Desafío Derco para proceso postulación Data Scientist

Por Diego Rivera

Santiago Enero, 2020

Tabla de Contenido

Contexto del Desafío

El desafío a resolver consiste en una *Prueba conocimientos analíticos.*

Donde se está estudiando el fenómeno de preferencia y recambio de automóviles en Chile.
El link adjunto (https://www.dropbox.com/s/52p5cthsj7otaw2/bbdd%20prueba%20corp.7z?dl=0) contiene un set de datos anonimizados de tenencia de vehículos histórica.

Sample:

```
PATENTE;MARCA;MODELO;AÑO;ID_CLIENTE;COMUNA;REGION;SEXO;ACTIVIDAD;TASACION;FEC_TRANSFERENC
IA;COLOR2;EDAD;VIGENCIA
393A4B4C-085;TOYOTA;RAV 4;2015;50838335;TEMUCO;9;M;NULL;11947500;NULL;BLANCO;NULL;N
B1F3DB7E-F67;MAZDA;CX 5;2016;46322649;VILLA ALEMANA;5;M;NULL;NULL;NULL;NEGRO;38;S
F6DBB2E6-A76;GREAT WALL;HAVAL NEW H3 2.0;2014;36226810;ANTOFAGASTA;DE
ANTOFAGASTA;NULL;NULL;NULL;NULL;NEGRO;49;S
EBCF63CD-99D;JEEP;COMPASS SPORT 2.4;2013;43482783;NULL;NULL;NULL;NULL;NULL;20160418;GRIS;40;S
FF38B368-B0F;SUBARU;FORESTER 2.0;2017;25657273;NULL;NULL;NULL;NULL;NULL;NULL;GRIS;62;S
BC985B84-711;KIA MOTORS;NEW SORENTO EX 2.2;2013;14210133;CONCON;5;NULL;NULL;0;NULL;BLANCO;64;S
D88B232D-2CB;HONDA;PILOT EXL 4X4 3.5 AUT;2014;21198752;RANCAGUA;DEL LIBERTADOR BERNARDO
OHIGGINS;NULL;NULL;NULL;NULL;CAFE;63;S
AC1A5003-5B6;KIA MOTORS;NEW CARENS LX 1.7;2014;24699069;PROVIDENCIA;METROPOLITANA DE
SANTIAGO;NULL;NULL;NULL;20141205;BLANCO;61;N
A9F4BE58-817;DODGE;DURANGO SLT 4X4 5.7;2011;32883814;QUILICURA;METROPOLITANA DE
SANTIAGO;;0;0;20120215;BLANCO;50;N
CA31F279-2F4;MERCEDES BENZ;ML350 BLUE TEC;2014;23348761;LOS ANGELES;DEL BIO
BIO;NULL;NULL;NULL;20140801;GRIS;56;N
```

Objetivo: Aplicando técnicas de ETL, enriquecimiento, exploración, descubrimiento y analítica para explicar y caracterizar los fenómenos.

El conjunto de datos presenta las siguientes características:

Variable	Descripción	Tipo de Dato	Datos Nulos
PATENTE	Placa patente del vehículo	Categórico	0
MARCA	Fabricante del vehículo	Categórico	0
MODELO	Versión del vehículo	Categórico	0
AÑO	Año de fabricación del modelo correspondiente	Entero	0
ID_CLIENTE	Identificador único de cliente	Entero	0
COMUNA	Comuna de la sucursal de la transacción	Categórico	0
REGION	Región de la sucursal de la transacción	Categórico	0

SEXO	Género del cliente	Categórico	0
ACTIVIDAD	Actividad del cliente	Entero	809190
TASACION	Valor de tasación del vehículo	Categórico	0
FEC_TRANSFERENCIA	Fecha de transferencia del vehículo al cliente	Entero	665817
COLOR2	Color del vehículo	Categórico	0
EDAD	Edad del cliente	Entero	283479
VIGENCIA	Vigencia de la transacción del vehículo	Categórico	0

Tabla 1: Descripción de variables (atributos).

