SALESIANOS UNIVERSIDAD DON BOSCO

FACULTAD DE INGENIERÍA

ESCUELA DE COMPUTACIÓN

CICLO 02-2020

“TERCER DESAFÍO PRÁCTICO”

GRUPO DE LABORATORIO:

01

CARRERA:

INGENIERÍA EN CIENCIAS DE LA COMPUTACIÓN.

PRESENTADO POR:

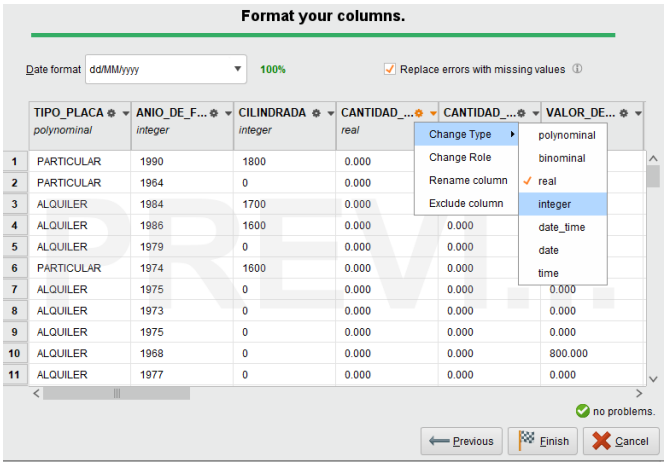
Carnet	Nombre	Apellido
VC190544	Francisco José	Valle Cornejo
AV190086	César Adilson	Ayala Vásquez

DOCENTE:

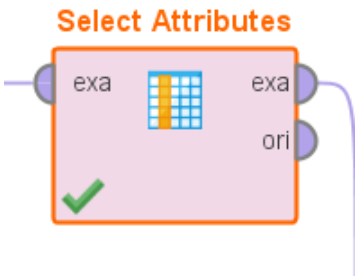
Alexander Alberto Sigüenza Campos

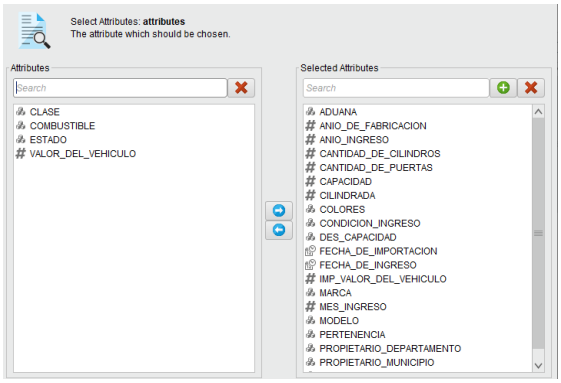
EJERCICIO PARQUE VEHICULAR

Paso 1: Agregar un lector de archivo .csv y poner como origen en archivo “parque_vehicular_datos_abiertos_13NOV2018”, seleccionar que los datos son separados por punto y coma, y transformar su tipo de variables



Paso 2: Insertar la herramienta “Select Atributes” y quitar campos que no son muy relevantes para el análisis, la herramienta de k-means tarda mucho en procesar, y tarda más si mantenemos todos los datos, por otro lado, el árbol de decisión tiene que ser legible y bien presentable, es decir, lo más pequeño posible sin por eso dañar su precisión.

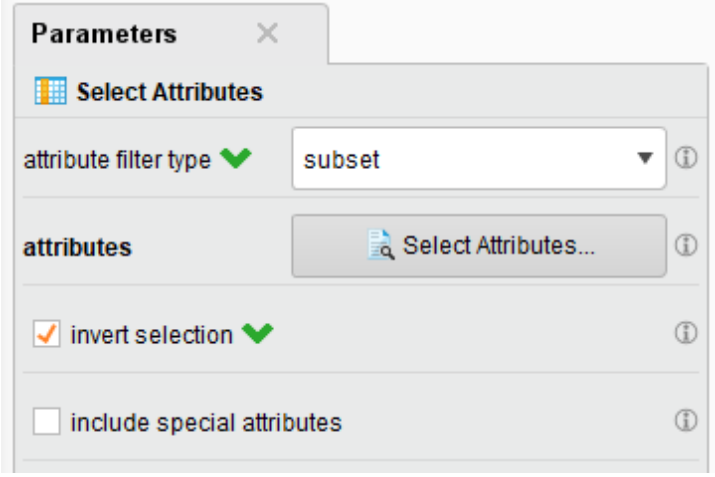




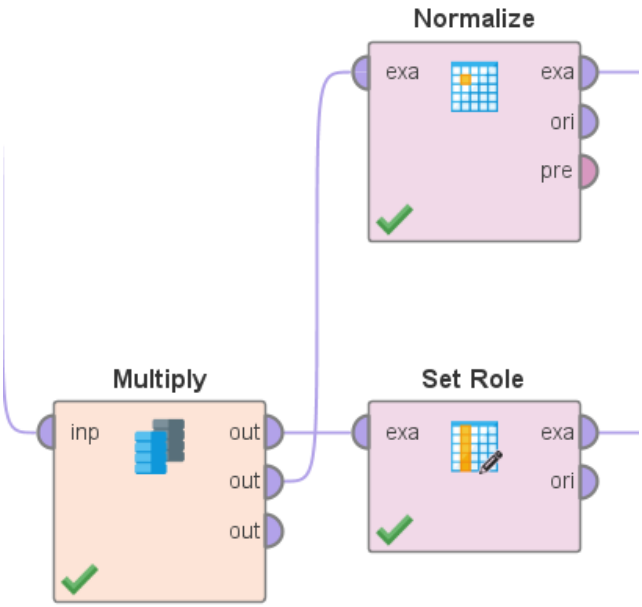
Se selecciona únicamente la clase vehículo, su tipo de combustible, su estado y el valor monetario.

Las demás se obvian por ser de tipo “polinomial” y tener muchas subclasificaciones, por ser valores monetarios referentes a exportaciones, o información extra que vuelve demasiado específico y complejo el árbol de decisiones.

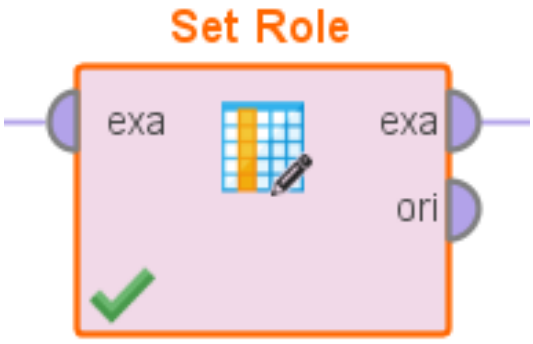
Ya que son columnas que se quieren omitir, se selecciona la propiedad “subset” e “invertir selección”

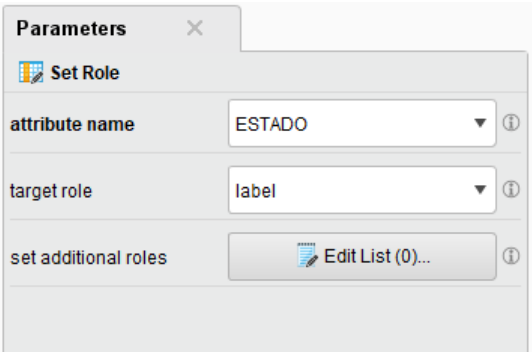


Paso 3: Se pone un elemento “Multiply” este copia su origen de datos (Con los atributos ya omitidos) y lo envía a cuantos elementos se quiera, en este caso aquí es cuando el ejercicio se parte en dos para la parte del K-Means y la parte del Árbol de Decisiones.

