Exploración de la base de datos ENOE

Amanda Balderas M.

Junio 2015

Objetivo

- Describir el proceso mediante el cual se realizó la obtención y limpieza de los datos de la `base_enoe` .
- Hacer una descripción de la `base_enoe` .
- Presentar un análisis exploratorio de la información que nos permitirá identificar algunas caracteristicas de los datos.

Introducción

Para cualquier análisis es importante contar con la información de interés en un formato correcto y con datos limpios, en este documento se describe el proceso realizado para obtener la información que se utilizará para el análisis.

Asimismo se presenta un breve análisis exploratorio de la información.

En este documento se desarrollan los siguientes puntos:

1. Obtención de los datos
2. Descripción de la base de datos
3. Limpieza y transformación de los datos
4. Análisis exploratorio de los datos
5. Anexo

---

1. Obtención de los datos

Los datos utilizados para este proyecto se obtuvieron de la página oficial del Instituto de Nacional de Estadística y Geografía INEGI (http://www.inegi.org.mx/). La información corresponde es resultado de la Encuesta Nacional de Ocupación y Empleo (ENOE).

Encuestas en Hogares

ENOE

Encuesta Nacional de Ocupación y Empleo (ENOE), población de 15 años y más de edad

Productos y servicios

La ENOE tiene como objetivo obtener información estadística sobre las características ocupacionales de la población a nivel nacional, así como otras variables demográficas y económicas que permitan profundizar en el análisis de los aspectos laborales.

Dicha encuesta se realiza trimestralmente a nivel nacional, para las 32 entidades federativas. El marco muestral utilizado para la ENOE es el marco nacional de viviendas 2012 del INEGI, el tamaño de la muestra es de 120,260 viviendas.

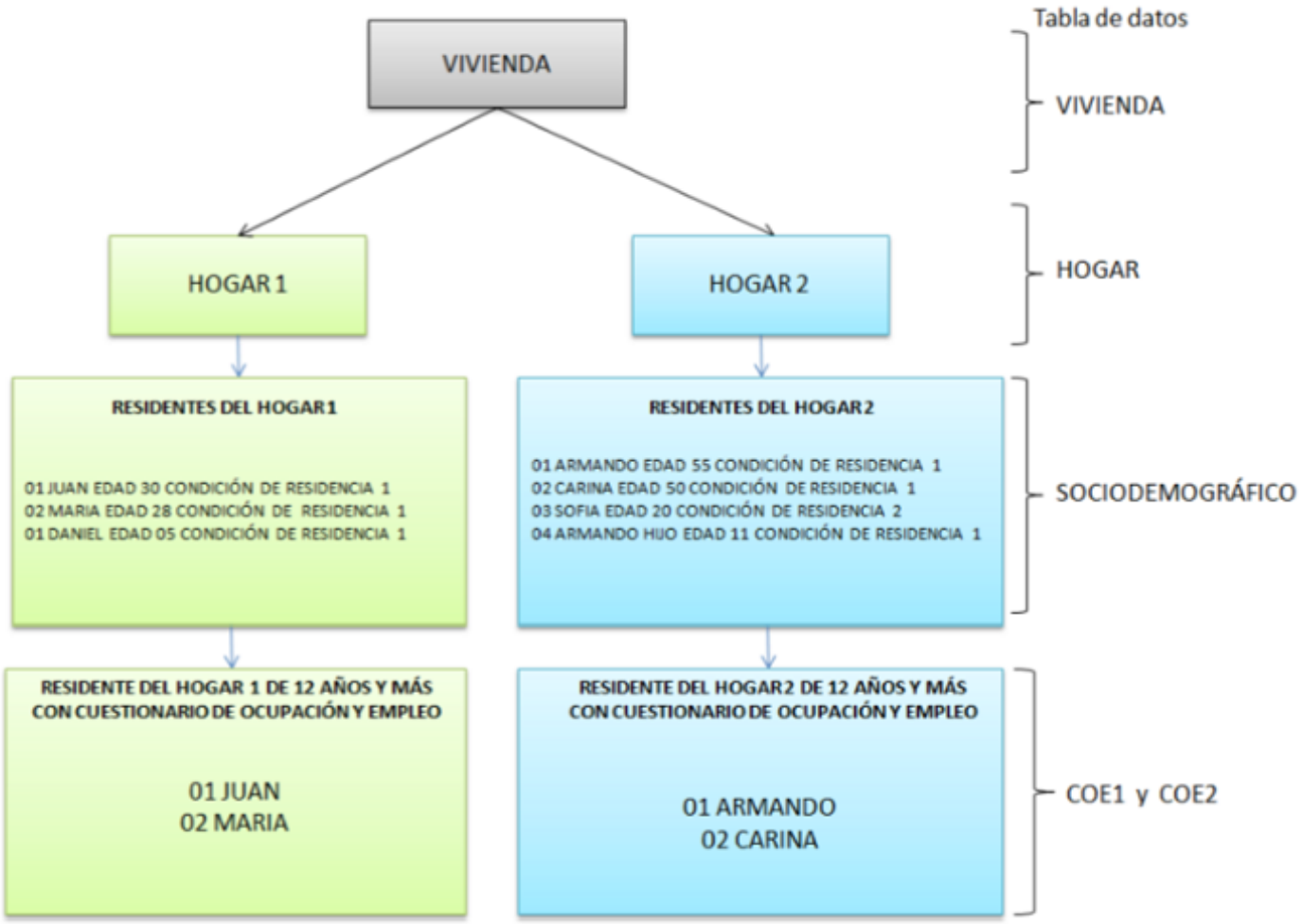
La unidad de análisis es el hogar y la población objetivo son las personas residentes habituales de las viviendas seleccionadas.

El INEGI presenta entre sus productos las bases correspondientes a los microdatos que contienen la información recabada en la encuesta para cada persona que fue entrevistada.

La base de datos de la ENOE, se conforma por cinco tablas de datos en formato “DBF”, las tablas son:

Tabla	Iniciales
1. Tabla de vivienda	VIV
2. Tabla de hogares	HOG
3. Tabla de sociodemográfico	SDEM
4. Tabla de cuestionario de ocupación y empleo I	COE1
5. Tabla de cuestionario de ocupación y empleo II	COE2

La siguiente imagen representa la relación que existe entre las tablas y sus registros. Para cada vivienda existe uno o más hogares, para cada hogar existe uno o más residentes, para cada residente de 12 años y más existe un cuestionario de ocupación y empleo.



Para el desarrollo de este proyecto se utilizará en particular la información de la tabla *Sociodemográfico* y se realizará el análisis de la información correspondiente al último levantamiento publicado de la ENOE que tiene como periodo de referencia el primer trimestre de 2015.

---

2. Descripción de la base de datos.

En la tabla *Sociodemográfico* se almacenan las características de los residentes del hogar, como es la condición de residencia, la edad, el sexo, estado civil, etc.

Además en esta tabla se incluyen una serie de campos llamados *precodificados*, los cuales son de gran utilidad para el procesamiento y consulta de datos, permitiendo generar información en forma rápida y oportuna sobre las características sociodemográficas y de ocupación directamente desde esta tabla, sin tener que implementar procesos adicionales de codificación.

