Analisis

February 3, 2017

Encuesta sobre lenguajes de programación y género

Introducción

una breve introducción de la idea con link a la encuesta. También cargamos los módulos que usaremos

Hace un tiempo en el chat de Télegram de OSL de la UGR (Oficina de Software libre de la Universidad) estuvimos hablando del género en el mundo de la programación y me surgió una duda.

¿Hay diferencias de género en el uso de los lenguajes de programación?

Me pareció muy interesante, así que cree un formulario en Google Docs.

La idea era ver como se distribuye el género en los lenguajes de programación. O al menos hacerme una idea ya que con este tipo de encuesta siempre se tienen muchos sesgos. Por ejemplo quién participa ya que al enviarla a personas amigas es muy probable que sean muy similares a mi y no muestren toda la diversidad que hay. También la forma en que está hecha seguro que tiene multiples fallos. Al menos limité a que te tengas que identificar para participar evitando un poco que alguien la haga varias veces.

A pesar de los posibles errores puede servir para hacerse una idea, una primera aproximación.

También he querido aprovechar para aprender Jupyter y su notebook. Así el análisis de los datos de la encuesta lo podéis reproducir todos y modificar a vuestro antojo, ¡¡¡¡viva la ciencia libre!!!!. Los datos los podéis hallar en esta hoja de cálculo de Google

Lo que veís a continuación es la parte donde cargo las librerías que voy a usar.

```
In [110]: # -*- coding: utf-8 -*-
          import requests
          import io
          import operator
          import  csv
          import pandas as pd
          import numpy as np
          import matplotlib.pyplot as plt
          import matplotlib.cm as cm
          import plotly.plotly as py
          import cufflinks as cf
          from scipy import stats
          cf.set_config_file(offline=False, world_readable=True, theme='ggplot')
          %matplotlib inline
```

Descripción de los datos obtenidos

La encuesta consistía en tres preguntas que se mostraban en un orden alatorio, para cada pregunta mostraremos un gráfico con los datos obtenidos. Primero cargamos los datos de la encuesta desde el google docs.