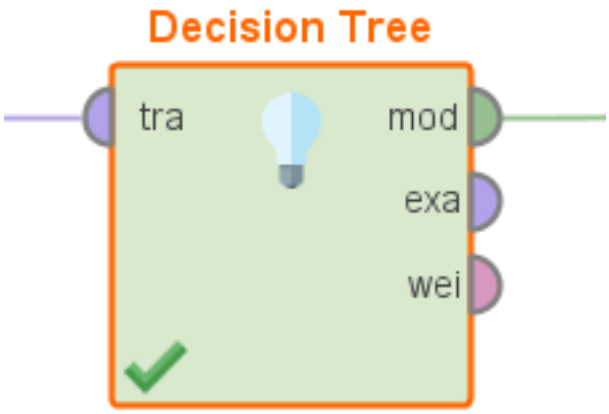


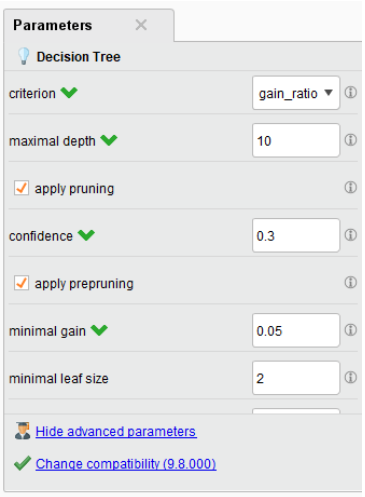
Paso 4 (Árbol de Decisiones): Se necesita un elemento objetivo, de preferencia que sea binomial, en nuestro análisis es la columna “ESTADO” que solo puede ser “ALTA/BAJA”, para seleccionarla como tal, debemos usar la herramienta de “Set Role” para identificarla



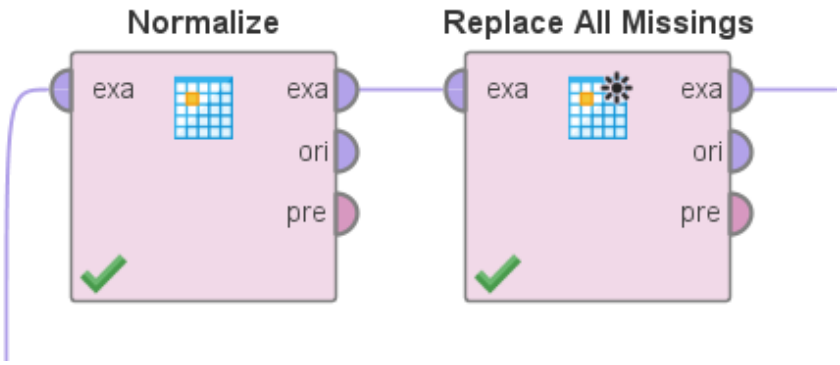


Paso 5 (Árbol de Decisiones): Se agrega el “Decision Tree” y se deja con configuraciones intermedias, sin configuraciones el árbol es de aproximadamente unos 16 niveles, haciendo unos ajustes como limitando su profundidad, subiendo su ganancia minima, y su confianza se obtiene un diagrama más pequeño.

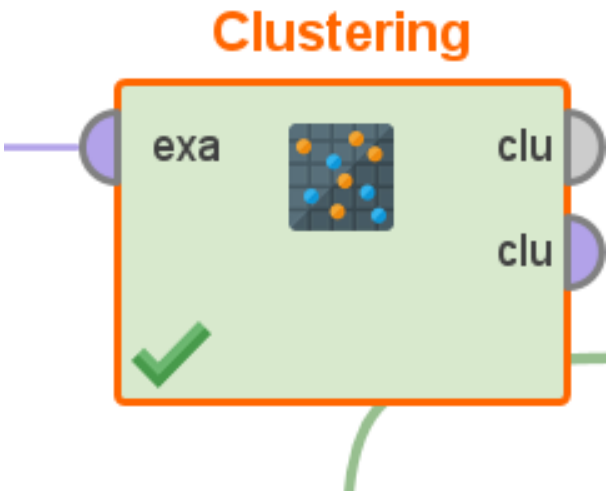


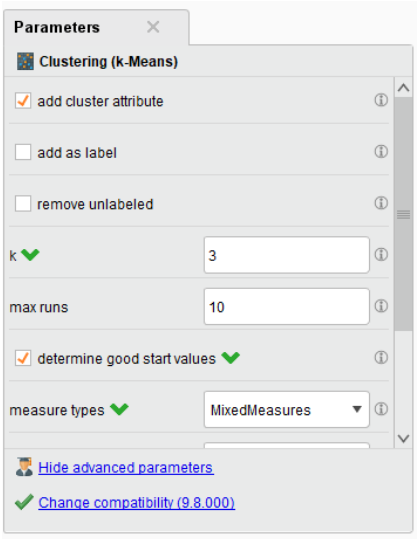


Paso 6 (K-Means): Se agrega un “Normalize” y un “Replace All Missings”, el Normalize convierte todo a una distribución normal para que la desviación estándar sea similar en todos los grupos, a la hora de obtener los datos, es normal que ocurran errores, sobre todo con archivos con una carga superior al millón de filas, estos se reemplazan por “vacío” los cuales el k-means no es capaz de soportar, por lo que agregamos el Reemplazador para que esos “Vacíos” cambien a ser el valor promedio.



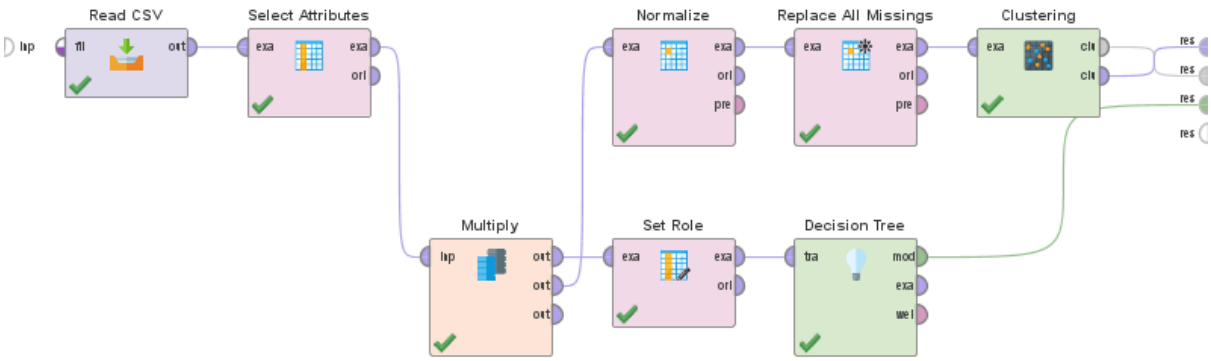
Paso 7 (K-Means): Se agrega un “cluster” por K-Means, probando se llegó a la conclusión que 3 era la cantidad de clusters más óptima, por lo que se selecciona y se da a “Ejecutar” En aproximadamente 30 minutos el proceso habrá finalizado





