UniFi & IMT | Master II - Data Science and Statistical Learning

Programming and Algorithm in Python & R

Gianmarco Santoro

Python | "Guinea baboons" dataset analysis

Python 3.10 | MacBook Pro 2014[1] | Local run code output synthesis and comments:

```
1) Answer: The Actor & Recipient with highest involvement is: NEKKE, with 964 events
Function time: 0.011709 s. Number of times where Nekke is Actor and number of times where Nekke is Recipient.

2) Answer: The Actor with highest involvement is: MAKO, with 479 events
Function time: 0.008172 s

3) Answer: The Recipient with highest involvement is: NEKKE, with 499 events
Function time: 0.009473 s

4) Answer: The Date with highest involvement is: 04/07/2019, with 369 events
Function time: 0.023041 s

5a) Answer: The hour, considering a precision of hour, with highest involvement is: 10, with 1480 events
Function time: 0.009007 s
5b) Answer: The hour, considering a precision of minutes, with highest involvement is: 10:30, with 45 events
Function time: 0.024616 s

6) Answer: The Behavior with highest involvement is: Resting, with 1834 events
Function time: 0.008676 s

7) Answer: The Actor & Recipient with highest involvement as couple is: ('MAKO', 'NEKKE'), with 180 events
Function time: 0.010171 s. The function recognizes the couple also when roles are switched.

8) Answer: The Actor & Recipient with highest time involvement is: ANGELE, with 24026 seconds
Function time: 0.009757 s. Amount of time where Angele is Actor and amount of time where Angele is Recipient.

9) Answer: The Actor with highest time involvement is: ANGELE, with 16468 seconds
Function time: 0.008795 s

10) Answer: The Recipient with highest time involvement is: FELIPE, with 12154 seconds
Function time: 0.008419 s

11) Answer: The Actor & Recipient with highest time involvement as couple is: ('KALI', 'PIPO'), with 8811 seconds
Function time: 0.011219 s. The function recognizes the couple also when roles are switched, e.g. ('KALI', 'PIPO')
is counted also when ('PIPO', 'KALI').
```

---

[1] MacBook Pro Retina (13-inch, Mid 2014), 2.6GHz dual-core Intel Core i5 processor (Turbo Boost up to 3.1GHz) with 3MB shared L3 cache, 8GB of 1600MHz DDR3L onboard memory.

12) Answer: Primate with most different behavior is: FELIPE with 17 different behaviours. (As well as LOME)
Function time: 0.008803 s. For each primate, the function counts the number of different behaviours just once for each kind of behaviour, both for Actor and Recipient roles.

13) Answer: Primate couple with most different behavior is: ('ANGELE', 'FELIPE') with 11 different behaviours
Function time: 0.007609 s. For each primate couple, the function counts the number of different behaviours just once for each kind of behaviour, no matter who is Recipient and who Actor.

14) Answer: Day with most different behavior is: 19/06/2019 with 16 different behaviours. (As well as 25/06/2019, 09/07/2019)
Function time: 0.007447 s. For each day, the function counts the number of different behaviours just once for each kind of behaviour.

15) Answer: Primate couple with most interactions is: ('ANGELE', 'FELIPE'), with 5764 interactions
Function time: 0.226859 s. The function recognizes the couple also when roles are switched.

16) Answer: Primate with most interactions is: EWINE, with 17019 interactions
Function time: 0.147620 s. The primate is counted every time it has interaction as Actor or Recipient.

17) Answer: There are 11 primates with 28 interaction days. Primate with most interaction days are: ARIELLE, FANA, VIOLETTE, HARLEM, FELIPE, ANGELE, BOBO, EWINE, FEYA, KALI, PETOULETTE.
Function time: 0.260151 s. The days are counted as one when Primate have interaction as Actor or Recipient at least once in that day.

18) Answer: There are 33 primate couples with 28 interaction days. Primate couples with most interaction days are: ('ARIELLE', 'FANA'), ('ARIELLE', 'VIOLETTE'), ('FANA', 'HARLEM'), ('ANGELE', 'FELIPE'), ('FANA', 'VIOLETTE'), ('FEYA', 'VIOLETTE'), ('HARLEM', 'VIOLETTE'), ('ANGELE', 'EWINE'), ('EWINE', 'HARLEM'), ('ARIELLE', 'HARLEM'), ('FANA', 'FEYA'), ('ANGELE', 'FEYA'), ('ANGELE', 'HARLEM'), ('EWINE', 'FANA'), ('EWINE', 'FELIPE'), ('ANGELE','FANA'), ('ARIELLE', 'FELIPE'), ('FELIPE', 'FEYA'), ('FANA', 'FELIPE'), ('FEYA', 'PETOULETTE'), ('ANGELE', 'PETOULETTE'), ('FANA', 'PETOULETTE'), ('EWINE', 'PETOULETTE'), ('FELIPE', 'HARLEM'), ('KALI', 'PETOULETTE'), ('BOBO', 'FANA'), ('EWINE', 'FEYA'), ('ARIELLE', 'EWINE'), ('BOBO', 'PETOULETTE'), ('ARIELLE', 'PETOULETTE'), ('ARIELLE', 'FEYA'), ('HARLEM', 'PETOULETTE'), ('ANGELE', 'ARIELLE')
Function time: 0.142141 s. The function recognizes the couple also when roles are switched, and the day counts as one if the couple has at least an interaction in that day.

