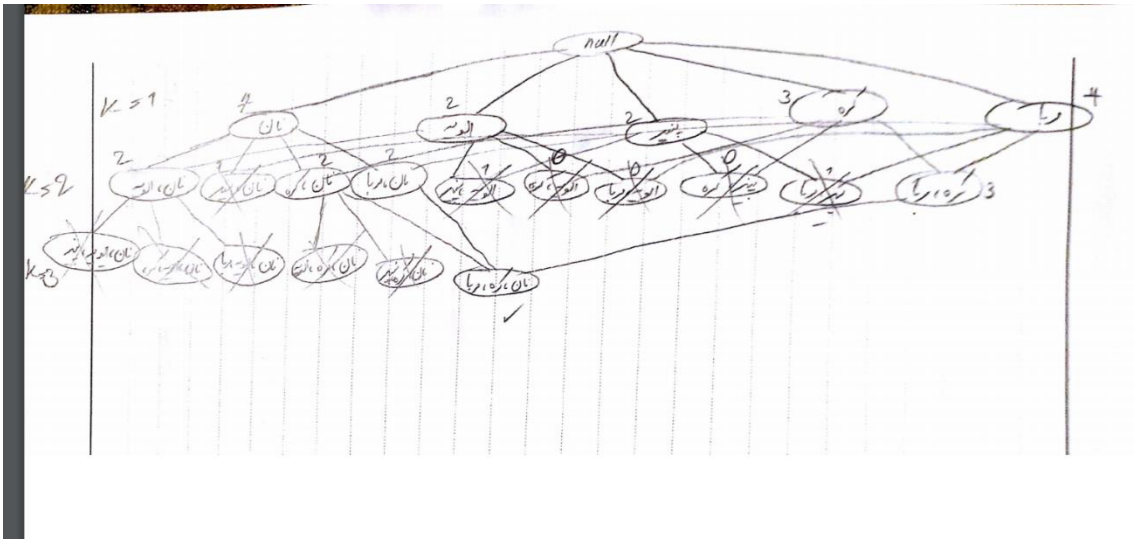
[کره] → [نان، مربا] = 2/2 = 100% ①

[مربا] → [نان، کره] = 2/2 = 100% ①

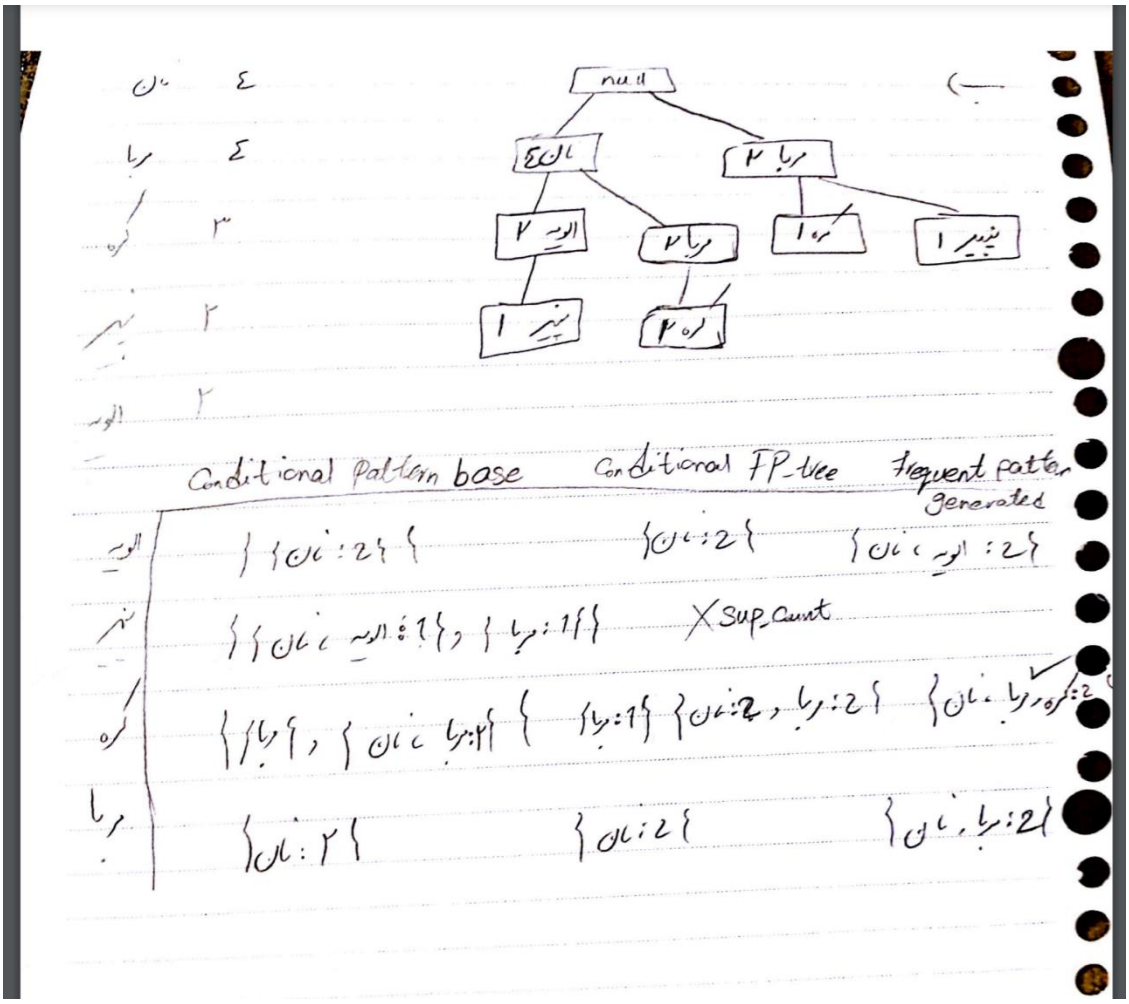
[نان] → [کره، مربا] = 2/3 = 66% ②

همه قوی اند.

7. الف)

