

**Universidad Nacional del Centro de la
Provincia de Buenos Aires**

Facultad de Ciencias Exactas

Ingeniería de Sistemas

Materia: Fundamento de la Ciencia de Datos



Grupo 6

Integrantes:

Nicolas Angladette

nicoruso37@outlook.com

Luciano Adaglio

luchiadaglio7@gmail.com

Agustin Lopez

agustinlopez1244@gmail.com

1. Introducción

En este informe se hablará de la experiencia de los alumnos mencionados trabajando y analizando el dataset relacionado al Arbolado Público de Buenos Aires (Censo 2017-2018). El objetivo de nuestra investigación, teniendo en cuenta la información brindada en internet y nuestro análisis, fue utilizar estos datos recolectados para plantear hipótesis relacionadas a problemáticas que se enfoquen en una parte mas urbanística que medioambiental, ya que, según la información de la página web de la ciudad de Buenos Aires, el dataset incluye detalles sobre la ubicación georreferenciada de cada ejemplar, especie, altura, diámetro y otras características relevantes para el análisis y gestión de los espacios verdes en el ámbito urbano. Creemos que la utilidad del conjunto es para la toma de decisiones de plantación de árboles teniendo en cuenta como estos se distribuyen y cómo pueden afectar al tránsito de la ciudad y a sus ciudadanos, así como también a la parte visual de la misma.

Identificamos problemas respecto a los datos y a la falta de información de los mismos, ya que algunas variables del dataset no estaban del todo claras a que se refieren sus datos o qué valores podrían tomar, lo que nos llevó a una investigación a parte, donde primero debemos volvernos “expertos” en cuanto a arbolado y urbanismo, para lograr entender mejor qué quiere decirnos cada valor que aparece en nuestro dataset.

El trabajo fue inicialmente desarrollado en Google Colab para posteriormente correrlo en una Jupyter Notebook.

La manera de estructurarlo fue la siguiente:

1. **Análisis exploratorio de los Datos (sección 2):**

Nuestro primer desafío fue el análisis de los datos, debido a que es una tarea muy importante, ya que, de no limpiar y procesar bien los datos que uno tiene y además no entenderlos, los estudios y conclusiones que podemos sacar de los mismos no serán tan buenas.

Esta parte del proyecto se explica en la sección [EDA](#).

2. Planteo y pruebas de hipótesis (sección 3):

Luego de la limpieza y procesamiento adecuado del dataset, iniciamos la tarea del planteamiento y prueba de las hipótesis solicitadas.

Esta parte del proyecto fue la más entretenida y a su vez, la más complicada, ya que no podíamos plantear hipótesis que no tengan un objetivo o algún sentido real que vaya de acuerdo al porqué se tomaron estos datos.

Esta parte del proyecto se explica en la sección [Hipótesis](#), siendo cada hipótesis una subsección donde se muestra la hipótesis planteada, la manera de abordarla y los resultados obtenidos y la conclusión de los resultados.

3. Conclusiones del trabajo (sección 4):

Para finalizar el informe, se incluye una sección la cual contiene reflexiones tanto de nuestro análisis del dataset dado para realizar el trabajo, como de la materia y sus contenidos y además, de la importancia de la Ciencia de Datos para la toma de decisiones ante una problemática o cuestión.

Esta parte del proyecto se explica en la sección [Conclusiones](#).

2. Análisis Exploratorio de los Datos (EDA)

Para iniciar el análisis de los datos, lo primero que hicimos fue descargar el dataset para empezar a ver los tipos de las variables y la cantidad de valores no nulos, además de la cantidad total de filas y demas, esto lo hicimos con la función `info()` sobre el data frame donde guardamos el dataset:

```
dataset.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 370180 entries, 0 to 370179
Data columns (total 18 columns):
#   Column                Non-Null Count  Dtype  
---  -
0   long                   354838 non-null float64
1   lat                    354838 non-null float64
2   nro_registro           370180 non-null object
3   tipo_activ             370180 non-null object
4   comuna                 370180 non-null int64  
5   manzana                224140 non-null object
6   calle_nombre           370087 non-null object
7   calle_altura           364677 non-null object
8   calle_chapa            363721 non-null object
9   direccion_normalizada  355941 non-null object
10  ubicacion              361884 non-null object
11  nombre_cientifico      370180 non-null object
12  ancho_acera            367083 non-null object
13  estado_plantera        370180 non-null object
14  ubicacion_plantera     368776 non-null object
15  nivel_plantera         368828 non-null object
16  diametro_altura_pecho  369894 non-null float64
17  altura_arbol           365858 non-null float64
dtypes: float64(4), int64(1), object(13)
memory usage: 50.8+ MB
```

Gráfico 1: Resultado de aplicar la función `info()` sobre el dataset, vemos la cantidad de filas, los tipos de cada columna y la cantidad de filas no nulas del dataset.

Lo primero que analizamos fue la cantidad total de filas del dataset y los tipos de las variables, sobre todo de las variables que se declaran como `float64` para ver si podía tratarse en realidad de alguna variable para la que podemos usar menos bits y las que son `object` para ver qué valores toman, siendo las candidatas a la limpieza.

De este resultado, pudimos ver que la variable con más cantidad de datos faltantes era `manzana`, donde solo el 60% de las filas tenían valor en este campo, el 40% restante era nulo. Luego, comenzamos a analizar una por una las variables y que tipo (no de programación) de variables eran, si eran cuantitativas (continuas/discretas), cualitativas (ordinales/nominales) o dicotómicas.

Antes de iniciar la limpieza, hicimos una copia del dataset original para evitar dañar los datos originales y tener que descargar el dataset de nuevo, y también aplicamos la función `head()` para ver los primeros cinco registros y los valores que toman para cada variable.

