

Datawarehouse y Minería de Datos

CATEDRATICO KAREN MEDRANO

FASE 2

|  |  |  |
| --- | --- | --- |
| N° | ESTUDIANTES | CARNET |
| 1 | MENA CORTEZ, BRYAN ANTONIO | MC211787 |
| 2 | LOPEZ GUTIERREZ, CALEB VERENICE | LG211551 |

1. De diciembre DEL AÑO 2022
2. **OBJETIVOS**

* Analizar la información del Ministerio de Obras Públicas (MOP) mediante las estrategias de minería de datos.
* Proporcionar la información y datos mediante K Means y Power BI.

1. **MARCO TEORICO**

**TEMA: USO DE POWER BI**

Para introducir el archivo CSV del cual queremos extraer los datos, debemos irnos a inicio y seleccionamos obtener datos, en ese apartado seleccionamos CSV.

Graphical user interface, application, Teams

Description automatically generated

Luego nos aparecerá una vista previa de nuestra tabla de datos, ahí podemos hacer cambios de formato o modificaciones, en este caso lo utilizaremos tal y como se nos entregaron los datos.

Graphical user interface, text, application, table

Description automatically generated

Para las consultas de datos que deseamos realizar necesitamos ejecutar diversos scripts, luego de un análisis estos son los que nosotros hemos creado:

Estado = COUNT(Esquelas[Ves\_Estado\_Descripcion]) //Para obtener las cantidades de cada estado

Suma total recaudada = SUM(Esquelas[Ves\_Valor]) // obtener el dinero total de todas las multas

Numero de Esquelas = COUNT(Esquelas[Ves\_Departamento]) //El numero total de multas impuestas

Estos scripts los hemos implementado ya que, si bien Power BI nos implementa ciertas consultas de datos de manera automática, siempre existirán diversas consultas para obtener un “Query” muy específico de algo que deseamos saber.

A la hora de mostrar datos, tenemos la pestaña de diseño, donde seleccionaremos el más adecuado según los datos que deseamos mostrar.

Diagram, icon

Description automatically generated

**Los primeros datos que mostraremos serán las cantidades de esquelas impuestas en cada departamento del país, siendo San Salvador el departamento que más multas acumula.**

Chart, sunburst chart

Description automatically generated

**La siguiente consulta la hemos enfocado en conocer los estados de las más de un millón de multas impuestas, como podemos observar, mas de 800 mil multas ya han sido canceladas, representando la mayoría.**

Graphical user interface

Description automatically generated

**En cuanto al numero de esquelas por años, 2006, 2007 y 2008 encabezan la lista.**

Chart, sunburst chart

Description automatically generated

**En cuanto a los motivos de las multas, podemos observar que las razones son muy diversas, sin embargo, conducir a una velocidad mayor a la permitida es la razón más frecuente de multas.**