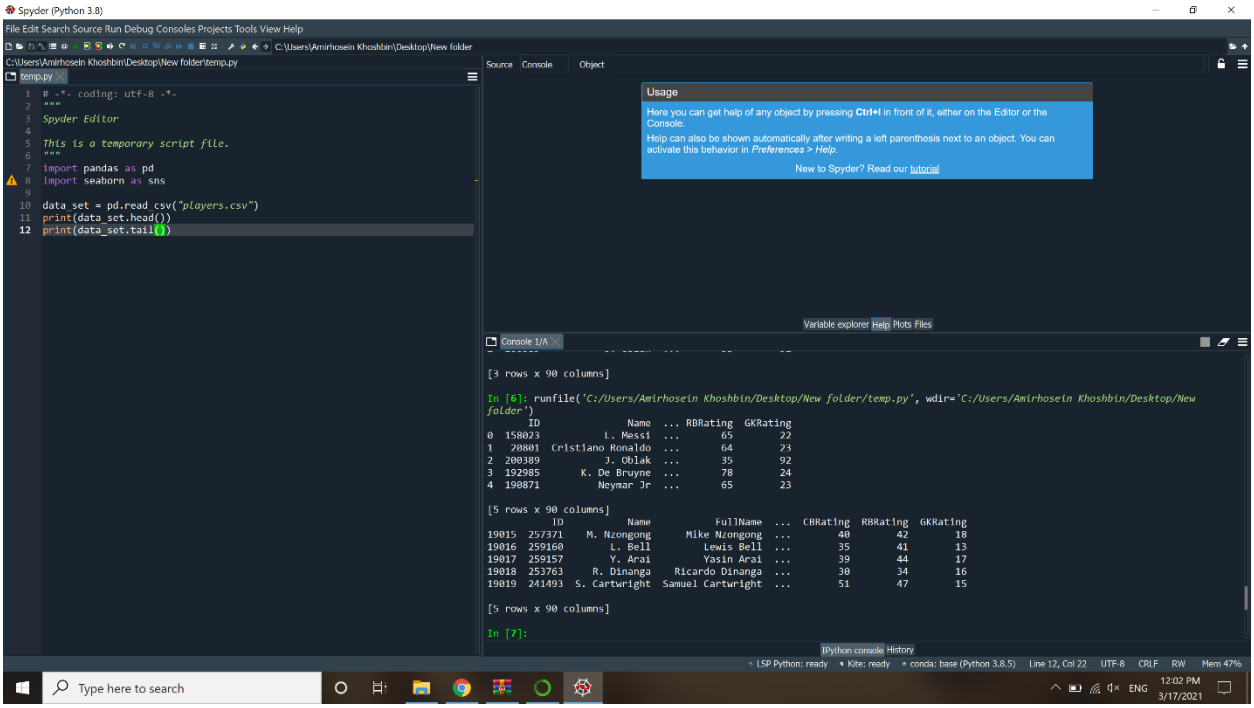

null
K = 1
K = 2
K = 3

ب)

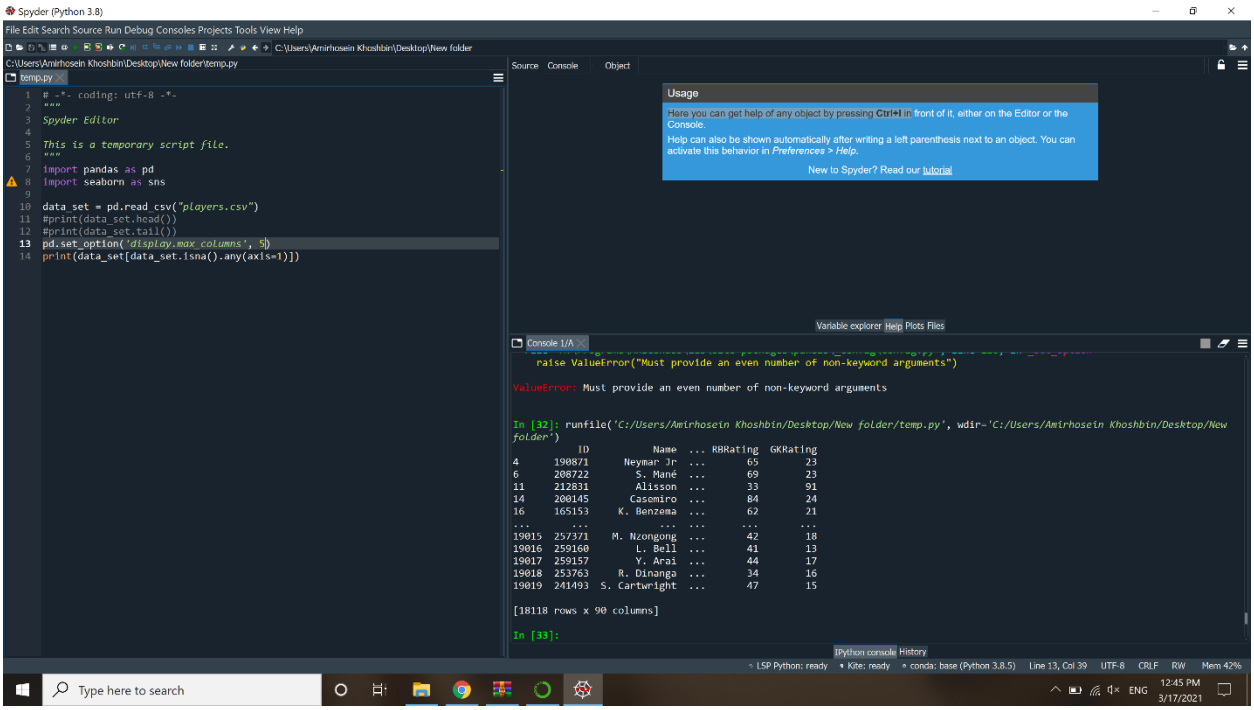

null
Conditional pattern base
Conditional FP-tree
Frequent pattern generated
X sup_count

بخش عملی:

