Foulques GERAUD

MOOC S4

Python for data science

Planning

- Intro
- Getting started with data science
- Background in python and Unix
- Jupyter notebooks, Numpy
- Pandas
- Conclusion

Why this MOOC

- Discover python
- Mathematic side
- Concrete computing application
- Study process
- Global vision

Advances

- 80 h MOOC
- 8 chapters
- 1 project
- Progression $\rightarrow$ 5th week

Tools

- Mainly Jupyter notebook
- Python
- Libraries : numpy, matplotlib, pandas

- San Diego Supercomputer Center courses
- Forum topics

Getting started with data science

- Some about data science utility
- Why pyhton ?
- Study case : soccer data analysis

How to proceed into data science

- Getting started with the problem
- Know to ask the good question
- 6 steps : acquire, explore, preprocess, analyze, report, act

Python : Basics

- Useful because very furnished of libraries
- Dynamic variables. Types classification
- Background objects
- Loops with ranges, functions
- scope

Key data structures

- String usage
- Lists usage : remove, pop, append, copy…
- Tuples, dictionaries
- Syntax in lists

- Unix

```
list = [f(i) for i in range(0,n)]
```

Jupyter Notebooks

- Good presentation tool
- Collaborative
- From the beginning to the end

- Markdown : Latex, HTML
- In- and out- cells
- Presentation of tables

Exploring Data

We will start our data exploration by generating simple statistics of the data.

Let us look at what the data columns are using a pandas attribute called "columns".

In [17]:
```
df.columns
```

Out[17]:
```
Index(['id', 'player_fifa_api_id', 'player_api_id', 'date', 'overall_rating',
       'potential', 'preferred_foot', 'attacking_work_rate',
       'defensive_work_rate', 'crossing', 'finishing', 'heading_accuracy',
       'short_passing', 'volleys', 'dribbling', 'curve', 'free_kick_accuracy',
       'long_passing', 'ball_control', 'acceleration', 'sprint_speed',
       'agility', 'reactions', 'balance', 'shot_power', 'jumping', 'stamina',
       'strength', 'long_shots', 'aggression', 'interceptions', 'positioning',
       'vision', 'penalties', 'marking', 'standing_tackle', 'sliding_tackle',
       'gk_diving', 'gk_handling', 'gk_kicking', 'gk_positioning',
       'gk_reflexes'],
      dtype='object')
```

Next will display simple statistics of our dataset. You need to run each cell to make sure you see the outputs.

In [18]:
```
df.describe().transpose()
```

Out[18]:

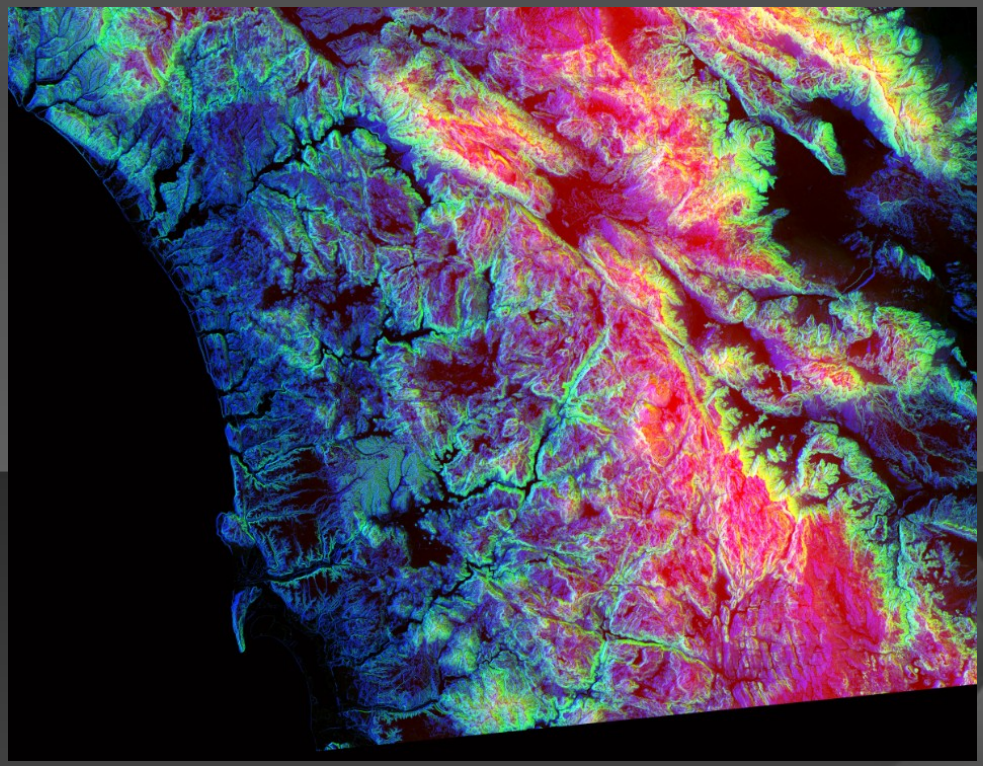
	count	mean	std	min	25%	50%	75%	ma
id	183978.0	91989.500000	53110.018250	1.0	45995.25	91989.5	137983.75	183978.
player_fifa_api_id	183978.0	165671.524291	53851.094769	2.0	155798.00	183488.0	199848.00	234141.
player_api_id	183978.0	135900.617324	136927.840510	2625.0	34763.00	77741.0	191080.00	750584.
overall_rating	183142.0	68.600015	7.041139	33.0	64.00	69.0	73.00	94.
potential	183142.0	73.460353	6.592271	39.0	69.00	74.0	78.00	97.
crossing	183142.0	55.086883	17.242135	1.0	45.00	59.0	68.00	95

Numpy

- Standard numeric library for arrays
- Scientific computing included
- Ndarrays : functions, mutability
- Slicing, Boolean indexing
- Native methods : operation and basic stats
- Broadcast
- speed

Satellite image application

- Images are arrays
- Colors
- Numpy’s importation
- Operations



Working with pandas

- Numpy but better
- Operations on data sets
- Time series, combinations, separation of data sets
- structures

Data structures

- Pandas series : dictionaries with sets of words as index
- Access by word and numeral indices
- Contains data not the same type
- Mutable operators + and *

- Data frame : transorms dict. Or pandas series
- Natural merge
- Clear show
- Intern methods

Data with pandas

- Ingestion : pandas.read()
- Getting variance, mean, min, max...
- .describe()

- Data cleaning : why
- Dropna
- estimations

Data visualization

- Histograms for time series
- Sticks for comparison
- Box plots for pushed study of a distribution
- Line diagrams for continuous, points for spotted situations…

- Master groupby, del, loc and drop is important

Frequent operations

- Merging dataframes : concat/append, join parameter
- Problem with key values
- .merge()

- Call : .str.method()
- Split, extract, contains, replace
- Regular expressions

- Using time with time
- POSIX time VS dated formats

Conclusion

- MOOC VS school
- Large domains : images, data cleaning…
- Concrete application of programming
- Nice features from Jupyter