

Examen Data Science

Nivel avanzado

Instrucciones:

- Resuelve las preguntas de las secciones A, B y C.
- Envía tus respuestas a más tardar 48 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.

Toma en cuenta que a nosotros nos interesa conocer, sobre todo, tu capacidad de comprensión y planteamiento de problemas. Busca el 80-20 en cada pregunta (el 20% del trabajo que te da 80% del valor en la solución) y muestra tu intuición y creatividad. En promedio, la gente sólo resuelve el 70% del examen.

Sección A

A.1 Datos abiertos de la CDMX

La Agencia Digital de Innovación Pública tiene disponibles los datos georeferenciados de las **carpetas de investigación aportados por la PGJ**. La tabla está disponible aquí: <https://datos.cdmx.gob.mx/explore>

- 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).
 - a) Pinta un mapa con este indicador. Describe los resultados.
- 10) ¿Cómo diseñarías un indicador que midiera el nivel "inseguridad"? Diseñalo 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 (hint: puedes usar algoritmos de Machine Learning no supervisados).
 - a) ¿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 ruta de transporte público donde más ocurren estos delitos?

Sección B

B1. QQP

Descarga la Base de datos histórica de [Quién es Quién en los Precios](#) de Profeco y resuelve los siguientes incisos. Para el procesamiento de los datos y el análisis exploratorio debes usar Spark SQL con Python.

1. Procesamiento de los datos
 - a. ¿Cuántos registros hay?
 - b. ¿Cuántas categorías?
 - c. ¿Cuántas cadenas comerciales están siendo monitoreadas?
 - d. ¿Cómo podrías determinar la calidad de los datos? ¿Detectaste algún tipo de inconsistencia o error en la fuente?
 - e. ¿Cuáles son los productos más monitoreados en cada entidad?
 - f. ¿Cuál es la cadena comercial con mayor variedad de productos monitoreados?
2. Análisis exploratorio
 - a. Genera una canasta de productos básicos que te permita comparar los precios geográfica y temporalmente. Justifica tu elección y procedimiento
 - b. ¿Cuál es la ciudad más cara del país? ¿Cuál es la más barata?
 - c. ¿Hay algún patrón estacional entre años?
 - d. ¿Cuál es el estado más caro y en qué mes?
 - e. ¿Cuáles son los principales riesgos de hacer análisis de series de tiempo con estos datos?
3. Visualización
 - a. Genera un mapa que nos permita identificar la oferta de categorías en la zona metropolitana de León Guanajuato y el nivel de precios en cada una de ellas. Se darán puntos extra si el mapa es interactivo

Sección C

C.1 BOPS

Resuelve el siguiente [caso](#):

1. ¿Deberían expandirse a Canadá?
2. ¿Cuántos millones de dólares se ganaron o perdieron a partir del programa? Explica tu razonamiento y metodología.

Pista: Existen dos experimentos naturales.