

Caso 1: New York City Taxi

Este primer caso gira entorno a al sistema de transporte público de la Ciudad de New York (USA). Desde 2009 la *New York City Taxi & Limousine Commission* ha publicado datos acerca de los recorridos hechos por los distintos servicios de Taxi de la ciudad¹. Para resolver este caso tendréis suficiente con analizar los datos de dos meses cualquiera del año 2015. Tened en cuenta que los Yellow Cabs y los Green Cabs se rigen por normativas distintas en cuanto a recogida de pasajeros, por lo que deberéis combinar los datos de ambas flotas. Con estos datos se os propone que apliquéis técnicas de estadística, analítica, minería de datos y visualización para responder a las siguientes preguntas. No hay restricciones acerca de las técnicas ni tecnologías a utilizar siempre y cuando los resultados sean reproducibles y estén debidamente justificados. Explicitad y detallad todos los pasos hechos para responder a cada pregunta.

PRIMERA PARTE: ANÁLISIS CUANTITATIVO.

- 1.1 Primer examen preliminar del *dataset*. ¿Qué parámetros hay en el *dataset*? ¿Cuál es su significado? ¿Existen valores aparentemente incorrectos?
- 1.2 Empezamos por visualizar el *dataset*. Haced un *plot* de los puntos de recogida y otro con los puntos de llegada del *dataset* de los cuatro ficheros y extrae conclusiones preliminares. ¿Se aprecian diferencias entre los puntos de los Yellow Cabs y los de los Green Cabs? ¿Se aprecian diferencias entre un mes y otro?
- 1.3 Mejora la visualización anterior con un *heat map*.

SEGUNDA PARTE: ANÁLISIS CUALITATIVO.

- 2.1 ¿Cuál es el trayecto en el que la relación precio/km es más alta? ¿Cuál es el trayecto en el que la relación tiempo/km es más alta? ¿Cuál es el trayecto en el que la relación precio/tiempo es más alta?
- 2.2 ¿Cuál es el trayecto en el que la relación precio/km es más baja? ¿Cuál es el trayecto en el que la relación tiempo/km es más baja? ¿Cuál es el trayecto en el que la relación precio/tiempo es más baja?
- 2.3 Muestra la evolución del tiempo medio de trayecto a lo largo del día. Muestra la evolución de la distancia media de trayecto a lo largo del día.
- 2.4 En la película "Jungla de Cristal 3: La venganza", Bruce Willis y Samuel L. Jackson deben ir en menos de 30 minutos desde la 72 con Broadway hasta la parada de metro de Wall Street (Broadway con Rector St) para evitar que estalle una bomba². Calcula si esto es posible y en caso de que lo fuera, qué probabilidad de éxito tendrían (calculada como casos de éxito dividido por casos totales).

TERCERA PARTE: ANÁLISIS PREDICTIVO.

- 3.1 ¿Cuáles son las zonas donde es más probable coger un taxi en función de la hora del día?
- 3.2 ¿Cuál es la mejor hora del día para ir al aeropuerto?
- 3.3 Diseña un modelo que, dada una hora, unas coordenadas origen, y unas coordenadas destino, predice la duración del trayecto y su coste. Muestra la relevancia de los atributos del *dataset*.

¹ Los datos están disponibles aquí: <https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page>

² <https://www.youtube.com/watch?v=0m5QoYnKne0>