



# Análisis socioeducativo de los habitantes de la Ciudad de Buenos Aires

**Profesor:** Damian Dapuetto

**Tutor:** Héctor Alonso

**Grupo de Trabajo:** Lucia Buzzeo, Lucia Hukovsky,  
Jose Saint German, Juan Martín Carini

# Introducción

- Descubrir las principales variables intervinientes en el nivel máximo educativo alcanzado por la población de la Ciudad Autónoma de Buenos Aires (CABA).
- Generar un modelo de predicción aplicado a nuestra variable target “Nivel Máximo Educativo”, esto lo haremos implementando los siguientes modelos de clasificación:
  - **Árbol de decisión:** que construye un árbol durante el entrenamiento que es el que aplica a la hora de realizar la predicción.
  - **Bosque Aleatorio:** que es un conjunto (ensemble) de árboles de decisión combinados con bagging.

Para trabajar esta problemática, se ha recurrido a la [Encuesta Anual de Hogares](#) del Gobierno de la Ciudad de Buenos Aires para el año 2019. El dataset está disponible en la base de datos abiertos del GCBA. Esta encuesta contiene información demográfica, social, económica, educativa y de salud de 14319 habitantes de la Ciudad, la cual es una muestra representativa que permite obtener un vistazo de la población de la Ciudad.

# Estructura de los trabajos

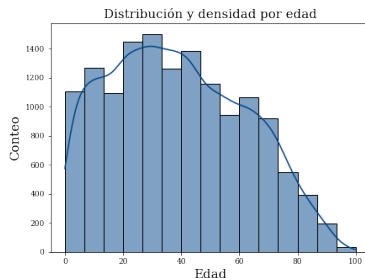
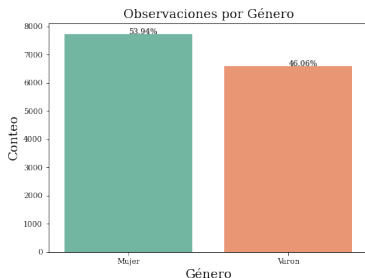
Este trabajo se ha dividido en 3 partes:

- 1 **Introducción a las variables del problema:** Se realiza un análisis de las variables del dataset. En el mismo se busca conocer su performance dentro del dataset. A la vez, se investiga cómo las variables interactúan entre sí. Esta parte es lo que se conoce como análisis univariado, bivariado y multivariado.
- 2 **Modelos analíticos:** En esta sección se entrenan diversos modelos analíticos y algoritmos que sirven para alcanzar los objetivos seteados para el presente proyecto. Como la variable objetivo es categórica, se realizan modelos de clasificación.
- 3 **Conclusión:** Se alcanzan conclusiones finales sobre los hallazgos. Además, se discuten posibles limitaciones y se plantean futuras líneas de análisis, a partir del análisis presente.

# Análisis exploratorio de los datos (EDA)

# Análisis univariado

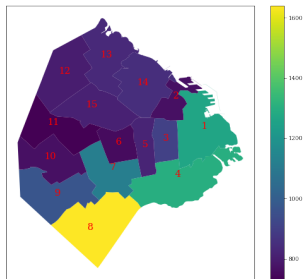
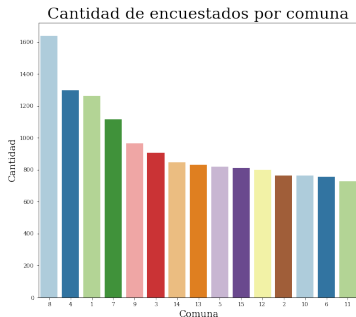
Comenzamos con un pantallazo general sobre las primeras cualidades de los datos, como muestra representativa para la EPH, sobre quiénes son los ciudadanos representados en el dataset.



En la variable género los datos parecen equilibrados en las categorías. Para el caso de la variable “edad”, la distribución se asemeja a la de una normal.

# Análisis univariado

Seguimos observando la variable “comuna”:

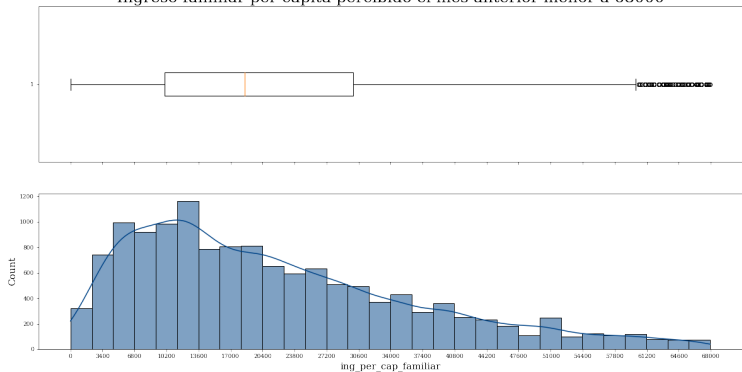


Observando ambos gráficos vemos que las comunas 1,4,7 y 8 tienen mayor cantidad de casos.

