CI-0115 Probabilidad y Estadística
Prof. Sebastián Ruiz Blais
I-2024
Grupos 2 y 3

Proyecto – Entrega 1 (5% de la nota)
*Fecha de entrega***:** *viernes 3 de mayo del 2024*

Objetivos:
1. **Escoger tema** para análisis de datos. Para esto, pueden comenzar con una idea general de lo que quisieran hacer. Como es posible que no encuentren una base de datos para exactamente el tema que quieran, se sugiere tener 3 temas candidatos y **explorar la disponibilidad** de la información en la web, siguiendo el criterio de riqueza de los datos (punto 2). Puede buscar bases de datos en los sitios http://w.ww.kaggle.com/, https://datasetsearch.research.google.com/, https://www.earthdata.nasa.gov/, https://datahub.io/collections, entre otras (ver https://careerfoundry.com/en/blog/data-analytics/where-to-find-free-datasets/).
2. Escoger base de datos cumpliendo criterio de **riqueza de los datos**. Esto es, deben haber suficientes **datos cuantitativos**, por lo menos 5-10 columnas y 100+ filas. Datos cuantitativos son datos numéricos, idealmente con un rango de valores suficientemente amplio (p.ej., algunas variables con valores enteros entre 1 y 5 pueden ser difíciles de analizar). **Consulten con el profesor** para corroborar que el tema y la base de datos es aceptable. Es posible usar más de una base de datos, siempre y cuando se puedan unir por medio de una columna común.
3. **Justificar por qué se eligió el tema y la base de datos** anterior, incluyendo qué es lo que quisieran investigar de este tema. Por ejemplo, se eligió una base de datos sobre salud mental y variables como nivel educativo, capacidad adquisitiva y alimentación para estudiar los factores que influyen en la salud mental. Esta explicación debe describir la base de datos, incluyendo el tipo de variables y datos que se tiene.
4. **Realizar visualizaciones iniciales** para explorar la base de datos. Es decir, mostrar en el cuaderno de Jupyter algunas porciones de la o las bases de datos escogidas. Aquí también deben mostrar un ejemplo de los datos (una fila) y decir a qué corresponde cada punto. Por ejemplo, la información para un país dado en un año dado.
5. Seleccionar **variables de interés** (por lo menos 5), modificar los nombres de las variables para utilizar nombres más breves/representativos. Esto implica ignorar el resto de la base de datos.
6. Realizar un **pre-procesamiento para filtrar los datos** no deseados/necesarios y excluir valores fuera del rango estudiado (p.ej., en el lab 2 se eliminaron valores previos a 2015 y filas con NaN en variables de interés). Justificar pasos realizados.
7. **Utilizar histogramas** para visualizar las distribuciones de las variables de interés. Describir estas distribuciones: ¿cómo se distribuyen los datos?
8. Obtener la **media y desviación estándar** de las variables de interés. Presentar esta información de manera clara.

9. **Obtener correlaciones entre las variables de interés** para explorar cómo se relacionan. Explique los resultados de las correlaciones.
10. Graficar variables de interés una en función de la otra (p.ej., X en función de Y)
10. **Identificar datos atípicos** (outliers) **y descartarlos**. Explique qué criterio utiliza en caso de descartarlos o no.
11. **Documentar** con claridad todas las preguntas que les gustaría poder contestar del tema (sin importar si van a poder contestarlas o no) y cuáles de ellas logran contestar en esta etapa de análisis de los datos.

Evaluación
- Calidad del análisis: logran implementar cada uno de los objetivos del proyecto con una base de datos lo suficientemente rica. Realizan gráficas correctamente etiquetadas. Realizan cada paso correctamente.
- Justifican cada etapa
- Calidad de las explicaciones: explican los pasos realizados y las preguntas realizadas. Contestan algunas de las preguntas basándose en el análisis de los datos. Presentan los pasos de manera ordenada y usando gráficas y visualizaciones para mejorar las explicaciones.
- Presentación de los resultados: explican al grupo de manera clara lo que realizaron y la motivación de su trabajo.
- Auto-evaluación: posterior a la presentación, cada integrante completa a conciencia la ficha de auto-evaluación y evaluación de los integrantes del grupo.

Indicaciones
- Desarrolle su trabajo usando los Jupyter notebooks y entregue en el espacio asignado en Mediación Virtual.
- A diferencia de los laboratorios, para el proyecto los grupos deben tomar decisiones de análisis dependiendo de las preguntas realizadas. Cada decisión debe estar debidamente justificada.
- Investigue los objetivos que no sabe cómo realizar.
- Consulte con el profesor a la hora de tomar decisiones importantes o para aclarar dudas.