UNIVERSIDAD DON BOSCO

FACULTAD DE INGENIERÍA

DEPARTAMENTO DE COMPUTACIÓN

DATAWAREHOUSE Y MINERÍA DE DATOS

DESAFIO 01

JOSÉ ROLANDO ÁLVAREZ MEJÍA AM232553

ENLACE DE REPOSITORIO

https://github.com/Josepo616/Desafio01_DMD

SAN SALVADOR, 11 de marzo de 2025

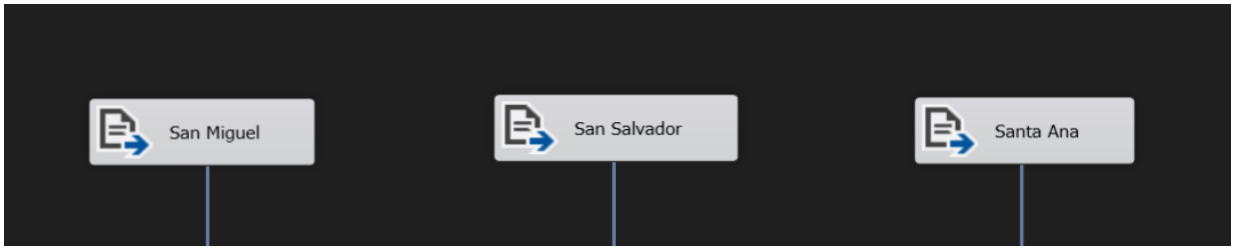
Contenido

Descripción del proyecto

Este proyecto busca como objetivos el desarrollo de un proceso ETL utilizando SQL server integration services.

Desarrollado mediante un objetivo que busca una empresa que se dedica a la floristería y la cual posee distintas fuentes de datos csv las cuales serán procesadas, transformadas y guardadas en unos archivos Excel mediante el requerimiento de la empresa

Los recursos principales serán los siguientes:



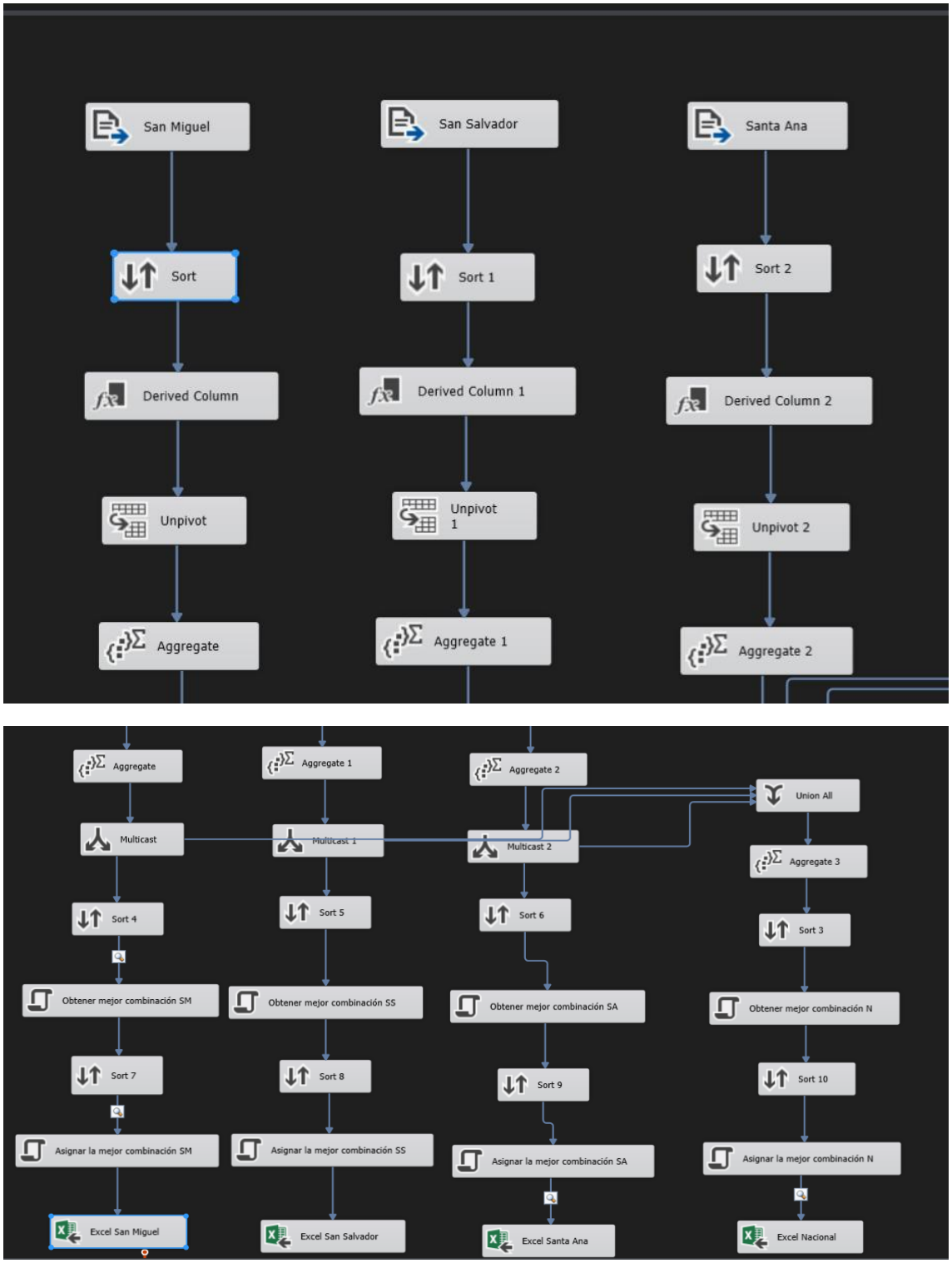
Los cuales representas 3 archivos csv con datos de ventas a nivel departamental y para los cuales la empresa requiere realizar estadísticamente cuántas han sido las ventas totales, las ventas por producto, tendencias de compras y mejores combinaciones

Así mismo la estructura de cada archivo es la siguiente

id,Rosas,Claveles,Macetas,Tierra,Girasoles,Hortensia,Globos,Tarjetas,fOrquídias, Carmesí,Lirios,Aurora,Tulipanes,Listón