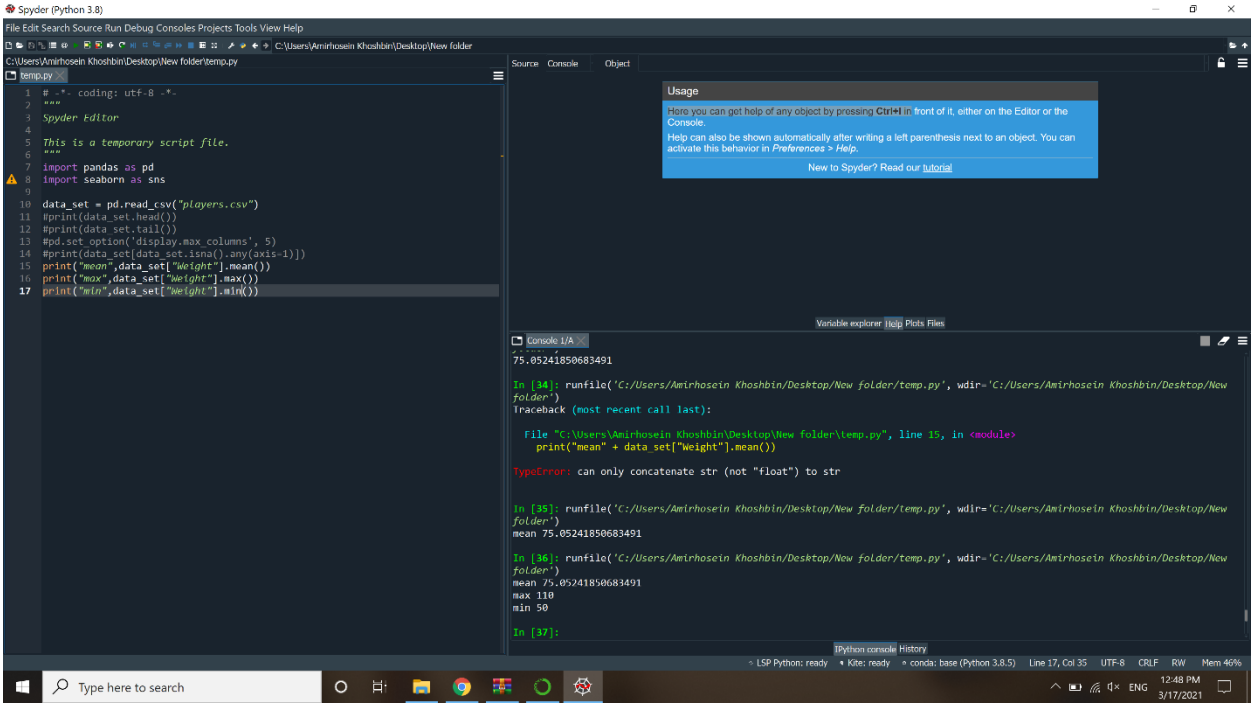
.1


Spyder (Python 3.8)
File Edit Search Source Run Debug Consoles Projects Tools View Help
C:\Users\Amirhosein Khoshbin\Desktop\New folder\temp.py
temp.py
-*- coding: utf-8 -*-
"""
Spyder Editor
This is a temporary script file.
"""
import pandas as pd
import seaborn as sns
data_set = pd.read_csv("players.csv")
print(data_set.head())
print(data_set.tail())
Usage
Here you can get help of any object by pressing Ctrl+I in front of it, either on the Editor or the Console.
Help can also be shown automatically after writing a left parenthesis next to an object. You can activate this behavior in Preferences > Help.
New to Spyder? Read our tutorial
Variable explorer Help Plots Files
Console 1/A
[3 rows x 90 columns]
In [6]: runfile('C:/Users/Amirhosein Khoshbin/Desktop/New folder/temp.py', wdir='C:/Users/Amirhosein Khoshbin/Desktop/New folder')
ID Name ... RBRating GKRating
0 158023 L. Messi ... 65 22
1 20801 Cristiano Ronaldo ... 64 23
2 200389 J. Oblak ... 35 92
3 192985 K. De Bruyne ... 78 24
4 190871 Neymar Jr ... 65 23
[5 rows x 90 columns]
ID Name FullName ... CBRating RBRating GKRating
19015 257371 M. Nzongong Mike Nzongong ... 40 42 18
19016 259160 L. Bell Lewis Bell ... 35 41 13
19017 259157 Y. Arai Yasin Arai ... 39 44 17
19018 253763 R. Dinanga Ricardo Dinanga ... 30 34 16
19019 241493 S. Cartwright Samuel Cartwright ... 51 47 15
[5 rows x 90 columns]
In [7]:
IPython console History
LSP Python: ready Kite: ready conda: base (Python 3.8.5) Line 12, Col 22 UTF-8 CRLF RW Mem 47%
Type here to search
ENG 12:02 PM 3/17/2021