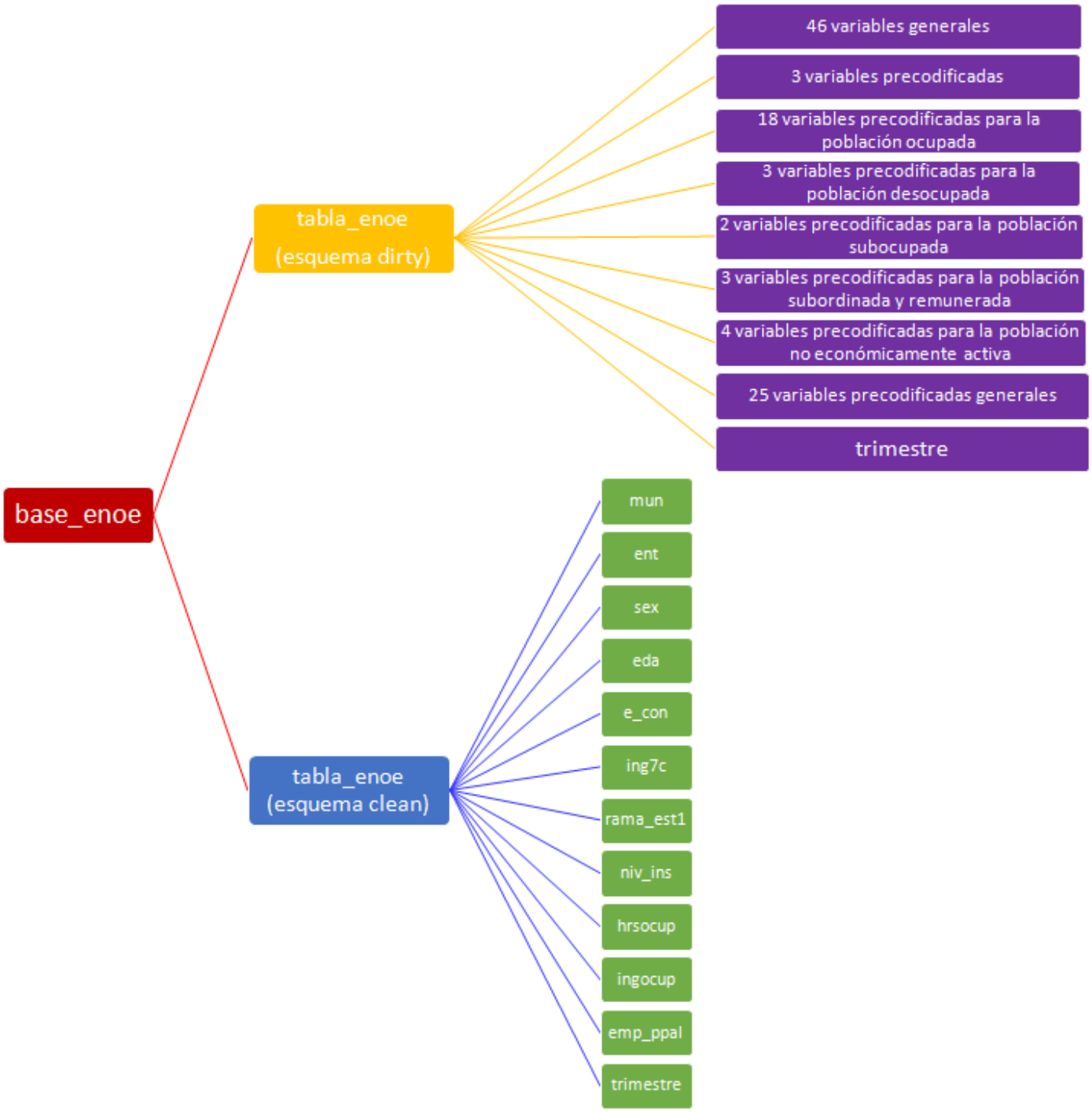
Las variables precodificadas están basadas en los principales grupos poblacionales de acuerdo con su condición de actividad y se clasifican de la manera siguiente:

- Variables precodificadas.
- Variables precodificadas para la población ocupada.
- Variables precodificadas para la población desocupada.
- Variables precodificadas para la población subocupada.

- Variables precodificadas para la población subordinada y remunerada.
- Variables precodificadas para la población no económicamente activa.
- Variables precodificadas generales.

La tabla *Sociodemográfico* correspondiente al primer trimestre de 2015, cuenta con 104 variables y 404,432 registros.

La `base_enoe` se conforma de las tablas llamadas `tabla_enoe` que se encuentran en el esquema *dirty* y en el esquema *clean*.



Variables seleccionadas

Para este ejercicio no se incluyeron todas las variables se la tabla *Sociodemográfico* en la `tabla_enoe` del esquema *dirty*, para un manejo más práctico de la información se seleccionaron algunas variables que se consideraron interesantes para el análisis.

Las variables seleccionadas son:

Nombre de la variable	Descripción de la variable	Tipo	Longitud	Códigos	Descripción del código
MUN	Municipio	Caracter	3	001 - 575	Número de municipio según entidad
ENT	Entidad	Caracter	2	01 - 32	Ver Anexo
SEX	Sexo	Caracter	1	1	Hombre
				2	Mujer
EDA	Edad	Caracter	2	00	Menores de 1 año
				01 - 96	Número de años cumplidos del integrante del hogar
				97	97 años y más
				98	Edad no especificada para mayores (12 años y más)
				99	Edad no especificada para menores (00 a 11 años)
E_CON	Estado conyugal	Caracter	1	1	Vive con su pareja en unión libre
				2	Está separado(a)
				3	Está divorciado(a)
				4	Esta viudo(a)
				5	Está casado(a)
				6	Está soltero(a)
				9	No sabe
ING7C	Clasificación de la población ocupada por nivel de ingreso	Numérico	1	1	Hasta un salario mínimo
				2	Más de 1 y hasta 2 salarios mínimos
				3	Más de 2 y hasta 3 salarios mínimos
				4	Más de 3 y hasta 5 salarios mínimos
				5	Más de 5 salarios mínimos
				6	No recibe ingresos
				7	No especificado
RAMA_EST1	Clasificación de la población ocupada según sector de actividad - Totales	Numérico	1	1	Primario
				2	Secundario
				3	Terciario
				4	No especificado
NIV_INS	Clasificación de la población de 5 años y más por nivel de instrucción	Numérico	1	1	Primaria incompleta
				2	Primaria completa
				3	Secundaria completa
				4	Medio superior y superior
				5	No especificado
HRSOCUP	Horas trabajadas a la semana	Numérico	3	1 - 168	Número de horas trabajadas a la semana
INGOCUP	Ingreso mensual	Numérico	6	1 - 999998	Valor del ingreso registrado en esta variable está en función al número de horas trabajadas a la semana; es decir, si tiene horas a la semana se coloca el valor del ingreso en esta variable del captado en la entrevista, de lo contrario se registra un cero.
EMP_PPAL	Clasificación de empleos formales e informales	Numérico	1	1	Empleo informal
				2	Empleo formal
trimestre	Trimestre al que corresponde la información	Caracter	6	01_2005 - 01 -2015	Ver Anexo

Dado lo anteior y considerando que trabajaremos con la información correspondiente al primer trimestre de 2015, tenemos que la `tabla_enoe_1_2015` del esquema *dirty_particion* que contiene la información en formato original, se conforma de 15 variables y 4010,432 observaciones, cada observación corresponde a una encuesta.