Question 19, answer: Day with most interactions is: 13/06/2019, with 3577 interactions
Function time: 0.099833 s

• Global time to run the whole script: 1.044429 seconds

• Big O notation: in the context of algorithmic complexity analysis, the notation O(x) represents the worst-case time complexity of an algorithm, which is connected to efficiency and time, where x represents the size of the input. So, in this script all calls to the main function have an efficiency of about:

$$O(n \cdot m), \text{ where } m \ll n \Rightarrow O(n \cdot m) \approx O(n)$$

Since n are the number of lines in the txt file (for cycle => O(n)) and m is number of elements in very short lists (some calls have nested “for cycle” so O(n · m)). There are some other operations which could be summed to the count of big-O, but they have very lower magnitude, so can be not considered, for example: append() is usually O(1).

R | “Math Marks” dataframe analysis

Synthetic output and comments:

A) Show the header of dataframe generated. I’ve created also an SchoolIndex column to answer question 1.2:

	School	Sex	MathMark	SchoolIndex
1	Tecnico	Female	6.5	7
2	Tecnico	Female	6.0	128
3	Tecnico	Female	5.7	132
4	Liceo	Female	8.0	162
5	Liceo	Male	6.9	45
6	Tecnico	Female	6.5	152

B) 1) Number of Female and Male students per kind of school. Creates a two-way contingency table:

	Female	Male
Liceo	75625	92375
Tecnico	104588	127412

1.2) Number of Female and Male students in each school:

	Female	Male
1	894	1106
2	923	1077
…	…	…
199	902	1098
200	896	1104

2) Calculate statistics of MathMark for Male and Female students per kind of school:

	School	Sex	MathMark.Mean	MathMark.Median	MathMark.L_Quart.25%	MathMark.U_Quart.75%	MathMark.sd
1	Liceo	Female	6.9978896	7.0000000	6.4000000	7.6000000	0.9000883
2	Tecnico	Female	6.0012267	6.0000000	5.4000000	6.6000000	0.8971419
3	Liceo	Male	7.9342214	8.0000000	7.0000000	9.0000000	1.3837229
4	Tecnico	Male	6.9872069	7.0000000	6.0000000	8.0000000	1.4710903

3) Plot to see Math Marks vs Sex:

RPlot_1 = Distribution of Math Marks by Sex – see attached png-file

4) Plot to see Math Marks vs School:

RPlot_2 = Distribution of Math Marks by School – see attached png-file

5) Create a new variable and column with categorical-Math-Marks, so update dataframe:

	School	Sex	MathMark	SchoolIndex	CatMark
1	Tecnico	Female	6.5	7	Sufficient
2	Tecnico	Female	6.0	128	Sufficient
3	Tecnico	Female	5.7	132	Insufficient
4	Liceo	Female	8.0	162	Good
5	Liceo	Male	6.9	45	Sufficient
6	Tecnico	Female	6.5	152	Sufficient

6.1) Absolute frequency table of CatMark by Sex:

	Female	Male
Severely Insufficient	13446	12792

```
  Insufficient           45808 26105
  Sufficient             66269 45445
  Good                   50360 78813
  Excellent               4330 56632
```

6.2) Relative frequencies distribution of Categorical Marks for each Sex:

```
                          Female       Male
  Severely Insufficient 0.07461171 0.05820180
  Insufficient          0.25418810 0.11877409
  Sufficient            0.36772597 0.20676837
  Good                  0.27944710 0.35858809
  Excellent             0.02402712 0.25766765
```

6.3) Cumulative absolute frequencies:

```
Female-SI  Female-I  Female-S  Female-G  Female-E   Male-SI    Male-I    Male-S    Male-G    Male-E
    13446     59254    125523    175883    180213    193005    219110    264555    343368    400000
```

6.4) Cumulative relative frequencies, both for Female and Male:

```
Female-SI   Female-I   Female-S   Female-G   Female-E    Male-SI     Male-I     Male-S     Male-G     Male-E
0.07461171 0.32879981 0.69652578 0.97597288 1.00000000 1.05820180 1.17697589 1.38374426 1.74233235 2.00000000
```

6.5) Plotting the absolute frequency distribution:

RPlot_3 = Frequency Distribution of Categorical Marks by Sex – see attached png-file

6.6) Plotting the cumulative absolute frequency distribution:

RPlot_4 = Cumulative Frequency Distribution of Categorical Marks by Sex – see attached png-file

6.7) Plotting the frequency distribution:

RPlot_5 = Frequency Distribution of Categorical Marks by Mark – see attached png-file

7) Create a binary variable for Marks:

```
   School    Sex MathMark SchoolIndex      CatMark     ExcelMath
1 Tecnico Female      6.5           7   Sufficient Not Excellent
2 Tecnico Female      6.0         128   Sufficient Not Excellent
3 Tecnico Female      5.7         132 Insufficient Not Excellent
4   Liceo Female      8.0         162         Good Not Excellent
5   Liceo   Male      6.9          45   Sufficient Not Excellent
6 Tecnico Female      6.5         152   Sufficient Not Excellent
```

Finally, frequency of binary marks conditioned to School:

```
                     Liceo    Tecnico
  Excellent     0.23393452 0.09336638
  Not Excellent 0.76606548 0.90663362
```