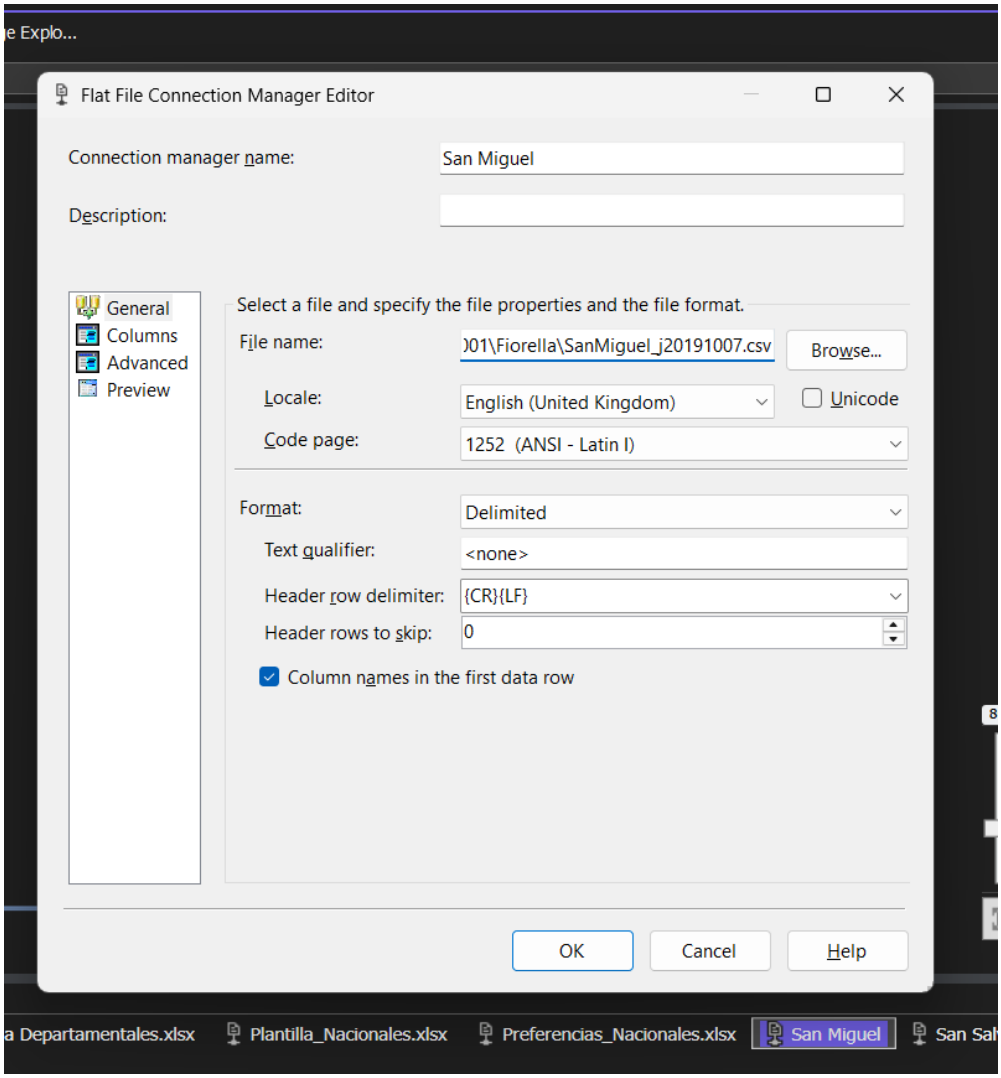
En el cual el id se involucra el nombre de un cliente pero para nuestro análisis no será necesario

La solución planteada será la siguiente

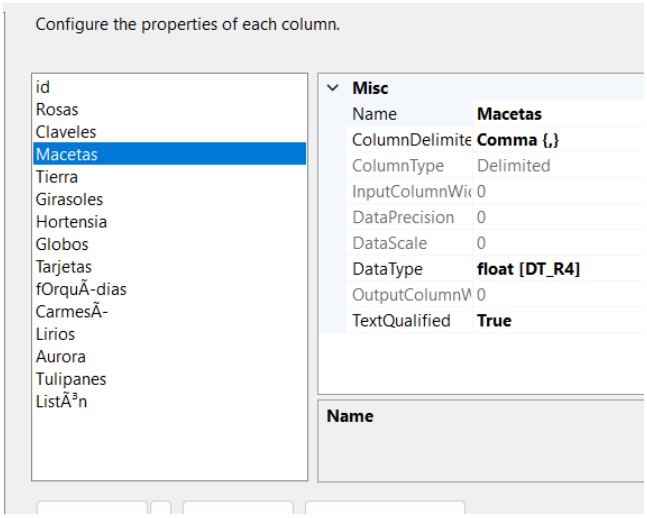


En la cual se tratan de cubrir todos los requisitos de software recomendados, como la extracción correcta desde el origen, la transformación de los tipos de datos necesarios, las operaciones lógicas o matemáticas necesarias para el correcto flujo de datos, posteriormente la validación en los archivos de destino y finalmente la inserción de todos los datos necesarios

Conexión con los datos de origen



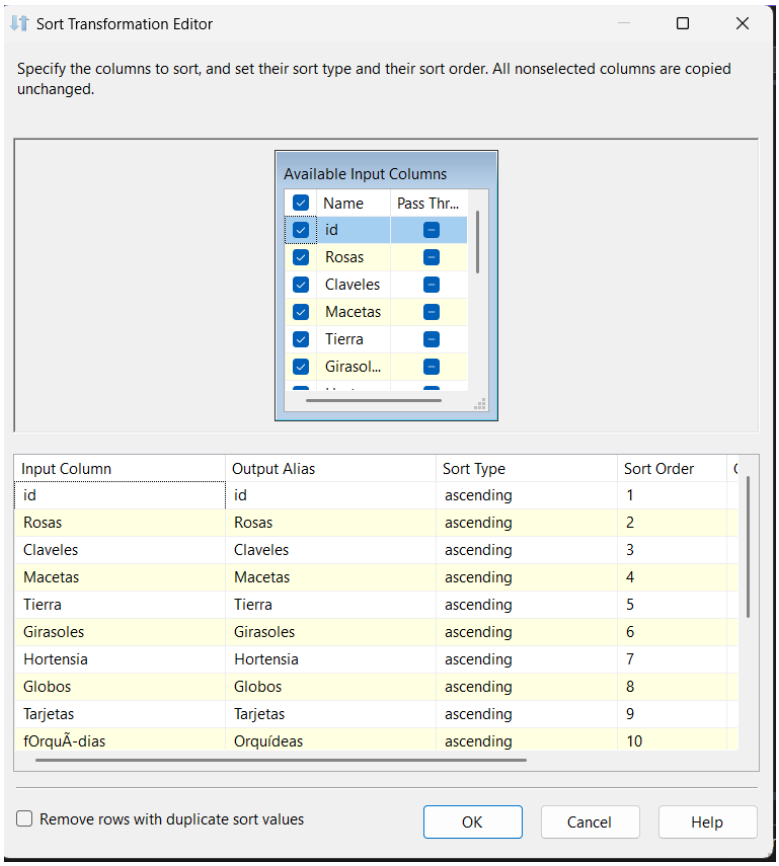
Colocándole el nombre de el departamento para una mejor comprensión, accedemos mediante la ruta de el archivo csv previamente almacenada en una ubicación de fácil acceso



