# Proyecto fase 1:
# Análisis exploratorio

María Fernanda Estrada Cornejo 14198
Christopher Kevin Sandoval García 13660
Luis Estuardo Delgado Ordoñez 17187

Minería de Datos


Universidad del Valle de Guatemala
Facultad de Ingeniería
Marzo 2020

# I. DESCRIPCIÓN DEL PROBLEMA

**Situación problemática**

El uso de un medio de transporte se ha vuelto vital en la vida de las personas, ya sea un medio público o personal. Sin embargo, debido al incremento de la violencia en el transporte público en Guatemala, muchos han optado por alternativas como Uber, taxis o tener carros propios. Ésta última alternativa representa una gran responsabilidad, ya que la persona debe escoger el automóvil adecuado para las calles en Guatemala. Pueden surgir preguntas como: ¿comprar una minivan para tener comodidad?, ¿comprar un carro pequeño por ahorro de gasolina?, ¿comprar un carro grande para evitar accidentes?, etc. Es evidente, entonces, que el tipo de un carro afectará cómo será nuestra movilidad en las calles de Guatemala.

**Problema científico**

El problema que el proyecto tratará de resolver es: ¿existe una relación entre los tipos de carros que se han importado y la cantidad de accidentes de ese año?

**Objetivos**

1. Determinar si existe una relación entre los tipos de carros que se han importado y la cantidad de accidentes de ese año.
2. Determinar la característica más importante en el conjunto de datos que describen al vehículo.
3. Determinar la mayor cantidad de información posible del vehículo involucrado en un accidente de tránsito.

# II. DESCRIPCIÓN DE LOS DATOS

El conjunto de datos posee variables que detallan la información de importación de varios vehículos. El dataset en total posee 1848193 filas con 21 variables; ofrece información como país de proveniencia, modelo del vehículo, marca, tipo de vehículo, entre otros.

Al descargar y juntar todos los datos en un dataset, se observó que había una columna extra que hacía que no correspondiera la columna con sus datos. Se eliminó la última columna y se ordenó el conjunto. Por otro lado, habían datos mal colocados, por ejemplo un año 3015 en lugar del 2015. Toda esta limpieza de datos se hizo para tener un mejor análisis exploratorio.

Por último, a continuación se indica el tipo de cada una de las variables:

- Pais.de.Proveniencia: cualitativa nominal
- Aduana.de.Ingreso: cualitativa nominal
- Fecha.de.la.Poliza: cualitativa ordinaria
- Partida.Arancelaria: cualitativa ordinaria
- Modelo.del.Vehiculo: cualitativa ordinaria
- Marca: cualitativa nominal
- Linea: cualitativa nominal
- Centimetros.Cubicos: cuantitativa discreta
- Distintivo: cualitativa nominal
- Tipo.de.Vehiculo: cualitativa nominal
- Tipo.de.Importador: cualitativa nominal
- Tipo.Combustible: cualitativa nominal
- Asientos: cuantitativa discreta
- Puertas: cuantitativa discreta
- Tonelaje: cuantitativa continua
- Valor.CIF: cuantitativa continua
- Impuesto: cuantitativa continua
- Anio: cualitativa ordinaria
- Mes: cualitativa ordinaria
- Dia: cualitativa ordinaria
- DiaSem: cualitativa nominal