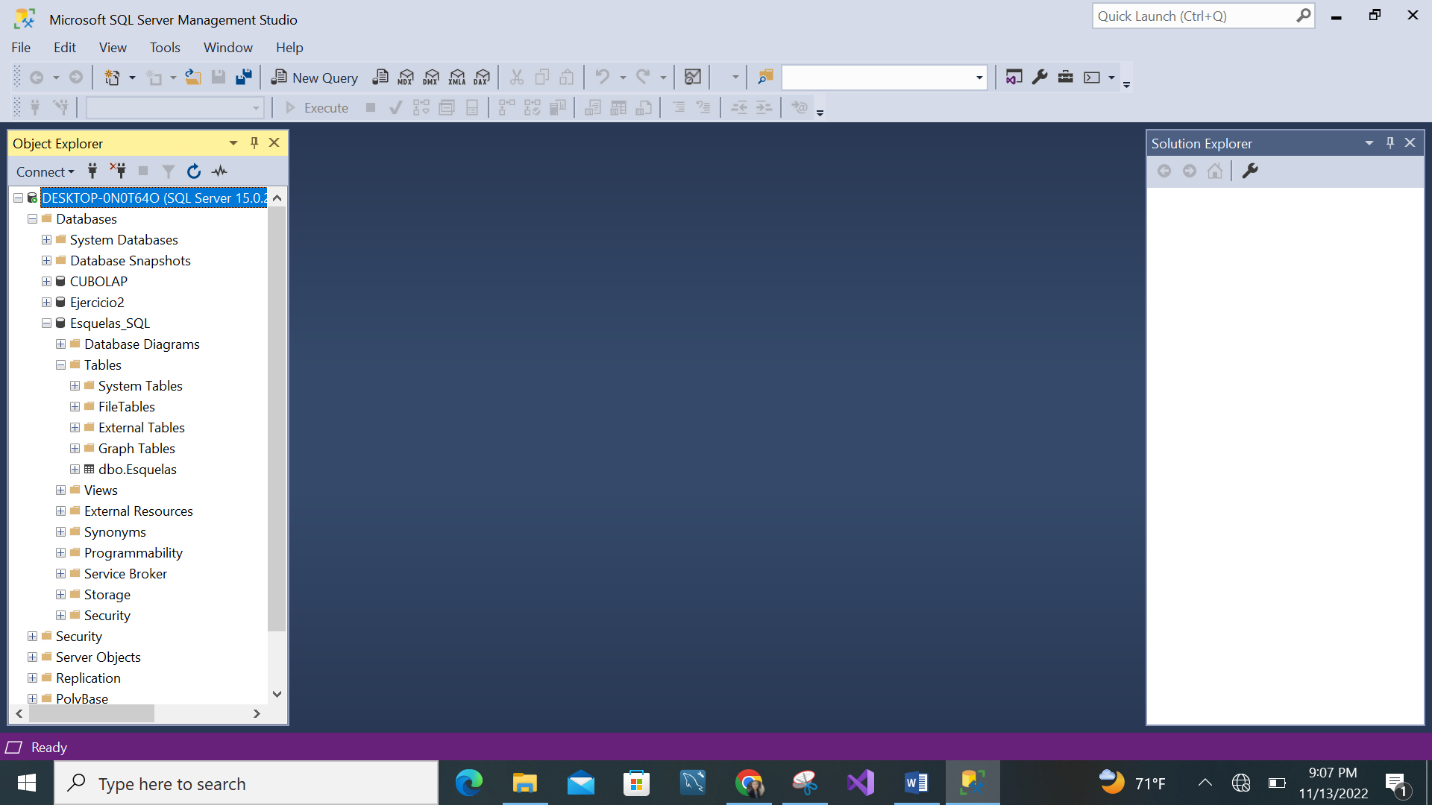
Chart, sunburst chart

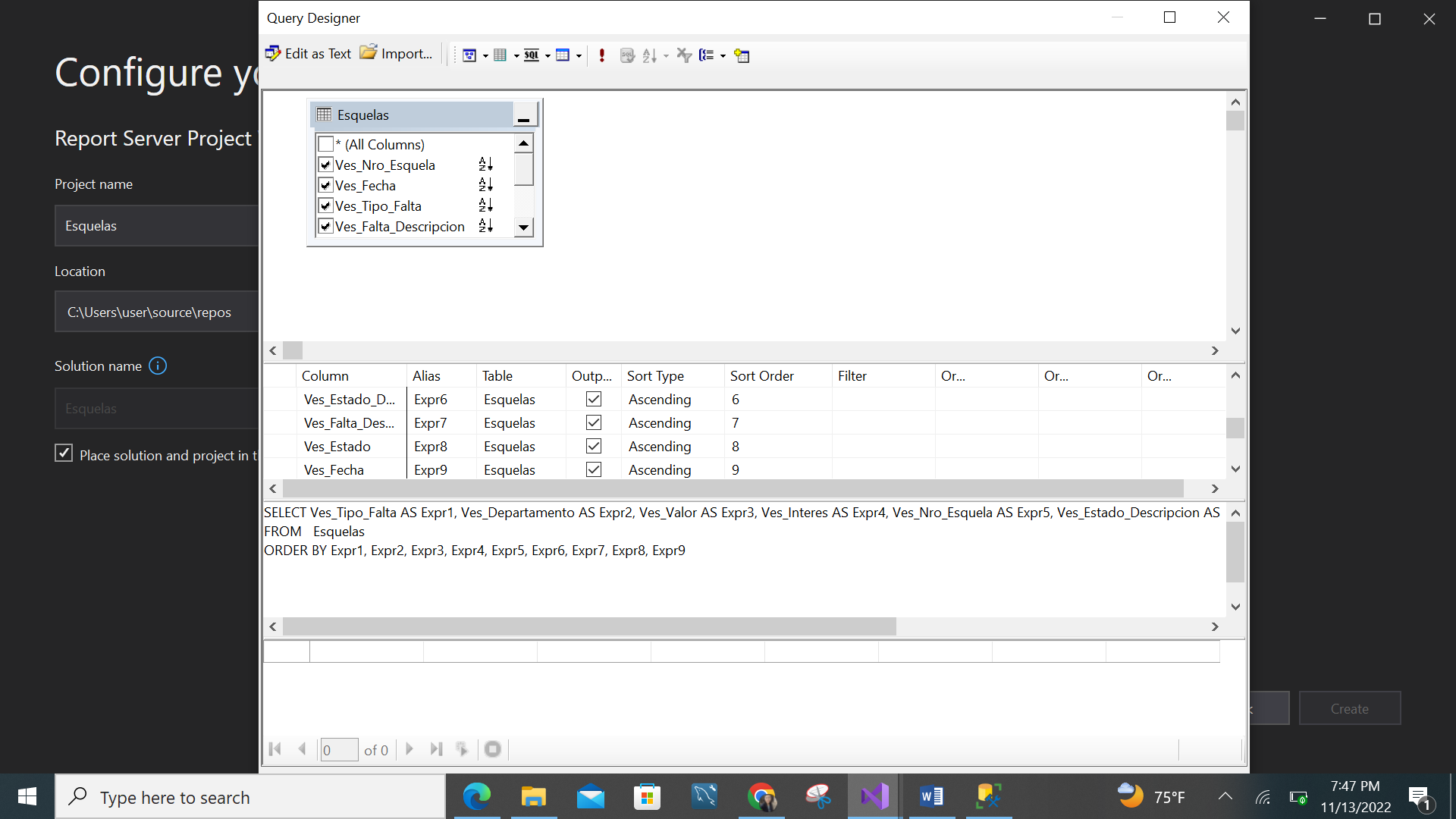
Description automatically generated

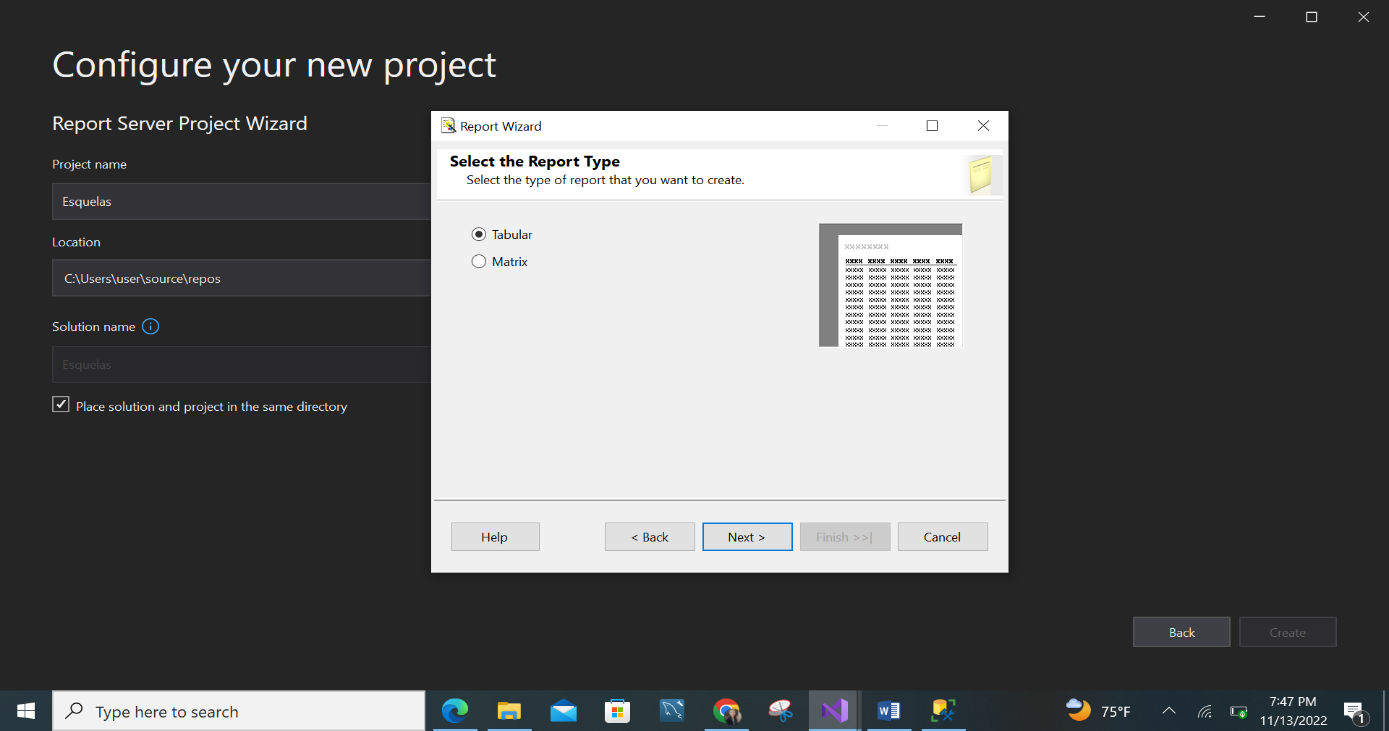
**En cuanto a valor monetario, hemos clasificado el grafico de pastel por los montos en que se puede recibir una multa y vemos que la multa mas aplicada es la de $57.14 y también hemos totalizado el valor de las multas, por lo que obtuvimos que el gobierno ha recaudado más de $34 millones de dólares en multas.**