Como se observa en la Tabla 1, en la descripción de variables o atributos se aprecian una gran cantidad de datos nulos para la variable ACTIVIDAD, correspondiente al 80.4% de los registro de ese tipo, por lo cual se pondrá atención en el tratamiento e impacto de la propiedad explicativa de esta variable para el objetivo del desafío.

Profundizando la observación de la estructura y resumen de los atributos del set de datos, se pudo rectificar que el atributo "ACTIVIDAD" no aporta dato relevante alguno, ya que cuenta con min, máx, media, mediana y 1er y 3er cuartil igual a cero y el resto de valores son NA, por lo cual será eliminada al ser de nulo aporte estadístico y explicativo, observación posible de validar en la Tabla 2. Este último hecho resulta por decirlo menos llamativo ya que el tipo de dato que debiese entregar la variable actividad debe ser *categórica* y no numérica, en este caso *entero*. Lo mismo ocurre para la variable TASACION, donde el dato debe ser *numérico* al tratarse de valores monetarios y no *categóricos* como entrega el set de datos.

Debido a lo anterior, esta última variable, TASACION, será evaluada al final.

	Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	**NA's**
AÑO	2010	2012	2013	2014	2015	2018	0
ID_CLIENTE	0	29023307	40627585	56500043	50742038	329820688	0
ACTIVIDAD	0	0	0	0	0	0	809190
FEC_TRANSFERENCIA	20091015	20140704	20151202	20151804	20170116	20180111	665817
EDAD	3.00	39.00	48.00	48.93	58.00	134.00	283479

Tabla 2: Estadísticos de Resumen.

