Análisis del Concentrado Hogar ENIGH 2018

Joel Alejandro Zavala Prieto

Contents

Información de contacto

Mail: alejandro.zavala1001@gmail.com

Facebook: https://www.facebook.com/AlejandroZavala1001

Git: https://github.com/AlejandroZavala98

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

Introducción

Para este análisis que se realizo se tomo información del la ENIGH "Encuesta Nacional de Ingresos y Gastos de los Hogares" 2018, para mas detalle de cada variable que conforma este dataset se puede consultar en:

https://www.inegi.org.mx/contenidos/programas/enigh/nc/2018/doc/enigh18_descriptor_archivos_fd_ns.pdf

Se pretende realizar un análisis exploratorio de este dataset recabado y hacer algunas inferencias a modelos deseados

1. ¿Qué es la ENIGH?

La Encuesta Nacional de Ingresos y Gastos del Hogar (ENIGH) se publica por el INEGI cada dos años y proporciona los ingresos y gastos de los hogares de México, con información sobre la procedencia de sus ingresos, elreparto de sus gastos y las diferencias en los gastos dependiendo de la entidad donde los hogares están establecidos, su estrato socioeconómico, el sexo del jefe de familia asi como otras características sociodemográficas.

Objetivo

De acuerdo a la ENIGH su objetivo:

> "La ENIGH tiene como objetivo dar respuesta a los requerimientos de aquellos usuarios especializados, con un interés particular en el estudio de micro datos, permitiendo un análisis más detallado del monto, la estructura y la distribución de los ingresos de los hogares y del destino de los gastos del hogar en bienes de consumo duradero y no duradero. También se obtiene información sobre la infraestructura de las viviendas, la composición familiar de los hogares, así como de la actividad económica de cada uno de sus integrantes."

Población objetivo

La población de principal interés, son los integrantes que sean residentes de los hogares de 12 o más años, que residen habitualmente en viviendas dentro del territorio nacional y que cuentan con un trabajo, y contemplan un salario, y por ende son la población que tiene el poder adquisitivo de llevar acabo un gasto. Cobertura temática.

Cobertura temática

La cobertura temática de ENIGH, pretende abarcar aquellas reflexiones e investigaciones que versen sobre: hogares, gastos monetarios y no monetarios que realizo el hogar en el periodo de referencia, las erogaciones financieras, el gasto de tarjetas gastos realizados por el hogar que fueron cubiertos mediante alguna tarjeta de crédito bancario o comercial., características y ocupacionales de los integrantes del hogar, ingresos y percepciones de capital de cada uno de los integrantes del hogar, gasto por persona, la actividad de los integrantes del hogar, ingresos y gastos de los negocios del hogar dedicados tanto a las actividades agrícolas, forestales y de tala, como a actividades de cría, explotación y productos derivaos de la pesca y caza, ingresos y gastos de los negocios del hogar dedicados a las actividades industriales, comerciales y de servicios y sus características propias.

Cobertura temporal

El periodo temporal que se lleva acabo se consideran planes a mediano plazo ya que su periodo de tiempo es de un año, para este caso sería el periodo del año 2018.

Tamaño de muestra

El tamaño de la muestra efectiva es de 87,826 viviendas.