Transformamos los datos necesarios a conveniencia, en nuestro caso de float para todas las plantas y Unicode string para el campo id

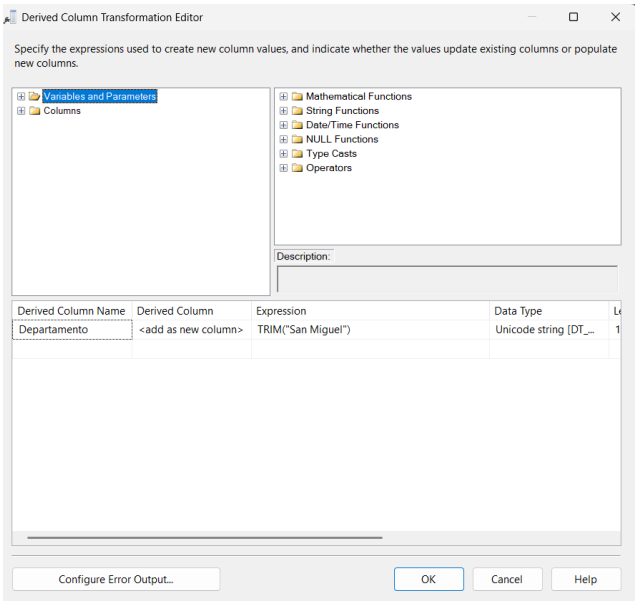
Componentes del flujo de datos

Sort



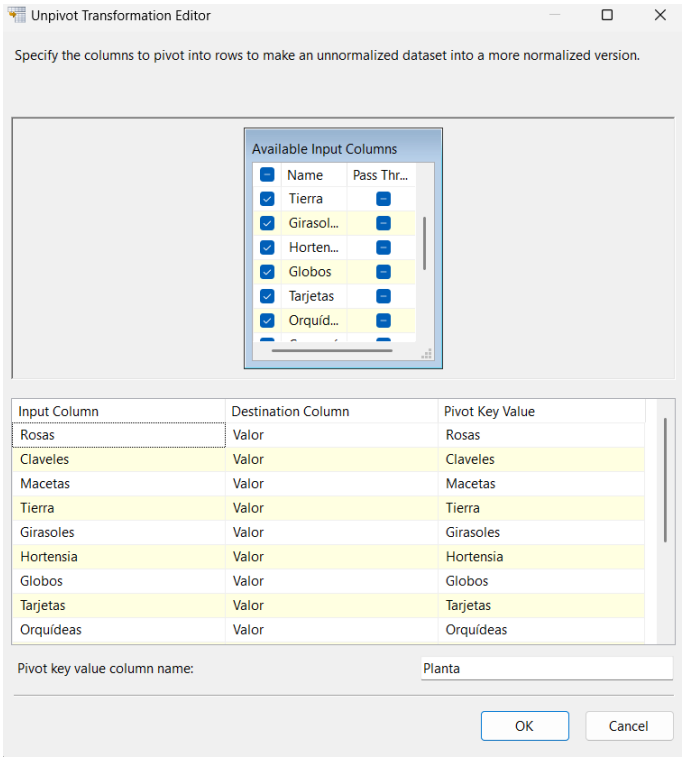
Procedemos con una herramienta sort para ordenar nuestros datos comenzando por id, lo que permitirá que estén ordenados alfabéticamente

Derived Column



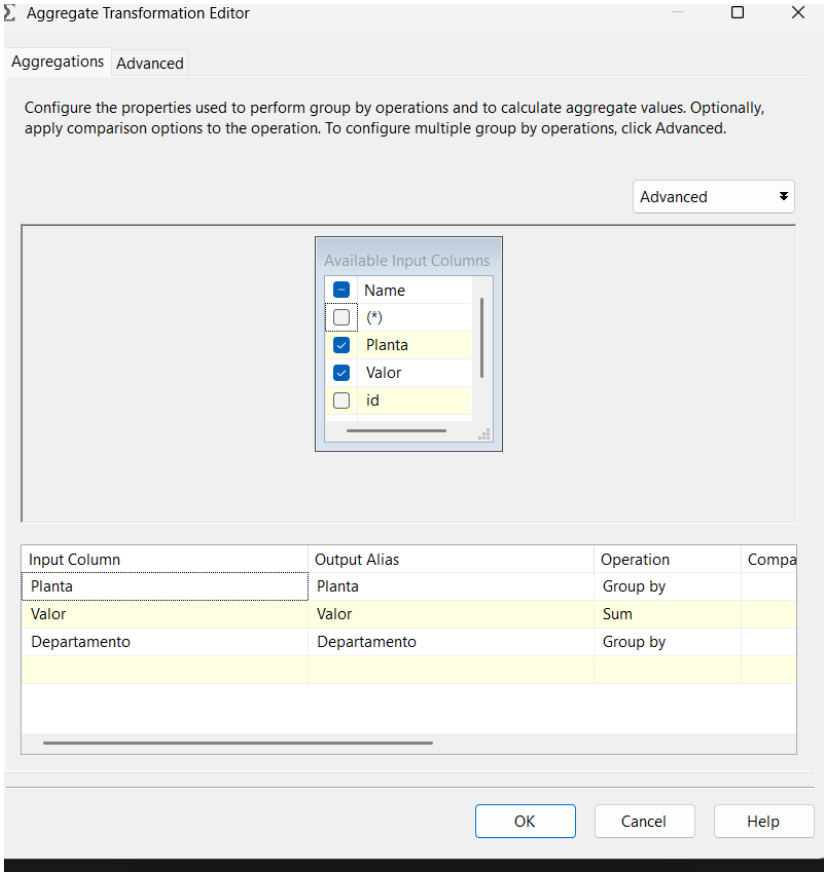
Mediante una herramienta derived column añadimos una nueva columna que establecerá el departamento

Unpivote



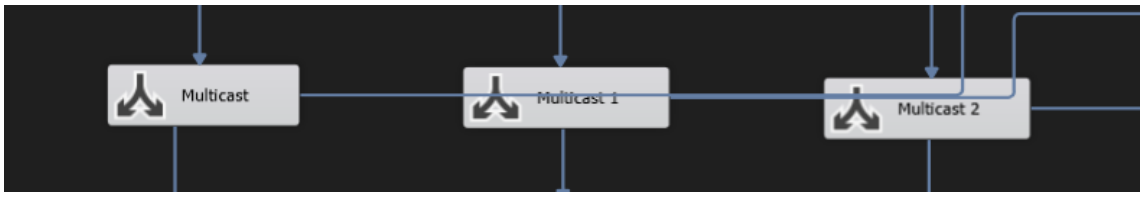
Con una herramienta unpivot realizamos un intercambio de columnas a filas para todas las plantas que antes eran columnas, ahora se convirtieron en filas para el encabezado “Planta” y el dato que poseían se coloca en la columna con encabezado “Valor”

