Laboratorio de Datos

Primer Cuatrimestre 2024

Trabajo Práctico N° 1

El Trabajo Práctico deberá ser resuelto en grupos de dos o tres personas. No se aceptarán entregas individuales. La entrega se realizará a través del campus (pestaña Trabajos Prácticos). La fecha límite es el 28/05 a las 23:59. Deben entregar un Notebook con los nombres de les integrantes del equipo, la resolución de los ejercicios y los informes pertinentes.
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.

Trabajaremos con el dataset `sube-2023.csv`[1] que contiene datos sobre la utilización de la SUBE durante el año 2023 a nivel nacional. En este link pueden consultar el tipo y la descripción de cada variable.

Procesamiento de datos [2 pts.]

1. **(a)** Visualizar el tipo de datos de cada columna. Transformar la columna `DIA_TRANSPORTE` para que sea reconocida como una fecha.

Sugerencia: investigar la función `to_datetime` de `pandas`. Para completar el argumento `format`, revisar la documentación de `datetime`

(b) Agregar tres columnas al DataFrame:

i. `FECHA_DIA` : debe indicar el nombre del día de la semana correspondiente a `DIA_TRANSPORTE`

ii. `FECHA_ORDINAL` : debe indicar el ordinal correspondiente a `DIA_TRANSPORTE` (por ejemplo, a 2023-01-01 le corresponde 1, a 2023-01-02 le corresponde 2 y así sucesivamente). Debe ser un entero (`int`).

iii. `FECHA_MES` : debe indicar el mes correspondiente a `DIA_TRANSPORTE`

Sugerencia: investigar el método `apply` de `DataFrame`.

2. Crear el DataFrame `datos_amba`, el cual sólo debe tener datos de AMBA y debe excluir datos preliminares. Además, al ejecutar `datos_amba.head()` debe observarse el siguiente orden y formato de columnas:

	fecha	fecha_dia	fecha_mes	fecha_ordinal	jurisdiccion	linea	pasajeros	tipo_transporte
0	2023-01-01	Sun	01	1	MUNICIPAL	1	61	COLECTIVO
1	2023-01-01	Sun	01	1	MUNICIPAL	2B	11	COLECTIVO
3	2023-01-01	Sun	01	1	PROVINCIAL	BS_AS_LINEA_326	438	COLECTIVO
5	2023-01-01	Sun	01	1	MUNICIPAL	BS_AS_LINEA_514	3067	COLECTIVO
6	2023-01-01	Sun	01	1	MUNICIPAL	BS_AS_LINEA_522	332	COLECTIVO

3. Utilizando `datos_amba`, identificar:

[1] Fuente: https://datos.transporte.gob.ar/dataset/sube-cantidad-de-transacciones-usos-por-fecha

(a) la proporción de la cantidad total anual de pasajeros que le corresponde a cada medio de transporte

(b) la tupla (mes, línea de subte) donde viajó la mayor cantidad de pasajeros

(c) el día hábil[2] con menor desvío estándar en cantidad de pasajeros

Análisis Exploratorio [5 pts.]

4. *Análisis exploratorio.* 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`) y de resumen de datos (media, mediana, desvío estándar, operaciones sobre el DataFrame, etc.).

El objetivo es entender, comparar y/o estudiar aspectos o patrones en la cantidad de pasajeros del transporte público. Algunas preguntas disparadoras pueden ser:

- ¿Cómo varía el uso del transporte público a lo largo del año? ¿Se observa el mismo efecto en AMBA y en el interior del país? ¿Y en todos los medios de transporte por igual?
- ¿Cómo difiere el uso del transporte público durante los días hábiles en comparación a los fines de semana?
- ¿Hay algún día hábil que resulte un outlier? ¿Pueden explicarlo?

No es necesario que respondan a cada una de esas preguntas, pueden explorar por donde se les ocurra. Tampoco están limitades a usar sólo el dataset de 2023 (en el link hay datasets desde 2020 a 2024). Pueden hacer comparaciones interanuales, enfocarse en un solo tipo de transporte, etc.

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

Modelado [3 pts.]

5. *Modelo de regresión.* En este ítem, intentaremos ajustar la cantidad de pasajeros que viajan por día en una línea de colectivos utilizando la información de pasajeros por día de otras líneas. Para esto, utilizaremos el dataset `sube-2023-regresion.csv`

(a) Generar un DataFrame en el que las columnas sean las líneas de colectivo (`TIPO_TRANSPORTE == "COLECTIVO"`) de AMBA de jurisdicción nacional (`PROVINCIA == "JN"`) y las observaciones sean los días del año. Es decir, cada fila del DataFrame corresponde a un día del año, y en esa fila deben figurar la cantidad de pasajeros que viajaron en cada línea de colectivo en el día correspondiente.

Para generar el DataFrame pueden completar el siguiente código:

[2] Consideramos como día hábil a cualquier día, salvo Sábado y Domingo, sin importar si es feriado.

```
datos_ColectivoJN = datos_AMBA[???]
cols = datos_ColectivoJN.LINEA.unique()  # Los nombres de las
   lineas de colectivo

pasajeros_por_linea = pd.DataFrame()
for col in cols:
    datos_linea = datos_ColectivoJN[datos_ColectivoJN.LINEA ==
   col][["DIA_TRANSPORTE", "CANTIDAD"]]
    datos_linea =
   datos_linea.set_index("DIA_TRANSPORTE").rename(columns =
   {"CANTIDAD" : col})
    pasajeros_por_linea = pd.concat([pasajeros_por_linea,
   datos_linea], axis = 1)
```

(b) Eliminar las columnas correspondientes a líneas de colectivo que tengan datos faltantes.

(c) Se quiere ajustar la cantidad de pasajeros en la línea `BSAS_LINEA_009` en función de los pasajeros en otras líneas. Proponer tres modelos de regresión distintos. En cada modelo, pueden utilizarse la información solo de otras 5 líneas de colectivos (pueden ser distintas líneas de colectivo en los distintos modelos). Los criterios de selección de esas 5 columnas los determinan ustedes, y deben estar explicitados en el informe.

(d) Si alguno de los modelos es Regresión Ridge, determinar mediante un esquema de validación el hiperparámetro α.

(e) Proponer un esquema de validación de los modelos y utilizarlo para seleccionar el mejor de los tres modelos propuestos.

(f) Para el modelo elegido, indicar la fórmula final de modelo.

(g) Los modelos propuestos por los distintos grupos serán testeados en un conjunto de testeo conteniendo días distintos a los utilizados para la construcción del modelo. Los equipos que tengan los mejores desempeños obtendrán un punto bonus en la nota del TP.