



*Analyzing School Census Data*

José Daniel Carrera Bolaños

Martín García Chagüezá

David Melo Valbuena

Juan Andrés Ruiz Muñoz

Santiago de Cali, Colombia

2024

# Introducción

El **U.S. Census at School** es un proyecto educativo diseñado para involucrar a estudiantes de los Estados Unidos en actividades relacionadas con la recolección, análisis y comprensión de datos. En este documento buscamos indagar más en profundidad algunas de las variables otorgadas por el cuestionario que van desde datos sobre edad, género, dominancia de mano, métodos de transporte, tiempos de reacción, actividades favoritas, entre otros.

Aplicando un análisis estadístico sobre el dataset que proviene de este censo, aplicaremos conceptos relacionados con el desarrollo de análisis descriptivos y visualizaciones de datos, como histogramas, ojivas, y diagramas de cajas.

Como muestras de datos, recogimos información de 500 estudiantes para 4 estados: **California, Florida, Nueva York y Texas**; su respectivo proceso de limpieza será descrito en la siguiente sección.

## Limpieza

- En un [notebook](#) dedicado únicamente para la limpieza se realizan la carga de las 5 muestras descargadas mediante Pandas.

The screenshot shows a Jupyter Notebook interface with the following code:

```
^ Cargando los dataframes
[100]: df_ca = pd.read_csv('../data/sample_data/california.csv', encoding='cp1252')
[101]: df_fl = pd.read_csv('../data/sample_data/florida.csv', encoding='cp1252')
[102]: df_ny = pd.read_csv('../data/sample_data/new_york.csv', encoding='cp1252')
[103]: df_tx = pd.read_csv('../data/sample_data/texas.csv', encoding='cp1252')
[104]: df_ca
```

Below the code, the first few rows of the `df_ca` DataFrame are displayed:

	Country	Region	DataYear	ClassGrade	Gender	Ageyears	Handed	Height_cm	Footlength_cm	Armspan_cm	...	Watching_TV_Hours	Paid_Work_Hours	Work_At_Home_Hours	Schoolwork_Pressure	Planned_Education_Level	Favorite_Music	Superpower	Preferred_Status
0	USA	CA	2017	11	Male	16.0	Right-Handed	174	26	170	..	6	0	10	A lot	Undergraduate degree	Pop	Freeze time	Rich
1	USA	CA	2014	5	Male	10.0	Right-Handed	146	23	140	..	7	0	7	Very little	Graduate degree	Classical	Invisibility	Happy
2	USA	CA	2017	12	Male	17.0	Right-Handed	180	25	61	..	1	0	6	Very little	Some college	Tech/Industrial	Fly	Happy
3	USA	CA	2023	9	Male	14.0	Right-Handed	173	31	174	..	0	0	7	A lot	Other	Rap/Hip hop	Freeze time	Healthy
4	USA	CA	2013	12	Male	17.0	Right-Handed	160	23.5	58	..	10	0	4	Some	Graduate degree	Pop	Super strength	Happy
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	
495	USA	CA	2021	12	Female	16.0	Right-Handed	158	21	153	..	0	0	0	Some	Graduate degree	Pop	Telepathy	Rich
496	USA	CA	2022	8	Female	13.0	Nan	S'10	15	20	..	?	?	?	NaN	Other	Other	Invisibility	Famous
497	USA	CA	2018	8	Female	13.0	Right-Handed	164	32	166	..	4	0	2	Very little	Some college	Pop	Telepathy	Famous
498	USA	CA	2022	7	Female	12.0	Right-Handed	155	22	153	..	0	0	3	Some	Graduate degree	Other	Telepathy	Rich
499	USA	CA	2023	12	Male	16.0	Right-Handed	168	22	160	..	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

500 rows × 60 columns

- Se seleccionan **sólo 12 columnas de las 60 existentes** en los 4 datasets de muestra. A partir de esa selección realizamos la fusión de los 4 muestras en una sola: en este nuevo dataframe van a haber 2.000 registros con datos de los 4 estados.

```

df_ca.columns # Todos los dataframes tienen las mismas columnas

Index(['Country', 'Region', 'DataYear', 'ClassGrade', 'Gender', 'Ageyears',
       'Handed', 'Height_cm', 'Footlength_cm', 'Armspan_cm',
       'Languages_spoken', 'Travel_to_School', 'Travel_time_to_School',
       'Reaction_time', 'Score_in_memory_game', 'Favourite_physical_activity',
       'Importance_reducing_pollution', 'Importance_recycling_rubbish',
       'Importance_conserving_water', 'Importance_saving_energy',
       'Importance_owning_computer', 'Importance_Internet_access',
       'Left_Footlength_cm', 'Longer_foot', 'Index_Fingerlength_mm',
       'Ring_Fingerlength_mm', 'Longer_Finger_lefthand', 'Birth_month',
       'Favorite_Season', 'Allergies', 'Vegetarian', 'Favorite_Food',
       'Beverage', 'Favorite_School_Subject', 'Sleep_Hours_Schoolnight',
       'Sleep_Hours_Non_Schoolnight', 'Home_Occupants', 'Home_Internet_Access',
       'Communication_With_Friends', 'Text_Messages_Sent_Yesterday',
       'Text_Messages_Received_Yesterday', 'Hanging_Out_With_Friends_Hours',
       'Talking_On_Phone_Hours', 'Doing_Homework_Hours',
       'Doing_Things_With_Family_Hours', 'Outdoor_Activities_Hours',
       'Video_Games_Hours', 'Social_Websites_Hours', 'Texting_Messaging_Hours',
       'Computer_Use_Hours', 'Watching_TV_Hours', 'Paid_Work_Hours',
       'Work_At_Home_Hours', 'Schoolwork_Pressure', 'Planned_Education_Level',
       'Favorite_Music', 'Superpower', 'Preferred_Status', 'Role_Model_Type',
       'Charity_Donation'],
      dtype='object')

def seleccionando_columnas(df):
    return df[['Region', 'Gender', 'Handed', 'Favourite_physical_activity', \
               'Importance_reducing_pollution', 'Birth_month', 'Beverage', 'Favorite_School_Subject', 'Sleep_Hours_Non_Schoolnight', \
               'Paid_Work_Hours', 'Work_At_Home_Hours', 'Planned_Education_Level']]

