

#### Máster en Big Data

## Asignatura: Casos de analítica

## Fecha: 19/03/2024

Apellidos: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Nombre: \_\_\_\_\_\_\_\_\_\_ Login \_\_\_\_\_\_\_\_\_\_ Curso: **2023/24**

**Caso 1: New York City Taxi**

Este primer caso gira entorno 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 ciudad[[1]](#footnote-1). Para resolver este caso tendréis suficiente con analizar los datos de dos meses cualquiera del año 2023. 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.

Parte 0: Propuesta de Diseño del *Pipeline* de Datos.

1. Propón y justifica una arquitectura completa (end-to-end) que permita soportar todo el ciclo de vida de los datos de este caso de uso:
   1. Adquisición de datos: ¿Qué dispositivo se podría utilizar para capturar la información (origen, destino, tiempo, ocupantes, precio, …) acerca de los trayectos de cada uno de los taxis de New York?
   2. Infraestructura para el almacenamiento y procesamiento de datos: Detalla qué tecnologías y servicios (*cloud!*) se podrían utilizar para almacenar esta información. Estima el coste mensual que tendría tu solución.
   3. Tecnologías para el almacenamiento y procesamiento de datos: Detalla qué tecnologías de almacenamiento y procesado de datos se podrían utilizar para almacenar toda esta información.
   4. Tecnologías para la visualización. ¿Qué tecnología permitiría mostrar un *dashboard* a tiempo real con la ubicación de cada taxi?

Primera Parte: Análisis Cuantitativo.

* 1. Primer examen preliminar del *dataset*. ¿En qué formato está el dataset y qué tiene que ver este formato con Big Data? ¿Qué parámetros hay en el *dataset*? ¿Cuál es su significado? ¿Existen valores aparentemente incorrectos?
  2. Empezamos por visualizar el *dataset*. Haced un *plot* de las zonas de recogida y otro con las zonas 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?
  3. Mejora la visualización anterior con un *heat map*. Ayúdate de los *Zone Map…*

Segunda Parte: Análisis Cualitativo.

* 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. ¿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?
  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.
  4. Elige dos zonas cualquiera de la ciudad y calcula la probabilidad de desplazarse de una zona a otra en menos de X minutos. (El valor X así como las zonas deben ser fácilmente modificables).
  5. Repite los apartados 1.1, 1.2, 1.3, 2.1, 2.2, 2.3 y 2.4 con un dataset del mismo mes pero del año 2009. Comenta todas las diferencias que vayas encontrando.

Tercera Parte: Análisis Predictivo.

* 1. ¿Cuáles son las zonas donde es más probable coger un taxi en función de la hora del día?
  2. ¿Cuál es la mejor hora del día para ir al aeropuerto?
  3. Diseña un modelo que, dada una hora, una zona origen, y una zona 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 [↑](#footnote-ref-1)