Análisis de datos de los resultados Prueba Saber Pro 2023

El objetivo de este proyecto es realizar un análisis que permita identificar el comportamiento de los resultados de las pruebas saber pro para el año 2023. Se emplean herramientas y paquetes para el análisis de datos como Pandas, numpy, Matplotlib y Seaborn.

Librerias

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
```

Datos

Se monta el modulo de drive para leer el archivo.

```
from google.colab import drive
drive.mount('/content/drive')
ruta_data = '/content/drive/MyDrive/data_icfes_saber_pro_2023.TXT'
```

```
Drive already mounted at /content/drive; to attempt to forcibly
remount, call drive.mount("/content/drive", force_remount=True).
```

La base de datos utilizada, se encuentra disponible en el portal del Instituto Colombiano para la Evaluación de la Educación. El archivo se guardó en google drive, y cuenta con 100 variables, dentro de las que vale la pena mencionar el genero del o la estudiante, el municipio y departamento de residencia, así como el municipio y departamento de presentación del examen. Además, otras variables como el caracter y origen de las instituciones de educación superior en las que estudia cada uno de los que presenta la prueba, y logicamente los resultados individuales para cada prueba de cada uno de los estudiantes, sus niveles de desempeño, los percentiles correspondientes al nivel nacional y a su propio nucleo basico del conocimiento.

```
df_saber_pro = pd.read_table(ruta_data, sep='¬')
```

```
<ipython-input-3-c8e8c3eb2e0c>:1: ParserWarning: Falling back to the
'python' engine because the separator encoded in utf-8 is > 1 char
long, and the 'c' engine does not support such separators; you can
avoid this warning by specifying engine='python'.
  df_saber_pro = pd.read_table(ruta_data, sep='¬')
```

```
df_saber_pro
```

```
{"type":"dataframe","variable_name":"df_saber_pro"}
```

Explorando los datos

```
# El df cuenta con las siguientes observaciones y variables.
print(df_saber_pro.shape)
```

```
(139288, 100)
```

```
# Las 100 variables, o columnas, del df son:
df_saber_pro.columns
```

```
Index(['ESTU_TIPODOCUMENTO', 'ESTU_NACIONALIDAD', 'ESTU_GENERO',
       'ESTU_FECHANACIMIENTO', 'ESTU_EXTERIOR', 'PERIODO',
'ESTU_CONSECUTIVO',
       'ESTU_ESTUDIANTE', 'ESTU_PAIS_RESIDE', 'ESTU_DEPTO_RESIDE',
       'ESTU_COD_RESIDE_DEPTO', 'ESTU_MCPIO_RESIDE',
'ESTU_COD_RESIDE_MCPIO',
       'ESTU_AREARESIDE', 'ESTU_ESTADOCIVIL',
'ESTU_TITULOOBTENIDOBACHILLER',
       'ESTU_VALORMATRICULAUNIVERSIDAD', 'ESTU_PAGOMATRICULABECA',
       'ESTU_PAGOMATRICULACREDITO', 'ESTU_PAGOMATRICULAPADRES',
       'ESTU_PAGOMATRICULAPROPIO', 'ESTU_COMOCAPACITOEXAMENSB11',
       'ESTU_CURSODOCENTESIES', 'ESTU_CURSOIESAPOYOEXTERNO',
       'ESTU_CURSOIESEXTERNA', 'ESTU_SIMULACROTIPOICFES',
       'ESTU_ACTIVIDADREFUERZOAREAS', 'ESTU_ACTIVIDADREFUERZOGENERIC',
       'ESTU_TIPODOCUMENTOSB11', 'ESTU_SEMESTRECURSA',
'FAMI_EDUCACIONPADRE',
       'FAMI_EDUCACIONMADRE', 'FAMI_OCUPACIONPADRE',
'FAMI_OCUPACIONMADRE',
       'FAMI_ESTRATOVIVIENDA', 'FAMI_TIENEINTERNET',
'FAMI_TIENECOMPUTADOR',
       'FAMI_TIENELAVADORA', 'FAMI_TIENEHORNOMICROOGAS',
       'FAMI_TIENESERVICIOTV', 'FAMI_TIENEAUTOMOVIL',
'FAMI_TIENEMOTOCICLETA',
       'FAMI_TIENECONSOLAVIDEOJUEGOS', 'FAMI_TRABAJOLABORPADRE',
       'FAMI_TRABAJOLABORMADRE', 'ESTU_HORASSEMANATRABAJA',
       'FAMI_CUANTOSCOMPARTEBAÑO', 'ESTU_PAGOMATRICULA',
       'ESTU_VLRULTIMOSEMESCURSADO', 'ESTU_PRESENTACIONCASA',
       'ESTU_PRESENTACIONSABADO', 'INST_COD_INSTITUCION',
       'INST_NOMBRE_INSTITUCION', 'ESTU_PRGM_ACADEMICO',
       'ESTU_SNIES_PRGMACADEMICO', 'GRUPOREFERENCIA',
'ESTU_PRGM_CODMUNICIPIO',
       'ESTU_PRGM_MUNICIPIO', 'ESTU_PRGM_DEPARTAMENTO',
       'ESTU_NIVEL_PRGM_ACADEMICO', 'ESTU_METODO_PRGM',
'ESTU_NUCLEO_PREGRADO',
       'ESTU_NUCLEO_PREGRADO_1', 'ESTU_INST_CODMUNICIPIO',
       'ESTU_INST_MUNICIPIO', 'ESTU_INST_DEPARTAMENTO',
       'INST_CARACTER_ACADEMICO', 'INST_ORIGEN',
'ESTU_PRIVADO_LIBERTAD',
       'ESTU_COD_MCPIO_PRESENTACION', 'ESTU_MCPIO_PRESENTACION',
       'ESTU_DEPTO_PRESENTACION', 'ESTU_COD_DEPTO_PRESENTACION',
```