Analizando cada variable utilizando diferentes funciones como `unique()`, `isna()`, entre otras, decidimos tomar algunas decisiones:

- Eliminación de las filas donde las variables `long` y `lat` eran `null`, ya que estos registros tampoco contaban con información de la ubicación del árbol censado y además eran solo el 4,41% de las filas del dataset.
- Limpieza de la columna `tipo_activ` para dejarla con dos valores únicos ya que presentaba diferencias en cómo se escriben estos valores únicamente.
- Preprocesamiento de la columna `manzana`, indicando si es partida o no, es decir, si está subdividida en partes más chicas.
- Limpieza de la columna `ubicación`, estandarizamos todo a mayúsculas y solo dejamos una palabra.
- Preprocesamiento de columnas `calle_altura` y `calle_chapa`.
- Eliminación de columna `direccion_normalizada` ya que se deriva de `calle_altura` y `calle_chapa` y no aporta nueva información.
- Preprocesamiento de `ancho_acera`, decidimos pasar a `nan` ciertos valores ya que no sabemos que pueden significar, además de que no representan una cantidad importante de filas, decidimos no considerarlos.
- Limpieza de `estado_plantero`, `ubicacion_plantero` y `nivel_plantear`, lo que hicimos fue estandarizar sus valores.
- Por último, pasamos las columnas `diametro_altura_pecho` y `altura_arbol` a `float`.

Luego de todo esto, hicimos un `info()` del nuevo dataset limpio y quedó así:

```

<class 'pandas.core.frame.DataFrame'>
Index: 354838 entries, 0 to 370172
Data columns (total 17 columns):
#   Column                Non-Null Count  Dtype
---  -
0   long                   354838 non-null float64
1   lat                   354838 non-null float64
2   nro_registro          354838 non-null object
3   tipo_activ            354838 non-null object
4   comuna                354838 non-null int64
5   manzana              214188 non-null Int64
6   calle_nombre         354838 non-null object
7   calle_chapa          354838 non-null Int64
8   ubicacion            347948 non-null object
9   nombre_cientifico     354838 non-null object
10  ancho_acera           352677 non-null float64
11  estado_plantera       354838 non-null object
12  ubicacion_plantera    354035 non-null object
13  nivel_plantera        354014 non-null object
14  diametro_altura_pecho 354562 non-null float64
15  altura_arbol          350660 non-null float64
16  manzana_partida       354838 non-null int64
dtypes: Int64(2), float64(5), int64(2), object(8)
memory usage: 49.4+ MB

```

Gráfico 2: Dataset limpio, vemos que se agrega la columna manzana_partida y que ahora tenemos menos registros en total que antes.

En base a toda esta exploración de los datos, comenzaron a aparecer ciertas preguntas, como por ejemplo, cuáles son los árboles más altos (si además el diámetro tiene alguna relación con la altura, es decir, si los más altos además son los más anchos), cuáles especies son las más dominantes, que tan ocupadas están las planteras de Buenos Aires, cual es la diferencia entre la distribución de árboles por las distintas comunas de la ciudad, teniendo en cuenta diferencias de tamaño entre comunas, preguntas las cuales, nos llevaron a plantear y responder las hipótesis planteadas en la sección 3.

3. Hipótesis planteadas y resolución

3.1. Hipótesis 1 (univariada):

3.1.1. Definición de la hipótesis

La mayoría de las planteras de la Ciudad de Buenos Aires se encuentran “ocupadas” o “sobreocupadas”, lo que indica una alta presión sobre el arbolado urbano existente.

Datos requeridos: estado_plantera

Serviría para planificar nuevas plantaciones o reemplazos.

3.1.2. Estrategia de abordaje

Para comprobar la hipótesis, calculamos los porcentajes de las categorías de estado_plantera, o sea, dividimos la cantidad total de planteras en: vacía, subocupada, ocupada, sobreocupada, sobreocupada parc. cerrada, parcialmente cerrada, cerrada.

3.1.3. Resultados obtenidos y discusión

El resultado nos da el siguiente gráfico:

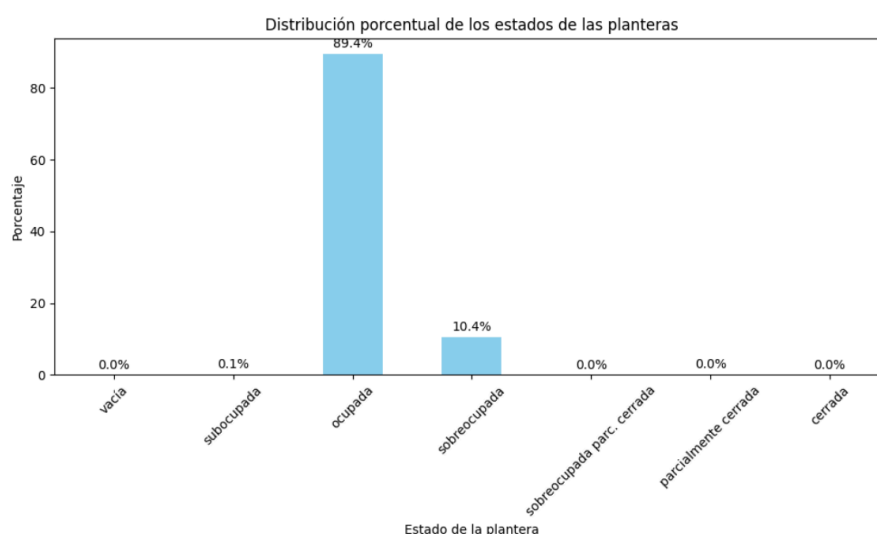


Gráfico 1: gráfico de barras con los porcentajes de las planteras de la ciudad.

Se puede apreciar claramente que la gran mayoría de las planteras se encuentran ocupadas o sobreocupadas, lo que **se confirma la hipótesis**.

Por consiguiente, se sugiere que la ciudad debería planificar nuevas plantaciones o reemplazos para equilibrar el uso de las planteras y evitar la saturación.

3.2. Hipótesis 2 (univariada):

3.2.1. Definición de la hipótesis:

Algunas especies dominan el arbolado urbano de Buenos Aires, causando alergia, como el plátano de sombra.

Datos requeridos: nombre_cientifico

3.2.2. Estrategia de abordaje

Para comprobar esta hipótesis, tomamos las 10 especies más frecuentes del dataset y el resto de especies las agrupamos en una categoría llamada otros y mostrar qué porcentaje abarca cada especie sobre el total de árboles.