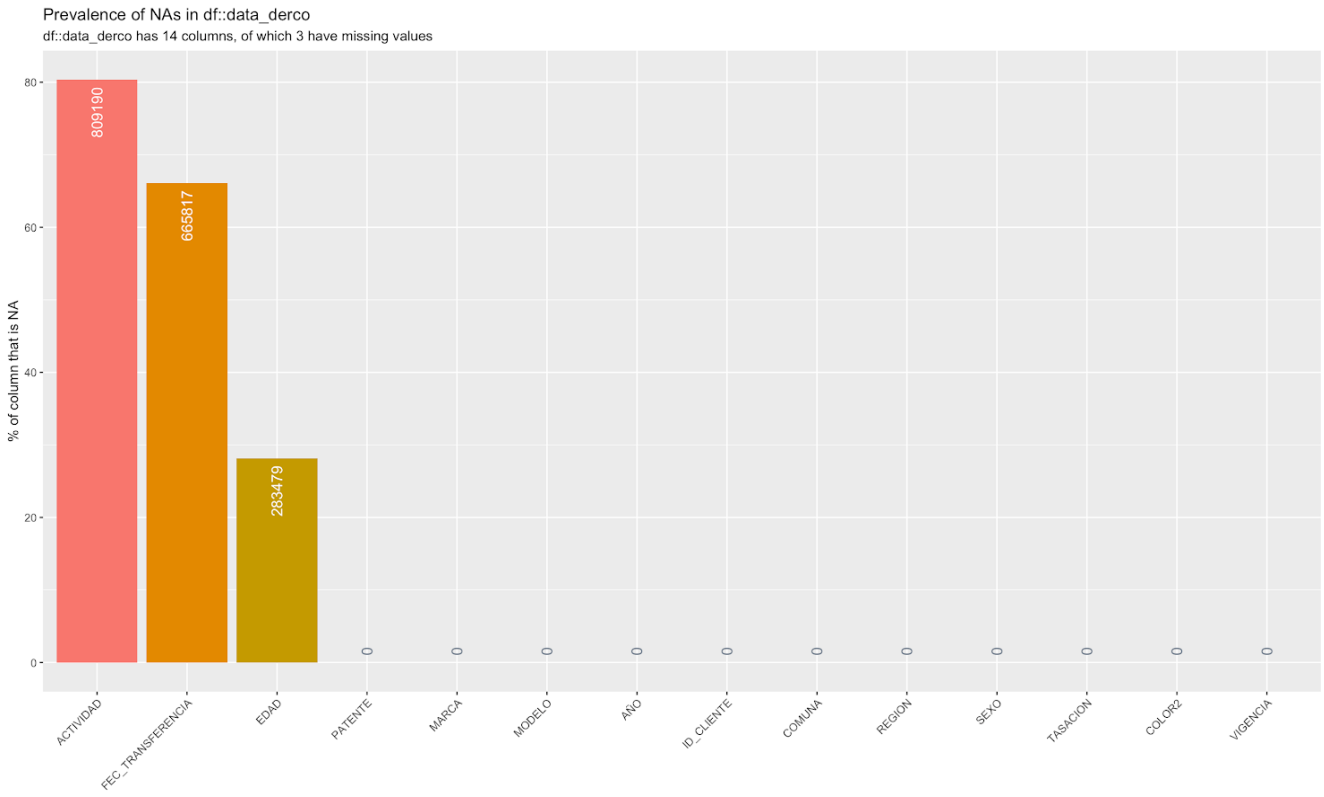

Gráfico 1: Prevalencia de NA's en set de datos de estudio.

Aunque en Gráfico 1 se puede observar que la variable TASACION no presentan valores faltantes, esto no necesariamente es cierto. Dada la inconsistencia del tipo de dato en el que fue presentado, al momento de corregirlo habrá que ver qué es lo pasa al momento de corregir el tipo de dato.

En función de esta información, y considerando los atributos que corresponden a variables categóricas mencionados anteriormente, se decide realizar una inspección de los atributos uno a uno.

Tratamiento de Variables

Primero me enfocaré en el tratamiento de los valores NA.

"Los patrones de datos faltantes representan relaciones matemáticas genéricas entre los datos observados y los ausentes.

Clasificación de los datos perdidos:

***MCAR** (Missing Completely At Random): La probabilidad de que una respuesta a una variable sea dato faltante es independiente tanto del valor de esta variable como del valor de otras variables del conjunto de datos.*

MAR *(Missing At Random): La probabilidad de que una respuesta sea dato faltante es independiente de los valores de la misma variable pero es dependiente de los valores de otras variables del conjunto de datos.*

NMAR *(Not Missing At Random): La probabilidad de que una respuesta a una variable sea dato faltante es dependiente de los valores de la variable."* [1]

Asumiendo FEC_TRANSFERENCIA y EDAD como MCAR (Missing Completely At Random), además de ser variables numéricas enteras, se realizará una imputación por la media para cada variable. Donde se obtuvo el siguiente resultado:

	DATA ORIGINAL			**DATA IMPUTADA**	
	FEC_TRANSFERENCIA	**EDAD**		**FEC_TRANSFERENCIA**	**EDAD**
Min.	20091015	3.00		20091015	3.00
1st Qu.	20140704	39.00		20151804	43.00
Median	20151202	48.00		20151804	48.93
Mean	20151804	48.93		20151804	48.93
3rd Qu.	20170116	58.00		20151804	54.00
Max.	20180111	134.00		20180111	134.00
NA's	665817	283479		-	-

Tabla 3: Estadísticos de Resumen comparativos para las variables FEC_TRANSFERENCIA y EDAD previo y posterior a la imputación por la media de datos realizada.

[1] http://rpubs.com/ydmarinb/429757. **Imputación de datos -** Yubar Daniel Marín Benjumea.