```
       'MOD_RAZONA_CUANTITAT_PUNT', 'MOD_RAZONA_CUANTITAT_DESEM',
       'MOD_RAZONA_CUANTITATIVO_PNAL', 'MOD_RAZONA_CUANTITATIVO_PNBC',
       'MOD_LECTURA_CRITICA_PUNT', 'MOD_LECTURA_CRITICA_DESEM',
       'MOD_LECTURA_CRITICA_PNAL', 'MOD_LECTURA_CRITICA_PNBC',
       'MOD_COMPETEN_CIUDADA_PUNT', 'MOD_COMPETEN_CIUDADA_DESEM',
       'MOD_COMPETEN_CIUDADA_PNAL', 'MOD_COMPETEN_CIUDADA_PNBC',
       'MOD_INGLES_PUNT', 'MOD_INGLES_DESEM', 'MOD_INGLES_PNAL',
       'MOD_INGLES_PNBC', 'MOD_COMUNI_ESCRITA_PUNT',
       'MOD_COMUNI_ESCRITA_DESEM', 'MOD_COMUNI_ESCRITA_PNAL',
       'MOD_COMUNI_ESCRITA_PNBC', 'PUNT_GLOBAL', 'PERCENTIL_GLOBAL',
       'PERCENTIL_NBC', 'ESTU_INSE_INDIVIDUAL', 'ESTU_NSE_INDIVIDUAL',
       'ESTU_NSE_IES', 'ESTU_ESTADOINVESTIGACION'],
      dtype='object')
```

```
# Primeras cinco observaciones del df
df_saber_pro.head()
```


```
{"type":"dataframe","variable_name":"df_saber_pro"}
```


```
# Últimas cinco observaciones del df
df_saber_pro.tail()
```


```
{"type":"dataframe"}
```


```
# Dada la cantidad de filas y columnas, no es posible visualizar el
dataframe
# completo. Pasando el argumento de longitud a la función head, el
resultado
# es la vista de las 5 primeras y últimas observaciones, y de las 10
primeras
# y últimas variables.
df_saber_pro.head(len(df_saber_pro))
```


```
{"type":"dataframe","variable_name":"df_saber_pro"}
```


```
# Los valores NA existentes en variables como las relacionadas con la
# preparación previa a la prueba no es posible imputarlos ya que
# no son propiamente datso faltantes, sino que una opción de respuesta
# era NA.
df_saber_pro.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 139288 entries, 0 to 139287
Data columns (total 100 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   ESTU_TIPODOCUMENTO              139288 non-null  object
 1   ESTU_NACIONALIDAD               139288 non-null  object
 2   ESTU_GENERO                     139288 non-null  object
 3   ESTU_FECHANACIMIENTO            139288 non-null  object
 4   ESTU_EXTERIOR                   139288 non-null  object
```

```
5   PERIODO                          139288 non-null  int64
6   ESTU_CONSECUTIVO                 139288 non-null  object
7   ESTU_ESTUDIANTE                  139288 non-null  object
8   ESTU_PAIS_RESIDE                 139288 non-null  object
9   ESTU_DEPTO_RESIDE                139288 non-null  object
10  ESTU_COD_RESIDE_DEPTO            139288 non-null  int64
11  ESTU_MCPIO_RESIDE                139288 non-null  object
12  ESTU_COD_RESIDE_MCPIO            139288 non-null  int64
13  ESTU_AREARESIDE                  136582 non-null  object
14  ESTU_ESTADOCIVIL                 138965 non-null  object
15  ESTU_TITULOOBTENIDOBACHILLER     136515 non-null  object
16  ESTU_VALORMATRICULAUNIVERSIDAD   136515 non-null  object
17  ESTU_PAGOMATRICULABECA           136472 non-null  object
18  ESTU_PAGOMATRICULACREDITO        136467 non-null  object
19  ESTU_PAGOMATRICULAPADRES         136477 non-null  object
20  ESTU_PAGOMATRICULAPROPIO         136483 non-null  object
21  ESTU_COMOCAPACITOEXAMENSB11      136577 non-null  object
22  ESTU_CURSODOCENTESIES            7718 non-null    object
23  ESTU_CURSOIESAPOYOEXTERNO        7715 non-null    object
24  ESTU_CURSOIESEXTERNA             7714 non-null    object
25  ESTU_SIMULACROTIPOICFES          7719 non-null    object
26  ESTU_ACTIVIDADREFUERZOAREAS      7719 non-null    object
27  ESTU_ACTIVIDADREFUERZOGENERIC    7718 non-null    object
28  ESTU_TIPODOCUMENTOSB11           135151 non-null  object
29  ESTU_SEMESTRECURSA               136515 non-null  object
30  FAMI_EDUCACIONPADRE              136377 non-null  object
31  FAMI_EDUCACIONMADRE              136374 non-null  object
32  FAMI_OCUPACIONPADRE              136580 non-null  object
33  FAMI_OCUPACIONMADRE              136580 non-null  object
34  FAMI_ESTRATOVIVIENDA             134207 non-null  object
35  FAMI_TIENEINTERNET               136417 non-null  object
36  FAMI_TIENECOMPUTADOR             136135 non-null  object
37  FAMI_TIENELAVADORA               136119 non-null  object
38  FAMI_TIENEHORNOMICROOGAS         136052 non-null  object
39  FAMI_TIENESERVICIOTV             136331 non-null  object
40  FAMI_TIENEAUTOMOVIL              135982 non-null  object
41  FAMI_TIENEMOTOCICLETA            136043 non-null  object
42  FAMI_TIENECONSOLAVIDEOJUEGOS     135445 non-null  object
43  FAMI_TRABAJOLABORPADRE           135314 non-null  object
44  FAMI_TRABAJOLABORMADRE           135429 non-null  object
45  ESTU_HORASSEMANATRABAJA          136087 non-null  object
46  FAMI_CUANTOSCOMPARTEBAÑO         135658 non-null  object
47  ESTU_PAGOMATRICULA               135459 non-null  object
48  ESTU_VLRULTIMOSEMESCURSADO       135871 non-null  object
49  ESTU_PRESENTACIONCASA            4 non-null       object
50  ESTU_PRESENTACIONSABADO          136576 non-null  object
51  INST_COD_INSTITUCION             139288 non-null  int64
52  INST_NOMBRE_INSTITUCION          139288 non-null  object
53  ESTU_PRGM_ACADEMICO              139288 non-null  object
```