3.2.3. Resultados obtenidos y discusión

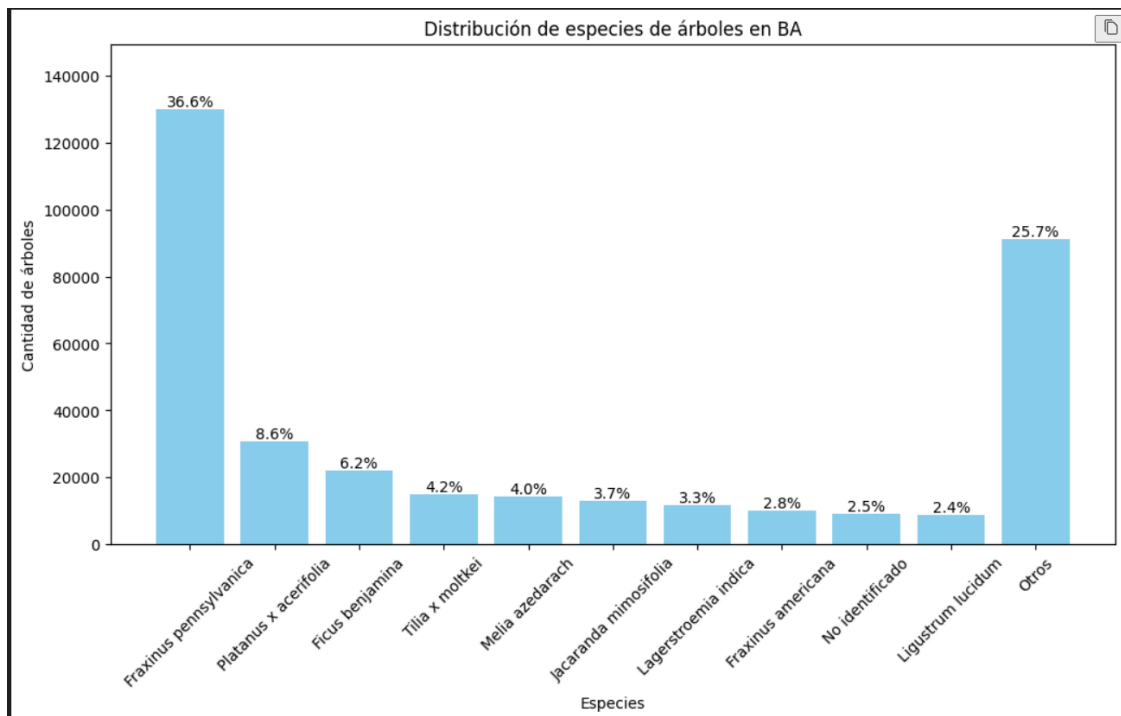


Gráfico 1: Gráfico de barras que muestra las 10 especies más abundantes de la ciudad de Buenos Aires.

Vemos claramente que el 8,6% del total de árboles de la ciudad de Buenos Aires corresponden a la especie *Platanus x acerifolia*, conformando el segundo grupo más abundante en la ciudad. Con estos resultados podemos decir que el *Platanus* es una especie abundante en la ciudad, por lo tanto, podemos decir que **se confirma la hipótesis**.

Se sugiere la revisión y planificación de políticas de gestión del arbolado para mitigar el impacto en la salud de los ciudadanos alérgicos, ya que la gran presencia de *Platanus x acerifolia* en el arbolado urbano implica una alta exposición poblacional al polen alergénico que produce esta especie.

A continuación, se muestran imágenes de algunas de las especies más abundantes en la ciudad:



Fraxinus Pennsylvanica, también conocido como fresno rojo americano, carácter invasor, con raíces superficiales y agresivas. Contiene un polen que causa alergia.



Platanus x acerifolia, mejor conocido como plátano de sombra (el objetivo de nuestra hipótesis).

Contiene un polen alérgico.



Ficus Benjamina, raíces invasivas comprometiendo cañerías y veredas, produce higos no comestibles que ensucian las veredas cuando caen.



Jacaranda mimosifolia, mejor conocida como jacaranda, una de las especies de árboles más emblemáticas de la ciudad de Buenos Aires.

3.3. Hipótesis 3 (univariada):

3.3.1. Definición de la hipótesis

La distribución del arbolado urbano en la Ciudad Autónoma de Buenos Aires no es equitativa entre comunas; algunas presentan una mayor concentración de árboles que otras.

Datos requeridos: comuna

Esta hipótesis puede servir para que, en caso de que haya algún proyecto de forestación, saber que comuna pueda tener prioridad.

3.3.2. Estrategia de abordaje

Para comprobar esta hipótesis, hicimos dos cálculos: la densidad de árboles por comuna (cantidad de árboles en la comuna/superficie(km²)) (*gráfico 1*) y el porcentaje de árboles totales que tiene cada comuna (*gráfico 2*). Los datos de las superficies los obtuvimos de [acá](#).

El primero es el más útil, ya que lo que nos interesa es la concentración, no la cantidad. El segundo nos da una idea de cómo los árboles están distribuidos por la ciudad.

3.3.3. Resultados obtenidos y discusión

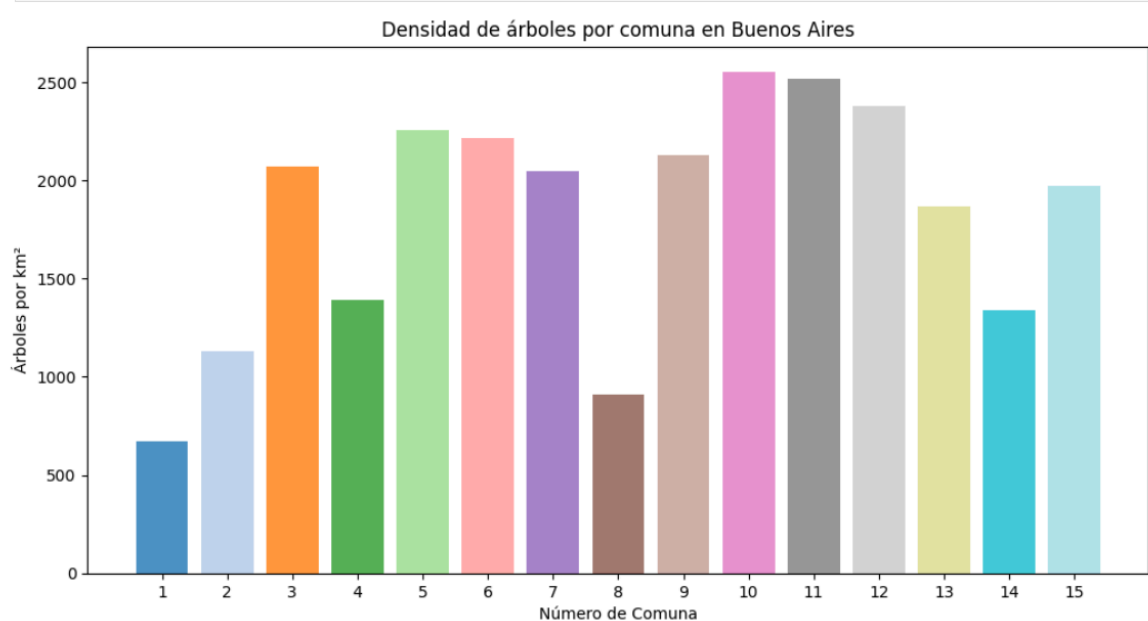


Gráfico 1: gráfico de barras que muestra la comparativa de cantidad de árboles por km² por comuna.

