# Taller 2 - Logística Internacional

## Descripción

El archivo `sales.csv` contiene un amplio registro de transacciones realizadas en una empresa de retail, con detalles logísticos respecto a las transacciones:

- `Region`: región o continente a donde se enviará el pedido (región destino)
- `Country`: país a donde se enviará el pedido (país destino)
- `Item_Type`: categoría de producto enviado
- `Sales_Channel`: Canal de ventas del producto (canal en donde fue ordenado)
- `Order_Priority`: Nivel de prioridad de envío (H: alta, L: baja)
- `Order_Date`: Fecha en que se realizó la orden por canal de ventas
- `Order_ID`: Id de la orden
- `Ship_Date`: Fecha en que fue enviado a destino
- `Units_Sold`: Unidades de producto
- `Unit_Price`: Precio Unitario
- `Unit_Cost`: Costo Unitario
- `Total_Revenue`: Ingreso total (precio unitario $\times$ unidades)
- `Total_Cost`: Costo total (costo unitario $\times$ unidades)
- `Total_Profit`: Utilidad total (Ingreso total $-$ Costo total)

Junto con ello, el archivo `countries.csv` incluye información de los países (como un código identificador y su posición en un mapa):

- `country`: código del país (2 caracteres)
- `latitude`: Latitud en grados
- `longitude`: Longitud en grados
- `name`: Nombre del país

Cargar los datos en una tabla en SQL Server y elaborar las consultas para analizar cada uno de los siguientes problemas

# 1 Distancias al centro de distribución

En esta parte del taller asumiremos que nuestra empresa de retail tiene su centro de distribución en Estados Unidos, una gran bodega desde donde realiza los envíos a todo el mundo. Como tenemos las coordenadas de los países a donde se hacen los envíos, podremos saber a que distancia se encuentran de Estados Unidos utilizando una *función de Semiverseno* ampliamente aprobada para calcular la distancia entre dos puntos de nuestro planeta utilizando sus coordenadas.

(1) Construir una consulta de cruce entre la tabla `sales` y la tabla `countries` que nos entregue los siguientes campos para cada registro de `sales`:

   - `order_id (sales)`
   - `country (sales)`
   - `latitude (countries)`
   - `longitude (countries)`
   - Diferencia en días entre fecha de envío y fecha de la orden `DATEDIFF(d,sales.order_date,sales.ship_date)` que llamaremos `dias_t`
   - `units_solds (sales)`
   - `unit_price(sales)`

(2) Crear una función de *distancia* que recibirá 4 parámetros (2 latitudes y 2 longitudes) y entregará como output la distancia en kilómetros entre 2 puntos: `distancia(@lat_1 float, @long_1 float, @lat_2 float,@long_2 float)`→`distancia[Km]`. La función viene lista para ser implementada en el archivo `semiverseno.sql`.

(3) Incluir un campo de distancia en la consulta elaborada en (1) calculando la distancia a Estados Unidos (que llamaremos `distancia_us`) y usando para esto la función implementada en (2) (Coordenadas Estados Unidos lat=37.09024/long=-95.712891). Cargar esta nueva consulta en una tabla temporal.

# 2 Correlación entre variables

La correlación es una medida estadística que expresa hasta qué punto dos variables están relacionadas linealmente (esto es, cambian conjuntamente a una tasa constante). Por ejemplo, si dos variables tienen correlación positiva quiere decir que se mueven en la misma dirección y linealmente en magnitud, en promedio, mientras que si dos variables tienen correlación negativa, se mueven en sentidos contrarios y linealmente en magnitud, en promedio. No confundir con causalidad (por ejemplo si una variable sube y esto causa que otra variable suba, desde la correlación no debemos inferir esto). Si la correlación es cercana a cero, no se puede concluir que exista una relación lineal entre ambas variables.
Para calcular la correlación en SQL entre un `campo1` y un `campo2` podemos usar las funciones de agregado `AVG` y `STDEVP` de la siguiente manera:

```
SELECT
(AVG(campo1 * campo2) - (AVG(campo1) * AVG(campo2))) / (STDEVP(campo1) * STDEVP(campo2))
FROM tabla
```

(1) Calcular la correlación entre `distancia_us` y `dias_t`, a modo de ver si mayor distancia a Estados Unidos se relaciona con mayores días de envío. ¿Su resultado es positivo? ¿cercano a cero?¿Podríamos concluir que a mayor distancia del centro de distribución mayor tiempo de envío?

(2) Se supone que una orden con mayor prioridad debiera tener un envío más rápido, ¿Es así? Calcular con una consulta agrupada el promedio de dias de envío (`AVG(dias_t)`) según prioridad de la orden (`sales.order_priority`)

(3) Calcular de igual manera que en (1) la correlación entre dias de envío (`dias_t`) y unidades enviadas (`sales.units_sold`) y también la correlación entre dias de envío (`dias_t`) y utilidad generada de la orden (`sales.total_profit`), ¿se evidencia mayor correlación en estos últimos casos?

# Entrega y Plazo

En un archivo ppt (Presentación de Power Point) enviar 1 captura de pantalla por cada pregunta (por ejemplo sección 1 y 2 son 3 capturas lo que hace un total de 6 capturas. Enviar el archivo con su nombre (nombreyapellido.ppt) a `jcarrasco@programbi.cl`.
Plazo de recepción: quinta (y penúltima) clase.