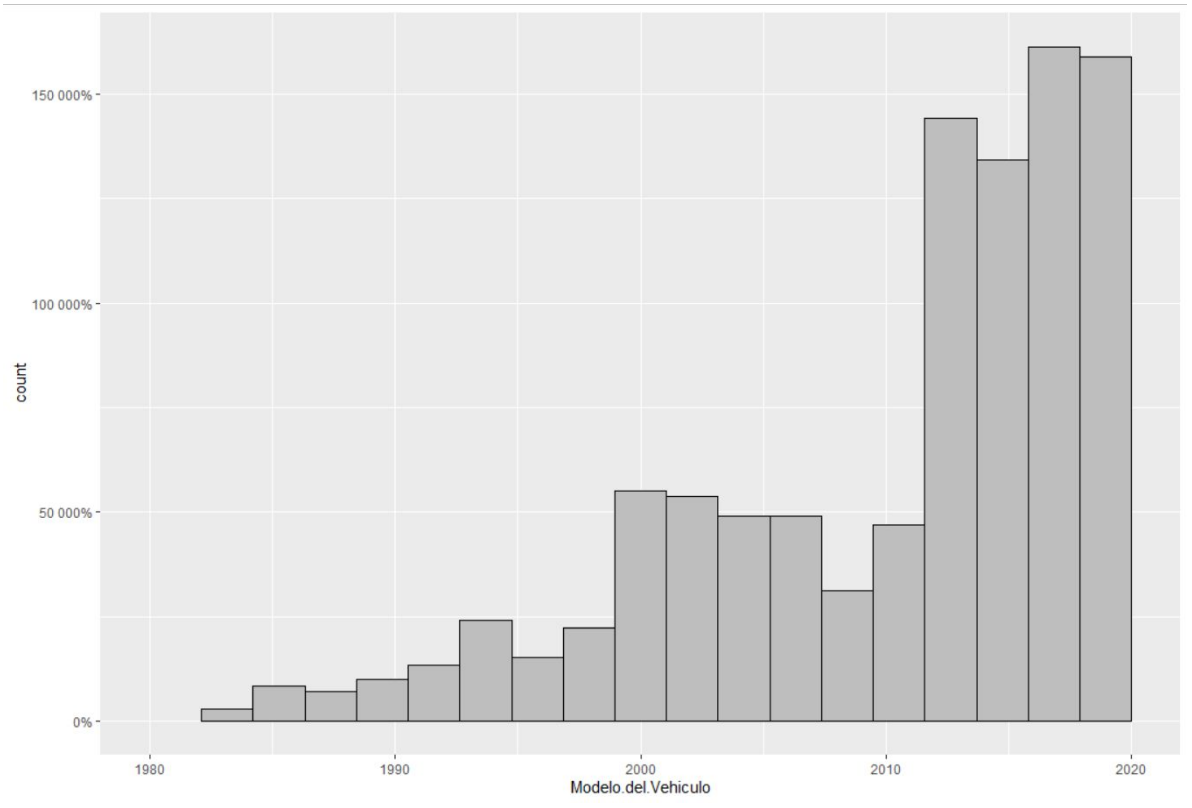
# III. ANÁLISIS EXPLORATORIO

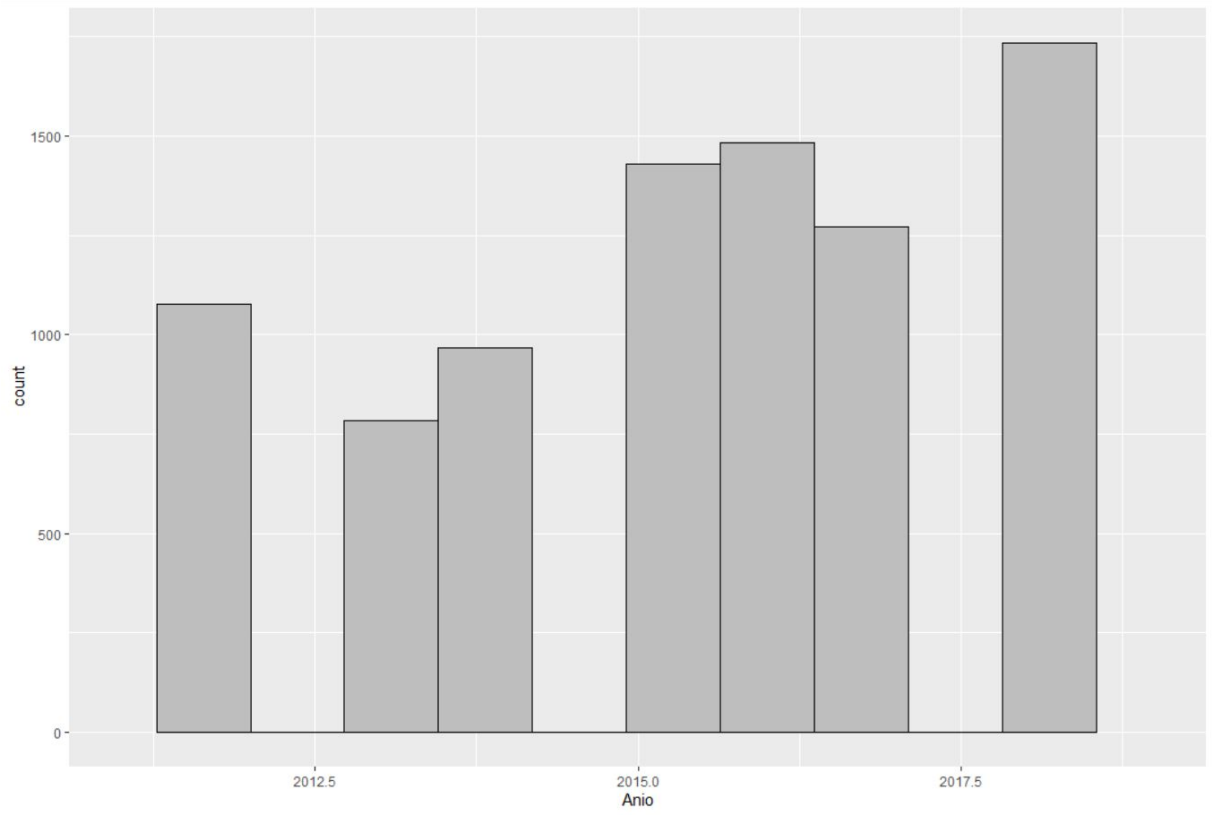
**Variables cuantitativas**

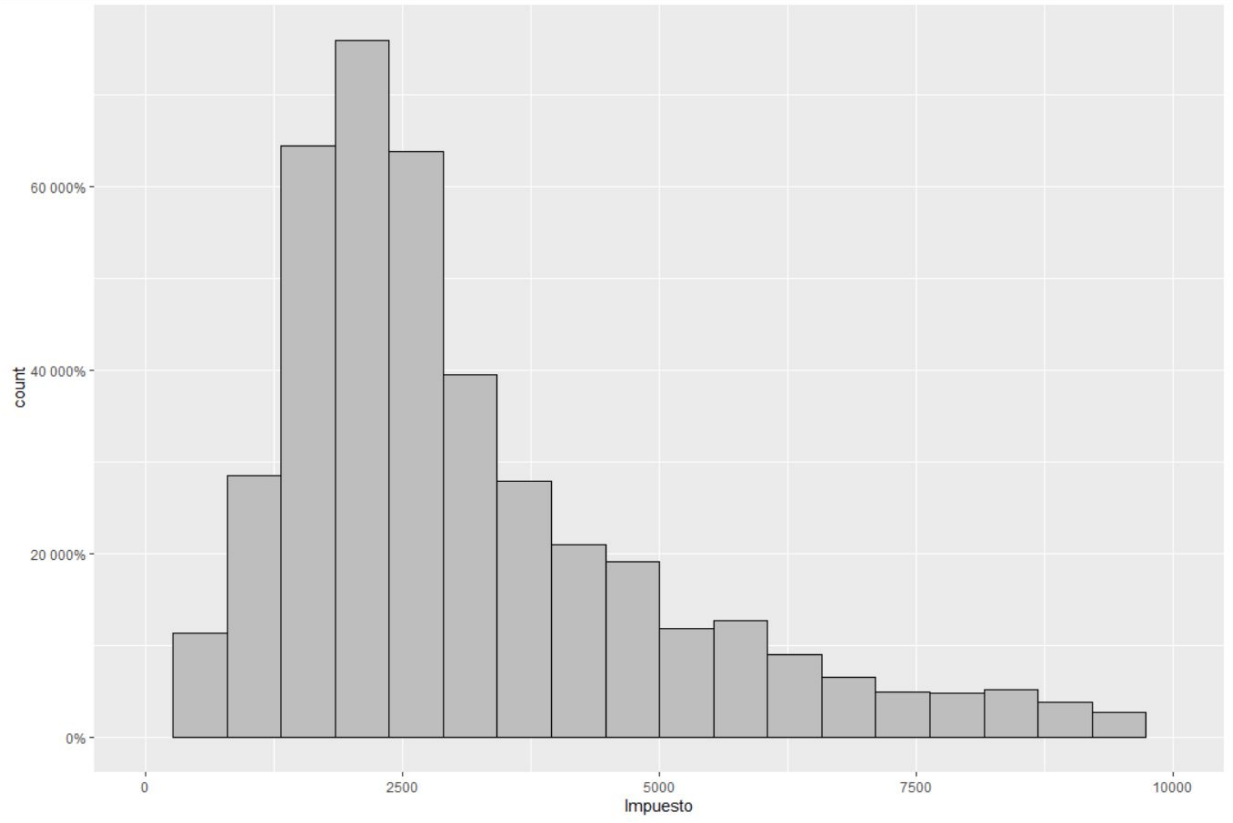
Antes de comenzar, se realizó un breve resumen de las variables cuantitativas para observar y determinar si hay casos atípicos o algún otro caso extraño. Abajo se muestran unos ejemplos de estos resúmenes.