RESULTADOS:

Diagrama Completo



Arbol de Desicion:

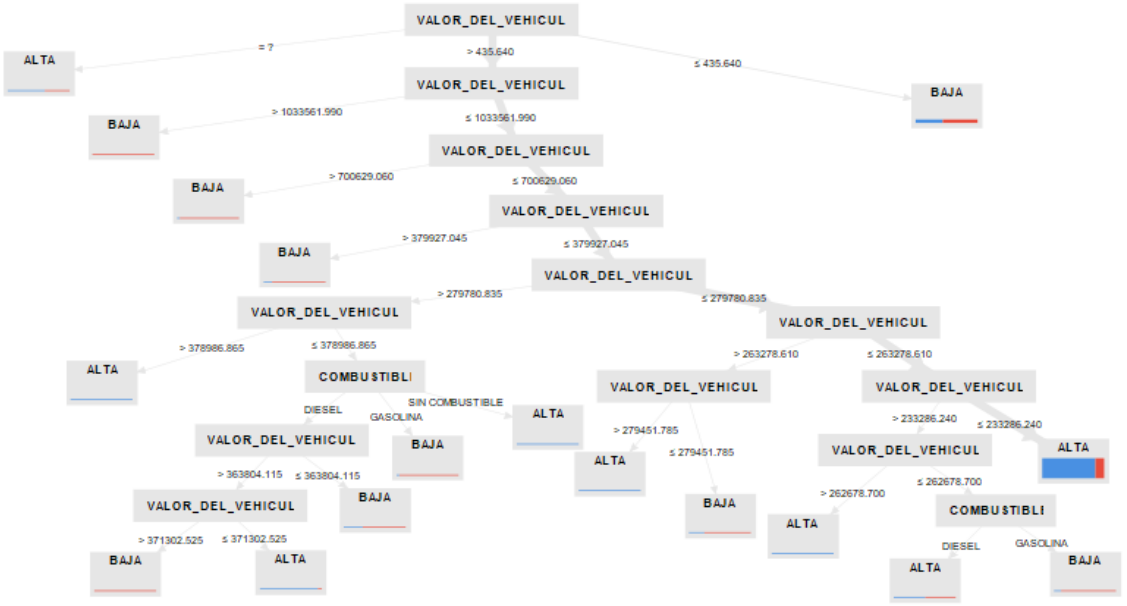
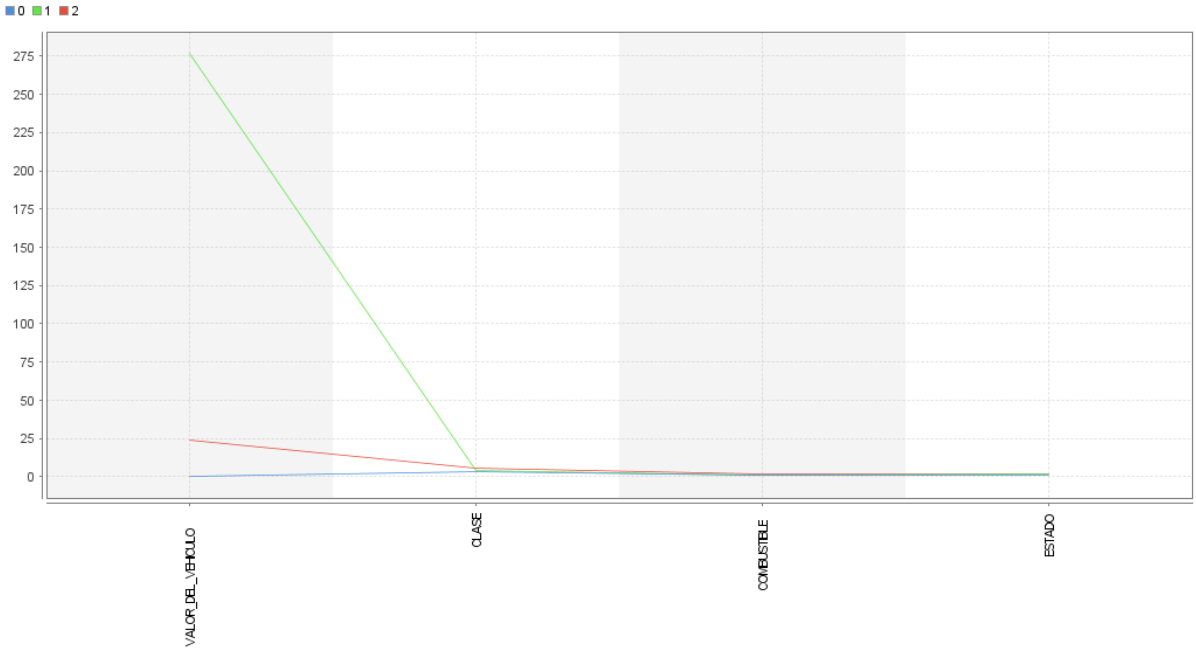
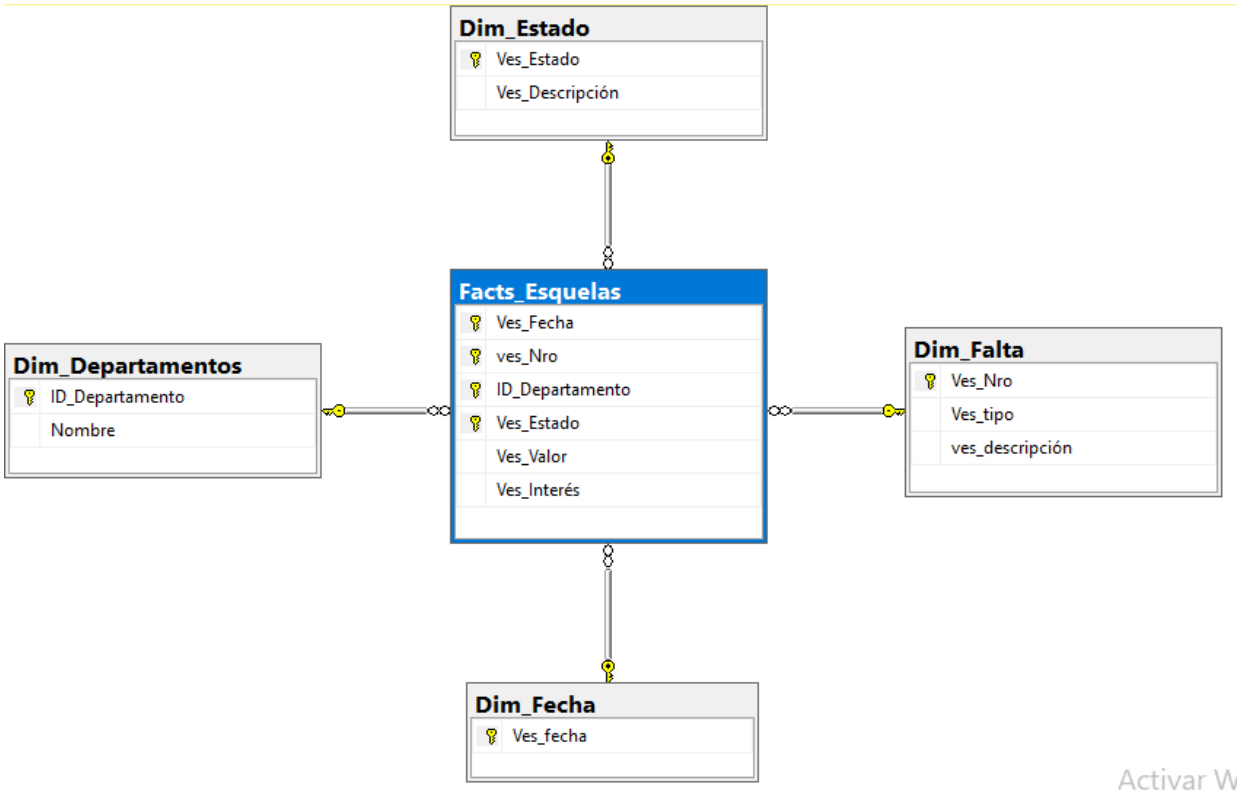