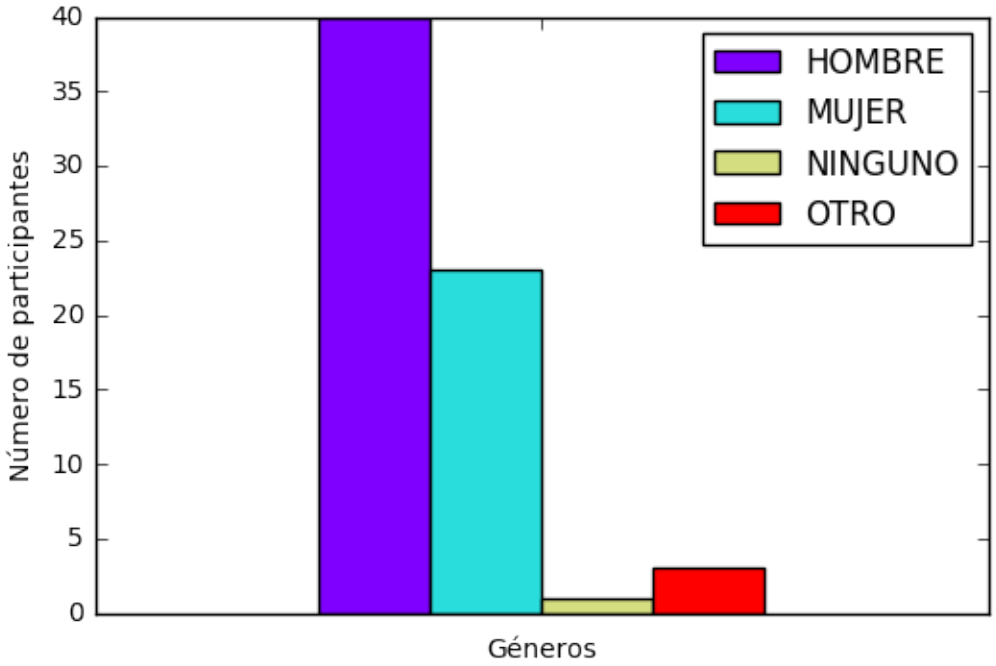
```
In [111]: headers={}
          headers["User-Agent"]= "Mozilla/5.0 (Windows NT 6.2; WOW64; rv:22.0) Geck
          headers["DNT"]= "1"
          headers["Accept"] = "text/html,application/xhtml+xml,application/xml;q=0.
          headers["Accept-Encoding"] = "deflate"
          headers["Accept-Language"]= "es-ES,en;q=0.5"
          headers["Content-Type"]= "application/x-www-form-urlencoded; charset=UTF-
          lines = []

          file_id="1xFpri9AF6N23vo6J5UUA_d5rY_oeVwpQ-EUBani0O7M"
          url = "https://docs.google.com/spreadsheets/d/{0}/export?format=csv".form

          r = requests.get(url)
          r.encoding ='utf-8'



          data = {}
          cols = []
          genero_lenguaje = {}
          residencia_lenguaje = {}
          lenguaje = {}
          sio = io.StringIO( r.text, newline=None)
          reader = csv.reader(sio, dialect=csv.excel)
          rownum = 0

          for row in reader:
              if rownum == 0:
                  for col in row:
                      col=col.upper();
                      col=col.strip();
                      data[col] = ''
                      cols.append(col)


              else:
                  orden_columnas=[0,1,3,2]
                  for i in orden_columnas:
                      col=row[i];
                      col=col.upper()
                      col=col.strip();
                      data[cols[i]] = col
                      if (i==1):
                          if (not col in genero_lenguaje):
```

```
                        genero_lenguaje[col]={}
                elif (i==2):
                    genero=data[cols[i-1]]
                    residencia=data[cols[i+1]]
                    if (col in genero_lenguaje[genero]):
                        genero_lenguaje[genero][col]+=1
                    else:
                        genero_lenguaje[genero][col]=1;
                    if (col in residencia_lenguaje[residencia]):
                        residencia_lenguaje[residencia][col]+=1
                    else:
                        residencia_lenguaje[residencia][col]=1;
                    if (not col in lenguaje):
                        lenguaje[col]=1;
                    else:
                        lenguaje[col]+=1;
                elif (i==3):
                    if (not col in residencia_lenguaje):
                        residencia_lenguaje[col]={}

        rownum = rownum + 1

generos=genero_lenguaje.keys();
residencias=residencia_lenguaje.keys();
lenguajes=lenguaje.keys();
#ordenamos alfabéticamente
generos=sorted(generos)
residencias=sorted(residencias)
lenguajes=sorted(lenguajes)
```

Género

Las posibilidades de género eran Hombre, Mujer, Otro, Ninguno. En la encuesta las opciones se mostraban en un orden aleatorio.

