Limpieza de datos con pandas

Silvia Vacacela

Estudiante de Maestría en Sistemas Inteligentes de EPN

@silviiVS

SilviaVacacela

Gabriela Guamán

Developer backend python

@_gabugm

gabygm

Agenda

- ETL(Extract-Transform-Load)
- Extracción
- Transformación(Limpieza de datos)
- Tareas de Limpieza de datos
- Pandas
- Ejemplo Práctico

ETL

EXTRACTION ⇨ **TRANSFORM** ⇨ **LOAD** ⇨ **ANALITYCS**

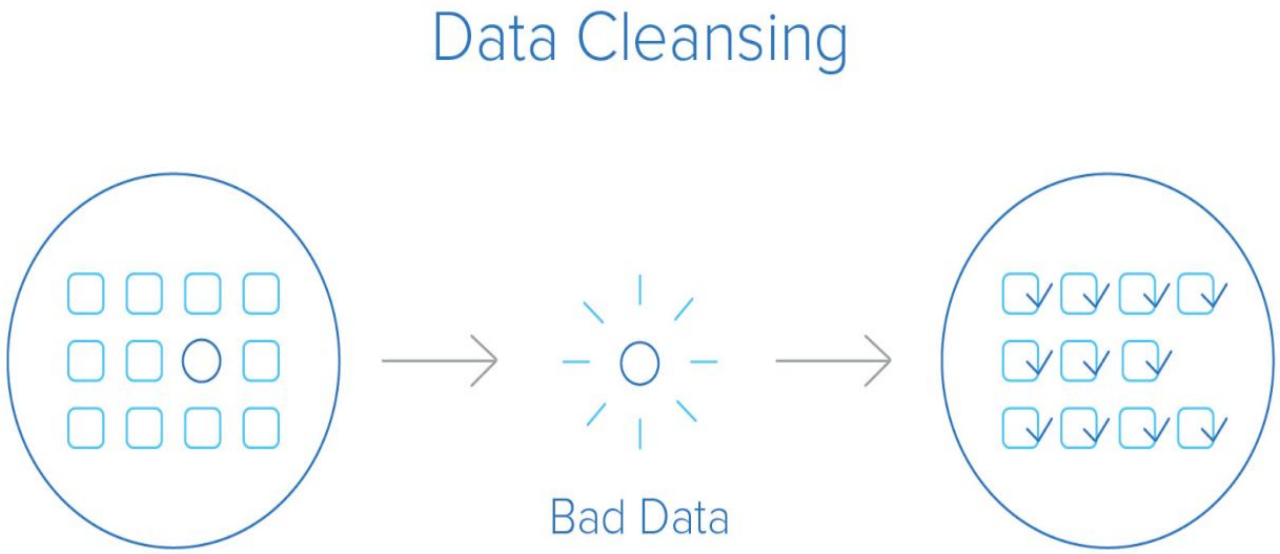
Extraction(Fuentes de datos)

- ✓ Web
- ✓ Base de datos
- ✓ Archivos CSV
- ✓ APIs
- ✓ Datasets publicos

Kaggle

http://www.datosabiertos.gob.ec/

Transform(Limpieza de datos)