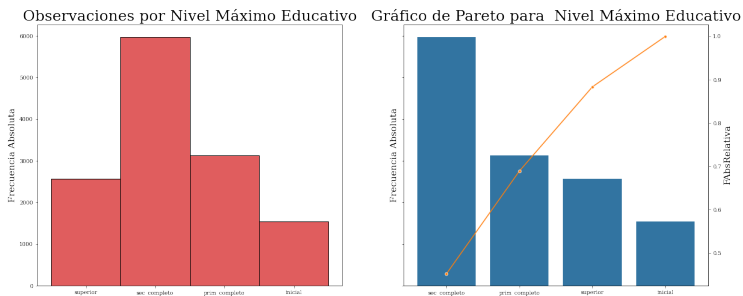


# Análisis univariado

Ingreso familiar per capita percibido el mes anterior menor a 68000



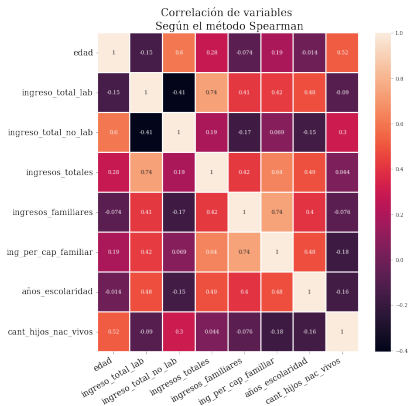
# Análisis univariado



Podemos observar que el nivel máximo educativo más alcanzado es el secundario completo, seguido por el primario. Contrario de lo que habíamos intuido anteriormente, el nivel superior quedó en tercer lugar. Adicionalmente, el nivel secundario y primario explican casi el 77% de los datos.

# Análisis bivariado

Para comenzar el análisis bivariado del problema, realizamos diferentes heat maps para ver si algo nos llama la atención entre las variables numéricas.



- No se observan fuertes correlaciones.
- “años\_escolaridad” correlaciona moderadamente bien con variables relacionadas al ingreso.
- La principal correlación positiva es “años\_escolaridad” con ingreso familiar per cápita (“ing\_per\_cap\_familiar”).

# Análisis bivariado

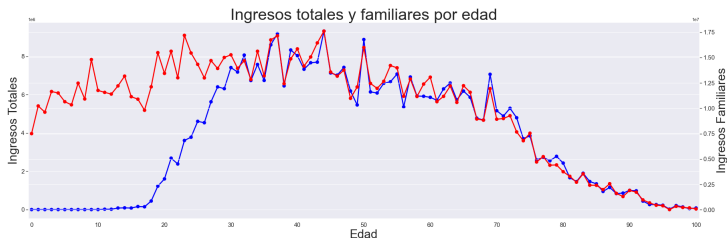
Corremos una tabla de correlación y filtramos las de valores más altos:

	Variable 1	Variable 2	Correlación
2	ingreso_total_lab	ingresos_totales	0.80
6	ingresos_familiares	ing_per_cap_familiar	0.76
4	ingresos_totales	ing_per_cap_familiar	0.62
5	ingresos_totales	años_escolaridad	0.60
1	edad	ingreso_total_no_lab	0.57
7	años_escolaridad	Target	0.57
3	ingreso_total_lab	años_escolaridad	0.54

- Como es esperable, hay alta correlación entre las variables relacionadas al ingreso.
- A su vez, encontramos una alta correlación (66 %) entre los ingresos y los años de escolaridad.
- Y también observamos una relación positiva entre la edad y los ingresos totales.

# Análisis bivariado

Observaremos la relación entre los ingresos totales de cada hogar y los ingresos familiares por edad:

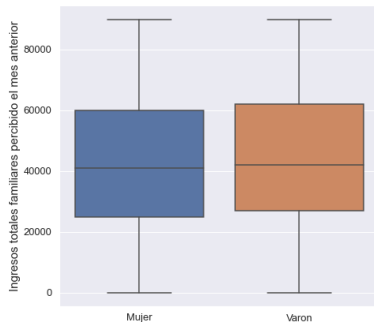
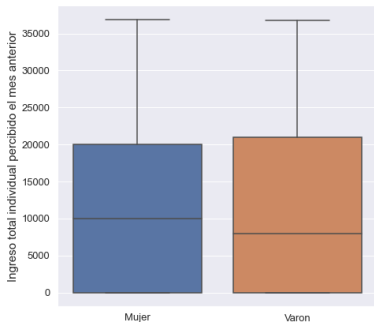


Se puede ver que desde los 30 años en adelante el ingreso total de la persona se corresponde con el ingreso familiar. Por ende suele haber un único ingreso fuerte por grupo familiar.