Aggregate



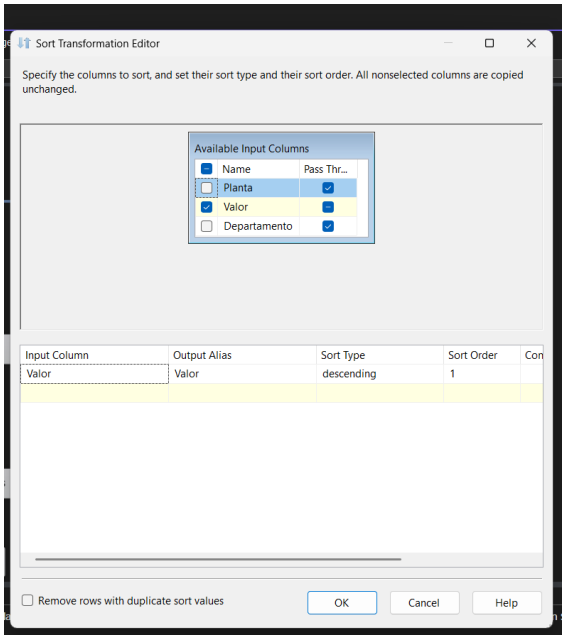
Ahora con la herramienta de aggregate hacemos un agrupamiento por planta y más importante la columna valor la sumamos considerando dicho agrupamiento previo realizado

Multicast



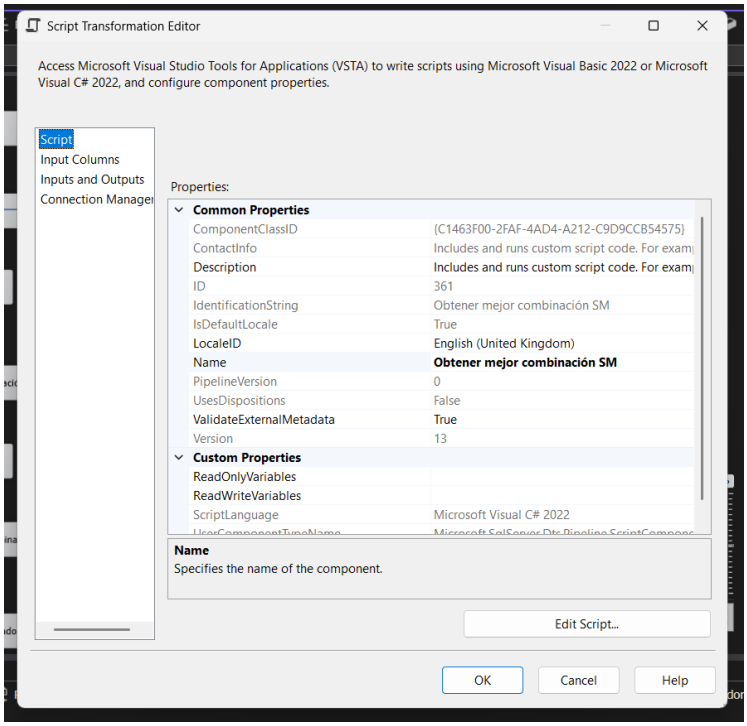
En este punto añadimos herramientas multicast para permitirnos distintas salidas que explicaremos en los demás puntos

Sort



Nuevamente con una herramienta sort realizamos un reordenamiento mediante el valor más alto hasta el más bajo

Script component



Ahora mediante un script componente ejecutaremos un script que nos permitirá seleccionar y concatenar las primeras 2 plantas considerando que el sort anterior debe de mostrarnos comenzando la de mayor cantidad hasta la última

Parte relevante de el código:

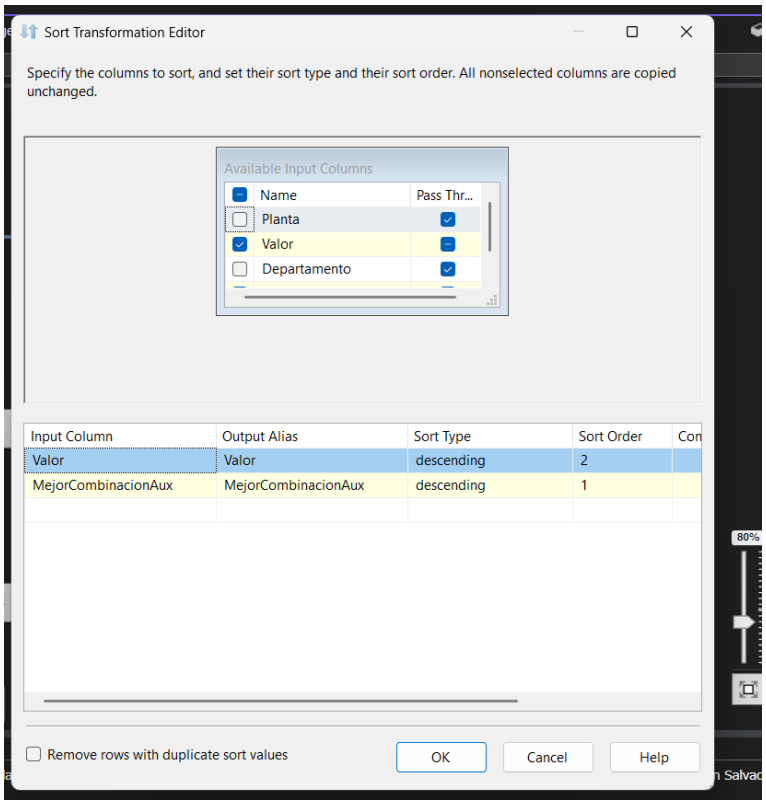
```
/// <param name="Row">The row that is currently passing through the component</param>
2 references
public override void Input0_ProcessInputRow(Input0Buffer Row)
{
    if (contador < 2)
    {
        combinacion += (contador == 0 ? "" : " y ") + Row.Planta;
        Row.MejorCombinacionAux = combinacion;
        contador++;

    }
    else
    {
        combinacion = "";
    }

}
}
```

