

UNIVERSIDAD DON BOSCO



DIRECCIÓN DE EDUCACIÓN A DISTANCIA

Asignatura: Datawarehouse y Minería de Datos

Docente: Magister Karens Medrano

Proyecto de cátedra fase II

Presentado por:

Rigoberto Alcides Villalta - VV000329

El Salvador, 10 de diciembre de 2022

Análisis de parque vehicular

Hemos realizado el ETL y análisis en el software RapidMiner. Primero tratamos de cargar los datos en RapidMiner, pero el archivo nos generaba errores en dos columnas, dado que por la cantidad de datos se imposibilita el uso de LibreOffice o Excel para modificarlo, ocupamos el lenguaje de programación *Python* y el módulo *csv* para poder modificar los datos, el *script* ocupado es el siguiente (está ampliamente comentado para explicar el proceso):

```
from tempfile import NamedTemporaryFile
import shutil
import csv

# Creamos un archivo temporal con el modo escribir "w"
parque_vehicular_modificado = NamedTemporaryFile(mode="w", delete=False)
fields = ["TIPO_PLACA", "ANIO_DE_FABRICACION", # 1
          "CILINDRADA", #2
          "CANTIDAD_DE_CILINDROS", "CANTIDAD_DE_PUERTAS", "VALOR_DEL_VEHICULO",
          "COLORES", "FECHA_DE_IMPORTACION", "IMP_VALOR_DEL_VEHICULO",
          "FECHA_DE_INGRESO", "ANIO_INGRESO", "MES_INGRESO", "CLASE",
          "PERTENENCIA", "MARCA", "MODELO", "CAPACIDAD", "DES_CAPACIDAD",
          "COMBUSTIBLE", "ADUANA", "CONDICION_INGRESO", "PROPIETARIO_DEPARTAMENTO",
          "PROPIETARIO_MUNICIPIO", "ESTADO"]

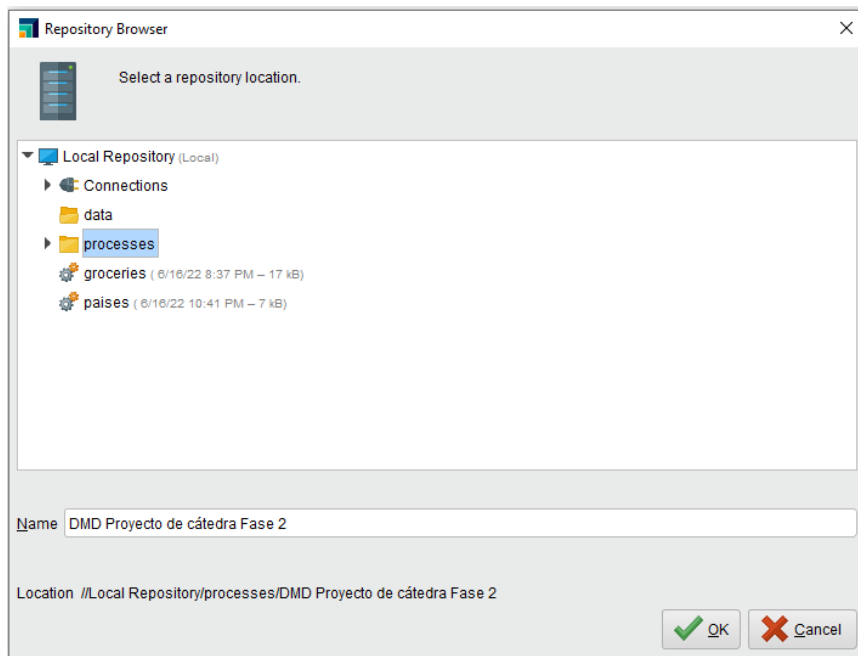
# Abrimos nuestro archivo en modo escribir
with open("parque_vehicular.csv", "r") as parque_vehicular, parque_vehicular_modificado:
    # Creo un nuevo archivo csv en modo escritura, ahí guardamos lo ocupado
    lectura_de_archivo = csv.DictReader(parque_vehicular, fieldnames=fields)
    writer = csv.DictWriter(parque_vehicular_modificado, fieldnames=fields)

    # En Python cada fila se vuelve una lista, iteramos el archivo en cada lista
    for fila in lectura_de_archivo:
        # ahora vamos a tratar de modificar de que cada campo que queremos limpiar
        # lo haremos con los campos que nos han dado algún problema, CILINDRADA y ANIO_DE_FABRICACION:
        try:
            # convertimos el dato en entero
            nueva_cilindrada = int(fila[1])
            nuevo_anio_de_fabricacion = float(fila[2])
            nueva_fila = [
                fila[0], nueva_cilindrada, nuevo_anio_de_fabricacion,
                fila[3], fila[4], fila[5], fila[6], fila[7],
                fila[8], fila[9], fila[10], fila[11], fila[12],
                fila[13], fila[14], fila[15], fila[16], fila[17],
                fila[18], fila[19], fila[20], fila[21], fila[22],
                fila[23]
            ]
            writer.writerow(nueva_fila)
        except:
            # si falla se ignora la fila
            pass

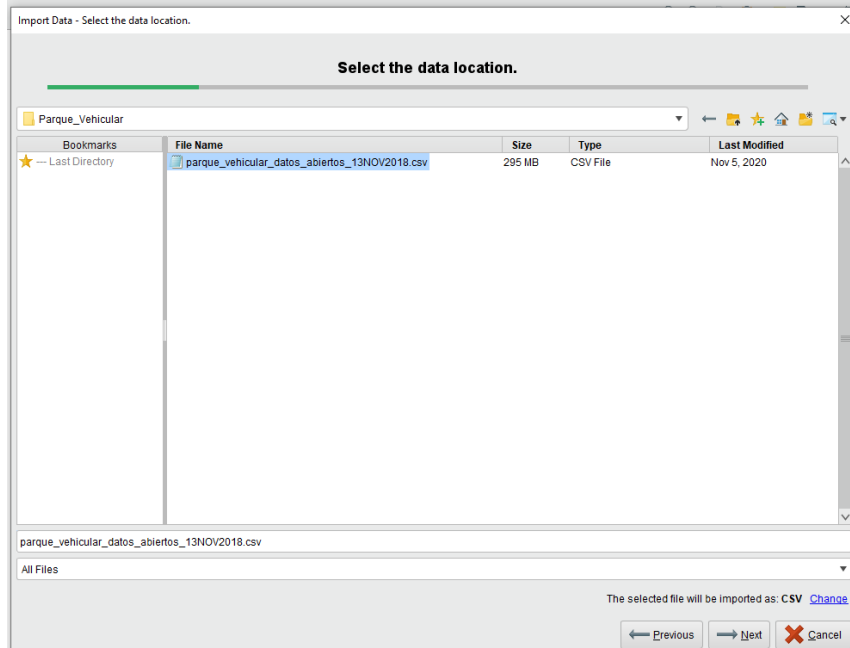
#Al terminar la iteracion guardamos el nuevo archivo
shutil.move(parque_vehicular_modificado.name, "parque_vehicular_modificado.csv")
```

