Laboratorio de Datos

Verano 2026

Trabajo Práctico N° 1

Grupos: 3 o 2 integrantes. No se aceptarán entregas individuales.

Entrega: notebook a través del Campus (pestaña Trabajos Prácticos) con los nombres de les integrantes del equipo. Incluir también otros archivos relevantes para el flujo de trabajo.

Fecha límite: 23 de febrero 23.59 h

Bonus: Se valorará que el Notebook y el código tengan un formato prolijo: ejercicios separados por títulos (Ejercicio 1, Ejercicio 2, etc.), nombres descriptivos para las variables, comentarios, etc. Sería ideal que hubiera texto acompañando los distintos bloques de código y/o acciones de forma que el lector pueda seguir el razonamiento del grupo. Además, recomendamos que antes de entregar el TP, corroboren que todas las celdas corran sin errores:

```
Kernel -> Restart Kernel and Run All Cells
```

En este trabajo vamos a analizar el servicio de EcoBicis en Ciudad Universitaria, para conocer mejor el comportamiento de los usuarios del servicio y ver posibles mejoras.
Trabajaremos con el dataset `full_data.csv` del repo https://github.com/FedeGiovannetti/Eco_bici_Ciudad_Universitaria que contiene información sobre la cantidad de bicis y espacios disponibles en los dos puntos ubicados dentro de Ciudad Universitaria. La información se almacena cada una hora y está activo desde mayo de 2025.

Pueden usar los datos directamente desde el repositorio:

```
datos =
   pd.read_csv('https://raw.githubusercontent.com/FedeGiovannetti/
Eco_bici_Ciudad_Universitaria/refs/heads/main/data/full_data.csv')
```

O a partir del archivo disponible en el campus:

```
datos = pd.read_csv('full_data.csv')
```

Procesamiento de datos [2 pts.]

En esta primera sección, vamos a implementar algunas acciones básicas de selección y procesamiento de nuestros datos. La idea será quedarnos con un objeto llamado `datos_limpios`

1. Nos quedamos solo con las columnas:
   - station_id,
   - num_bikes_available,
   - num_bikes_disabled,
   - num_docks_available,
   - num_docks_disabled,
   - Date,
   - hora,
   - dia.
2. Pasar todos los nombres de columna a español utilizando un diccionario. Los nombres deberían ser:
   - estacion,
   - bicis_disponibles
   - bicis_rotas,
   - puertos_disponibles,
   - puertos_rotos,
   - fecha,
   - horario,
   - dia_semana.
3. Pasar los días de la variable `dia` a español (pueden usar `map` con un diccionario o alguna función específica para fechas).
4. ¿Cuáles son los distintos tipos de variables que tenemos ahora?
5. ¿La base de datos contiene datos faltantes? ¿Cuántos?
6. Los datos de fechas y hora son bastante complejos de manejar. Una forma bastante simple de hacerlo sería generar las variables "anio", "mes", "dia" y "hora" extrayendo la información consecuente de los strings de "fecha" y "horario". Verificá que las variables resultantes sean numéricas (o convertilas a numéricas si fuera necesario).
7. Otra forma de manejar las fechas y horas es pasarlas al formato `datetime`. Para esto, usaremos el siguiente bloque de código:

```
 datos_limpios["fecha"] =
pd.to_datetime(datos_limpios["fecha"])

 datos_limpios["hora"] =
pd.to_datetime(datos_limpios["hora"]).dt.time
```

8. Implementá una función llamada `determinar_estacion_anio` que permita construir una variable llamada `estacion_anio` al pasarle nuestros datos. La variable puede estar construida a partir de las variables "dia" y "mes" o de la variable "fecha" (opción para valientes).

Análisis descriptivos [2 pts.]

9. Quisiéramos analizar la cantidad de bicis disponibles por hora, la cantidad de bicis rotas, y la cantidad de puertos disponibles en todo Ciudad Universitaria (es decir, sumando la información de las dos estaciones). Para eso, generen un nuevo dataframe agrupado por fecha y hora llamado `datos agrupados` que contenga las variables "cantidad_bicis_disponibles", "cantidad_bicis_rotas", "cantidad_puertos_disponibles", y conserve las variables "fecha", "hora", "dia_semana", "dia", "mes" y "anio".

