

# Taller de Agrupamiento Automático

---

## Objetivo

Utilizar algoritmos de agrupamiento automático para analizar la estructura de la colección de datos de relaciones extra-maritales.

---

## Procedimiento

### 1. Abrir Python

- Para abrir python siga las instrucciones del ítem número 1 de la **Guía de conexión** e ingrese a su carpeta personal dando doble clic sobre ella.

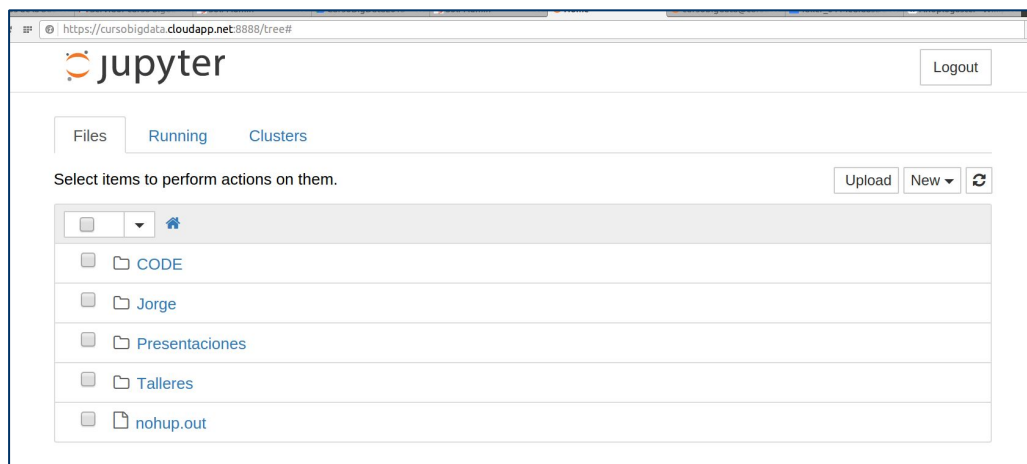


Figura 1

Una vez dentro de su carpeta de clic en **New** (recuadro de color rojo) situado al costado superior derecho de la ventana y luego de clic en **Python 2** (recuadro de color verde) como se ilustra en la **Figura 2**. Posteriormente se abrirá una pestaña en la que aparece una ventana como la mostrada en la **Figura 3**.