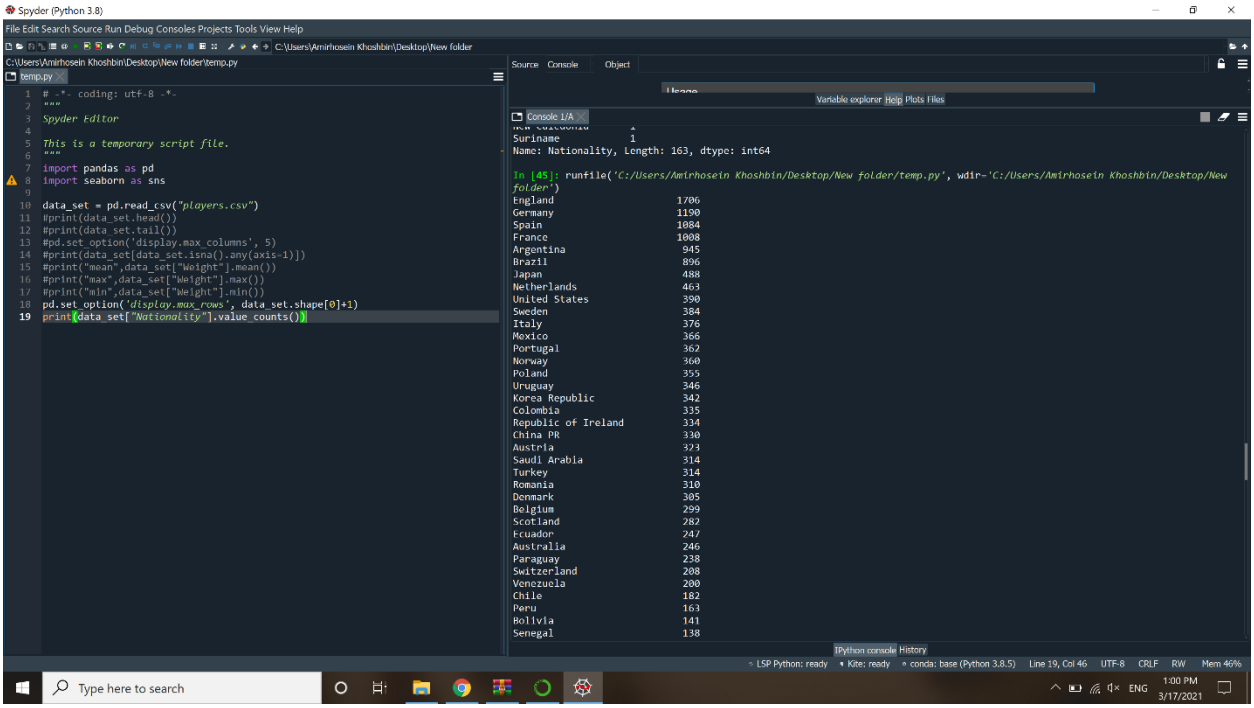
.2

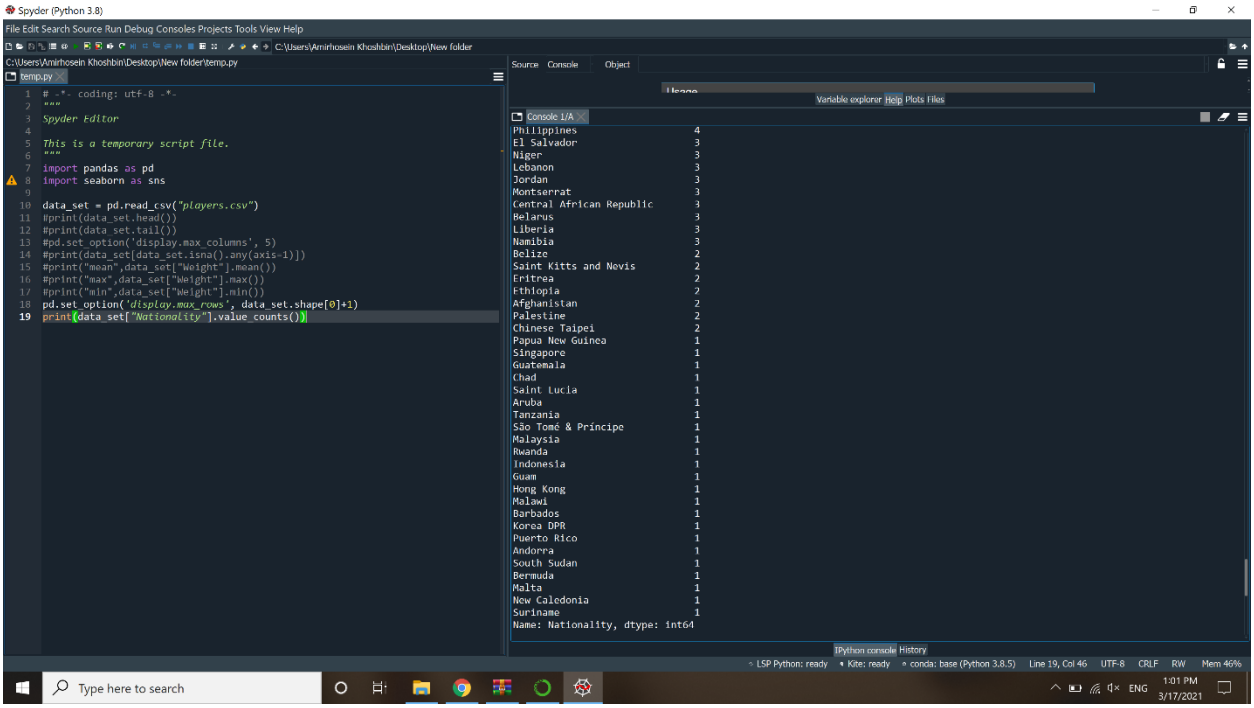

Spyder (Python 3.8)
File Edit Search Source Run Debug Consoles Projects Tools View Help
C:\Users\Amirhosein Khoshbin\Desktop\New folder\temp.py
temp.py
-*- coding: utf-8 -*-
"""
Spyder Editor
This is a temporary script file.
"""
import pandas as pd
import seaborn as sns
data_set = pd.read_csv("players.csv")
#print(data_set.head())
#print(data_set.tail())
pd.set_option('display.max_columns', 5)
print(data_set[data_set.isna().any(axis=1)])
Usage
Here you can get help of any object by pressing Ctrl+I in front of it, either on the Editor or the Console.
Help can also be shown automatically after writing a left parenthesis next to an object. You can activate this behavior in Preferences > Help.
New to Spyder? Read our tutorial
Variable explorer Help Plots Files
Console 1/A
raise ValueError("Must provide an even number of non-keyword arguments")
ValueError: Must provide an even number of non-keyword arguments
In [32]: runfile('C:/Users/Amirhosein Khoshbin/Desktop/New folder/temp.py', wdir='C:/Users/Amirhosein Khoshbin/Desktop/New folder')
ID Name ... RBRating GKRating
4 190871 Neymar Jr ... 65 23
6 208722 S. Mané ... 69 23
11 212831 Alisson ... 33 91
14 200145 Casemiro ... 84 24
16 165153 K. Benzema ... 62 21
... ...
19015 257371 M. Nzongong ... 42 18
19016 259160 L. Bell ... 41 13
19017 259157 Y. Arai ... 44 17
19018 253763 R. Dinanga ... 34 16
19019 241493 S. Cartwright ... 47 15
[18118 rows x 90 columns]
In [33]:
IPython console History
LSP Python: ready Kite: ready conda: base (Python 3.8.5) Line 13, Col 39 UTF-8 CRLF RW Mem 42%
Type here to search
ENG 12:45 PM 3/17/2021

.3

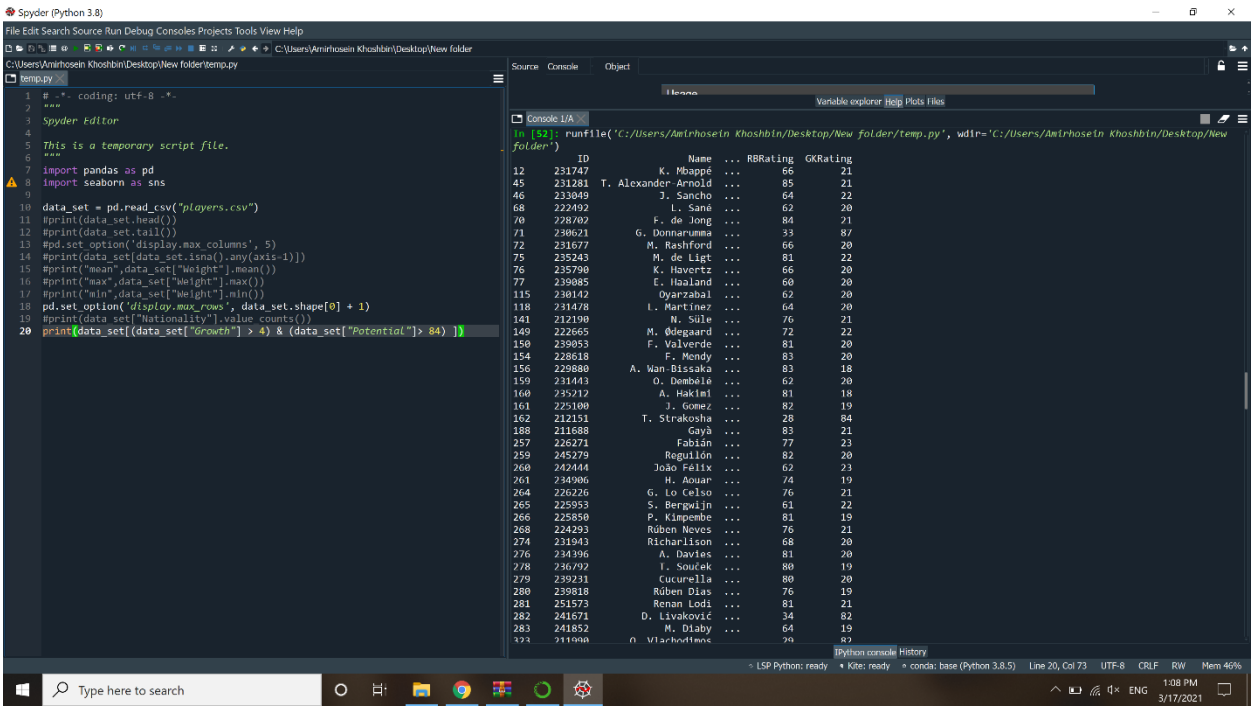


4. انگلیس: 1706، و تعداد زیادی از کشور ها که پایین مشخص است 1.

