PROYECTO DE ANALISIS DE DATOS

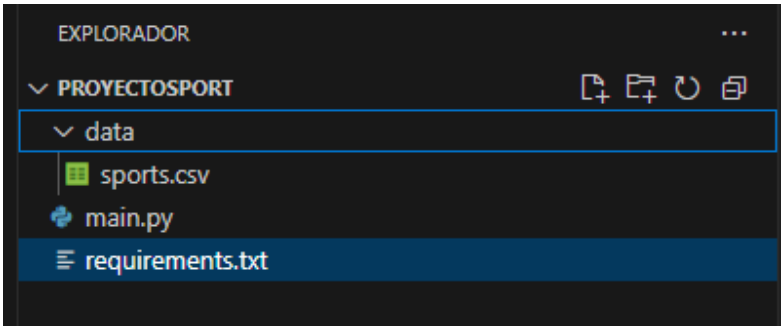
Jorge Alfredo Kirkconell Urbina
jkirkconell@econsultores.biz
Ricardo Antonio Soliman Satani
rsoliman@perdomoyasociados.com

Introducción al Proyecto

Este proyecto se desarrolló en el marco del Bootcamp de Análisis de Datos, con el objetivo de evaluar la equidad de género en el financiamiento del deporte universitario en Estados Unidos, utilizando un conjunto de datos amplio proveniente de instituciones académicas.

Estructura de Proyecto

```
PROYECTOSPORT/
|
└── main.py
└── otrosArchivos.py
└── requirements.txt
└── data/
    └── sports.csv
```



Requisitos del Proyecto (`requirements.txt`)

Contiene las bibliotecas necesarias para ejecutar los scripts del análisis. Asegura la reproducibilidad del entorno de desarrollo.

```
requirements.txt
requirements.txt
1  pandas
2  matplotlib
3  seaborn
4  jupyter
5  notebook
```

Descripción de Datos (`main.py`)

Se carga el archivo sports.csv con pandas. Durante la carga, se emite una advertencia de tipo de dato mixto en una de las columnas:

```
main.py
main.py > ...
1   import pandas as pd
2   import matplotlib.pyplot as plt
3   import seaborn as sns
4
5   # Cargar el archivo CSV
6   df = pd.read_csv("data/sports.csv")
7
8   # Vista general
9   print(df.head())
10  print(df.info())
```

Resultado Ejecución main.py

DtypeWarning: Columns (8) have mixed types. Specify dtype option on import or set low_memory=False.

```
  df = pd.read_csv("data/sports.csv")

   year  unitid  ... total_exp_menwomen            sports

0  2015  100654  ...          397818.0          Baseball

1  2015  100654  ...         1560328.0        Basketball

2  2015  100654  ...          498133.0        All Track Combined
```

```
3  2015  100654  ...        3059353.0          Football
4  2015  100654  ...          83913.0             Golf

[5 rows x 28 columns]
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 132327 entries, 0 to 132326
Data columns (total 28 columns):
 #   Column              Non-Null Count   Dtype
---  ------              --------------   -----
 0   year                132327 non-null  int64
 1   unitid              132327 non-null  int64
 2   institution_name    132327 non-null  object
 3   city_txt            132282 non-null  object
 4   state_cd            132282 non-null  object
 5   zip_text            132282 non-null  float64
 6   classification_code 132327 non-null  int64
 7   classification_name 132327 non-null  object
 8   classification_other 1679 non-null   object
 9   ef_male_count       132327 non-null  int64
 10  ef_female_count     132327 non-null  int64
 11  ef_total_count      132327 non-null  int64
 12  sector_cd           132327 non-null  int64
 13  sector_name         132282 non-null  object
 14  sportscode          132327 non-null  int64
 15  partic_men          61865 non-null   float64
 16  partic_women        68885 non-null   float64
 17  partic_coed_men     767 non-null     float64
 18  partic_coed_women   767 non-null     float64
 19  sum_partic_men      132327 non-null  int64
```

```
 20  sum_partic_women      132327 non-null  int64
 21  rev_men               61865 non-null   float64
 22  rev_women             68883 non-null   float64
 23  total_rev_menwomen    87134 non-null   float64
 24  exp_men               61865 non-null   float64
 25  exp_women             68885 non-null   float64
 26  total_exp_menwomen    87136 non-null   float64
 27  sports                132327 non-null  object
dtypes: float64(11), int64(10), object(7)
memory usage: 28.3+ MB
None
```

Enfoques evaluados en el Proyecto

1. ¿Hay equidad de género en la inversión deportiva universitaria?

