Taller de análisis de datos usando Python

Kevin Martinez-Folgar, MD. kmartinezfolgar@gmail.com
Sebastian Oliva datos@sebastianoliva.com

2017-11-07

CheatSheet Python y Pandas

Usamos import pandas as pd ; df como un dataframe de ejemplo y s como series y np como numpy, etc.

Cargar Datos

pd.read_csv(ruta_archivo) – DataFrame de un CSV
pd.read_excel(ruta_archivo) – De un Excel
pd.read_sql(query, obj_conexion) – Desde una tabla o query SQL
pd.read_json(string_json)
pd.read_html(ruta_archivo) - o mejor usa la libreria rows
pd.read_clipboard()
pd.DataFrame(obj) – De un dict de Python

Exportar Datos

```
df.to_csv(ruta_archivo)
df.to_excel(ruta_archivo)
df.to_sql(tabla, obj_conexion)
# Generar rango de fechas.
pd.date_range('1900/1/30',
periods=12*44, freq='M')
```

Revisar

df.head(n) – top n filas
df.tail(n)
df.describe() – forma y tipos de columnas
df.info()
s.value_counts(dropna=False)

Selección y Filtrado

df[col] – serie de la columna
df[[col1, col2]] – dataframe columnas
s.iloc[0] – selección por posición
s.loc[0] – selección por valor de índice
df.iloc[0,:] - primera fila, todas columnas
df.iloc[0,0] – primera fila, primer valor
df.iloc[:,[1,2,3]] – todas las filas, primeras tres columnas
df[df['Col']>1234] – todas donde el valor de Col sea mayor a 1234

Limpieza y manipulación

```
df.columns = ['a','b','c']
df.dropna()
df.fillna(x)
s.astype(float)
s.replace([1,3],
     ['primera','tercera'])
df.set_index('id_col')
pd.melt(df) – de fila a columna
pd.pivot_table(df, values='D',
index=['A', 'B'],
columns=['C'], aggfunc=np.sum) -
pivota sobre la tabla df, con los valores D
agrupados en suma
df.groupby(by='col').agg(sum)
df.groupby(level='ind').agg(avg)
df.groupby(by=['col1','col2']).agg
(sum)
```