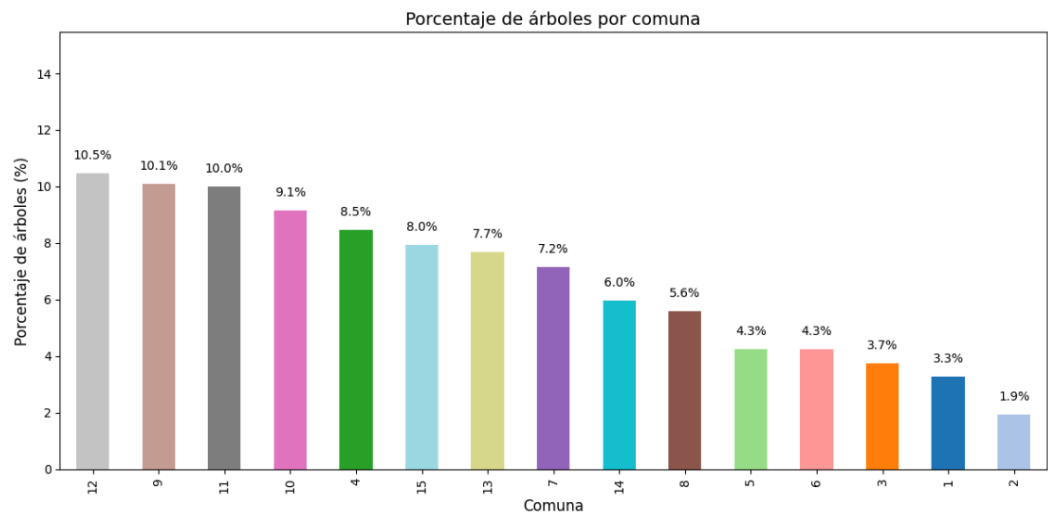


Gráfico 2: gráfico de barras donde vemos el porcentaje de árboles que posee cada comuna sobre el total de árboles registrados.

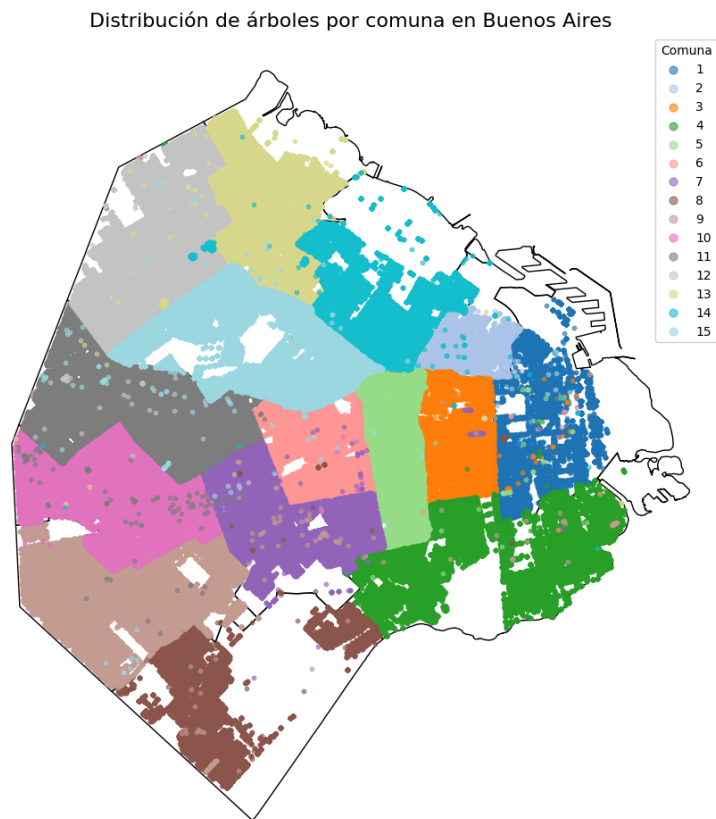


Gráfico 3: mapa de la ciudad de Buenos Aires donde se pinta con cada color los árboles en cada comuna, los espacios en blanco son espacios verdes.

Toda decisión tomada está sujeta a un cierto error, debido a algún error en las mediciones, ya que hay árboles que por su color deberían pertenecer a una comuna, pero por sus coordenadas se ubica en otra.

Vemos una clara diferencia entre la densidad arbórea entre las comunas, habiendo comunas como la 11 y 12 las cuales rozan los 2500 árboles por km² y comunas como la 1, la cual apenas pasa los 500. Con esta información, podemos decir que **se confirma la hipótesis** planteada.

3.4. Hipótesis 4 (bivariada):

3.4.1. Definición de la hipótesis:

Algunas comunas tienen desbalance, con pocas especies “eficientes” frente a especies de menor rendimiento ambiental.

Datos requeridos: nombre_cientifico, comuna

3.4.2. Estrategia de abordaje:

Para estudiar esta hipótesis fuimos desarrollando paso por paso el estudio, primero buscamos ver si existe algún tipo de relación entre el tamaño de las comunas (en km²) y la cantidad de árboles:

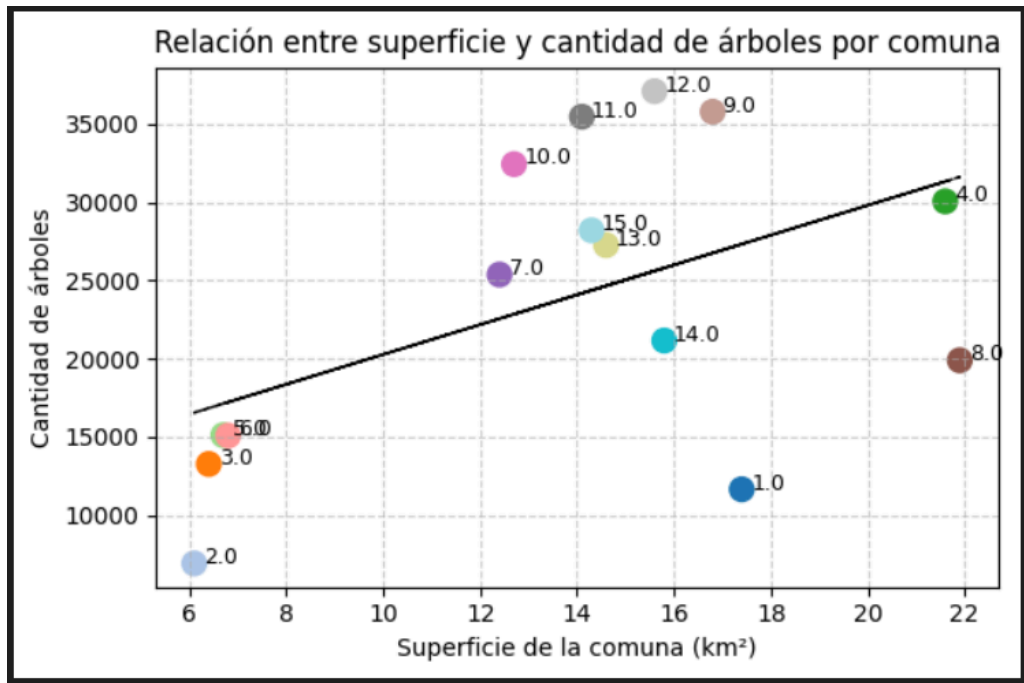


Gráfico 1: Scatter-Plot donde cada punto representa cada comuna de la ciudad. Se ve una clara tendencia que, a mayor superficie de la comuna, mayor cantidad de árboles hay en ella.

El análisis lo continuamos con el gráfico 1 de la hipótesis 3 analizamos la densidad de árboles para cada comuna, es decir, la cantidad de árboles por km² en cada comuna, el cual es un valor muy importante de analizar para saber si la zona tiene un número “sano” de árboles o está “pelada”.

Luego, vemos una comparativa de las distribuciones de especies eficientes entre comunas:

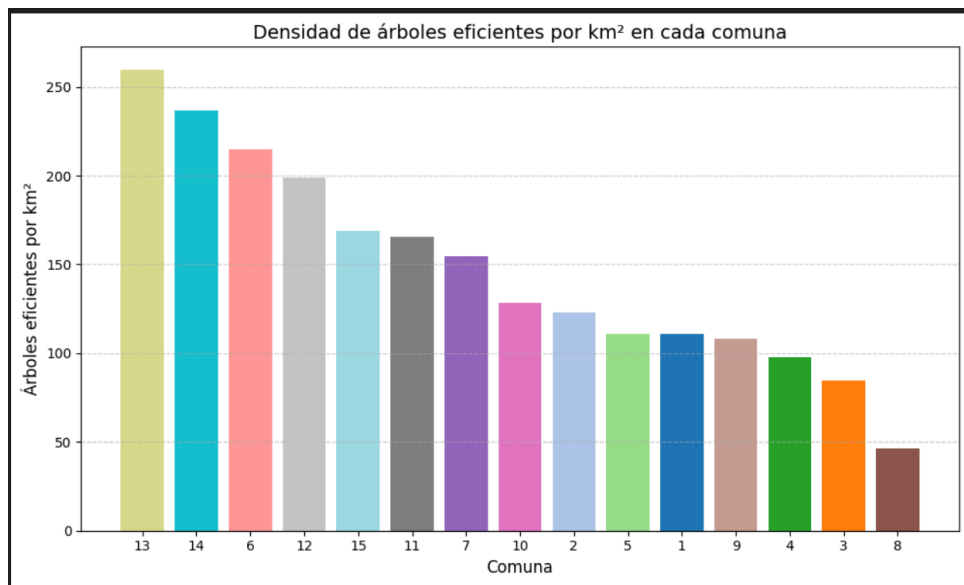


Gráfico 2: gráfico de barra, compara la cantidad de árboles eficientes en cuanto a la toma de CO₂ por km².

Además de esto, decidimos utilizar un test de hipótesis para verificar si los resultados obtenidos gráficamente son además, correctos estadísticamente.

Primero, formulamos las hipótesis estadísticas:

- **Hipótesis nula:** No hay diferencia en la proporción de especies eficientes entre comunas.
- **Hipótesis alternativa:** Sí hay diferencia, es decir, algunas comunas tienen más especies eficientes que otras.

Segundo:

- Hacemos un test no paramétrico:
 - Prueba Chi-cuadrado de independencia entre variables categóricas: sirve para verificar si hay asociación entre dos variables categóricas, es decir, si la proporción de especies eficientes depende de la comuna.
- Verificamos que se cumplan los supuestos:
 - Datos Categóricos
 - Independencia de las variables
 - Frecuencias esperadas: Cada celda en la tabla de contingencia debe tener una frecuencia esperada de al menos 5

categoria comuna	eficiente	otra
1	961.201303	10685.798697
2	570.267108	6339.732892
3	1094.896341	12172.103659
4	2481.280877	27584.719123
5	1247.160135	13864.839865
6	1244.766829	13838.233171
7	2094.720656	23287.279344
8	1642.138193	18255.861807
9	2954.165180	32841.834820
10	2677.614517	29767.385483
11	2927.921339	32550.078661
12	3062.689238	34048.310762
13	2250.285561	25016.714439
14	1746.453334	19415.546666
15	2328.439389	25885.560611

Gráfico 3: tabla de contingencia con las frecuencias de cada combinación

Al saber que son datos categóricos, hay independencia de las variables ya que cada observación es de un árbol único y al ver que las celdas de la tabla de contingencia son números mayores que 5, podemos concluir que se puede realizar el test de chi cuadrado.

Luego, realizamos el test habiendo cumplido todos los supuestos:

```
Chi-cuadrado: 6703.389
Grados de libertad: 14
p-valor: 0.00000
```

Gráfico 4: resultados del test Chi-cuadrado para ver si existen diferencias significativas en las proposiciones de las comunas.

