INTELIGENCIA DE NEGOCIOS
Dr. Mario Garza Fabre
Mayo – Agosto 2018

PROYECTO UNIDAD 2: Exploración de datos

El objetivo de este proyecto es que el alumno ponga en práctica sus habilidades de exploración y análisis de datos para dar respuesta a una serie de preguntas planteadas. El alumno trabajará con datos reales (disponibles de manera gratuita para el público en general) sobre reportes de crímenes en el Reino Unido.

ESPECIFICACIONES GENERALES

- Se plantearán preguntas que el alumno deberá responder mediante la exploración, procesamiento y análisis de los datos.

- El alumno deberá utilizar tablas numéricas y/o métodos gráficos apropiados, que permitan ilustrar claramente los datos y proporcionar respuesta a la información solicitada.

- El alumno podrá utilizar el lenguaje de programación y/o herramientas de análisis de su preferencia.

- Las gráficas deben presentar colores, tipos y tamaños de letra adecuados. También deben de ser suficientemente descriptivas para que puedan interpretarse con facilidad (indicar qué información se presenta, en qué unidades se mide, qué entidades/elementos se comparan, etc, según corresponda).

- Los resultados deberán ser presentados acompañados por su correspondiente interpretación (cuando así se requiera). Es decir, el alumno analizará los resultados obtenidos y escribirá un texto breve donde describa sus observaciones y conclusiones. El texto deberá presentarse en lenguaje formal, con uso adecuado de signos de puntuación, y deberá responder claramente a la pregunta: ¿qué podemos observar y aprender de las tablas y/o gráficas generadas (en relación a la pregunta planteada)?

DATOS DISPONIBLES

Se utilizarán datos reales sobre reportes de crímenes en el Reino Unido. Estos datos se encuentran disponibles para el público en general y pueden descargarse de la siguiente dirección: https://data.police.uk/

Más específicamente, se utilizarán datos para una selección de 42 fuerzas policiacas en Inglaterra y Gales. Los datos a utilizarse corresponden a crímenes reportados entre enero de 2011 y diciembre de 2017.

Adicionalmente, se utilizarán los reportes de la operación "Stop and Search" disponibles para el periodo de abril de 2014 a diciembre de 2017. Es importante notar, que estos reportes no se encuentran disponibles para todas las 42 fuerzas policiacas, y algunas fuerzas policiacas adicionales son consideradas ahora. También, distintas fuerzas policiacas comenzaron a publicar estos datos en diferentes fechas, por lo que el alumno deberá tomar ésto en cuenta al momento de procesar estos archivos para investigar la información solicitada.

Finalmente, se proporcionará al alumno un conjunto de archivos KML (Keyhole Markup Language), a través de los cuales el alumno podrá obtener información sobre el área de cobertura de cada fuerza policiaca (información necesaria para contestar algunas preguntas).

IMPORTANTE: No es posible garantizar por parte del instructor que los datos necesarios para responder las preguntas planteadas están completos. Cuando el dato relevante para responder una pregunta no se encuentra disponible en un registro específico, **el alumno deberá ignorar dicho registro** al momento de generar la respuesta para esta pregunta (para el resto de las preguntas el registro podrá ser utilizado, siempre y cuando los datos relevantes en cada caso se encuentren disponibles).

PREGUNTAS A INVESTIGAR

La siguiente tabla presenta la lista de preguntas de interés, los resultados que se esperan para dar respuesta a cada una de ellas, y finalmente su ponderación con respecto a la calificación de este proyecto.

#	Pregunta	Resultado esperado	% Eval.
1	¿Cuántos crímenes se reportaron en total desde enero de 2011 hasta diciembre de 2017? (considerando las 42 fuerzas policiacas)	1. Cantidad total.	**2**
2	¿Cómo ha cambiado el número total anual de crímenes desde 2011 hasta 2017?	1. Gráfica de barras o una curva (o alguna otra alternativa gráfica que considere conveniente) contrastando los totales de cada uno de los 7 años. 2. Interpretación de resultados.	**4**
3	¿Cómo ha cambiado el número total mensual de crímenes desde 2011 hasta 2017?	1. Gráfica de barras o una curva (o alguna otra alternativa gráfica que considere conveniente) contrastando los totales de cada uno de los 84 meses. 2. Interpretación de resultados.	**4**