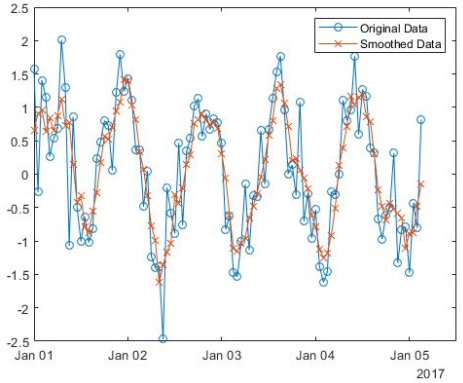
Tareas de Limpieza de datos

Llenar los valores faltantes

-

Identificar outliers y suavizar los datos ruidosos

-



Corregir los datos inconsistentes

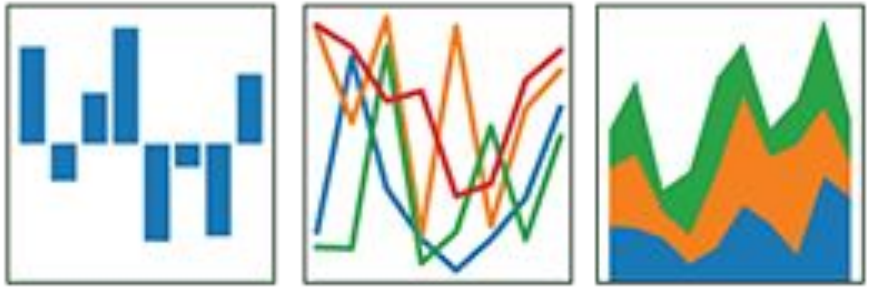
-

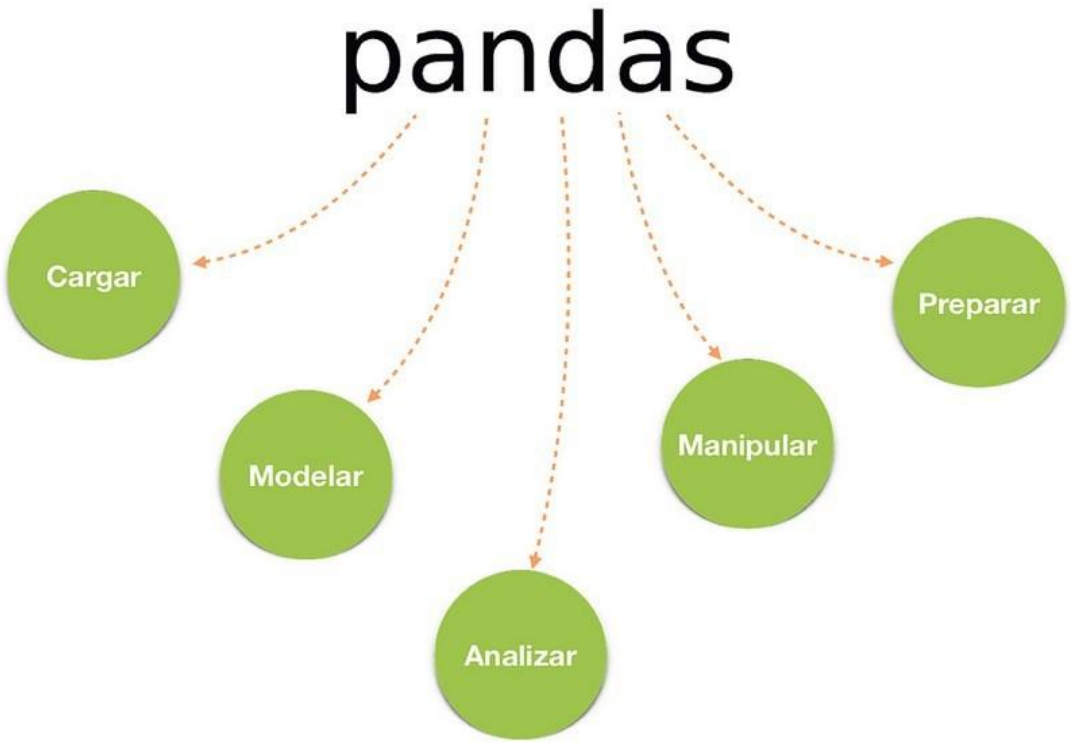
'Luis Medina'
'Luis medina'

¿Por qué Pandas?

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$

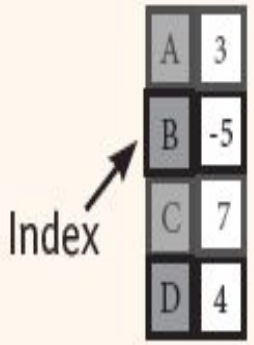




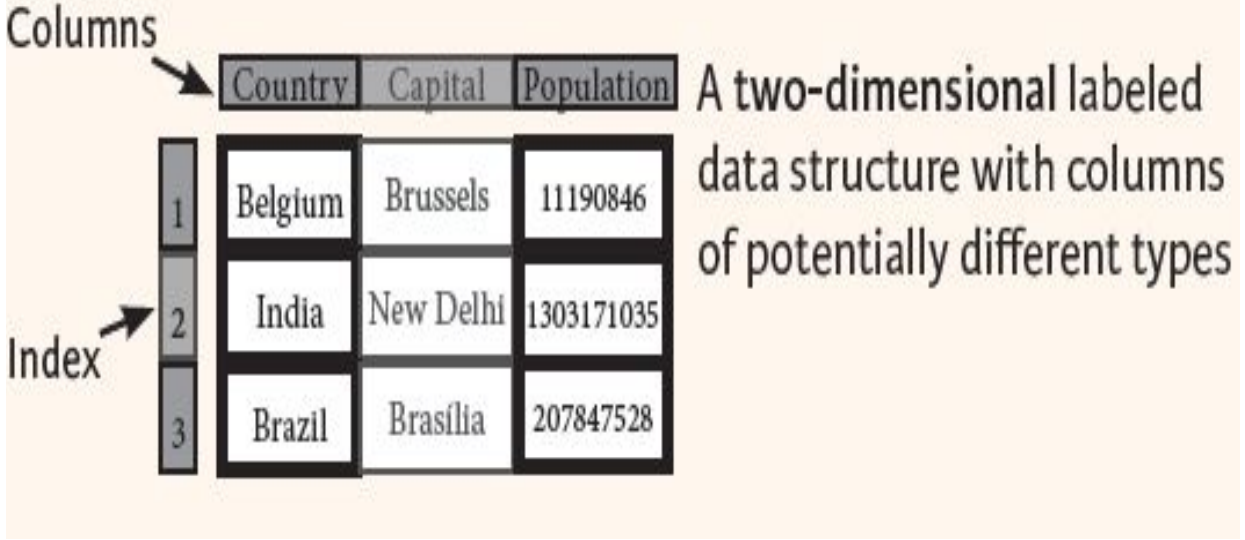
PANDAS DATA STRUCTURE

Series

A one-dimensional labeled array capable of holding any data type



DataFrame



EJEMPLO PRÁCTICO

Extract data meetup

Limpieza de datos con Pandas

In [ ]:
```
1 import requests
2 import pandas as pd
3 import hashlib
```

Extrar data api meetup

In [ ]:
```
1 members = requests.get('https://api.meetup.com/pyladiesEc/members')
```

In [ ]:
```
1 members?
```

In [ ]:
```
1 members.json()
```

Convert DataFrame

Convertir Json a DataFrame

In [ ]:
```
1 membersDataFrame = pd.DataFrame(members.json())
```

In [ ]:
```
1 membersDataFrame
```

In [ ]:
```
1 membersDataFrame.info()
```

Limit the number of rows to show

In [ ]:
```
1 pd.options.display.max_rows = 12
```

In [ ]:
```
1 #membersDataFrame.tail() #Five Last rows
```

In [ ]:
```
1 #membersDataFrame.head()#Five first rows
```

Ejercicio completo en:

https://github.com/pyladies-ecuador/data-cleaning/tree/master/dataCleaning