HEALTH INSURANCE

EDA

ENTENDIMIENTO DEL NEGOCIO Y ORIGEN DE LOS DATOS

Objetivo del proyecto: Plasmar los hallazgos más importantes del proceso de limpieza y preparación de los datos para identificar la calidad de los datos recolectados.

- Cliente: Empresa de Seguros
- Objetivo: Encontrar variables relacionadas a tener seguro médico

Origen de los datos:

- Suministrados por una firma de mercados.
- Muestra representativa de la población
- Datos recolectados en octubre del 2016

IDENTIFICACIÓN DE LA ESTRUCTURA DE DATOS

```
## 'data.frame':    1002 obs. of  12 variables:
##  $ date            : Factor w/ 996 levels "10/24/2016 10:02:16",..: 386 387 388 389 390 391 392 393 394 395
...
##  $ custid          : int  2068 2073 2848 5641 6369 8322 8521 12195 14989 15917 ...
##  $ sex             : int  1 1 2 2 1 1 2 2 2 1 ...
##  $ is.employed     : logi  NA NA TRUE TRUE TRUE TRUE ...
##  $ annual_incomeUSD: int  11300 0 4500 20000 12000 180000 120000 40000 9400 24000 ...
##  $ marital.stat    : Factor w/ 4 levels "Divorced/Separated",..: 2 2 3 3 3 3 3 2 2 1 ...
##  $ health.ins      : logi  TRUE TRUE FALSE FALSE TRUE TRUE ...
##  $ housing.type    : Factor w/ 4 levels "Homeowner free and clear",..: 1 4 4 3 4 2 1 4 4 1 ...
##  $ vehicle         : Factor w/ 2 levels "NO","YES": 2 2 2 1 2 2 2 2 2 2 ...
##  $ num.vehicles    : int  2 3 3 4 1 1 1 3 2 1 ...
##  $ age             : Factor w/ 78 levels "0","18","19",..: 34 25 6 6 16 25 24 33 29 56 ...
##  $ state.of.res    : Factor w/ 54 levels " New York","Alabama",..: 24 11 12 33 11 36 14 24 15 37 ...
```

DETECCIÓN DE ANOMALÍAS

- Formatos incorrectos: fecha, sexo, **edad** y tener vehículo.
- Duplicados:

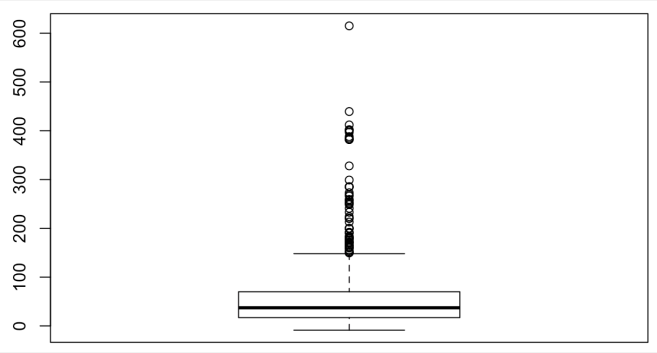
	date	custid	sex	is.employed	annual_incomeUSD	marital.stat	health.ins	housing.type	vehicle	num.vehicles	age	state.of.res
876	2016-10-25	1238436	Female	NA	0	Married	FALSE	Rented	TRUE	2	25	Tennessee
877	2016-10-25	1238436	Female	NA	0	Married	FALSE	Rented	TRUE	2	25	Tennessee
1001	2016-10-27	1414286	Female	FALSE	20900	Married	FALSE	Rented	TRUE	2	36	New York
1002	2016-10-27	1414286	Female	FALSE	20900	Married	FALSE	Rented	TRUE	2	36	New York

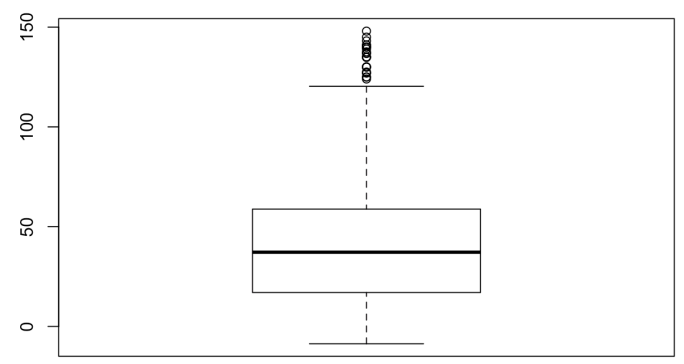
- **Campos vacíos**

Nombre	Cantidad	Tratamiento
Sexo	1	Borrado
Es o no empleado	329 (33%)	Nueva variable Otro
Ingreso Anual	1	Borrado
Estado Civil	1	Borrado
Tipo de Vivienda	57 (6%)	Borrados
Tiene o no Vehículo	1	Borrado
Número de Vehículos	112	Llenado con cero. Relacionado con Vehículo
Edad	3	Reemplazado por Cero. Después imputado.