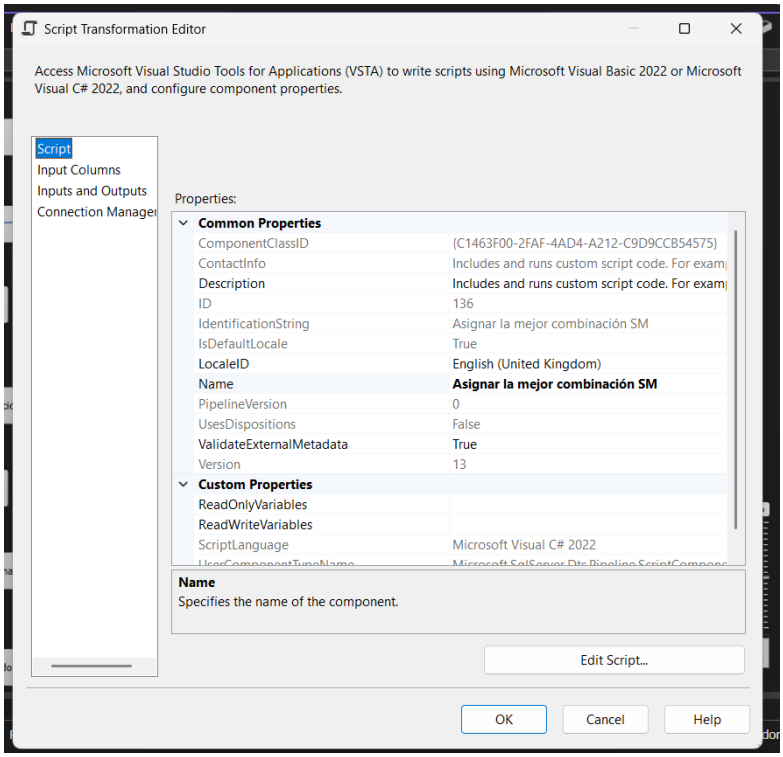
Inicializamos un contador a 0 y cuando procese la primera fila será el Row.Planta lo que nos dará el primer nombre, al momento de procesar la segunda lo concatenaremos con “y” y asignaremos la siguiente planta. Esto se realizará siempre y cuando el contador sea menor que 2 por lo que en otras palabras utilizará solo las primeras 2 filas, y todas para todas las demás las dejará como vacío

Sort



Ejecutamos nuevamente un sort para reordenar las primeras 2 columnas y así poder obtener en primera posición las plantas guardadas en el script anterior

Script component



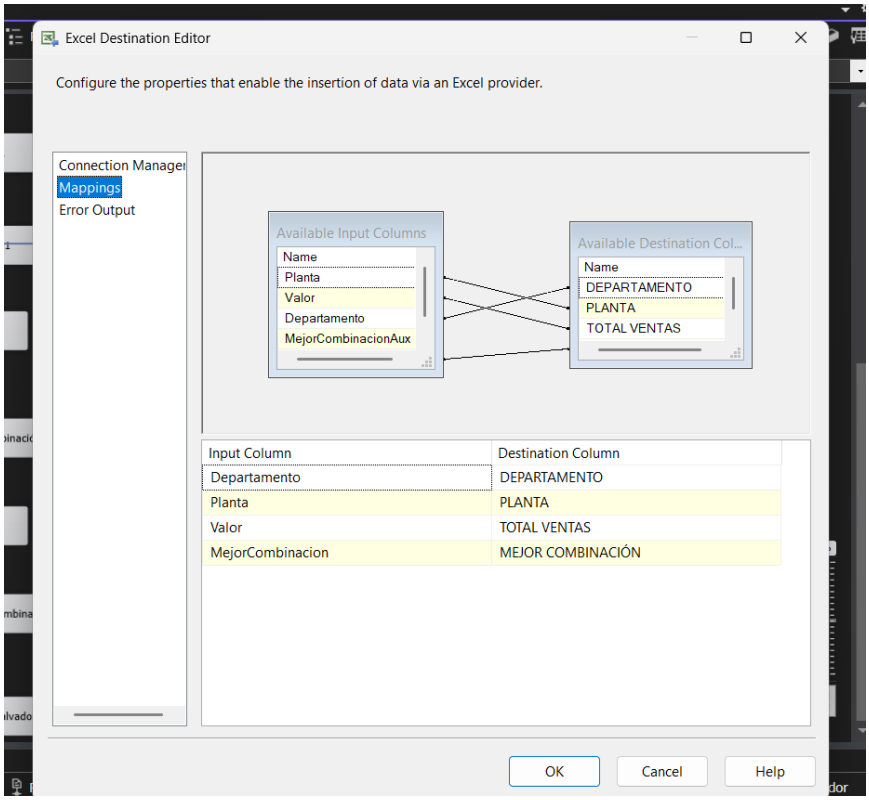
Mediante otro script component procedemos a asignar la primera fila de la columna de “MejorCombinacionAux” a la columna final que se llamará “MejorCombinación”

Parte relevante:

```
    /// <param name="Row">The row that is currently passing through the component
    public override void Input0_ProcessInputRow(Input0Buffer Row)
    {
        if (contador == 0)
        {
            Row.MejorCombinacion = Row.MejorCombinacionAux;
            contador++;
        }
        else
        {
            Row.MejorCombinacion = ""; // Resto de filas vacías
        }
    }
}
```

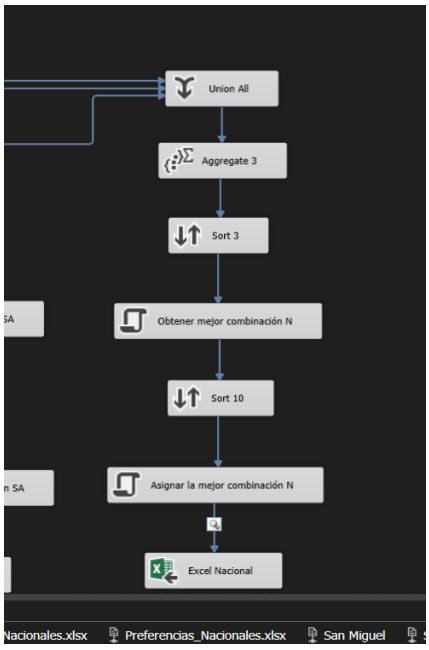
Cuando el contador previamente establecido sea igual a 0 me indica que se está procesando la primera fila, por lo que procedo con la asignación de la columna aux a la final, de lo contrario todas quedarán de manera vacía

Excel File Destination



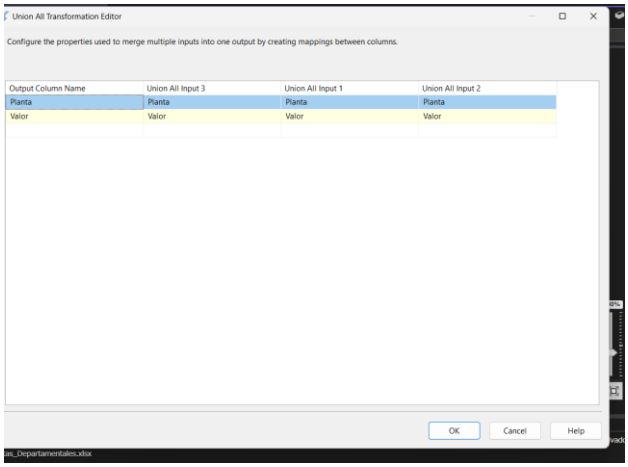
Finalmente mediante un archivo Excel de destino realizamos la inserción de los datos de manera limpia y ordenada conforme a los requerimientos previamente establecidos