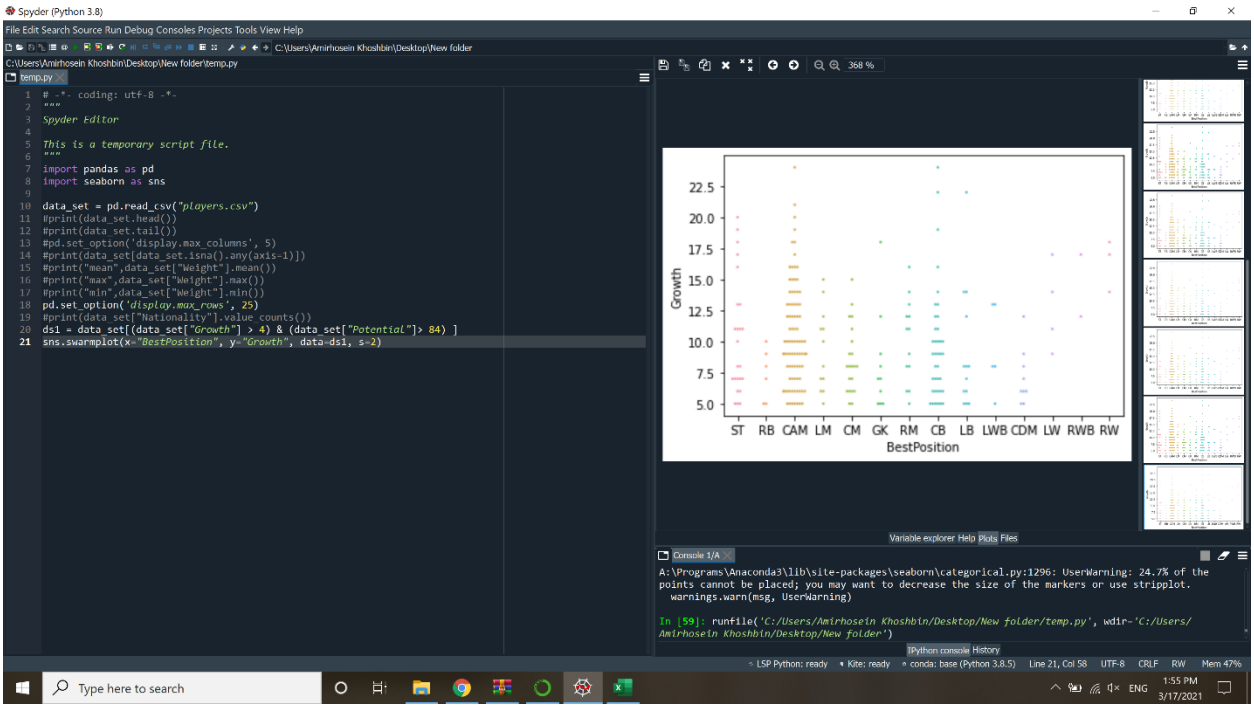




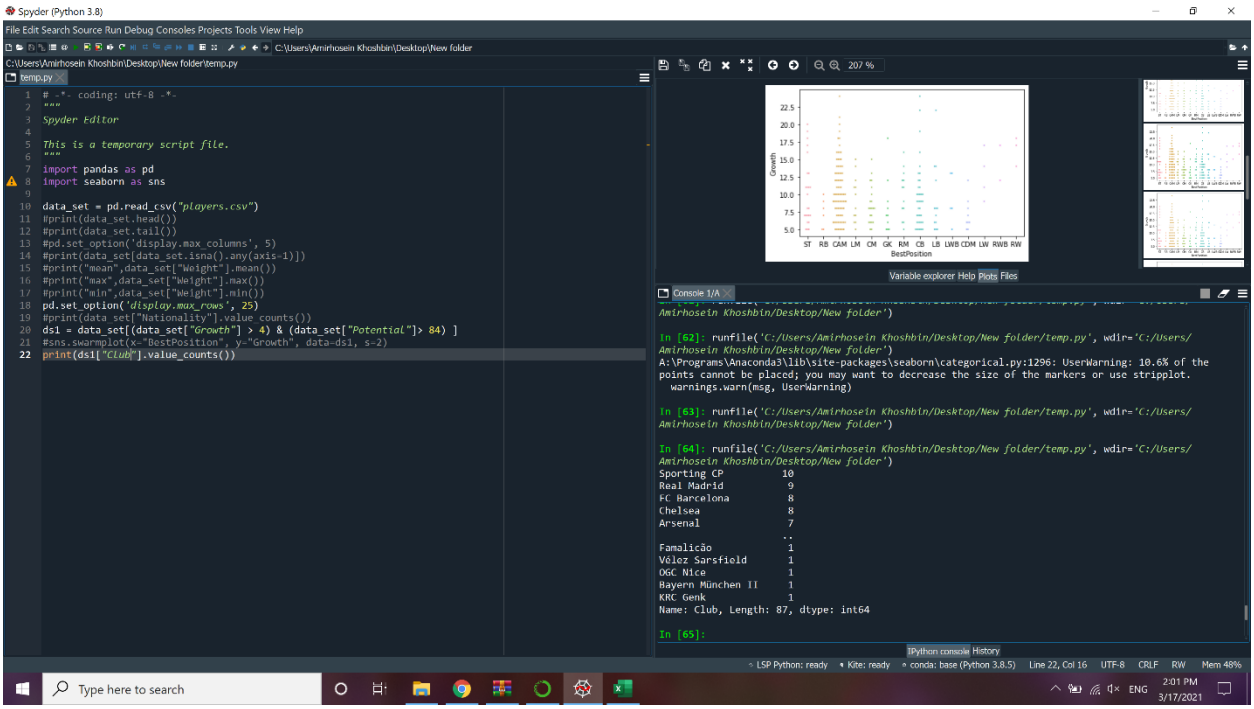
5. مشخصه به ترتیب، امباپه، ارنولد، سانچو و...