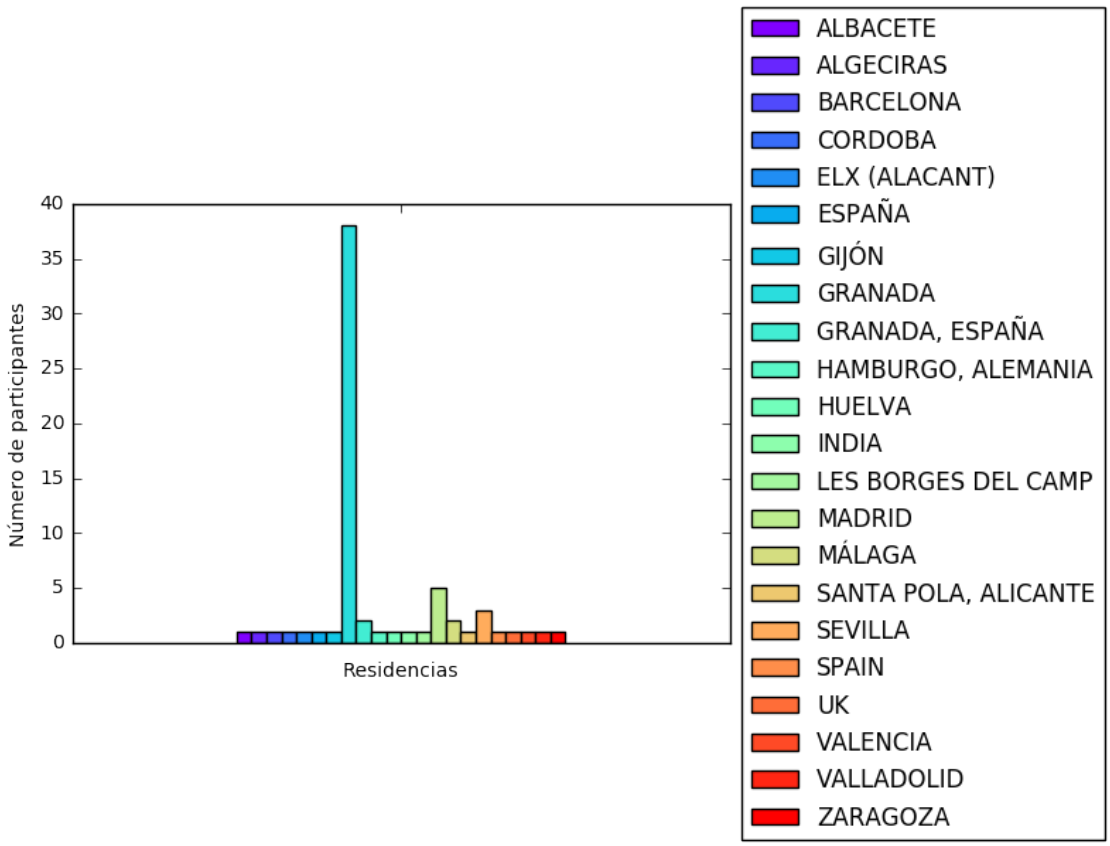
```
In [112]: #gráfico de barras de cada genero
          cantidad_genero={};
          colors = cm.rainbow(np.linspace(0, 1, len(generos)))
          for g in generos:
              datas=genero_lenguaje[g];
              total=sum(datas.values());
              cantidad_genero[g]=total;
          df = pd.DataFrame(data=cantidad_genero,index=[""],columns=generos)
          df.head()
          genero_plot = df.plot(kind='bar',label='Gráfico de barras de cada genero'
          genero_plot.set_ylabel("Número de participantes")
          genero_plot.set_xlabel("Géneros")
          pass
```



Residencia

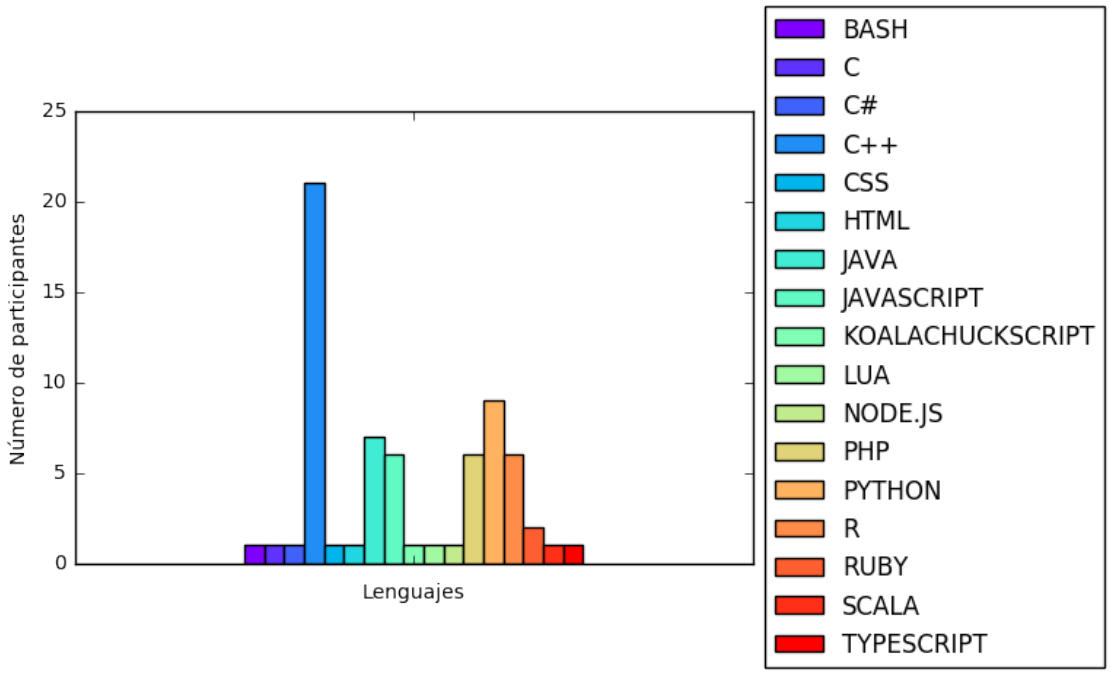
Esta pregunta era demasiado abierta, así que hay personas que han respondido su ciudad o su país. No he filtrado ni he agrupado lugares. Son datos en bruto.

```
In [113]: #gráfico de barras de cada residencia
          cantidad_residencia={};
          colors = cm.rainbow(np.linspace(0, 1, len(residencias)))
          for g in residencias:
              datas=residencia_lenguaje[g];
              total=sum(datas.values());
              cantidad_residencia[g]=total;
          df = pd.DataFrame(data=cantidad_residencia,index=[""],columns=residencias
          df.head()
          residencia_plot = df.plot(kind='bar',label='Gráfico de barras de cada res
          residencia_plot.set_ylabel("Número de participantes")
          residencia_plot.set_xlabel("Residencias")
          residencia_plot.legend(loc='center left', bbox_to_anchor=(1.0, 0.5))
          pass
```



