Examen Data Scientist

Instrucciones:

- Resuelve las preguntas de cada una de las secciones. No es necesario que contestes el examen completo. Explica todas tus respuestas para ver tu razonamiento, capacidad de comprensión. Muestra tu intuición y tu creatividad. Nos interesa conocer tu forma de plantear los problemas.
- Envía tus respuestas a más tardar 72 horas después de recibir el correo.
- Las respuestas podrán ser enviadas en un PDF, un notebook, o con un link a tu repositorio de GitHub.
- Debes especificar claramente las respuestas de cada inciso, e indicar el código usado.

Sección 1: SQL

Ejercicio 1a: SQL básico

1. Lee los datos en la ruta `data/lyft/lyft_rides_log.csv` y `data/lyft/lyft_users.csv`
2. Crea un notebook de python y utiliza la función `sqldf` del paquete `pandasql` para correr tus queries.
3. Encuentra los 10 usuarios principales que han recorrido la mayor distancia. Muestra sus nombres y la distancia total recorrida.

Ejercicio 1b: Rendimiento del algoritmo

Facebook ha desarrollado un algoritmo de búsqueda que analizará los comentarios de los usuarios y presentará los resultados de la búsqueda a un usuario. Para evaluar el rendimiento del algoritmo, se nos proporciona una tabla que consta del resultado de búsqueda en el que el usuario hizo clic, la consulta de búsqueda del usuario y la posición de búsqueda resultante que se devolvió para el comentario específico. Cuanto más alta sea la posición, mejor, ya que estos comentarios eran exactamente lo que buscaba el usuario.

1. Utiliza los datos de facebook de la ruta `data/facebook/fb_search_events.csv` y `data/facebook/fb_search_results.csv`.

2. Escribe una consulta que evalúe el rendimiento del algoritmo de búsqueda frente a la consulta de cada usuario.

Sección 2: Análisis exploratorio de datos

La Agencia Digital de Innovación Pública tiene disponibles los datos georeferenciados de las carpetas de investigación aportados por la Procuraduría General de Justicia de la Ciudad de México La tabla está disponible aquí: https://datos.cdmx.gob.mx/explore, o bien, en la ruta `data/pgj/carpetas-de-investigacion-pgj-cdmx.csv`. Esta tabla consiste de delitos a nivel de calle de la PGJ desde enero de 2016 hasta junio de 2019.

Contesta las siguientes preguntas:

1. ¿Qué pruebas identificarías para asegurar la calidad de estos datos? No es necesario hacerlas. Sólo describe la prueba y qué te dice cada una.

2. ¿Cuántos delitos registrados hay en la tabla? ¿Qué rango de tiempo consideran los datos?

3. ¿Cómo se distribuye el número de delitos en la CDMX? ¿Cuáles son los 5 delitos más frecuentes?

4. Identifica los delitos que van a la alza y a la baja en la CDMX en el último año (ten cuidado con los delitos con pocas ocurrencias).

5. ¿Cuál es la alcaldía que más delitos tiene y cuál es la que menos?. ¿Por qué crees que sea esto?

6. Dentro de cada alcaldía, cuáles son las tres colonias con más delitos

7. ¿Existe alguna tendencia estacional en la ocurrencia de delitos (mes, semana, día de la semana, quincenas)?

8. ¿Cuales 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.

9. Calcula el número de homicidios dolosos por cada 100 mil habitantes anual para cada Área Geoestadística Básica (AGEB) del INEGI. (Hint: no importa que el dato de población no esté actualizado).

    - Pinta un mapa con este indicador.
    - Describe los resultados.

10. ¿Cómo diseñarías un indicador que midiera el nivel "inseguridad"? Diséñalo al nivel de desagregación que te parezca más adecuado (ej. manzana, calle, AGEB, etc.).

11. Con alguna de las medidas de crimen que calculaste en los incisos anteriores, encuentra patrones de concentración geográfica de delitos.

    - Utiliza un algoritmo de Machine Learning no supervisado.
    - ¿Qué caracteriza a cada punto de concentración de delitos y qué tienen en común?

12. Toma los delitos clasificados como "Robo a pasajero a bordo de transporte público con y sin violencia". ¿Cuáles son las rutas de transporte público donde más ocurren estos delitos?

Sección 3: Spark SQL

Vas a utilizar los datos históricos de la Encuesta Nacional sobre Confianza del Consumidor (ENCO). Los diccionarios de datos se pueden consultar en: (https://www.inegi.org.mx/rnm/index.php/catalog/636/data-dictionary)[https://www.inegi.org.mx/rnm/index.php/catalog/636/data-dictionary]. Los datos están disponibles en (https://www.inegi.org.mx/programas/enco/?ps=microdatos)[https://www.inegi.org.mx/programas/enco/?ps=microdatos].

Resuelve los siguientes incisos. Para el preprocesamiento de datos deberás utilizar R, Python, o bien, un script de shell. Para el procesamiento de los datos y el análisis exploratorio debes debes usar Spark SQL en el lenguaje de programación de tu elección.

1. Preprocesamiento de datos

    a. Escribe una función que descargue los datos de la página de microdatos desde 2001 hasta 2021.
    b. Crea un dataframe que resuma por entidad, municipio, el número de hogares que están planeando comprar un automóvil nuevo o usado en los próximos 2 años, el número de hogares que están planeando comprar, construir o remodelar una casa en los próximos 2 años, el número de hogares en condición de ahorrar, y número de hogares con expectativa de vacaciones. (Considera los factores de expansión de los hogares muestreados.)
2. Procesamiento de datos

    a. ¿Cuántos registros hay?
    b. ¿Cuántos municipios se muestrearon?
    c. ¿Cuántas entidades federativas están en la muestra?
    d. ¿Cómo podrías determinar la calidad de los datos? ¿Cómo detectarías algún tipo de inconsistencia o error en la fuente? (Responde a esta pregunta de manera general.)
3. Análisis exploratorio

    a. ¿Cómo se relacionan las expectativas para comprar un automóvil y para compra/remodelación de una casa?
    b. ¿Cuál es el municipio más caro del país entre los municipios encuestados? ¿Cuál es el más barato?
    c. ¿Hay algún patrón estacional entre años?
    d. ¿Cuál es el estado más caro y en qué mes? (Hacer la estimación para estados utilizando factores de expansión.)
4. Visualización

    a. Genera un mapa que nos permita identificar en qué municipios de la CDMX y el área metropolitana hay mayor expectiva de adquirir un automóvil.

Sección 4: Análisis de un caso

Resuelve el caso que se encuentra en el directorio `data/bops`.

1. ¿Deberían expandirse a Canadá?
2. ¿Cuántos millones de dólares se ganaron o perdieron a partir del programa? Explica tu razonamiento y metodología. (Hint: Existen dos experimentos naturales. Canadá y las tiendas que se encuentran lejos. Utilízalos.)