```

```

df_merged = pd.concat([seleccionando_columnas(df_ca), seleccionando_columnas(df_fl), seleccionando_columnas(df_ny), seleccionando_columnas(df_tx)])
df_merged = df_merged.reset_index(drop=True)

df_merged

  Region Gender Handed Favourite_physical_activity Importance_reducing_pollution Birth_month Beverage Favorite_School_Subject Sleep_Hours_Non_Schoolnight
0   CA     Male Right-Handed Running/logging                      1000        April    Water  Computers and technology          11
1   CA     Male Right-Handed Walking/Hiking                      1000        May     Water           Science          12
2   CA     Male Right-Handed Soccer                           1000        June    Water  Mathematics and statistics          10
3   CA     Male Right-Handed Swimming                         1000        June    Water  Physical education          8-7
4   CA     Male Right-Handed Basketball                        900         July    Water           English            9
...   ...     ...
1995  TX     Male Left-Handed          NaN                  600        July  Soft drink (caffeinated)          Music            9
1996  TX     Male Right-Handed          NaN                  NaN        NaN     NaN           NaN          NaN
1997  TX     Male Right-Handed Running/logging                   332        May     Water           History          10
1998  TX    Female Right-Handed Martial Arts                     893        August    Water           History            7
1999  TX     Male Right-Handed Baseball/Softball                 NaN        October  Soft drink (non-caffeinat...

```

2000 rows × 12 columns

- Limpiamos los registros que contengan datos nulos, después de esta limpieza quedan 1480 registros: **una disminución del 26% de los registros**.

Revisando los tipos de datos en el dataframe nuevo											
	Region	Gender	Handed	Favourite_physical_activity	Importance_reducing_pollution	Birth_month	Beverage	Favorite_School_Subject	Sleep_Hours_Non_Schoolnight		
0	CA	Male	Right-Handed	Running/Jogging	1000	April	Water	Computers and technology	11		
1	CA	Male	Right-Handed	Walking/Hiking	1000	May	Water	Science	12		
2	CA	Male	Right-Handed	Soccer	1000	June	Water	Mathematics and statistics	10		
3	CA	Male	Right-Handed	Swimming	1000	June	Water	Physical education	8-7		
4	CA	Male	Right-Handed	Basketball	900	July	Water	English	9		
...	...	...	...	...	...	...	...	...	...	...	...
1988	TX	Female	Right-Handed	Walking/Hiking	1000	August	Water	Mathematics and statistics	8		
1990	TX	Male	Right-Handed	Football (American)	439	July	Water	Physical education	10		
1991	TX	Female	Right-Handed	Swimming	1000	March	Water	History	10		
1993	TX	Female	Right-Handed	Basketball	5	December	Water	Computers and technology	8		
1997	TX	Male	Right-Handed	Running/Jogging	332	May	Water	History	10		

1480 rows × 12 columns

- Ahora, transformamos las columnas que contienen valores numéricos. Este es uno de los pasos más importantes, pues para realizar análisis sobre las variables ordinales **todos sus valores** deben ser de tipo entero o decimal (*flotante*).

Teniendo esto en cuenta revisamos los tipos de datos (*Dtypes*) dentro de nuestro dataframe: encontramos que ningún dato es de tipo numérico, representados con **int64** o **float64**.

```
df_merged.info()

<class 'pandas.core.frame.DataFrame'>
Index: 1480 entries, 0 to 1997
Data columns (total 12 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   Region          1480 non-null    object  
 1   Gender           1480 non-null    object  
 2   Handed           1480 non-null    object  
 3   Favourite_physical_activity  1480 non-null    object  
 4   Importance_reducing_pollution 1480 non-null    object  
 5   Birth_month      1480 non-null    object  
 6   Beverage         1480 non-null    object  
 7   Favorite_School_Subject    1480 non-null    object  
 8   Sleep_Hours_Non_Schoolnight 1480 non-null    object  
 9   Paid_Work_Hours     1480 non-null    object  
 10  Work_At_Home_Hours   1480 non-null    object  
 11  Planned_Education_Level 1480 non-null    object  
dtypes: object(12)
memory usage: 150.3+ KB
```