Lenguajes de programación
Aquí no he agrupado lenguajes que podrían ser el mismo como Node.js y javascript.

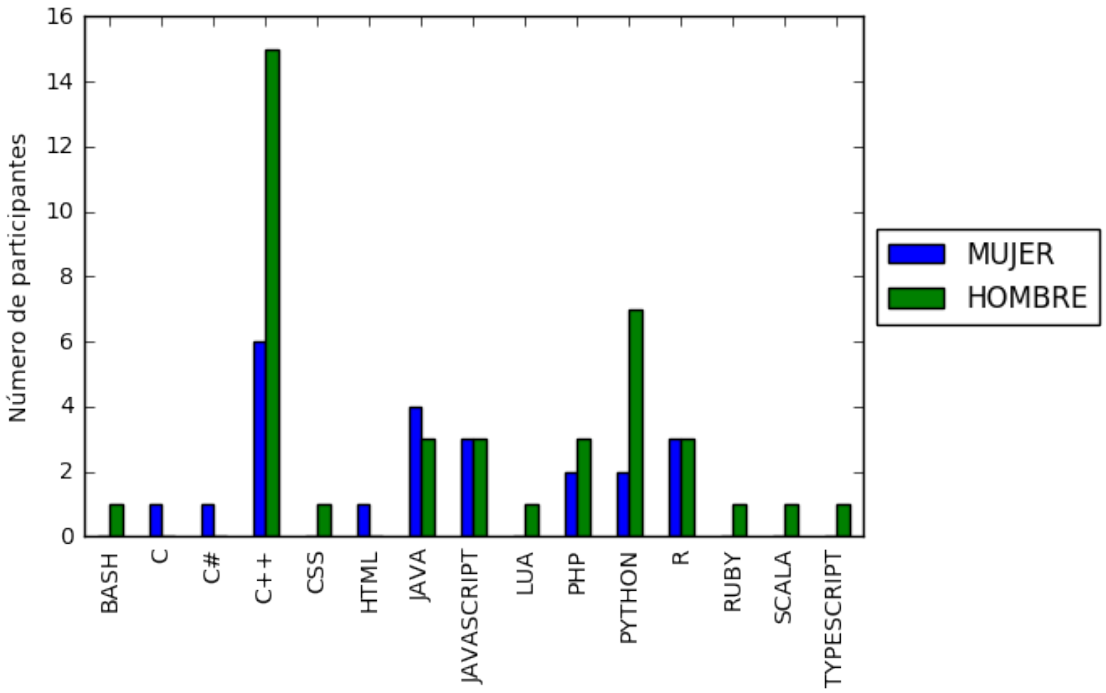
```
In [114]: #gráfico de barras de lenguajes de programación
          colors = cm.rainbow(np.linspace(0, 1, len(lenguajes)))
          df = pd.DataFrame(data=lenguaje,index=[""],columns=lenguajes)
          df.head()
          lengua_plot = df.plot(kind='bar',label='Gráfico de barras de cada lengua
          lengua_plot.set_ylabel("Número de participantes")
          lengua_plot.set_xlabel("Lenguajes")
          lengua_plot.legend(loc='center left', bbox_to_anchor=(1.0, 0.5))
          pass
```