4	¿Cómo se comparan los 7 años (2011-2017) en términos de sus totales mensuales de reportes de crímenes? ¿Existen patrones comunes entre los diferentes años? ¿Cuáles son los años que exhiben un comportamiento más parecido?	1. Gráfica de curvas, donde cada curva represente los 12 valores para un año (7 curvas en total). 2. Matriz de correlación, donde se muestre el coeficiente de correlación calculado para cada pareja de años. Utilizar correlación de Spearman. 3. Representación gráfica que facilite la visualización e interpretación de la matriz de correlación del punto anterior. Utilizar, por ejemplo, mapas de calor (heatmaps). 4. Matriz de dispersión (scatter matrix), donde se muestre una gráfica de dispersión (scatter plot) para cada pareja de años. 5. Interpretación de resultados.	**13**
5	¿Cuántos y qué porcentaje de los crímenes reportados entre 2011 y 2017 corresponden a cada fuerza policiaca?	1. Gráfica de pastel o de barras, o alguna otra alternativa que considere conveniente. 2. Interpretación de resultados.	**4**
6	¿Cuántos y qué porcentaje de los crímenes reportados entre 2011 y 2017 corresponden a cada tipo de crimen?	1. Gráfica de pastel o de barras, o alguna otra alternativa que considere conveniente. 2. Interpretación de resultados.	**4**
7	¿Cuántos y qué porcentaje de los crímenes reportados entre 2011 y 2017 corresponden a cada combinación de fuerza policiaca y tipo de crimen?	1. Tabla de contingencia que detalle la información solicitada (tipo de crimen vs. fuerza policiaca). 2. Representación gráfica que facilite la visualización e interpretación de la tabla de contingencia del punto anterior. Utilizar, por ejemplo, mapas de calor (heatmaps). 3. Interpretación de resultados.	**8**

8	¿Algunos tipos de crimen ocurren con mayor o menor frecuencia dependiendo de la temporada (diferentes meses del año)? **Nota: para cada mes del año, considerar todos los reportes de crímenes que se tienen registrados (2011-2017).**	1. Tabla de contingencia que detalle la información solicitada (tipo de crimen vs. mes del año). 2. Representación gráfica que facilite la visualización e interpretación de la tabla de contingencia del punto anterior. Utilizar, por ejemplo, mapas de calor (heatmaps). 3. Interpretación de resultados.	**8**
9	¿Existe una relación entre el número de crímenes reportados por cada fuerza policiaca y el tamaño de su área de cobertura?	1. Cálculo de los coeficientes de correlación de Pearson y Spearman (calcular y reportar los dos coeficientes). 2. Gráfica de dispersión (scatter plot) donde se visualicen los datos en cuestión. 3. Agregar una “linea/recta de mejor ajuste” (best fit line) a la gráfica de dispersión del punto anterior (no es necesario presentar una figura adicional, modificar la figura original). 4. Interpretación de resultados.	**12**
10	¿Cuántos y qué porcentaje de los crímenes reportados por cada fuerza policiaca han ocurrido en una ubicación fuera de su área de cobertura?	1. Gráfica de barras, o alguna otra alternativa que considere conveniente. 2. Interpretación de resultados.	**5**
11	¿Qué fuerzas policiacas exhiben niveles de actividad más parecidos de acuerdo con sus números mensuales de reportes de crímenes? **Nota: considerar únicamente las 10 fuerzas policiacas con mayor cantidad total de crímenes reportados (desde 2011 hasta 2017).**	1. Gráfica de curvas, donde cada curva represente los 84 meses para cada fuerza policiaca (10 curvas en total). 2. Matriz de distancias, donde se muestre la distancia calculada entre cada pareja de curvas de las fuerzas policiacas. Utilizar RMSD (root-mean-square deviation) como medida de distancia. 3. Representación gráfica que facilite la visualización e interpretación de la matriz de distancias del punto anterior. Utilizar, por ejemplo, mapas de calor (heatmaps). 4. Interpretación de resultados.	**10**

12	¿Qué fuerzas policiacas exhiben un comportamiento más parecido en términos del cambio (con el paso del tiempo) en su número mensual de reportes de crímenes? **Nota: considerar únicamente las 10 fuerzas policiacas con mayor cantidad total de crímenes reportados (desde 2011 hasta 2017).**	0. Incluir gráfica de curvas generada para la pregunta 11 (es necesario incluirla, pero no hay calificación por hacerlo). 1. Matriz de correlación, donde se muestre el coeficiente de correlación calculado para cada pareja de curvas de las fuerzas policiacas. Utilizar correlación de Spearman. 2. Representación gráfica que facilite la visualización e interpretación de la matriz de correlación del punto anterior. Utilizar, por ejemplo, mapas de calor (heatmaps). 3. Interpretación de resultados.	**8**
13	¿Cuántas aplicaciones de la estrategia “Stop and Search” se reportaron desde abril de 2014 hasta diciembre de 2017?	1. Cantidad total.	**2**
14	¿Qué porcentaje de las aplicaciones de la estrategia “Stop and Search” se enfocaron en los diferentes tipos de inspección (personas, vehículos, personas y vehículos)?	1. Gráfica de pastel, barras o curva (u otra alternativa que considere conveniente). 2. Interpretación de resultados.	**4**
15	¿La estrategia “Stop and Search” se aplica con mayor frecuencia para la inspección de personas con un cierto rango de edad?	1. Gráfica de pastel, barras o curva (u otra alternativa) contrastando el porcentaje de aplicaciones de esta estrategia a personas de los diferentes rangos de edad. 2. Interpretación de resultados.	**4**
16	¿La estrategia “Stop and Search” se aplica de manera imparcial para la inspección de personas de los diferentes grupos étnicos? **Nota: obtener grupo étnico de “Officer-defined ethnicity”.**	1. Gráfica de pastel, barras o curva (u otra alternativa) contrastando el porcentaje de aplicaciones de esta estrategia a personas de los diferentes grupos étnicos. 2. Interpretación de resultados.	**4**
17	¿La estrategia “Stop and Search” se aplica de manera imparcial para personas de los diferentes géneros (hombres, mujeres, ...)?	1. Gráfica de pastel, barras o curva (u otra alternativa) contrastando el porcentaje de aplicaciones de esta estrategia a personas de los diferentes géneros. 2. Interpretación de resultados.	**4**