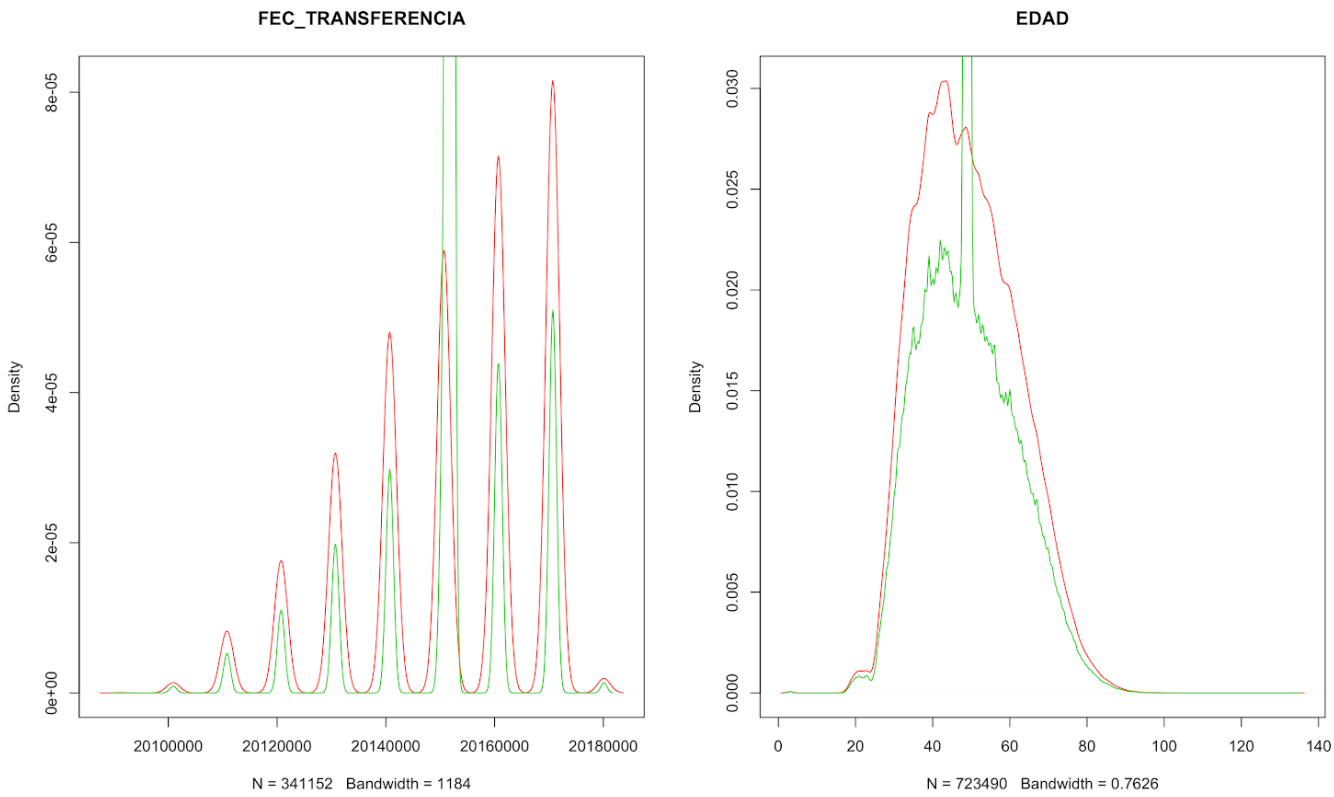

Gráfico 2: Distribución de densidad para la distribución de las variables FEC_TRANSFERENCIA y EDAD previo y posterior a la imputación por la media de datos realizada. En color rojo distribución de datos de variables originales, en color verde de los datos imputados.

Continuando revisión para resto de los datos. PATENTE, AÑO, VIGENCIA se encuentran con datos completos y coherentes.

Se pasarán los campos en blanco para MARCA, MODELO a "OTRO". Dada la funcionalidad del atributo ID_CLIENTE, que es de identificación de cliente único, se cambiará el tipo de dato de numérico a factor y se reemplazarán los valores 0 por "OTRO". De momento la principal funcionalidad de ID_CLIENTE será identificar el comportamiento de recambio de vehículos a nivel de cada cliente.

Para la variable dicotómica de SEXO se aplicará la misma lógica anterior para los campos sin información, primero pasando el tipo de dato factor a character, revisando que no haya espacios en blanco en los campos vacíos, para luego reemplazar éstos y posibles NA a “OTRO”, para luego retornar el tipo de dato a factor.

