Examen Ciencia de Datos
L3/L4

Observaciones generales:

- Envía tus respuestas a más tardar 48 horas después de recibir el correo.
- Las respuestas deberán ser enviadas como un link a tu repositorio de GitHub.
- El examen tiene una sección práctica (sección A), que implica el análisis de datos, y una teórica (sección B), para la cual no es necesario generar código.
- Resuelve la **sección A** en **Python**. Tus respuestas deberán contener, en cada paso, una descripción del razonamiento que seguiste, el código con el que generaste tus resultados y tus resultados.
- Para la **sección B** no tienes que generar código, puedes entregar tus resultados en PDF o en el formato que consideres más conveniente.
- Toma en cuenta que a nosotros nos interesa conocer, sobre todo, tu capacidad de comprensión y planteamiento del problema.
- Busca el 80-20 en cada pregunta (el 20% del trabajo que te da 80% del valor en la solución).

Sección A
Datos abiertos de la CDMX

La Agencia Digital de Innovación Pública tiene disponibles datos de las **carpetas de investigación aportados por la PGJ**. La tabla está disponible aquí: https://datos.cdmx.gob.mx/dataset/carpetas-de-investigacion-fgj-de-la-ciudad-de-mexico/resource/48fcb848-220c-4af0-839b-4fd8ac812c0f
Utilizando estos datos, responde las siguientes preguntas:

1. ¿Qué pruebas identificarías para asegurar la calidad de estos datos? No es necesario hacerlas, sólo describe la prueba y lo que te dice cada una.
2. Identifica los delitos que van a la alza y a la baja en la CDMX (ten cuidado con los delitos con pocas ocurrencias).
3. ¿Cuál es la alcaldía que más delitos tiene y cuál es la que menos? ¿Por qué crees que sea esto?
4. ¿Existe alguna tendencia estacional en la ocurrencia de delitos (mes, semana, día de la semana, quincenas) en la CDMX? ¿A qué crees que se deba?
5. ¿Cuáles son los delitos que más caracterizan a cada alcaldía? Es decir, delitos que suceden con mayor frecuencia en una alcaldía y con menor frecuencia en las demás.
6. Diseña un indicador que mida el nivel de “inseguridad”. Genéralo al nivel de desagregación que te parezca más adecuado (ej. manzana, calle, AGEB, etc.). Analiza los resultados ¿Encontraste algún patrón interesante? ¿Qué decisiones se podrían tomar con el indicador?

Sección B
La Michoacana (Ésta es una sección teórica, no es necesario generar código para resolverla)

Hace algunos años, la nevería La Michoacana decidió hacer un cambio radical en su forma de operar. En lugar de neverías tradicionales, La Michoacana decidió instalar 4,000 máquinas expendedoras de paletas dentro de la Ciudad de México. Además de eso, decidió implementar un sistema de suscripción para las máquinas expendedoras. Es decir, en lugar de pagar por cada paleta, los clientes pagan una suscripción mensual fija por tener acceso a cuántas paletas quieran, cuando quieran. Las paletas se hacen en una planta que tiene un congelador central, desde el cual las paletas se distribuyen a las máquinas expendedoras en camiones refrigerados de una capacidad muy grande.

Las máquinas expendedoras tienen la peculiaridad de que el gasto energético depende de la cantidad de paletas que estén en la máquina, pues cada una se congela en un pequeño cajón. El costo de mantener una paleta por un día es de $1. El transporte de las paletas desde el congelador central a cada máquina expendedora tiene un costo fijo de $100 por viaje a una máquina y cada máquina no puede ser surtida más de una vez al día.