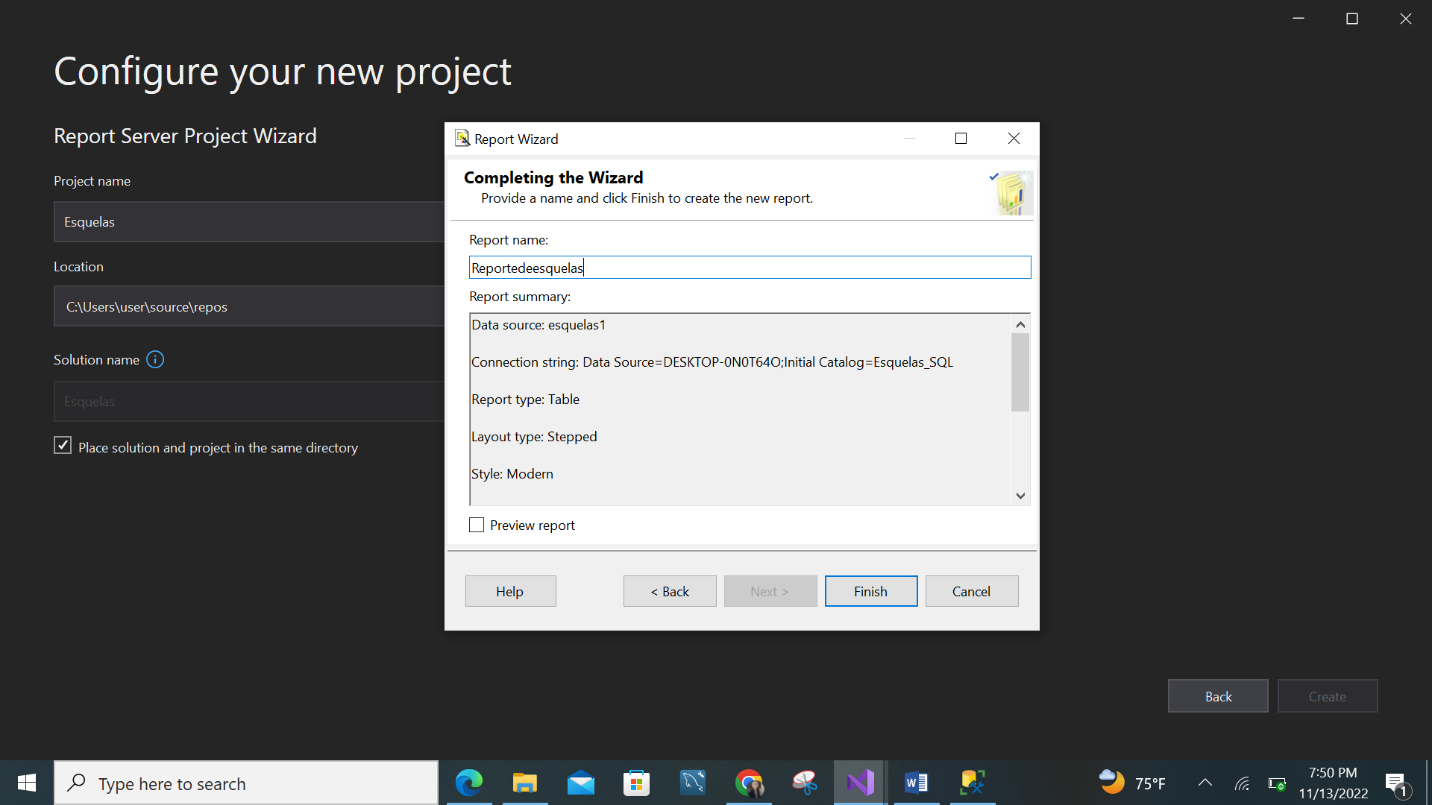
Chart, pie chart

Description automatically generated

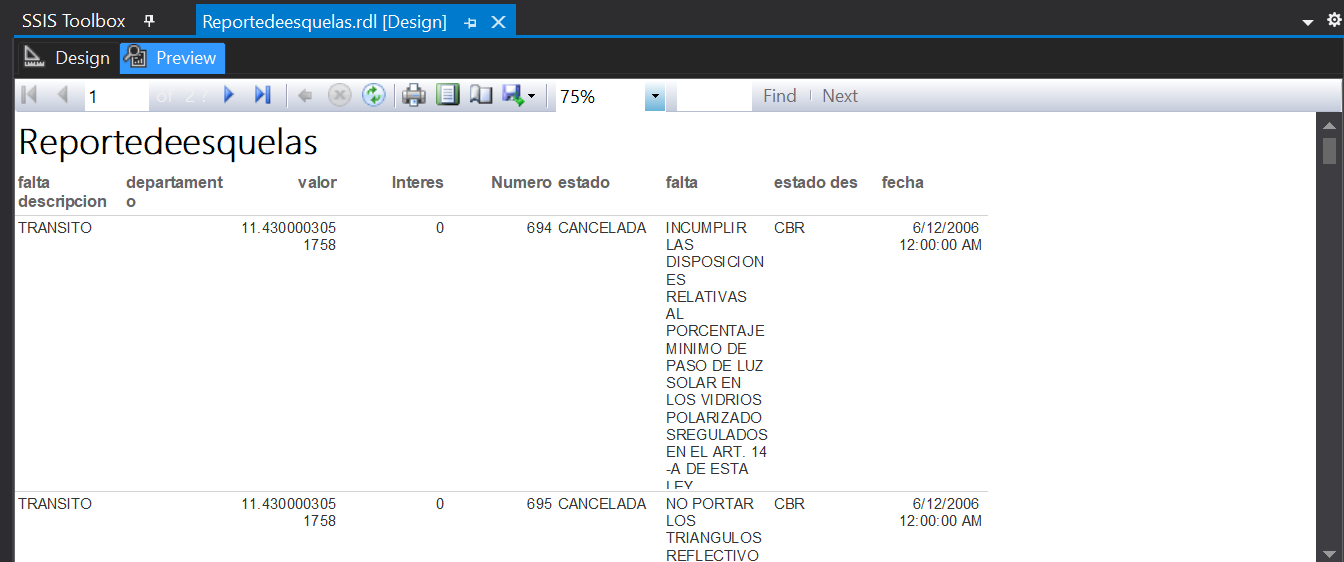
**TEMA: REPORTING SERVICES**

Para poder hacer un mayor análisis y uso de reporting services hemos importado el archivo cvs, el cual nos proporcionara un conocimiento mejor de los datos de Esquelas en el MOP.

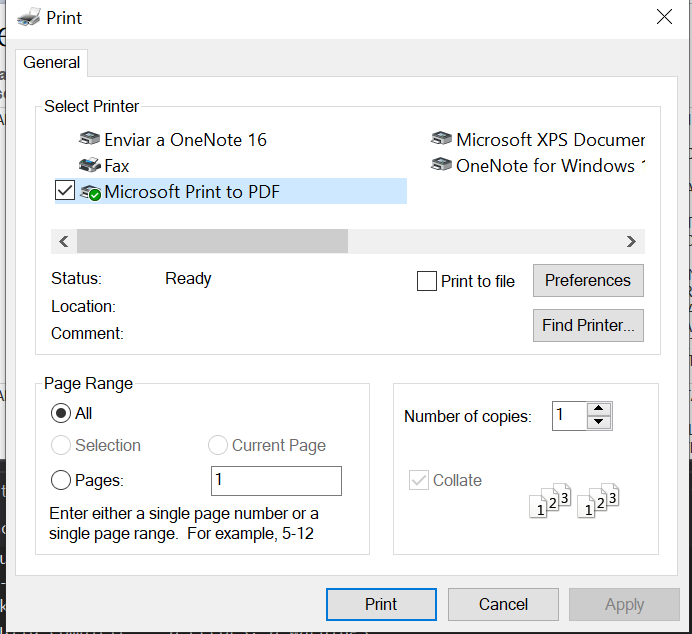
Haciendo uso de visual studio creamos un proyecto en el cual diseñamos las características de nuestro reporte, esto para poder organizar los datos e información que se encontraban en SQL Server.

Hacemos una selección del tipo de reporte, en el que tabulamos los datos.

Completamos de organizar nuestro reporte y finalizamos.



Podemos observar el reporte de esquelas, basado en los datos que se encontraban en nuestro gestor de una manera más gráfica y con alias que ayudan a reconocer los valores internos que se encontraban a modo de dar a conocer la información pertinente.

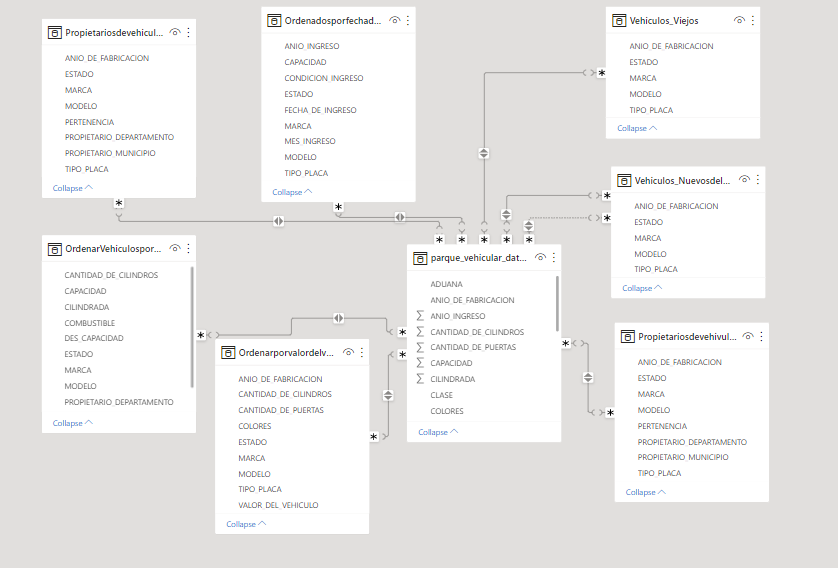


Y obtener un pdf del reporte, de tal forma que los datos se puedan obtener de una forma más ordenada y presentable.

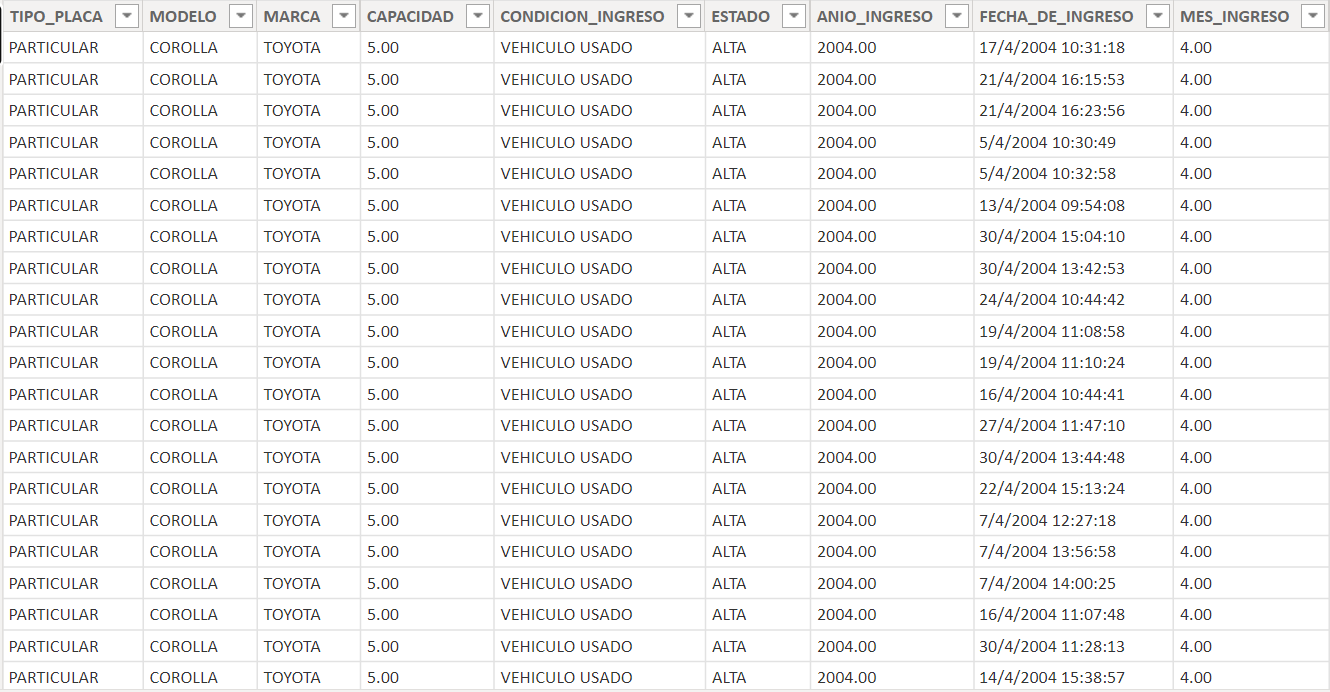
**ANALISIS DE PARQUE VEHICULAR**

**TEMA: CUBO OLAP HACIENDO USO DE POWER BI**

* Crearemos las vistas en SQL Server:
* Luego de cargar el archivo SQL procedemos a hacer el análisis y hacemos la optimización por medio del modelo estrella:



* Organizamos los datos por medio de tablas:



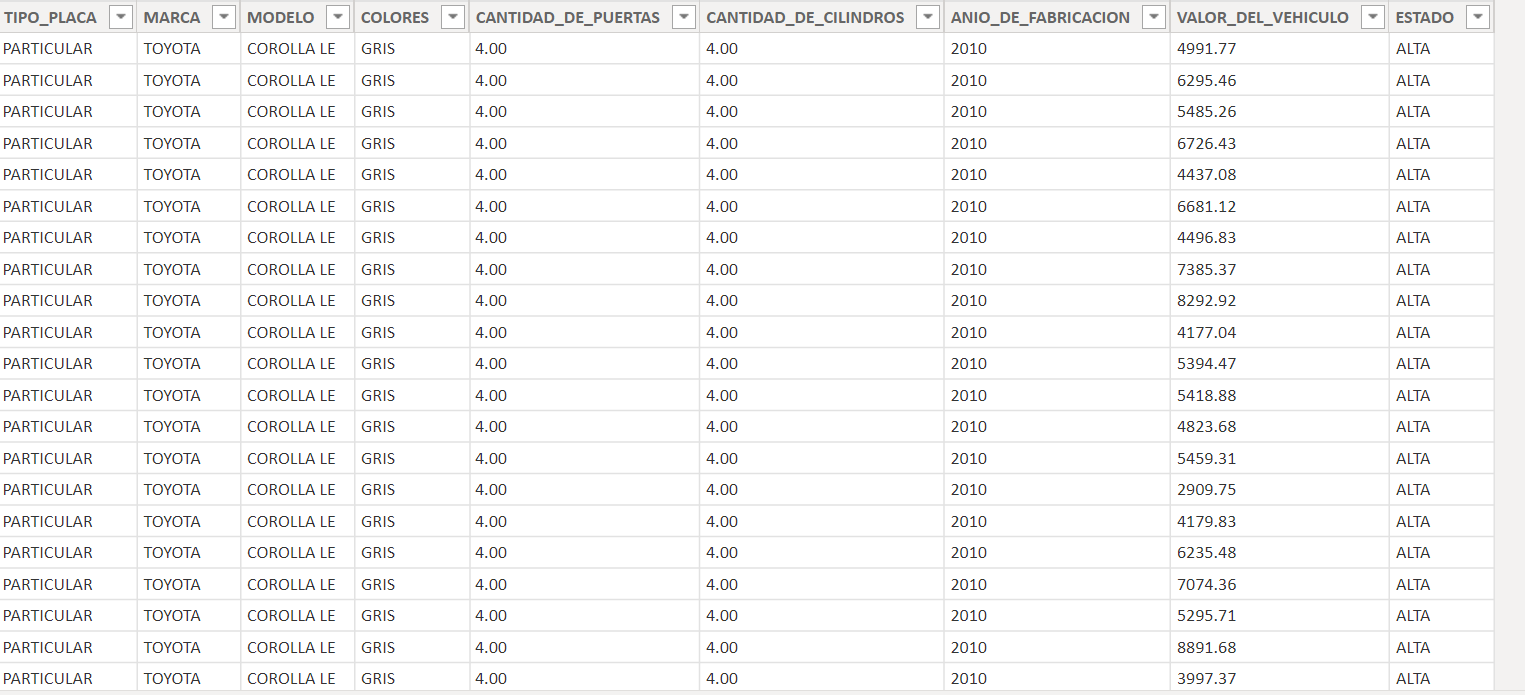
Tabla 1: Ordenados por fecha de ingreso

Tabla 2: Ordenadas por valor de Vehículo

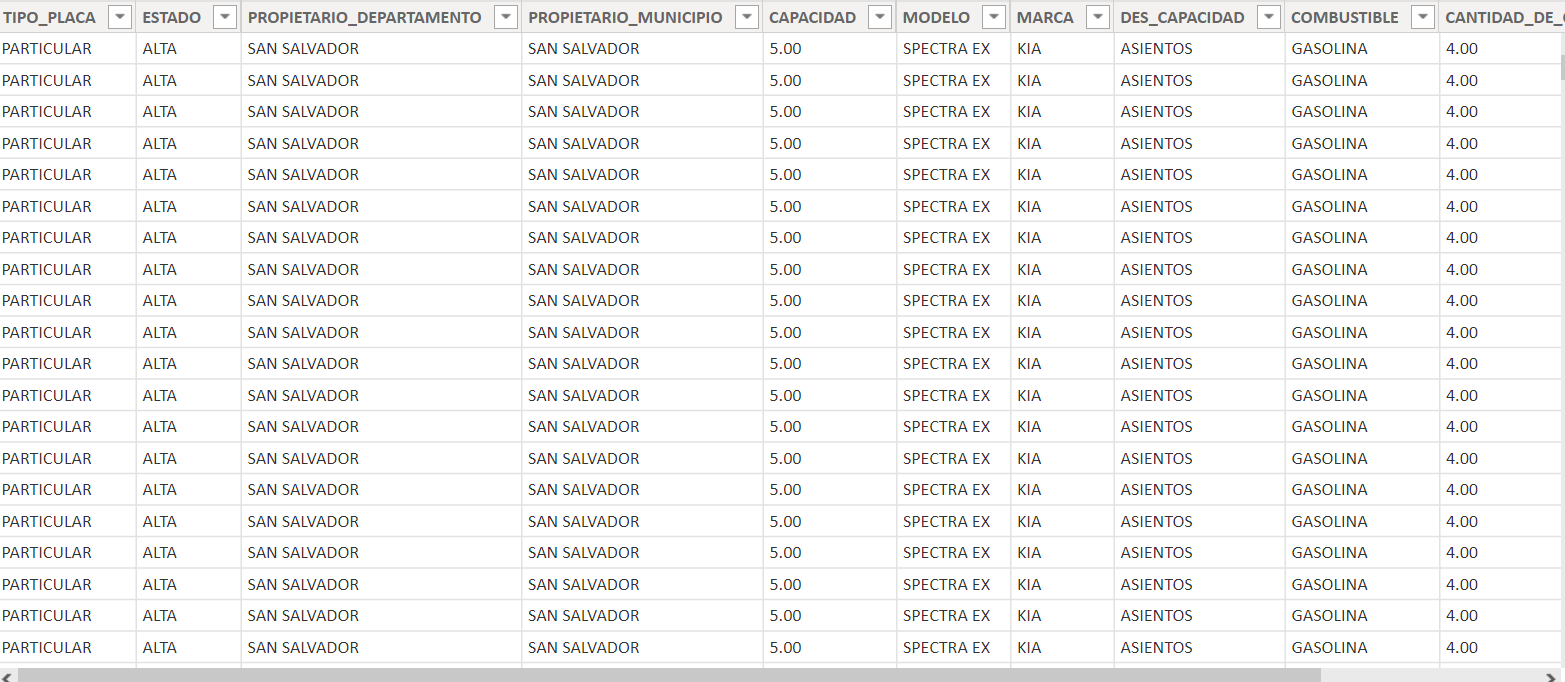


Tabla 3: Ordenar vehículos por capacidad