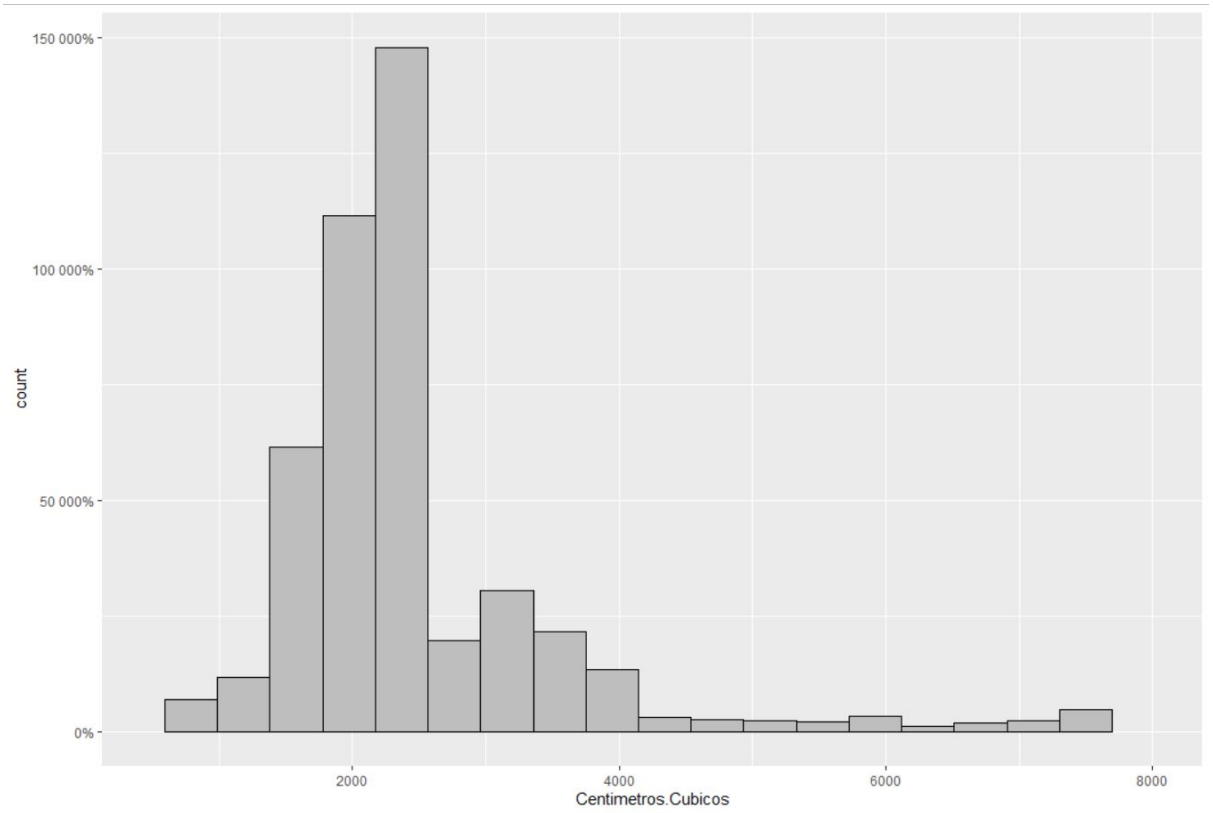
```
> summary(data_sample$Modelo.del.Vehiculo)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1958    2005    2013    2010    2016    2020

> summary(data_sample$Impuesto)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    164    2932   30111   59846   71138 4940711

> summary(data_sample$Centimetros.Cubicos)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     12     125     200    1399    2400   15000

> summary(data_sample$Anio)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2011    2013    2015    2015    2017    2019
```

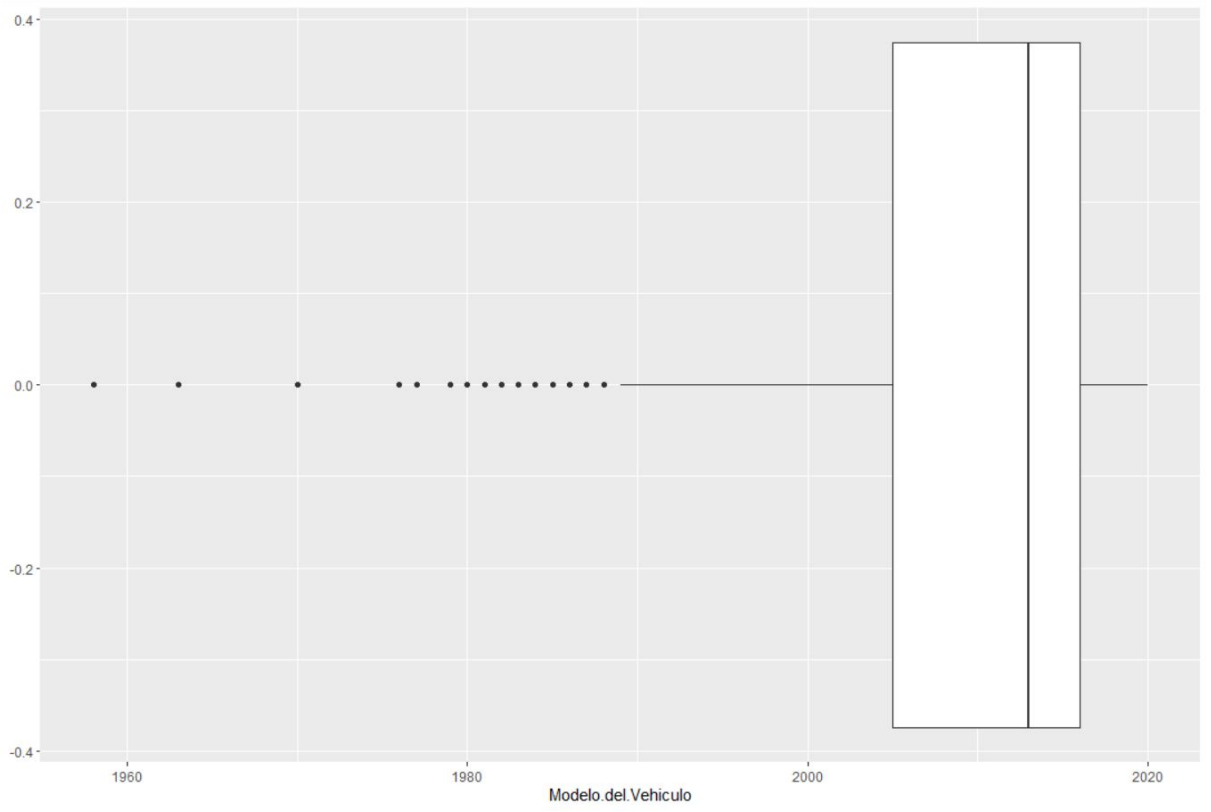
Luego, para determinar si las variables cuantitativas siguen una distribución normal, se realizaron histogramas y diagramas de caja y bigotes de cada uno. En la mayoría de variables, no se encontró una distribución normal de los datos.