Género y lenguajes

Vamos al tema, primero podemos comparar cuantos hombres y mujeres han respondido para cada lenguaje. Así podemos, a simple vista, hacer una comparativa. No miro las otras opciones de género porque hay muy poca muestra.

```
In [115]: df = pd.DataFrame(data=genero_lenguaje,columns=['MUJER','HOMBRE'])
          df.head()
          plot_lenguaje_genero = df.plot(kind='bar')
          plot_lenguaje_genero.set_ylabel("Número de participantes")
          plot_lenguaje_genero.legend(loc='center left', bbox_to_anchor=(1.0, 0.5))
          pass
```



Así, a primera vista, y si los datos no cambian, solo en 6 lenguajes hay de ambos géneros. Podemos ver una gráfica de cada género de forma separada para los lenguajes donde hay de los dos. Para eso buscamos los lenguajes comunes y nos quedamos con los géneros que tengan más de 1 en común.

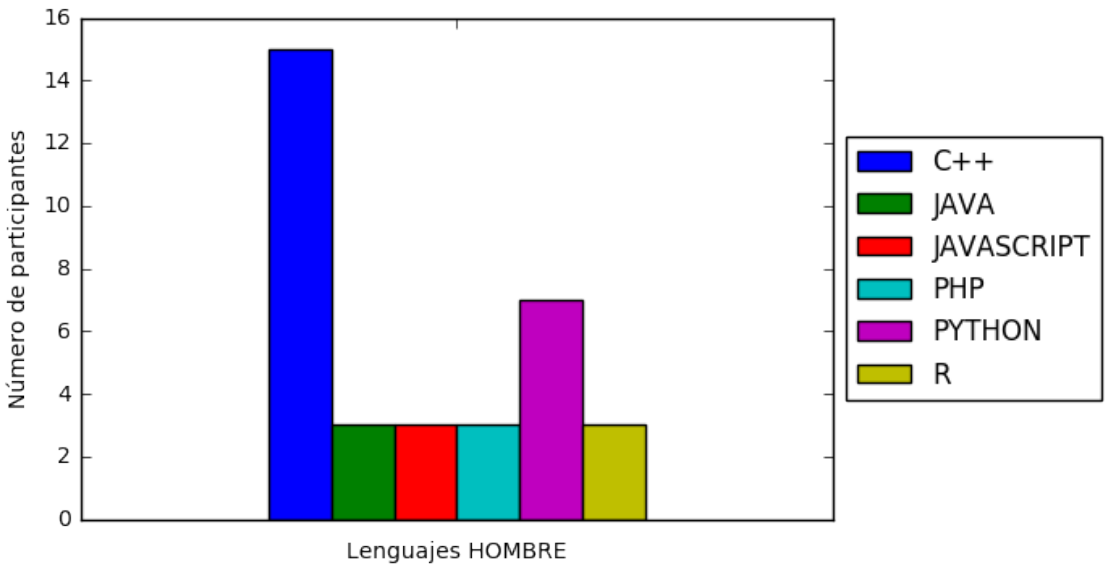
```
In [116]: genero_ambos_lenguaje={}
          for g1 in generos:
              lenguajes_g1=set(genero_lenguaje[g1].keys())
              comunes={}
              for g2 in generos:
                  if (g1!=g2):
                      lenguajes_g2=set(genero_lenguaje[g2].keys());
                      encomun=sorted(list(lenguajes_g1&lenguajes_g2));
                      if (len(encomun)>1):
                          comunes[g2]=encomun;
              if (len(comunes)>0):
                  genero_ambos_lenguaje[g1]=comunes;
          personas_genero_ambos_lenguaje={}

          for g1 in genero_ambos_lenguaje.keys():
              datas_g1=genero_ambos_lenguaje[g1];
              for g2 in datas_g1.keys():
                  lenguajes_g2=datas_g1[g2];
                  lenguaje={}
                  for l in genero_lenguaje[g1].keys():
```