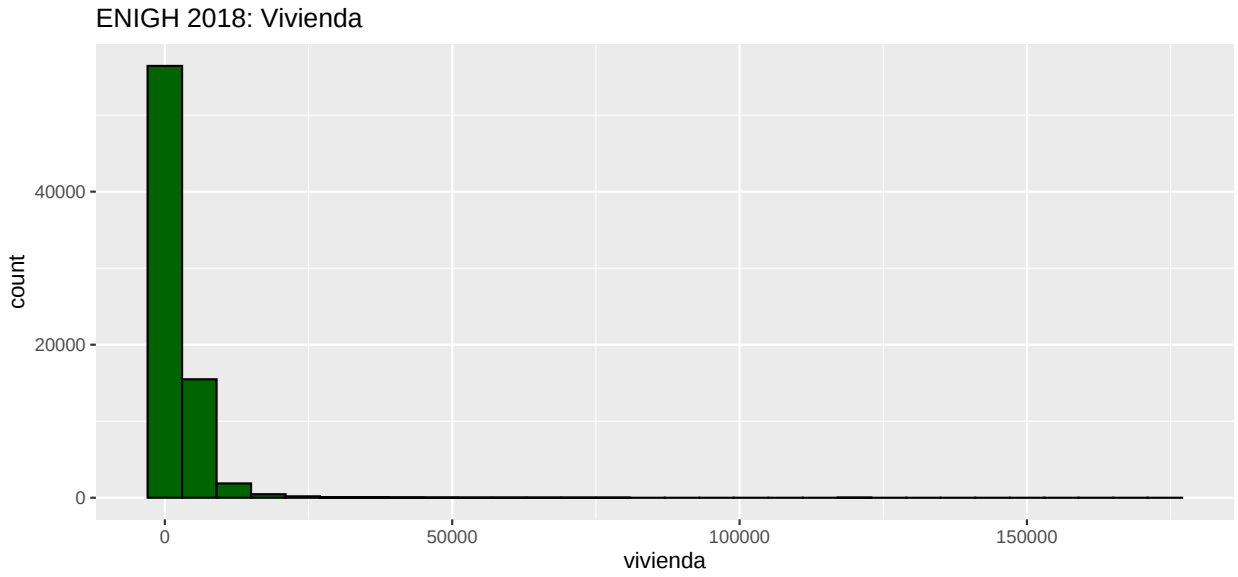
2. Generacion de variables

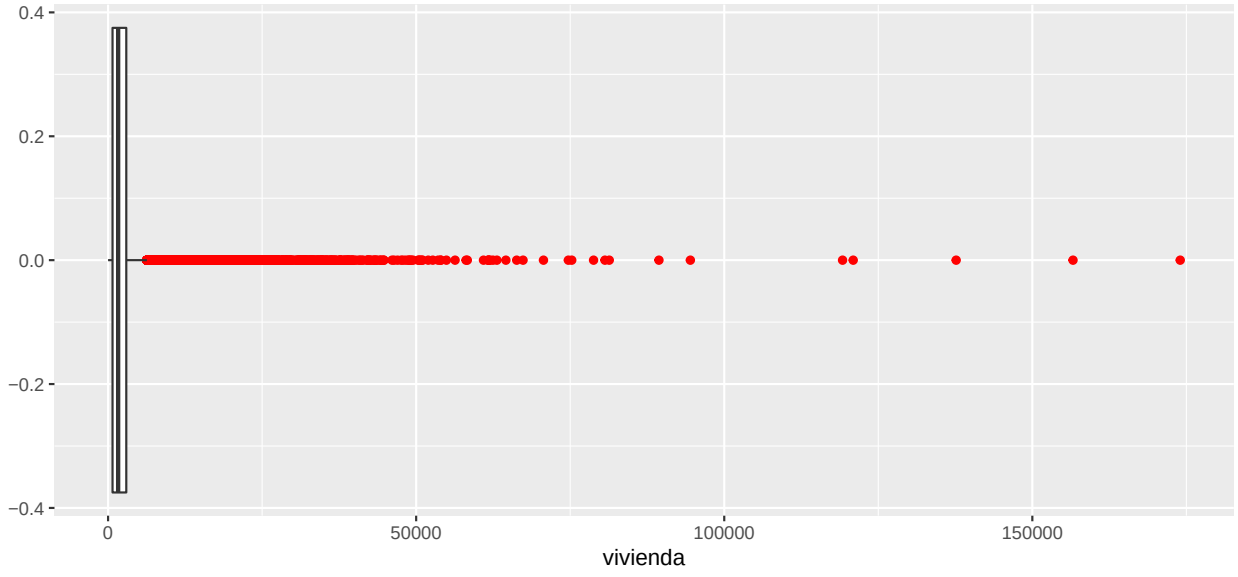
En el documento “VariablesGeneradas-Analisis_concentradoHogarENIGH2018.R” se anexa el código que se utilizo para generar las variables con la base de datos “ConcentradoHogarENIGH2018.xlsx” a un archivo llamado “VariablesExtra.csv” donde únicamente se anexan las variables del ejercicio a generar.