Gráfico K-Means:



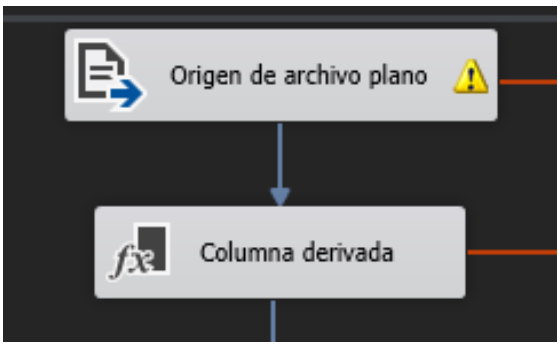
EJERCICIO ESQUELAS

Paso 1: Crearemos una base de datos para realizar un ETL y organizar nuestra base de datos con un modelo dimensional para la realización del cubo. En este caso, obtendremos 4 dimensiones y una tabla hechos. El query de la base de datos está adjunto en github, la base de datos queda de la siguiente manera:

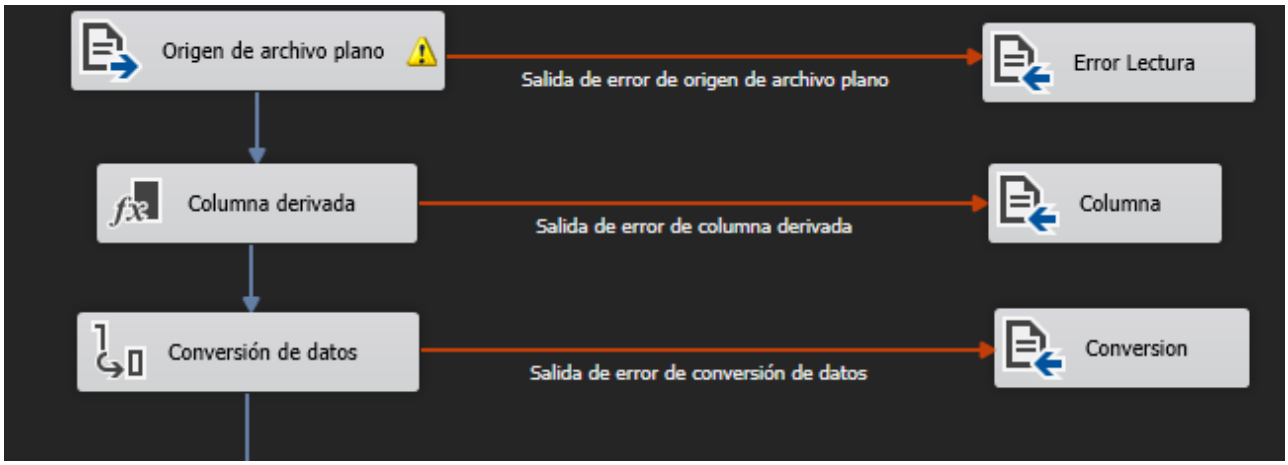


Paso 3: Agregamos una columna derivada para tomar letras de el nombre de los departamentos para obtener el ID, que será conformado por las primeras tres letras del nombre. Esto se realiza con un SUBSTRING, empezando desde el primer carácter y terminando en el tercero.

SUBSTRING([Ves Departamento],1,3)

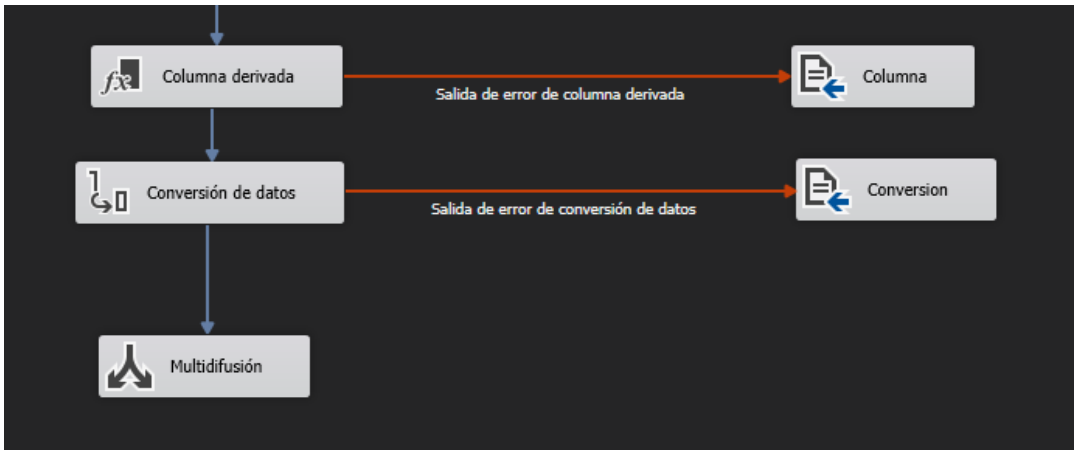


Paso 2: Hacemos una conversión de datos para darle formato a la fecha y también a las cifras que representan dinero, estas son: “Ves_Valor” y “Ves_Interes”.

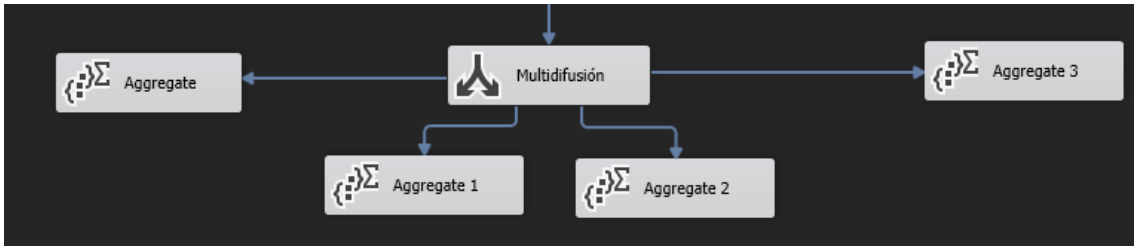


Asignamos a cada componente una salida de error para identificar si algo falla en la ejecución del flujo de datos, así podremos resolver los problemas más fácilmente, si es que estos se llegan a presentar.