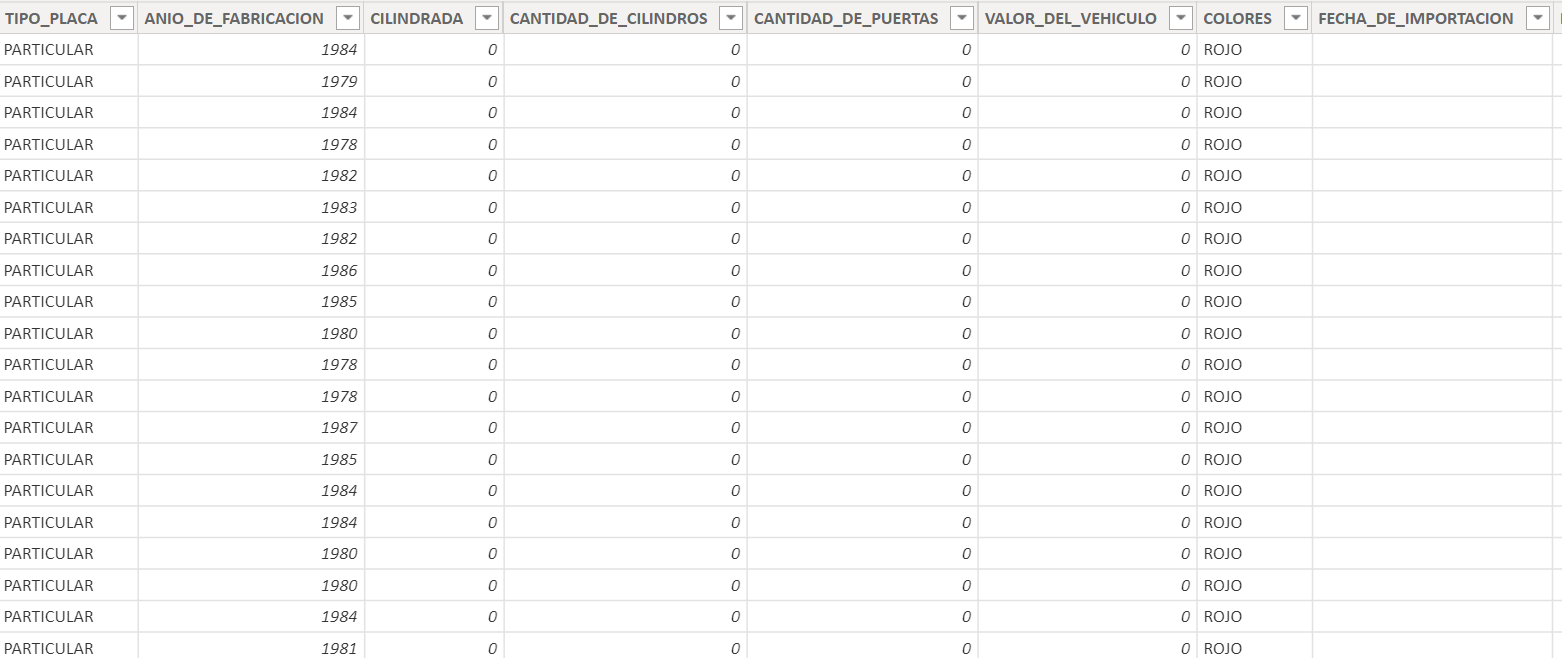
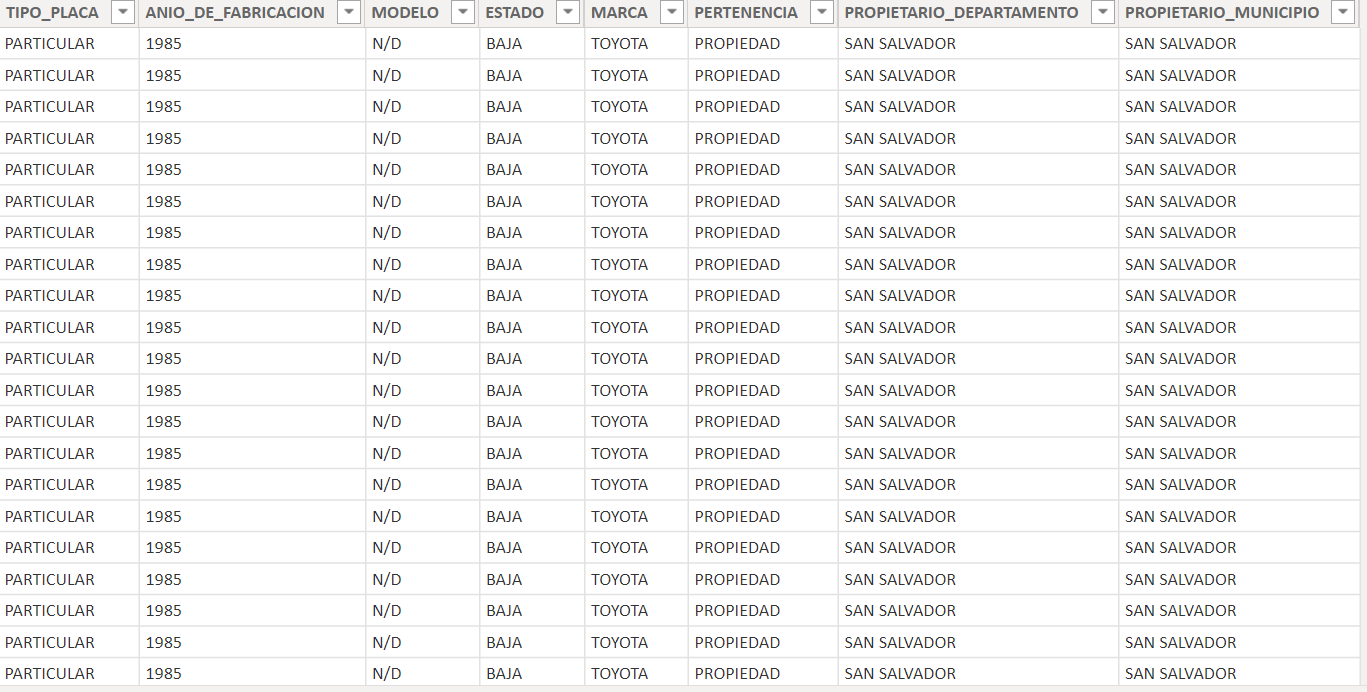


Tabla 4: Parque vehicular datos abiertos



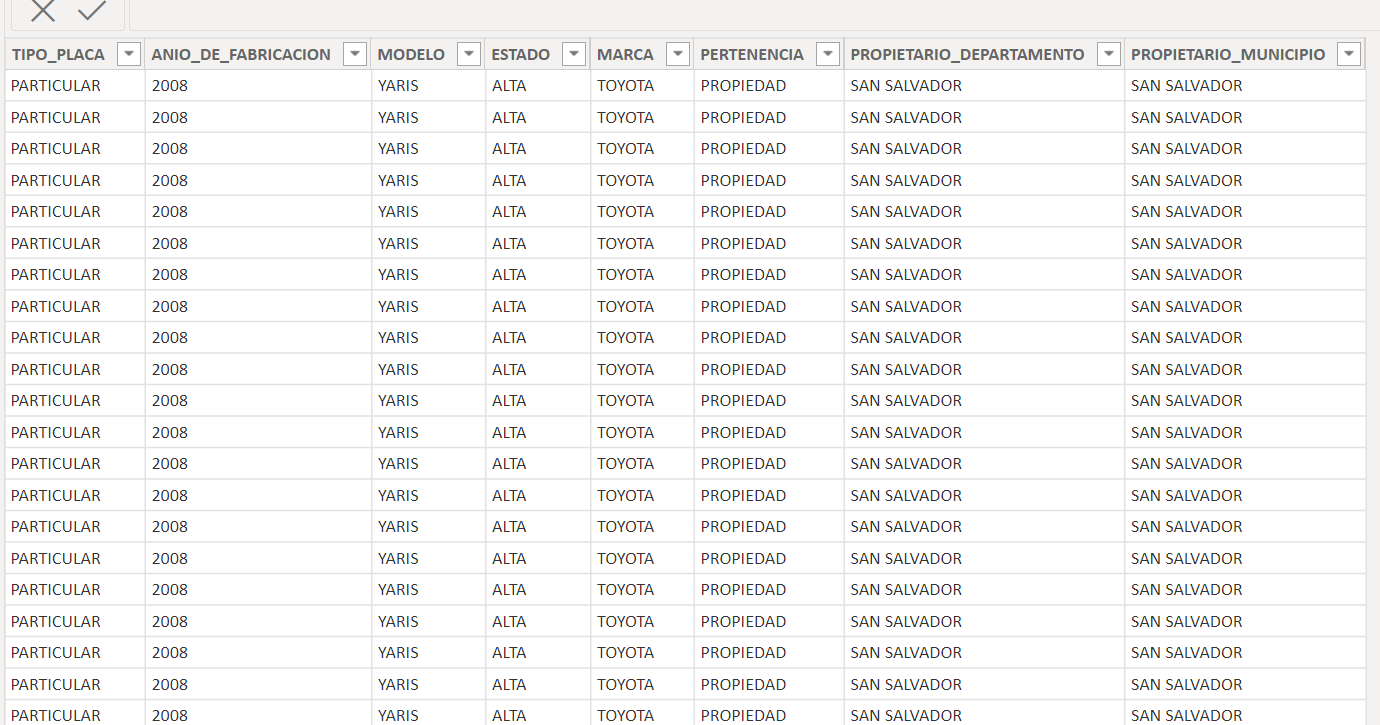
Tabla 5: Propietarios de Vehículos viejos