- Comparar la participación femenina vs masculina con el gasto e ingreso asignado a cada grupo.
- *¿Reciben las atletas mujeres la misma proporción de recursos que los hombres en relación a su participación?*

Contenido equidadGenero.py

equidadGenero.py

equidadGenero.py > ...

```python
import pandas as pd
import os

# Ruta del archivo CSV
data_path = os.path.join("data", "sports.csv")

# Cargar datos
df = pd.read_csv(data_path)

# Crear una copia del dataframe para trabajar
df_copy = df.copy()

# Paso 1: Filtrar filas que tengan datos de participación y/o ingresos/gastos por género
df_gendered = df_copy[
    (df_copy["sum_partic_men"] > 0) | (df_copy["sum_partic_women"] > 0)
].copy()

# Paso 2: Calcular totales por fila
df_gendered["partic_total"] = df_gendered["sum_partic_men"] + df_gendered["sum_partic_women"]
df_gendered["rev_total"] = df_gendered["rev_men"].fillna(0) + df_gendered["rev_women"].fillna(0)
df_gendered["exp_total"] = df_gendered["exp_men"].fillna(0) + df_gendered["exp_women"].fillna(0)

# Paso 3: Calcular proporciones por género
df_gendered["partic_men_pct"] = df_gendered["sum_partic_men"] / df_gendered["partic_total"]
df_gendered["partic_women_pct"] = df_gendered["sum_partic_women"] / df_gendered["partic_total"]
df_gendered["rev_men_pct"] = df_gendered["rev_men"].fillna(0) / df_gendered["rev_total"]
df_gendered["rev_women_pct"] = df_gendered["rev_women"].fillna(0) / df_gendered["rev_total"]
df_gendered["exp_men_pct"] = df_gendered["exp_men"].fillna(0) / df_gendered["exp_total"]
df_gendered["exp_women_pct"] = df_gendered["exp_women"].fillna(0) / df_gendered["exp_total"]

# Paso 4: Promedios anuales
summary_by_year = df_gendered.groupby("year")[
    ["partic_women_pct", "rev_women_pct", "exp_women_pct"]
].mean().reset_index()

# Mostrar resumen en consola
print("Resumen anual de participación e inversión femenina:")
print(summary_by_year.head())
```

Resultado Ejecución equidadGenero.py

	year	partic_women_pct	rev_women_pct	exp_women_pct
0	2015	0.519238	0.527164	0.528684
1	2016	0.522245	0.530729	0.532237
2	2017	0.522585	0.530567	0.532034
3	2018	0.522331	0.530493	0.531797
4	2019	0.525097	0.534069	0.536046

Esto define lo siguiente:

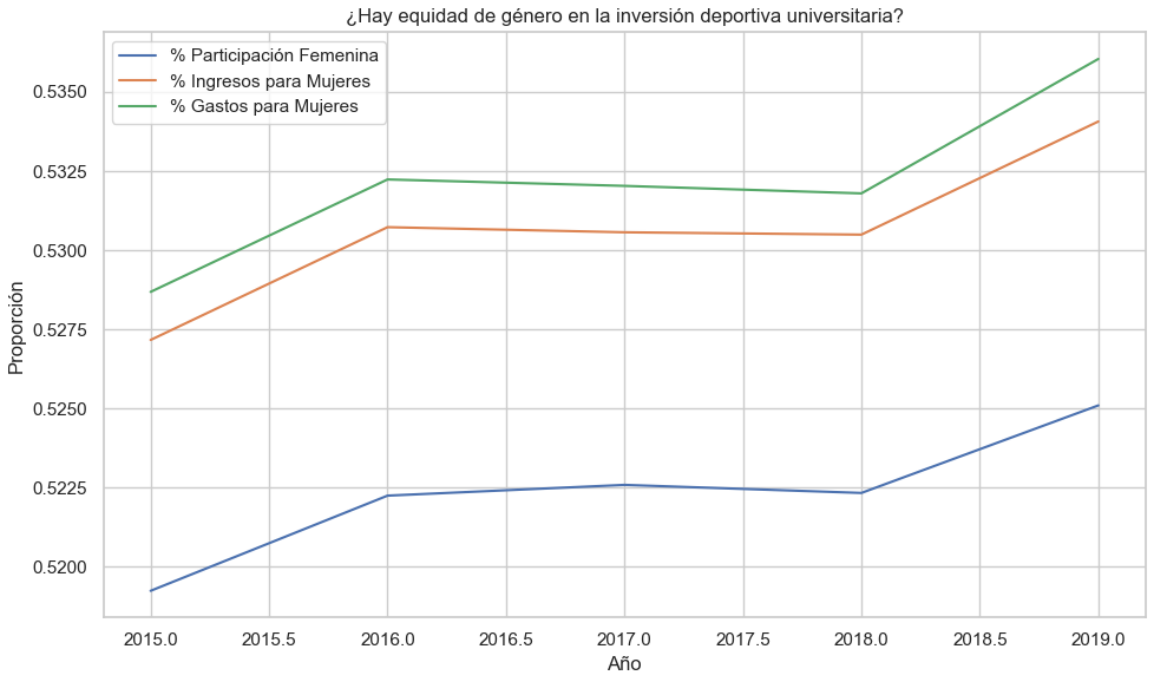
Año	% Participación Femenina	% Ingresos para Mujeres	% Gastos para Mujeres
2015	51.9%	52.7%	52.9%
2016	52.2%	53.1%	53.2%
2017	52.3%	53.1%	53.2%
2018	52.2%	53.0%	53.2%
2019	52.5%	53.4%	53.6%

Diagrama de equidad

generarDiagramaEquidad.py > ...

```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import os

# Asegurarse de que matplotlib use una fuente limpia
sns.set(style="whitegrid")

# Ruta al archivo CSV
DATA_PATH = os.path.join("data", "sports.csv")

# Cargar los datos
df = pd.read_csv(DATA_PATH)

# Filtrar filas con participación registrada
df_gendered = df[(df["sum_partic_men"] > 0) | (df["sum_partic_women"] > 0)].copy()

# Calcular totales y proporciones por género
df_gendered["partic_total"] = df_gendered["sum_partic_men"] + df_gendered["sum_partic_women"]
df_gendered["rev_total"] = df_gendered["rev_men"].fillna(0) + df_gendered["rev_women"].fillna(0)
df_gendered["exp_total"] = df_gendered["exp_men"].fillna(0) + df_gendered["exp_women"].fillna(0)

df_gendered["partic_women_pct"] = (df_gendered["sum_partic_women"] / df_gendered["partic_total"])
df_gendered["rev_women_pct"] = df_gendered["rev_women"].fillna(0) / df_gendered["rev_total"]
df_gendered["exp_women_pct"] = df_gendered["exp_women"].fillna(0) / df_gendered["exp_total"]

# Agrupar por año y calcular promedios
summary_by_year = df_gendered.groupby("year")[
    ["partic_women_pct", "rev_women_pct", "exp_women_pct"]
].mean().reset_index()

# Visualización
plt.figure(figsize=(10, 6))
plt.plot(summary_by_year["year"], summary_by_year["partic_women_pct"], label="% Participación Femenina")
plt.plot(summary_by_year["year"], summary_by_year["rev_women_pct"], label="% Ingresos para Mujeres")
plt.plot(summary_by_year["year"], summary_by_year["exp_women_pct"], label="% Gastos para Mujeres")

plt.title("¿Hay equidad de género en la inversión deportiva universitaria?")
plt.ylabel("Proporción")
plt.xlabel("Año")
plt.legend()
plt.tight_layout()

# Crear carpeta outputs si no existe
os.makedirs("outputs", exist_ok=True)
plt.savefig(os.path.join("outputs", "equidad_genero.png"))

plt.show()
```

Resultado



Interpretación

- Las mujeres representan aproximadamente el **52% de la participación deportiva** universitaria.
- En general, reciben una **proporción ligeramente mayor** de ingresos y gastos (~53%-54%).
- Lo que **rompe con la idea preconcebida** de que las mujeres podrían estar recibiendo menos recursos que los hombres en el deporte universitario.