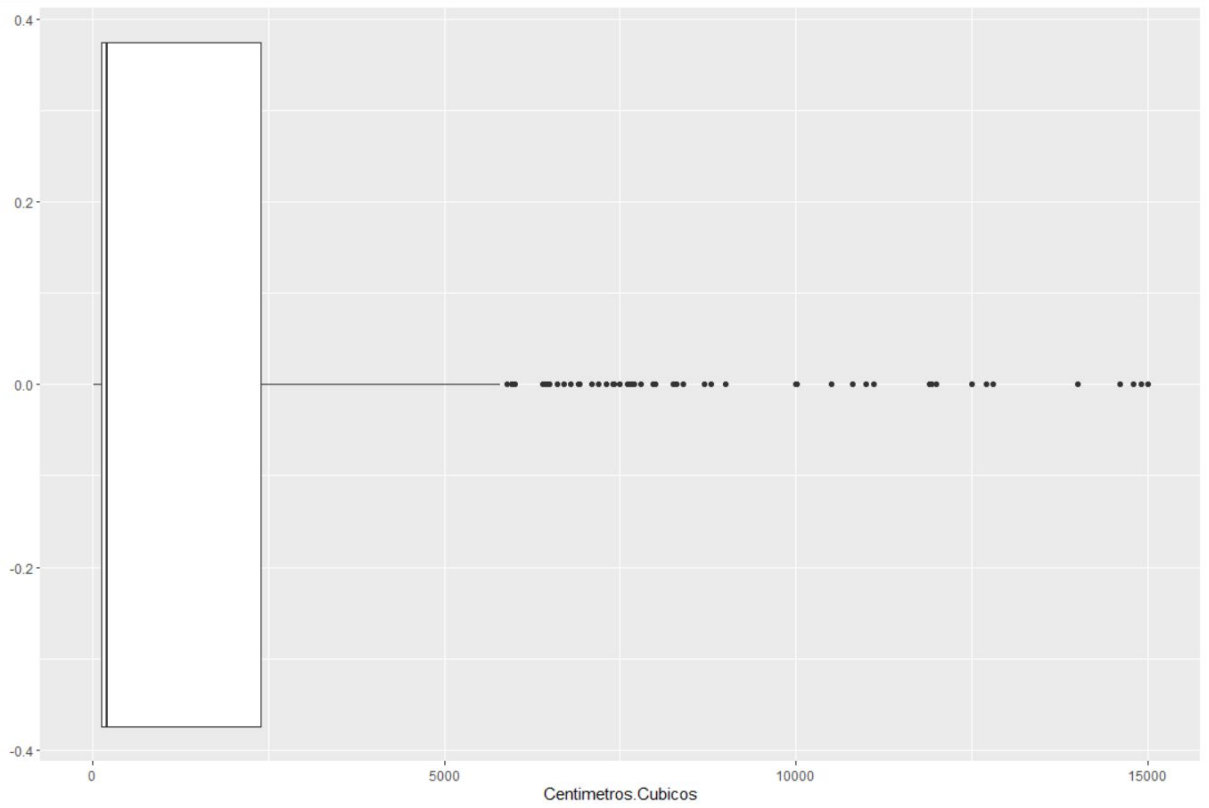


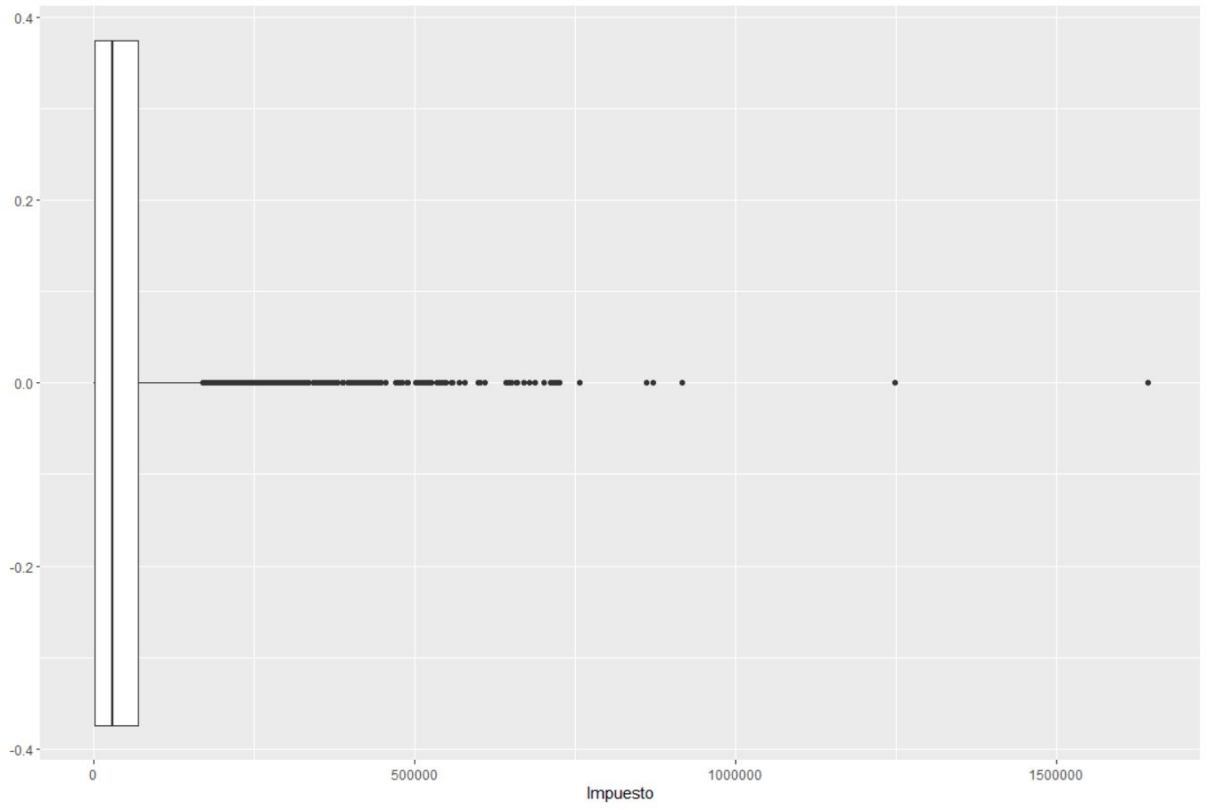

1500
1000
500
0
count
2012.5
2015.0
2017.5
Anio


60 000%
40 000%
20 000%
0%
count
0
2500
5000
7500
10000
Impuesto


150 000%
100 000%
50 000%
0%
count
2000
4000
6000
8000
Centimetros.Cubicos

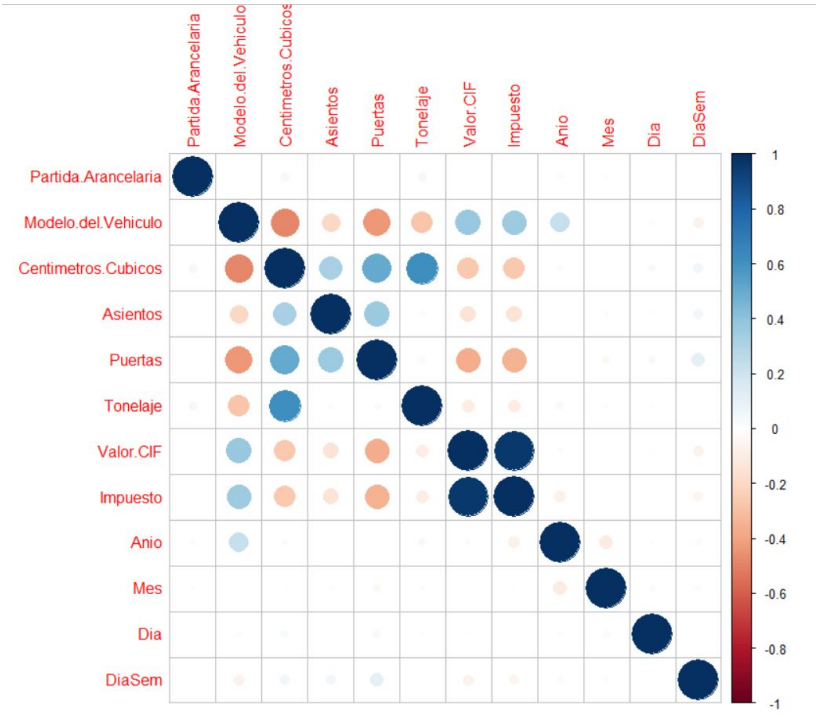

0.4
0.2
0.0
-0.2
-0.4
1960
1980
2000
2020
Modelo.del.Vehiculo




Por último, se debía determinar si existía relación alguna entre las variables. Según el siguiente gráfico, solamente existe una relación fuerte entre los centímetros

cúbicos y el tonelaje del vehículo. Las demás variables tienen relación entre 0 a -0.4.




**Variables cualitativas**

Frecuencia de los 30 países con más importaciones. Los tres países con más importaciones son China, Japón y Estados Unidos.

| Var1 | Freq |
|---|---|
| CHINA | 707511 |
| JAPON | 336924 |
| ESTADOS UNIDOS | 317258 |
| INDIA | 224197 |
| COREA DEL SUR | 75265 |
| CANADA | 44056 |
| TAILANDIA | 35700 |
| MEXICO | 31546 |
| ALEMANIA REP. FED. | 19978 |
| BRASIL | 15459 |
| REINO UNIDO | 11490 |
| TAIWAN | 7924 |
| INDONESIA | 3863 |
| ARGENTINA | 3685 |
| COLOMBIA | 1833 |
| ITALIA | 1790 |
| FRANCIA | 1649 |
| SUECIA | 1564 |
| AUSTRIA | 816 |
| HONDURAS | 815 |
| ESPANA | 778 |
| SUDAFRICA | 580 |
| BELGICA | 516 |
| HUNGRIA | 492 |
| KOREA DEL NORTE (REPUBLICA DEMOCRATICA) | 319 |
| SLOVAKIA | 275 |
| POLONIA | 265 |
| TURQUIA | 262 |
| SUIZA | 218 |
| EL SALVADOR | 108 |

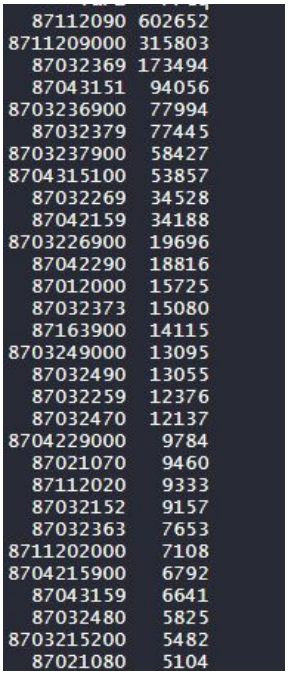
Frecuencia de la aduana donde entra el auto. Las aduanas que reciben mayor cantidad de autos son Puerto Quetzal, El Carmen y Tecún Umán.