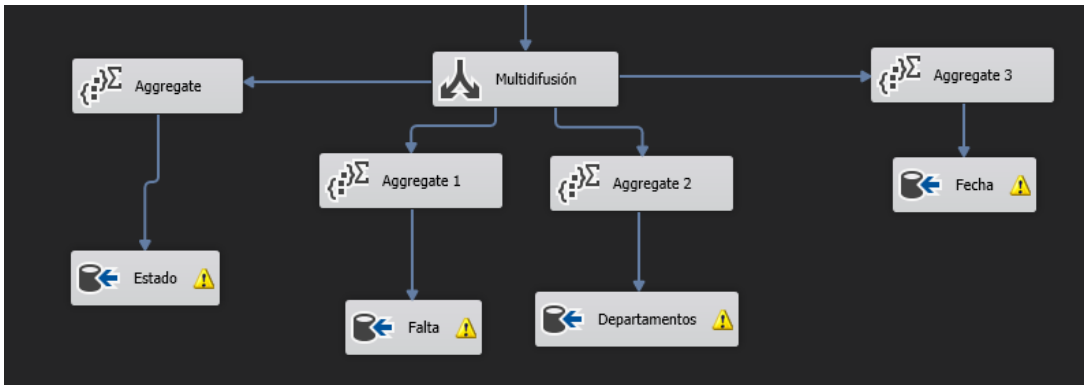
Paso 3: Agregaremos una multidifusión para hacer la división de los datos, así dirigiremos cada a su respectiva tabla en la base de datos.



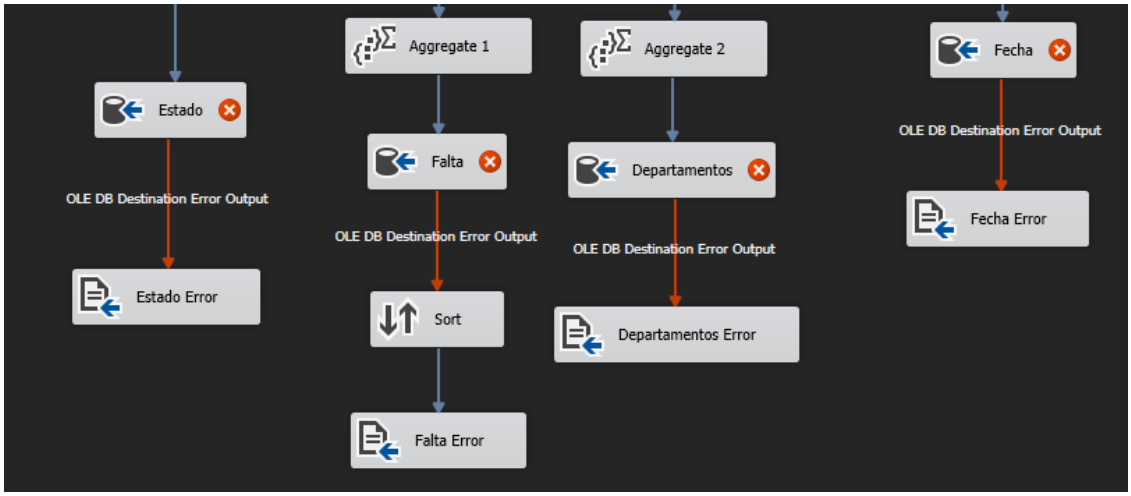
Paso 4: Del multidifusión realizamos 4 salidas, cada una de estas a un agreggate para ordenar cada dato y agruparlos y toda la información esté mejor organizada.



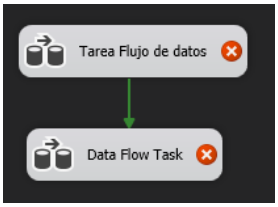
Paso 5: Justo luego de los aggregate, definiremos los destinos, en este caso, es la base de datos antes creada, por ahora solo enviaremos información a las dimensiones, debido a que la tabla hechos hace referencia a las tablas dimensiones.



Paso 6: Por último, definiremos una salida de error para cada destino para identificar fácilmente algún problema, para darle solución de manera eficaz. Incluiremos un sort en Faltas para tener los errores ordenados en caso de que se den.



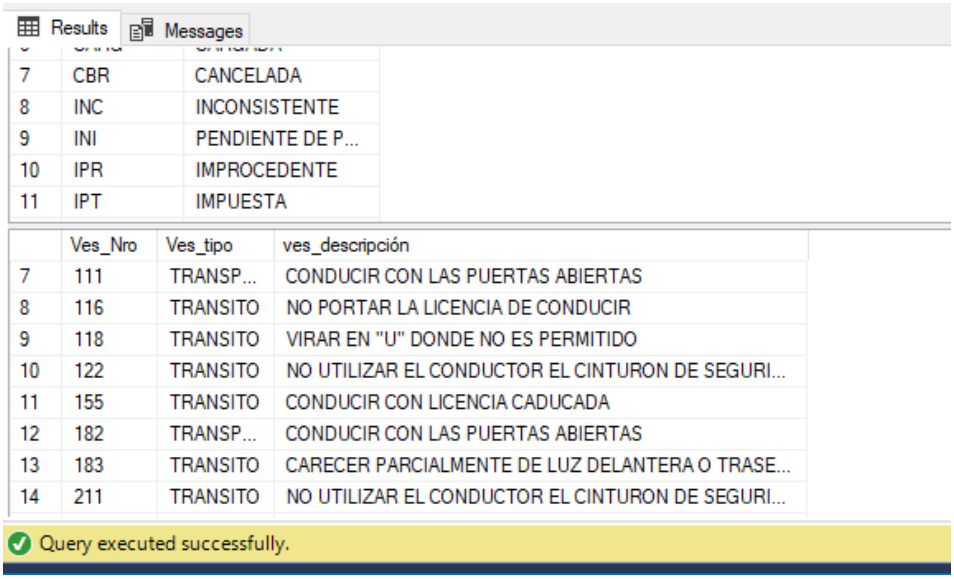
Paso 7: Este mismo proceso se realizará (con unos ligeros cambios) para la tabla de hechos, pero este se hará en otro flujo de datos para evitar problemas en la base de datos con las llaves.



Paso 8: El proceso en este flujo de datos es esencialmente el mismo que el anterior, el origen es el mismo, y la transformación de los datos prácticamente es

igual también, el destino será la tabla de hechos en la base de datos, siempre con salidas de error para identificar los problemas de manera más sencilla.

Paso 9: Ejecutaremos el ETL y luego veremos todos los resultados en la base de datos, ahora está lista para aplicar estrategias de minería de datos, específicamente Cubo OLAP y Power Bi. Los resultados podemos verlos con una consulta SQL.