18	¿La estrategia “Stop and Search” se aplica con mayor o menor frecuencia dependiendo de la temporada (diferentes meses del año)?	1. Gráfica de barras o curva donde se muestre, para cada mes, el promedio del número de operaciones reportadas (promediar el total mensual de los años para los que se tiene información). 2. Agregar “bigotes” (gráfica de errores, error bar) a las barras o curva del punto anterior. Los bigotes inferiores y superiores representarán, respectivamente, los valores mínimos y máximos del conjunto de valores usado para calcular el promedio (no es necesario presentar una figura adicional, modificar figura original). 3. Interpretación de resultados.	**7**
19	¿Qué fuerzas policiacas presentan una mayor cantidad de datos faltantes? **Nota: en un mismo registro, cada campo vacío será tomado en cuenta como un dato faltante.**	1. Gráfica de barras (o alguna otra alternativa gráfica que considere conveniente) contrastando los totales de datos faltantes para cada fuerza policiaca (tomando en cuenta todos los registros desde enero de 2011 a diciembre de 2017). 2. Interpretación de resultados.	**4**
20	¿Cuáles son los datos (campos) que faltan con mayor frecuencia en los reportes de crímenes? ¿En qué porcentaje del total de registros falta cada dato?	1. Gráfica de barras (o alguna otra alternativa gráfica que considere conveniente) comparando el porcentaje de veces que cada dato ha faltado (tomando en cuenta todos los registros desde enero de 2011 a diciembre de 2017). 2. Interpretación de resultados.	**4**
21	Pregunta Extra 1. Plantear una pregunta interesante que sea diferente a las planteadas por el instructor. **Nota: confirmar con el instructor si la pregunta es admisible antes de incluirla.**	1. Gráfica(s). 2. Interpretación de resultados.	**5**
22	Pregunta Extra 2. Plantear una pregunta interesante que sea diferente a las planteadas por el instructor. **Nota: confirmar con el instructor si la pregunta es admisible antes de incluirla.**	1. Gráfica(s). 2. Interpretación de resultados.	**5**
		TOTAL:	**125**

ENTREGABLES DEL PROYECTO

Cada equipo de trabajo deberá entregar un único archivo (.zip o tar.gz) con el siguiente contenido:

1. **Reporte de resultados.**
    - Portada. Importante incluir nombres de los integrantes del equipo.
    - Resultados. Responder las preguntas en el orden establecido (de acuerdo con la tabla de preguntas). Para cada pregunta, presentar: número y enunciado de la pregunta, resultados obtenidos.
2. **Código fuente.** Deberá entregar el código fuente (o equivalente) utilizado para generar la(s) respuesta(s) de cada pregunta. Indicar claramente (mediante el nombre de los archivos) qué pregunta se resuelve con cada archivo fuente. Es posible que un mismo archivo fuente se utilice para generar diferentes respuestas; en tal caso, el código debe de estar documentado para indicar claramente qué parte del código corresponde a cada una de ellas.

EVALUACIÓN

- El 100% del valor de este proyecto será distribuido y evaluado como se indica en la tabla de preguntas.
- Cada pregunta tiene asignado un porcentaje de la puntuación máxima para este proyecto. Para preguntas en las que la respuesta esperada consiste de dos o más elementos (tablas, gráficas, interpretación, etc), la puntuación de dicha pregunta será dividida entre el número de elementos.
- También se han incluido preguntas opcionales que el alumno podrá responder para obtener puntos adicionales, mismos que podrá utilizar en la unidad de su preferencia.
- **IMPORTANTE:** Responder una pregunta no garantiza que se otorgará la puntuación máxima que ésta puede alcanzar. Para que se otorgue la puntuación máxima, deberán presentarse **TODOS** los elementos solicitados, y cada uno de ellos deberá cumplir con los **ESTÁNDARES DE CALIDAD** establecidos.

FECHA Y HORA LÍMITE DE ENTREGA

La fecha y hora límite de entrega será: **Viernes 15 de junio de 2018, 23:55 horas.**

NOTA: Recuerden que la puntuación máxima que puede obtener un proyecto se reducirá 10 % por cada día hábil de retraso en su entrega.