Para el atributo COMUNA, los campos en blanco pasarán a "SIN COMUNA" además de los valores "\xa5U\xa5OA", y corregir los siguientes valores para las siguientes comunas, en base a verificaciones en internet, aunque lo ideal sería seguir patrones de acuerdo a los requerimientos del solicitante, en base a lo anterior se propone lo siguiente:

- Revisando posibles casos de alcances de nombre de comunas de distintas regiones.
    - ver regiones entre PLACILLA y PLACILLA DE PENUELAS y PENUELAS.
    - ver a cual comuna corresponden SAN JOSE, SAN JOSE DE M, y MARIQUINA en base a la región.
    - ver a cual comuna corresponden SAN PEDRO, SAN PEDRO DE, en base a la región.

El resto de los alcances de nombres de comunas que fueron chequeados, se encuentra documentado como comentarios en el archivo adjunto “Derco.R”, que contiene el script de desarrollo del caso.

Para la variable REGION, se tomó la misma lógica bajo el siguiente marco:

- para 01, 1, DE TARAPACA -> I DE TARAPACA
- para 02, 2, DE ANTOFAGASTA -> II DE ANTOFAGASTA
- para 03, 3, DE ATACAMA -> III DE ATACAMA
- para 04, 4, DE COQUIMBO -> IV DE COQUIMBO
- para 05, 5, DE VALPARAISO -> V DE VALPARAISO
- para 06, 6, DEL LIBERTADOR BERNARDO OHIGGINS -> VI DEL LIBERTADOR GENERAL BERNARDO OHIGGINS
- para 07, 7, DEL MAULE -> VII DEL MAULE
- para 08, 8, DEL BIO BIO -> VIII DEL BIO BIO
- para 09, 9, DE LA ARAUCANIA -> IX DE LA ARAUCANIA
- para 10, DE LOS LAGOS -> X DE LOS LAGOS
- para 11, AYSEN DEL GENERAL CARLOS IBANEZ -> XI DE AYSEN DEL GENERAL CARLOS IBANEZ DEL CAMPO
- para 12, DE MAGALLANES Y ANTARTICA CHILENA -> XII DE MAGALLANES Y ANTARTICA CHILENA
- para 13, METROPOLITANA DE SANTIAGO -> METROPOLITANA DE SANTIAGO
- para 14, DE LOS RIOS -> XIV DE LOS RIOS
- para 15, DE ARICA y PARINACOTA -> XV DE ARICA y PARINACOTA

Teniendo como resultado resultado la correcta categorización de las regiones.

```
                                  I DE TARAPACA                                  II DE ANTOFAGASTA
                                           6197                                           29254
                                 III DE ATACAMA                                   IV DE COQUIMBO
                                           9995                                           18128
                              IX DE LA ARAUCANIA                       METROPOLITANA DE SANTIAGO
                                          16107                                          302218
                                      SIN REGION                                V DE VALPARAISO
                                          465775                                         48579
VI DEL LIBERTADOR GENERAL BERNARDO OHIGGINS                                    VII DEL MAULE
                                          19009                                          15544
                                 VIII DEL BIO BIO                                  X DE LOS LAGOS
                                          45589                                          14855
XI DE AYSEN DEL GENERAL CARLOS IBANEZ DEL CAMPO          XII DE MAGALLANES Y ANTARTICA CHILENA
                                           2007                                           5868
                                 XIV DE LOS RIOS                          XV DE ARICA y PARINACOTA
                                           5459                                           2385
```

Ilustración 1: Respuesta de consola de tabla de variable REGION.