Adjunto al proyecto el Script.

Ahora si creamos un nuevo proyecto de RapidMiner:



Importo lo datos desde el archivo proporcionado:



Import Data - Specify your data format

Specify your data format

☒ Header Row File Encoding ☒ Use Quotes

Start Row Escape Character ☐ Trim Lines

Column Separator Decimal Character ☒ Skip Comments

| 1 | TIPO_PL... | ANIO_DE... | CILINDR... | CANTIDA... | CANTIDA... | VALOR_... | COLORES | FECHA_... | IMP_VA... | FECHA_... | ANIO_IN... | MES_IN... | CT |
|----|------------|------------|------------|------------|------------|-----------|------------|-----------|-----------|-------------|------------|-----------|----|
| 2 | PARTIC... | 1990 | 1800 | 0.00 | 4.00 | 4094.56 | AMARILLO | 16/9/1994 | 4094.56 | 17/11/19... | 1994.00 | 11.00 | AL |
| 3 | PARTIC... | 1964 | 0 | 0.00 | 0.00 | 0.00 | AMARILLO | | 0.00 | 16/2/1989 | 1989.00 | 2.00 | AL |
| 4 | ALQUILER | 1984 | 1700 | 0.00 | 0.00 | 0.00 | AMARILL... | | 0.00 | 20/3/1985 | 1985.00 | 3.00 | AL |
| 5 | ALQUILER | 1986 | 1600 | 0.00 | 0.00 | 0.00 | AMARILL... | | 0.00 | 18/2/1988 | 1988.00 | 2.00 | AL |
| 6 | ALQUILER | 1979 | 0 | 0.00 | 0.00 | 0.00 | AMARILLO | | 0.00 | 26/10/19... | 1982.00 | 10.00 | AL |
| 7 | PARTIC... | 1974 | 1600 | 0.00 | 0.00 | 0.00 | AMARILLO | | 0.00 | 26/10/19... | 1982.00 | 10.00 | AL |
| 8 | ALQUILER | 1975 | 0 | 0.00 | 0.00 | 0.00 | AMARILLO | | 0.00 | 23/5/1984 | 1984.00 | 5.00 | AL |
| 9 | ALQUILER | 1973 | 0 | 0.00 | 0.00 | 0.00 | AMARILL... | | 0.00 | 3/6/1988 | 1988.00 | 6.00 | AL |
| 10 | ALQUILER | 1975 | 0 | 0.00 | 0.00 | 0.00 | AMARILLO | | 0.00 | 4/2/1988 | 1988.00 | 2.00 | AL |
| 11 | ALQUILER | 1968 | 0 | 0.00 | 0.00 | 800.00 | AMARILLO | 29/7/1983 | 0.00 | 29/8/1983 | 1983.00 | 8.00 | AL |
| 12 | ALQUILER | 1977 | 0 | 0.00 | 0.00 | 0.00 | AMARILLO | | 0.00 | 12/6/1984 | 1984.00 | 6.00 | AL |
| 13 | ALQUILER | 1965 | 0 | 0.00 | 0.00 | 0.00 | AMARILLO | | 0.00 | 26/10/19... | 1982.00 | 10.00 | AL |
| 14 | ALQUILER | 1983 | 1587 | 0.00 | 4.00 | 4183.76 | AMARILL... | 24/2/1995 | 4183.76 | 24/3/1995 | 1995.00 | 3.00 | AL |
| 15 | ALQUILER | 1978 | 0 | 0.00 | 0.00 | 0.00 | AMARILLO | | 0.00 | 12/4/1989 | 1989.00 | 4.00 | AL |
| 16 | PARTIC... | 1981 | 1770 | 0.00 | 0.00 | 4413.71 | AMARILLO | 20/9/1993 | 4413.71 | 27/10/19... | 1993.00 | 10.00 | AL |
| 17 | ALQUILER | 1967 | 0 | 0.00 | 0.00 | 0.00 | AMARILLO | | 0.00 | 26/10/19... | 1982.00 | 10.00 | AL |

no problems.

