

1lnpm0f4w

January 23, 2025

```
[16]: import pandas as pd
import numpy as np
```

```
[18]: # rutas
p1 = "D:/one_drive_unal/OneDrive - Universidad Nacional de Colombia/2025/
      prueba_JEP/Prueba_Tecnica/input/"
p2 = "C:/Users/IVAN CORREDOR/CODE_PYTHON/PRUEBA_GRAI/output/"
```

```
[19]: # cargar los datos
entidad_a = pd.read_excel(p1 + "A.xlsx", decimal=",")
entidad_b = pd.read_excel(p1 + "B.xlsx", decimal=",")
```

```
[20]: entidad_a.head()
```

```
[20]:
```

	NOMBRE1	NOMBRE2	APELLIDO1	APELLIDO2	DEPARTAMENTO	MUNICIPIO \
0	JOSE	JAVIER	HURTADO	VERGARA	CAQUETA	CARTAGENA DEL CHAIRA
1	ISMAEL	ALBERTO	OROZCO	RAMIREZ	RISARALDA	PEREIRA
2	CARLOS	ALFONSO	FLOREZ	NaN	CAQUETA	CARTAGENA DEL CHAIRA
3	EFREM	NaN	SALDARRIAGA	NaN	ANTIOQUIA	PUERTO BERRIO
4	JHON	JAIR	RESTREPO	AGUILAR	ANTIOQUIA	URAMITA

	FECHA_HECHOS	COD_MUNICIPIO	NUMERO_DOCUMENTO	EDAD	SEXO
0	1998-03-03	18150	10001293	22.0	HOMBRE
1	1996-08-08	66001	10001366	19.0	HOMBRE
2	1998-03-03	18150	10001916	21.0	MUJER
3	2004-06-01	5579	10004010	NaN	HOMBRE
4	2000-04-03	5842	10007592	25.0	HOMBRE

```
[46]: entidad_a.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10 entries, 0 to 9
Data columns (total 11 columns):
#   Column          Non-Null Count  Dtype
---  -
0   NOMBRE1         10 non-null    object
1   NOMBRE2         9 non-null     object
2   APELLIDO1       10 non-null    object
```

```

3  APELLIDO2          7 non-null    object
4  DEPARTAMENTO       10 non-null    object
5  MUNICIPIO         10 non-null    object
6  FECHA_HECHOS      10 non-null    datetime64[ns]
7  COD_MUNICIPIO     10 non-null    int64
8  NUMERO_DOCUMENTO  10 non-null    int64
9  EDAD              6 non-null     float64
10 SEXO              10 non-null    object
dtypes: datetime64[ns](1), float64(1), int64(2), object(7)
memory usage: 1012.0+ bytes

```

```
[24]: entidad_b.head()
```

```

[24]:      NOMBRE1  NOMBRE2 APELLIDO1 APELLIDO2 DEPARTAMENTO MUNICIPIO \
0      JOSE    JAVIER  HURTADO   VERGARA    CAQUETA  CARTAGENA DEL CHAIRA
1      CARLOS  ALFONSO   LOPEZ      NaN    CAQUETA      LA MONTANITA
2      CARLOS  ALFONSO   FLOREZ      NaN    CAQUETA  CARTAGENA DEL CHAIRA
3  ALEXANDER      NaN  PORTELA    OZUNA      SUCRE      SAN ONOFRE
4      DIEGO  LEONARDO  GONZALEZ  RAMIREZ  ANTIOQUIA      COCORNA

      FECHA_HECHOS  COD_MUNICIPIO  NUMERO_DOCUMENTO  EDAD  SEXO
0  1998-03-01          18150          10001293  22.0  HOMBRE
1  1998-03-01          18410          10001916  21.0  MUJER
2  1998-03-03          18150          10001916  21.0  MUJER
3  2000-11-08          70713          1000588161   NaN  HOMBRE
4  2003-01-01           5197          1001663514  10.0  HOMBRE

```

```
[52]: entidad_b.info()
```

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10 entries, 0 to 9
Data columns (total 11 columns):
#   Column                Non-Null Count  Dtype
---  -
0   NOMBRE1                10 non-null    object
1   NOMBRE2                9 non-null     object
2   APELLIDO1              10 non-null    object
3   APELLIDO2              8 non-null     object
4   DEPARTAMENTO            10 non-null    object
5   MUNICIPIO              10 non-null    object
6   FECHA_HECHOS           10 non-null    datetime64[ns]
7   COD_MUNICIPIO          10 non-null    int64
8   NUMERO_DOCUMENTO       10 non-null    int64
9   EDAD                   9 non-null     float64
10  SEXO                   10 non-null    object
dtypes: datetime64[ns](1), float64(1), int64(2), object(7)
memory usage: 1012.0+ bytes

```