| Var1 | Freq |
|---|---|
| PUERTO QUETZAL | 930652 |
| EL CARMEN | 263742 |
| TECUN UMAN | 200461 |
| PUERTO BARRIOS | 192144 |
| CENTRAL DE GUATEMALA | 124114 |
| SANTO TOMAS DE CASTILLA | 71531 |
| EXPRESS AEREO | 34664 |
| G8, CENTRALSA | 12483 |
| PEDRO DE ALVARADO | 6591 |
| G1, INTEGRADA | 4035 |
| G4, ALSERSA | 1765 |
| ADUANA INTEGRADA AGUA CALIENTE | 1740 |
| SAN CRISTOBAL | 1571 |
| MELCHOR DE MENCOS | 660 |
| AGUA CALIENTE | 621 |
| ADUANA INTEGRADA EL FLORIDO | 324 |
| ADUANA INTEGRADA CORINTO | 273 |
| LA MESILLA | 237 |
| VALLE NUEVO | 164 |
| G3, ALPASA | 140 |
| G5, CEALSA | 82 |
| G2, ALMINTER | 75 |
| G7, ALCORSA | 48 |
| LA ERMITA | 44 |
| EL CEIBO | 16 |
| EL FLORIDO | 14 |
| G6, ALMAGUATE | 2 |

Frecuencia de las 30 fechas donde entraron más vehículos. Se puede observar que en los últimos meses es cuando entran más vehículos.

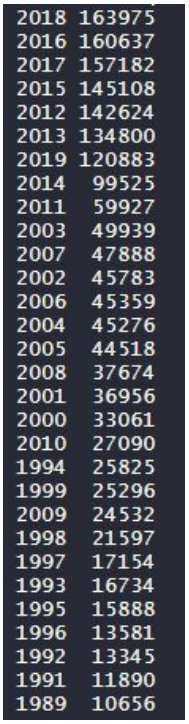
| | |
|---|---|
| 14/12/2015 | 4869 |
| 17/07/2017 | 4633 |
| 23/11/2016 | 4156 |
| 18/12/2014 | 4077 |
| 26/08/2015 | 4075 |
| 06/09/2016 | 3893 |
| 07/07/2016 | 3854 |
| 12/06/2012 | 3832 |
| 21/10/2015 | 3778 |
| 06/01/2016 | 3733 |
| 04/06/2018 | 3710 |
| 10/08/2016 | 3709 |
| 23/04/2018 | 3660 |
| 10/01/2017 | 3564 |
| 14/08/2017 | 3494 |
| 22/06/2016 | 3442 |
| 24/01/2018 | 3395 |
| 29/09/2015 | 3391 |
| 22/08/2017 | 3383 |
| 04/09/2015 | 3292 |
| 03/04/2018 | 3233 |
| 12/12/2017 | 3233 |
| 09/04/2012 | 3220 |
| 08/12/2016 | 3198 |
| 06/10/2016 | 3116 |
| 06/08/2012 | 3099 |
| 13/11/2015 | 3081 |
| 19/06/2015 | 3072 |
| 21/12/2012 | 3059 |
| 06/07/2015 | 3043 |