Results | Messages

7	CBR	CANCELADA
8	INC	INCONSISTENTE
9	INI	PENDIENTE DE P...
10	IPR	IMPROCEDENTE
11	IPT	IMPUESTA

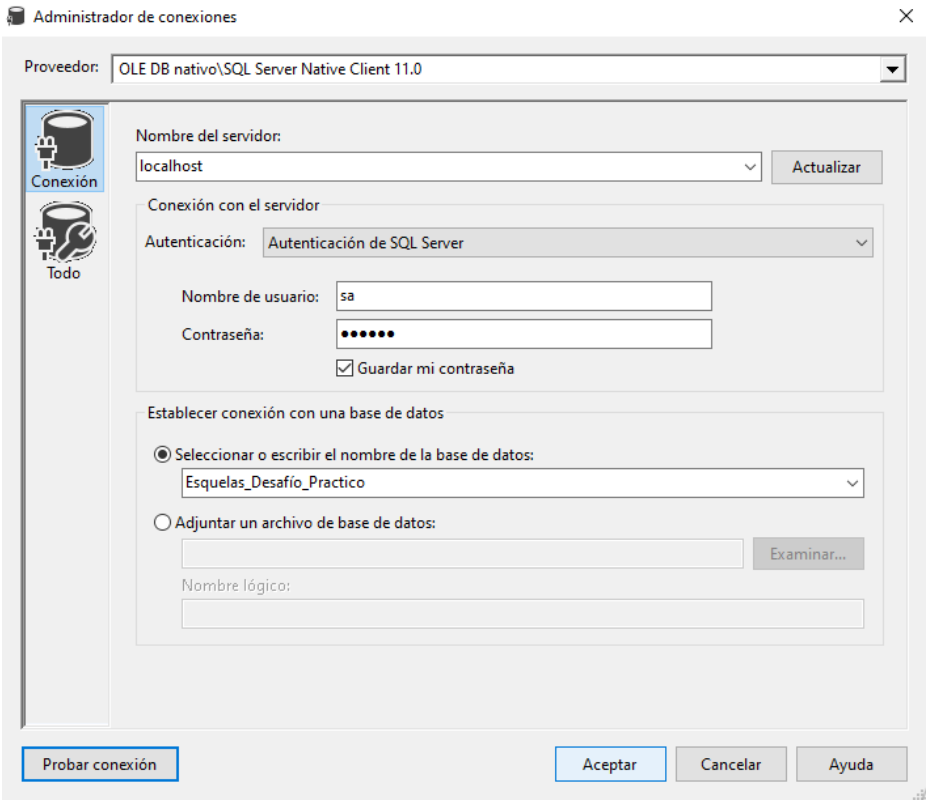
	Ves_Nro	Ves_tipo	ves_descripción
7	111	TRANSP...	CONDUCIR CON LAS PUERTAS ABIERTAS
8	116	TRANSITO	NO PORTAR LA LICENCIA DE CONDUCIR
9	118	TRANSITO	VIRAR EN "U" DONDE NO ES PERMITIDO
10	122	TRANSITO	NO UTILIZAR EL CONDUCTOR EL CINTURON DE SEGURI...
11	155	TRANSITO	CONDUCIR CON LICENCIA CADUCADA
12	182	TRANSP...	CONDUCIR CON LAS PUERTAS ABIERTAS
13	183	TRANSITO	CARECER PARCIALMENTE DE LUZ DELANTERA O TRASE...
14	211	TRANSITO	NO UTILIZAR EL CONDUCTOR EL CINTURON DE SEGURI...

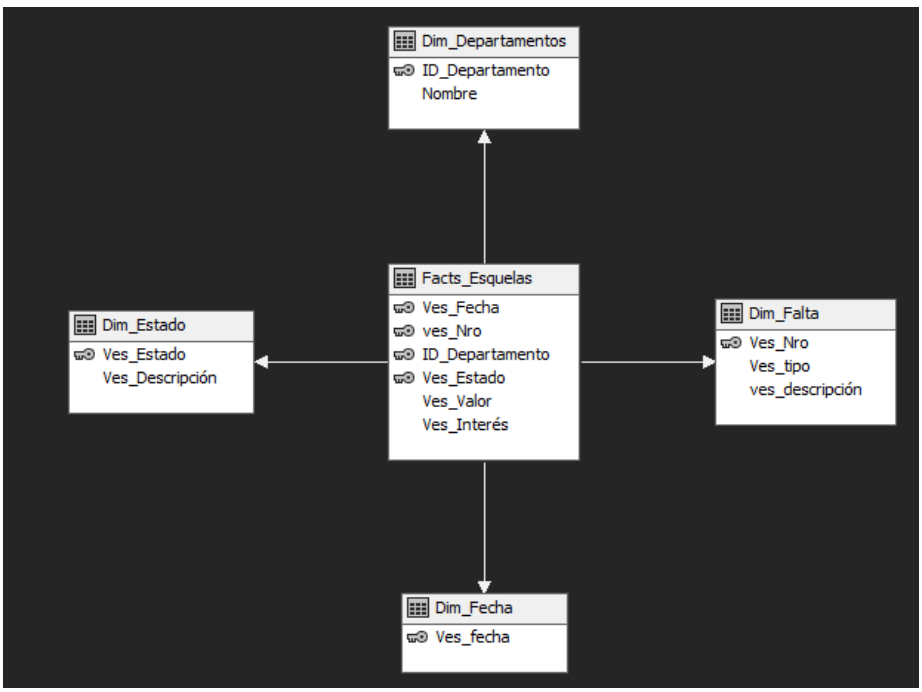
Query executed successfully.

Nota: Se ha creado un .bak para la base de datos, de tal manera que se pueda tener sin necesidad de correr el ETL, este archivo .bak está adjunto el el repositorio

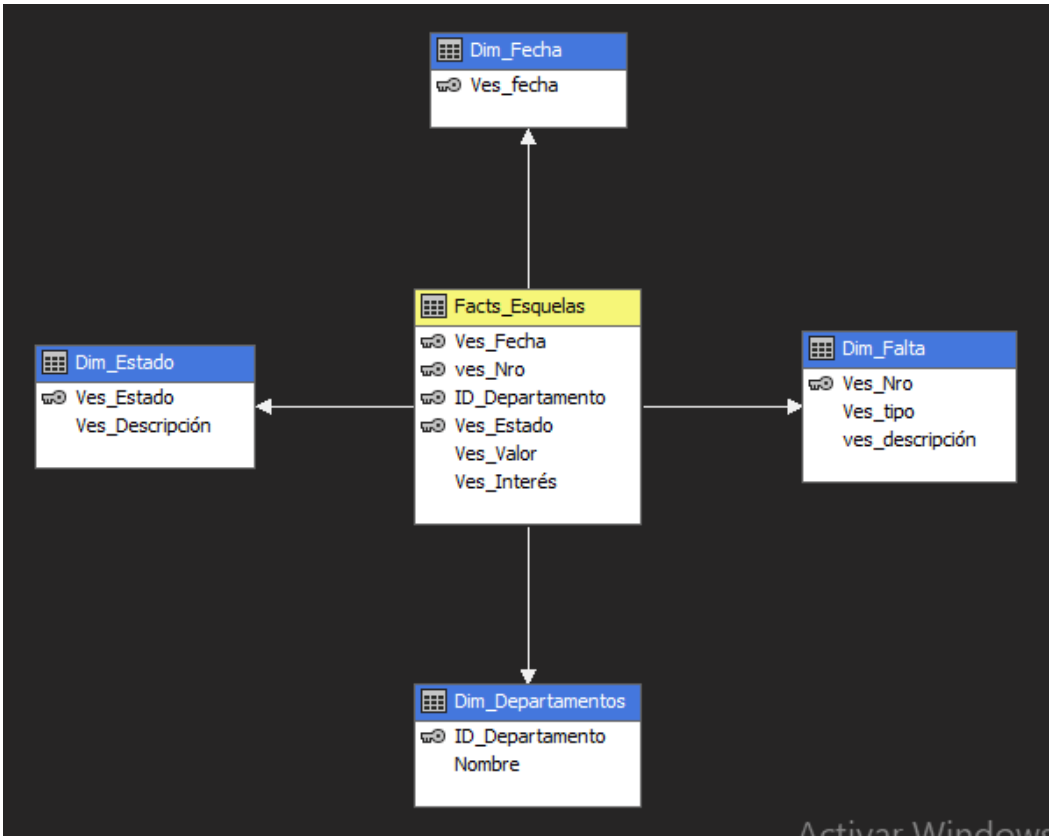
de Github, Esto debido al proceso extremadamente lento de procesar las casi 900,000 filas del documento plano.