De las 15 variables que contiene la tabla, 14 son variables originales:

1. R_DEF
2. MUN

3. ENT
4. C_RES
5. SEX
6. EDA
7. E_CON
8. ING7C
9. CLASE2
10. RAMA_EST1
11. NIV_INS
12. HRSOCUP
13. INGOCUP
14. EMP_PPAL

Y una variable se incorpora a la tabla para identificar el periodo al que corresponde la información y que además sirve como parámetro para la partición de las tablas en *Postgresql*.

15. trimestre

Verificamos la estructura de la `tabla_enoe` en el esquema *dirty_particion*, considerando la información del primer trimestre de 2015.

```
# Nos conectamos a base_UFO en Postgresql
drv <- dbDriver("PostgreSQL")
cone <- dbConnect(drv, dbname="base_enoe", host="localhost", port=5432, user="postgres",
password="bameam29")

# Verificamos que la tabla existe
dbExistsTable(cone, c("dirty", "tabla_enoe"))
```

```
## [1] TRUE
```

```
# Leeemos la tabla
d_tabla_enoe_1_2015 <- dbReadTable(cone, c("dirty_particion", "tabla_enoe_1_2015"))
#d_tabla_enoe_1_2015 <- dbGetQuery(cone, "select * from dirty.tabla_enoe where trimestre
= '1_2015'")

# Checamos dimensión y estructura
dim(d_tabla_enoe_1_2015)
```

```
## [1] 404432     15
```

```
str(d_tabla_enoe_1_2015)
```

```
## 'data.frame':    404432 obs. of  15 variables:
##  $ r_def    : chr  "00" "00" "00" "00" ...
##  $ mun      : chr  "002" "002" "002" "002" ...
##  $ ent      : chr  "09" "09" "09" "09" ...
##  $ c_res    : chr  "1" "1" "1" "1" ...
##  $ sex      : chr  "2" "1" "2" "2" ...
##  $ eda      : chr  "27" "45" "48" "24" ...
##  $ e_con    : chr  "6" "1" "1" "6" ...
##  $ clase2   : int  1 1 1 3 1 1 1 1 1 1 ...
##  $ ing7c    : int  4 3 3 0 3 3 2 6 2 5 ...
##  $ rama_est1: int  3 2 3 0 3 3 3 3 3 3 ...
##  $ niv_ins  : int  4 4 3 4 2 4 3 3 1 4 ...
##  $ hrsocup  : int  40 24 32 0 45 48 45 16 48 24 ...
##  $ ingocup  : int  0 6000 4300 0 0 4800 3440 0 3000 0 ...
##  $ emp_ppal : int  2 1 2 0 1 1 1 1 1 2 ...
##  $ trimestre: chr  "1_2015" "1_2015" "1_2015" "1_2015" ...
```

Por otra parte, la `tabla_enoe_1_2015` del esquema *clean_particion*, que contiene la información limpia y transformada, se forma de 12 variables y un total de 167,959 observaciones, está tabla es la que se utilizará para el análisis de los datos.

---

3. Limpieza y transformación de los datos.

Selección de registros

Para seleccionar los registros que son utilizados en el análisis, se utilizó como base el criterio definido por el INEGI para obtener la información correspondiente a población ocupada, se decidió seleccionar estos registros, ya que con esta información podremos garantizar que la base corresponde a personas que en el momento de la entrevista se encontraban trabajando, además de que son los criterios que el INEGI considera para obtener las estadísticas oficiales que se publican como resultados de la ENOE.

Tenemos que en la tabla *Sociodemográfico* existen los campos:

`R_DEF` Esta variable almacena el resultado definitivo de la entrevista del hogar.

`C_RES` Esta variable determina la condición de residencia del ocupante del hogar.

`CLASE2` Esta variable clasifica a la población ocupada y desocupada; disponible y no disponible.

Estas variables se incluyen en la `tabla_enoe` del el esquema *dirty* ya que se utilizarán para la selección de registros:

Nombre de la variable	Descripción de la variable	Tipo	Longitud	Códigos	Descripción del código
R_DEF	Resultado definitivo de la entrevista	Caracter	2	00	Entrevista completa
				Tipo "A"	Vivienda habitada
				01	Nadie en el momento de la entrevista
				02	Ausente temporal
				03	Se negó a dar información
				04	Informante inadecuado
				05	Otro motivo
				14	El hogar se mudó
				15	Entrevista suspendida
				Tipo "B"	Vivienda deshabitada
				06	Adecuada para habitarse
				07	De uso temporal
				08	Inadecuada para habitarse
				09	De uso temporal para fines diferentes de habitación
				Tipo "C"	Vivienda fuera de muestra
				10	Demolida
				11	Cambió de sitio (móvil)
				12	Uso permanente para fines diferentes a los de habitación
				13	Otro motivo
C_RES	Condición de residencia	Caracter	1	1	Residente habitual
				2	Ausente definitivo
				3	Nuevo residente
CLASE2	Clasificación de la población en ocupada y desocupada; disponible y no	Numérico	1	1	Población ocupada
				2	Población desocupada
				3	Disponibles
				4	No disponibles

Además en las el grupo de variables seleccionadas tenemos:

`EDA` Esta variable almacena los años cumplidos del residente. La reciente reforma constitucional define los 15 años como la edad legal mínima para trabajar.

Con las variables anteriores y dadas las definiciones que tenemos para cada variable se obtiene el criterio general para selección de la población ocupada con ingresos, que queda de la siguiente manera:

```
(R_DEF = 00) y (C_RES = (1 o 3)) y (EDA > 14 y EDA < 97) y (CLASE2 = 1)
```

Para la selección de los registros se realizan los siguientes pasos:

Pasos:		Total de registros	Se eliminan
1.	Eliminar todos los registros que en el campo "R_DEF" sean diferentes a "00".	404,432	40
2.	Eliminar todos los registros que en el campo "C_RES" sean iguales a "2".	404,392	7,580
3.	Eliminar todos los registros que en el campo "EDA" contengan valores menores a 15 y mayores a 96.	396,812	105,936
4.	Eliminar todos los registros que en el campo "CLASE2" sean diferentes a "1".	290,876	122,917
5.	Registros finales en la base.	167,959	

Transformación de los datos

```
c_tabla_enoe_1_2015 <- d_tabla_enoe_1_2015
```

Se verifica cada una de las variables descritas anteriormente y que serán utilizadas en el análisis.

Variables para la selección de registros

`r_def` y `c_res`

Se verifica que las variables son de tipo caracter y que las categorías corresponden a las que se tienen definidas.

Tenemos que estas variables sólo se utilizarán para la selección de registros por lo que no se realiza ninguna transformación en ellas.

```
# Resultado entrevista
class(c_tabla_enoe_1_2015$r_def)
```

```
## [1] "character"
```

```
c_tabla_enoe_1_2015$r_def <- as.factor(c_tabla_enoe_1_2015$r_def)
unique(c_tabla_enoe_1_2015$r_def)
```

```
## [1] 00 15
## Levels: 00 15
```

```
# Condición de residencia
class(c_tabla_enoe_1_2015$c_res)
```

```
## [1] "character"
```

```
c_tabla_enoe_1_2015$c_res <- as.factor(c_tabla_enoe_1_2015$c_res)
unique(c_tabla_enoe_1_2015$c_res)
```

```
## [1] 1 3 2
## Levels: 1 2 3
```

`clase2`

Se verifica que la variable es de tipo numérico y cumple el rango definido, dado que esta variable sólo se utilizará para la selección de registros, no se realiza ninguna transformación en ella.

```
# Clasificación de la población
class(c_tabla_enoe_1_2015$clase2)
```

```
## [1] "integer"
```

```
unique(c_tabla_enoe_1_2015$clase2)
```

```
## [1] 1 3 2 4 0
```

`eda` Verificamos que la variable es de tipo caracter, para facilitar su uso en el análisis se modifica a tipo numérica.