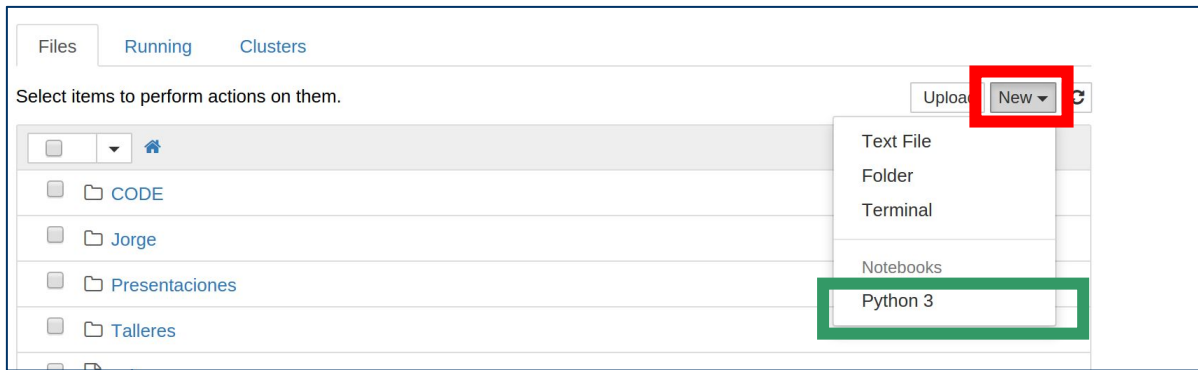


Figura 2

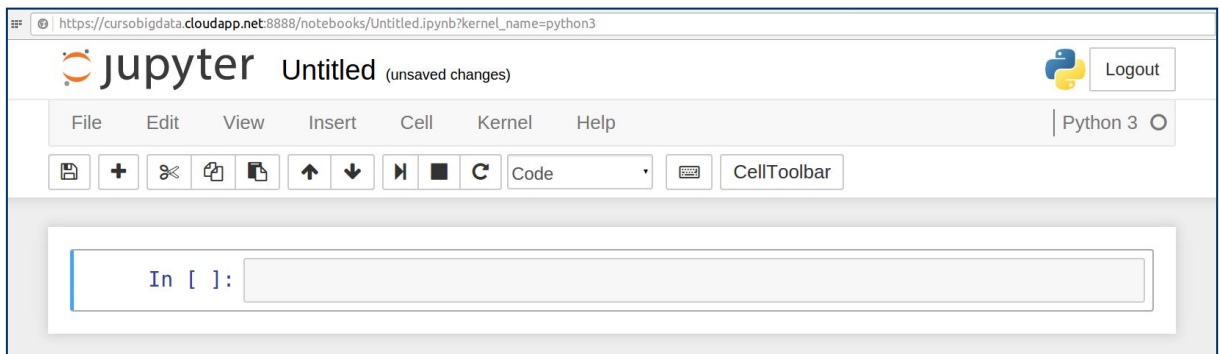


Figura 3

## 2. Renombrar el Notebook

- Para renombrar el Notebook abierto de clic en **File** situado al costado superior izquierdo de la ventana y luego de clic en **Rename...** (recuadro de color verde) como se ilustra en la **Figura 4**. Posteriormente se abrirá una ventana emergente como la mostrada en la **Figura 5**.

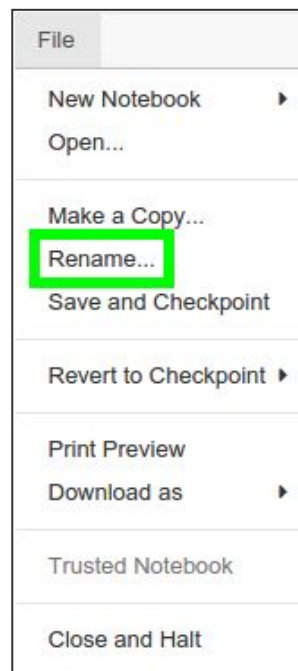


Figura 4



Figura 5

Reemplace **Untitled** por **Taller\_06** y de clic en **OK**.

## 3. Desarrollo del código

- Tenga en cuenta los botones señalados de color rojo y de color verde de la siguiente barra de herramientas.

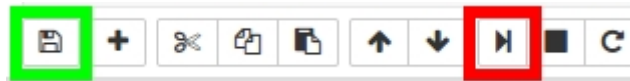




Figura 6

- Importar los módulos SciPy, NumPy y Pyplot de Python, empleados en matemáticas, ciencia e ingeniería, la manipulación de matrices y creación de gráficas, respectivamente. Escriba en la celda las instrucciones que aparecen a continuación y de clic en  (recuadro señalado de color rojo en la **Figura 6**) para ejecutar esa instrucción.

```
In [11]: %matplotlib inline
import scipy
import numpy as np
import matplotlib as plt
```

Bloque de código 1


Cargue el dataset ubicado en **/home/cursobigdata/Notebooks/Files/ExtramaritalAffairs\_data.csv** a la variable **A**, que será una matriz. Escriba en la celda el comando que aparece a continuación y de clic en .

```
A=np.loadtxt('/home/cursobigdata/Notebooks/Files/ExtramaritalAffairs_data.csv',
delimiter=',', skiprows= 2)
```

```
In [ ]: A=np.loadtxt('/home/cursobigdata/Notebooks/Files/ExtramaritalAffairs_dat
```


```
In [ ]: otebooks/Files/ExtramaritalAffairs_data.csv', delimiter=',', skiprows= 2)
```

Bloque de código 2

Cree una matriz de ceros con las mismas dimensiones de la matriz **A** (6366 filas, 9 columnas) y escoja las 8 primeras columnas de la matriz de ceros **X**. Escriba en la celda el comando que aparece a continuación y de clic en .


```
In [14]: X = np.zeros(A.shape)
X = X[:, 0:8]
```