10. ¿En qué estación del año suele haber más bicis disponibles y puertos disponibles? Justificá tu respuesta visualmente. ¿Cómo explicarías esa tendencia?

11. ¿Cómo cambia la cantidad promedio de bicis y puertos disponibles a lo largo del día para cada día de la semana? Justificá tu respuesta visualmente (por ejemplo, un gráfico donde pueda verse una línea para cada día de la semana representando la cantidad promedio de bicis disponibles a cada hora). ¿Cómo explicarías esa tendencia?

12. En los gráficos que realizaste hay algunos valores que parezcan incorrectos o mal medidos? ¿Cómo lo podrías justificar?

Análisis exploratorio [4 pts.]

13. La idea de este ítem es que realicen un análisis exploratorio de los datos, aplicando las herramientas de visualización (seaborn.objects, seaborn y/o matplotlib), de resumen de datos (media, mediana, desvío estándar, operaciones sobre el DataFrame, etc.) y/o de Regresión.

El objetivo es entender, comparar y/o estudiar el uso de las bicis en Ciudad Universitaria. Algunas preguntas disparadoras pueden ser:

- ¿Cuáles son las semanas con mayor uso de bicis? Corresponden con los inicios de los cuatrimestres de la FCEyN / FADU / CBC?
- ¿Cómo afecta el clima el uso de las bicis? (días de lluvia, días de mucho frío, días de mucho calor) ¿Hay un comportamiento distinto en días de semana y fines de semana?
- ¿Cómo es el mantenimiento de las estaciones? ¿Reparan o retiran las bicis rotas a alguna hora en particular?
- ¿Hay diferencias en el uso de cada una de las dos estaciones? ¿Suele llenarse primero una que otra?

No es necesario que respondan a cada una de esas preguntas (ni se limiten a eso), lo mejor es que exploren por donde se les ocurra. Alentamos que se planteen hipótesis y usen los datos para corroborarlas o rechazarlas. Pueden aplicar cualquiera de las herramientas que hemos visto hasta ahora. Asimismo, pueden utilizar información de otras fuentes si consideran que puede ser útil.

Importante: en el Notebook, las visualizaciones y resúmenes de datos que realicen deben estar acompañados por las conclusiones que obtengan a partir de ellos.

14. Como posible información adicional, entre los materiales del TP, tienen un archivo que se llama `clima.csv` que contiene información sobre las condiciones meteorológicas desde marzo a diciembre del 2025. Pueden explorar que significa cada variable acá:

https://meteostat.net/en/station/87582?t=2025-03-01/2026-02-16

Regresión Lineal [2 pts.]

En esta sección, queremos analizar el uso de bicis en otras estaciones de la Ciudad. Para eso vamos a utilizar el dataset `viajes-por-dia.csv` con la cantidad de viajes por hora realizados desde cada estación y cantidad de viajes por hora realizados hacia cada estación.

15. Se quiere ajustar la cantidad de viajes con origen en la estación 005 - Plaza Italia en función de viajes originados en otras estaciones o con destino en distintas estaciones. Es decir, queremos hacer un modelo para ajustar la variable `origen_5` en función de otras variables del DataFrame.

16. Propongan tres modelos de regresión distintos. En cada modelo, pueden utilizarse a lo sumo 5 variables explicativas (pueden ser variables distintas en los distintos modelos). Los criterios de selección de esas 5 columnas los determinan ustedes arbitrariamente, y deben estar explicitados en el informe.

17. Propongan un esquema de validación de los modelos y utilizarlo para seleccionar el mejor de los tres modelos propuestos.

Importante: Pueden utilizar cualquier criterios para la elección de las variables de cada modelo, no se pide buscar las 5 mejores variables entre todas las disponibles. Lo que sí tienen que hacer es explicar entre los 3 modelos que proponen cómo seleccionan el mejor de ellos.

18. Para el modelo elegido, indiquen la fórmula final de modelo.