Nivel nacional



Para el flujo de datos a nivel nacional ocupamos los multicast mencionados anteriormente y será exactamente la misma lógica

Unnion All

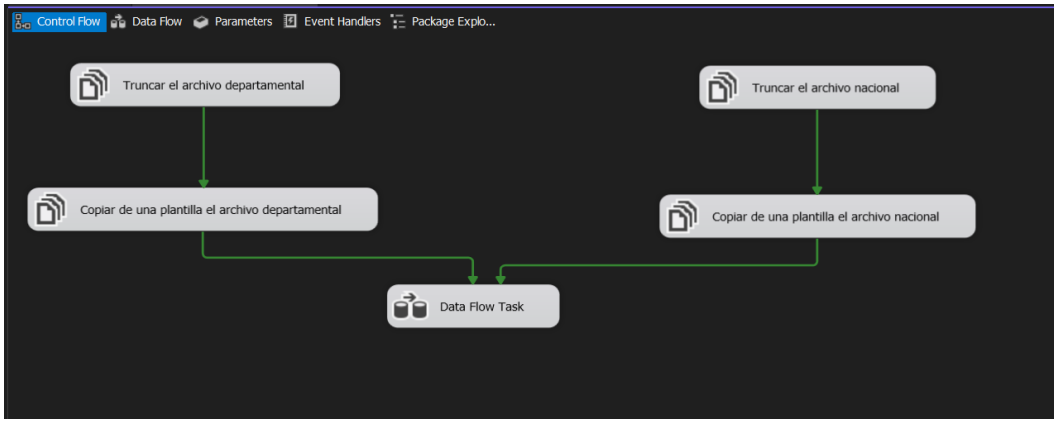


Esta herramienta nos permite añadir nuevas filas a las mismas columnas y de esta manera obtener estadísticas a nivel nacional

Todo este flujo de datos y herramientas previamente establecidas ser repetirá para los 3 departamentos, únicamente cambiando la columna departamento y los resultados finales

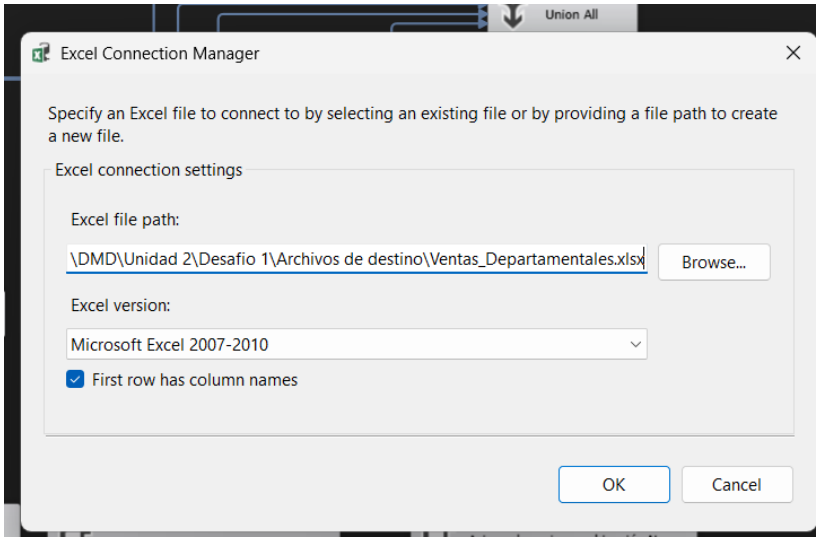
Configuraciones Excel

Para la buena práctica y ejecución de el flujo de datos se realizó un borrado de cualquier dato previamente existente en Excel

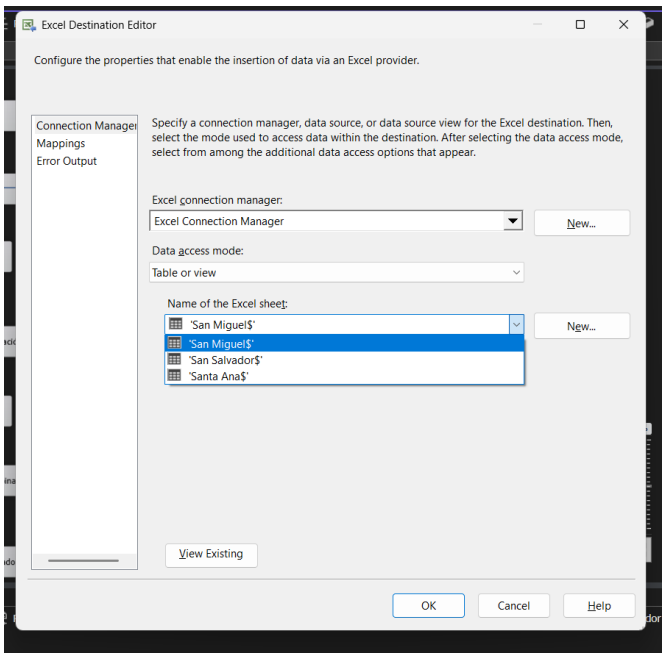


Y posteriormente una copia de un archivo plantilla hacia el original que será en donde se trabajarán