3. Análisis descriptivo de las variables generadas

Haciendo estadística descriptiva para la variable **vivienda**

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



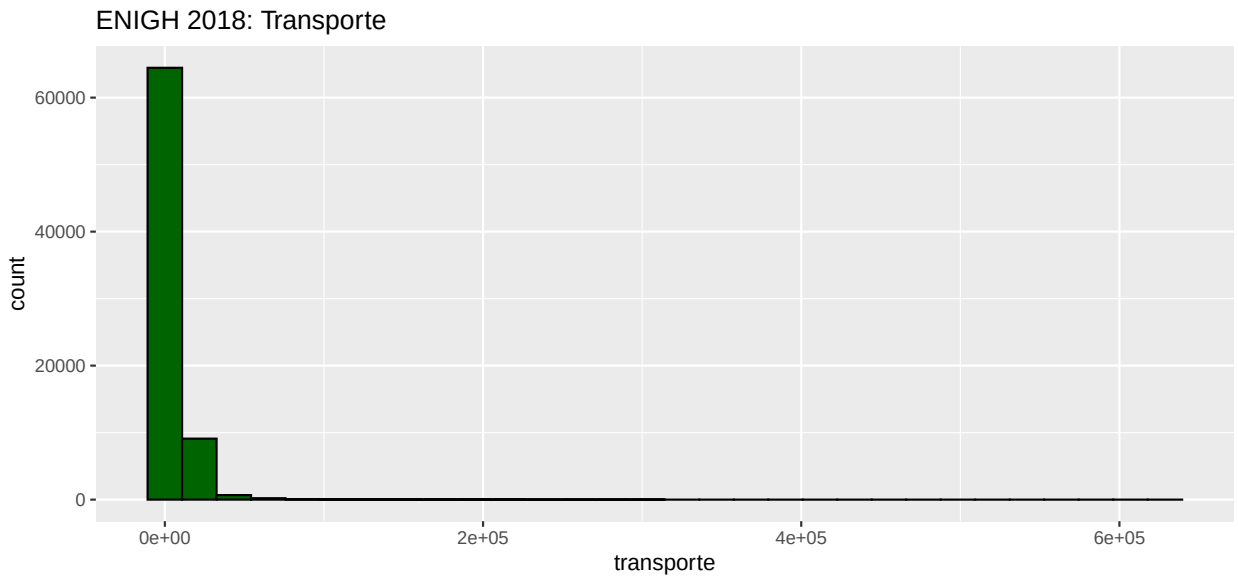


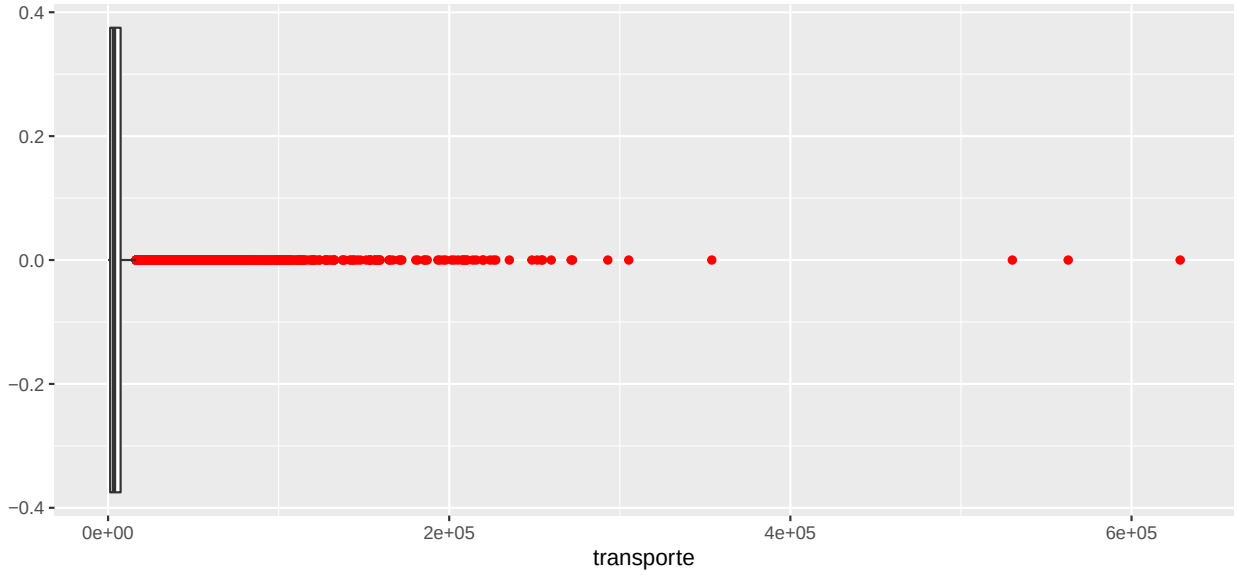

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0     725    1636    2526    2956  174000
```