```
 54  ESTU_SNIES_PRGMACADEMICO         139288 non-null  int64
 55  GRUPOREFERENCIA                  138289 non-null  object
 56  ESTU_PRGM_CODMUNICIPIO           139288 non-null  int64
 57  ESTU_PRGM_MUNICIPIO              139288 non-null  object
 58  ESTU_PRGM_DEPARTAMENTO           139288 non-null  object
 59  ESTU_NIVEL_PRGM_ACADEMICO        139288 non-null  object
 60  ESTU_METODO_PRGM                 139288 non-null  object
 61  ESTU_NUCLEO_PREGRADO             139288 non-null  object
 62  ESTU_NUCLEO_PREGRADO_1           139288 non-null  object
 63  ESTU_INST_CODMUNICIPIO           139288 non-null  int64
 64  ESTU_INST_MUNICIPIO              139288 non-null  object
 65  ESTU_INST_DEPARTAMENTO           139288 non-null  object
 66  INST_CARACTER_ACADEMICO          139288 non-null  object
 67  INST_ORIGEN                      139288 non-null  object
 68  ESTU_PRIVADO_LIBERTAD            139288 non-null  object
 69  ESTU_COD_MCPIO_PRESENTACION      139288 non-null  int64
 70  ESTU_MCPIO_PRESENTACION          139288 non-null  object
 71  ESTU_DEPTO_PRESENTACION          139288 non-null  object
 72  ESTU_COD_DEPTO_PRESENTACION      139288 non-null  int64
 73  MOD_RAZONA_CUANTITAT_PUNT        139288 non-null  int64
 74  MOD_RAZONA_CUANTITAT_DESEM       139288 non-null  int64
 75  MOD_RAZONA_CUANTITATIVO_PNAL     139288 non-null  int64
 76  MOD_RAZONA_CUANTITATIVO_PNBC     139288 non-null  int64
 77  MOD_LECTURA_CRITICA_PUNT         139288 non-null  int64
 78  MOD_LECTURA_CRITICA_DESEM        139288 non-null  int64
 79  MOD_LECTURA_CRITICA_PNAL         139288 non-null  int64
 80  MOD_LECTURA_CRITICA_PNBC         139288 non-null  int64
 81  MOD_COMPETEN_CIUDADA_PUNT        139288 non-null  int64
 82  MOD_COMPETEN_CIUDADA_DESEM       139288 non-null  int64
 83  MOD_COMPETEN_CIUDADA_PNAL        139288 non-null  int64
 84  MOD_COMPETEN_CIUDADA_PNBC        139288 non-null  int64
 85  MOD_INGLES_PUNT                  139198 non-null  float64
 86  MOD_INGLES_DESEM                 139198 non-null  object
 87  MOD_INGLES_PNAL                  139198 non-null  float64
 88  MOD_INGLES_PNBC                  139198 non-null  float64
 89  MOD_COMUNI_ESCRITA_PUNT          139288 non-null  int64
 90  MOD_COMUNI_ESCRITA_DESEM         135808 non-null  float64
 91  MOD_COMUNI_ESCRITA_PNAL          139288 non-null  int64
 92  MOD_COMUNI_ESCRITA_PNBC          139288 non-null  int64
 93  PUNT_GLOBAL                      139288 non-null  int64
 94  PERCENTIL_GLOBAL                 139198 non-null  float64
 95  PERCENTIL_NBC                    139198 non-null  float64
 96  ESTU_INSE_INDIVIDUAL             134393 non-null  float64
 97  ESTU_NSE_INDIVIDUAL              134393 non-null  float64
 98  ESTU_NSE_IES                     139288 non-null  int64
 99  ESTU_ESTADOINVESTIGACION         139288 non-null  object
dtypes: float64(8), int64(26), object(66)
memory usage: 106.3+ MB
```

