PROGRAMACIÓN tics100

FACULTAD DE INGENIERÍA Y CIENCIAS.
UNIVERSIDAD ADOLFO IBAÑEZ

Primer semestre 2019

Pandas (II)

Ejemplo

- Para aprender Pandas de forma más práctica, iremos paso a paso usando datos de los juegos olímpicos obtenidos de https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results
  - Una copia de los datos se encuentra en webcursos para que los bajen
- Lo que haremos es:
  - Cargar los datos
  - Analizar un dataframe
  - Acceder a datos
  - Realizar filtros
  - Obtener estadísticas

Agrupando datos

```
import pandas as pd

df = pd.read_csv('athlete_events.csv')

grupos = df.groupby('Year')
```

groupby () nos permite crear agrupaciones por un campo

```
print(grupos['Weight', 'Height'].agg(['min', 'max', 'mean']))
```

Y sobre esos grupos podemos sacar estadísticas de ciertos campos, en este caso de peso y altura

	Weight			Height		
	min	max	mean	min	max	mean
Year						
1896	45.0	106.0	71.387755	154.0	188.0	172.739130
1900	51.0	102.0	74.556962	153.0	191.0	176.637931
1904	43.0	115.0	72.197279	155.0	195.0	175.788732
1906	52.0	114.0	75.917073	165.0	196.0	178.206226

Ejercicio

Tiempo : 10 minutos

1. Compare el peso y altura promedio de los atletas de invierno con los de verano.
Para ello, primero debe agrupar los atletas para cada temporada, y luego seleccionar altura y peso y pedir los valores promedio solamente (mean)
2. Ahora realice la comparación anterior, pero solamente para los juegos olímpicos desde 1996 a la fecha

Recuerde que tiene los siguientes campos:

- Height - Altura en centímetros
- Weight - Peso en kilogramos
- Season - Temporada (Summer o Winter)

Ejercicio

Tiempo : 10 minutos

```
import pandas as pd

df = pd.read_csv('athlete_events.csv')

grupos = df.groupby('Season')
print(grupos['Weight', 'Height'].agg(['mean']))

desde_1996 = df[df['Year']>=1996].groupby('Season')
print(desde_1996['Weight', 'Height'].agg(['mean']))
```

Graficando datos

```
import pandas as pd
import matplotlib.pyplot as mpl

df = pd.read_csv('athlete_events.csv')

grupos = df.groupby('Year')
datos = grupos['Weight', 'Height'].agg(['min', 'max', 'mean'])

mpl.scatter(datos.index, datos ['Weight']['mean'])
```

scatter() nos permite graficar puntos, en este caso usamos index, y la media de los pesos

Graficando datos

```
import pandas as pd
import matplotlib.pyplot as mpl

df = pd.read_csv('athlete_events.csv')

grupos = df.groupby('Year')
datos = grupos['Weight', 'Height'].agg(['min', 'max', 'mean'])

mpl.scatter(datos.index, datos ['Weight']['mean'])
```

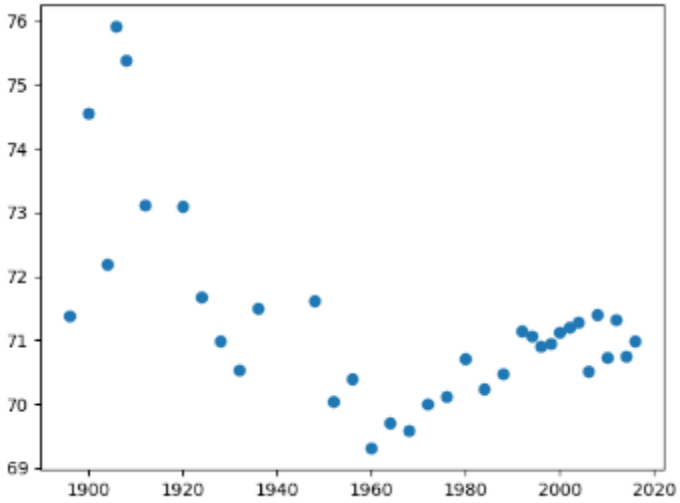
index posee los valores con los que agrupamos, en este caso los años de los juegos

```
mpl.savefig('grafico.png')
```

savefig() permite guarder el gráfico en una imagen

Inspeccionando datos

- Observe el gráfico que obtuvimos ¿Qué puede deducir de este gráfico?



Graficando datos

```
import pandas as pd
import matplotlib.pyplot as mpl

df = pd.read_csv('athlete_events.csv')

mujeres_2016 = df[(df['Sex']=='F') & (df['Year'] == 2016)]
alturas = mujeres_2016['Height'].dropna()
```

dropna() elimina los registros que no tienen dato de Altura

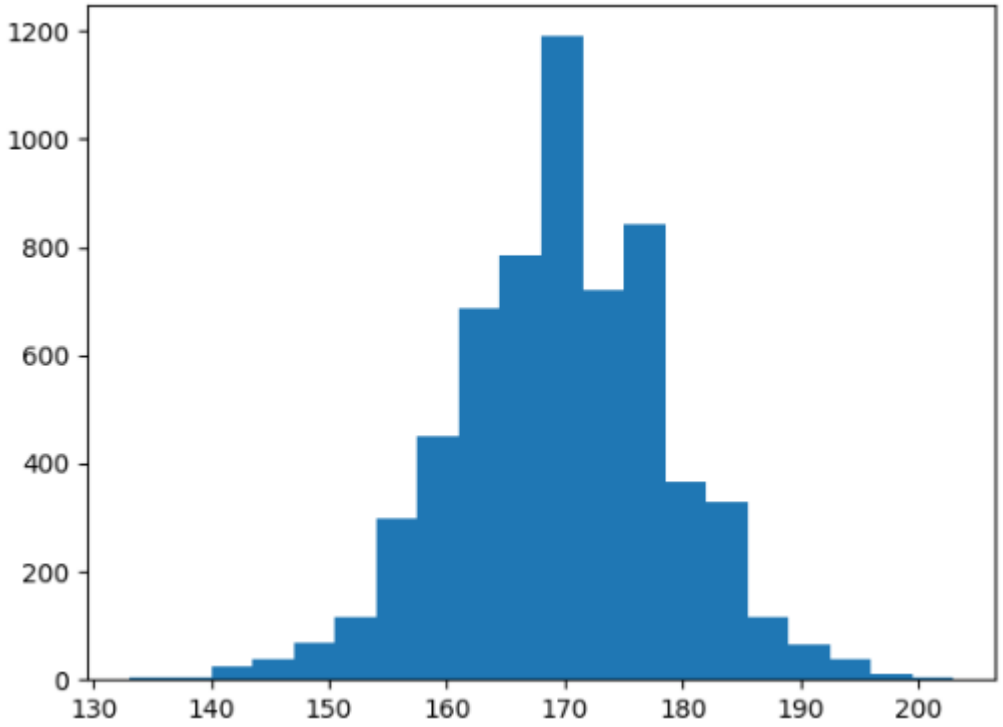
```
mpl.hist(alturas, bins=20)
```

hist() nos permite graficar histogramas (agrupación por rango de Alturas en este caso)

```
mpl.savefig('histograma.png')
```

Inspeccionando datos

- Observe el gráfico que obtuvimos ¿Qué puede deducir de este gráfico?



Ejercicio

Tiempo : 30 minutos

1. Liste todos los deportes de los juegos olímpicos de verano
2. Seleccione un deporte y analice los datos de los atletas en el tiempo:
   - Peso
   - Altura
   - Edad
3. Grafique los datos que juzgue relevantes