Bloque de código 3

Normalice los valores de la matriz **A** y almacénelos en la matriz **X**. Escriba en la celda las instrucciones que aparecen a continuación y de clic en .


```
In [16]: for i in range(8):
         X[:,i] = (A[:,i] - np.min(A[:,i])) / (np.max(A[:,i]) - np.min(A[:,i]))
```

#### Bloque de código 4

Extraiga y almacene en **X** los 100 primeros registros de ésta matriz (**X**, que contiene los datos de la matriz **A** normalizados), y almacene en **Y** los datos de los primeros 100 registros de la matriz **A** correspondientes a la última columna (número de infidelidades por año). Escriba en la celda las instrucciones que aparecen a continuación y de clic en .

```
In [17]: X = X[0:100,:]
         Y = A[0:100,-1]
```

#### Bloque de código 5

Visualice en un histograma (representación gráfica de una distribución de frecuencias por medio de barras) el **número de infidelidades por año** vs el **número de individuos** de los primeros 100 registros del dataset (almacenados en **Y**). Escriba en la celda las instrucciones que aparecen a continuación y de clic en . Visualice el histograma en la figura 7.

```
In [ ]: plt.figure(figsize=(12,5))
         plt.title('Relaciones Extramaritales')
         plt.xlabel(u'Infidelidades por año')
         plt.ylabel(u'Número de individuos')
         h = plt.hist(Y,100)
```

#### Bloque de código 6

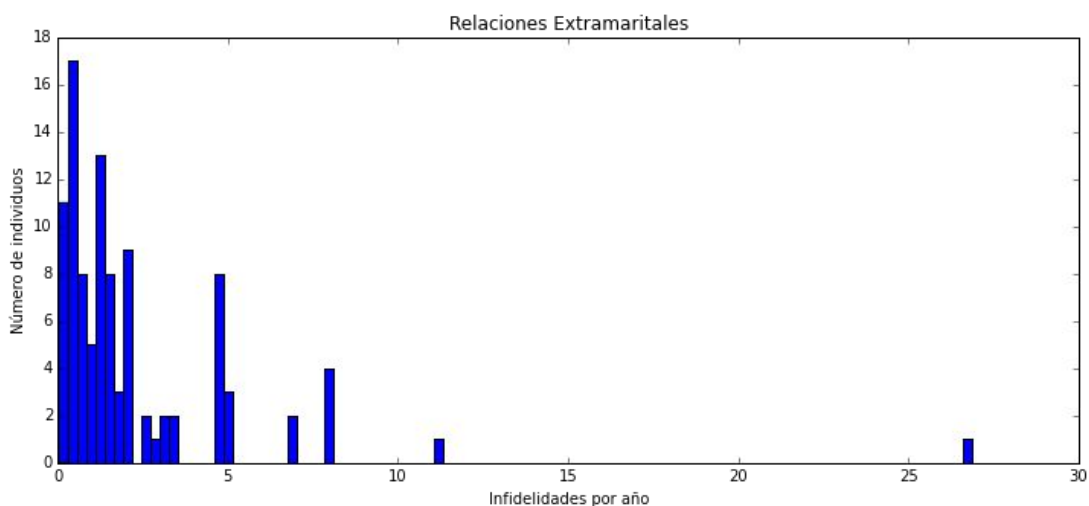



Figura 7

Visualice en un gráfico de dispersión la relación entre la **edad de cada individuo** presente en el dataset (matriz **A**) y el **número de infidelidades por año**. Escriba en la celda la instrucción que aparece a continuación y de clic en . Visualice el resultado en la figura 8.

```
In [ ]: plt.figure(figsize=(12,5))
plt.title('Relaciones extramaritales por edad')
plt.xlabel(u'Edad')
plt.ylabel(u'Infidelidades por año')
plt.scatter(A[:,1],A[:,1],s=150)
```