Frecuencia de las 30 partidas arancelarias en las que entraron más vehículos.



| | |
|---|---|
| 87112090 | 602652 |
| 8711209000 | 315803 |
| 87032369 | 173494 |
| 87043151 | 94056 |
| 8703236900 | 77994 |
| 87032379 | 77445 |
| 8703237900 | 58427 |
| 8704315100 | 53857 |
| 87032269 | 34528 |
| 87042159 | 34188 |
| 8703226900 | 19696 |
| 87042290 | 18816 |
| 87012000 | 15725 |
| 87032373 | 15080 |
| 87163900 | 14115 |
| 8703249000 | 13095 |
| 87032490 | 13055 |
| 87032259 | 12376 |
| 87032470 | 12137 |
| 8704229000 | 9784 |
| 87021070 | 9460 |
| 87112020 | 9333 |
| 87032152 | 9157 |
| 87032363 | 7653 |
| 8711202000 | 7108 |
| 8704215900 | 6792 |
| 87043159 | 6641 |
| 87032480 | 5825 |
| 8703215200 | 5482 |
| 87021080 | 5104 |

Frecuencias de los 30 modelos (año) más comunes en importaciones. Se observa que los modelos más recientes son los más importados.



| | |
|---|---|
| 2018 | 163975 |
| 2016 | 160637 |
| 2017 | 157182 |
| 2015 | 145108 |
| 2012 | 142624 |
| 2013 | 134800 |
| 2019 | 120883 |
| 2014 | 99525 |
| 2011 | 59927 |
| 2003 | 49939 |
| 2007 | 47888 |
| 2002 | 45783 |
| 2006 | 45359 |
| 2004 | 45276 |
| 2005 | 44518 |
| 2008 | 37674 |
| 2001 | 36956 |
| 2000 | 33061 |
| 2010 | 27090 |
| 1994 | 25825 |
| 1999 | 25296 |
| 2009 | 24532 |
| 1998 | 21597 |
| 1997 | 17154 |
| 1993 | 16734 |
| 1995 | 15888 |
| 1996 | 13581 |
| 1992 | 13345 |
| 1991 | 11890 |
| 1989 | 10656 |

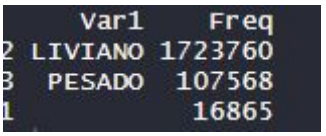
Frecuencia de las 30 marcas más importadas de vehículos. Las tres marcas más importadas son Toyota, Honda y Suzuki. Las marcas menos conocidas o de mayor precio son las menos importadas.

| | |
|---|---|
| TOYOTA | 291595 |
| HONDA | 285203 |
| SUZUKI | 209079 |
| ITALIKA | 131117 |
| MAZDA | 95315 |
| BAJAJ | 86128 |
| FREEDOM | 81090 |
| YAMAHA | 59915 |
| SERPENTO | 49920 |
| HYUNDAI | 45215 |
| NISSAN | 39007 |
| KIA | 37393 |
| MITSUBISHI | 34358 |
| MOVESA | 28368 |
| HERO | 26437 |
| CHEVROLET | 21520 |
| FORD | 20761 |
| ASIA HERO | 19571 |
| FREIGHTLINER | 15748 |
| ISUZU | 15323 |
| INTERNATIONAL | 14792 |
| TVS | 14774 |
| BMW | 13986 |
| VOLKSWAGEN | 12648 |
| HAOJUE | 11064 |
| HINO | 9875 |
| GENESIS | 9520 |
| UM | 9067 |
| JIALING | 8566 |
| KYMCO | 7149 |

Frecuencia de las 30 líneas de vehículos más importadas en el periodo.

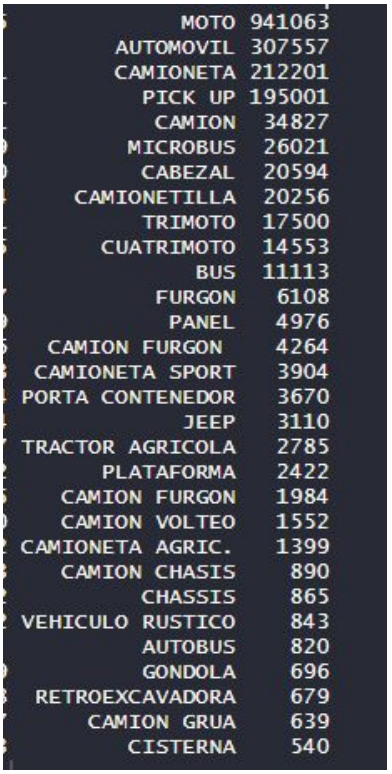
| | |
|---|---|
| GN125F | 63735 |
| CGL125 | 36192 |
| GN125H | 35151 |
| XR150L | 24219 |
| FIRE125 | 22978 |
| AN125HK | 20812 |
| YARIS | 20270 |
| AX100 | 19948 |
| PULSAR 135 LS | 19152 |
| CIVIC LX | 16433 |
| CS125 | 15210 |
| EN125-2A | 14634 |
| NAVI 110 | 11752 |
| CIVIC EX | 11392 |
| SUPER LIFE | 11162 |
| 3 | 10945 |
| GTK125 | 10444 |
| COROLLA LE | 10140 |
| COROLLA CE | 9841 |
| V-MEN | 9368 |
| YBR125G | 9305 |
| COROLLA S | 8890 |
| HI LUX | 8879 |
| CR-V EX 4WD | 8267 |
| CB1 | 8244 |
| 4X2 STD | 7533 |
| CIVIC | 7191 |
| PROTEGE LX | 7103 |
| FIRE 150 | 7033 |
| 4X4 DLX | 6857 |

Frecuencia del distintivo del vehículo. Se importan más vehículos livianos que pesados.



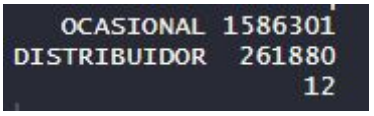
| | Var1 | Freq |
|---|---|---|
| 2 | LIVIANO | 1723760 |
| 3 | PESADO | 107568 |
| 1 | | 16865 |

Frecuencia de los 30 tipos de vehículos que son más importados. Los tipos más importados son motos y automóviles, ya que son de uso más común.