3.4.3. Resultados obtenidos y discusión

Los resultados obtenidos se muestran a continuación un gráfico de barras apiladas que muestra la situación de los árboles censados en cuanto al porcentaje de árboles eficientes contra los no eficientes:

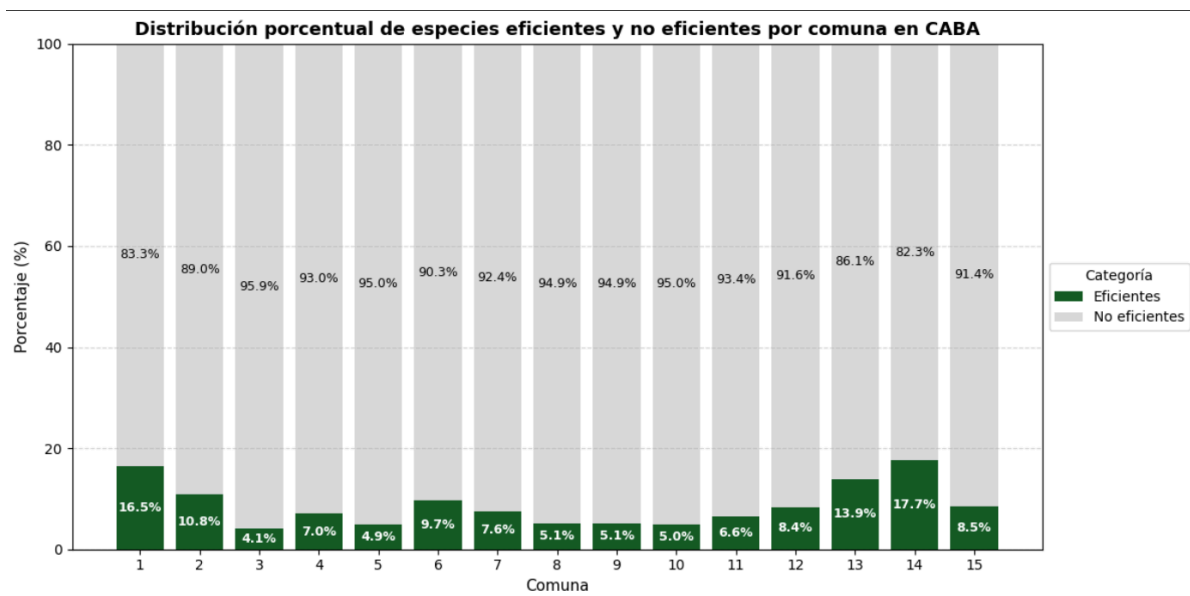


Gráfico 5: gráfico de barras apiladas marcando los porcentajes eficientes y no eficientes para cada comuna.

Podemos observar una cantidad muy escasa de árboles eficientes en la mayoría de las comunas y además, como se planteó en la hipótesis, existe un desbalance en la distribución de especies eficientes entre las comunas, teniendo comunas con una cantidad aceptable de árboles como la comuna 14 (17,7%) y otras como la 3, siendo la que menos tiene, con un 4,1%. Contrastamos la hipótesis además, al ver que $p < 0.05$, rechazamos la hipótesis nula y concluimos que hay diferencias significativas entre la cantidad de especies eficientes por comunas.

Es importante usar los gráficos 2 y 5 de esta hipótesis y el gráfico 1 de la hipótesis 3 en conjunto para realmente ver la situación forestal, ya que ese valor tan alto en el gráfico 5 para la comuna 1 nos dice que tiene un nivel aceptable de especies eficientes, pero si vemos los demás gráficos, nos damos cuenta de que esto sucede ya que la cantidad de árboles por km^2 que tiene esta comuna es mucho más baja que otras, es decir, tiene mayor proporción de eficientes pero por el hecho de que tiene menos densidad arbórea.

Con todos estos resultados, **se confirma la hipótesis.**

Se recomienda una forestación focalizada priorizando comunas con menos del 10% de la cantidad de árboles eficientes, buscando a futuro, llegar a niveles del 30% o más en todas las comunas.

3.5. Hipótesis 5 (bivariada):

3.5.1. Definición de la hipótesis:

Algunas especies de árboles alcanzan alturas promedio mayores que otras.

Datos requeridos: altura_arbol, nombre_cientifico

3.5.2. Estrategia de abordaje:

Para estudiar esta hipótesis utilizamos un test de hipótesis, para lo cual planteamos lo siguiente:

- **Hipótesis nula:** No existen diferencias significativas en la altura promedio de los árboles entre las distintas especies.
- **Hipótesis alternativa:** Al menos una especie presenta una altura promedio significativamente diferente de las demás.

Variable dependiente: altura_arbol.

Variable independiente: nombre_cientifico.

Y además hacemos una exploración con un box-plot:

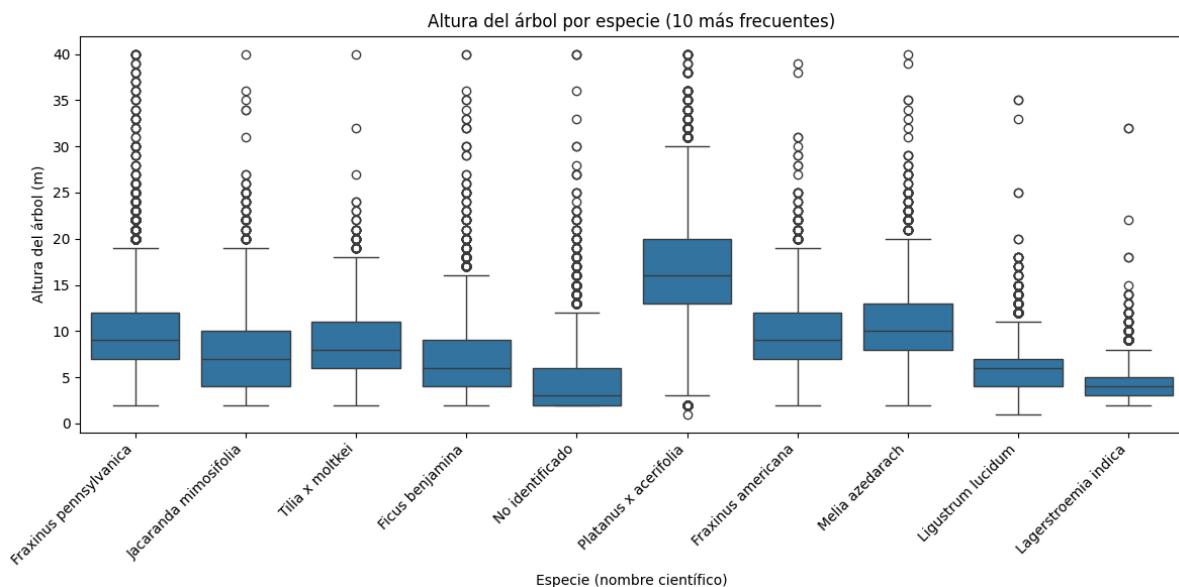


Gráfico 1: Box-plot comparando las alturas de los árboles por especie, tomando las 10 más frecuentes.