Con lo anterior, tenemos que el rango de la variable va de 1 a 100, por lo que se conserva el formato de la variable sin discretizar.

Además sabemos que, de acuerdo a la definición de la variable, los valores menores a 1 y mayores a 96 no son edades específicas; sin embargo, no se realiza ningún cambio ya que en la selección de registros se eliminan esos casos.

```
# Edad
class(c_tabla_enoe_1_2015$eda)
```

```
## [1] "character"
```

```
c_tabla_enoe_1_2015$eda <- as.integer(c_tabla_enoe_1_2015$eda)
min(c_tabla_enoe_1_2015$eda, na.rm = TRUE)
```

```
## [1] 0
```

```
max(c_tabla_enoe_1_2015$eda, na.rm = TRUE)
```

```
## [1] 99
```

Resto de las variables seleccionadas

`mun` , `ent` , y `sex`

Se verifica que las variables son de tipo caracter y que las categorías corresponden a las definidas.

```
# Municipio
class(c_tabla_enoe_1_2015$mun)
```

```
## [1] "character"
```

```
c_tabla_enoe_1_2015$mun <- as.factor(c_tabla_enoe_1_2015$mun)
unique(c_tabla_enoe_1_2015$mun)
```

```
##   [1] 002 006 007 010 013 015 005 003 011 012 014 016 017 004 009 008 031
##  [18] 033 058 023 020 057 109 121 029 039 060 099 104 070 044 037 108 025
##  [35] 030 081 059 091 100 028 024 092 122 120 098 101 097 021 026 046 048
##  [52] 018 019 049 114 119 140 041 136 090 034 125 035 050 038 133 123 193
##  [69] 001 053 088 106 118 051 054 076 067 055 027 390 293 375 385 399 553
##  [86] 107 519 115 350 087 409 083 403 157 227 174 056 042 036 040 043 032
## [103] 078 089 077 102 184 156 022 131 093 069 112 052 079 071 132 074 075
## [120] 094 061 065 086 045 066 085 073 082 124 103 110 062 047 068 515 397
## [137] 482 208 154 164 206 207 143 204 128 160 096 084 064 072 401 565 441
## [154] 327 525 295 551 505 177 203 194 211 138 105 161 129 148 111 546 185
## [171] 384 520 324 217 137 180 189 186 168 117 149 116 063 113 333 340 135
## [188] 507 424 470 491 473 386 242 513 198 317 234 439 422 334 414 431 485
## [205] 190 496 533 166 346 134 266 202 178 172 197 150 191 142 126 095 200
## [222] 141 175 151 210 155 188
## 227 Levels: 001 002 003 004 005 006 007 008 009 010 011 012 013 014 ... 565
```

```
# Entidad
class(c_tabla_enoe_1_2015$ent)
```

```
## [1] "character"
```

```
c_tabla_enoe_1_2015$ent <- as.factor(c_tabla_enoe_1_2015$ent)
unique(c_tabla_enoe_1_2015$ent)
```

```
##  [1] 09 15 14 19 21 11 24 31 08 28 30 12 01 16 05 27 07 02 25 26 10 18 04
## [24] 17 20 32 06 22 29 03 23 13
## 32 Levels: 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 ... 32
```

```
# Sexo
class(c_tabla_enoe_1_2015$sex)
```

```
## [1] "character"
```

```
c_tabla_enoe_1_2015$sex <- as.factor(c_tabla_enoe_1_2015$sex)
unique(c_tabla_enoe_1_2015$sex)
```

```
## [1] 2    1    <NA>
## Levels: 1 2
```

e_con

Se verifica que la variable es de tipo caracter y que las categorías corresponden a las definidas.

```
# Estado conyugal
class(c_tabla_enoe_1_2015$e_con)
```

```
## [1] "character"
```

```
c_tabla_enoe_1_2015$e_con <- as.factor(c_tabla_enoe_1_2015$e_con)
unique(c_tabla_enoe_1_2015$e_con)
```

```
## [1] 6    1    5    4    <NA> 2    3    9
## Levels: 1 2 3 4 5 6 9
```

ing7c

Verificamos que la variable es de tipo numérico, podemos observar que además de los valores definidos para esta variable se tienen casos con valor cero.

Aunque la variable está definida originalmente como numérica, sabemos que realmente representa una variable categórica, por lo que para este ejercicio se convertirá a tipo factor.

```
# Clasificación del ingreso
class(c_tabla_enoe_1_2015$ing7c)
```

```
## [1] "integer"
```

```
c_tabla_enoe_1_2015$ing7c <- as.factor(c_tabla_enoe_1_2015$ing7c)
unique(c_tabla_enoe_1_2015$ing7c)
```

```
## [1] 4 3 0 2 6 5 1 7
## Levels: 0 1 2 3 4 5 6 7
```

`rama_est1`

Se verifica que la variable es de tipo númerico y sus valores corresponden a los definidos. Dentro de esos valores se tiene el valor “0” que corresponde a casos en los que no aplica un valor para la variable o un valor perdido, por lo que se modifica ese valor para considerar en general esos casos como valores faltantes.

Aunque la variable está definida originalmente como numérica, sabemos que realmente representa una variable categórica, por lo que para este ejercicio se convertirá a tipo factor.

```
class(c_tabla_enoe_1_2015$rama_est1)
```

```
## [1] "integer"
```

```
unique(c_tabla_enoe_1_2015$rama_est1)
```

```
## [1] 3 2 0 1 4
```

```
c_tabla_enoe_1_2015$rama_est1[c_tabla_enoe_1_2015$rama_est1 == 0] <- NA
c_tabla_enoe_1_2015$rama_est1 <- as.factor(c_tabla_enoe_1_2015$rama_est1)
unique(c_tabla_enoe_1_2015$rama_est1)
```

```
## [1] 3    2    <NA> 1    4
## Levels: 1 2 3 4
```

`niv_ins`

Se verifica que la variable es de tipo numérico y sus valores corresponden a lo descrito en la definición.

En la definición de la variable tenemos que los valores “0” que se cambian a valores faltantes, por ser casos donde no aplica un valor o valores pedidos.