Previous Next Cancel

Ajustamos los tipos de datos a los que corresponden:

Aquí es muy importante hacer notar que los tipos polinomiales y binomiales son de suma importancia, ya que de no estar correctamente configurados y de tener datos incorrectos no se tendrá las agregaciones y agrupaciones de forma correcta

Import Data - Format your columns.

Format your columns.

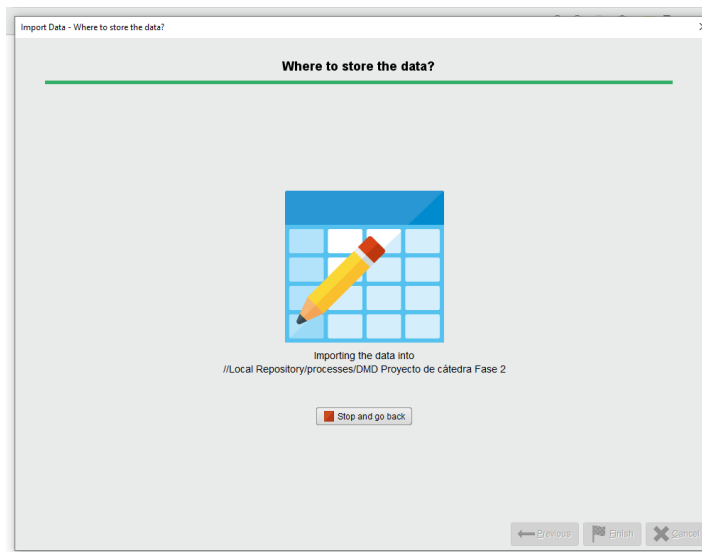
Date format 100% ☐ Replace errors with missing values

| | CANTIDAD_... | CANTIDAD_... | VALOR_DE... | COLORES | FECHA_DE... | IMP_VALO... | FECHA_DE... | ANIO_INGI |
|----|--------------|--------------|-------------|------------------|-------------|-------------|-------------|-----------|
| | real | real | real | polynomial | polynomial | polynomial | polynomial | real |
| 1 | 0.000 | 4.000 | 4094.560 | AMARILLO | 16/9/1994 | | | 1994.000 |
| 2 | 0.000 | 0.000 | 0.000 | AMARILLO | ? | | | 1989.000 |
| 3 | 0.000 | 0.000 | 0.000 | AMARILLO F/C B/N | ? | | | 1985.000 |
| 4 | 0.000 | 0.000 | 0.000 | AMARILLO F/C B/N | ? | 0.000 | | 1988.000 |
| 5 | 0.000 | 0.000 | 0.000 | AMARILLO | ? | 0.000 | | 1982.000 |
| 6 | 0.000 | 0.000 | 0.000 | AMARILLO | ? | 0.000 | | 1982.000 |
| 7 | 0.000 | 0.000 | 0.000 | AMARILLO | ? | 0.000 | | 1984.000 |
| 8 | 0.000 | 0.000 | 0.000 | AMARILLO F/C B/N | ? | 0.000 | | 1988.000 |
| 9 | 0.000 | 0.000 | 0.000 | AMARILLO | ? | 0.000 | | 1988.000 |
| 10 | 0.000 | 0.000 | 800.000 | AMARILLO | 29/7/1983 | 0.000 | | 1983.000 |
| 11 | 0.000 | 0.000 | 0.000 | AMARILLO | ? | 0.000 | | 1984.000 |
| 12 | 0.000 | 0.000 | 0.000 | AMARILLO | ? | 0.000 | | 1982.000 |
| 13 | 0.000 | 4.000 | 4183.760 | AMARILLO F/C B/N | 24/2/1995 | 4183.760 | | 1995.000 |
| 14 | 0.000 | 0.000 | 0.000 | AMARILLO | ? | 0.000 | | 1989.000 |
| 15 | 0.000 | 0.000 | 4413.710 | AMARILLO | 20/9/1993 | 4413.710 | | 1993.000 |
| 16 | 0.000 | 0.000 | 0.000 | AMARILLO | ? | 0.000 | | 1982.000 |
| 17 | 0.000 | 4.000 | 4619.440 | AMARILLO NEG... | 26/4/1994 | 4619.440 | | 1994.000 |

no problems.

Previous Next Cancel

La importación se realiza correctamente y sin problemas.:



Dado el trabajo previo realizado, ahora RapidMiner nos despliega una gran cantidad de información estadística y de datos:

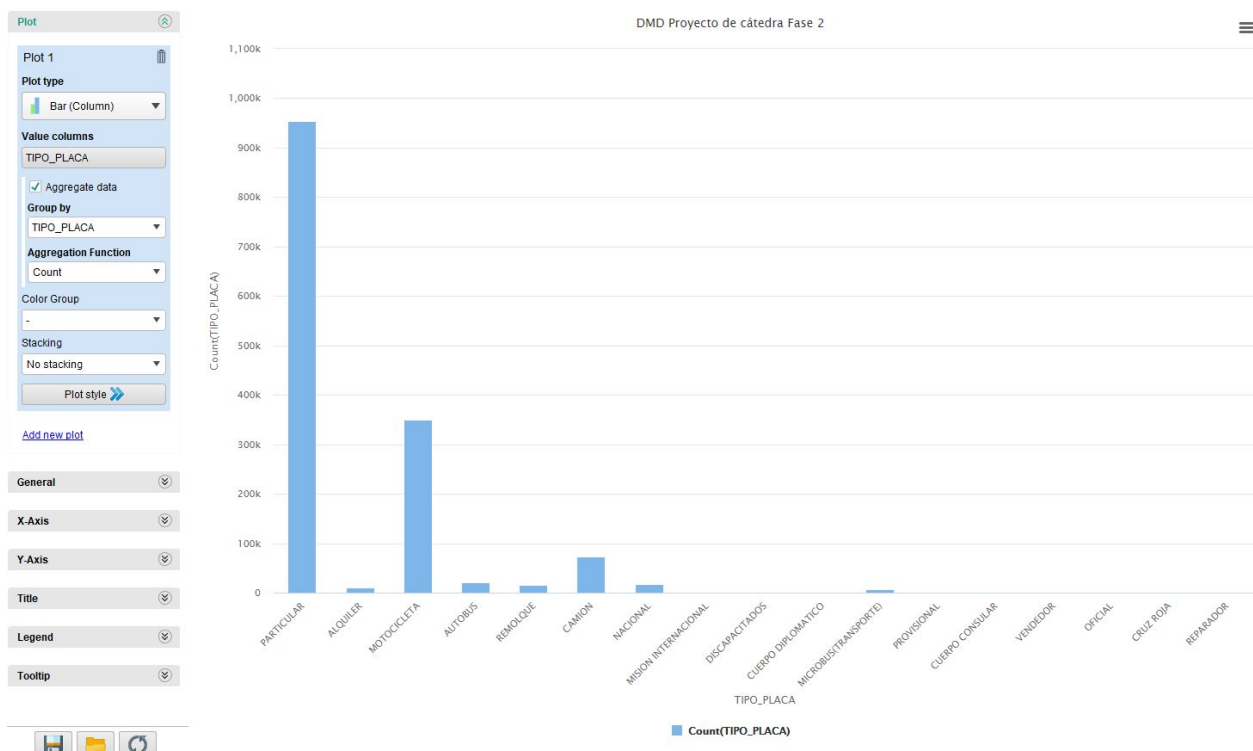


Gráfico 1: Tipo de vehículo por total (Count)

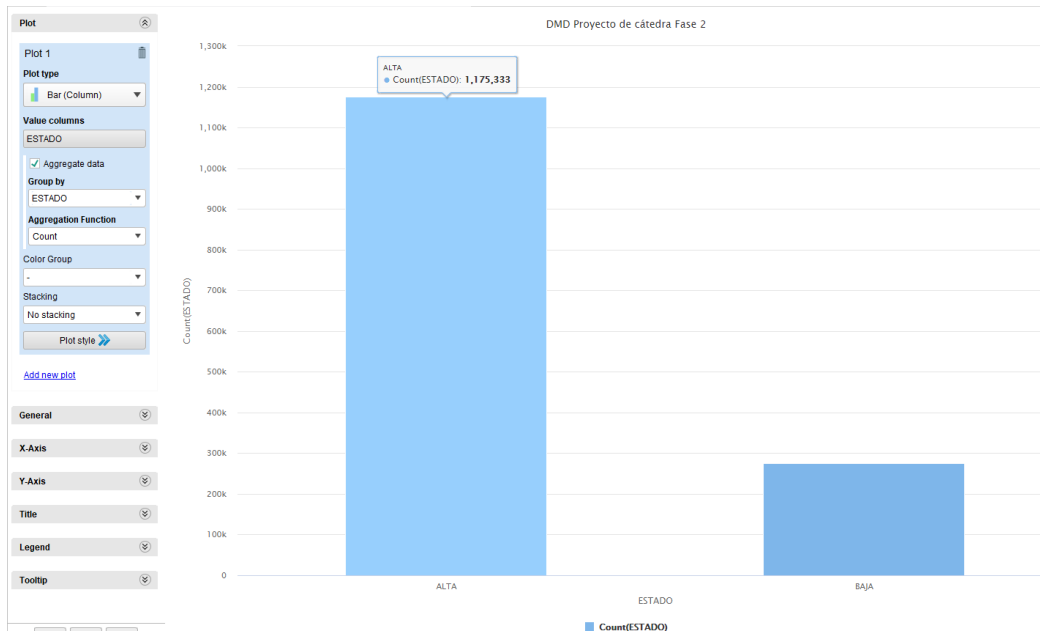


Gráfico 2: Vehículos en alta y baja (ya no están en el paquete vehicular)