Sin embargo, este proceso se nos dificulta debido a la falta de un formato único para las respuestas, como se observa en la siguiente imagen:

```
df_merged["Sleep_Hours_Non_Schoolnight"].value_counts()

Sleep_Hours_Non_Schoolnight
10      333
9       324
8       262
7       185
12      98
11      79
6       76
5       54
4       30
3       19
13      10
8.5     10
8-9     7
14      6
9.5     5
15      5
2       5
10.5    3
1       3
0       3
7.5     3
9-10    2
4-5     2
16      2
...
5.5     1
25      1
9-11    1
12 hr   1
Name: count, dtype: int64
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...
```

Para ello, se define un proceso de limpieza generalizado para ciertos aspectos:

- La remoción de caracteres no numéricos (líneas diagonales, guiones, puntos mal ubicados).
- La conversión de valores literales o *strings* a enteros, dependiendo de si contiene un punto decimal o no.
- Convertir los valores flotantes a enteros.

En algunos casos donde existen intervalos de hora, se extrae el promedio de los dos valores y luego se redondea a un entero, como se observa en la lógica de la función *process\_sleep\_hours*.

```
# Define las columnas numéricas
numeric_columns = ['Importance_reducing_pollution', 'Sleep_Hours_Non_Schoolnight', 'Paid_Work_Hours', 'Work_At_Home_Hours']

# Redefinir la función para procesar los intervalos de horas en Sleep_Hours_Non_Schoolnight
def process_sleep_hours(value):
    if isinstance(value, str) and ('-' in value or '/' in value):
        hours = re.split(['-|/'], value)
        return sum(map(int, hours)) / len(hours)
    return value

# Redefinir el pipeline completo
def clean_value(value):
    if isinstance(value, str):
        value = re.sub(r'[^d.]*', '', value) # Eliminar caracteres no numéricos
        if '-' in value:
            value = value.replace('-', '')
        if '.' in value:
            return round(float(value))
        if value.isdigit():
            return int(value)
    elif isinstance(value, (int, float)):
        return int(round(value))
    return None

# Procesar intervalos en Sleep_Hours_Non_Schoolnight
df_merged['Sleep_Hours_Non_Schoolnight'] = df_merged['Sleep_Hours_Non_Schoolnight'].apply(process_sleep_hours)

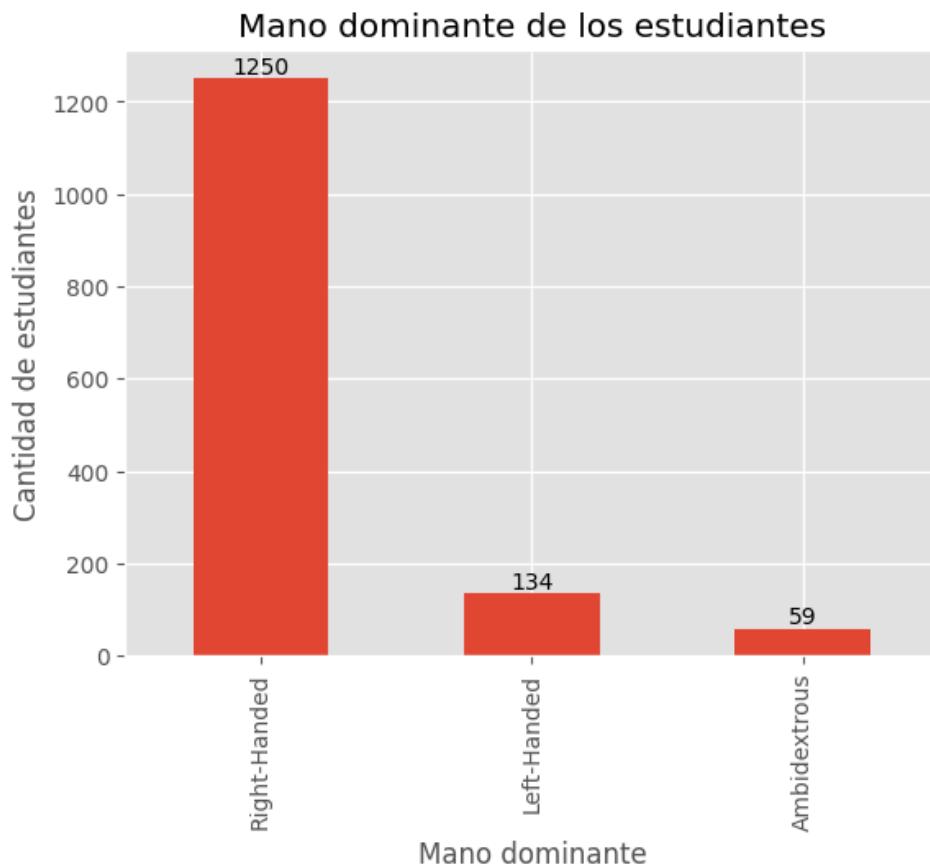
# Aplicar limpieza a columnas numéricas
for col in numeric_columns:
    df_merged[col] = df_merged[col].apply(clean_value)

# Filtrar valores fuera de rango usando pd.DataFrame.query
df_merged = df_merged.query('0 <= Importance_reducing_pollution <= 1000')
df_merged = df_merged.query('1 <= Sleep_Hours_Non_Schoolnight <= 15')
df_merged = df_merged.query('0 <= Paid_Work_Hours <= 64')
df_merged = df_merged.query('0 <= Work_At_Home_Hours <= 56')

df_merged = df_merged.reset_index(drop=True).copy()
```

## 1. ¿Cuántos estudiantes son diestros, zurdos o ambidiestros?

- **Datos:** Se tiene información sobre la mano dominante de los estudiantes, categorizada como diestros, zurdos o ambidiestros.
- **Distribución de los Datos:**
  - 1.250 estudiantes son diestros.
  - 134 estudiantes son zurdos.
  - 59 estudiantes son ambidiestros.