Aunque la variable está definida originalmente como numérica, sabemos que realmente representa una variable categórica, por lo que para este ejercicio se convertirá a tipo factor.

```
# Nivel de instrucción
class(c_tabla_enoe_1_2015$niv_ins)
```

```
## [1] "integer"
```

```
unique(c_tabla_enoe_1_2015$niv_ins)
```

```
## [1] 4 3 2 1 0 5
```

```
c_tabla_enoe_1_2015$niv_ins[c_tabla_enoe_1_2015$niv_ins == 0] <- NA
c_tabla_enoe_1_2015$niv_ins <- as.factor(c_tabla_enoe_1_2015$niv_ins)
unique(c_tabla_enoe_1_2015$niv_ins)
```

```
## [1] 4    3    2    1    <NA> 5
## Levels: 1 2 3 4 5
```

HRSOCUP Se verifica que la variable es de tipo numérico. El rango de la variable va de 0 a 168, dentro de ese rango se encuentran los valores definidos para la variable, pero se tiene el valor “0” que corresponde a casos en los que no aplica un valor para la variable o un valor perdido, estos casos se tratarán en el análisis de valores faltantes.

```
# Horas ocupadas
class(c_tabla_enoe_1_2015$hrsocup)
```

```
## [1] "integer"
```

```
min(c_tabla_enoe_1_2015$hrsocup)
```

```
## [1] 0
```

```
max(c_tabla_enoe_1_2015$hrsocup)
```

```
## [1] 168
```

ingocup

Se verifica que la variable es de tipo numérica. Tenemos que el rango de la variable va de 0 a 180,000.

También tenemos que en los valores definidos para esta variable no se considera el valor “0”; estos casos se modifican a valores faltantes, además de que considerando el criterio definido para la selección de registros, al aplicar los filtros correspondientes se espera que estos casos se eliminen, lo cual se verificará al realizar el análisis de valores faltantes.

```
# Ingreso mensual
class(c_tabla_enoe_1_2015$ingocup)
```

```
## [1] "integer"
```

```
min(c_tabla_enoe_1_2015$ingocup)
```

```
## [1] 0
```

```
max(c_tabla_enoe_1_2015$ingocup)
```

```
## [1] 180000
```

```
c_tabla_enoe_1_2015$ingocup[c_tabla_enoe_1_2015$ingocup == 0] <- NA
min(c_tabla_enoe_1_2015$ingocup, na.rm = TRUE)
```

```
## [1] 16
```

```
max(c_tabla_enoe_1_2015$ingocup, na.rm = TRUE)
```

```
## [1] 180000
```

emp_ppal

Se verifica que la variable es de tipo numérico y sus valores corresponden a los definidos para dicha variable. También se observa el valor “0” que corresponde a casos en los que no aplica un valor para la variable o un valor perdido, por lo que se modifica ese valor para considerar esos casos como valores faltantes.

Aunque la variable está definida originalmente como numérica, sabemos que realmente representa una variable categórica, por lo que para este ejercicio se convertirá a tipo factor.

```
# Tipo de empleo
class(c_tabla_enoe_1_2015$emp_ppal)
```

```
## [1] "integer"
```

```
unique(c_tabla_enoe_1_2015$emp_ppal)
```

```
## [1] 2 1 0
```

```
c_tabla_enoe_1_2015$emp_ppal[c_tabla_enoe_1_2015$emp_ppal == 0] <- NA
c_tabla_enoe_1_2015$emp_ppal <- as.factor(c_tabla_enoe_1_2015$emp_ppal)
unique(c_tabla_enoe_1_2015$emp_ppal)
```

```
## [1] 2    1    <NA>
## Levels: 1 2
```

trimestre

Esta variable se cro durante la carga, para identificar el periodo al que corresponden los datos.

Se verifica que la variable es de tipo caracter y sus valores corresponden a los definidos para dicha variable.

```
# Trimestre
class(c_tabla_enoe_1_2015$trimestre)
```

```
## [1] "character"
```

```
unique(c_tabla_enoe_1_2015$trimestre)
```

```
## [1] "1_2015"
```

Ahora, obtenemos la base con los registros correspondientes a población ocupada de acuerdo a los criterios que ya se describieron.

```
# Realizamos el primer filtro
c_tabla_enoe_1_2015 <- subset(c_tabla_enoe_1_2015, r_def == '00')
eliminados1 <- nrow(d_tabla_enoe_1_2015) - nrow(c_tabla_enoe_1_2015)
eliminados1
```

```
## [1] 40
```

```
# Realizamos el segundo filtro
c_tabla_enoe_1_2015 <- subset(c_tabla_enoe_1_2015, c_res == '1' | c_res == '3')
eliminados <- nrow(d_tabla_enoe_1_2015) - nrow(c_tabla_enoe_1_2015)
eliminados2 <- eliminados - eliminados1
eliminados2
```

```
## [1] 7580
```

```
# Realizamos el tercer filtro
c_tabla_enoe_1_2015 <- subset(c_tabla_enoe_1_2015, eda > 14 & eda < 97)
eliminados <- nrow(d_tabla_enoe_1_2015) - nrow(c_tabla_enoe_1_2015)
eliminados3 <- eliminados - eliminados2 - eliminados1
eliminados3
```

```
## [1] 105936
```

```
# Realizamos el cuarto filtro
c_tabla_enoe_1_2015 <- subset(c_tabla_enoe_1_2015, clase2 == 1)
eliminados <- nrow(d_tabla_enoe_1_2015) - nrow(c_tabla_enoe_1_2015)
eliminados4 <- eliminados - eliminados3 - eliminados2 - eliminados1
eliminados4
```

```
## [1] 122917
```

```
# Seleccionamos las variables de interes
variables_f <- c('ent', 'mun', 'sex', 'eda', 'e_con', 'ing7c', 'rama_est1', 'niv_ins', 'h
rsocup', 'emp_ppal', 'ingocup', 'trimestre')
c_tabla_enoe_1_2015 <- subset(c_tabla_enoe_1_2015, select = variables_f)

# Dimensión de la base
dim(c_tabla_enoe_1_2015)
```

```
## [1] 167959     12
```

```
# Revisamos la estructura
str(c_tabla_enoe_1_2015)
```

```
## 'data.frame':    167959 obs. of  12 variables:
##  $ ent      : Factor w/ 32 levels "01","02","03",..: 9 9 9 9 9 9 9 9 9 9 ...
##  $ mun      : Factor w/ 227 levels "001","002","003",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ sex      : Factor w/ 2 levels "1","2": 2 1 2 1 2 1 2 2 1 1 ...
##  $ eda      : int  27 45 48 54 49 20 18 91 50 17 ...
##  $ e_con    : Factor w/ 7 levels "1","2","3","4",..: 6 1 1 5 5 6 6 6 5 6 ...
##  $ ing7c    : Factor w/ 8 levels "0","1","2","3",..: 5 4 4 4 4 3 7 3 6 3 ...
##  $ rama_est1: Factor w/ 4 levels "1","2","3","4": 3 2 3 3 3 3 3 3 3 2 ...
##  $ niv_ins  : Factor w/ 5 levels "1","2","3","4",..: 4 4 3 2 4 3 3 1 4 2 ...
##  $ hrsocup  : int  40 24 32 45 48 45 16 48 24 40 ...
##  $ emp_ppal : Factor w/ 2 levels "1","2": 2 1 2 1 1 1 1 1 2 1 ...
##  $ ingocup  : int  NA 6000 4300 NA 4800 3440 NA 3000 NA NA ...
##  $ trimestre: chr  "1_2015" "1_2015" "1_2015" "1_2015" ...
```

** Valores faltantes**

Para finalizar el proceso de limpieza, verificaremos los registros que contengan valores faltantes en alguna de las variables seleccionadas.