Tabla 6: Propietarios de vehículos nuevos

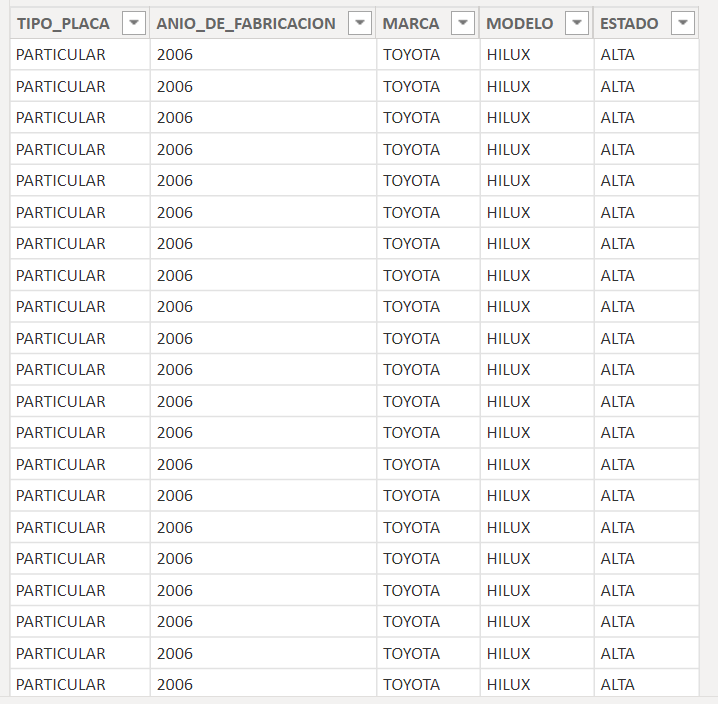


Tabla 7: Vehículos nuevos del 2000

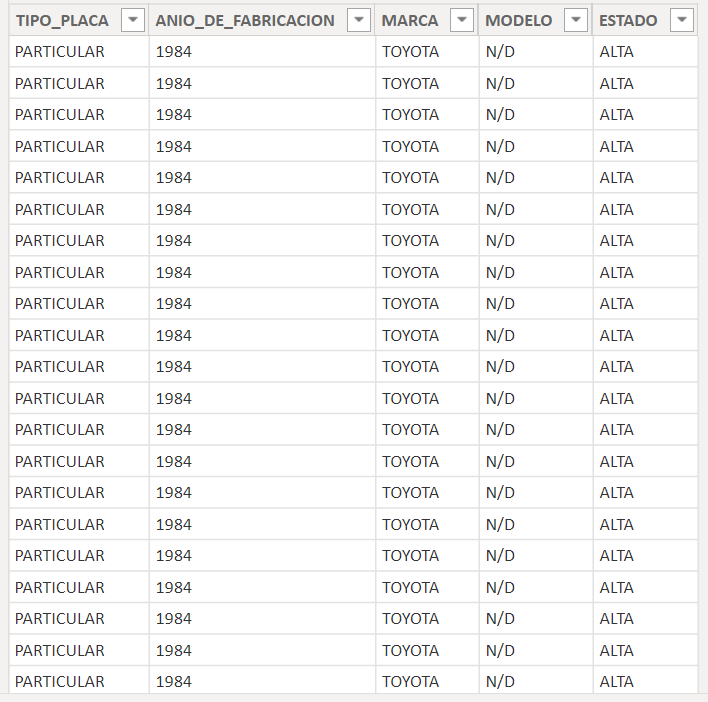
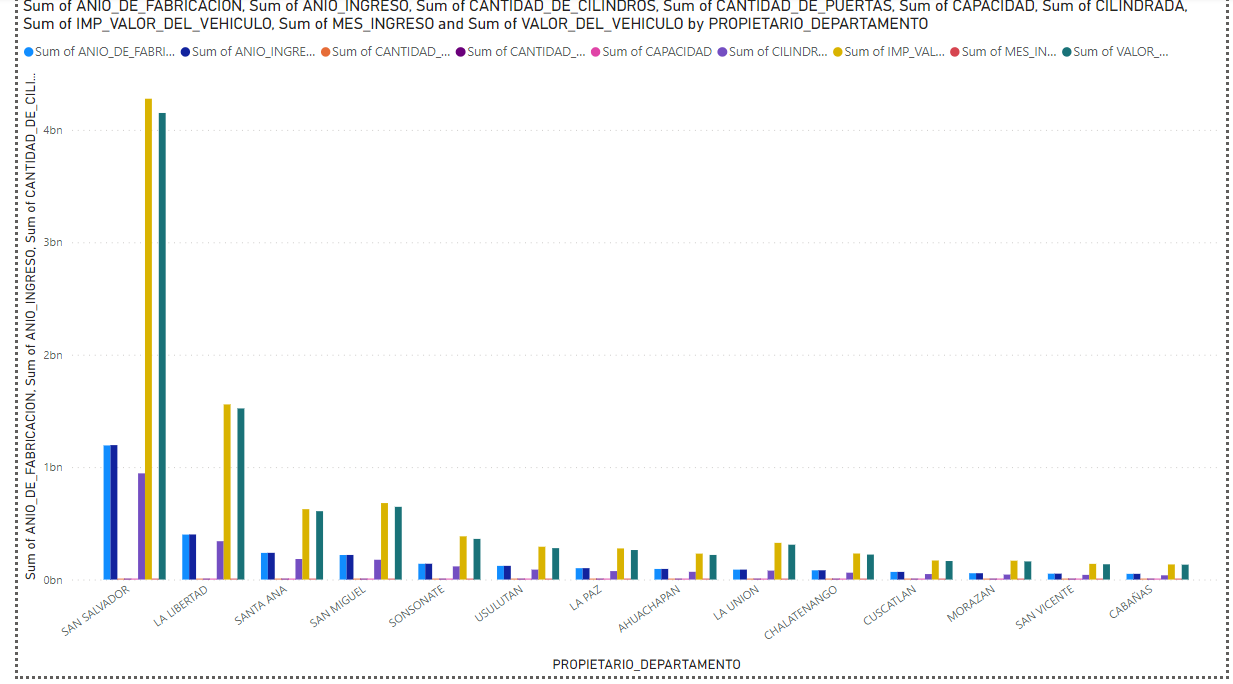


Tabla 8: Vehículos Viejos del 2000

* Los siguientes gráficos son basados en las vistas y las tablas.