Dada la gran variedad de colores y de subcategorías, para facilitar un análisis con el atributo COLOR2 se reasignan a:

AMARILLO, AZUL, BEIGE, BLANCO, CAFE, CELESTE, DORADO, GRIS, NARANJO, NEGRO, PLATEADO, ROJO, ROSADO, VERDE, VIOLETA

Teniendo como resultado:

```
AMARILLO     AZUL    BEIGE   BLANCO     CAFE  CELESTE   DORADO     GRIS  NARANJO    NEGRO     OTRO PLATEADO     ROJO   ROSADO
     882    47373    18126   223317    11516     4108    11691   248942     4997   129004    12569   219043    66474        2
   VERDE  VIOLETA
    8147      778
```

Ilustración 2: Respuesta de consola de tabla de variable COLOR2.

Finalmente el atributo TASACIÓN corresponde a la tasación fiscal del vehículo que permite entre otras cosas estimar lo que cuesta; es decir, el valor del vehículo, por ejemplo para fines tributarios, conocer el valor del permiso de circulación de un auto, calcular el valor del impuesto a la transferencia en caso de que sea menor que el precio del auto.

Dado lo anterior este atributo se trabajará como tipo numérico, obteniendo el siguiente resultado:

```
    TASACION
Min.   :       0
1st Qu.:       0
Median :       0
Mean   : 4958653
3rd Qu.: 9110000
Max.   :91880000
NA's   :495391
```

Ilustración 3: Respuesta de consola del resumen estadístico de la variable TASACION.

Para el caso de la variable TASACION, se decidió eliminar los valores NA, principalmente por la gran variabilidad de estos ya que resultaría poco reprsentativo hacer una imputación y dado el contexto del set de datos y la variable misma hace baste sentido visualizar esta variabilidad dado los modelos que contiene el set de datos. Por lo cual no se considerarán datos atípicos para esta variable.

```
                    MODELO      MARCA MAX(TASACION)
1 RANGE ROVER VOGUE 5.0 LAND ROVER      91880000
```

Ilustración 4: Respuesta de consola de la query al data frame sobre el MODELO y MARCA del vehículo de mayor TASACION.

Análisis Gráfico de los Datos

Luego de haber realizado el tratamiento de los datos, se buscará entender el comportamiento de cada una de las variables que pueda ayudar a comprender y visualizar de mejor forma el objetivo del desafío. Para esto se presentan las distribuciones de las distintas variables numéricas contenidas en los datos.