Podemos observar que hay una importante cantidad de valores faltantes en la variable `ingocup`, ya que verificando algunos registros podemos observar que hay casos donde se reporta valor para la variable `ing7c` y/o para la variable `hrsocup`, pero parece ser que no se quiso reportar el dato específico de ingreso, lo cual suele ser muy común en las encuesta, particularmente para este tipo de variables como el ingreso.

```
# Número de valores faltantes por variable
faltantes <- lapply(c_tabla_enoe_1_2015, function(x) sum(is.na(x)))
faltantes
```

```
## $ent
## [1] 0
##
## $mun
## [1] 0
##
## $sex
## [1] 0
##
## $eda
## [1] 0
##
## $e_con
## [1] 0
##
## $ing7c
## [1] 0
##
## $rama_est1
## [1] 0
##
## $niv_ins
## [1] 0
##
## $hrsocup
## [1] 0
##
## $emp_ppal
## [1] 0
##
## $ingocup
## [1] 43825
##
## $trimestre
## [1] 0
```

```
# Verificamos casos sin respuesta en la variable de Ingreso
subset(d_tabla_enoe_1_2015, ingocup == 0 & hrsocup > 0 & (ing7c < 6 & ing7c > 0), select
= c(ing7c, hrsocup, ingocup))[1:15,]
```

```
##    ing7c hrsocup ingocup
## 1      4      40       0
## 5      3      45       0
## 10     5      24       0
## 14     2      40       0
## 17     5      32       0
## 20     3      37       0
## 22     3      40       0
## 23     4      32       0
## 32     5      28       0
## 40     2      84       0
## 46     2      40       0
## 48     2      72       0
## 83     2      36       0
## 85     2      36       0
## 86     2      45       0
```

Checamos el número de observaciones incompletas (con valores faltantes en alguna de las variables).

```
sum(!complete.cases(c_tabla_enoe_1_2015))
```

```
## [1] 43825
```

Dado que tenemos una importante cantidad de registros con algún valor faltante en la `ingocup` y para tratar de conservar la mayor cantidad de información haremos imputaciones para los valores faltantes.

`ingocup`

Vamos a verificar algunos casos:

- En la variable `ing7c` tenemos la categoria *6* que corresponde a “No percibe ingresos”, entonces los registros con esta categoria, se les asigna el valor cero a `ingocup` .

```
c_tabla_enoe_1_2015$ingocup[c_tabla_enoe_1_2015$ing7c == '6'] <- 0
```

- En la variable `hrsocup` tenemos la categoria valores con cero lo que nos indica que esa persona pudo no haber percibido ingresos, por lo que a los casos con *NA* en la variable `ingocup` se les asigna el valor cero.

```
c_tabla_enoe_1_2015$ingocup[c_tabla_enoe_1_2015$hrsocup == 0 & is.na(c_tabla_enoe_1_201
5$ingocup)] <- 0
```

- En la variable `ing7c` tenemos la categoria *7* que corresponde a “No especificado”, entonces los registros con esta categoria, se les imputará el promedio del ingreso para la variable `ingocup` .

```
prom_ing <- round(mean(c_tabla_enoe_1_2015$ingocup, na.rm = TRUE),0)
c_tabla_enoe_1_2015$ingocup[c_tabla_enoe_1_2015$ing7c == '7'] <- prom_ing
```

- Verificamos cuantos casos continuan sin dato en `ingocup` y tenemos aún una importante cantidad, que después de lo verificado, resultan ser los casos donde se dió un rango de ingresos `ing7c` y un número de horas trabajadas `hrsocup`, pero no se quiso dar el dato de ingreso `ingocup`, por lo que a estos casos también asignamos el promedio del ingreso.

```
sum(is.na(c_tabla_enoe_1_2015$ingocup))
```

```
## [1] 13622
```

```
c_tabla_enoe_1_2015$ingocup[is.na(c_tabla_enoe_1_2015$ingocup)] <- prom_ing

sum(is.na(c_tabla_enoe_1_2015$ingocup))
```

```
## [1] 0
```

Verificamos que ya no hay valores faltantes en ninguna de las variables.

```
faltantes <- lapply(c_tabla_enoe_1_2015, function(x) sum(is.na(x)))
faltantes
```

```
## $ent
## [1] 0
##
## $mun
## [1] 0
##
## $sex
## [1] 0
##
## $eda
## [1] 0
##
## $e_con
## [1] 0
##
## $ing7c
## [1] 0
##
## $rama_est1
## [1] 0
##
## $niv_ins
## [1] 0
##
## $hrsocup
## [1] 0
##
## $emp_ppal
## [1] 0
##
## $ingocup
## [1] 0
##
## $trimestre
## [1] 0
```

Finalmente, verificamos estructura de la información que se obtuvo después del proceso de limpieza y transformación.

```
# Checamos dimensión y estructura
str(c_tabla_enoe_1_2015)
```

```
## 'data.frame':    167959 obs. of  12 variables:
##  $ ent      : Factor w/ 32 levels "01","02","03",..: 9 9 9 9 9 9 9 9 9 9 ...
##  $ mun      : Factor w/ 227 levels "001","002","003",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ sex      : Factor w/ 2 levels "1","2": 2 1 2 1 2 1 2 2 1 1 ...
##  $ eda      : int  27 45 48 54 49 20 18 91 50 17 ...
##  $ e_con    : Factor w/ 7 levels "1","2","3","4",..: 6 1 1 5 5 6 6 6 5 6 ...
##  $ ing7c    : Factor w/ 8 levels "0","1","2","3",..: 5 4 4 4 4 3 7 3 6 3 ...
##  $ rama_est1: Factor w/ 4 levels "1","2","3","4": 3 2 3 3 3 3 3 3 3 2 ...
##  $ niv_ins  : Factor w/ 5 levels "1","2","3","4",..: 4 4 3 2 4 3 3 1 4 2 ...
##  $ hrsocup  : int  40 24 32 45 48 45 16 48 24 40 ...
##  $ emp_ppal : Factor w/ 2 levels "1","2": 2 1 2 1 1 1 1 1 2 1 ...
##  $ ingocup  : num  5161 6000 4300 5161 4800 ...
##  $ trimestre: chr  "1_2015" "1_2015" "1_2015" "1_2015" ...
```

Cargamos la información limpia a nuestra `tabla_enoe` en el esquema *clean*.

```
dbWriteTable(cone, c("clean", "tabla_enoe"), value = c_tabla_enoe_1_2015, append = TRUE,
row.names = FALSE)
```

Verificamos en "*psql*" que la información se cargó completa.

```
SELECT COUNT (*) FROM clean.tabla_enoe;
```

Nos desconectamos de *Postgresql*.

```
dbDisconnect(cone)
dbUnloadDriver(drv)
```

---

4. Análisis exploratorio de los datos.

Para el análisis exploratorio utilizamos la información que se cargo a la `tabla_ufo` del esquema *clean*, que ya contiene la información que se obtuvo despúes del proceso de limpieza y transformación.