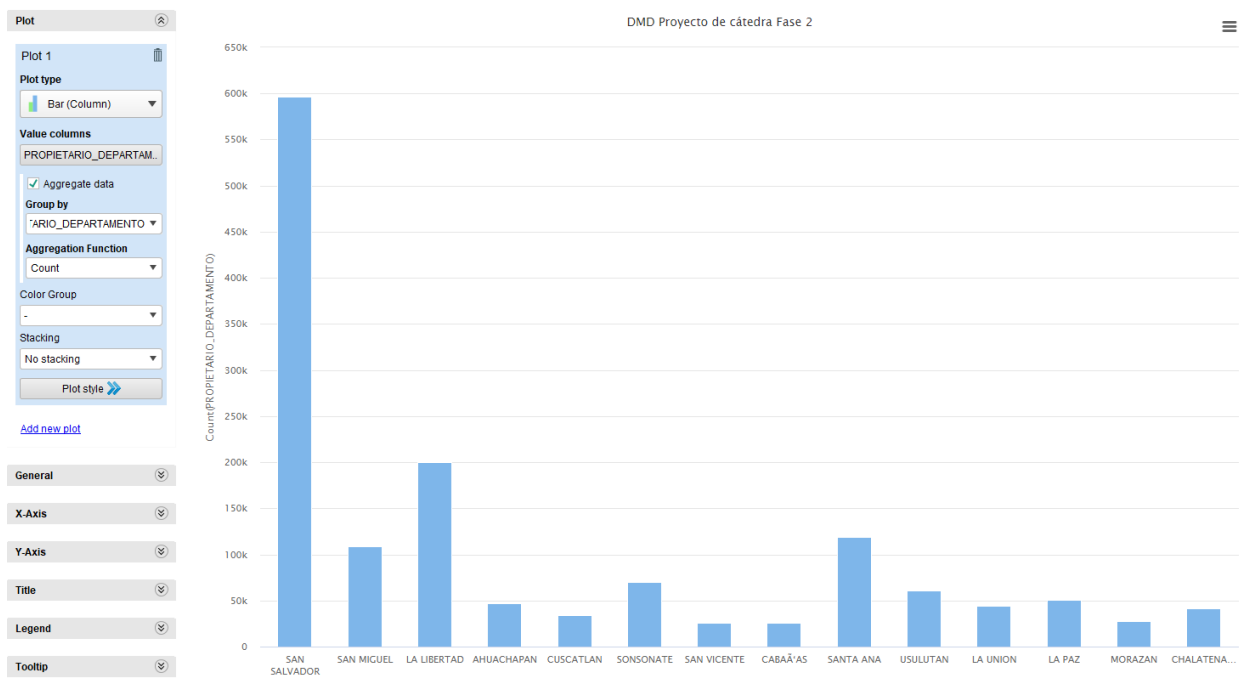


Gráfico 3: Vehículo por departamento

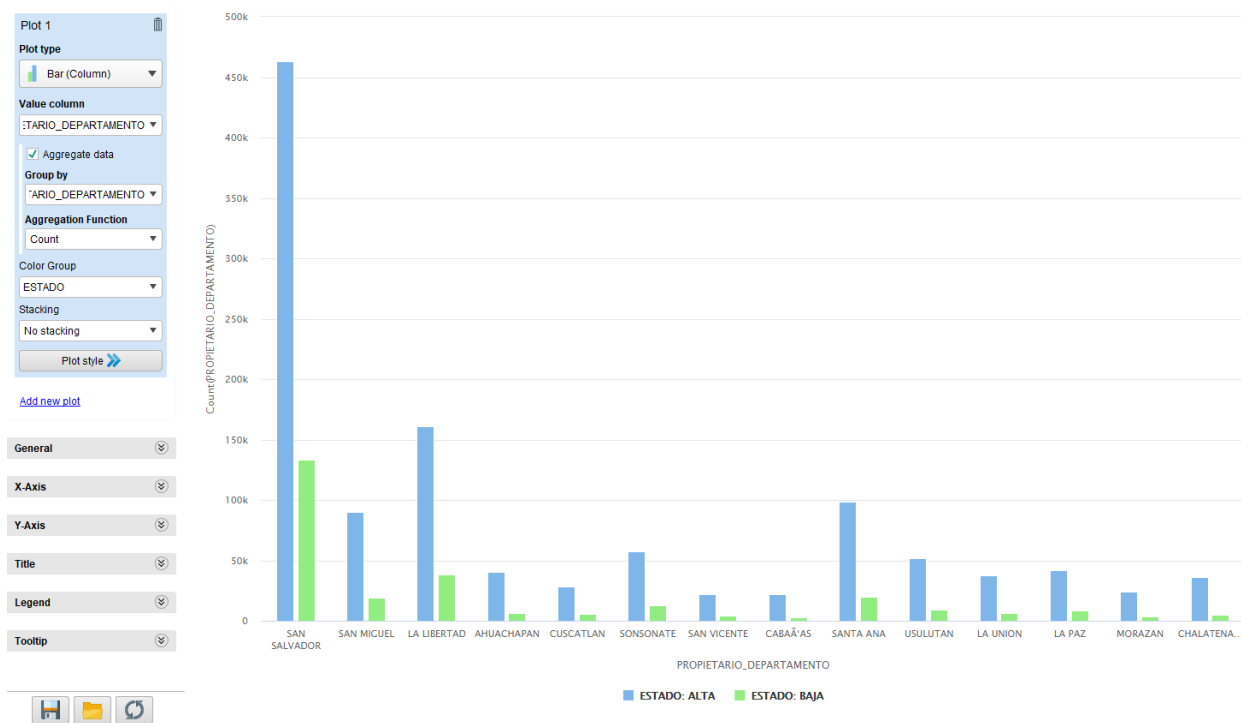



Gráfico 4: Comparativa vehiculos de alta y baja por departamento



| Index | Nominal value | Absolute count | Fraction |
|-------|---------------|----------------|----------|
| 1 | SAN SALVADOR | 596434 | 0.411 |
| 2 | LA LIBERTAD | 199935 | 0.138 |
| 3 | SANTA ANA | 118535 | 0.082 |
| 4 | SAN MIGUEL | 109071 | 0.075 |
| 5 | SONSONATE | 70141 | 0.048 |
| 6 | USULUTAN | 60744 | 0.042 |
| 7 | LA PAZ | 50509 | 0.035 |
| 8 | AHUACHAPAN | 47143 | 0.032 |
| 9 | LA UNION | 44089 | 0.030 |
| 10 | CHALATENANGO | 41054 | 0.028 |
| 11 | CUSCATLAN | 33964 | 0.023 |
| 12 | MORAZAN | 28021 | 0.019 |
| 13 | SAN VICENTE | 26103 | 0.018 |
| 14 | CABAÑAS | 25357 | 0.017 |

Gráfico 5: Conteo de vehículos por departamento

Dentro del contenido del proyecto se pueden ver muchos más gráficos y estadísticas.

