Introducción a la Ciencia de Datos

Checklist para antes de terminar nuestra producción

Gráficos

- ¿Los ejes tienen nombre? ¿Los nombres de los ejes explican de forma clara y específica qué es cada variable?
- Cuando corresponda, colocar entre paréntesis o entre corchetes las unidades de las variables.
- ¿El gráfico tiene título? ¿El título representa lo que muestra el gráfico? ¿Es lo suficientemente específico? ¿Es fácil de leer o es demasiado largo?
- Si el gráfico usa colores para diferenciar categorías, ¿está la leyenda que explica cada color? ¿Están presentes todas las categorías? ¿Cada clase tiene un nombre lo suficientemente corto como para entrar sin problemas en el espacio destinado a la leyenda? ¿El nombre de cada categoría describe precisamente esa categoría? El objetivo es que sólo con ver la leyenda y sin conocer el dataset, se debe entender la diferencia entre los datos de diferentes colores.
- Revisar los límites del gráfico. A veces para mostrar ciertos comportamientos es necesario recortar parte de la información (¿outliers tal vez?). Esta decisión dependerá 100% de lo que queramos mostrar con este gráfico. También se podría elegir mostrar dos veces el mismo gráfico con diferentes límites haciendo un "zoom" en la zona de interés.
- Si dibujaron alguna línea/curva sobre el gráfico explicar qué es (ajuste, media, mediana, media+std, etc.)
- Revisar el tamaño de la tipografía de los textos que aparecen en el gráfico. Todos tienen que ser legibles pero seguir una cierta jerarquía en cuanto a la importancia.

Gráfico de dispersión o *scatter plot*

- La elección de ejes para cada variable, ¿fue realizada a conciencia? ¿Ilustra la información que querían mostrar y/o explicar? Si hay dudas, corroborar con otra persona qué se entiende a partir del gráfico y/o invertir ejes y comprobar cuál de las opciones representa la información deseada.
- ¿El tamaño elegido para los puntos permite la visualización de todos los datos y tendencia de los mismos? Verificar que no haya muchos puntos que se superpongan y que los puntos se vean enteros. Recuerden que en los gráficos existe la variable de opacidad (*alpha*) que permite transparentar los datos.
- Si el tamaño de los puntos o el símbolo representa una variable, detallarlo en la leyenda.
- Si le colocan una etiqueta a los datos, expliquen qué variable representa.
- ¿El/los colores elegidos para los puntos contrastan bien con el fondo y entre ellos? Pueden investigar con distintos mapas de colores. Es importante distinguir cada punto del fondo y de sus vecinos.
- Chequeen que toda la información que están agregando al gráfico (ya sea con colores, símbolos o tamaños) sirva para mostrar algo. Recuerden: **si no suma, resta**.
- ¿Sería útil agregar una línea que una los putos del *scatterplot*? Para decidir esto pueden probar hacerlo y comprobar si este agregado permite apreciar mejor la tendencia de los

datos.

Gráfico de barras

- Elegir la mejor forma de mostrar los datos según lo que quieran explicar: ¿columnas apiladas o separadas? ¿totales o porcentajes? Debe quedar clara la elección que tomaron.
- ¿Las columnas del gráfico tienen nombre?
- Si hubiera columnas que no se llegan a ver, agregar una etiqueta con el valor.
- ¿El ancho de las columnas y la separación entre las columnas permite ver y distinguir bien su comportamiento?

Gráficos de distribuciones: histogramas, *violin plots*, *box plots*, *density plots*

- ¿Elegí la mejor opción para representar esta distribución? *Para esto pensar: ¿qué tipo de datos tengo? ¿qué quiero visualizar de ellos? ¿Es este el gráfico que mejor nos lo muestra?*
- Para histograma y density plot: ¿está bien elegido el tamaño de los bines? Para decidir esto revisar: ¿el tamaño permite tener una cantidad apreciable de datos en cada bin? ¿el número de bines es suficiente como para representar el comportamiento del dataset?

Descripción del gráfico

- Una buena práctica es comenzar explicando qué variables se están graficando, y explicar de dónde salieron los datos. Es importante ser lo más específicxs posibles: por ejemplo, si realizamos algún filtro sobre los datos originales explicar esa particularidad.
- Escriban conclusiones que realmente se puedan extraer de los gráficos: no presenten intuiciones o premisas que se argumentan con otra información como si surgieran del gráfico. Si esperaban encontrar algo pero no se visualiza aclararlo. ¿Qué hipótesis tenían? ¿Por qué creen que no se aprecia en este análisis?
- Relean lo que escribieron y decidan si se entiende unívocamente lo que quisieron decir. Recuerden que sólo tenemos el escrito como información. No sabemos qué pensaron más allá de eso. Si lo lee alguien que no sabe del tema, ¿entenderá lo que buscan explicar?
- Mantener la persona y número con la que se redacta durante todo el escrito.
- ¡Pasenle un corrector ortográfico y gramático al texto cuando lo tengan escrito! (Documento de Google)

Limpieza y adecuación de datos

- Antes que nada familiarizarse con el dataset y el dominio del problema
- Verificar errores de importación
- Verificar tipos de datos de cada columna, que sean acorde a lo que representan.
- Verificar valores de los datos: outliers, faltantes, etc.
- Para los valores problemáticos decidir qué estrategia utilizar para reemplazarlos o borrar
- Verificar datos duplicados (ojo, entender si son duplicados o no)
- **Todas las decisiones se tienen que documentar y justificar**

Archivo

- Elegir como formato de archivo a enviar uno que no pueda ser modificado como word. Por el contrario, pueden elegirse: .PDF, .PNG, .JPEG, etc.
- Elegir como nombre del archivo uno que represente la entrega y que contenga su nombre y/o apellido.
- Enviar un archivo con formato Apellido_nombre_entrega.pdf/png/jpeg etc. NO EDITABLES (como word) ni carpetas con varios archivos.