## Insights y Conclusiones

- **Distribución de la Mano Dominante:**
  - La mayoría de los estudiantes son diestros, lo cual es consistente con la distribución general de la población global, donde se estima que aproximadamente el 90% de las personas son diestras.
  - El número de ambidiestros es relativamente bajo, lo que es esperado, ya que esta característica es menos común en la población.
- **Importancia del Análisis:**

Este análisis básico permite entender mejor la composición de la población estudiantil en términos de una característica física que podría influir en otros aspectos del aprendizaje y comportamiento. Además, se podrían realizar análisis adicionales para ver si hay correlaciones entre la mano dominante y otras variables, como el rendimiento académico, deportes favoritos, o métodos de transporte usados.

**2. ¿Cuál es el mes donde nacieron más estudiantes de los que participan en el Censo escolar en cada estado?**

- **Descripción:** Se determinó el mes más común en el que nacieron los estudiantes en cada estado participante.
  - **Insight Principal:** Septiembre es el mes donde más estudiantes nacieron en todos los estados evaluados (*CA, FL, NY, TX*).
  - **Interpretación:** Esto podría estar relacionado con factores socioculturales o escolares, por ejemplo, que los padres planifiquen los nacimientos en este mes para que los niños comiencen la escuela a una edad específica.

```
conteo_nacimientos_por_mes = (df.groupby(['Region', 'Birth_month']).size()
|                                .reset_index(name='Count'))

mes_más_comun_por_estado = conteo_nacimientos_por_mes.groupby('Region').agg(
    Mes_Nacimiento=('Birth_month', lambda x: x.loc[x.idxmax()]),
    Conteo=('Count', 'max')
).reset_index()
```

mes más común por estado

Region	Mes_Nacimiento	Conteo
CA	September	38
FL	September	38
NY	September	41
TX	September	38

**3. ¿Cuál es la bebida favorita de los estudiantes que participan en el Censo escolar en cada estado?**

- **Descripción:** Se analizó cuál es la bebida más consumida por los estudiantes en cada estado

- **Insight Principal:** En todos los estados (*CA, FL, NY, TX*), la bebida favorita es el agua.
- **Interpretación:** El hecho de que el agua sea la bebida favorita entre los estudiantes sugiere una conciencia sobre la salud y la hidratación, lo cual puede estar influenciado por programas de educación nutricional o concientización sobre la salud.

```

conteo_bebidas_favoritas = (df.groupby(["Region", "Beverage"]).size()
                             .reset_index(name='Count'))

bebida_favorita_por_estado = conteo_bebidas_favoritas.groupby('Region').agg(
    Bebida_Favorita=('Beverage', lambda x: x.loc[x.idxmax()]),
    Conteo=('Count', 'max')
).reset_index()

bebida_favorita_por_estado

```

✓ 0.0s ┌ Abrir "bebida\_favorita\_por\_estado" en Data Wrangler

Region	Bebida_Favorita	Conteo
0 CA	Water	280
1 FL	Water	255
2 NY	Water	260
3 TX	Water	235

## 4. ¿Cuál es el deporte / actividad favorita de los estudiantes que participan en el Censo escolar en cada estado?

- **Descripción:** Se investigó cuál es la actividad física favorita de los estudiantes en cada estado.
- **Insight Principal:** Caminar o hacer senderismo es la actividad física favorita de los estudiantes en todos los estados.
- **Interpretación:** Esta preferencia puede reflejar la facilidad de estas actividades, en especial caminar, que no requiere equipo especial y puede ser realizada al aire libre. Probablemente, esto hace que estas actividades sean populares entre los jóvenes.

```

conteo_deporte_favorito = (df.groupby(["Region", "Favourite_physical_activity"]).size()
                           .reset_index(name='Count'))

deporte_favorito_por_estado = conteo_deporte_favorito.groupby('Region').agg(
    Deporte_Favorito=('Favourite_physical_activity', lambda x: x.loc[x.idxmax()]),
    Conteo=('Count', 'max')
).reset_index()

deporte_favorito_por_estado

```

✓ 0.0s Abrir "deporte\_favorito\_por\_estado" en Data Wrangler

Region	Deporte_Favorito	Conteo
0 CA	Walking/Hiking	49
1 FL	Walking/Hiking	60
2 NY	Walking/Hiking	48
3 TX	Walking/Hiking	54

## 5. ¿Cuál es el nivel más alto de educación que planean alcanzar los estudiantes que participan en el Censo escolar en cada estado?