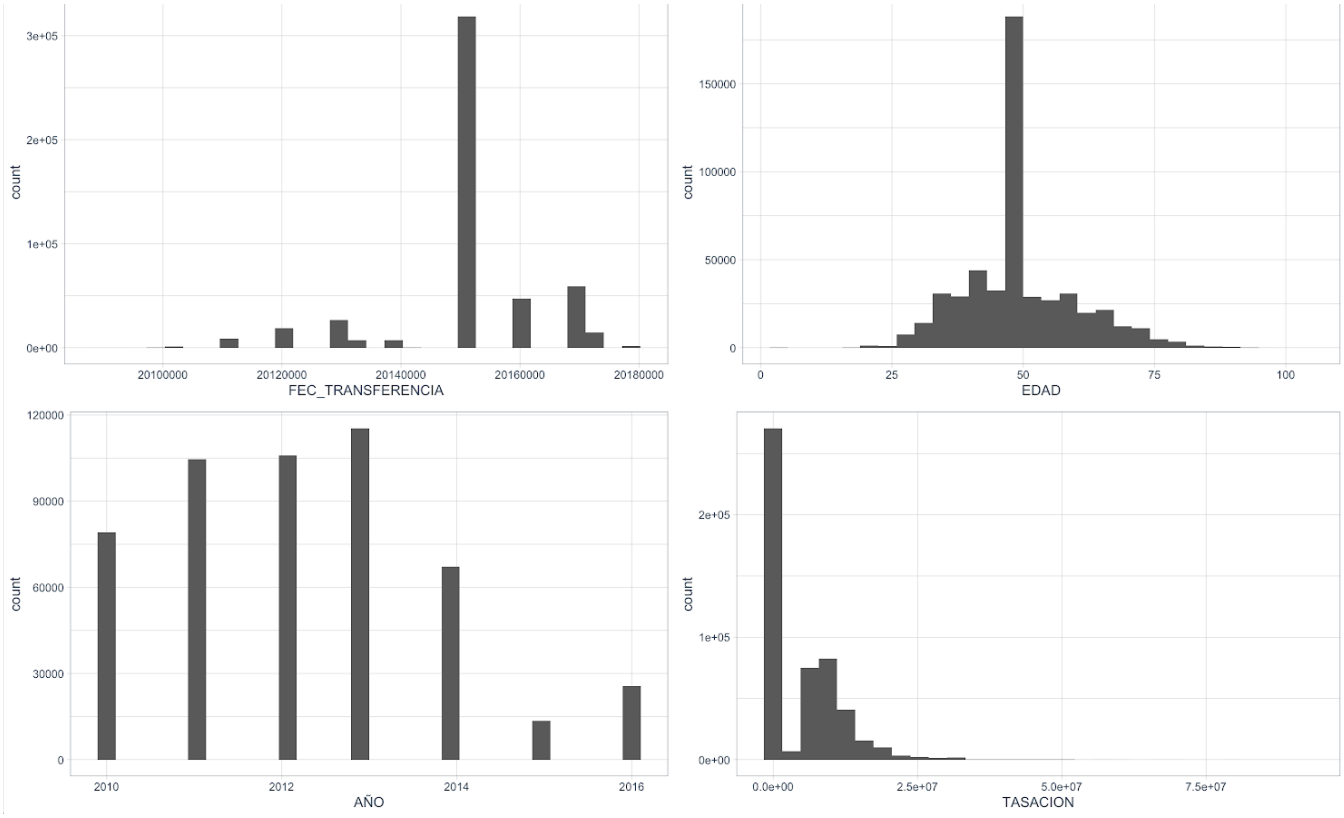

Gráfico 3: Distribución de variables FEC_TRANSFERENCIA, EDAD, AÑO y TASACION.

La verdad es que visualmente no es muy simple emitir conclusiones, lo que se puede apreciar simplemente es la naturaleza discreta y continua de las variables.

El siguiente análisis realizado corresponde a la visualización de outliers o datos atípicos, Gráfico 4, para las variables numéricas en cuestión, donde como se comentó anteriormente para la variable TASACION no se considerarán como valores atípicos, lo mismo para FEC_TRANSFERENCIA. En la variable AÑO no se aprecian valores atípicos.

Sin embargo para la variable EDAD habrá que considerar dentro del contexto de clientes que renueven vehículos a aquellos que estén entre los 18 y 75 años de edad, edad mínima para conducir y edad idóneamente máxima para hacerlo.