```
[54]: # crear variable FUENTE
entidad_a["FUENTE"] = "A"
entidad_b["FUENTE"] = "B"
```

```
[62]: entidad_a.duplicated().sum()
```

```
[62]: 0
```

```
[94]: # crear llave única en A
entidad_a["IDENTIFICADOR"] = entidad_a[['NOMBRE1', 'NOMBRE2', 'APELLIDO1',
↳ 'APELLIDO2', 'NUMERO_DOCUMENTO']].apply(
    #concatena las variables de identificación y crea un valor de 4 dígitos
    lambda row: 1000 + abs(hash('_'.join(row.dropna().astype(str)))) % 9000,
↳ axis=1
)
```

```
[26]: entidad_a.head()
```

```
[26]:
```

	NOMBRE1	NOMBRE2	APELLIDO1	APELLIDO2	DEPARTAMENTO	MUNICIPIO \
0	JOSE	JAVIER	HURTADO	VERGARA	CAQUETA	CARTAGENA DEL CHAIRA
1	ISMAEL	ALBERTO	OROZCO	RAMIREZ	RISARALDA	PEREIRA
2	CARLOS	ALFONSO	FLOREZ	NaN	CAQUETA	CARTAGENA DEL CHAIRA
3	EFREM	NaN	SALDARRIAGA	NaN	ANTIOQUIA	PUERTO BERRIO
4	JHON	JAIR	RESTREPO	AGUILAR	ANTIOQUIA	URAMITA

	FECHA_HECHOS	COD_MUNICIPIO	NUMERO_DOCUMENTO	EDAD	SEXO
0	1998-03-03	18150	10001293	22.0	HOMBRE
1	1996-08-08	66001	10001366	19.0	HOMBRE
2	1998-03-03	18150	10001916	21.0	MUJER
3	2004-06-01	5579	10004010	NaN	HOMBRE
4	2000-04-03	5842	10007592	25.0	HOMBRE

```
[96]: # crear llave única en B
entidad_b["IDENTIFICADOR"] = entidad_b[['NOMBRE1', 'NOMBRE2', 'APELLIDO1',
↳ 'APELLIDO2', 'NUMERO_DOCUMENTO']].apply(
    #concatena las variables de identificación y crea un valor de 4 dígitos
    lambda row: 1000 + abs(hash('_'.join(row.dropna().astype(str)))) % 9000,
↳ axis=1
)
```

```
[28]: ## append
tabla_c = pd.concat([entidad_a, entidad_b], axis = 0).reset_index(drop=True)
```

```
[110]: tabla_c["IDENTIFICADOR"].value_counts()
```

```
[110]: IDENTIFICADOR
      9893      2
      7169      2
      2271      2
      9902      2
      2389      1
      2068      1
      4849      1
      6078      1
      5908      1
      9407      1
      3735      1
      8964      1
      1099      1
      4502      1
      5711      1
      6730      1
      Name: count, dtype: int64
```

```
[160]: # generar tabla resumen
resumen = tabla_c.groupby(["IDENTIFICADOR"])["FUENTE"].size().
    ↪reset_index(name="CANTIDAD_FUENTES").sort_values(by="CANTIDAD_FUENTES",
    ↪ascending =False).reset_index(drop=True)
resumen
```

```
[160]: IDENTIFICADOR  CANTIDAD_FUENTES
0          2271          2
1          7169          2
2          9893          2
3          9902          2
4          1099          1
5          2068          1
6          2389          1
7          3735          1
8          4502          1
9          4849          1
10         5711          1
11         5908          1
12         6078          1
13         6730          1
14         8964          1
15         9407          1
```