Bloque de código 7

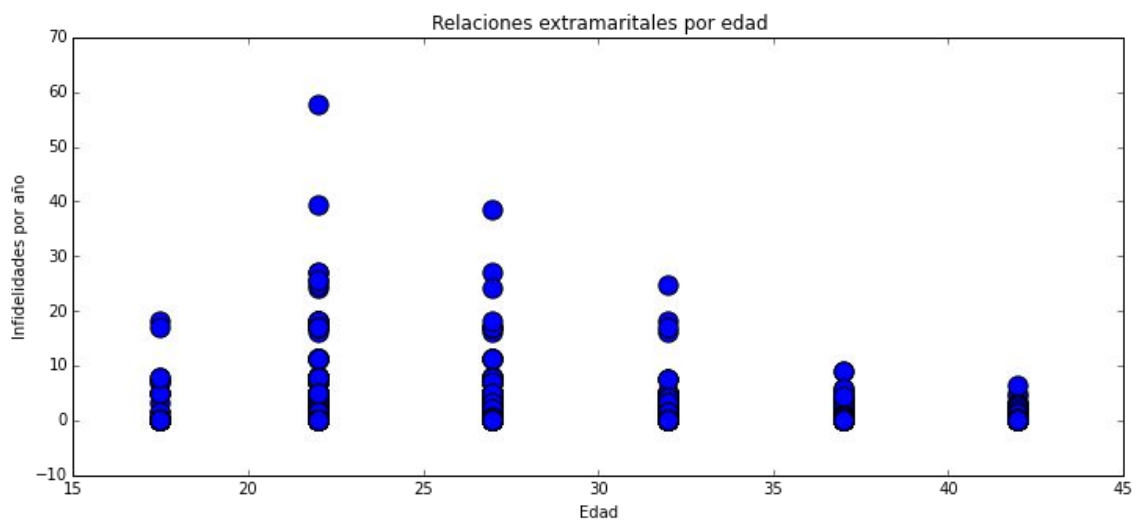


Figura 8

Al analizar los resultados graficados tenga en cuenta los siguientes valores para identificar las edades:

17.5 → Por debajo de 20;


22 → De 20 a 24;

27 → De 25 a 29;

32 → De 30 a 34;

37 → De 35 a 39;

42 → 40 o más.

Visualice en un gráfico de dispersión la relación entre el **nivel de educación de cada individuo** presente en el dataset (matriz **A**) y el **número de infidelidades por año**. Escriba en la celda la instrucción que aparece a continuación y de clic en . Visualice el resultado en la figura 9.

```
In [ ]: plt.figure(figsize=(12,5))
plt.title('Relaciones extramaritales por educación')
plt.xlabel(u'Nivel de educación')
plt.ylabel(u'Infidelidades por año')
plt.scatter(A[:,1],A[:,1],s=150)
```

Bloque de código 8

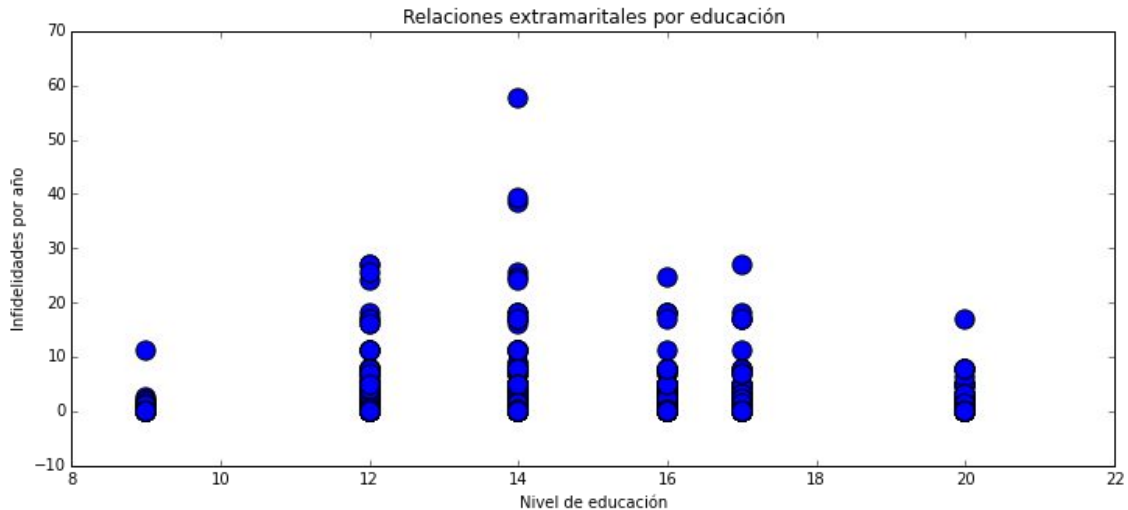



Figura9


Al analizar los resultados graficados tenga en cuenta los siguientes valores para identificar el nivel de educación de los individuos:

- 9 → Escuela primaria;
- 12 → Escuela secundaria;
- 14 → Educación superior (formal o no formal);
- 16 → Graduado de la universidad;
- 17 → Estudios de postgrado;
- 20 → Graduado de postgrado.

Para hacer un dendrograma o diagrama de árbol que agrupe los registros de acuerdo a las similitudes existentes entre sí (determinadas por la distancia entre los datos), se importan los siguientes módulos escribiendo en la celda la instrucción que aparece a continuación y de clic en .

```
In [38]: from scipy.cluster.hierarchy import dendrogram, linkage
```

Bloque de código 9

Para agrupar y graficar los los 100 primeros registros identificados por el número de individuo, es decir, la fila a la cual corresponde en el dataset, escriba en cada celda las instrucciones que aparecen a continuación y de clic en .

```
In [40]: Z = linkage(X,'ward')
```

Bloque de código 10

```
In [ ]: plt.figure(figsize=(12,5))
plt.title('Hierarchical clustering dendrogram')
plt.xlabel(u'Individuos')
plt.ylabel(u'Distancia')
dendrogram(Z, leaf_rotation=90, leaf_font_size= 9.5)
plt.show()
```

Bloque de código 11

El dendrograma jerárquico resultante deberá verse como el de la siguiente figura:

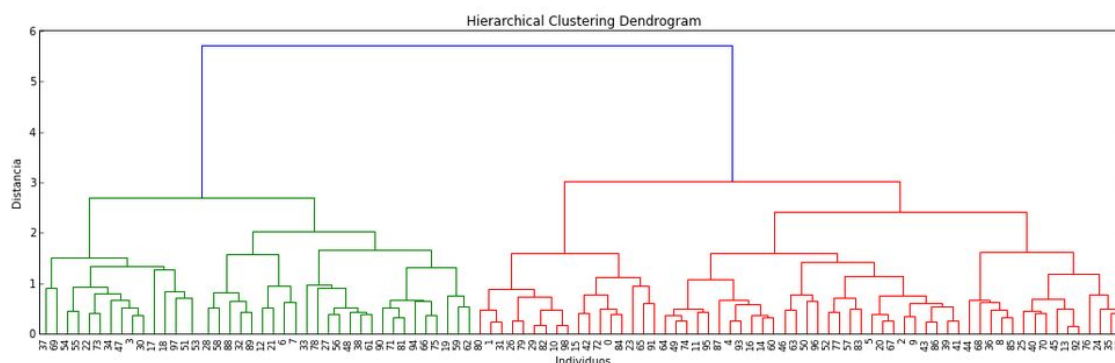


Figura 10

Compare los individuos 3 y 7 ambos presentes en el grupo de color verde y en diferente subgrupo. Analice sus diferencias y similitudes en variables tales como el número de infidelidades por año, la edad y el nivel de educación.

```
In [47]: print (u'\t\t Grupo \t Infidelidades por año \t Edad \t Nivel de educación \n')
print (u'\t\t Individuo 3: \t Verde', Y[3], '\t\t', A[3,1], '\t\t', A[3,5], '\n')
print (u'\t\t Individuo 7: \t Verde', Y[7], '\t\t', A[7,1], '\t\t', A[7,5], '\n')
```

Grupo	Infidelidades por año	Edad	Nivel de educación
Individuo 3:	Verde 0.7272727	37.0	16.0
Individuo 7:	Verde 1.826086	37.0	12.0

Bloque de código 12

Compare los individuos 6 y 7 ambos presentes en el grupo de color verde y en el mismo subgrupo. Analice sus diferencias y similitudes en variables tales como el número de infidelidades por año, la edad y el nivel de educación.



```
In [48]: print (u'\t\t Grupo \t Infidelidades por año \t Edad \t Nivel de educación \n')
print (u'\t\t Individuo 6: \t Verde', Y[6], '\t', A[6,1], '\t', A[6,5], '\n')
print (u'\t\t Individuo 7: \t Verde', Y[7], '\t', A[7,1], '\t', A[7,5], '\n')
```