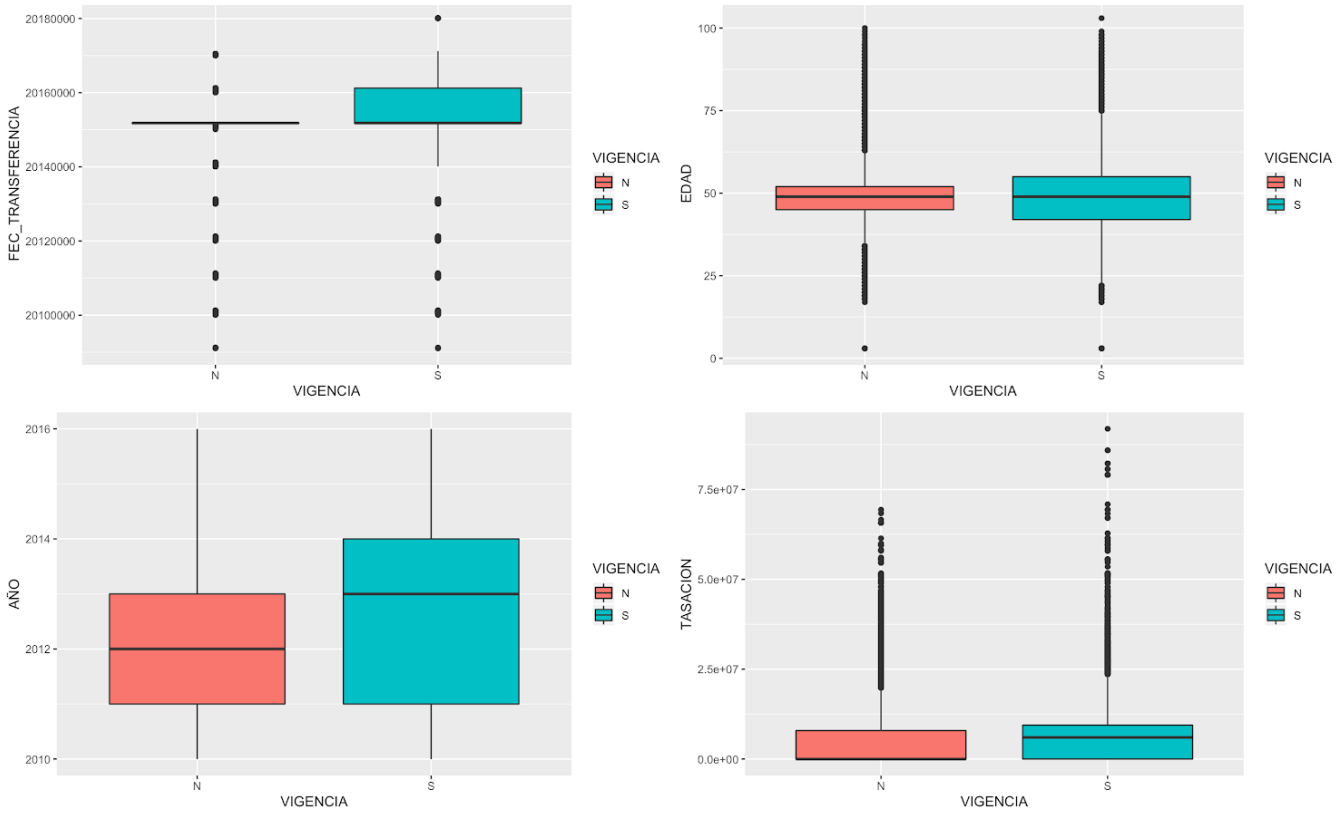

Gráfico 4: Boxplot de variables FEC_TRANSFERENCIA, EDAD, AÑO y TASACION vs variable dicotómica VIGENCIA.

Luego de remover los outliers para la variable EDAD y los valores cero para TASACION, se obtuvieron los siguientes resúmenes estadísticos para dichas variables, con el muy buen número de observaciones de 237659.

```
> dim(derco_dataFinal3)
[1] 237659     13
> summary(derco_dataFinal3$EDAD)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   42.00   48.93   47.93   52.00   75.00
> summary(derco_dataFinal3$TASACION)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
 1460000  7263000  9274000 10515235 11930000 91880000
```

Ilustración 5: Respuesta de consola del total de la dimensión del data frame final de trabajo y del resumen estadístico de las variables EDAD y TASACION.

Análisis de Correlación

Se realizó un análisis de correlación de las variables en cuestión, con el objetivo de evaluar el comportamiento entre ellas y evaluar posibles mitigaciones de alguna multicolinealidad entre las variables que pudiese afectar al potenciales modelos predictivos entre otras conclusiones. En el Ilustración 6 se observa el mapa de correlación elaborado para las variables numéricas del set de datos, omitiendo todas aquellas que sean categóricas. Como se puede observar, no existe multicolinealidad entre las variables. En general correlaciones bastante bajas tanto positivas como negativas.

Ilustración 6: Mapa de correlación.

Podrían obtenerse muchos más resultados y conclusiones por medio de técnicas de visualización de datos y de consulta de datos al data frame, como segmentar por género, región, comuna, prácticamente por cada variable categórica, por periodos de tiempos, modelos específicos, marcas, etc. Lo que técnicamente se traduce a lo ya realizado en el código y presentado en este resumen.