- **Descripción:** Se evaluaron las aspiraciones educativas de los estudiantes en términos del nivel educativo más alto que planean alcanzar.
- **Insight Principal:** El nivel educativo más alto que planean alcanzar la mayoría de los estudiantes en todos los estados es un título de licenciatura (*undergraduate degree*).
- **Interpretación:** Con estos conteos podemos mostrar que existe un alto nivel de aspiración académica entre los estudiantes, lo cual es alentador y sugiere un fuerte enfoque en la educación superior dentro de la población estudiantil.

```

conteo_nivel_educacion_deseado = (df.groupby(["Region", "Planned_Education_Level"]).size()
                                   .reset_index(name='Count'))

nivel_educacion_deseado_por_estado = conteo_nivel_educacion_deseado.groupby('Region').agg(
    Nivel_Educacion_Deseado=('Planned_Education_Level', lambda x: x.loc[x.idxmax()]),
    Conteo=('Count', 'max')
).reset_index()

nivel_educacion_deseado_por_estado

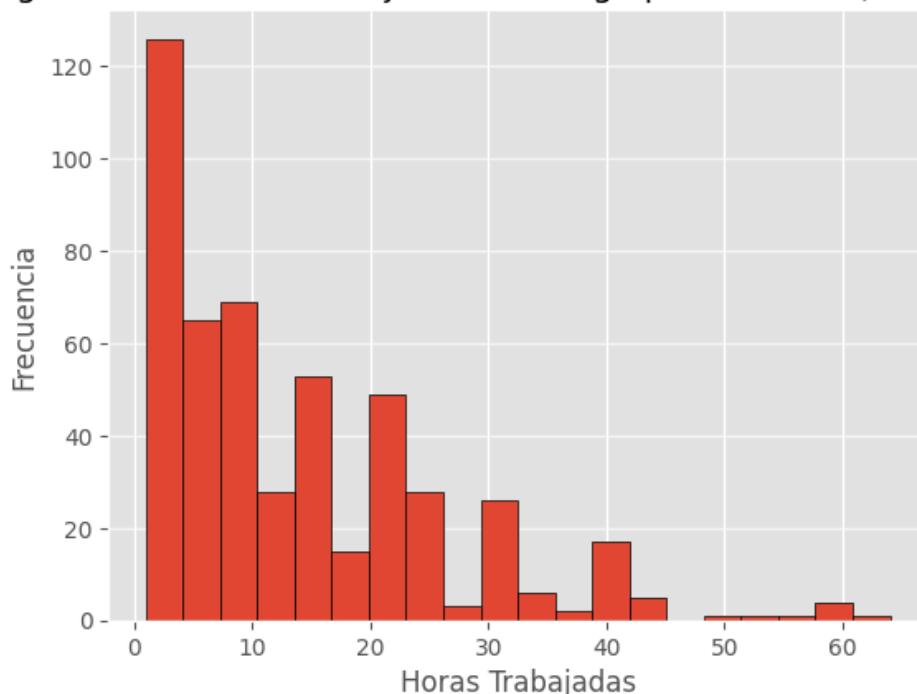
```

✓ 0.0s Abrir "nivel\_educacion\_deseado\_por\_estado" en Data Wrangler

Region	Nivel_Educacion_Deseado	Conteo
0 CA	Undergraduate degree	257
1 FL	Undergraduate degree	270
2 NY	Undergraduate degree	265
3 TX	Undergraduate degree	215

## 6. Tiempo que trabaja con pago el estudiante en la semana

Histograma de Horas Trabajadas con Pago por Semana (Excluyendo 0)



Análisis de la variable "Paid\_Work\_Hours" (Horas de trabajo remunerado)

Variable: Paid\_Work\_Hours

Descripción: Esta variable mide las horas que los estudiantes trabajan por semana de manera remunerada.