K-means

Para el análisis de los datos primero corrimos un K-means:

Load DataSelect TaskPrepare TargetSelect InputsModel TypesResults

RESTARTBACKNEXT

Select Data for a New Model

Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Connections
data
processes
DMD Proyecto de cédula Fase 2 (12/10/22 2:35 PM - 147 MB)
DMD Proyecto de cédula Fase 2 (12/10/22 4:31 PM - 3 MB)
groceries (6/10/22 8:37 PM - 17 MB)
paises (6/10/22 10:41 PM - 7 MB)
Temporary Repository (Local)
DB (Legacy)

Information
Name: DMD Proyecto de cédula Fase 2
Number of rows: 1,451,100
Number of columns: 24

Attributes / Columns
TIPO_PLACA, AÑO_DE_FABRICACIÓN, CILINDRADA, CANTIDAD_DE_CILINDROS, CANTIDAD_DE_PUERTAS, VALOR_DEL_VEHICULO, COLORES, FECHA_DE_IMPORTACION, IMP_VALOR_DEL_VEHICULO, FECHA_DE_INGRESO, AÑO_INGRESO, MES_INGRESO, CLASE, PERTENENCIA, MARCA, MODELO, CAPACIDAD, DES, CAPACIDAD, COMBUSTIBLE, AQUEENA, CONDICION_INGRESO, PROPIETARIO_DEPARTAMENTO, PROPIETARIO_MUNICIPIO, ESTADO

Load DataSelect TaskPrepare TargetSelect InputsModel TypesResults

RESTARTBACKNEXT

Predict
Want to predict the values of a column?

Clusters
Want to identify groups in your data?

Outliers
Want to detect outliers in your data?

| ... | VALOR_DEL_... | COLORES | FECHA_DE_I... | IMP_VALOR_... | FECHA_DE_I... | ANIO_INGRE... | MES_INGRESO |
|-----|---------------|------------------|---------------|---------------|---------------|---------------|-------------|
| | Number | Category | Date / Time | Number | Date / Time | Number | Number |
| | 4094.560 | AMARILLO | Sep 16, 1994 | 4094.560 | Nov 17, 1994 | 1994 | 11 |
| | 0 | AMARILLO | ? | 0 | Feb 16, 1989 | 1989 | 2 |
| | 0 | AMARILLO F/C ... | ? | 0 | Mar 20, 1985 | 1985 | 3 |

Seleccionamos solo las columnas que nos interesan:

Select AllDeselect All

| Selected | Status ↑ | Quality | Name | Correlation | ID-ness | Stability | Missing | Text-ness |
|-------------------------------------|-------------|-------------|------------|-------------|---------|-----------|---------|-----------|
| <input checked="" type="checkbox"/> | <div></div> | <div></div> | TIPO_PLACA | ? | 0.00% | 65.58% | 0.00% | 4.50% |

Ya en esta para este momento podemos ver varios datos nos brinda el asistente.

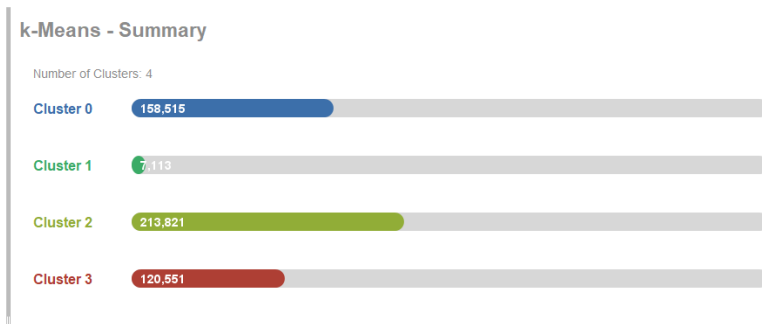
Creamos 4 clusters:

Load DataSelect TaskPrepare TargetSelect InputsModel TypesResults

RESTARTBACKRUN

Models
☒ k-Means Clustering
Number of Clusters: 4
☐ v-Means Clustering

Data Preparation
☒ Remove Columns with Too Many Values
Maximum Number of Values: 50
☐ Extract Data Information



Con esto ya podemos ver algunas correlaciones:

Correlations

| Attribut... | CLASE ... | CLASE ... | CLASE ... | CLASE ... | CLASE ... | CLASE ... | CLASE ... | CLASE ... | CLASE ... | CLASE ... | CLASE ... | CLASE ... | CLASE ... | CLASE ... | CLASE ... | CLASE ... | CLASE ... | CLASE ... | CLASE ... | CLASE ... |
|-------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| CLASE #... | 1 | -0.001 | -0.005 | -0.010 | -0.072 | -0.009 | -0.014 | -0.018 | -0.002 | -0.012 | -0.006 | -0.048 | -0.009 | -0.041 | -0.007 | -0.002 | -0.004 | -0.001 | -0.005 | -0.002 |
| CLASE #... | -0.001 | 1 | -0.001 | -0.002 | -0.013 | -0.002 | -0.002 | -0.003 | -0.000 | -0.002 | -0.001 | -0.008 | -0.002 | -0.007 | -0.001 | -0.000 | -0.001 | -0.000 | -0.001 | -0.000 |
| CLASE #... | -0.005 | -0.001 | 1 | -0.008 | -0.052 | -0.007 | -0.010 | -0.013 | -0.002 | -0.008 | -0.005 | -0.035 | -0.007 | -0.030 | -0.005 | -0.001 | -0.003 | -0.000 | -0.004 | -0.001 |
| CLASE #... | -0.010 | -0.002 | -0.008 | 1 | -0.104 | -0.014 | -0.020 | -0.025 | -0.003 | -0.017 | -0.009 | -0.069 | -0.013 | -0.060 | -0.011 | -0.003 | -0.006 | -0.001 | -0.007 | -0.002 |
| CLASE #... | -0.072 | -0.013 | -0.052 | -0.104 | 1 | -0.094 | -0.134 | -0.174 | -0.023 | -0.115 | -0.062 | -0.474 | -0.090 | -0.410 | -0.074 | -0.020 | -0.040 | -0.005 | -0.048 | -0.017 |
| CLASE #... | -0.009 | -0.002 | -0.007 | -0.014 | -0.094 | 1 | -0.018 | -0.023 | -0.003 | -0.015 | -0.008 | -0.062 | -0.012 | -0.054 | -0.010 | -0.003 | -0.005 | -0.001 | -0.006 | -0.002 |
| CLASE #... | -0.014 | -0.002 | -0.010 | -0.020 | -0.134 | -0.018 | 1 | -0.033 | -0.004 | -0.022 | -0.012 | -0.089 | -0.017 | -0.077 | -0.014 | -0.004 | -0.007 | -0.001 | -0.009 | -0.003 |
| CLASE #... | -0.018 | -0.003 | -0.013 | -0.025 | -0.174 | -0.023 | -0.033 | 1 | -0.006 | -0.028 | -0.015 | -0.116 | -0.022 | -0.100 | -0.018 | -0.005 | -0.010 | -0.001 | -0.012 | -0.004 |
| CLASE #... | -0.002 | -0.000 | -0.002 | -0.003 | -0.023 | -0.003 | -0.004 | -0.006 | 1 | -0.004 | -0.002 | -0.015 | -0.003 | -0.013 | -0.002 | -0.001 | -0.001 | -0.000 | -0.002 | -0.001 |
| CLASE #... | -0.012 | -0.002 | -0.008 | -0.017 | -0.115 | -0.015 | -0.022 | -0.028 | -0.004 | 1 | -0.010 | -0.077 | -0.015 | -0.066 | -0.012 | -0.003 | -0.006 | -0.001 | -0.008 | -0.003 |
| CLASE #... | -0.006 | -0.001 | -0.005 | -0.009 | -0.062 | -0.008 | -0.012 | -0.015 | -0.002 | -0.010 | 1 | -0.041 | -0.008 | -0.036 | -0.006 | -0.002 | -0.003 | -0.000 | -0.004 | -0.001 |
| CLASE #... | -0.048 | -0.008 | -0.035 | -0.069 | -0.474 | -0.062 | -0.089 | -0.116 | -0.015 | -0.077 | -0.041 | 1 | -0.060 | -0.273 | -0.049 | -0.013 | -0.026 | -0.004 | -0.032 | -0.011 |
| CLASE #... | -0.009 | -0.002 | -0.007 | -0.013 | -0.090 | -0.012 | -0.017 | -0.022 | -0.003 | -0.015 | -0.008 | -0.060 | 1 | -0.052 | -0.009 | -0.003 | -0.005 | -0.001 | -0.006 | -0.002 |
| CLASE #... | -0.041 | -0.007 | -0.030 | -0.060 | -0.410 | -0.054 | -0.077 | -0.100 | -0.013 | -0.066 | -0.036 | -0.273 | -0.052 | 1 | -0.043 | -0.012 | -0.023 | -0.003 | -0.028 | -0.010 |
| CLASE #... | -0.007 | -0.001 | -0.005 | -0.011 | -0.074 | -0.010 | -0.014 | -0.018 | -0.002 | -0.012 | -0.006 | -0.049 | -0.009 | -0.043 | 1 | -0.002 | -0.004 | -0.001 | -0.005 | -0.002 |
| CLASE #... | -0.002 | -0.000 | -0.001 | -0.003 | -0.020 | -0.003 | -0.004 | -0.005 | -0.001 | -0.003 | -0.002 | -0.013 | -0.003 | -0.012 | -0.002 | 1 | -0.001 | -0.000 | -0.001 | -0.000 |
| CLASE #... | -0.004 | -0.001 | -0.003 | -0.006 | -0.040 | -0.005 | -0.007 | -0.010 | -0.001 | -0.006 | -0.003 | -0.026 | -0.005 | -0.023 | -0.004 | -0.001 | 1 | -0.000 | -0.003 | -0.001 |