Paso 10: Crearemos un nuevo proyecto de minería de datos en visual studio, definiremos el origen con nuestra base de datos y definiremos las vistas para la creación del cubo OLAP.

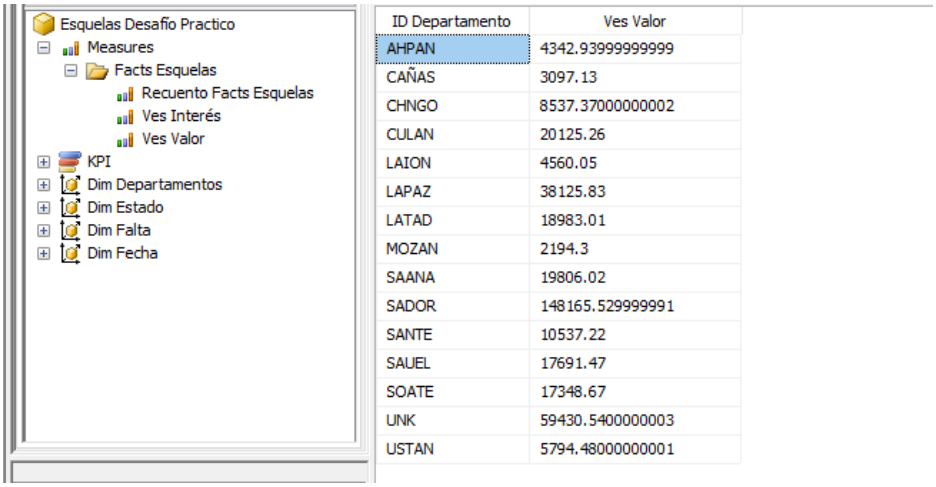




Paso 11: Definimos el cubo con su respectiva dimensión. La tabla hechos será la tabla del grupo de media, las restantes serán las dimensiones.



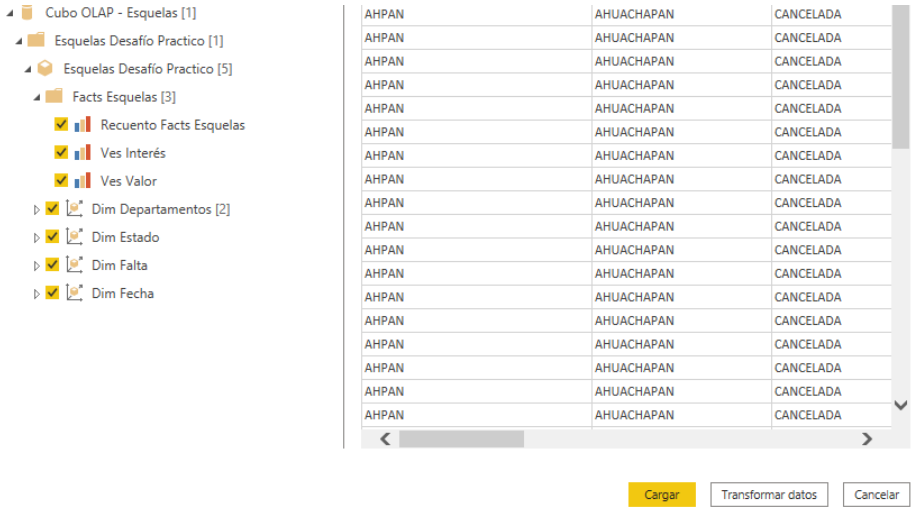
Paso 12: Lo que resta es la implementación del cubo, así que procesaremos el cubo para poder empezar a realizar consultas a través del cubo.



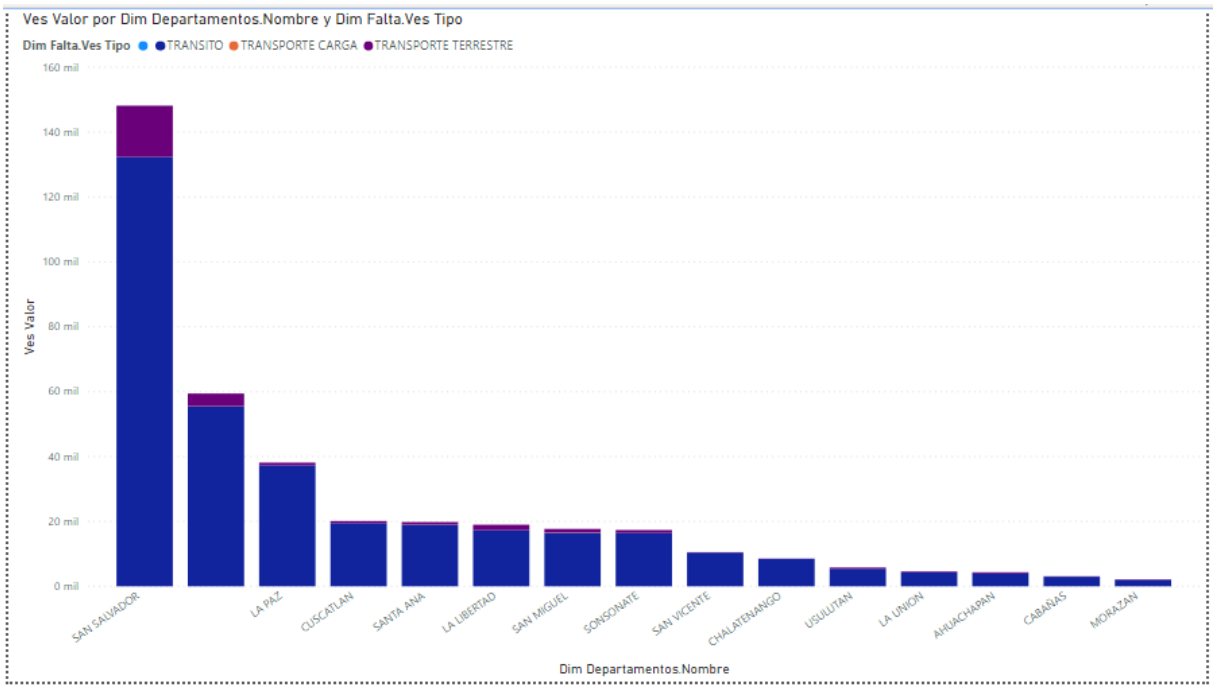
ID Departamento	Ves Valor
AHPAN	4342.93999999999
CAÑAS	3097.13
CHNGO	8537.37000000002
CULAN	20125.26
LAION	4560.05
LAPAZ	38125.83
LATAD	18983.01
MOZAN	2194.3
SAANA	19806.02
SADOR	148165.529999991
SANTE	10537.22
SAUEL	17691.47
SOATE	17348.67
UNK	59430.5400000003
USTAN	5794.48000000001

Esta es una consulta de ejemplo, así que el cubo ya fue implementado correctamente y está listo para su uso. A este cubo se le puede aplicar el servicio de informes o Power Bi para la presentación de toda la información.