Análisis Realizado

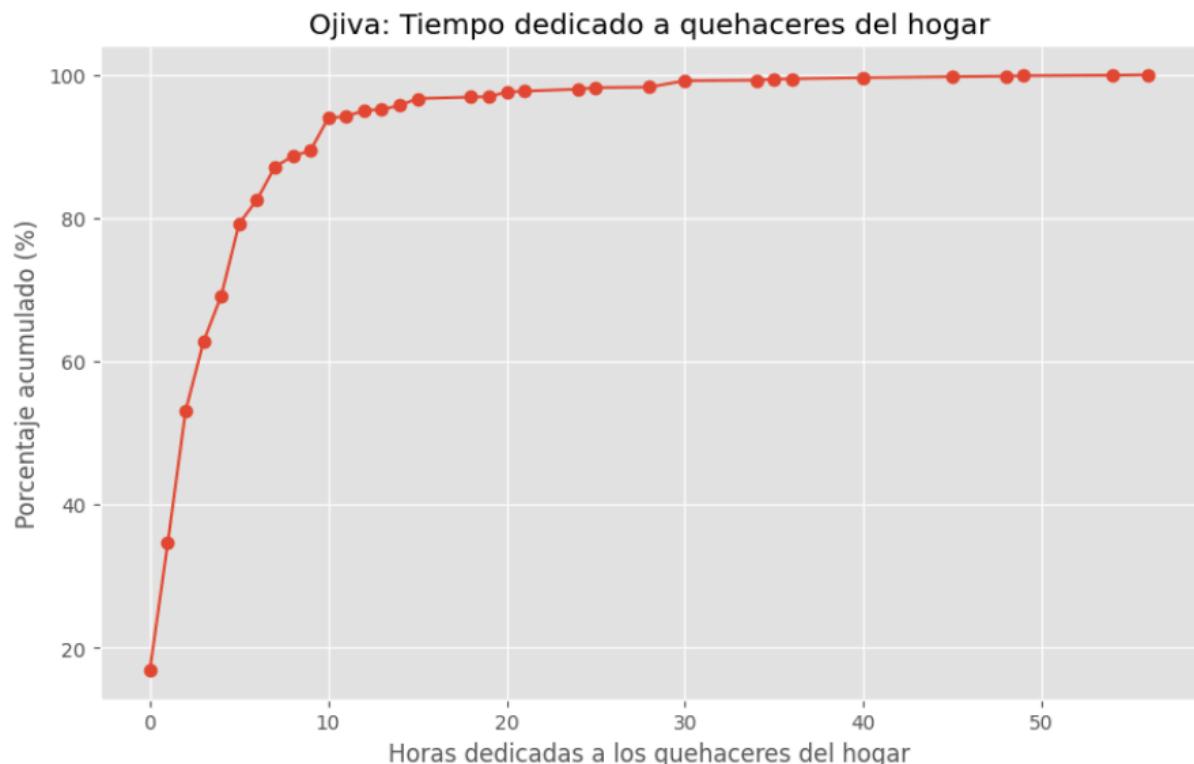
- Histograma: Se creó un histograma para visualizar la distribución de las horas de trabajo remunerado por semana entre los estudiantes.
- Medidas Clave Analizadas:
  - Porcentaje de estudiantes que trabajan máximo 13 horas a la semana:
    - Se calculó el porcentaje de estudiantes que trabajan hasta 13 horas por semana para entender cuántos estudiantes tienen una carga de trabajo ligera.
  - Porcentaje de estudiantes que trabajan mínimo 3 horas a la semana:
    - Se analizó el porcentaje de estudiantes que trabajan al menos 3 horas por semana, indicando cuántos tienen algún tipo de empleo remunerado.
  - Percentiles y Moda:
    - Se determinó el número mínimo y máximo de horas trabajadas por el 20% y 50% de los estudiantes, respectivamente.

- Se identificó el intervalo de moda, que es donde se encuentran la mayoría de las horas trabajadas.
- Promedio y Desviación Estándar:
  - Se calculó el promedio de horas trabajadas y la desviación estándar para comprender la tendencia central y la variabilidad en las horas de trabajo remunerado.
- Asimetría:
  - Se evaluó el tipo de asimetría (positiva, negativa o simétrica) en la distribución de las horas trabajadas, para entender si la mayoría de los estudiantes trabaja pocas horas o muchas horas.

#### Resultados e Insights:

- Distribución del Trabajo Remunerado: La mayoría de los estudiantes trabajan menos de 13 horas a la semana, indicando que la mayoría tiene trabajos a tiempo parcial o trabaja menos horas debido a sus responsabilidades académicas.
- Moda y Tendencia Central: El intervalo de moda muestra que hay un pico en el número de horas trabajadas, lo cual indica un rango común de trabajo entre los estudiantes.
- Variabilidad: La desviación estándar y el promedio sugieren que hay una variedad significativa en las horas de trabajo, con algunos estudiantes trabajando muchas más horas que otros.
- Asimetría: La asimetría de la variable puede indicar que la mayoría de los estudiantes trabaja menos horas, con algunos pocos trabajando muchas más horas.

## 7. Tiempo que dedican los estudiantes para ayudar a los quehaceres del Hogar



Análisis de la variable "Work\_At\_Home\_Hours" (Horas de trabajo en casa)

Variable: Work\_At\_Home\_Hours

Descripción: Esta variable mide las horas que los estudiantes dedican semanalmente a ayudar con los quehaceres del hogar.

Análisis Realizado

- Ojiva (Gráfica de Frecuencias Acumuladas): Se construyó una ojiva para visualizar cómo se distribuyen las horas dedicadas a los quehaceres del hogar entre los estudiantes.
- Medidas Clave Analizadas:
  - Porcentaje de estudiantes que dedican mínimo 15 horas a la semana:
    - Se determinó el porcentaje de estudiantes que dedican al menos 15 horas semanales a los quehaceres, indicando el compromiso significativo con las tareas del hogar.
  - Porcentaje de estudiantes que dedican máximo 5 horas a la semana:
    - Se analizó el porcentaje de estudiantes que dedican hasta 5 horas a la semana, mostrando aquellos que realizan pocas tareas domésticas.
  - Percentiles y Moda:
    - Se calcularon el número máximo y mínimo de horas de ayuda para el 15% y 35% de los estudiantes, respectivamente.
    - Se identificó el intervalo de moda, donde se concentra la mayoría de las horas de ayuda.
  - Promedio y Desviación Estándar:

- Se determinó el promedio de horas dedicadas a los quehaceres y la desviación estándar para entender la tendencia central y la variabilidad.
- Asimetría:
  - Se evaluó la asimetría en la distribución de las horas dedicadas a los quehaceres para identificar tendencias hacia la dedicación de más o menos tiempo.