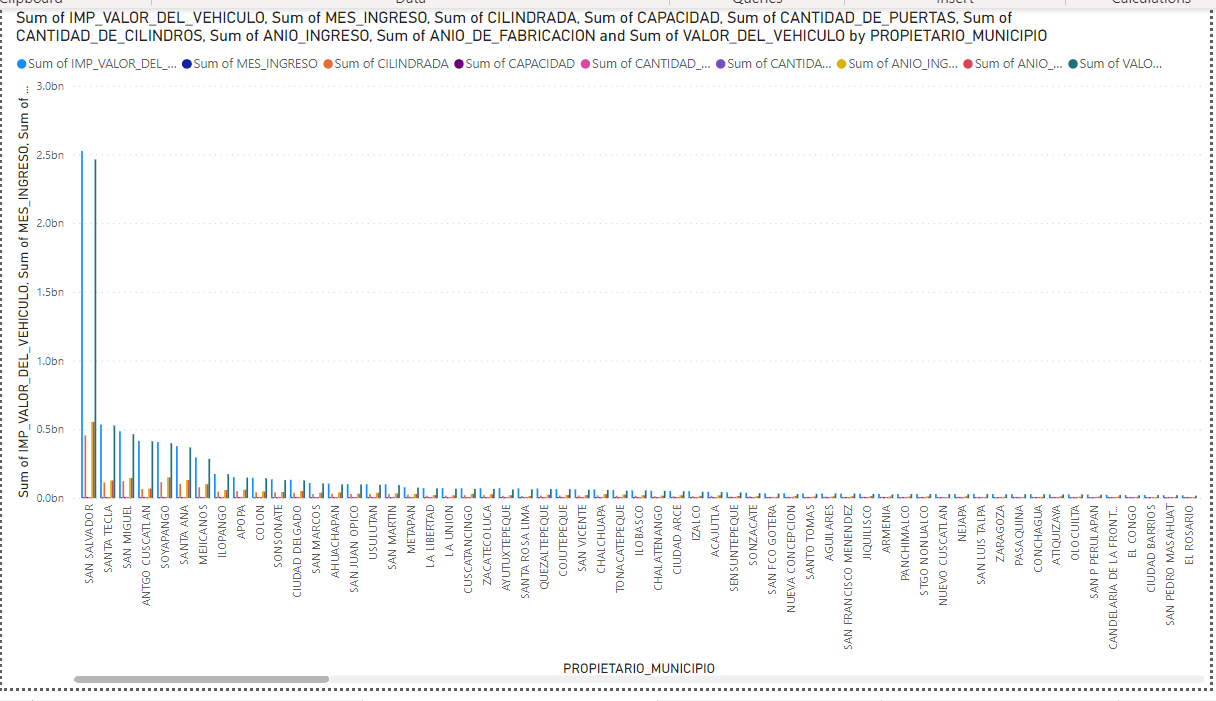
GRAFICO 1: Gráficos por departamentos

Grafico 2: Gráficos por municipio

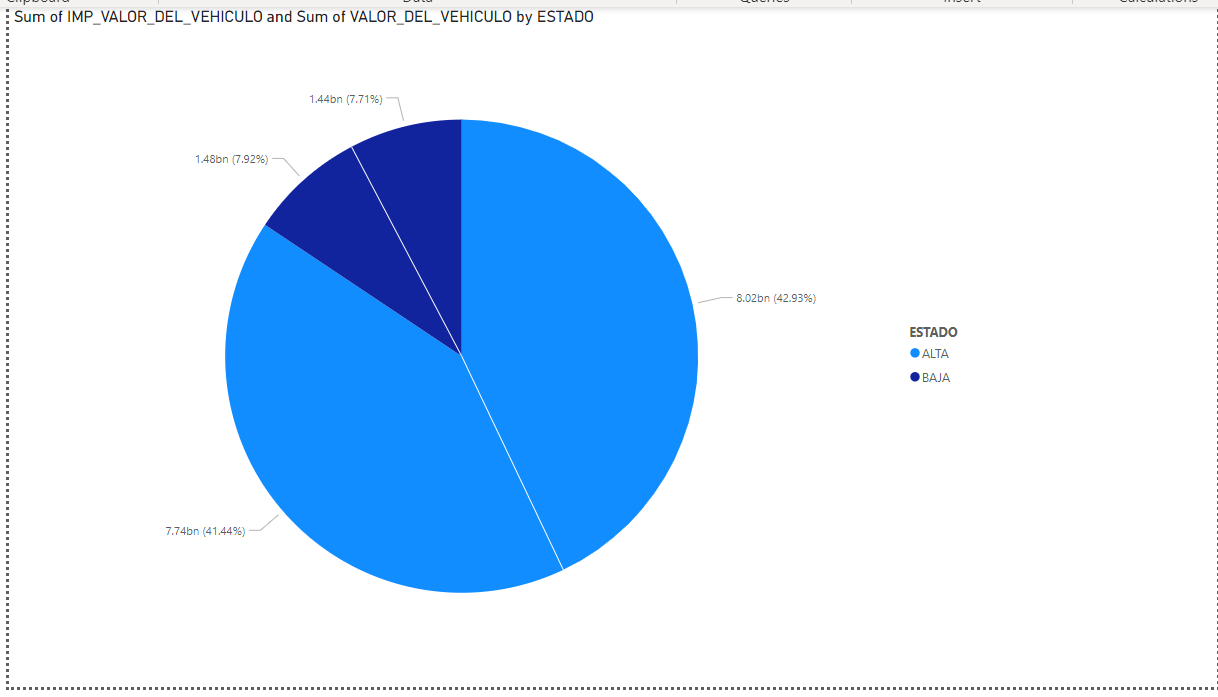
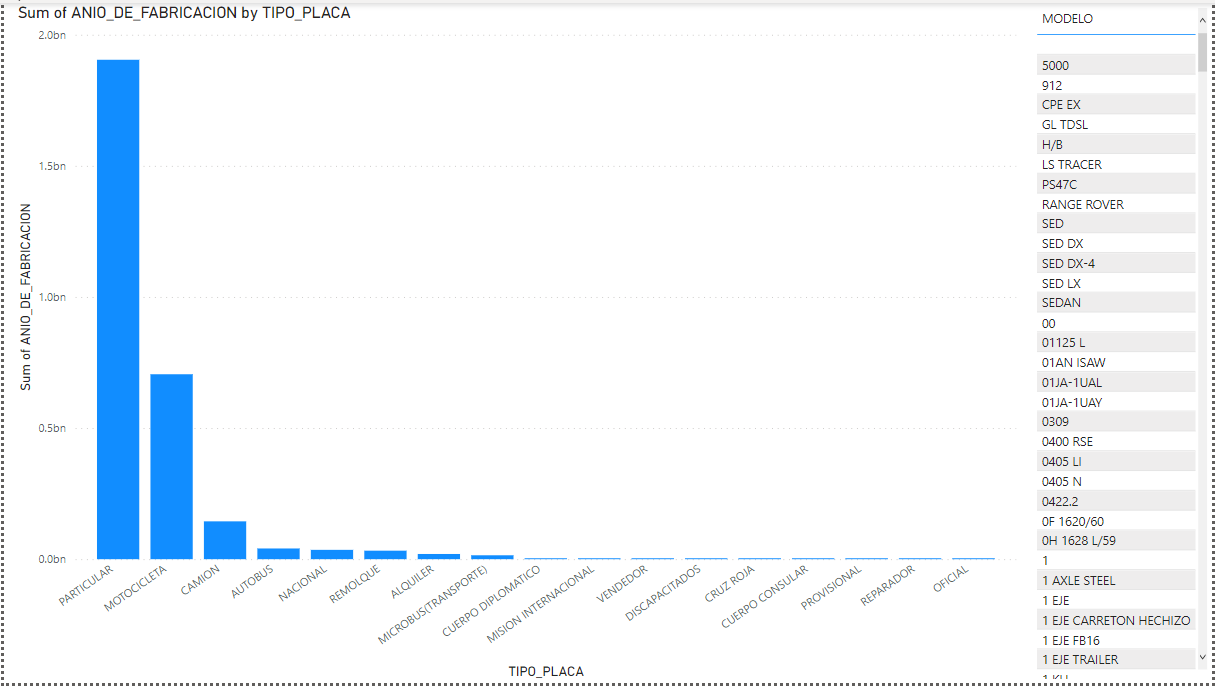
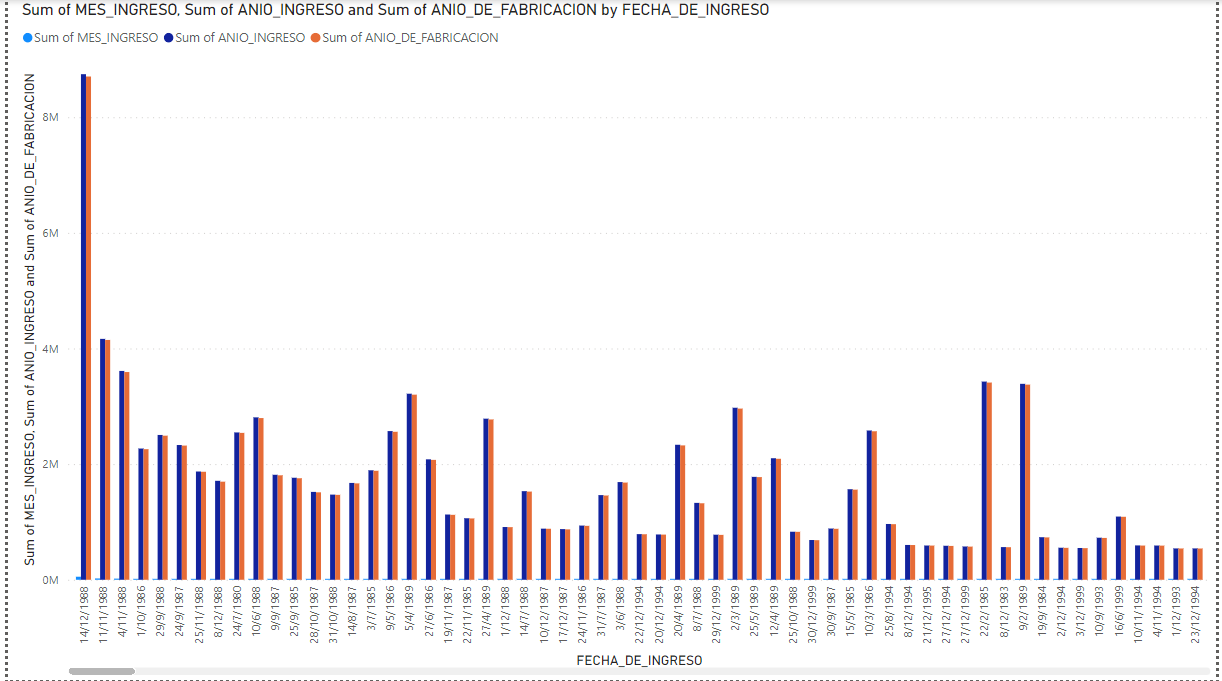
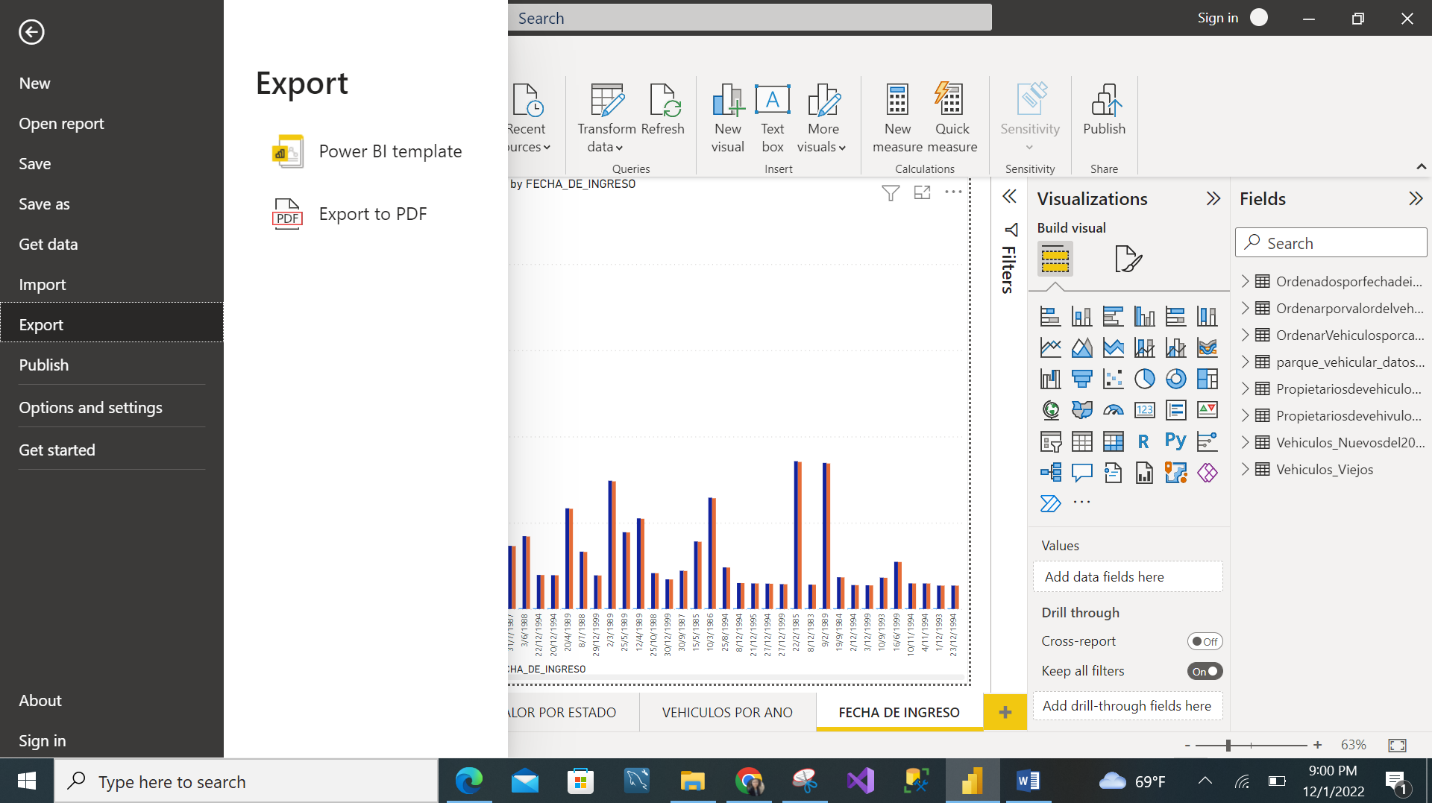
Grafico 3: Valores por estado

Grafico 4: Vehículos por año

Grafico 5: Valores de egreso 

* Impresión del reporte:

**TEMA: K-MEANS USANDO HERRAMIENTAS DE RAPIDMINER.**

Para la realización de análisis K-means podemos hacer uso de diversas herramientas, en nuestro caso usaremos el software Rapidminer y en especifico su asistente llamado automodel, en el cual iremos seleccionando la información que representaremos por medio del modelo k-menas.