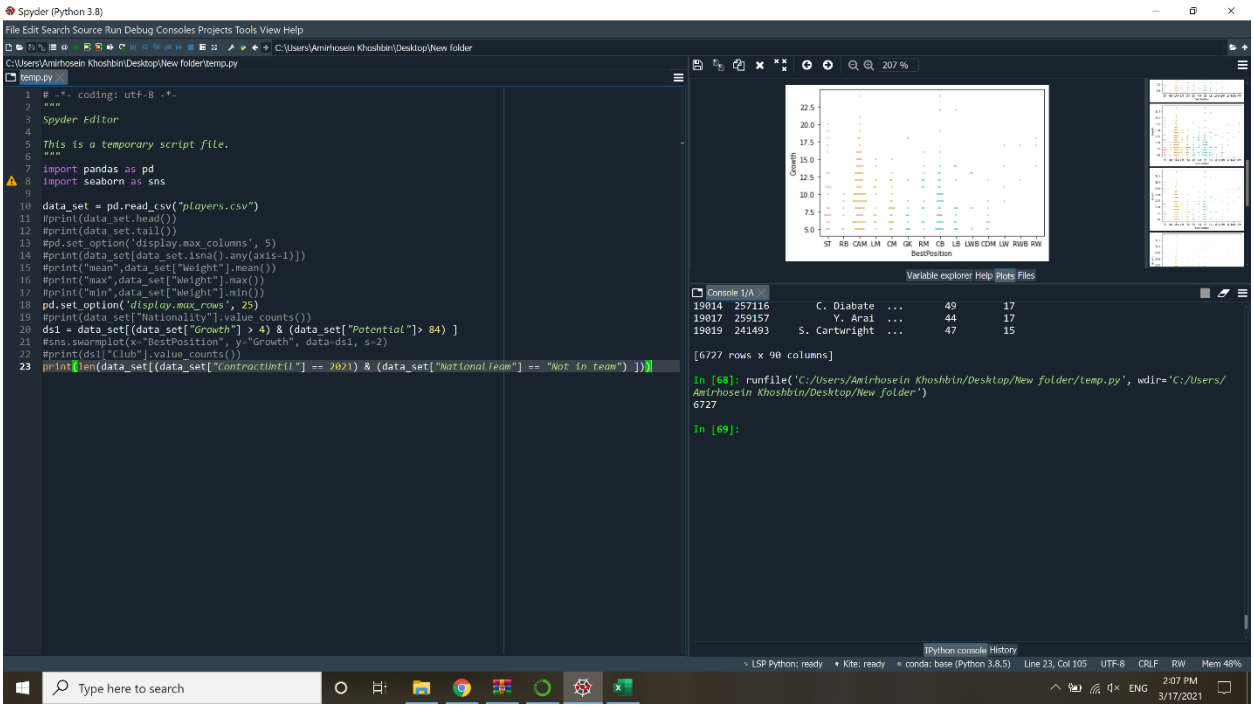
.6



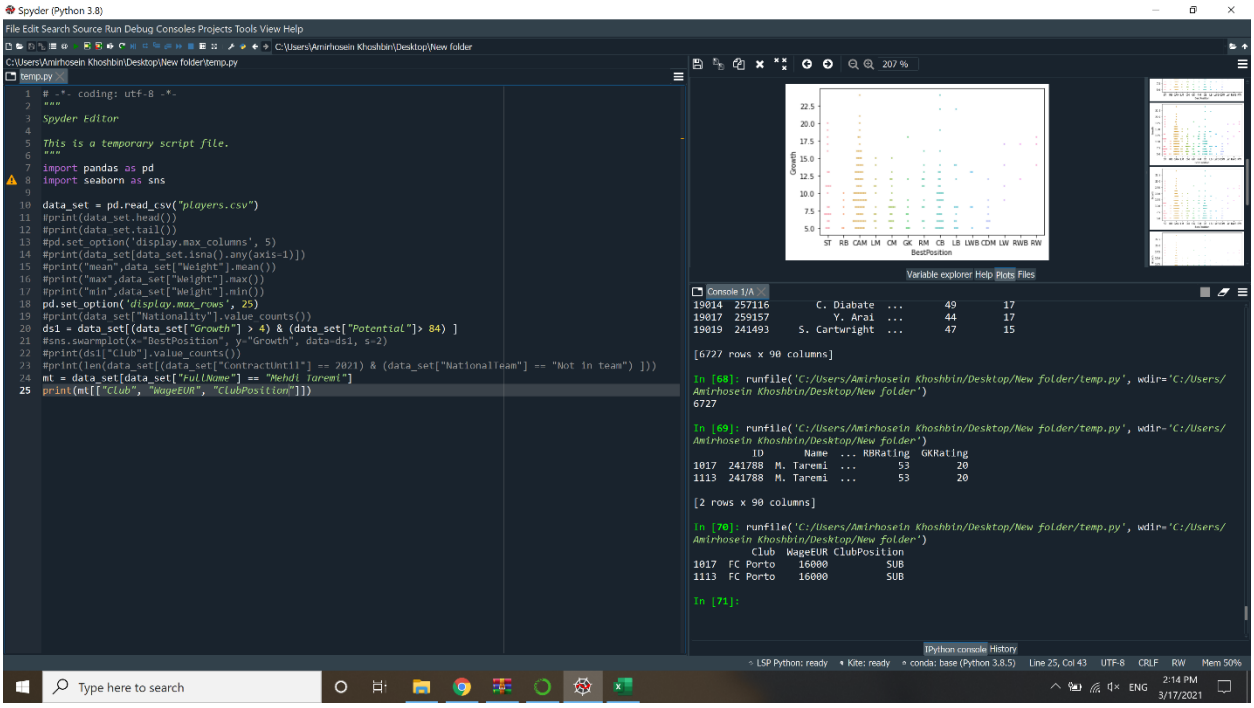
7.اسپورتینگ لیسبون با 10 بازیکن



9. 6727 تعداد این بازیکنان است.



10. موقعیت SUB، درامد 16000 باشگاه پورتو



2. upperBoundMinSupport: الگوریتم با این معیار شروع میکند 100% = 1.0 و با delta در هر دور کم میکند support رو. در واقع حد بالا برای حداقل support.

Delta: معیاری که در هر دور با آن حداقل support کم میشود.

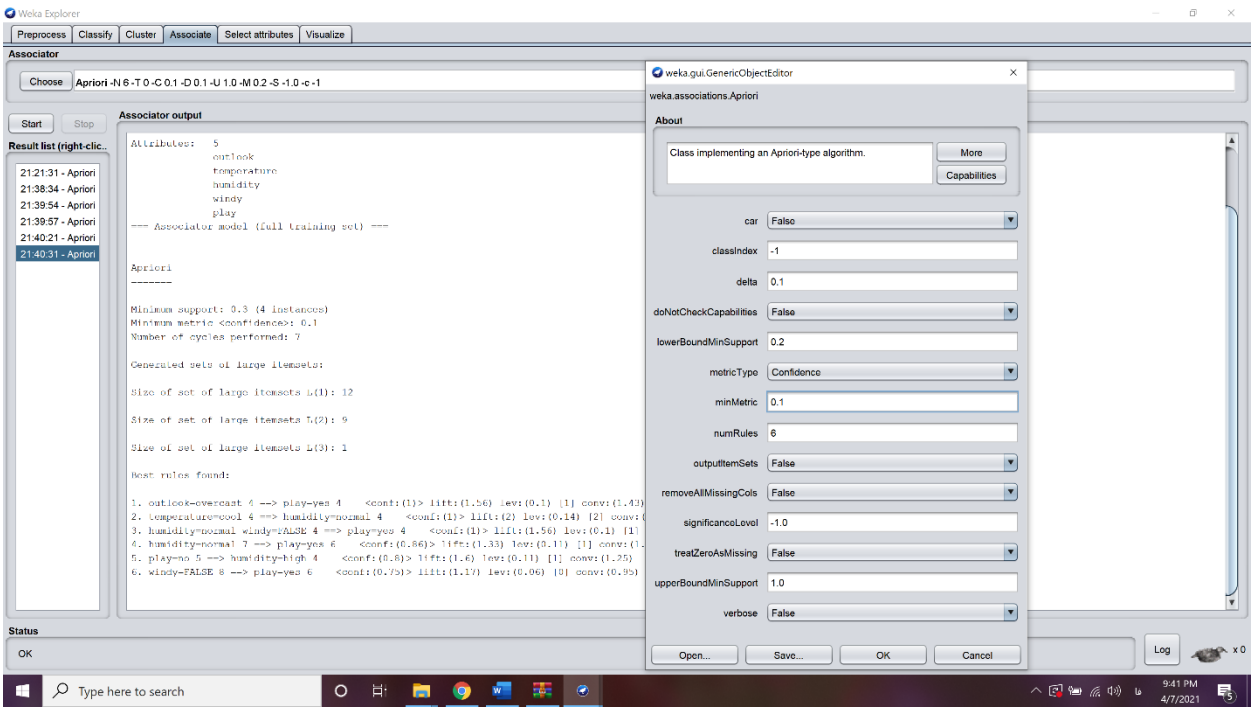
lowerBoundMinSupport: الگوریتم زمانی که support به این مقدار برسد، متوقف میشود.