Resultados e Insights:

- Distribución del Trabajo en Casa: Muchos estudiantes dedican menos de 5 horas a los quehaceres del hogar, sugiriendo que sus responsabilidades domésticas son limitadas o que priorizan sus estudios y actividades extraescolares.
- Compromiso con los Quehaceres: Un porcentaje menor de estudiantes dedica más de 15 horas a la semana, lo que indica un alto nivel de responsabilidad en el hogar.
- Tendencia y Variabilidad: El promedio y la desviación estándar sugieren una variabilidad moderada, con la mayoría de los estudiantes contribuyendo de manera similar en sus hogares.
- Asimetría: La distribución asimétrica puede indicar que hay una mayor concentración de estudiantes que dedican menos tiempo a los quehaceres, con algunos pocos que dedican mucho más tiempo.

## 8. Importancia de Reducir la Contaminación por Estado y Género

### Introducción

El análisis estadístico presentado se basa en un diagrama de cajas que compara la opinión sobre la importancia de reducir la contaminación en diferentes estados de los Estados Unidos, segmentada por género. Se utilizaron datos de cuatro estados: California (CA), Florida (FL), Nueva York (NY) y Texas (TX), y las opiniones se clasificaron en función de su importancia percibida en una escala de 0 a 1000.

### Resumen de Hallazgos

#### Importancia General por Género:

- **Mujeres:** En términos generales, las mujeres consideran más importante reducir la contaminación en comparación con los hombres. Esto se evidencia en la mediana más alta y en el rango intercuartílico (IQR) que se encuentra más cercano al valor máximo en la mayoría de los estados.
- **Hombres:** Aunque los hombres también consideran importante reducir la contaminación, sus opiniones muestran una mayor dispersión en algunos estados y una mediana más baja en comparación con las mujeres.

#### Importancia por Estado:

- **California (CA):** Es el estado donde tanto hombres como mujeres consideran más importante reducir la contaminación, con el IQR para ambos géneros

- ubicado cerca del valor máximo. Esto indica un fuerte consenso sobre la importancia de este tema.
- **Nueva York (NY):** Las mujeres en Nueva York muestran una alta importancia para reducir la contaminación, similar a California, pero con una mediana ligeramente más alta, sugiriendo un compromiso aún mayor en promedio.
- **Florida (FL):** Es el estado más heterogéneo en términos de opiniones, presentando bigotes superiores e inferiores largos para ambos géneros, lo que indica una alta variabilidad en las respuestas.
- **Texas (TX):** Similar a otros estados en términos de importancia, pero con menos variabilidad en las respuestas de los hombres, lo que sugiere una menor dispersión en las opiniones.

### **Homogeneidad y Heterogeneidad:**

- **Mujeres:** Muestran mayor homogeneidad en sus opiniones sobre la importancia de reducir la contaminación en 3 de los 4 estados, donde los bigotes superiores son más cortos (casi imperceptibles) y la dispersión de los datos es menor.
- **Hombres:** Presentan mayor heterogeneidad en Florida, con una amplia dispersión de opiniones reflejada en los bigotes más largos y un IQR más amplio.

### **Simetría en las Opiniones:**

- **Mayor Simetría (Mujeres en Florida):** En Florida, las mujeres presentan la mayor simetría en las opiniones, con bigotes superior e inferior largos y un IQR centrado, sugiriendo un equilibrio en la dispersión de las opiniones.
- **Mayor Asimetría (Hombres en California):** Los hombres en California muestran la mayor asimetría negativa, con un bigote inferior largo y un bigote superior casi indetectable, indicando que la mayoría de las opiniones se concentran entre la escala de 600 y 1000.

### **Tipos de Variables y Definiciones Importantes**

- **Variables Categóricas:** Estado (CA, FL, NY, TX) y Género (Hombre, Mujer). Estas variables son cualitativas y se utilizan para segmentar los datos y realizar comparaciones entre grupos.
- **Variables Cuantitativas:** Importancia de reducir la contaminación, medida en una escala de 0 a 1000. Esta variable es continua y cuantitativa, permitiendo medir la intensidad de la opinión sobre el tema.
- **Mediana:** El valor central que divide la distribución en dos partes iguales. Es menos sensible a los valores extremos y proporciona una buena medida de tendencia central cuando la distribución es asimétrica.
- **Rango Intercuartílico (IQR):** Mide la dispersión del 50% central de los datos. Es útil para entender la variabilidad de las opiniones dentro de cada grupo.
- **Bigotes:** Representan la extensión de los datos más allá del IQR, excluyendo los valores atípicos.

### **Conclusiones**

El análisis revela diferencias significativas en la percepción de la importancia de reducir la contaminación según el género y el estado. Las mujeres tienden a mostrar una mayor preocupación por el tema en la mayoría de los estados, con opiniones más homogéneas en comparación con los hombres, quienes presentan una mayor dispersión y asimetría en algunas regiones. Florida destaca por su alta variabilidad en las opiniones de ambos géneros, mientras que California muestra un fuerte consenso hacia la importancia del tema, especialmente entre las mujeres.

Este tipo de análisis es crucial para entender las percepciones y actitudes hacia cuestiones ambientales en diferentes segmentos de la población, lo que puede informar políticas y estrategias de comunicación más efectivas.