Sumario Estadístico

Obtenemos un resumen estadístico de los datos.

```
summary(c_tabla_enoe_1_2015)
```

```
##       ent              mun         sex             eda        e_con
##  15     :  7598   001    :  8741   1:101943   Min.   :15.00   1:26902
##  11     :  7258   004    :  8415   2: 66016   1st Qu.:28.00   2: 7625
##  14     :  6396   005    :  8306              Median :38.00   3: 3535
##  21     :  6396   002    :  7433              Mean   :38.91   4: 4376
##  20     :  6028   006    :  6186              3rd Qu.:48.00   5:77644
##  19     :  5911   017    :  6158              Max.   :96.00   6:47863
##  (Other):128372   (Other):122720                              9:   14
##      ing7c       rama_est1  niv_ins      hrsocup       emp_ppal
##  2      :39881   1: 15169   1:18075   Min.   :  0.00   1:88712
##  3      :35252   2: 41606   2:29665   1st Qu.: 32.00   2:79247
##  4      :29499   3:110350   3:57731   Median : 45.00
##  7      :20255   4:   834   4:62375   Mean   : 41.72
##  1      :19059              5:  113   3rd Qu.: 50.00
##  5      :14493                        Max.   :168.00
##  (Other): 9520
##     ingocup        trimestre
##  Min.   :     0   Length:167959
##  1st Qu.:  3010   Class :character
##  Median :  5160   Mode  :character
##  Mean   :  5194
##  3rd Qu.:  6000
##  Max.   :180000
##
```

Destacan los siguientes resultados:

- En la bse final hay una mayor proporción de hombres.
- La edad promedio de las personas ocupadas es de 39 años.
- La mayor proporción de personas coupadas están casadas o solteros.
- La mayoria de las personas coupadas de la muestra ganan entre 1 y hasta 3 salarios mínimos .
- La mayoria de la población ocupada de la muestra labora en el sector terciario, que basicamente corresponde a comercio y servicios.
- La mayor proporción de personas ocupadas tienen un nivel de instrucción de “Medio superior y Superior”, pero casi la misma proporción cuenta solo con “Secundaria completa”.
- Vemos que un poco más de la mitad de personas ocupadas de la muestra, trataban en un empleo informal.
- El promedio de horas trabajadas a la semana es de 42, con un máximo de 168 que correspondería a una persona que trabaja durante todo el día, los 7 días de la semana lo cual pudiera ser un dato erroneo.
- Vamos a verificar los casos de personas que trabajan más de 12 horas díarias, considerando que trabajan los 7 días de la semana.

```
sum(c_tabla_enoe_1_2015$hrsocup > 84)
```

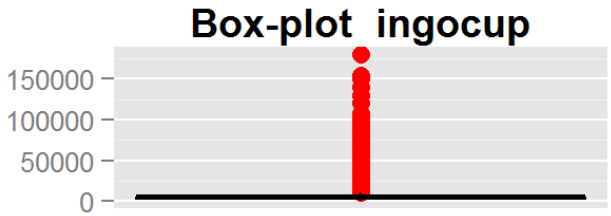
```
## [1] 1350
```

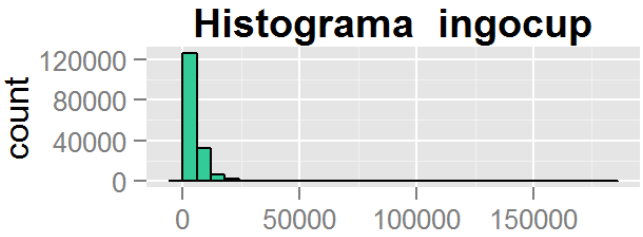
Análisis gráfico

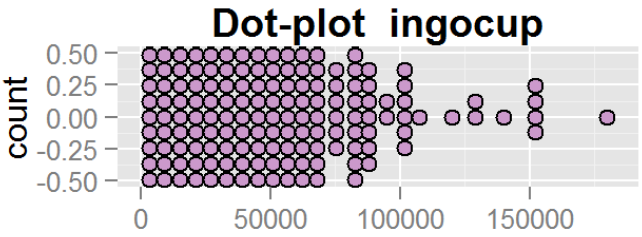
Ahora haremos un análisis gráfico de la variable `ingocup`, que en este caso es nuestra única variable numérica.

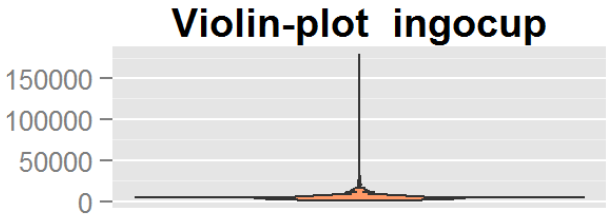
```
# Obtenemos gráficas de la variable duracion
base <- c_tabla_enoe_1_2015
x1 <- 11

  base$id <- c(1:nrow(base))
  grafica_0a <- ggplot(base, aes(x = base$id, y = base[,x1]))
  grafica_0b <- ggplot(base, aes(x = base[,x1]))

  # Box-plot
  grafica_1 <- grafica_0a +
  geom_boxplot(fill = '#3399CC', colour = 'black', outlier.colour = 'red', outlier.size =
3) +
  ggtitle(paste('Box-plot ', names(base)[x1])) +
  scale_y_continuous(name = '') +
  scale_x_continuous(name = '', breaks = NULL) +
  theme(plot.title = element_text(lineheight = .8, face = 'bold'))

  # Histograma
  grafica_2 <- grafica_0b +
  geom_histogram(fill = '#33CC99', colour = 'black') +
  ggtitle(paste('Histograma ', names(base)[x1])) +
  scale_x_continuous(name = '') +
  theme(plot.title = element_text(lineheight = .8, face = 'bold'))

  # Dot-plot
  grafica_3 <- grafica_0b +
  geom_dotplot(stackdir = 'centerwhole', fill = '#CC99CC') +
  ggtitle(paste('Dot-plot ', names(base)[x1])) +
  scale_x_continuous(name = '') +
  theme(plot.title = element_text(lineheight = .8, face = 'bold'))

  # Violin-plot
  grafica_4 <- grafica_0a +
  geom_violin(fill = '#FF9966') +
  ggtitle(paste('Violin-plot ', names(base)[x1])) +
  scale_y_continuous(name = '') +
  scale_x_continuous(name = '', breaks = NULL) +
  theme(plot.title = element_text(lineheight = .8, face = 'bold'))

  # Densidad
```

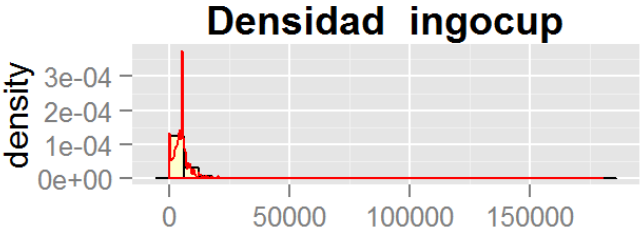
```r
grafica_5 <- grafica_0b +
  geom_histogram(aes(y = ..density..), fill = '#FFFFCC', colour = 'black') +
  geom_density(color = 'red') +
  ggtitle(paste('Densidad ', names(base)[x1])) +
  scale_x_continuous(name = '') +
  theme(plot.title = element_text(lineheight = .8, face = 'bold'))

# QQ-plot

# Variables que nos serviran para la qqline
yy <- quantile(base[,x1][!is.na(base[,x1])], c(0.25, 0.75))
xx <- qnorm(c(0.25, 0.75))
slope <- diff(yy) / diff(xx)
int <- yy[1L] - slope * xx[1L]

# Generamos la gráfica qqnorm y qqline
grafica_6 <- ggplot(base, aes(sample = base[,x1])) +
  ggtitle(paste('QQ-plot ', names(base)[x1])) +
  stat_qq(shape = 1, size = 4) +
  geom_abline(slope = slope, intercept = int, colour = 'red', size = 1) +
  theme(plot.title = element_text(lineheight = .8, face = 'bold'))

# Presentamos las gráficas en la misma pantalla
grid.newpage()
pushViewport(viewport(layout = grid.layout(3, 2)))
vplayout <- function(x, y) viewport(layout.pos.row = x, layout.pos.col = y)
print(grafica_1, vp = vplayout(1, 1))
print(grafica_2, vp = vplayout(1, 2))
print(grafica_3, vp = vplayout(2, 1))
print(grafica_4, vp = vplayout(2, 2))
print(grafica_5, vp = vplayout(3, 1))
print(grafica_6, vp = vplayout(3, 2))
```