Notemos que para esta variable la mayor parte de los gastos por vivienda se concentran debajo del 3er quantil es decir por debajo de los 2956 pesos

Haciendo estadistica descriptiva para la variable **transporte**

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



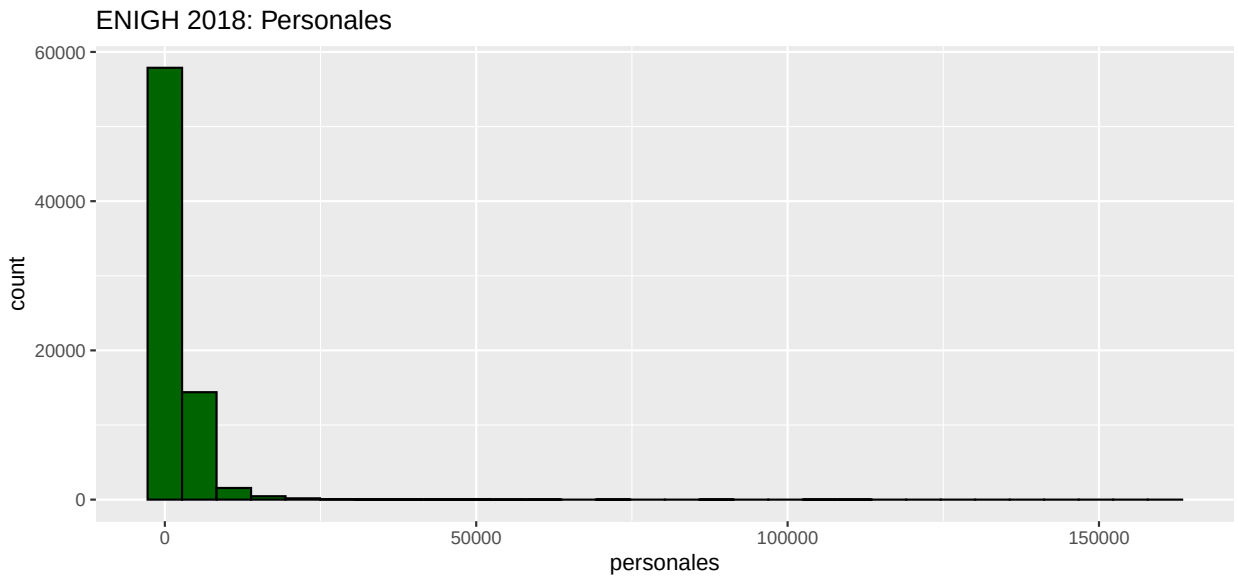


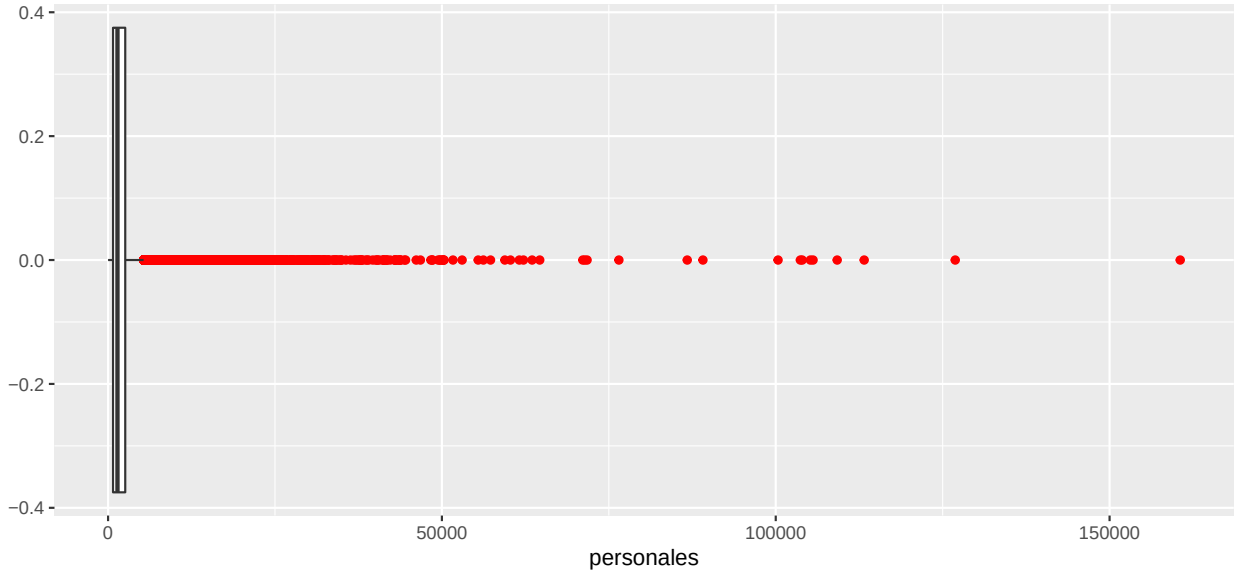

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    1228    3484    5918    7353  628556
```