- **Outliers**

Ingreso Anual: Imputado con mediana

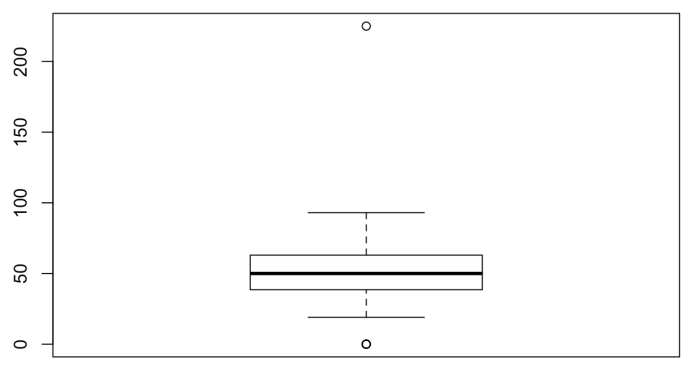


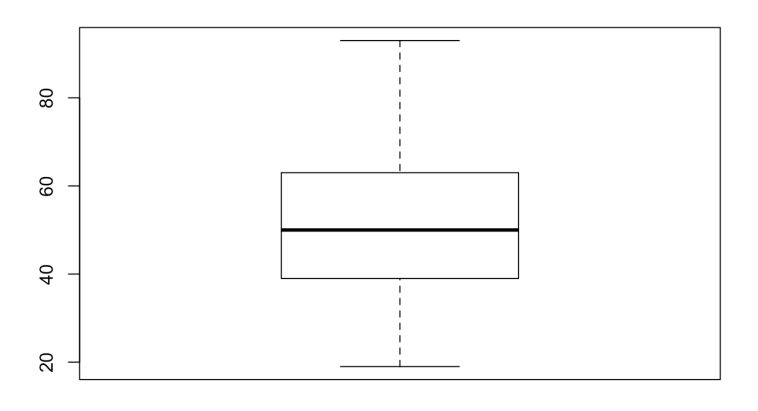


Antes

Edad: Imputado con mediana

Después





- **Datos Inválidos**

Nombre	Tratamiento
Ingreso Annual <= 0	Imputar con mediana
Vehículo = False & N° Vehículos > 0	Registro Borrado
Typo en estados de residencia	Ajustados a reales
Estado Civil	Borrado
Tipo de Vivienda	Borrados
Tiene o no Vehículo	Borrado
Número de Vehículos	Llenado con cero. Relacionado con Vehículo
Edad	Reemplazado por Cero. Después imputado.

PREPARACIÓN DE LOS DATOS

- Elimina variable fecha pues los valores van del 24, 25, 26 y 27 octubre de 2016
- Se borra el id de los usuarios pues es único
- Creación variable is home owner
- Creación columna que agrupa por regiones a los estados
- Eliminar variable estados pues la anterior los agrupa
- Reordenamiento de columnas

Summary

```
     sex            age                    marital.stat
Female:424    Min.   :19.0   Divorced/Separated:149
Male  :518    1st Qu.:39.0   Married            :506
              Median :50.0   Never Married      :197
              Mean   :51.5   Widowed            : 90
              3rd Qu.:63.0
              Max.   :93.0
                        housing.type is.house.owner  bureau.cardinal.point
Homeowner free and clear    :156    Mode :logical   Midwest  :270
Homeowner with mortgage/loan:412    FALSE:374       Northeast:224
Occupied with no rent       : 10    TRUE :568       South    :287
Rented                      :364                    West     :161


      is.employed  annual_incomeUSD  num.vehicles   health.ins
Employee    :589   Min.   :  0.03   Min.   :0.000   Mode :logical
Not Employee: 73   1st Qu.: 21.60   1st Qu.:1.000   FALSE:136
Other       :280   Median : 37.20   Median :2.000   TRUE :806
                   Mean   : 43.71   Mean   :1.918
                   3rd Qu.: 58.90   3rd Qu.:2.000
                   Max.   :148.00   Max.   :6.000
```

STR

```
'data.frame':    942 obs. of  10 variables:
 $ sex                    : Factor w/ 2 levels "
 $ age                    : num  49 40 22 31 40
 $ marital.stat           : Factor w/ 4 levels "
 $ housing.type           : Factor w/ 4 levels "
 $ is.house.owner         : logi  TRUE FALSE FAL
 $ bureau.cardinal.point: Factor w/ 4 levels "
 $ is.employed            : Factor w/ 3 levels "
 $ annual_incomeUSD       : num  11.3 37.2 4.5 1
 $ num.vehicles           : num  2 3 3 1 1 1 3 2
 $ health.ins             : logi  TRUE TRUE FALS
```

- **Head**

	sex	age	marital.stat	housing.type	is.house.owner	bureau.cardinal.point	is.employed	annual_incomeUSD	num.vehicles	health.ins
1	Female	49	Married	Homeowner free and clear	TRUE	Midwest	Other	11.3	2	TRUE
2	Female	40	Married	Rented	FALSE	South	Other	37.2	3	TRUE
3	Male	22	Never Married	Rented	FALSE	South	Employee	4.5	3	FALSE
5	Female	31	Never Married	Rented	FALSE	South	Employee	12.0	1	TRUE
6	Female	40	Never Married	Homeowner with mortgage/loan	TRUE	Northeast	Employee	37.2	1	TRUE
7	Male	39	Never Married	Homeowner free and clear	TRUE	West	Employee	120.0	1	TRUE

CONCLUSIONES

- Se recomienda:
  - Recolectar más datos.
  - Corregir variable anual income en la medida de lo posible
  - Tener más opciones para la variable is employed
  - Tener cuidado con los typo
- A pesar de lo anterior, gracias a las fases de limpieza y procesamiento de datos, es posible construir la primera versión del modelo.