Laboratorio 1

Exploración de datos

En este laboratorio realizaremos un trabajo de exploración sobre un dataset dado. Hemos tomado como inspiración el siguiente kernel de Kaggle: We are from our childhood. El conjunto de datos utilizado son las respuestas de gente joven a una encuesta, con la que les proponemos trabajar. En esta notebook en particular, la autora visualiza distintos aspectos de los datos tratando de encontrar factores de variación relacionados a la respuesta “Crecí en la ciudad” o “Crecí en el campo”. La consigna para este laboratorio es realizar un trabajo similar, aunque más simple, analizando algunas de las variables provistas por la encuesta.

Pueden utilizar otro dataset, previamente aprobado por las docentes. Si eligen esta opción, en el informe describir el dataset o los aspectos relevantes del mismo que se utilizarán durante el laboratorio. Es importante especificar por qué seleccionaron esos datos y qué tipo de información esperan extraer.

Entrega

En el repositorio de su grupo deben subir una notebook (recomendamos usar jupyter) con el código y las visualizaciones utilizadas para extraer la información que les pedimos a continuación. Si la notebook no la presentan en un github, suban también un archivo html para que podamos ver las imágenes sin necesidad de ejecutar el código. Recuerden que entre las celdas de código puede agregar markdown y latex para explicar las hipótesis que plantean y conclusiones que obtienen.

El informe sólo es necesario para aclarar decisiones de diseño que hayan tomado, librerías extras utilizadas, etc.

Implementación

Se deberá presentar la siguiente información

- Estadísticas descriptivas
    - Calcular estadísticos como la moda, media, mediana y desviación estándar del peso y de la edad. ¿Responden a alguna distribución conocida?
    - Realizar un análisis de outliers.
    - Explicar cómo varían las métricas cuando desglosamos por género. ¿Responden a alguna distribución conocida? Comparar cualitativamente y

gráficamente ambas distribuciones. Calcular la correlación entre todas estas variables y mostrarla con un gráfico conjunto.

- Calcular la probabilidad marginal y conjunta, y la correlación entre otras dos variables, por ejemplo consumo de alcohol y tabaquismo.
    - Representar visualmente la probabilidad conjunta entre los valores posibles de las variables elegidas.
- Responda a la siguientes preguntas: ¿Qué pasaría con los niveles de tabaquismo si se prohíbe fumar en los bares? ¿Qué pasaría con la cantidad de consumidores de alcohol si disminuye la cantidad de fumadores que consumen alcohol?

Se evaluarán los siguientes aspectos:

- Estructura legible de la notebook.
- Los tipos de gráficos son adecuados para la información representada

Otros datasets para inspirarse

- En el portal de datos abiertos de Argentina pueden encontrar datasets, la mayoría de ellos sobre agricultura. En particular encontramos datos sobre femicidios en Argentina, similar al que trabajaremos en clase de violencia institucional.
- En GapMinder existe una colección de datasets que describen World wide wellfare variables.
- En el github Fivethirtyeight datasets hay una colección de datos utilizados para crear reportes periodísticos con alto contenido de estadística. Entre los destacados:
    - Comic characters.
    - Fandango.
    - Nutrition
    - Star wars
- En Kaggle también existen muchas opciones de datasets y código para trabajarlos.