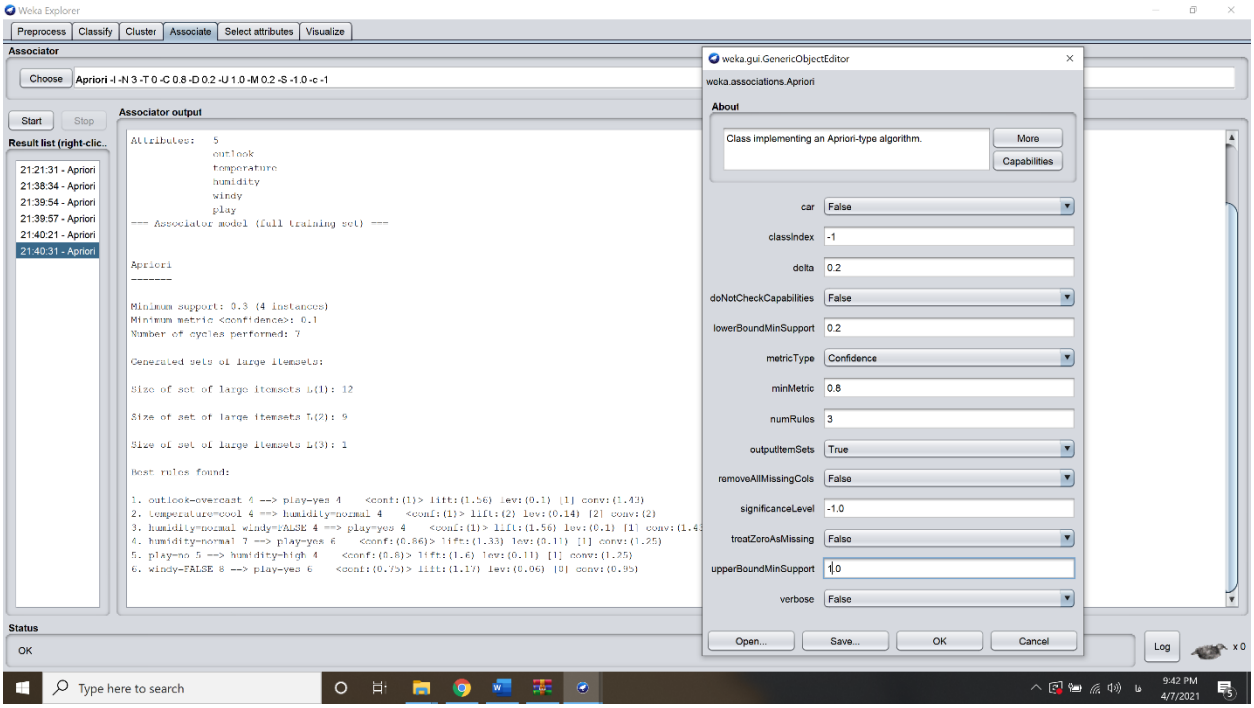
numRules: الگوریتم زمانی که به این تعداد بتواند قانون انجمنی تولید کند، متوقف میشود.

metricType: قوانین برا اساس این معیار مرتب میشوند.

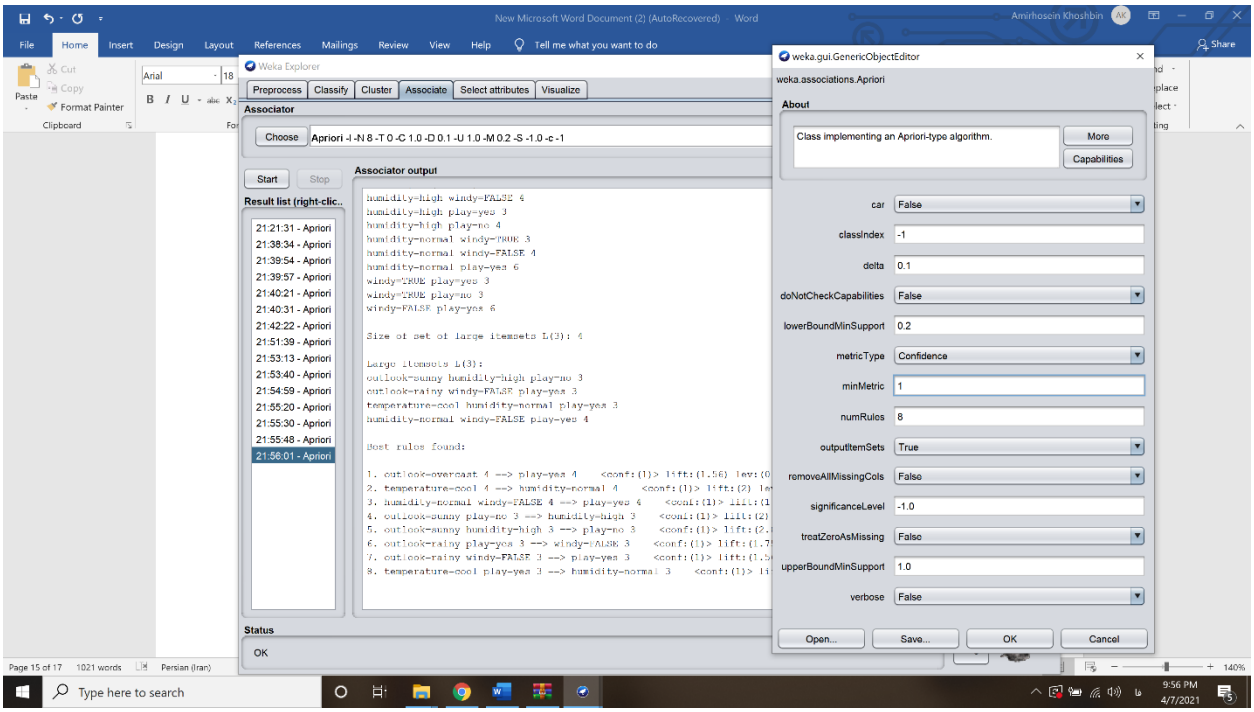
minMetric: قوانینی که از این عدد با معیار metricType بیشتر باشند به عنوان خروجی نمایش داده میشود.

outputItemSets: اگر میخواهیم تمام itemset های پرتکرار نمایش داده شوند، این باید true باشد.