Cada mañana, Luis, el jefe de operaciones, debe decidir qué máquinas expendedoras serán surtidas esa misma mañana y cuántas paletas deben ir a cada máquina expendedora que será surtida. Luis quiere mantener los costos de transporte y de energía bajos pero al mismo tiempo no quiere que existan muchos casos de máquinas sin paletas disponibles, pues los suscriptores podrían molestarse y abandonar el programa. Luis tiene el objetivo de mantener los costos de transporte y energía más bajos posibles y la cantidad de días/máquina donde hubo indisponibilidad de paletas menor a 2% al mes. Luis tiene total libertad todas las mañanas para elegir qué máquinas deben ser surtidas y cuántas paletas surtir en cada una. Para tomar esa decisión, Luis tiene, cada mañana, la siguiente información:

- La cantidad de paletas retiradas para cada máquina, cada día desde que empezó la operación (hace 5 años) hasta el día anterior.
- La cantidad de paletas disponibles en cada máquina, cada día (a final del día, medianoche) desde que empezó la operación (hace 5 años) hasta el día anterior.
- La capacidad, en cantidad de paletas, de cada máquina expendedora.
- El costo de surtir cada máquina ($100)
- El costo de mantener una paleta por un día en cualquier máquina expendedora ($1)
- No hay límite de carga en los camiones repartidores

Luis toma la decisión día con día de qué máquinas surtir y cuántas paletas surtir a cada una de ellas de forma bastante subjetiva, intuyendo que máquinas se quedarán sin paletas si no las surte y tratando de identificar una cantidad óptima de paletas para surtir cada una de ellas. Luis se ha dado cuenta de que se trata de un problema bastante complejo, donde existe una gran oportunidad de tomar mejores decisiones. Se ha dado cuenta de que, para cada máquina, la estrategia de llenar a tope la máquina expendedora y surtirla el día que observe que las paletas están apunto de acabarse para llenar de nuevo la máquina a tope tiene un costo muy alto de energía y un costo bajo de transporte. Se ha dado cuenta también de que, para cada máquina, hacer viajes diarios con el contenido de paletas necesarias para el siguiente día tiene un costo de energía bastante bajo pero un costo de transporte muy alto. Luis observa también que el consumo de paletas cambia mucho de máquina a máquina y día con día. Luis no tiene restricciones como mandar siempre la misma cantidad de paletas a una misma máquina o mandar paletas con frecuencia fija (e.g. cada 7 días). Dada esa libertad, entiende que se trata de

un problema complejo pero que es la clave para lograr sus objetivos de mantener los costos de transporte y energía bajos sin superar su cota de indisponibilidad.

1. Con la información anterior, diseña una solución al problema.
   No existe una única solución válida al problema, buscamos entender cómo tú entiendes el problema y saber cómo lo resolverías.
   La respuesta puede ser textual, utilizando notación matemática, utilizando diagramas, pseudocódigo, un video explicando en un pizarrón, notas en una servilleta, algún otro medio o una mezcla de los anteriores, lo que tú creas que va a comunicar mejor tu planteamiento.
   Puedes guiar tu respuesta por las siguientes preguntas:
   a. ¿De qué tipo de problema se trata?¿Tiene elementos comunes como regresión, clasificación, pronósticos de series de tiempo, clustering, optimización, etc?
   b. ¿Cómo te imaginas una solución funcional al problema?¿De qué partes está conformada?¿Cómo interactúan esas partes?¿Qué supuestos y riesgos ves en tu planteamiento?
   c. ¿Qué métodos o algoritmos utilizarías durante el desarrollo de esa solución?
   d. ¿Qué métricas evaluarías durante el desarrollo de la solución?
   e. ¿Cómo te imaginas el despliegue y la operación en producción de la solución?
   f. ¿Cómo evaluarías si la solución tuvo un impacto positivo y fue exitosa?

   No respondas puntualmente las preguntas anteriores, sólo son una guía que puedes usar para la narrativa de tu respuesta. Lo que más nos interesa es entender de forma coherente, completa y clara **tu** planteamiento del problema y **tu** diseño de solución. Esperamos poder leer tu respuesta y que nos comunique algo similar a que si hicieras una presentación de **tu** entendimiento del problema y diseño de solución al equipo de ciencia de datos de OPI.