	Grupo	Infidelidades por año	Edad	Nivel de educación
Individuo 6:	Verde	0.8521735	37.0	12.0
Individuo 7:	Verde	1.826086	37.0	12.0

### Bloque de código 13

Compare los individuos 1 y 99 ambos presentes en el grupo de color rojo pero en subgrupos diferentes. Analice sus diferencias y similitudes en variables tales como el número de infidelidades por año, la edad y el nivel de educación.

```
In [54]: print (u'\t\t Grupo \t Infidelidades por año \t Edad \t Nivel de educación \n')
print (u'\t\t Individuo 1: \t Rojo', Y[1], '\t', A[1,1], '\t', A[1,5], '\n')
print (u'\t\t Individuo 99: \t Rojo', Y[99], '\t', A[99,1], '\t', A[99,5], '\n')
```

	Grupo	Infidelidades por año	Edad	Nivel de educación
Individuo 1:	Rojo	3.2307692	27.0	14.0
Individuo 99:	Rojo	4.8999996	27.0	12.0

### Bloque de código 14

Compare los individuos 2 y 9 ambos presentes en el grupo de color rojo y en el mismo subgrupo. Analice sus diferencias y similitudes en variables tales como el número de infidelidades por año, la edad y el nivel de educación.

```
In [55]: print (u'\t\t Grupo \t Infidelidades por año \t Edad \t Nivel de educación \n')
print (u'\t\t Individuo 1: \t Rojo', Y[2], '\t', A[2,1], '\t', A[2,5], '\n')
print (u'\t\t Individuo 99: \t Rojo', Y[9], '\t', A[9,1], '\t', A[9,5], '\n')
```

	Grupo	Infidelidades por año	Edad	Nivel de educación
Individuo 1:	Rojo	1.3999996	22.0	16.0
Individuo 99:	Rojo	1.333333	27.0	16.0

### Bloque de código 15


Finalmente, compare los individuos 12 y 4 presentes en grupos diferentes (grupo de color verde y grupo de color rojo, respectivamente). Analice sus diferencias y similitudes en variables tales como el número de infidelidades por año, la edad y el nivel de educación.

```
In [16]: print (u'\t\t Grupo \t Infidelidades por año \t Edad \t Nivel de educación \n')
print (u'\t\t Individuo 2: \t Rojo', Y[2], '\t', A[2,1], '\t', A[2,5], '\n')
print (u'\t\t Individuo 9: \t Verde', Y[9], '\t', A[9,1], '\t', A[9,5], '\n')
```

	Grupo	Infidelidades por año	Edad	Nivel de educación
Individuo 2:	Rojo	1.3999996	22.0	16.0
Individuo 9:	Verde	1.333333	27.0	16.0

### Bloque de código 16

## 4. Guarde y cierre su Notebook

- Para guardar su Notebook de clic en  (recuadro señalado de color verde en la barra de herramientas de la **Figura 6**).
- Posteriormente de clic en **File** y luego en **Close and Halt** (recuadro señalado de color verde en la **Figura 7**).

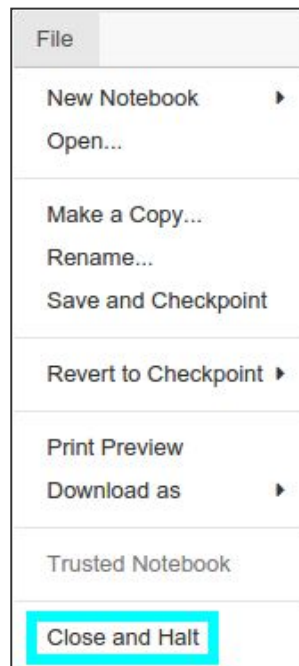


Figura 10

## 5. Desconectarse del servidor

- Para desconectarse del servidor, siga las instrucciones del ítem número 6 de la **Guía de conexión**.

---

**Fin de la Guía**