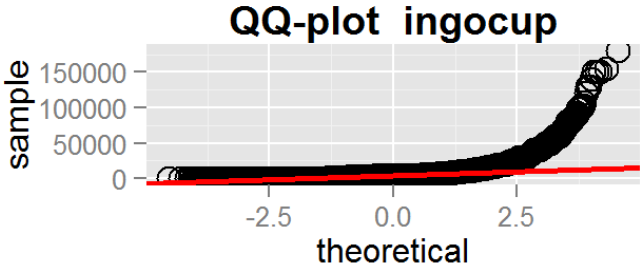




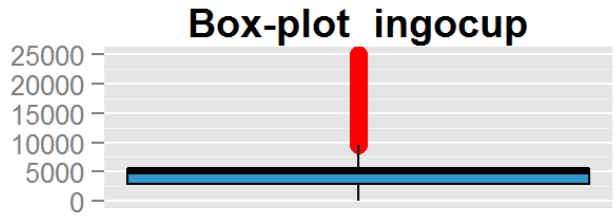


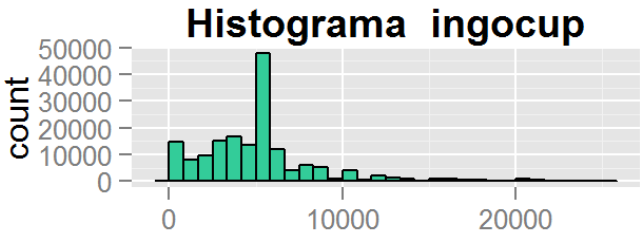


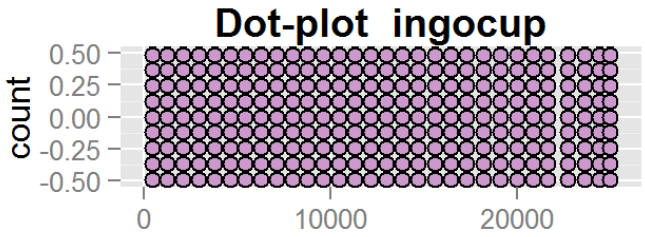


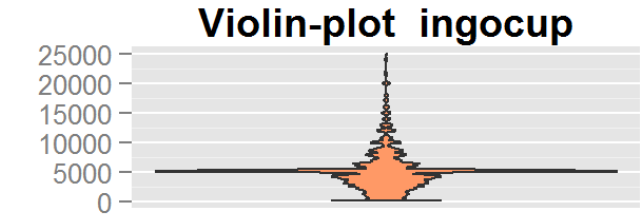


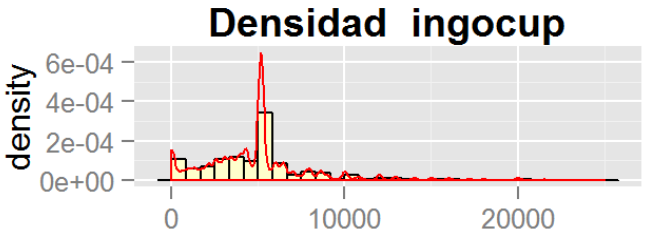
Ahora haremos un análisis gráfico de la variable `ingocup` , para los casos con valor menor o igual a 25,000 pesos.

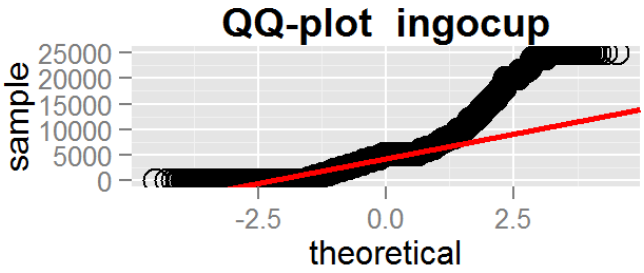












6. Anexo

- Tabla de categorías de la variable `ent`

Nombre de la variable	Códigos	Descripción del código	Códigos	Descripción del código
ENT	01	Aguascalientes	17	Morelos
	02	Baja California	18	Nayarit
	03	Baja California Sur	19	Nuevo León
	04	Campeche	20	Oaxaca
	05	Coahuila	21	Puebla
	06	Colima	22	Querétaro
	07	Chiapas	23	Quintana Roo
	08	Chihuahua	24	San Luis Potosí
	09	Distrito Federal	25	Sinaloa
	10	Durango	26	Sonora
	11	Guanajuato	27	Tabasco
	12	Guerrero	28	Tamaulipas
	13	Hidalgo	29	Tlaxcala
	14	Jalisco	30	Veracruz
	15	México	31	Yucatán
	16	Michoacán	32	Zacatecas

- Tabla de categorías de la variable `trimestre`

Nombre de la variable	Códigos	Descripción del código	Códigos	Descripción del código
trimestre	1_2005	Primer trimestre de 2005	1_2010	Primer trimestre de 2010
	2_2005	Segundo trimestre de 2005	2_2010	Segundo trimestre de 2010
	3_2005	Tercer trimestre de 2005	3_2010	Tercer trimestre de 2010
	4_2005	Cuarto trimestre de 2005	4_2010	Cuarto trimestre de 2010
	1_2006	Primer trimestre de 2006	1_2011	Primer trimestre de 2011
	2_2006	Segundo trimestre de 2006	2_2011	Segundo trimestre de 2011
	3_2006	Tercer trimestre de 2006	3_2011	Tercer trimestre de 2011
	4_2006	Cuarto trimestre de 2006	4_2011	Cuarto trimestre de 2011
	1_2007	Primer trimestre de 2007	1_2012	Primer trimestre de 2012
	2_2007	Segundo trimestre de 2007	2_2012	Segundo trimestre de 2012
	3_2007	Tercer trimestre de 2007	3_2012	Tercer trimestre de 2012
	4_2007	Cuarto trimestre de 2007	4_2012	Cuarto trimestre de 2012
	1_2008	Primer trimestre de 2008	1_2013	Primer trimestre de 2013
	2_2008	Segundo trimestre de 2008	2_2013	Segundo trimestre de 2013
	3_2008	Tercer trimestre de 2008	3_2013	Tercer trimestre de 2013
	4_2008	Cuarto trimestre de 2008	4_2013	Cuarto trimestre de 2013
	1_2009	Primer trimestre de 2009	1_2014	Primer trimestre de 2014
	2_2009	Segundo trimestre de 2009	2_2014	Segundo trimestre de 2014
	3_2009	Tercer trimestre de 2009	3_2014	Tercer trimestre de 2014
	4_2009	Cuarto trimestre de 2009	4_2014	Cuarto trimestre de 2014
			1_2015	Primer trimestre de 2015