```
# Distribución de las variables numéricas. De las 100 columnas, 34 son de tipo
# numérico (int64), por lo que son susceptibles de resumir estadisticamente.
df_saber_pro.describe()
```

```
{"type":"dataframe"}
```

```
# Se seleccionan algunas variables relevantes para el análisis descriptivo.
df_filter = df_saber_pro[['ESTU_GENERO', 'FAMI_ESTRATOVIVIENDA',
                          'INST_NOMBRE_INSTITUCION',
'ESTU_PRGM_ACADEMICO',
                          'GRUPOREFERENCIA', 'INST_ORIGEN',
                          'MOD_RAZONA_CUANTITAT_PUNT',
'MOD_RAZONA_CUANTITAT_DESEM',
                          'MOD_RAZONA_CUANTITATIVO_PNAL',
'MOD_RAZONA_CUANTITATIVO_PNBC',
                          'MOD_LECTURA_CRITICA_PUNT',
'MOD_LECTURA_CRITICA_DESEM',
                          'MOD_LECTURA_CRITICA_PNAL',
'MOD_LECTURA_CRITICA_PNBC',
                          'MOD_COMPETEN_CIUDADA_PUNT',
'MOD_COMPETEN_CIUDADA_DESEM',
                          'MOD_COMPETEN_CIUDADA_PNAL',
'MOD_COMPETEN_CIUDADA_PNBC',
                          'MOD_INGLES_PUNT', 'MOD_INGLES_DESEM',
'MOD_INGLES_PNAL',
                          'MOD_INGLES_PNBC',
'MOD_COMUNI_ESCRITA_PUNT',
                          'MOD_COMUNI_ESCRITA_DESEM',
'MOD_COMUNI_ESCRITA_PNAL',
                          'MOD_COMUNI_ESCRITA_PNBC', 'PUNT_GLOBAL',
'PERCENTIL_GLOBAL',
                          'PERCENTIL_NBC']]