Notemos que para esta variable la mayor parte de la variable transporte se concentran debajo del 3er quantil es decir por debajo de los 7353 pesos mientras que hay datos que se alejan demasiado como el maximo que es de 628,556

Haciendo estadistica descriptiva para la variable **personales**

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



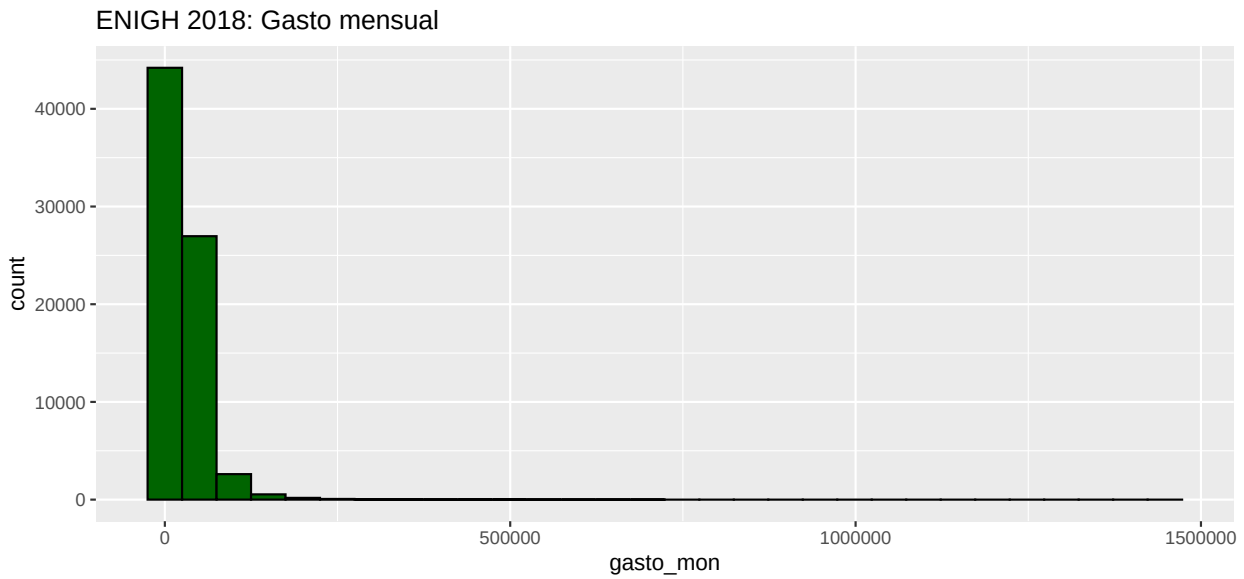


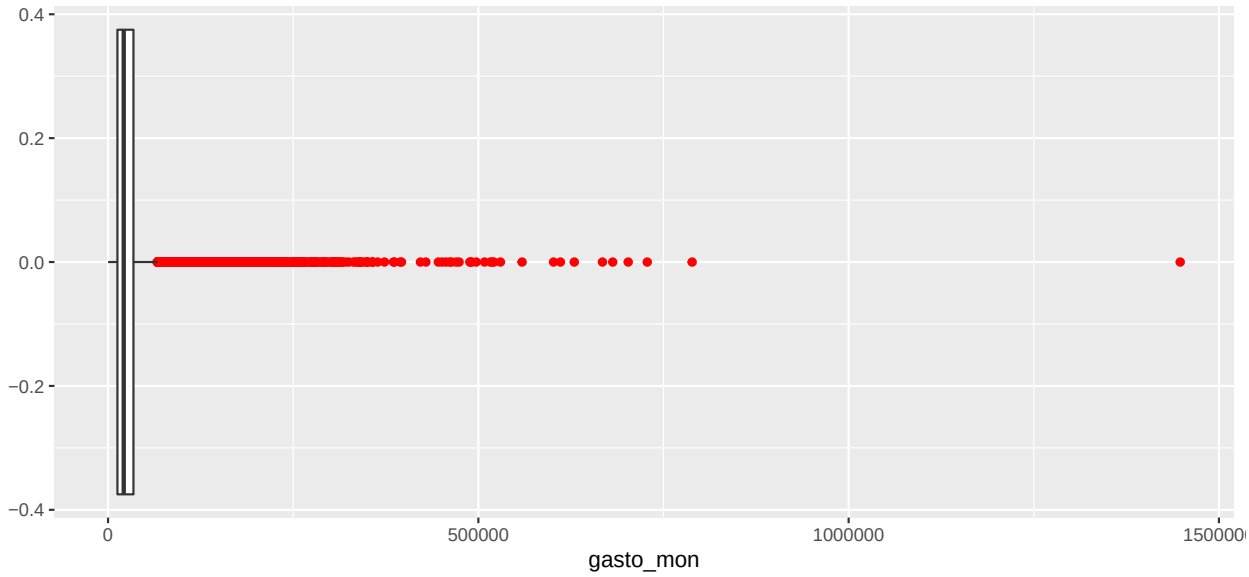