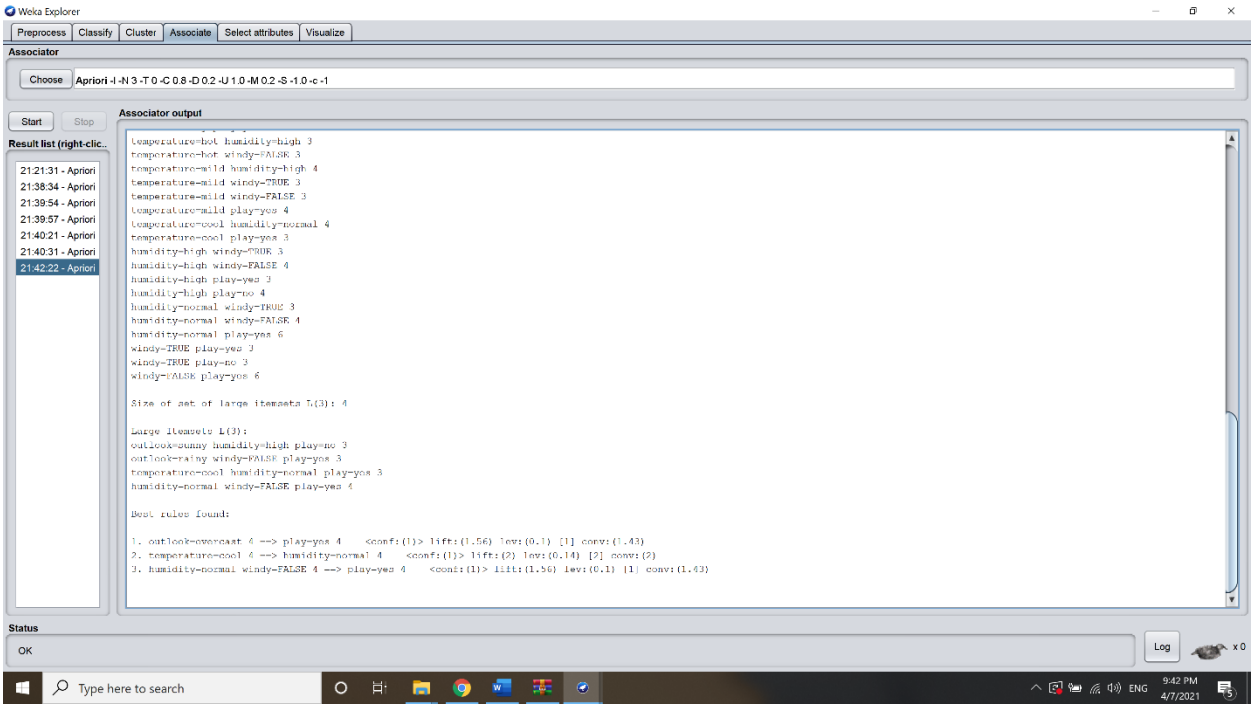


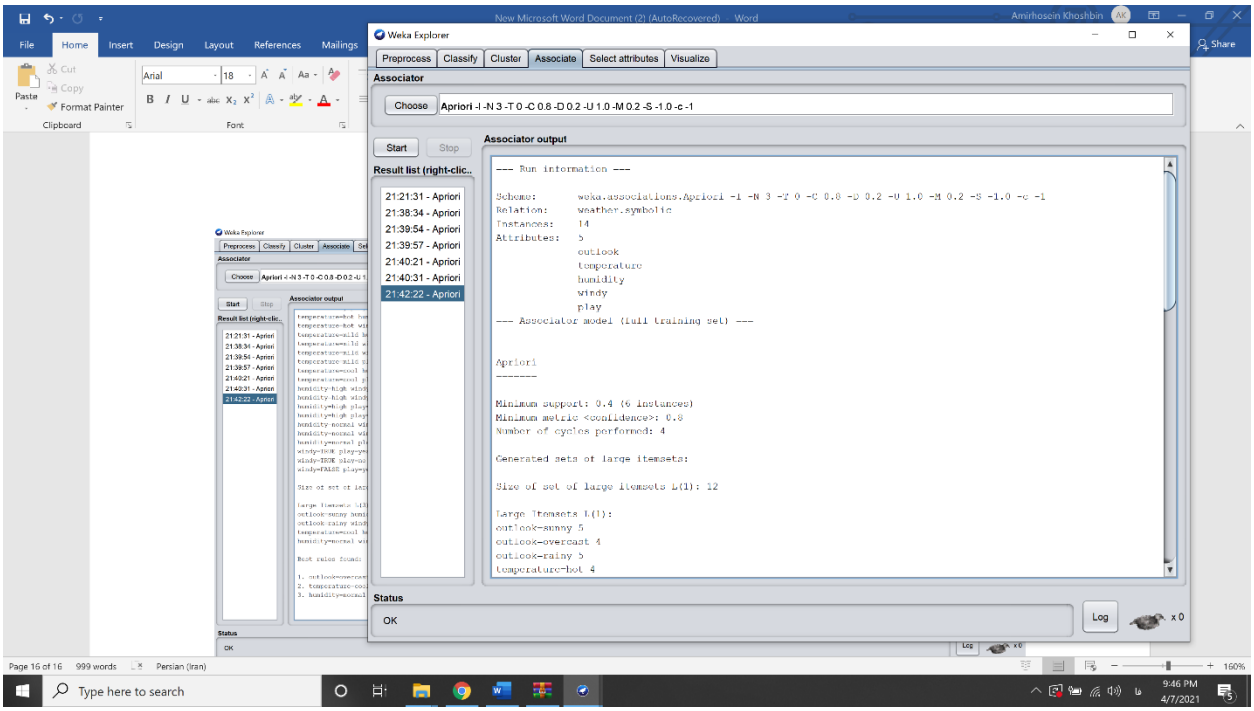


الان بر اساس confidence فیلتر میکنیم و metric هم 1 میذاریم و میبینیم که confidence های بزرگتر از 1 و برای confidence فقط 1 را نشان میدهد چون بزرگ تر از 1 معنی نمیدهد.



وقتی numrules را به 3 تغییر دادیم، فقط سه قانون تولید شد، و بر اساس confidence مرتب شده و outputItemSet، true شده و تمام set ها تولید شده نشان داده شده.




New Microsoft Word Document (2) (AutoRecovered) - Word
Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Associator
Choose
Apriori -I -N 3 -T 0 -C 0.8 -D 0.2 -U 1.0 -M 0.2 -S -1.0 -c -1
Start
Stop
Result list (right-clic..
21:21:31 - Apriori
21:38:34 - Apriori
21:39:54 - Apriori
21:39:57 - Apriori
21:40:21 - Apriori
21:40:31 - Apriori
21:42:22 - Apriori
Associator output
--- Run information ---
Scheme: weka.associations.Apriori -I -N 3 -T 0 -C 0.8 -D 0.2 -U 1.0 -M 0.2 -S -1.0 -c -1
Relation: weather.symbolic
Instances: 14
Attributes: 5
outlook
temperature
humidity
windy
play
--- Associator model (full training set) ---
Apriori
Minimum support: 0.4 (6 instances)
Minimum metric <confidence>: 0.8
Number of cycles performed: 4
Generated sets of large itemsets:
Size of set of large itemsets L(1): 12
Large Itemsets L(1):
outlook=sunny 5
outlook=overcast 4
outlook=rainy 5
temperature=hot 4
Status
OK
Log
x 0
Type here to search
9:46 PM
4/7/2021