df_filter
```

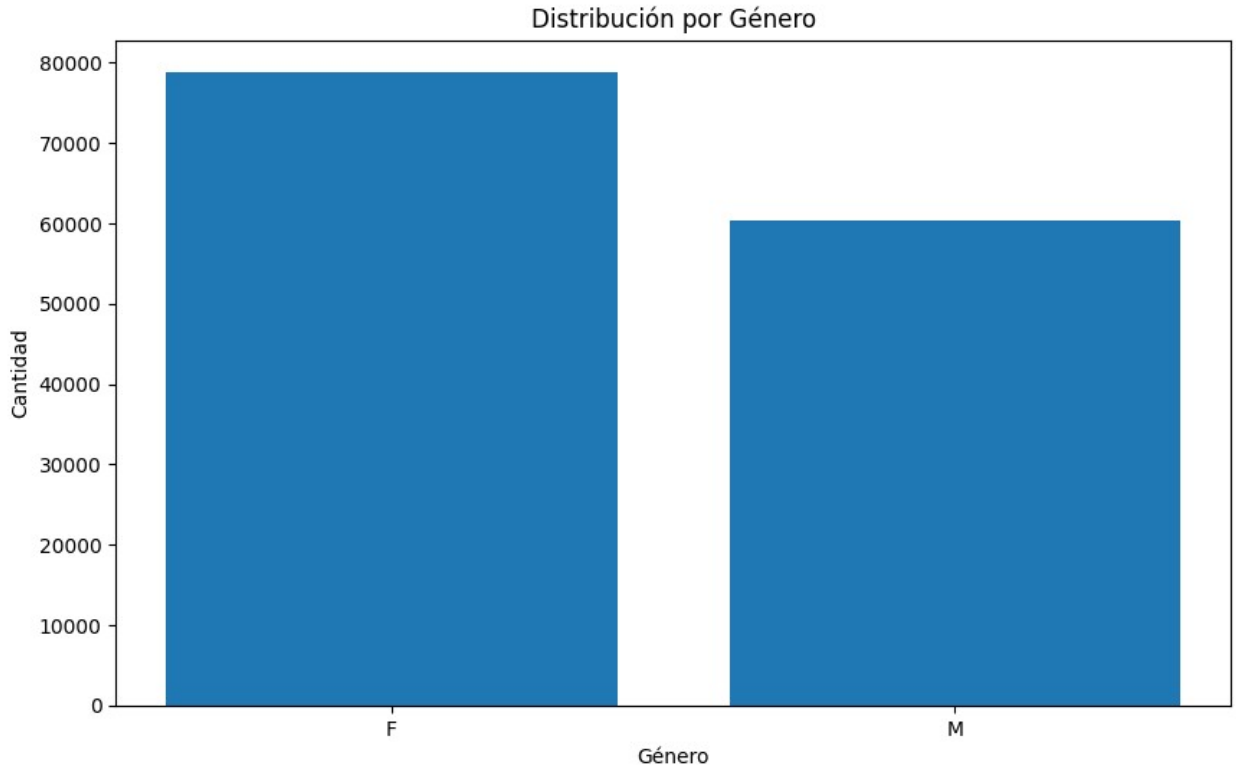
```
{"type":"dataframe","variable_name":"df_filter"}
```

Análisis Exploratorio

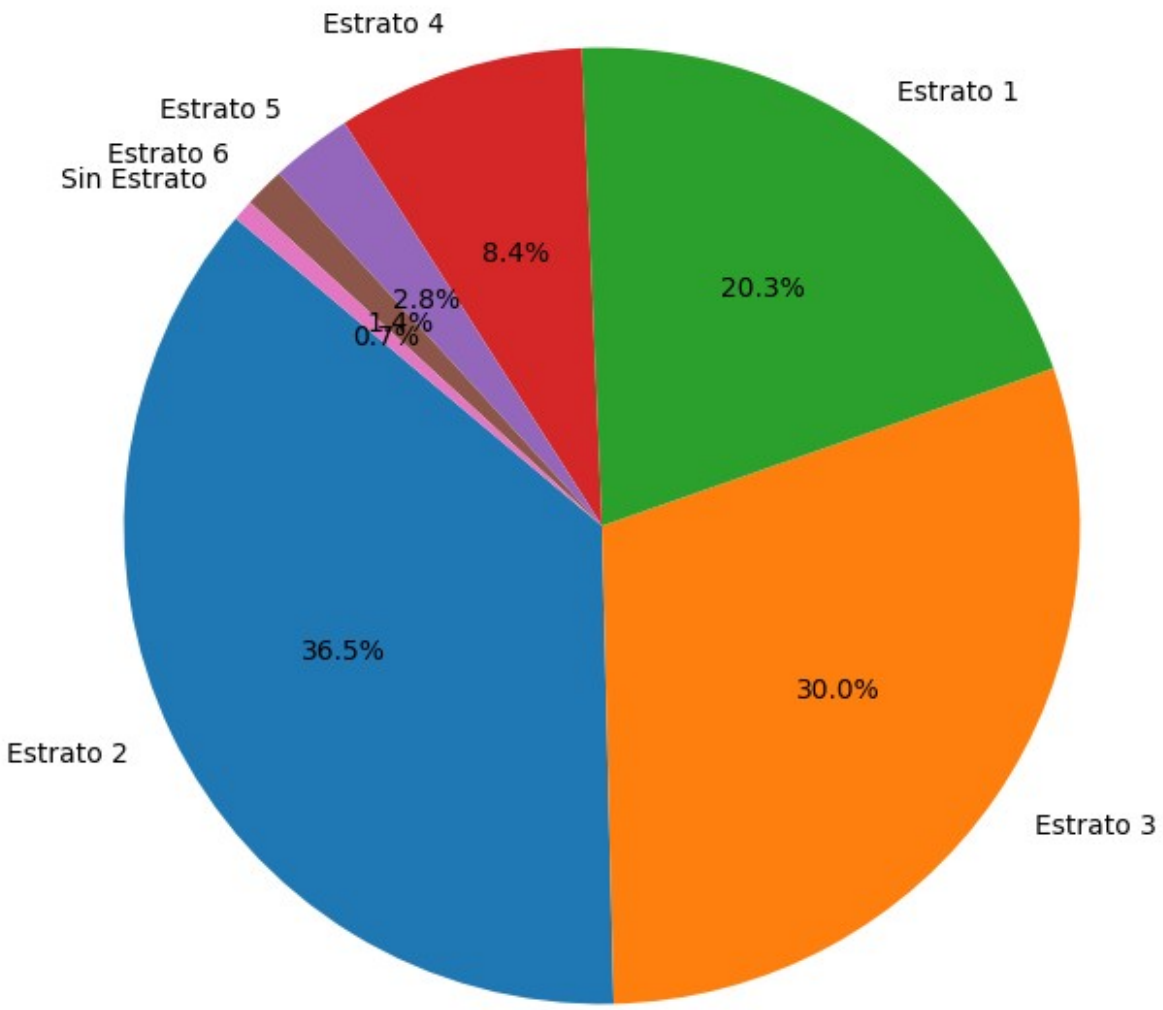
A continuación se presentan algunos elementos de estadística descriptiva de los datos.

```
# Distribución de los datos según la variable de Género
df_filter_genero = df_filter[['ESTU_GENERO']]
df_grouped =
df_filter_genero.groupby('ESTU_GENERO').size().reset_index(name='counts')
```