* Para empezar, escogeos la base de datos con la que trabajaremos, en este caso será: parque\_vehicular\_datos\_abiertos\_13NOV2018

Graphical user interface, text, application, email, website

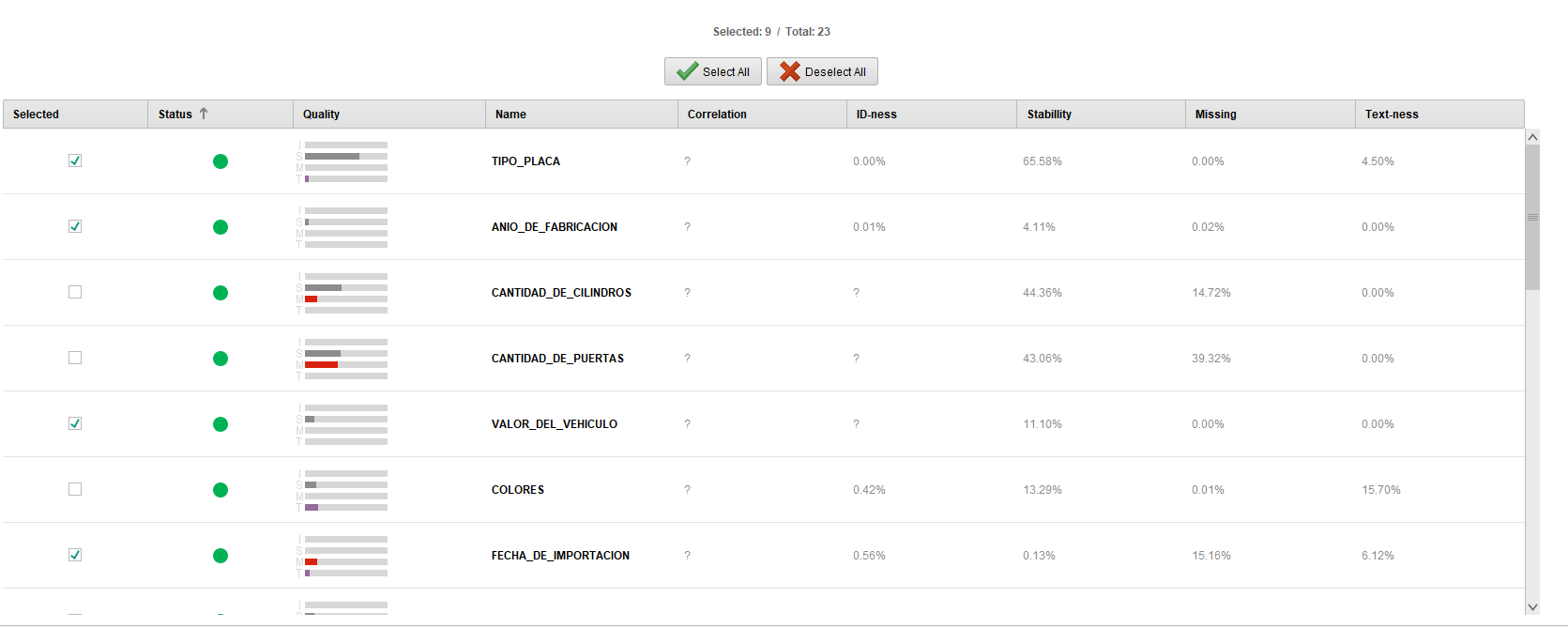
Description automatically generated

Rapidminer nos da la opción del análisis que deseamos hacer, ya sea: predicciones, agrupamientos o identificación de valores atípicos, para nuestro análisis usaremos los agrupamientos

Graphical user interface, text, application

Description automatically generated

Luego manualmente filtraremos los datos relevantes para nuestros agrupamientos.



Al tener claros los datos que analizaremos, debemos escoger el número de agrupamientos que deseamos, en este caso serán 3 ya que el tamaño de la base de datos es tan extenso que la computadora no soporta un análisis mayor.

Graphical user interface, application

Description automatically generated

Al finalizar el análisis obtenemos datos relevantes como por ejemplo en nuestro caso analizamos la relación entre el valor promedio de los vehículos y su año de ingreso al país.

Chart, scatter chart

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Con esto podemos observar que no necesariamente los vehículos mas nuevos son los mas costosos, ya que los datos arrojan que muchos vehículos antiguos tienen un valor muy elevado.

1. **ANTECEDENTES**

El análisis y la minería de datos se han convertido en una necesidad para las empresas e instituciones, esto debido a que la toma de decisiones tiene como uno de sus principales factores las predicciones basadas en las consecuencias de las acciones previas, por lo cual, no solamente se trata de llevar un registro de los datos, sino, de saber obtener información de relevancia por medio la interpretación de estos.

1. **SITUACION ACTUAL**

El “Ministerio de obras públicas” no cuenta con un análisis de los datos que tienen almacenados, únicamente poseen una base de datos con los registros de las esquelas, dicha base de datos ha sido entregada e n un documento CSV el cual consta de una única tabla con diversos campos relacionados a las esquelas impuestas, dicha base consta de más de un millón de registros los cuales deben ser analizados.

La información presentada de un “Parque Vehicular” mediante un archivo CSV el cual consta únicamente de una tabla y de un volumen de archivos grandes, ha sido presentada para ser analizados.

1. **METODOLOGIA**

A metodología utilizada en el presente proyecto será la de investigación descriptiva, ya que se utilizará dos bases de datos desorganizadas, las cuales serán analizadas por medio de diversas técnicas de minería de datos, con el fin de obtener información relevante las cuales serán presentadas de una manera organizada.

1. **FORMULACION DEL PROBLEMA**

La actual base de datos que posee el “Ministerio de obras públicas”, cuenta con registros de las esquelas impuestas sobre el territorio nacional, sin embargo, no tienen la capacidad de analizar dichos datos, por lo que no tienen un funcionamiento más que llevar el registro histórico, es necesario poder analizar la información que poseen ya que de esta manera se pueden tomar medidas de prevención y/o corrección.

En la base de datos presentada de un “Parque Vehicular” no existe una adecuada organización y optimización de los datos, por lo cual se necesita emplear un informe correcto y detallado para presentar un análisis y solución del problema.

1. **FACTIBILIDAD**

El análisis de los datos se realizará utilizando diversas herramientas de software y sus respectivas licencias, además se utilizará equipo con características de gamas media, esto debido a que el procesamiento de bases de datos tan extensas obliga a utilizar un poder de procesamiento medio-alto, luego del procesamiento de los datos, se espera que la información obtenida pueda ser utilizada para la creación de nuevas estrategias y planes por parte de la institución.

1. **JUSTIFICACION**

Hemos implementado **estrategias de minería de datos** para poder organizar una cantidad grande de datos y hacer un análisis de las esquelas de tránsito y del parque vehicular, lo cual son temas de los que se desean presentar reportes.

Usando **estrategias de minería de datos**, podemos visualizar los datos de una forma más entendible y gráfica, siendo esta una ayuda para el MOP, ya que podrán valorar los datos y dar una respuesta inmediata a situaciones que se presenten dentro de los resultados, por esa razón se presenta esta documentación.

Haciendo una aplicación correcta de las herramientas podremos conocer la manera correcta de adecuar los datos y beneficiar al parque vehicular, para encontrar respuestas a soluciones que se pueden tener mediante el orden de los datos.

1. **IMPORTANCIA**

La importancia de poder hacer uso de estas estrategias es el poder documentar un gran volumen de datos y poder brindar resultados más gráficos, siendo este el uso de Power BI, en el cual hicimos gráficos los resultados, análisis OLAP haciendo uso del programa Power BI, Reporting Services y k-means utilizando Rapidminer , herramientas que facilitan el manejo de grandes cantidades de datos, los cuales sería muy difícil analizar de manera manual , por lo cual, dichas herramientas nos ayudan a hacer la impresión de los datos para que se pueda tener acceso por quienes observaran los análisis realizados.

1. **ALCANCES Y LIMITACIONES**

Se debe tener equipos y herramientas técnicas adecuadas como así también una excelente gestión a la hora de hacer los procesos de ordenamiento de datos mediante las estrategias de minería, razón por la cual se puede tener una complicación cuando se tenga ya presente los documentos, si estos no han tenido un proceso correcto en el desarrollo del análisis de los datos, por eso se dice que la mayor complicación es que las herramientas son demasiadas complejas.

1. **PLANIFICACION DE RECURSOS**

A la hora de planificar se debe de tener presente el volumen, tipo de datos, variables, y todos aquellos aspectos necesarios antes de implementar la herramienta para que se puedan detallar la información y datos de una forma más clara, eso nos dará una mejor referencia de planificación.

1. **RECOMENDACIONES**

* Tener presente el formato de archivo para el uso de la herramienta.
* Hacer una investigación de la implementación de la herramienta.
* Tener presente la entrega de los datos físicos. Valorar los datos más importantes que serán expuestos de acuerdo al tipo de reporte que se quiere presentar.