# Análisis bivariado

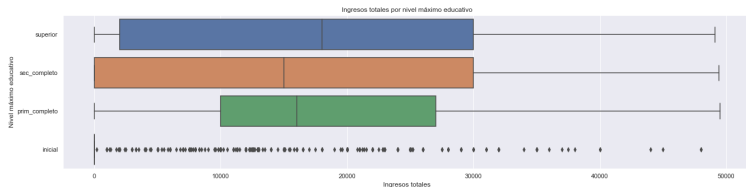
# Análisis bivariado

Veamos como se relacionan las categorías de ingresos con el genero de los encuestados:



Y resulta que con las variables de ingreso, no parece haber nada disruptivo. Salvo que los hombres parecieran tener ingresos totales y familiares mayores que las mujeres, pero no pareciera que haya distribuciones desiguales en los años de escolaridad.

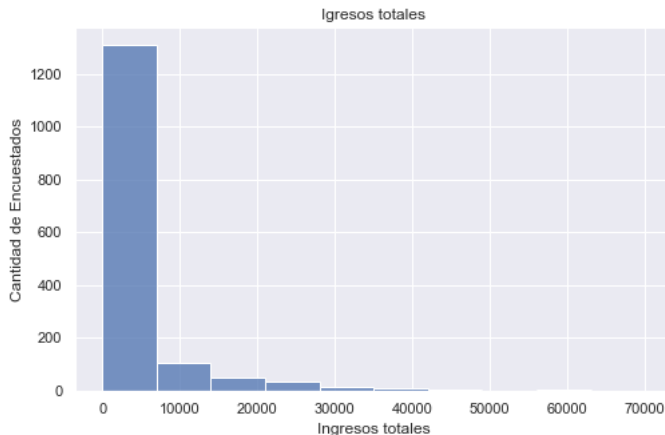
# Análisis bivariado



Parece que para el nivel inicial la remoción de outliers en otra categoría sigue siendo insuficiente para mostrar la distribución real de la variable. Echemos un vistazo a los valores de esta categoría:



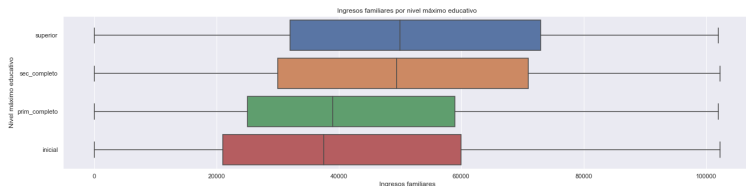
# Análisis bivariado



Lógicamente, la enorme mayoría de los ingresos tienen el valor inicial de 0, puesto que incluye a personas que en ese momento estaban cursando su educación inicial, por lo que tenían entre 2 y 6 años.

# Análisis bivariado

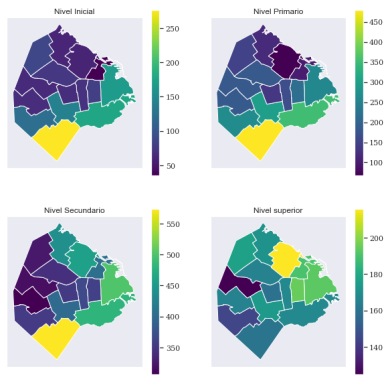
Entonces, es correspondiente analizar como se distribuyen los ingresos familiares con respecto a nuestro target:



En definitiva, se observa un corrimiento de los valores centrales (dentro de la caja) hacia la izquierda a medida que aumenta el nivel educativo.

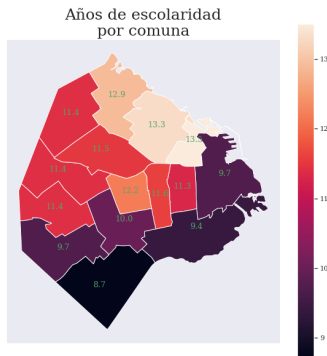
# Análisis bivariado

En primer lugar, analizaremos la relación de nuestro target con cada comuna:



Se observa que en el sur de la ciudad hay mayor cantidad de encuestados con niveles de inicial, primario y secundario completo, mientras que el norte (particularmente el barrio de Palermo) tiene mayor cantidad de personas con estudios superiores. En menor medida también las comunas del este (comúnmente llamado el “centro de la ciudad”) destacan por la cantidad de encuestados con nivel superior.