Se ve que la especie *Platanus x acerifolia* (Plátano) es, en promedio, la especie más alta, con una mediana de altura significativamente mayor (cerca de 15 metros) que el resto. En el

extremo opuesto, *Lagerstroemia indica* que los árboles "No identificados" son consistentemente los más bajos. La mayoría de las otras especies, como *Fraxinus pennsylvanica*, *Jacaranda* y *Melia azedarach*, presentan alturas medianas similares (agrupadas entre 7 y 12 metros), pero se caracterizan por una gran cantidad de valores atípicos, lo que indica la presencia común de individuos excepcionalmente altos que pueden alcanzar hasta 35 o 40 metros.

Para verificar la hipótesis del principio pensamos utilizar el test ANOVA, pero primero debemos verificar los siguientes supuestos:

- Normalidad de los residuos (usando test Shapiro Wilks)
- Homogeneidad de las varianzas (usando test de Levene)

```
Shapiro-Wilk Test:
  Estadístico = 0.9738, p = 0.000000

Levene Test:
  Estadístico = 2093.1895, p = 0.000000
```

Gráfico 2: resultados de los test de hipótesis mencionados.

Vemos que ambos tests nos devolvieron pValores muy chicos, los cuales caen fuera del intervalo de confianza, por lo que podemos rechazar ambas hipótesis nulas de ambos tests, por lo que no podemos utilizar ningún test paramétrico, entonces debemos pasar a utilizar algún test no paramétrico para verificar la hipótesis del principio, para esto, elegimos utilizar el test de Kruskal Wallis.

3.5.3. Resultados obtenidos y discusión

```
--- Kruskal-Wallis ---
H = 93406.49, p = 0.000000
```

Gráfico 3: Resultados del test de Kruskal Wallis

Como $p < 0.05$, rechazamos la hipótesis nula y podemos decir que la altura promedio difiere significativamente entre especies.

Concluimos que se presenta una diferencia significativa en la altura promedio de los árboles dependiendo de la especie de árbol. Primero lo podemos ver en el Box-Plot como el Platanus x Acerifolia tiene una concentración en árboles de mayor altura. Luego al hacer el ANOVA vimos lo mismo aunque los supuestos no se cumplieron por lo que tuvimos que descartar el test, y probar con el de Kruskal-Wallis y nos termina dando el resultado de que la altura promedio difiere significativamente entre especies. Esto es de gran ayuda para una futura deforestación y saber que si quieres árboles más chicos para zonas con veredas más angostas, podemos plantar árboles que sabemos que van a ser más bajos, mientras que en avenidas o comunas con aceras más anchas, podemos plantar árboles más altos para una mejor variación de especies. En conclusión, podemos decir que **se confirma la hipótesis 5**.

3.6. Hipótesis 6:

3.6.1. Definición de la hipótesis

En comunas con mayor diversidad de especies, los árboles presentan mayores promedios de diámetro y altura, en comparación con comunas con menor diversidad (más dominadas por pocas especies).

Datos requeridos: nombre_cientifico, diametro_altura_pecho, altura_arbol.

3.6.2. Estrategia de abordaje:

Lo primero que hacemos es crear un nuevo dataset el cual tenga solamente las columnas: comuna, nombre_cientifico, altura_arbol y diametro_altura_pecho:

```
<class 'pandas.core.frame.DataFrame'>
Index: 354838 entries, 0 to 370172
Data columns (total 4 columns):
#   Column                      Non-Null Count  Dtype
---  -
0   comuna                      354838 non-null  int64
1   nombre_cientifico          354838 non-null  object
2   diametro_altura_pecho      354562 non-null  float64
3   altura_arbol               350660 non-null  float64
dtypes: float64(2), int64(1), object(1)
memory usage: 13.5+ MB
```

Gráfico 1: Resultado de info() sobre el dataset para trabajar esta hipótesis.

Realizamos un elbow-plot para ver si existen grupos diferenciados que nos indique cierta relación entre la especie y sus medidas.

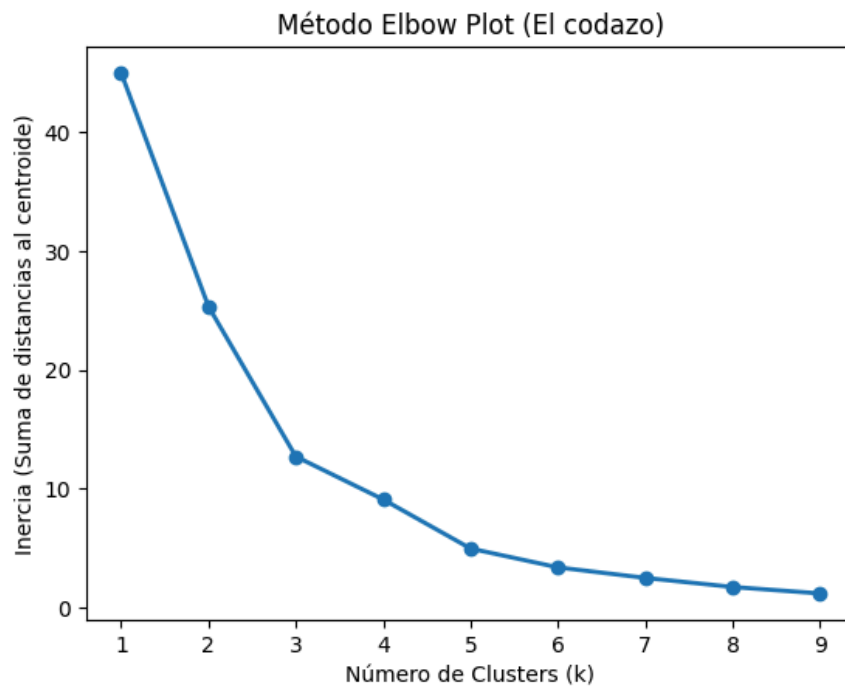


Gráfico 2: elbow-plot para encontrar un estimado de cuántos clusters debemos usar.