Generamos la conexión mediante el asistente de Excel file destination



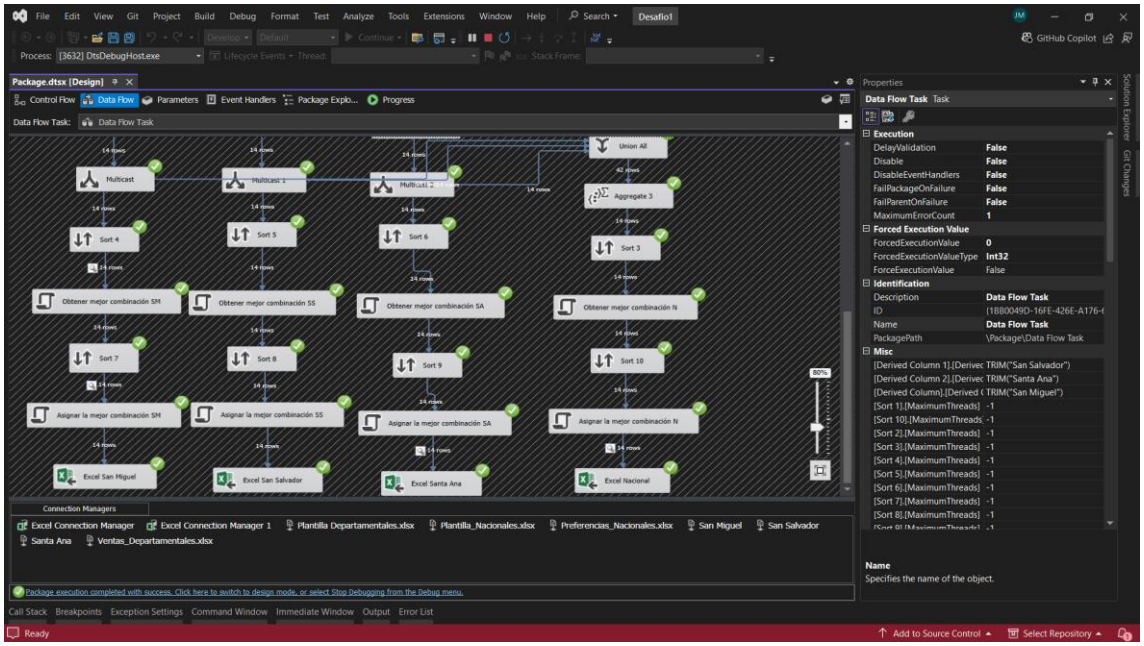
Hacia el archivo “Ventas_Departamentales.xlsl”



El cuál poseerá 3 distintas hojas para cada departamento y las columnas previamente establecidas en los puntos anteriores

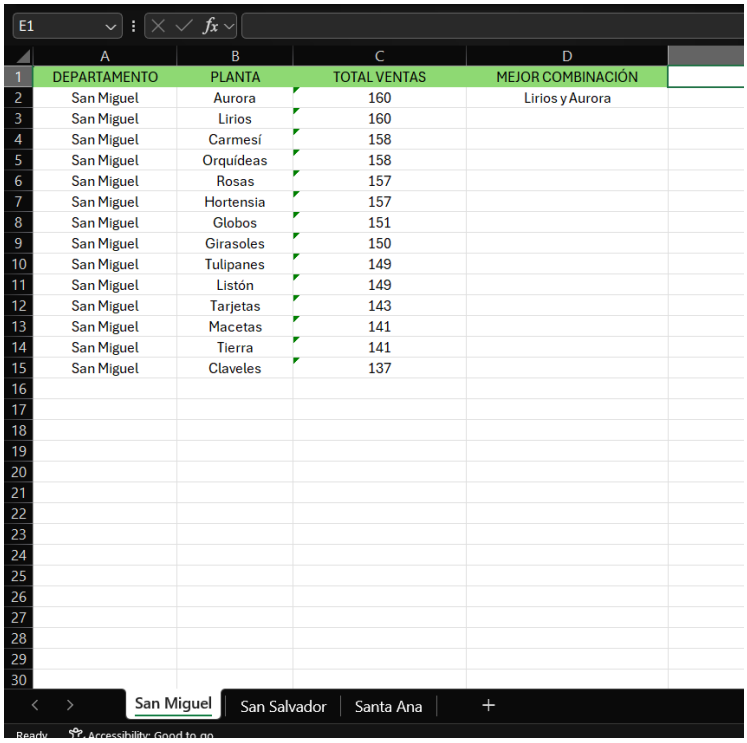
Misma configuración para el Excel a nivel nacional, cambiando únicamente la ruta de el archivo ahora llamado “Preferencias_Nacionales.xlsx”

Ejecución del flujo



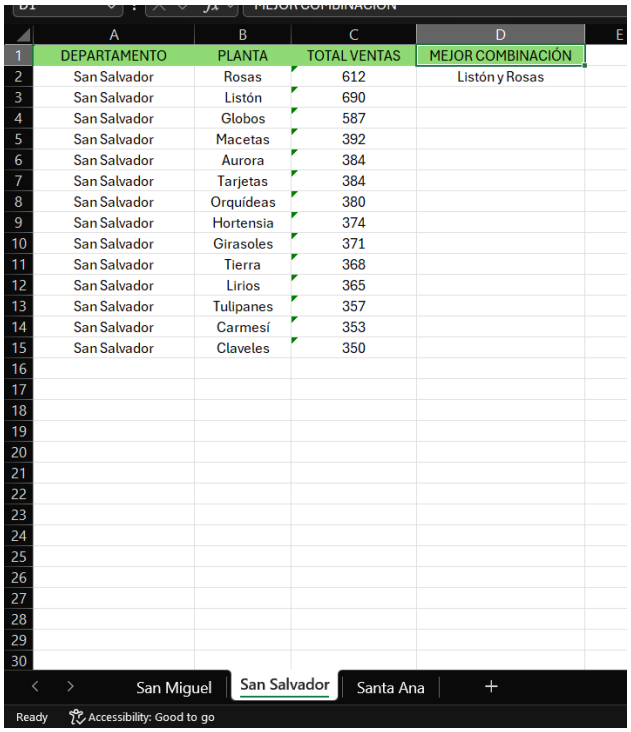
Como podemos observar todas las herramientas que conforman el flujo de datos han sido correctamente ejecutadas

Análisis de resultados



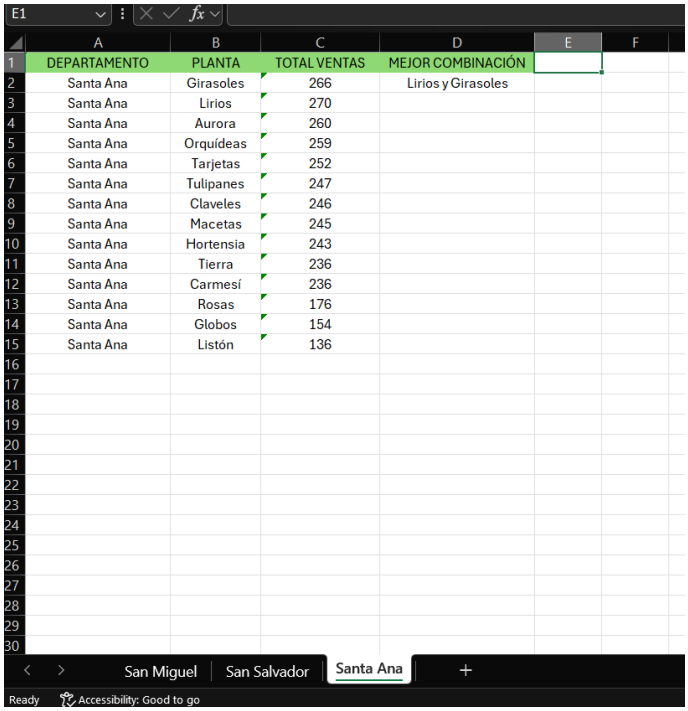
	A	B	C	D
1	DEPARTAMENTO	PLANTA	TOTAL VENTAS	MEJOR COMBINACIÓN
2	San Miguel	Aurora	160	Lirios y Aurora
3	San Miguel	Lirios	160	
4	San Miguel	Carmesí	158	
5	San Miguel	Orquídeas	158	
6	San Miguel	Rosas	157	
7	San Miguel	Hortensia	157	
8	San Miguel	Globos	151	
9	San Miguel	Girasoles	150	
10	San Miguel	Tulipanes	149	
11	San Miguel	Listón	149	
12	San Miguel	Tarjetas	143	
13	San Miguel	Macetas	141	
14	San Miguel	Tierra	141	
15	San Miguel	Claveles	137	

