Introducción a la Ciencia de Datos - 1C 2023

Guía de Trabajos Prácticos Nº 5

Un ejercicio de Ciencia de Datos

El objetivo de esta guía es utilizar las herramientas vistas hasta ahora para responder una pregunta sobre alguno de los *datasets* que ya conocemos. Esta guía es para empezar en clase y para entregar por el campus como entrega semanal.

Es importante destacar que **NO HAY UNA ÚNICA SOLUCIÓN O CAMINO** para responder cada una de las preguntas. El objetivo es que ganen experiencia en el diseño de un esquema de trabajo que les permita llegar a una respuesta y que se entrenen en los argumentos para convencer que los pasos tomados y los resultados obtenidos son adecuados.

Armen grupos de 3 personas, que van a analizar uno de los *datasets*. Lean la pregunta, piensen y planifiquen los pasos que deberían hacer para encontrar alguna respuesta. Tengan en mente la exposición de muestra que tuvimos en el aula. Pueden usar esa misma plataforma para realizar el esquema que van a tener que entregar: Miro.

Disparadores

Algunas de las cosas que podrían plantearse son:

- Ya conocen los *datasets* y los *features* de cada uno, eso ahorra mucho tiempo. Piensen una estrategia para llegar a la respuesta y cuáles de las variables de los datos pueden ser importantes para responder la pregunta y cuáles no.
- ¿Qué esperan encontrar a partir de las variables elegidas?, ¿necesitan buscar, filtrar, agrupar, comparar, buscar relaciones?
- Miren de nuevo las variables descartadas y piensen si alguna les puede servir para entender mejor las cosas del punto anterior.
- De las decisiones que tomaron y las cosas que planificaron ¿Qué datos/análisis podría ser interesante visualizar en un gráfico o tabla? Hagan dibujos a mano alzada de lo que suponen o esperan obtener.

Por Ejemplo: Tenemos que decidir dónde es mejor poner una tipuana tipu de 12 metros sobre Avenida Cabildo. Algunas opciones que podemos pensar son:

- *"Vamos a ponerla donde haya menos árboles"*

  Entonces vamos a necesitar ver la densidad de árboles a lo largo de toda la calle, para eso usaríamos la columna de calle (para poder filtrar) y la altura/chapa. Con esto ya podríamos hacer un gráfico que muestre la altura de la calle y la cantidad de árboles. Otra opción podría ser ver la cantidad por cuadra y buscar la cuadra más despoblada.

Ahora pensamos que vamos a necesitar una cantera para plantarla (nadie nos dijo que se hacían nuevas) entonces parece razonable agregar el estado de la plantera al análisis, para buscar las que no están ocupadas. Podemos agregar esta información al gráfico anterior o hacer uno nuevo. Entonces ya podemos ver lugares interesantes para plantar y elegir alguno con algún criterio justificado, por ejemplo: de lugares candidatos elegir alguno que tenga árboles más chicos, así nuestra tipuana recibe más sol.

- *"Vamos a ponerla donde haya otras Tipuanas, así se polinizan entre ellas y florecen más"*

  Entonces necesitamos filtrar por Av. Cabildo y por especie para encontrar la población de Tipuanas, una vez que tengamos esto volvemos a necesitar encontrar alguna plantera acorde, etc.

Es un proceso *iterativo* probablemente van a tener que ir y volver sobre las características del *dataset* hasta encontrar una forma que lxs convenza que sirve para responder la pregunta.

La idea es que hagan esta aproximación con un docente, validen el proceso que planearon y cuando hayan concluido **RECIÉN AHÍ** avancen con todo, no se apresuren.

Entrega

Armen un único documento (.pdf) donde muestren **la respuesta a la que llegaron**, y **cómo**. El documento debe incluir **el esquema** que armaron (pueden ser una imagen o texto, si prefieren), y una **argumentación de por qué el camino elegido es correcto**. Recuerden que **nos tienen que convencer** de que están dando la mejor solución que se puede dar **con los datos que tienen**, y que ésta se sigue de los datos y análisis. Si necesitan **otras fuentes de datos**, deben aparecer claramente citadas y explicadas en el documento Finalmente, pueden incluir (opcionalmente) los gráficos que consideren necesarios para mostrar lo que quieran. **La extensión máxima del trabajo es de tres carillas A4**.

Acá también, posiblemente, tengan que ensayar varias explicaciones y líneas de argumentación antes de dar con la correcta. No necesariamente la mejor explicación sea el orden en el que realizaron las tareas.

No se frustren, este proceso puede ser duro y lleno de idas y venidas. Charlen con otros equipos, intercambien ideas, hablen por Discord.

- Organicen los resultados obtenidos.
- Decidan qué gráficos o tablas son relevantes para incluir en el proceso de análisis, es importante sacar conclusiones de cada uno (y que estas se desprendan de los plots, claro).
- Expliciten si realizaron alguna suposición o citen la fuente si sacaron un dato de otro lado.
- Es interesante incluir si durante el análisis se encontraron con algo muy diferente de lo que esperaban o si apareció algo digno de mencionar.

- Recuerden **incluir el nombre de los integrantes** dentro del **.pdf** entregado. Alcanza con que lo entregue una persona por el campus. Por favor, que el nombre del archivo sea "Apellido1_Apellido2_Apellido3-Dataset_elegido.pdf".

Pregunta para el *Dataset* Árboles

El Jacarandá es un árbol nativo de la base de las yungas o selvas de montaña, en el noroeste de Argentina. Es considerado uno de los árboles indígenas más bellos del país. Los jacarandás porteños fueron incorporados al paisaje urbano por Carlos Thays. Según el último Censo del Arbolado Público Lineal de la Ciudad, en la actualidad hay poco más de 11.000 especímenes.