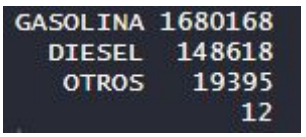
| | |
|---|---|
| MOTO | 941063 |
| AUTOMOVIL | 307557 |
| CAMIONETA | 212201 |
| PICK UP | 195001 |
| CAMION | 34827 |
| MICROBUS | 26021 |
| CABEZAL | 20594 |
| CAMIONETILLA | 20256 |
| TRIMOTO | 17500 |
| CUATRIMOTO | 14553 |
| BUS | 11113 |
| FURGON | 6108 |
| PANEL | 4976 |
| CAMION FURGON | 4264 |
| CAMIONETA SPORT | 3904 |
| PORTA CONTENEDOR | 3670 |
| JEEP | 3110 |
| TRACTOR AGRICOLA | 2785 |
| PLATAFORMA | 2422 |
| CAMION FURGON | 1984 |
| CAMION VOLTEO | 1552 |
| CAMIONETA AGRIC. | 1399 |
| CAMION CHASIS | 890 |
| CHASSIS | 865 |
| VEHICULO RUSTICO | 843 |
| AUTOBUS | 820 |
| GONDOLA | 696 |
| RETROEXCAVADORA | 679 |
| CAMION GRUA | 639 |
| CISTERNA | 540 |

Frecuencia del tipo de importador. Es más común que el importador sea ocasional.



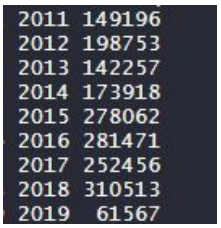
| | | |
|---|---|---|
| | OCASIONAL | 1586301 |
| | DISTRIBUIDOR | 261880 |
| | | 12 |

Frecuencia del tipo de combustible que utilizan los vehículos importados. Como la mayor cantidad de tipos de vehículos importados son motos y automóviles, la gasolina también debe ser el mayor tipo de combustible.

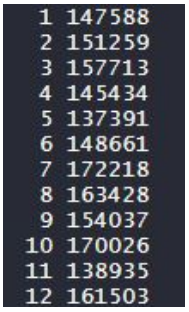


| | | |
|---|---|---|
| | GASOLINA | 1680168 |
| | DIESEL | 148618 |
| | OTROS | 19395 |
| | | 12 |

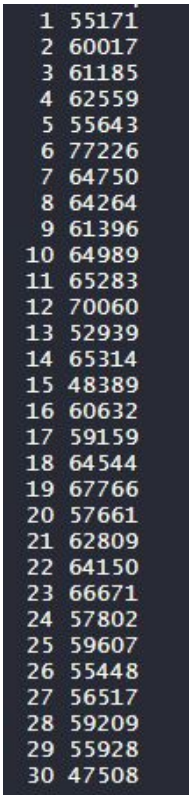
Frecuencia de la cantidad de autos importados por año. Del año 2013 al 2018 se ve un incremento considerable en la cantidad de vehículos importados.




Frecuencia de vehículos importados en cada mes de todos los años. Los meses con mayor cantidad de importaciones fueron Julio, Agosto, Octubre y Diciembre.




Frecuencia de vehículos importados en cada día de todos los meses en todos los años.

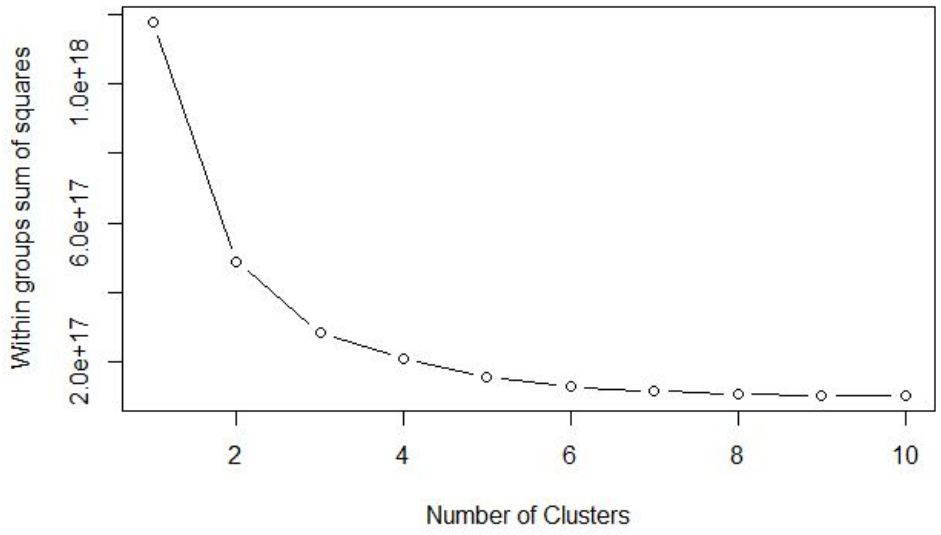


Frecuencia de la cantidad de vehículos importados por día de la semana en todos los años.

```
1  32916
2 364187
3 379232
4 330526
5 292113
6 337251
7 111968
```

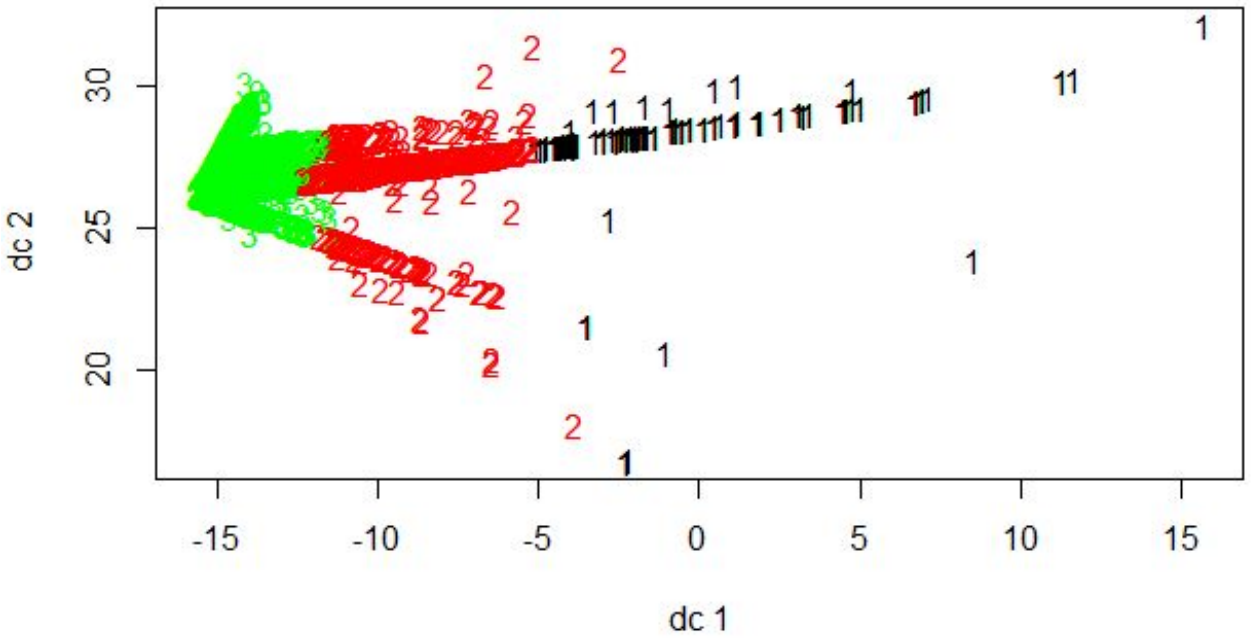
**Clusters**

Primero, se determina el mejor número de grupos para hacer clustering. El resultado de esta operación indica que el mejor número de grupos es 3. Las variables más significativas que se dejaron para el cluster son: "Modelo.del.Vehiculo", "Centimetros.Cubicos", "Asientos", "Puertas", "Valor.CIF", "Impuesto" y “Anio”.