San Miguel



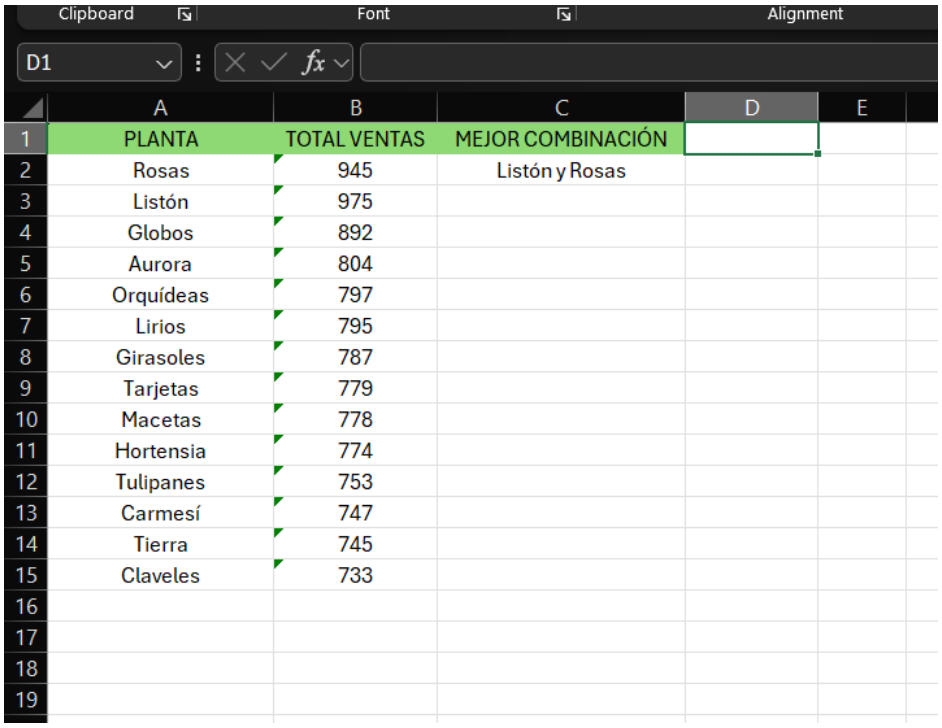
DEPARTAMENTO	PLANTA	TOTAL VENTAS	MEJOR COMBINACIÓN
San Salvador	Rosas	612	Listón y Rosas
San Salvador	Listón	690	
San Salvador	Globos	587	
San Salvador	Macetas	392	
San Salvador	Aurora	384	
San Salvador	Tarjetas	384	
San Salvador	Orquídeas	380	
San Salvador	Hortensia	374	
San Salvador	Girasoles	371	
San Salvador	Tierra	368	
San Salvador	Lirios	365	
San Salvador	Tulipanes	357	
San Salvador	Carmesí	353	
San Salvador	Claveles	350	

San Salvador



DEPARTAMENTO	PLANTA	TOTAL VENTAS	MEJOR COMBINACIÓN
Santa Ana	Girasoles	266	Lirios y Girasoles
Santa Ana	Lirios	270	
Santa Ana	Aurora	260	
Santa Ana	Orquídeas	259	
Santa Ana	Tarjetas	252	
Santa Ana	Tulipanes	247	
Santa Ana	Claveles	246	
Santa Ana	Macetas	245	
Santa Ana	Hortensia	243	
Santa Ana	Tierra	236	
Santa Ana	Carmesí	236	
Santa Ana	Rosas	176	
Santa Ana	Globos	154	
Santa Ana	Listón	136	

Santa Ana



	A	B	C	D	E
1	PLANTA	TOTAL VENTAS	MEJOR COMBINACIÓN		
2	Rosas	945	Listón y Rosas		
3	Listón	975			
4	Globos	892			
5	Aurora	804			
6	Orquídeas	797			
7	Lirios	795			
8	Girasoles	787			
9	Tarjetas	779			
10	Macetas	778			
11	Hortensia	774			
12	Tulipanes	753			
13	Carmesí	747			
14	Tierra	745			
15	Claveles	733			
16					
17					
18					
19					

Nivel Nacional

En base a los resultados podemos analizar diferentes situaciones:

1. San Miguel:
    a. Los Lirios y las Auroras son las plantas con mayor número de ventas y considerando esto, ambas se convierten en la mejor combinación
    b. Los claveles son las plantas con las ventas más bajas en este departamento
    c. La tendencia de ventas indica que la diferencia de unidades compradas entre los primeros 5 productos es mínima por lo que en ámbitos generales las personas de este departamento comparten gustos entre dichos productos y su popularidad en uno especifico podría ser mínima o podría no cambiar en un corto plazo
2. San Salvador
    a. Las Rosas y los Listones son los más vendidos por tanto la mejor combinación
    b. Los claveles son los menos vendidos nuevamente en este departamento
    c. La tendencia de ventas indica una diferencia significativa entre los primeros 3 productos y el resto, dándonos a interpretar que en San Salvador las personas prefieren comprar entre Rosas, Listones o Globos para realizar las combinaciones en base a los gustos que expone esta parte de la población

3. Santa Ana
    a. Parece ser que los Lirios y Las Girasoles han sido las más vendidas y por tanto la mejor combinación
    b. En Santa Ana los listones parecen ser los menos vendidos
    c. Parece haber una diferencia menor entre los primeros 4 productos más vendidos y la tendencia nos indica que en este departamento prefieren la compra de las plantas y no directamente los productos decorativos
4. Nacional
    a. Los Listones y Las Rosas encabezan nuevamente los productos más vendidos y por tanto la mejor combinación
    b. Los claveles que también habían sido los menos vendidos en otros departamentos, a nivel nacional lo vuelven a ser
    c. La tendencia nos puede indicar de manera simple y rápida que el departamento de San Salvador es dónde se encuentra su mayor productividad en ventas al ser la mayor zona donde se realizaron transacciones y en el momento en que se realizó la estadística nacional pudimos apreciar que se mantuvo una tendencia similar a la apreciada en ese departamento