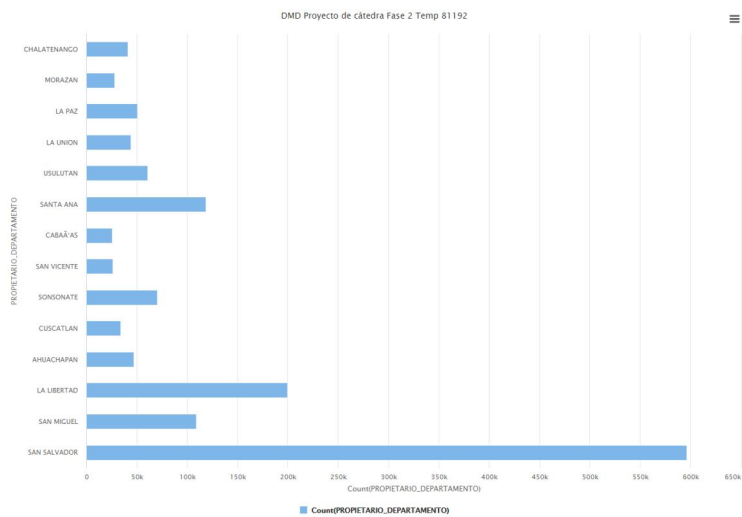
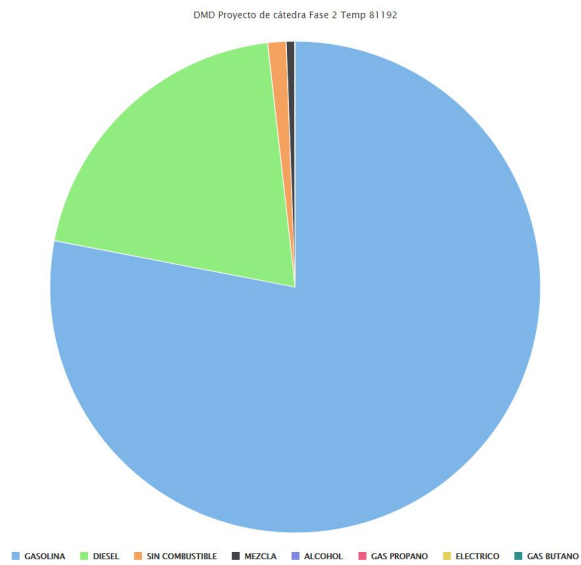
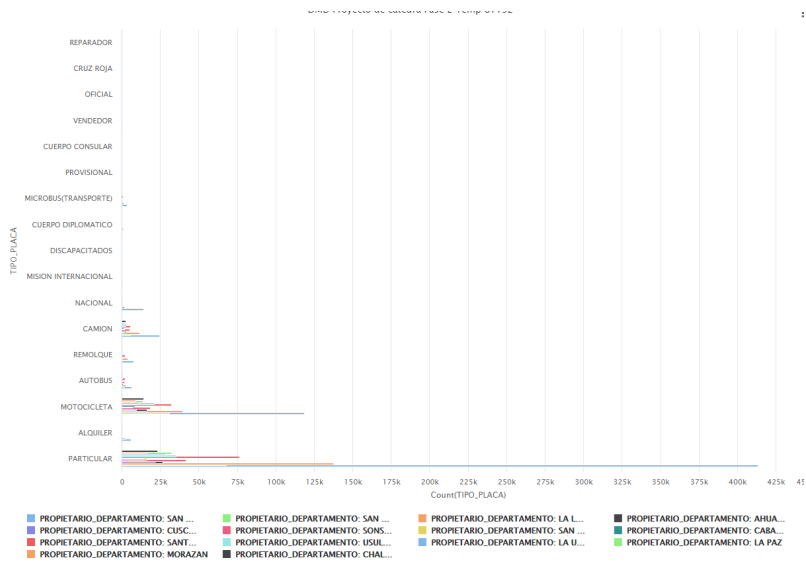
Luego de un análisis de las correlaciones (se pueden consultar estas en el proyecto en detalle) encontramos muy poca información útil, algunas cosas obvias como:

- Los autos con placas tipo “Autobus”, son de combustible Diesel.
- Los autos viejos son los que más se han dado de baja.
- La mayor concentración de vehículo de cualquier tipo (sea por combustible, por tipo de placa, por año, etc.) Están en San Salvador, no hay un dato que dispare.

Además hemos confirmado otras cosas que se mostraban los primeros datos:

- El parque vehicular está concentrado en San Salvador
- La distribución de todo tipo de vehículos es similar en todo el país (por ejemplo en todos los departamentos hay más vehículos particulares que de transporte público)
- El parque vehicular de gasolina es muy superior al de Diésel y el eléctrico es ínfimo.

Aquí podemos ver algunas gráficas que respaldan estos datos:



Conclusiones.

Francamente el análisis de los datos se torno sencillo con el programa, mostrando tendencias que nos pueden ayudar a tomar algunas decisiones como país y que demuestran muchos datos que conocemos, sea por intuición o por otros estudios que no son de parque vehicular pero que si reflejan realidades similares.

Por ejemplo algunas conclusiones de datos:

- El parque de vehículos privados es muy superior al de transporte público, de hecho si sumamos motocicletas y vehículos privados tenemos cerca de un millón trescientos mil vehículos, contra 30 mil que suman buses y micro buses, esto es una proporción más de 43 veces, lo cual nos indica muchos patrones como sociedad:
 - El transporte público, al ser ocupado por una gran proporción de la población, tiene serias deficiencias, principalmente debido a la saturación.
 - El tener un mal sistema de transporte publico hace que la gente busque a toda costa tener su propio vehículo.
 - Esto representa una buena oportunidad, un mal sistema es algo que al mejorar puede representar un incremento en la calidad de vida de la población, un nuevo sistema de transporte moderno y seguro.
- La fuerte concentración de vehículos en el área metropolitana de San Salvador (Entre los departamentos de San Salvador y La libertad se concentran la mitad de los vehículos) refleja la concentración de población y desarrollo que hay en el país. Si bien esto solo es una consecuencia, vale la pena discutir como nación la forma de hacer un desarrollo mejor distribuido en el país.
- La casi nula representación de vehículos eléctricos en el parque vehicular es una noticia relativamente mala, sin embargo puede ser una gran oportunidad. Además podemos ver que los vehículos Diesel (el cual se ha demostrado ser mas dañino para la salud que la gasolina) son relativamente pocos. Entonces se puede ocupar esto como una forma de gradualmente renovar el parque vehicular hacia vehículos eléctricos y ocupar el enorme potencial que tenemos en energía renovable (solar, geotérmica y eólica) para el desarrollo del país.
- Existen varios elementos dentro de los datos que están ausentes y que nos podrían ayudar a mejorar el análisis. Si yo fuera el consultor pediría aumentar los datos a los siguiente:
 - Género del dueño, en caso de ser persona natural, además de la división entre persona natural y jurídica. Esto nos podría brindar información valiosa sobre igualdad de género y participación de las empresas en el parque vehicular.
 - Año en que se dio de baja, para el caso de los vehículos en baja, para saber cuanto tiempo pasa entre la fabricación y la baja.