Podemos ver que el número indicado es $k = 3$ ya que cambia drásticamente la curva.

Luego aplicamos t-SNE para estudiar si existen clusters utilizando las variables planteadas al inicio.

3.6.3 Resultados obtenidos y discusión:

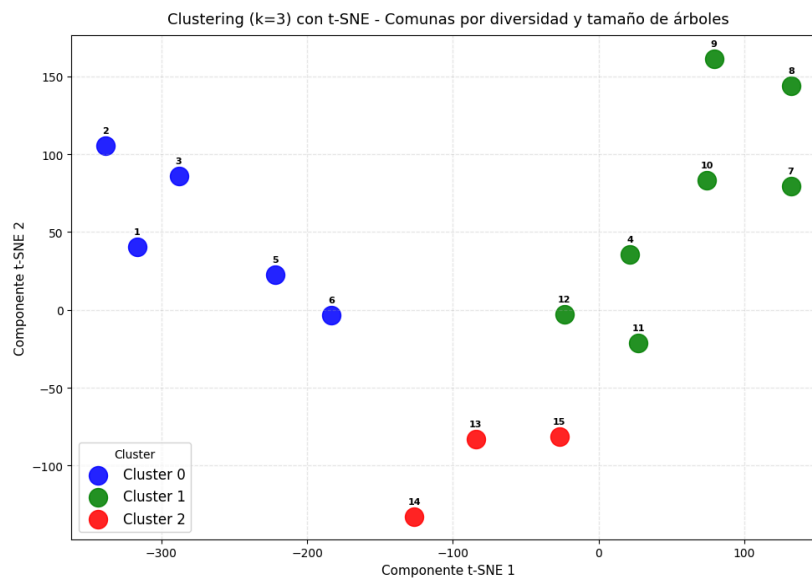


Gráfico 3: Resultado de visualización de clusters.

	diversidad	diametro_promedio	altura_promedio
cluster			
0	143.4	27.23	9.64
1	234.0	31.53	8.20
2	208.0	35.72	10.77

Gráfico 4: Tabla con los valores promedios de las variables de interés.

Aca podemos ver bien tres cluster que son:

- **Cluster 0**: Comuna 1, 2, 3, 5, 6 -- Diversidad Baja, diámetro bajo, altura Media.
- **Cluster 1**: Comuna 4, 7, 8, 9, 10, 11, 12 -- Diversidad Alta, diámetro Medio, altura Baja.
- **Cluster 2**: Comuna 13, 14, 15 -- Diversidad Media, diámetro Alto, altura Alta

Los datos nos muestran que:

- El **Cluster 1 (rojo)** es el que tiene la **mayor diversidad** de todos (234.0). Sin embargo, es el que tiene los árboles de **menor altura promedio** (8.20). Esto contradice directamente tu hipótesis.
- El **Cluster 2 (verde)** es el que tiene los árboles **más grandes** (mayor diámetro con 35.72 y mayor altura con 10.77). Sin embargo, su **diversidad** (208.0) no es la más alta, sino una **intermedia-alta**.

Los resultados **no confirman** la hipótesis. El análisis de clusters reveló que la relación no es lineal: las comunas con la máxima diversidad (Cluster 1) en realidad tienen los árboles de menor altura promedio. Por otro lado, las comunas con los árboles más altos y anchos (Cluster 2) no son las más diversas, sino un grupo con una diversidad media-alta.

En lugar de "más es mejor", el patrón sugiere que existe un equilibrio diferente. El clustering funcionó al identificar estos grupos distintos, pero los grupos no se alinean con nuestra suposición inicial.

4. Conclusión

Al finalizar este trabajo, pudimos lograr el objetivo planteado en la introducción: Utilizar el censo del Arbolado Público de Buenos Aires 2017-2018 para realizar un análisis enfocado en la **gestión urbanística**. Partimos de un conjunto de datos complejos, los cuales tuvimos que limpiar y procesar para poder hacer un uso correcto de los mismos y obtener resultados de calidad.

Las diferentes hipótesis nos dejaron en claro varios aspectos:

- La infraestructura se encuentra bajo una **presión considerable** ya que como pudimos ver en la **hipótesis 1**, casi en su totalidad, las planteras de la ciudad se encuentran ocupadas.
- Problemas de monocultivo y riesgo, los cuales fueron detectados en la **hipótesis 2** que **el arbolado está dominado** por especies como *Fraxinus pennsylvanica* (36.6%) y *Platanus x acerifolia* (8.6%), las cuales, a pesar de su abundancia, son conocidas por sus raíces invasivas y por ser fuertes alérgenos.
- **Distribución desigual** de árboles a lo largo de la ciudad estudiados en la **hipótesis 3**, donde hay comunas que presentan casi cinco veces una densidad arbórea mayor que otras.
- De la mano de la anterior, tenemos un **desbalance en la distribución de especies eficientes en la toma de dióxido de carbono** visto en la **hipótesis 4**, donde se nos muestra una baja calidad de arbolado en la mayoría de las comunas.
- **Diferencias significativas en las alturas de las especies** analizadas en la **hipótesis 5**. Este es un hallazgo interesante debido a que nos permite tomar una mejor decisión a la hora de elegir qué árbol plantar y en qué lugar, para evitar daños de infraestructura y cableado.
- Por último, para nuestra **última hipótesis**, no logramos demostrar que fuese verdadera, revelando que la relación entre la diversidad de los árboles y su tamaño no es lineal.

En conjunto, este trabajo permitió demostrar cómo el análisis riguroso de datos transforma información dispersa en insumos concretos para la toma de decisiones sobre políticas públicas de arbolado urbano. Los resultados enfatizan la necesidad de una gestión planificada,

equilibrada y sustentable del arbolado para mejorar la calidad ambiental y la infraestructura de la ciudad. Esperemos que este análisis le sirva a la persona que mandó a toda su familia a censar los árboles de la ciudad.

Referencias

Tamaño de las comunas:

https://es.wikipedia.org/wiki/Comunas_de_la_ciudad_de_Buenos_Aires

Mapa de la Ciudad donde se usó para graficar:

https://cdn.buenosaires.gob.ar/datosabiertos/datasets/secretaria-de-desarrollo-urbano/manzanas/manzanas_catastrales.geojson