```
# Crear el gráfico de barras
plt.figure(figsize=(10, 6))
plt.bar(df_grouped['ESTU_GENERO'], df_grouped['counts'])
plt.xlabel('Género')
plt.ylabel('Cantidad')
plt.title('Distribución por Género')
plt.show()
```

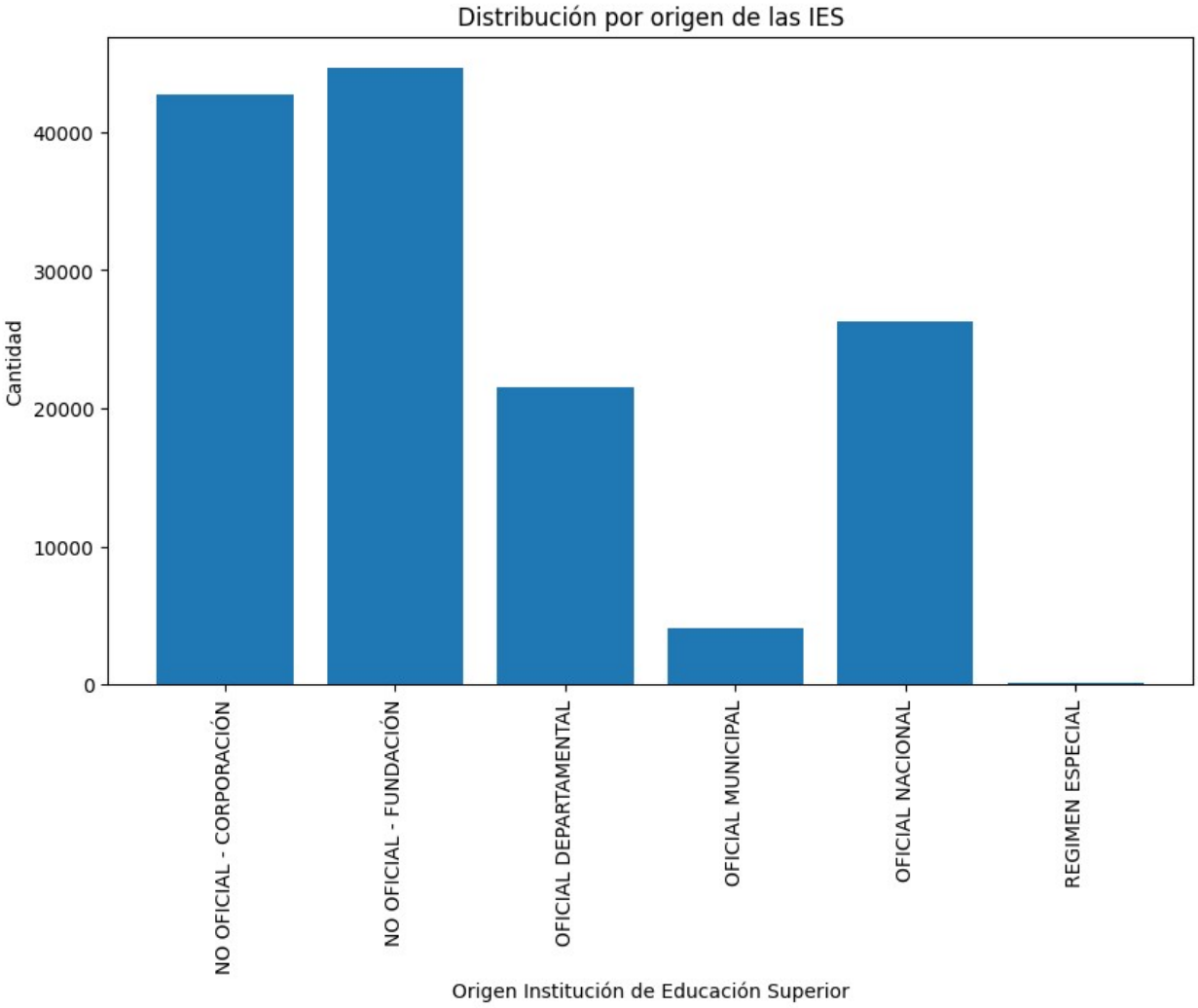



```
# Distribución por estrato socioeconómico
estrato_clean =
df_filter_estrato.dropna(subset=['FAMI_ESTRATOVIVIENDA'])
counts = estrato_clean['FAMI_ESTRATOVIVIENDA'].value_counts()
plt.figure(figsize=(8, 8))
plt.pie(counts, labels=counts.index, autopct='%1.1f%%',
startangle=140)
plt.title('Distribución por estrato socieconómico')
plt.show()
```

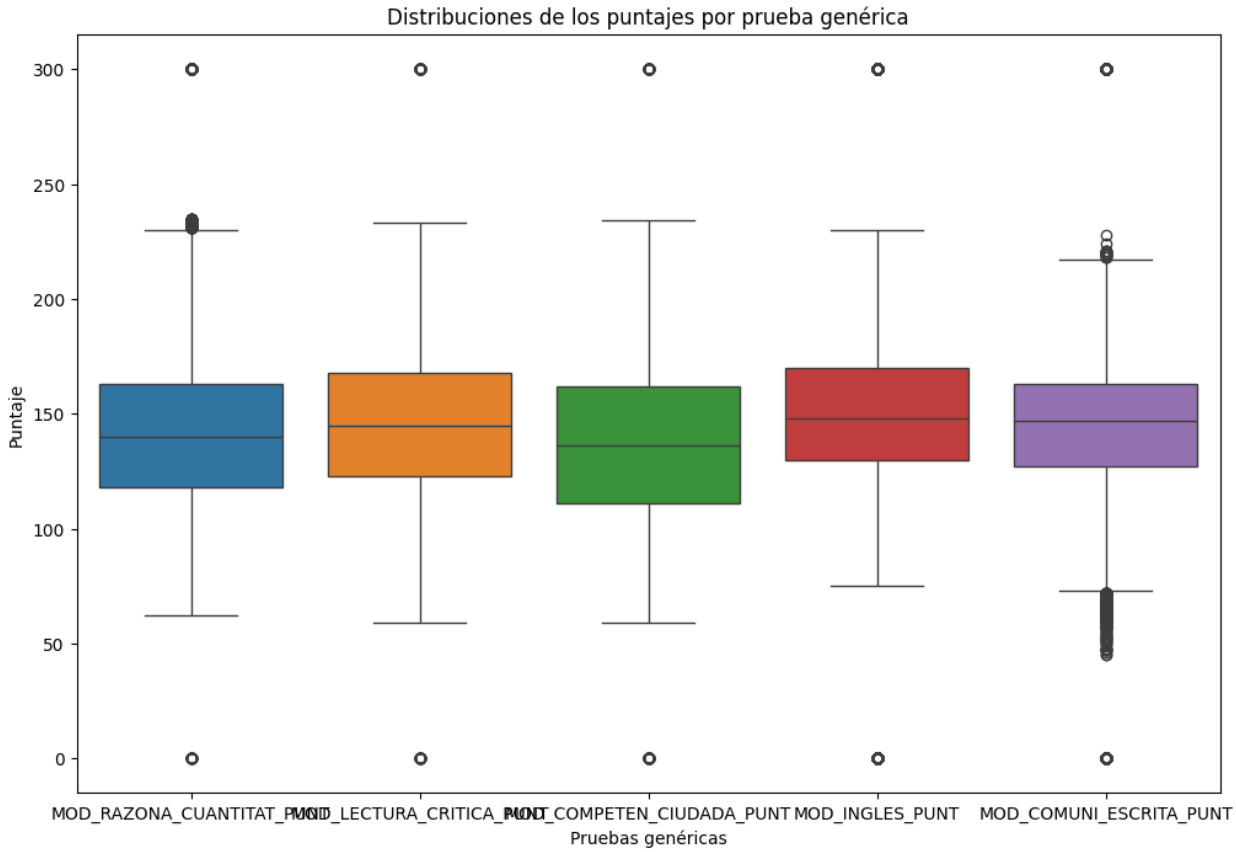