## 9. Análisis Estadístico de Horas de Sueño por Género

### Introducción

El análisis se centra en una tabla cruzada de indicadores que relaciona el número de horas de sueño durante las noches sin clases al día siguiente (fin de semana o días libres) con la variable de género. Este análisis permite observar patrones en los hábitos de sueño de hombres y mujeres y cómo difieren entre ellos.

### Resumen de Hallazgos

#### Distribución General de Horas de Sueño:

- **Promedio de Horas de Sueño:** En promedio, los estudiantes duermen 8.73 horas cuando no tienen clases al día siguiente. Este dato refleja una tendencia general hacia dormir más durante las noches sin responsabilidades académicas inmediatas.

#### Comparación por Género:

- **Porcentaje de Mujeres en la Muestra:** El 49.69% de los estudiantes de la muestra son mujeres, mostrando una distribución de género relativamente equilibrada en la muestra analizada.
- **Horas de Sueño por Género:**
  - **Mujeres:** En promedio, las mujeres duermen más tiempo que los hombres cuando no tienen clases al día siguiente. Además, las mujeres presentan un comportamiento de sueño más homogéneo, con una menor variabilidad en sus horas de sueño.
  - **Hombres:** Aunque los hombres duermen menos en promedio, presentan una mayor variabilidad en las horas de sueño. Esto podría reflejar patrones de sueño más inconsistentes o menos regulares entre los hombres.

#### Análisis de Curtosis:

- **Curtosis (Apuntamiento):** El análisis de curtosis revela que las mujeres tienen un histograma más puntiagudo (mayor curtosis), lo que sugiere que más

mujeres duermen un número específico de horas en comparación con una distribución más plana. En contraste, los hombres presentan un histograma más aplanado, indicando una mayor diversidad en las horas de sueño.

### **Visualización de la Distribución de Sueño:**

- El gráfico de distribución de horas de sueño muestra que tanto hombres como mujeres tienden a dormir entre 8 y 10 horas cuando no hay clases al día siguiente.

### **Tipos de Variables y Definiciones Importantes**

- **Variables Categóricas:**
  - **Género (Gender):** Variable categórica(Femenino, Masculino). Esta variable permite segmentar los datos para comparar los patrones de sueño entre hombres y mujeres.
- **Variables Cuantitativas:**
  - **Horas de Sueño (Sleep\_Hours\_Non\_Schoolnight):** Variable cuantitativa continua que representa el número de horas que los estudiantes duermen en noches sin clases al día siguiente.

### **Conclusiones**

El análisis sugiere que hay diferencias significativas en los patrones de sueño entre hombres y mujeres durante noches sin clases al día siguiente. Las mujeres tienden a dormir más horas en promedio sin tanta variabilidad. Por otro lado, los hombres presentan mayor variabilidad y una distribución más apuntada, indicando diferencias en los comportamientos de sueño que podrían ser relevantes para la planificación de horarios y políticas de salud en instituciones educativas.

Este tipo de análisis es crucial para entender los hábitos y necesidades de los estudiantes en términos de descanso, lo que puede contribuir a desarrollar políticas que promuevan un estilo de vida saludable y mejorar el rendimiento académico.

## **10. Análisis Estadístico de la Materia Preferida en el Colegio por Género**

### **Introducción**

El análisis presentado se basa en una tabla cruzada que relaciona la materia preferida en el colegio con la variable de género. Este tipo de análisis permite observar las preferencias académicas de estudiantes masculinos y femeninos en diversas asignaturas, y cómo estas preferencias se distribuyen entre los géneros.

### **Resumen de Hallazgos**

#### **Distribución General por Materia y Género:**

- **Arte (Art)**: Es notablemente preferida por las mujeres (116 mujeres frente a 29 hombres), lo que indica una tendencia significativa hacia las artes entre las estudiantes femeninas.
- **Computadoras y Tecnología (Computers and Technology)**: Preferida mayoritariamente por hombres (63 hombres frente a 21 mujeres), lo que refleja una inclinación hacia esta área entre los estudiantes masculinos.
- **Ciencias (Science)**: La distribución es relativamente equilibrada, aunque ligeramente inclinada hacia las mujeres (124 mujeres frente a 111 hombres), mostrando un interés casi igual entre ambos géneros en esta área.
- **Matemáticas y Estadísticas (Mathematics and Statistics)**: Es la materia con mayor preferencia entre los hombres (153 hombres frente a 101 mujeres), indicando una fuerte inclinación masculina hacia esta área académica.

### **Preferencias Específicas y Tendencias**

- **Mujeres**: Tienen una mayor preferencia por asignaturas como Arte, Inglés, Ciencias y Música. Este patrón sugiere una inclinación hacia disciplinas creativas y científicas.
- **Hombres**: Muestran una mayor preferencia por Matemáticas y Estadísticas, Historia, Educación Física, y Computadoras y Tecnología. Esto refleja un interés más pronunciado en áreas técnicas, históricas y físicas.

### **Conclusión General**

Este análisis proporciona una comprensión clara de las preferencias académicas de estudiantes de diferentes géneros, destacando áreas donde se observan diferencias significativas. Este tipo de información es valiosa para las instituciones educativas, ya que pueden adaptar sus programas y apoyo estudiantil para abordar estas preferencias y fomentar un ambiente académico inclusivo y diverso.

# Notebooks

- » 001\_cleaningData.ipynb
- » 002\_EDA\_SchoolCensus.ipynb