Se utilizará el agrupamiento por Kmeans con **3** clusters, ya que es el más sencillo y el más eficaz. Para verificar la calidad del agrupamiento generado, se utilizó el método de la silueta para Kmeans. Se obtuvo una silueta de **0.761**, indicando que la calidad de los grupos es bastante buena.

```
> silkm<-silhouette(km$cluster,dist(data_sample))
> mean(silkm[,3])
[1] 0.7611205
```




En el grupo 1 se pueden encontrar las siguientes características: modelos de vehículo recientes; elevados centímetros cúbicos; dos asientos en promedio; de 0 a 1 puertas; valor CIF elevado; impuesto elevado; año de ingreso reciente.

```
> summary(cluster1)
 Modelo.del.Vehiculo Centimetros.Cubicos    Asientos        Puertas
 Min.   :2012        Min.   :  100.0     Min.   :1.000   Min.   :0.0000
 1st Qu.:2015        1st Qu.:  135.0     1st Qu.:2.000   1st Qu.:0.0000
 Median :2016        Median :  150.0     Median :2.000   Median :0.0000
 Mean   :2016        Mean   :  483.5     Mean   :2.365   Mean   :0.2656
 3rd Qu.:2018        3rd Qu.:  150.0     3rd Qu.:2.000   3rd Qu.:0.0000
 Max.   :2019        Max.   :15200.0     Max.   :7.000   Max.   :5.0000
   Valor.CIF          Impuesto           Anio          Class
 Min.   :2822030   Min.   : 338644   Min.   :2011   Min.   :1
 1st Qu.:3123445   1st Qu.: 396243   1st Qu.:2015   1st Qu.:1
 Median :3762131   Median : 474708   Median :2016   Median :1
 Mean   :4080758   Mean   : 514190   Mean   :2016   Mean   :1
 3rd Qu.:4660636   3rd Qu.: 604225   3rd Qu.:2017   3rd Qu.:1
 Max.   :8135403   Max.   :1091566   Max.   :2019   Max.   :1
```

En el grupo 2 se pueden encontrar las siguientes características: modelos de vehículo recientes; medianos centímetros cúbicos; dos asientos en promedio; de 0 a 1 puertas; valor CIF intermedio; impuesto intermedio; año de ingreso reciente.

```
> summary(cluster2)
 Modelo.del.Vehiculo Centimetros.Cubicos    Asientos         Puertas
 Min.   :1900        Min.   :  100.0     Min.   : 1.000   Min.   :0.0000
 1st Qu.:2013        1st Qu.:  125.0     1st Qu.: 2.000   1st Qu.:0.0000
 Median :2015        Median :  135.0     Median : 2.000   Median :0.0000
 Mean   :2015        Mean   :  464.3     Mean   : 2.356   Mean   :0.4378
 3rd Qu.:2018        3rd Qu.:  200.0     3rd Qu.: 2.000   3rd Qu.:0.0000
 Max.   :2020        Max.   :12800.0     Max.   :15.000   Max.   :5.0000
   Valor.CIF          Impuesto          Anio          Class
 Min.   : 870097   Min.   :106332   Min.   :2011   Min.   :2
 1st Qu.:1155781   1st Qu.:144314   1st Qu.:2013   1st Qu.:2
 Median :1456181   Median :187133   Median :2015   Median :2
 Mean   :1551262   Mean   :208591   Mean   :2015   Mean   :2
 3rd Qu.:1840044   3rd Qu.:261131   3rd Qu.:2017   3rd Qu.:2
 Max.   :2793077   Max.   :915935   Max.   :2019   Max.   :2
```

En el grupo 3 se pueden encontrar las siguientes características: modelos de vehículo viejos; pocos centímetros cúbicos; tres asientos en promedio; de 1 a 2 puertas; valor CIF bajo; impuesto bajo; año de ingreso reciente.

```
> summary(cluster3)
 Modelo.del.Vehiculo Centimetros.Cubicos    Asientos         Puertas
 Min.   :1966        Min.   :    0       Min.   : 0.000   Min.   :0.000
 1st Qu.:2004        1st Qu.:  125       1st Qu.: 2.000   1st Qu.:0.000
 Median :2012        Median : 1497       Median : 2.000   Median :2.000
 Mean   :2009        Mean   : 1535       Mean   : 3.721   Mean   :1.818
 3rd Qu.:2016        3rd Qu.: 2400       3rd Qu.: 5.000   3rd Qu.:4.000
 Max.   :2020        Max.   :62000       Max.   :77.000   Max.   :5.000
   Valor.CIF          Impuesto           Anio          Class
 Min.   :  1970   Min.   :   236.4   Min.   :2011   Min.   :3
 1st Qu.: 20067   1st Qu.:  2623.4   1st Qu.:2013   1st Qu.:3
 Median : 88404   Median : 12030.6   Median :2015   Median :3
 Mean   :212573   Mean   : 28833.1   Mean   :2015   Mean   :3
 3rd Qu.:386318   3rd Qu.: 49372.0   3rd Qu.:2017   3rd Qu.:3
 Max.   :880965   Max.   :293442.7   Max.   :2019   Max.   :3
```

# IV. HALLAZGOS Y CONCLUSIONES

- Las Motos son el tipo de vehículo más importado al país.
- La gasolina es el principal combustible utilizado por los vehículos importados representando 90.91% de los vehículos importados.
- Las marcas de automóviles más importadas al país son de origen Japonés.
- El 93.27% de los vehículos importados se clasifican en la diferenciación de Liviano.
- La aduana más utilizada es la aduana de Puerto Quetzal recibiendo 50.92% de los vehículos que importados al país.
- Los tres países que proveen la mayor cantidad de vehículos importados a Guatemala son China (38.28%), Japón (18.22%) y Estados Unidos (17.17%).
- Las variables cuantitativas del dataset no tienen relación entre ellas, a excepción de los centímetros cúbicos y el tonelaje del vehículo.
- Los grupos obtenidos por el clustering, se dividen principalmente por su valor CIF y el impuesto pagado por el vehículo.
- El primer grupo de clustering es de valor CIF e impuesto altos; el segundo es de valor CIF e impuesto intermedio; el tercer grupo es de valor CIF e impuesto bajo.
- Lo siguientes pasos a tomar son: investigar los algoritmos de aprendizaje de máquinas más útiles para el proyecto y determinar si existe una relación entre el tipo de vehículo y la cantidad de accidentes por año.