```
df_filter_caracterins = df_filter[['INST_ORIGEN']]
df_group_caracterins =
df_filter_caracterins.groupby('INST_ORIGEN').size().reset_index(name='
counts')

# Crear el gráfico de barras
plt.figure(figsize=(10, 6))
plt.bar(df_group_caracterins['INST_ORIGEN'],
df_group_caracterins['counts'])
plt.xlabel('Origen Institución de Educación Superior')
plt.ylabel('Cantidad')
```

```
plt.title('Distribución por origen de las IES')
plt.xticks(rotation=90)
plt.show()
```

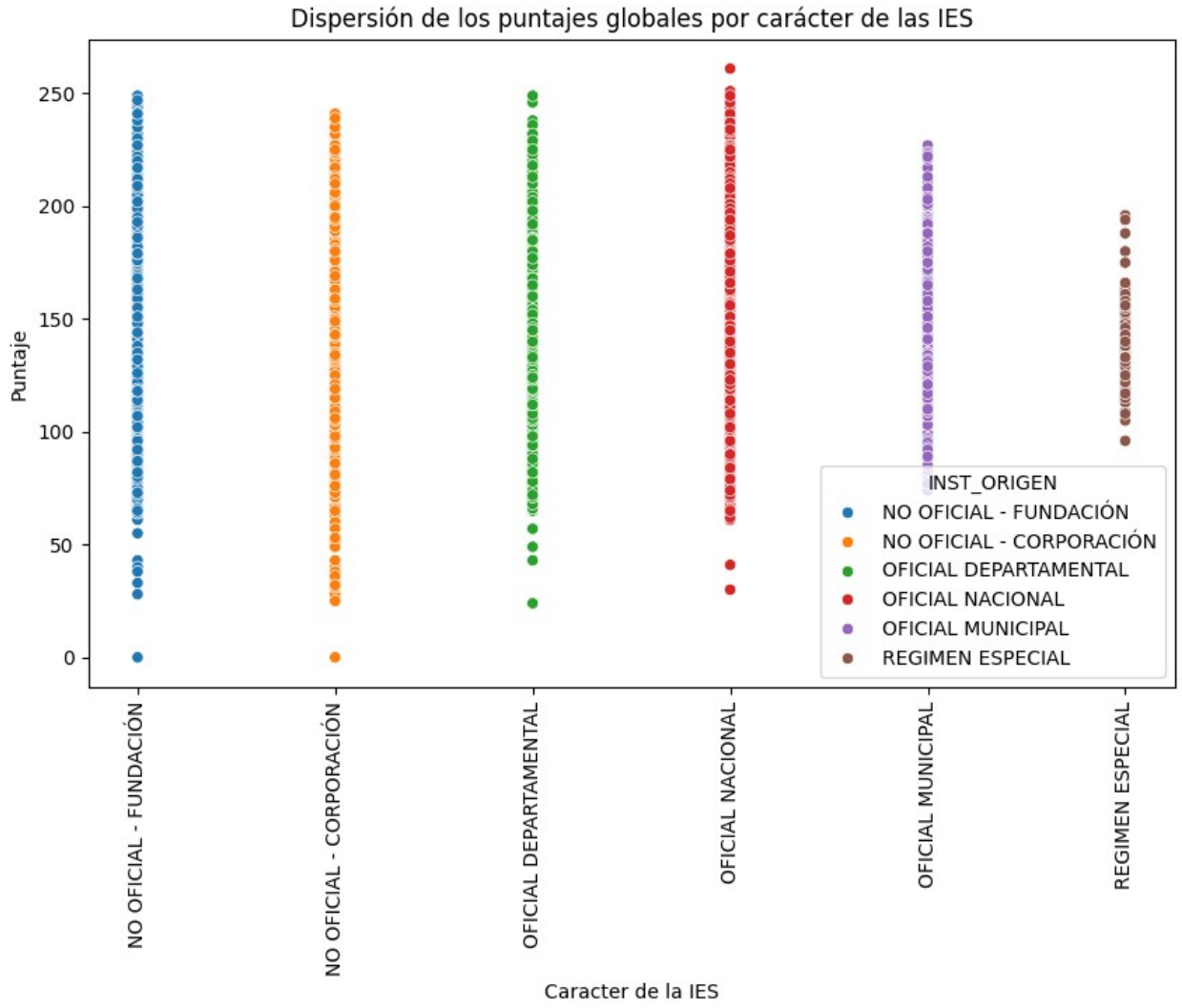