Paso 14: Estableceremos la conexión con power services a través de power bi, seleccionaremos sql server analysis services y buscaremos nuestro cubo, seleccionaremos nuestro cubo cargaremos los datos.

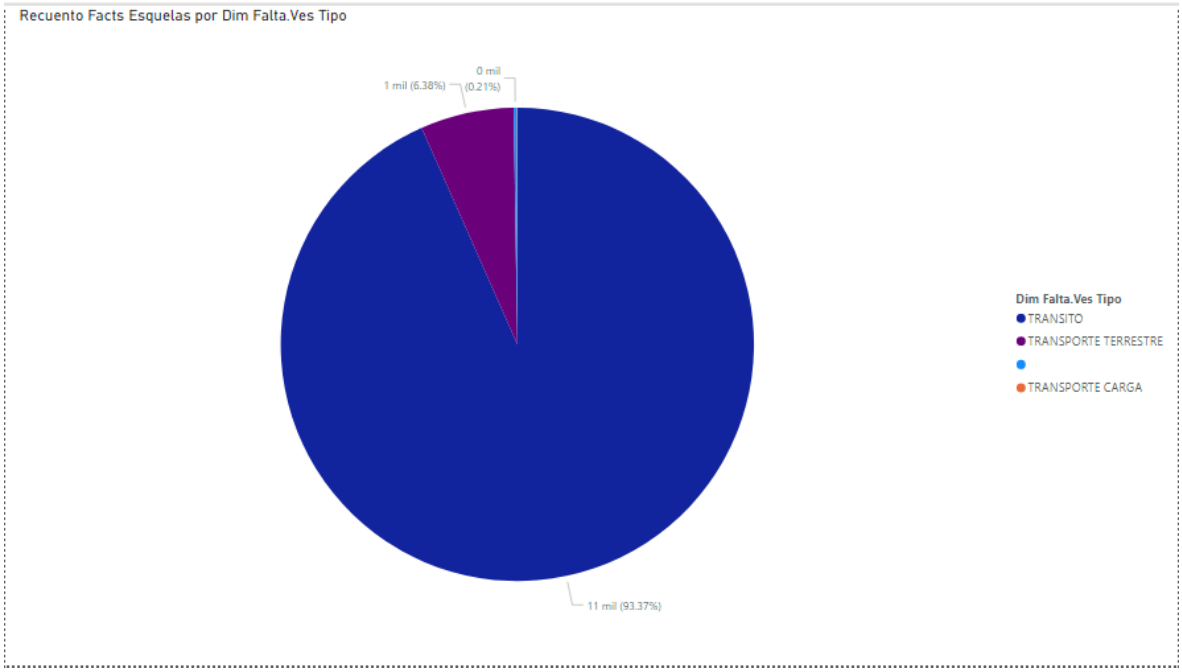


Paso 15: Seleccionaremos gráfico de barras apilado para mostrar nuestra información, en nuestro caso, mostraremos la cifra de dinero que recauda cada infracción, seccionada por su tipo y la dividirá por departamento, así que seleccionaremos la dimensión de departamento, el tipo de falta y de la tabla hechos el valor de la esquela. La gráfica quedará de la siguiente manera:

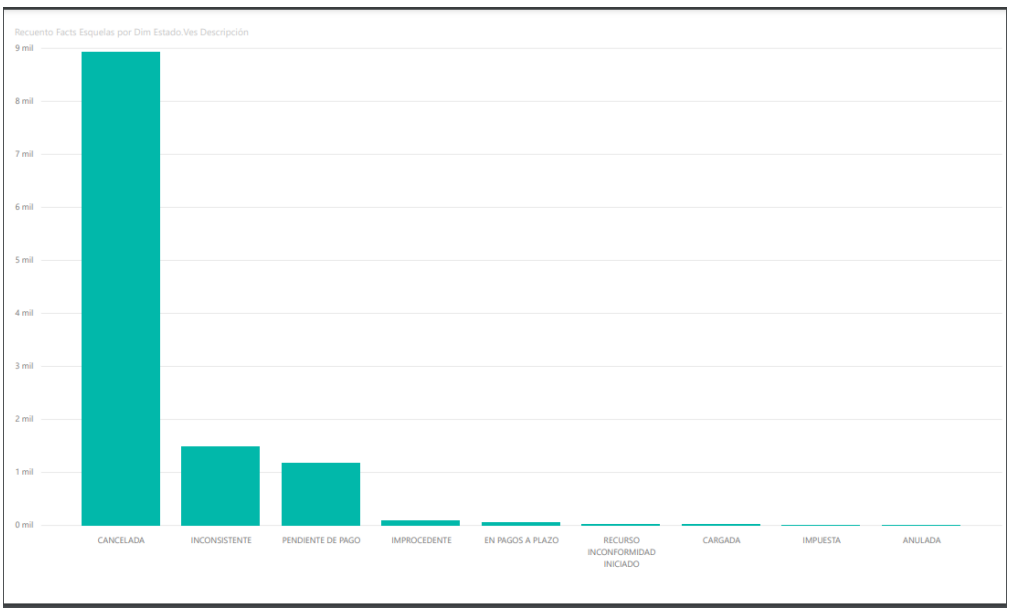


Podemos ver las infracciones divididas por su tipo en distintos colores, y cada columna representa un departamento, en el eje y se visualiza los valores monetarios.

Paso 16: Podemos visualizar el tipo de esquela con más frecuencia dentro del país, para esta vista, seleccionaremos el gráfico circular. Seleccionaremos primero de la dimensión Faltas, el tipo de falta, luego de la tabla hechos seleccionaremos el recuento de todas las faltas que se han hecho, se tendrá el siguiente gráfico.



Paso 17: Por último, podemos tener el recuento de todas las esquelas seccionadas por su estado, ya sea que haya sido cancelada, anulada, pagos en plazo, etc. Para esto seleccionaremos el recuento de esquelas de la tabla hechos, y posteriormente seleccionaremos la descripción del estado de la dimensión de estado. La gráfica se verá de la siguiente manera:



Todas estas gráficas están adjuntas al repositorio Github con formato pdf, esto para analizarlas con mayor calidad visual y mayor resolución, con estas gráficas se tiene una idea los ingresos que se tienen por esquelas, dónde son más usuales las esquelas, y también la cantidad de esquelas que se cancelan, se anulan,