En la mayoría de las fotografías que se utilizan para mostrar escenas típicas de la Ciudad Autónoma de Buenos Aires se observan varios de estos ejemplares. Sin embargo, una reciente encuesta reveló que hay vecinos y vecinas de la ciudad que no suelen ver esta especie en sus vidas diarias. A partir de este dato, el Ministerio de Ambiente y Espacio Público logró una partida presupuestaria para iniciar una campaña de plantación y posterior mantención de Jacarandás.

En la primera etapa de la campaña se espera poder plantar un total de 20 ejemplares con la proyección de continuar con esta estrategia varios años más. **El criterio que quiere tomar el Estado para la ubicación de estos árboles es que no importa la cantidad de peatones que transitan la calle donde se ubiquen los ejemplares, sino que los vecinos que pasen por allí noten lo más posible la presencia del jacarandá.**

El gobierno se contactó con ustedes para realizar una consultoría y los ayuden a definir a partir de los datos que tienen (`arbolado-publico-lineal-2017-2018.csv`) dónde ubicar estos 20 árboles. En primer lugar, les piden que piensen con la información disponible cómo definir qué significa que "un vecino note lo más posible la presencia del Jacarandá". Luego, para determinar la ubicación de los árboles deberán tener en cuenta las siguientes observaciones que realizó el Ministerio:

- Los ejemplares necesitan **que haya por lo menos a 3,5 metros de distancia entre ellos y las casas** para que las ramas que caen tengan menos probabilidades de causar daños.
- No se plantarán nuevos Jacarandás en cuadras que tengan 15 árboles o más (de cualquier especie, en ambas veredas: par e impar).
- Para hacer los estudios correspondientes a este análisis, puede considerarse que los árboles a plantar van a alcanzar una altura y diámetro similar a la que tienen los Jacarandás en esa misma comuna.

Quienes lxs contratan solicitan que les digan las primeras cuadras en las cuales deberían plantar los Jacarandá.

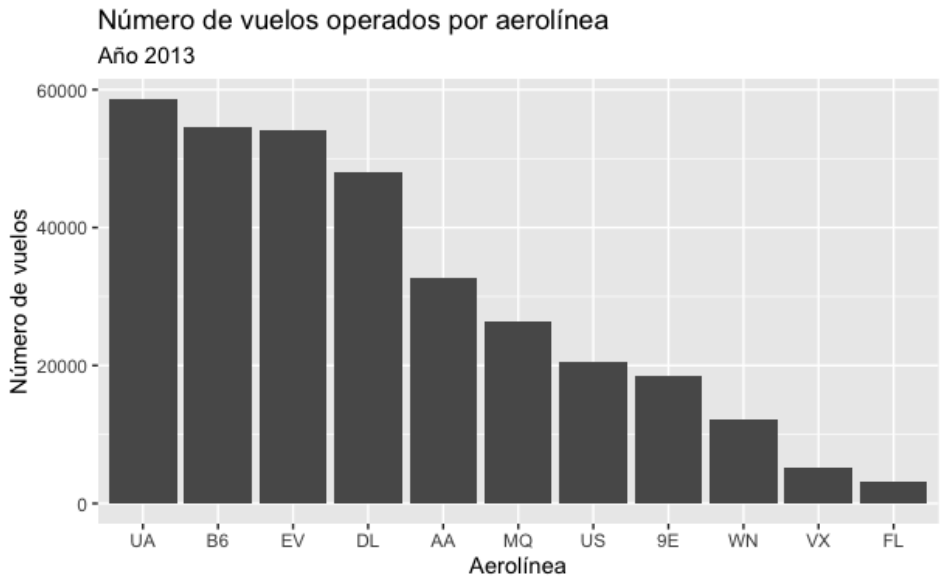
Fuente de los datos: https://data.buenosaires.gob.ar/dataset/arbolado-publico-lineal

Pregunta para el *Dataset* vuelos

Corre el año 2014, conseguimos nuestro primer trabajo como cientificxs de datos en una aerolínea de EEUU. En particular, trabajamos en el departamento que se ocupa de los vuelos que operan desde los tres aeropuertos de Nueva York (JFK: John F. Kennedy; LGA: La Guardia; EWR: Newark). A pocas semanas de arrancar con el trabajo llega una noticia: la empresa consiguió financiamiento para agregar tres aviones a la flota, y quieren que sean utilizados para mejorar la oferta de vuelos desde Nueva York para el 2014, así que te llaman para ayudar a tomar una decisión acerca de qué rutas hay que reforzar y te presentan el dataset con los vuelos que hizo la aerolínea en 2013. Por cuestiones administrativas, no se puede ofrecer vuelos a destinos con los que no estén operando.

En una reunión para decidir cómo encarar el problema, se presentaron tres visiones: 1. la de mejorar la seguridad de los vuelos; 2. la de mejorar la experiencia de los viajeros; 3. la comercial (infaltable) de mejorar las ganancias de la compañía.

Elijan alguno de estos caminos para plantear la pregunta de manera concreta; **elijan en qué aerolínea están trabajando**, entre las que volaron más de mil veces en 2013 (ver gráfico abajo); consideren el tamaño de su aerolínea y la relación con la competencia para dar una respuesta lo más precisa posible..



TIP

Recién estamos viendo la punta del iceberg del dataset de vuelos; si quieren pueden también consultar la tabla "planes" que tiene información sobre cada avión. Ya vamos a trabajar en detalle con datos relacionales, pero por ahora, les dejamos la línea para unir ambas tablas según el número del avión:

```
> inner_join(flights, planes, by='tailnum')
```