```
cols = ['MOD_RAZONA_CUANTITAT_PUNT', 'MOD_LECTURA_CRITICA_PUNT',
        'MOD_COMPETEN_CIUDADA_PUNT', 'MOD_INGLES_PUNT',
'MOD_COMUNI_ESCRITA_PUNT']
variables_puntajes = df_filter[cols]

# Crear el gráfico de boxplot usando seaborn
plt.figure(figsize=(12, 8))
sns.boxplot(data=variables_puntajes)
plt.title('Distribuciones de los puntajes por prueba genérica')
plt.xlabel('Pruebas genéricas')
plt.ylabel('Puntaje')
plt.show()
```




```
eje_x = 'INST_ORIGEN'
eje_y = 'PUNT_GLOBAL'
categoria = 'INST_ORIGEN'

# Crear el gráfico de dispersión usando seaborn
plt.figure(figsize=(10, 6))
sns.scatterplot(data=df_filter, x=eje_x, y=eje_y, hue=categoria)
plt.title('Dispersión de los puntajes globales por carácter de las
IES')
plt.xlabel('Caracter de la IES')
plt.ylabel('Puntaje')
plt.xticks(rotation=90)
plt.show()
```

```
/usr/local/lib/python3.10/dist-packages/IPython/core/
pylabtools.py:151: UserWarning: Creating legend with loc="best" can be
slow with large amounts of data.
  fig.canvas.print_figure(bytes_io, **kw)
```



Conclusiones

1. Si comparamos los datos de las bases anonimizadas publicadas por el ICFES, previo a 2023, la tendencia de presentación de la prueba Saber pro se mantiene en cuanto a un mayor número de mujeres. Para el caso del primero semestre del 2023 (teniendo en cuenta que desde 2022 la prueba se realiza dos veces al año), el número de mujeres fue de 78861, mientras el de hombres fue de 60427.
1. El creciente acceso a la educación superior de las clases sociales menos favorecidas historicamente también es importante mencionarlo, dado que si revisamos el estrato socioeconómico de quienes presentan la prueba, solamente entre los estratos 1, 2 y 3 se suma más del 80% de la población.
1. Pero por otra parte, también es necesario mencionar que un buen número de personas, exactamente 51796, de las que presentan la prueba, están inscritos en universidades o institucines de educación superior de caracter oficial nacional, departamental o municipal, es decir, instituciones públicas. No es menos cierto que las IES de tipo no oficial, ya sean fundaciones o corporaciones, cuentan también con un importante

número de estudiantes, siendo que en las IES privadas presentaron la prueba un total de 87377 personas.

1. Las distribuciones de los puntajes por cada una de las pruebas evidencia una cierta uniformidad en cuanto a las medidas de tendencia central, especialmente la media. Sin embargo, los rangos intercuatilicos muestran también que las distribuciones de las pruebas de competencias ciudadanas y razonamiento cuantitativo, son mayores en comparación al rango intercuartilico de la prueba de inglés y comunicación escrita. Para el caso de esta ultima prueba, la distribución también permite identificar una alta cantidad de valores cercanos a outliers, dado que la mayoría de los puntajes estan centrados en torno a la media, y que hay puntajes muy bajos, por fuera del primer cuartil. En suma, la prueba de comunicación escrita es la que presenta unos resultados mucho más dispersos en cuanto a valores atipicos, a pesar de que la mayoría se centra cerca a la media teorica, y es la prueba que requiere mayor atención en cuanto a las posibilidades y planes de mejoramiento.
1. Un diagrama de dispersión de los puntajes globales obtenidos de acuerdo al caracter de las IES, permite establecer que las instituciones de caracter oficial tienen una dispersión menor, ya que en la escala de 0 a 300, dichos puntajes inician en un punto más alto, se concentran hacia la media teórica (150) y se extienden hasta el 250, siendo estos ultimos valores si bien no atipicos, si bastante altos en comparación con los demás. Para el caso de las IES privadas del país, los puntajes globales están más dispersos, ya que inician en una escala mucho más baja y se extienden igualmente hacia el 250 con algunos casos de puntajes superiores. Las IES de caracter regimen especial, son interesantes para el análisis, ya que si bien no representan una muestra significativa dentro de la población total que presenta el examen en todo el país, sus resultados son bastante compactos en cuanto a la dispersión de los puntajes globales, iniciando en un escala cercana a los 100 puntos y extendiendose hasta los 200, es decir, una concentración bastante importante en torno a la media teórica.