```
[209]: import seaborn as sns
import matplotlib.pyplot as plt
# Generar la gráfica de barras
a = sns.countplot(x="IDENTIFICADOR", data= tabla_c)
```

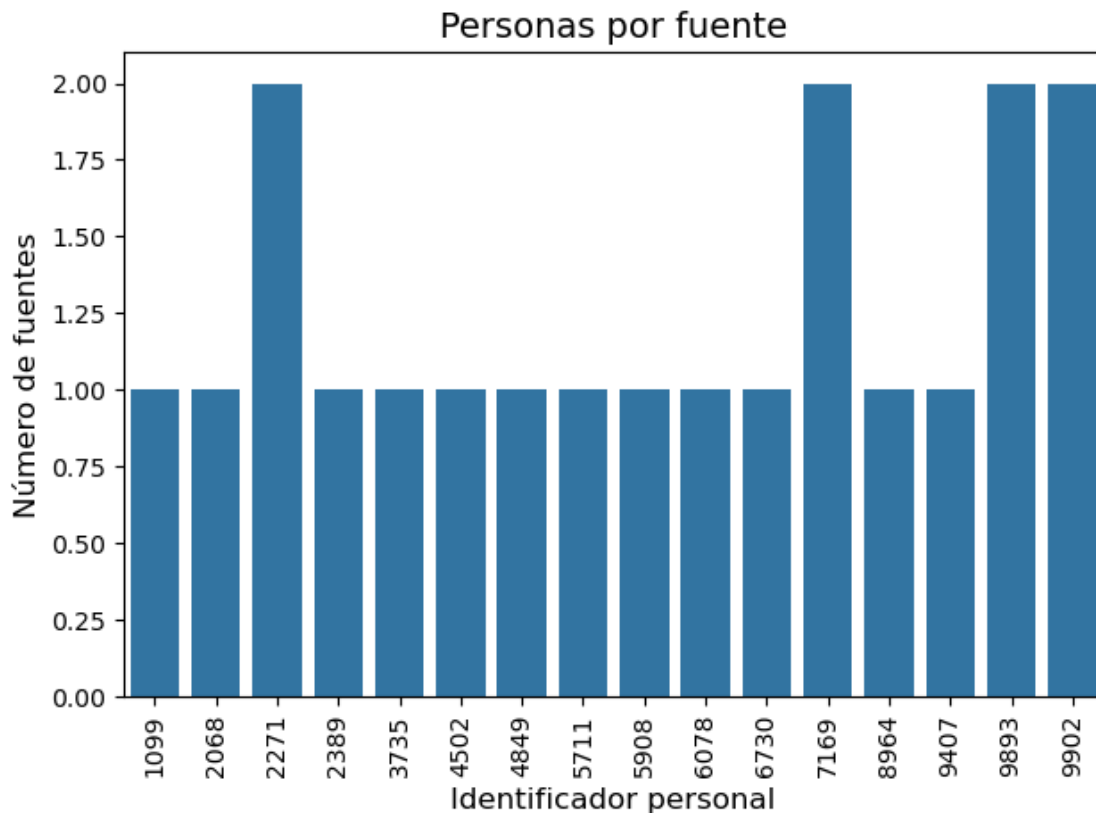
```

# Configurar títulos y etiquetas
plt.xlabel("Identificador personal", fontsize=12) # Título del eje X
plt.ylabel("Número de fuentes", fontsize=12)      # Título del eje Y
plt.title("Personas por fuente", fontsize=14)    # Título del gráfico

# Rotar etiquetas del eje X
plt.xticks(rotation=90, ha='center', fontsize=10)

# Mostrar el gráfico
plt.tight_layout()
# exportar la gráfica a pdf
plt.savefig(p1 + "personas_fuente.png")
plt.show()

```



```
[184]: tabla_c.info()
```

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20 entries, 0 to 19
Data columns (total 13 columns):
#   Column          Non-Null Count  Dtype

```

```

---  -----  -----  -----
0  NOMBRE1      20 non-null  object
1  NOMBRE2      18 non-null  object
2  APELLIDO1    20 non-null  object
3  APELLIDO2    15 non-null  object
4  DEPARTAMENTO  20 non-null  object
5  MUNICIPIO    20 non-null  object
6  FECHA_HECHOS 20 non-null  datetime64[ns]
7  COD_MUNICIPIO 20 non-null  int64
8  NUMERO_DOCUMENTO 20 non-null int64
9  EDAD         15 non-null  float64
10 SEXO         20 non-null  object
11 FUENTE       20 non-null  object
12 IDENTIFICADOR 20 non-null  int64
dtypes: datetime64[ns](1), float64(1), int64(3), object(8)
memory usage: 2.2+ KB

```

## 1 explicación de los resultados

Al combinar las fuentes A y B se obtiene un total de 20 registros. A nivel de personas, se identifican 16 personas únicas, de las cuales 4 se encuentran en las fuentes A y B. Los nombres de estas persona son: JOSE JAVIER HURTADO VERGARA, CARLOS ALFONSO FLOREZ, ROBERTO CARLOS CASTRO ALBORNOZ y JULIAN CAMILO HINCAPIEHINCAPIE.