# Análisis bivariado



Lo que podemos observar en los últimos dos gráficos es una clara división geográfica del nivel educativo:

- Las comunas del norte son las que tienen mayor nivel educativo.
- Las comunas del centro tienen niveles medios.
- Las comunas del sur (con las comunas 6 en el centro de la ciudad como outlier) y la comuna 1 en el este son las que tienen niveles más bajos.

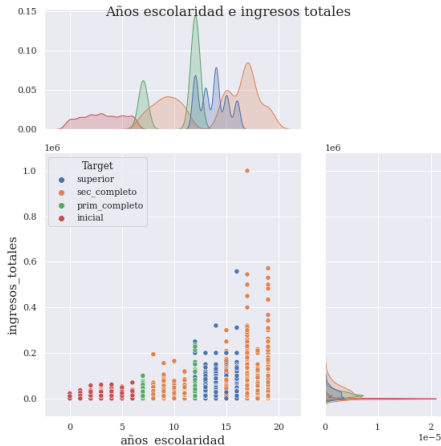
# Análisis multivariado

Probamos de cruzar años de escolaridad, nivel máximo educativo y los ingresos totales.

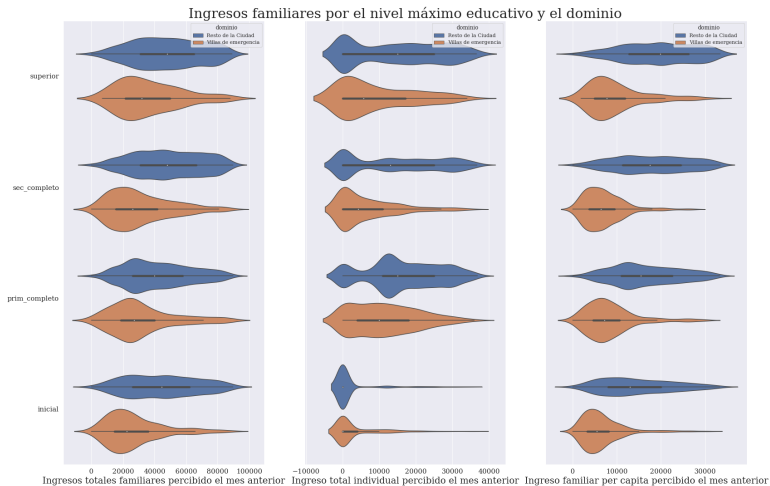
Al visualizar

el gráfico podemos observar que:

- Hasta los 6 años, como era esperable, todos los casos llegan al nivel inicial.
- Vemos dos años en que aparece el primario completo: 7 y 12 años. Estimamos que se debe a la división entre los que comenzaron su educación en la primaria y los que comenzaron en el nivel inicial.
- A partir de los 12 años vemos un aumento consistente de los ingresos totales.



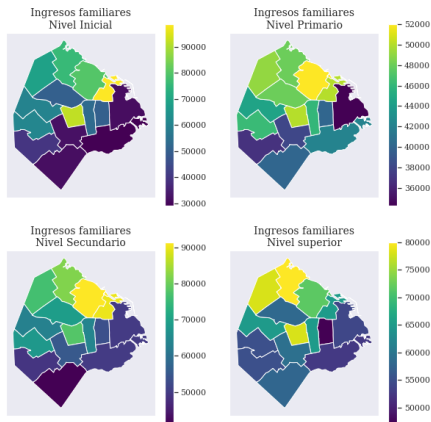
# Análisis multivariado



Aquí obtuvimos un descubrimiento interesante: no importa el nivel máximo educativo, los casos que no provienen de villas de emergencia (dominio="villas\_de\_emergencia") obtienen en promedio ingresos más altos en todos los niveles educativos. El alcanzar estudios superiores no parece homogeneizar ambos conjuntos. Esto se puede observar en el segundo gráfico, ya que el violín naranja acumula mayor cantidad de casos hacia la derecha, en comparación con los violines azules que tienen una mayor distribución.

# Análisis multivariado

Luego, observamos los ingresos familiares según el máximo nivel educativo alcanzado:



Aquí, podemos observar que a medida que avanza el nivel educativo máximo se atenúan levemente las diferencias de ingresos familiares entre comunas. Queda pendiente cruzar estos datos con la edad, para saber si el hecho de incluir a menores de edad está sesgando los valores para nivel inicial, primario y secundario.



Comenzamos transformando algunas variables para poder trabajar con los algoritmos:

- recategorizando la variables “Target” en variables numéricas, es decir, a cada nivel educativo le asignamos un valor numérico del 1 al 4:
  - **inicial**= 1,
  - **prim\_completo**= 2,
  - **sec\_completo**= 3,
  - **superior**= 4,
- reagrupamos la variable “comuna” por regiones para reducir la dimensionalidad (en norte, centro, sur y oeste),
- y por último renombramos algunas variables para que sean más cortas.

# Modelos analíticos