# Ejercicio – Data Science – Quilmes

**Contexto**

El equipo de Inteligencia Comercial está a cargo de la generación de campañas de marketing donde se ofrecen determinados productos bajo algún tipo de promoción.

El equipo de inteligencia comercial contacto al de Ciencia de Datos para entender que tan rentables están siendo sus campañas de Marketing, para lograr este objetivo solicitan calcular un Lift.

El lift se interpreta como: el % de incremental de ventas generado por haber efectuado alguna de estas campañas de marketing en comparación con un escenario donde no se hubiera efectuado ninguna de estas.

Estos cálculos de lift bajo este contexto son “*ex post*” se calculan con información de ventas generada en los últimos meses.

No hay posibilidad de efectuar un A/B testing en alguna campaña futura.

**Datos disponibles**

Se suministra la siguiente información:

Datos del cliente:

* clients.csv
* channel\_segmentation.csv
* La columna “chanel\_segmentation” se refiere al tipo de negocio donde se venden los productos: Autoservicios, kioscos y tradicional (almacén).
* La columna "business” se refiere al tipo de producto al que hace referencia la campaña (CZA: Cerveza y NABS: bebidas sin alcohol).

Datos de las promociones usadas en los clientes:

* coupons.csv
* Blacklist.csv
* El data set “coupons.csv” tiene una columna que se llama “sales\_with\_coupons”. Esta columna tiene la cantidad de ventas que se generaron bajo alguna promoción efectuada durante alguna campaña.
* Según el área de inteligencia comercial, se considera que el cliente fue participe de la campaña, si efectúo más de 5 compras (“sales\_with\_coupons” > 5).
* El data set “blacklist.csv” indica los clientes que no hay que considerar en el estudio por motivos varios.

Datos de las ventas del año 2022 y primer semestre de 2023:

* sales\_2022\_semestre\_1.csv
* sales\_2022\_semestre\_2.csv
* sales\_2023\_semestre\_1.csv

Para efectuar el ejercicio se va a tener que proceder a construir un data frame integrando esta información. Corresponde al Data Scientist identificar las claves y criterio de integración de datos.

**Consigna**

En base a esto se solicita calcular un nuevo número de Lift(beneficio generado) teniendo en cuenta:

* El lift debe referenciar a periodos de 2023, al menos 3 meses.
* Debería considerar solo productos del negocio CZA.
* Uno para cada uno de estos tres canales (“chanel\_segmentation”): AUTOSERVICIOS, KIOSCOS y TRADICIONAL.

Para efectuar el cálculo se recomienda generar dos muestras: una de tratamiento (adopción de la campaña) y otra de control (campaña no adoptada) y comparar el impacto en las ventas.

Corresponde al Data Scientist identificar el procedimiento para generar estas dos muestras como también identificar el período de tiempo a considerar para generar estas muestras.

Es un plus (no obligatorio) si las muestras son generadas bajo algún procedimiento estadístico (clustering, clasificación, etc).

El entregable debe incluir:

* Planteamiento del problema y solución
* Notebook de Python con la resolución.
* Tabla con el Lift resultado para CZA en cada uno de los tres canales solicitados.
* Un informe para el área de Inteligencia Comercial con una explicación no técnica de los resultados. Este informe hay que presentarlo en las dos entrevistas del proceso de selección. Siendo la segunda entrevista en inglés.
* Aclaración: en caso de que algún requerimiento sea ambiguo, tomar una asunción y aclararla en el entregable

**Ayuda**

* El ejercicio debe incluir un EDA. Exploración de datos. Se recomienda eliminar clientes outliers, duplicados o con un comportamiento anómalo.
* El cálculo de LIFT debe considerar la diferencia de performance de ventas entre el grupo tratamiento (adopción de la campaña) y el grupo de control (campaña no adoptada). Estas muestras deben construirse en base a los datos disponibles.
* Es deseable que esas dos muestras deben ser lo más similares posibles y la única diferencia entre ambas debe ser la adopción de la campaña (se recomienda aplicar algún tratamiento estadístico como limpieza de outliers o muestreo estratificado para que las muestras sean similares, esto es deseable no obligatorio).