2. Equidad de inversion de instituciones

Este análisis se enfoca en evaluar la equidad de género en la inversión deportiva universitaria a nivel institucional, centrándose específicamente en cuánto se gasta por atleta hombre vs. mujer.

Contenido brechaGeneroInstituciones.py

```python
# Cargar datos
df = pd.read_csv("data/sports.csv")

# Filtrar datos válidos
df_gendered = df[
    (df["sum_partic_men"] > 0) & (df["sum_partic_women"] > 0) &
    (df["exp_men"] > 0) & (df["exp_women"] > 0)
].copy()

# Calcular gasto por atleta
df_gendered["gasto_por_hombre"] = df_gendered["exp_men"] / df_gendered["sum_partic_men"]
df_gendered["gasto_por_mujer"] = df_gendered["exp_women"] / df_gendered["sum_partic_women"]
df_gendered["brecha_absoluta"] = df_gendered["gasto_por_hombre"] - df_gendered["gasto_por_mujer"]
df_gendered["razon_equidad"] = df_gendered["gasto_por_mujer"] / df_gendered["gasto_por_hombre"]

# Agrupación por institución
institucion_brecha = df_gendered.groupby("institution_name")[[
    "gasto_por_hombre", "gasto_por_mujer", "brecha_absoluta", "razon_equidad"
]].mean().reset_index()

# Mostrar top 5 instituciones con mayor brecha a favor de hombres
print("Top 5 instituciones con mayor brecha a favor de hombres:")
print(institucion_brecha.sort_values(by="brecha_absoluta", ascending=False).head())

# Mostrar top 5 instituciones con mayor equidad (razón cercana a 1)
print("\nTop 5 instituciones más equitativas en gasto:")
print(institucion_brecha.loc[(institucion_brecha["razon_equidad"] < 2)].sort_values(
    by="razon_equidad", ascending=False).head())

# Filtrar último año disponible
ultimo_anio = df_gendered["year"].max()
df_ultimo = df_gendered[df_gendered["year"] == ultimo_anio]

# Histograma de la razón de equidad por institución
plt.figure(figsize=(10, 6))
sns.histplot(df_ultimo["razon_equidad"], bins=40, kde=True, color="skyblue")
plt.axvline(1, color="red", linestyle="--", label="Equidad perfecta")
plt.title(f"Distribución de equidad en gasto por atleta - {ultimo_anio}")
plt.xlabel("Razón de Equidad (mujer/hombre)")
plt.ylabel("Cantidad de instituciones")
plt.legend()
plt.tight_layout()

# Guardar gráfico
os.makedirs("outputs", exist_ok=True)
plt.savefig("outputs/histograma_equidad_instituciones.png")
plt.show()
```

Resultado Ejecución brechaGeneroInstituciones.py

Top 5 instituciones con mayor brecha a favor de hombres:

	institution_name	gasto_por_hombre	gasto_por_mujer	brecha_absoluta	razon_equidad
1826	University of Kansas	285078.075771	84067.006746	201011.069024	0.786281
2059	West Virginia University	247543.213867	87505.459694	160037.754173	1.121441

1831 University of Louisville 260301.685974 106365.180768 153936.505205 0.930704

1827 University of Kentucky 270201.445448 118452.143251 151749.302197 0.997762

982 Marquette University 196899.945024 71828.476409 125071.468615 1.025841

 Top 5 instituciones más equitativas en gasto:

institution_name gasto_por_hombre gasto_por_mujer brecha_absoluta razon_equidad

827 Kansas Christian College 1613.013028 3239.190972 -1626.177944 1.988433

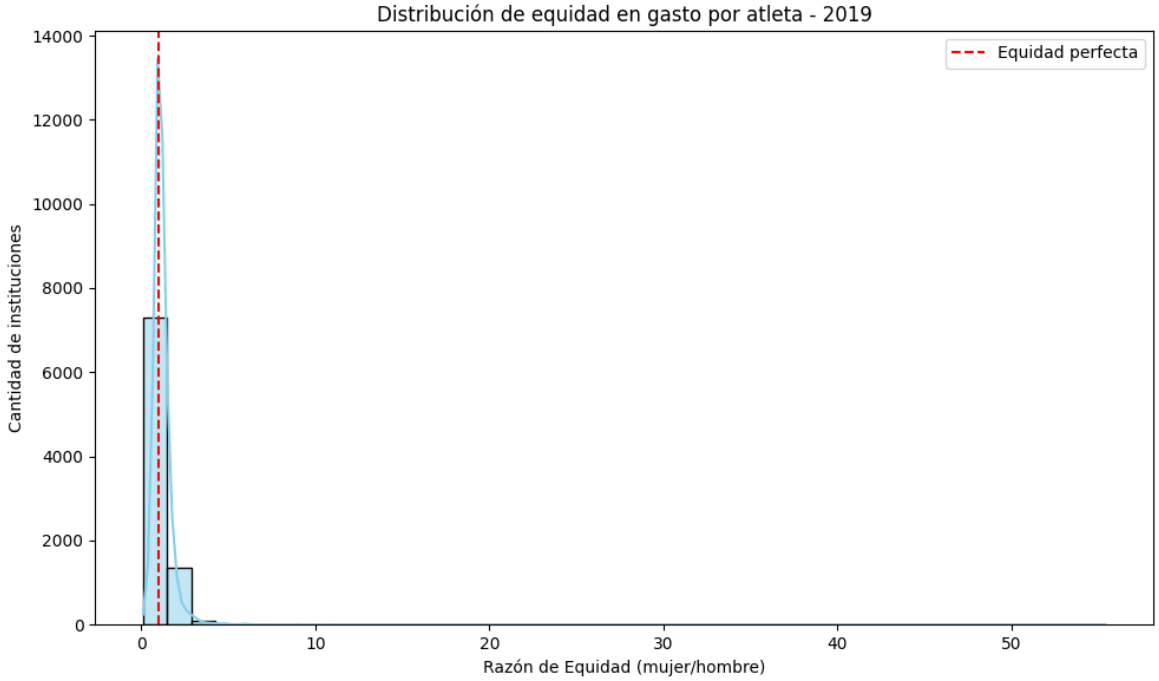
651 Glenville State College 9679.663506 9511.114235 168.549271 1.986139

168 Bryant & Stratton College-Solon 8882.933333 17559.727273 -8676.793939 1.976794

1487 Santa Monica College 3087.840299 5196.306162 -2108.465864 1.975863

580 Florida Agricultural and Mechanical University 28180.631238 34823.084675 -6642.453437 1.973582

Diagrama



Interpretación

- Existen instituciones con brechas importantes: algunas gastan cientos (o miles) de dólares más por atleta hombre que por atleta mujer.
- Sin embargo, también hay instituciones que presentan una equidad casi perfecta e incluso algunas donde la inversión por mujer es mayor.
- La distribución es amplia, lo que indica que el panorama es muy heterogéneo y no hay una política uniforme en todas las universidades.

Conclusiones Generales del Proyecto

Este proyecto de análisis de datos sobre equidad de género en el deporte universitario ha permitido identificar patrones clave en la asignación de recursos financieros por género, utilizando un enfoque cuantitativo riguroso sobre una base de datos amplia y detallada.

Principales hallazgos

- **Participación equitativa con asignación proporcional:**
  A nivel general, las mujeres representan aproximadamente el 52% de la participación en deportes universitarios. La inversión en términos de ingresos y gastos también ronda ese mismo porcentaje, lo que sugiere una **distribución proporcional a nivel agregado**.
- **Diferencias significativas a nivel institucional:**
  Aunque el promedio general muestra equidad, el análisis por institución revela **brechas importantes** en la asignación de recursos por género. Algunas universidades destinan mucho más dinero por atleta hombre que por mujer, mientras que otras incluso invierten más por atleta mujer.
- **Panorama heterogéneo:**
  La variabilidad entre instituciones indica que **no existe una política uniforme** sobre equidad de género en el deporte universitario. Las decisiones parecen estar descentralizadas y depender de factores internos de cada universidad (presupuesto, tradición deportiva, enfoque institucional, entre otros).
- **Utilidad del análisis de datos para decisiones políticas:**
  El proyecto demuestra el valor del análisis de datos como herramienta para **evaluar, transparentar y orientar políticas públicas o institucionales** en el ámbito deportivo y educativo.