Licenciatura en Ciencia de Datos
ECYT_UNSAM

Introducción a la Ciencia de Datos - 2C 2022

Guía de Trabajos Prácticos Nº 4

Un poco más 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. Las vamos a mirar y a hacerles una devolución (no se preocupen, es sin nota).

Es importante destacar que **NO HAY UNA ÚNICA SOLUCIÓN O CAMINO** para responder cada una de las preguntas, es todo un proceso en el que nos tienen que convencer con los pasos tomados y los resultados obtenidos.

Parte 1. En el aula

Armen grupos de 3 personas y entre todos van a analizar uno de los *datasets*, lean la pregunta, piensen y planifiquen los pasos que deberían hacer para encontrar alguna respuesta.

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 *features* de los datos pueden ser importantes para responder la pregunta y cuáles no.
- Con los *features* elegidos: ¿Qué esperan encontrar?, ¿necesitan buscar, filtrar, agrupar?
- Miren de nuevo los *features* descartados y piensen si alguno le 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 estén sobreocupadas. 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** vayan al laboratorio, no se apresuren.

Parte 2. En la computadora

- Ahora que tienen una idea de lo que están buscando usen los *features* que seleccionaron. Piensen el orden de selección, filtrado y las cosas que necesiten para obtener los gráficos que planearon. ¿Los resultados fueron muy diferentes de lo que esperaban encontrar?
- Si apareció algo muy diferente a lo esperado, ¿Lo pueden explicar? Quizás necesiten agregar algún otro *feature* para explicarlo mejor.
- OPCIONAL: ¿Les serviría buscar datos en otro lado? por ejemplo, tasa de crecimiento de alguna especie de árbol, esperanza de vida, etc. Pueden incluirlos y utilizarlos (citando la fuente)

Parte 3. Entrega

Armen un documento donde muestren todo el proceso que hicieron para llegar a una respuesta. Recuerden que nos tienen que convencer que dan la mejor solución y que esta se sigue de los datos y análisis. Acá también, posiblemente, tengan que ensayar varias explicaciones y líneas de argumentación antes de dar con la correcta. 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 en el **.pdf** entregado.

Por Ejemplo: Decidimos que vamos a poner la tipuana donde haya menos árboles.

- Justificamos nuestro criterio, eventualmente mostrando datos de otras fuentes. Por ejemplo, una encuesta vecinal que arroje como resultado que a la gente le gustan los árboles (invento).
- Hacemos un gráfico que muestre la altura de la calle y cantidad de árboles.
- Con el gráfico anterior elegimos una zona en particular, **con algún criterio que tenemos que mencionar**.
- Hacemos otro gráfico solo de esa zona donde mostramos la ocupación de las planteras, y elegimos la que esté menos ocupada. Si además tenemos en cuenta la altura de los árboles podemos agregar esta información al gráfico o hacer otro.
- Con los gráficos y un poco de texto explicando cómo fue el proceso pueden justificar su elección, no alcanza con decir “Este es el mejor lugar” eso se tiene que desprender de todo el análisis.

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 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 esa calle, 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, ya sea en vereda par o 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.

Como un extra: una vez terminado el análisis anterior, piensen si podrían estimar cuántos ejemplares plantarían en total para que todos los vecinos de la ciudad notaran más esta especie tan importante, pero sin desperdiciar recursos. En caso de no llegar a hacer el análisis, pueden explicar qué es lo que harían en palabras.

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

Pregunta para el *Dataset* Seguros

Recordemos los *features* de cada dato:

age	sex	bmi	children	smoker	region	charges
19	female	27,9	0	yes	southwest	16884,924
18	male	33,77	1	no	southeast	1725,5523
...	...	...	...	...	...	...

Para cada persona tenemos la edad, el sexo, el índice de masa corporal, la cantidad de hijos, si es fumador, en que región vive y lo que paga de seguro médico.

La empresa proveedora de los seguros médicos está planeando una campaña publicitaria en la vía pública y recurrió a nuestros servicios para buscar maximizar sus ya enormes ganancias. Tienen planificados 500 carteles para colocar y por cuestiones logísticas prefieren usarlos todos en una sola zona, pero podrían aceptar dividirlos en dos (si la justificación es buena).
Con los datos ya provistos en (`Insurance.csv`), ustedes tienen que decidir en qué zona/s le convendría a la empresa enfocar su campaña y si es mejor apuntarla a un sexo en particular o no.

Como un extra: Experiencias anteriores mostraron que con 500 carteles en una zona alcanzaron al 70% de la población, y de las personas que lo vieron el 25% contrató el seguro.
Suponiendo que cada zona tiene 10000 habitantes (si, son así de ordenados), prueben hacer una estimación de las ganancias que se esperan. Además pueden suponer estas cosas:

- Hay mitad de hombres y mujeres en cada zona.
- Si la campaña fue enfocada en un sexo en particular, el 75% de los nuevos clientes es de este sexo.

Si necesitan hacer alguna otra suposición consulten con el docente y explicítenla en el informe. En caso de no llegar a hacer el análisis, pueden explicar qué es lo que harían en palabras.

TIP
Si van a trabajar con Rstudio usen la siguiente línea para importar el .csv para que tome correctamente la coma como separador decimal:

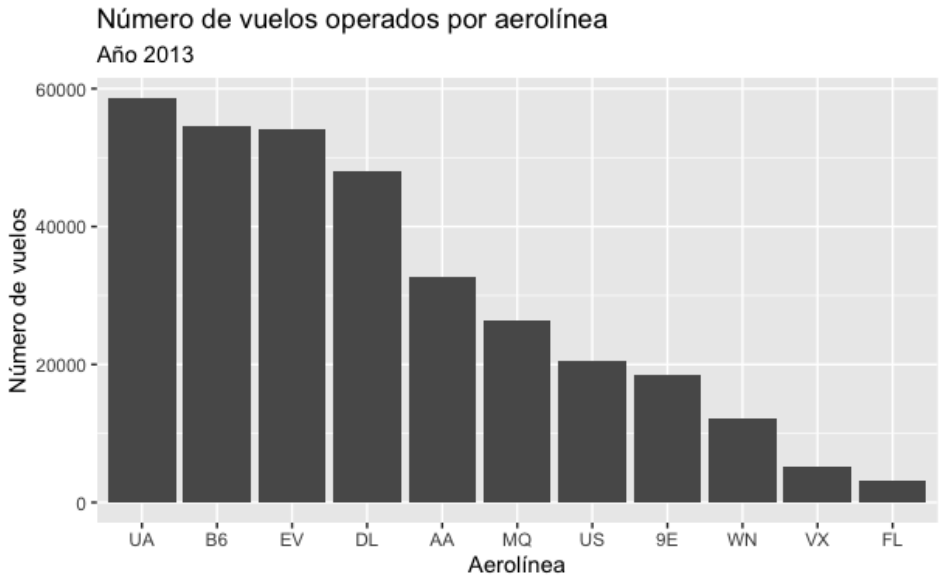
df_seguros <- read.csv(file = '... ...', dec=",")

Pregunta para el *Dataset* vuelos

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 la manera concreta; además, elijan en qué aerolínea están trabajando, entre las que volaron más de mil veces en 2013 (ver gráfico abajo); piensen cómo su respuesta se acomoda si trabajan para una empresa chica, mediana o grande.



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')
```