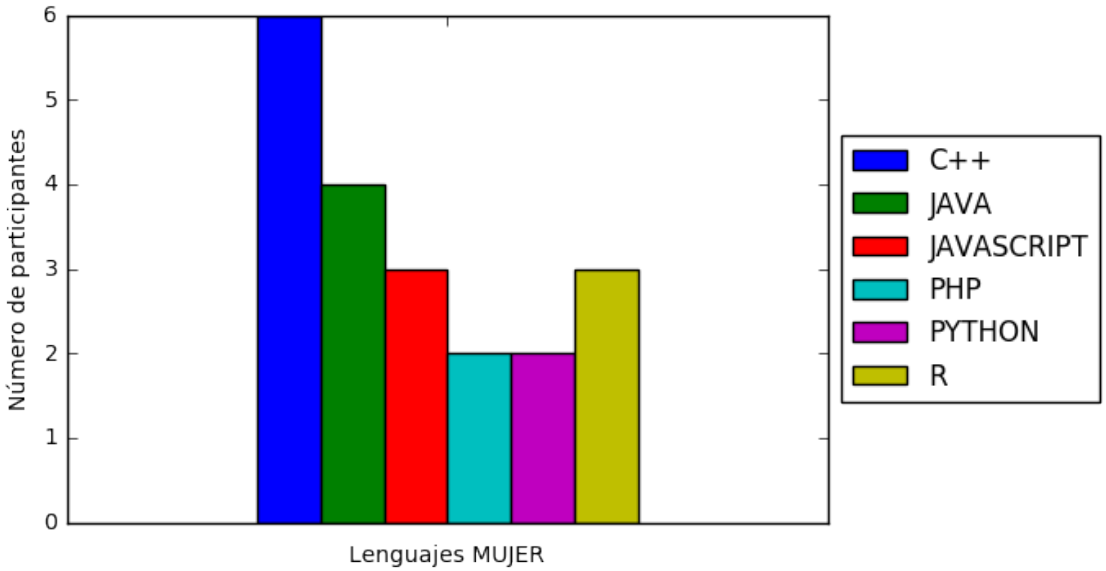
```
                    if (l in lenguajes_g2):
                        lenguaje[l]=genero_lenguaje[g1][l]
                nombres_lenguaje_ordenado=sorted(lenguaje.keys())

                personas_genero_ambos_lenguaje[g1]=lenguaje
        print(genero_ambos_lenguaje)
        print(personas_genero_ambos_lenguaje)
```

```
{'HOMBRE': {'MUJER': ['C++', 'JAVA', 'JAVASCRIPT', 'PHP', 'PYTHON', 'R']}, 'MUJER':
{'HOMBRE': {'R': 3, 'C++': 15, 'PYTHON': 7, 'JAVA': 3, 'PHP': 3, 'JAVASCRIPT': 3},
```

Podemos ver que tenemos un diccionario donde la información del nombre de los lenguajes es la misma en ambos elementos, redundancia que le llaman. Podría quedarme con uno de ellos pero de momento los dejo. A continuación hago una gráfica para cada género.

```
In [117]: for g in sorted(personas_genero_ambos_lenguaje.keys()):
              lenguajes_comun=sorted(personas_genero_ambos_lenguaje[g])
              df = pd.DataFrame(data=personas_genero_ambos_lenguaje[g],index=[""],c
              df.head()
              plot_lenguaje_genero_comun = df.plot(kind='bar')
              plot_lenguaje_genero_comun.set_xlabel("Lenguajes "+g)
              plot_lenguaje_genero_comun.set_ylabel("Número de participantes")
              plot_lenguaje_genero_comun.legend(loc='center left', bbox_to_anchor=
          pass
```





A simple vista parece que ambas distribuciones de datos son diferentes. Cabe preguntarse si realemnte lo són. Para eso hay varios test que pueden pasarse. Cualquier sugerencia es bien recibida.

Conclusión

Con los datos que se observan parece ser que si hay diferencias a la hora de elegir un lenguaje según tu género sea otro. Las causas se deberían mirar en otros estudios mejor hechos (Gender,2016) ya que este es demasiado simple y puede presentar multitud de sesgos.

Bibliografía

En el fondo este es un tema bastante estudiado si os pasais por google

[1] [Gender ratios of programmers, by language](http://blog.revolutionanalytics.com/2016/06/programmer gender.html)

```
In [ ]:
```