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##      0.0    743.2   1410.0   2219.8   2572.8 160580.6
```

Notemos que para esta variable la mayor frecuencia se encuentra debajo del 3er quantil es decir por debajo de los 2572.8 pesos mientras que hay datos que se alejan demasiado como el maximo que es de 160580.6 o nulos.

Haciendo estadistica descriptiva para la variable **gasto_mon**

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



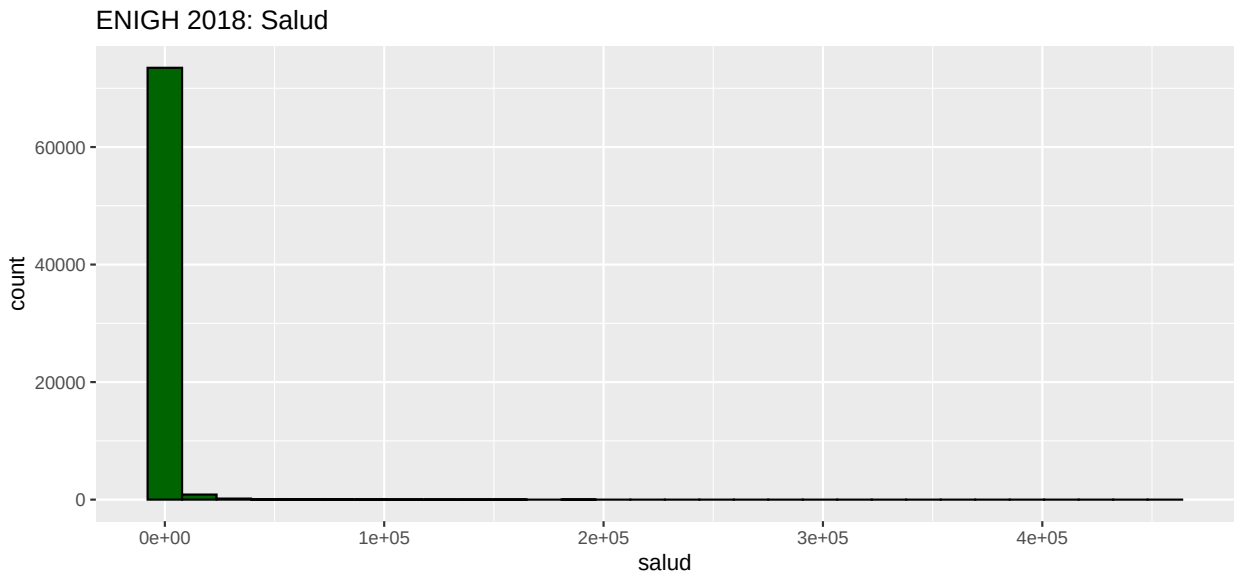


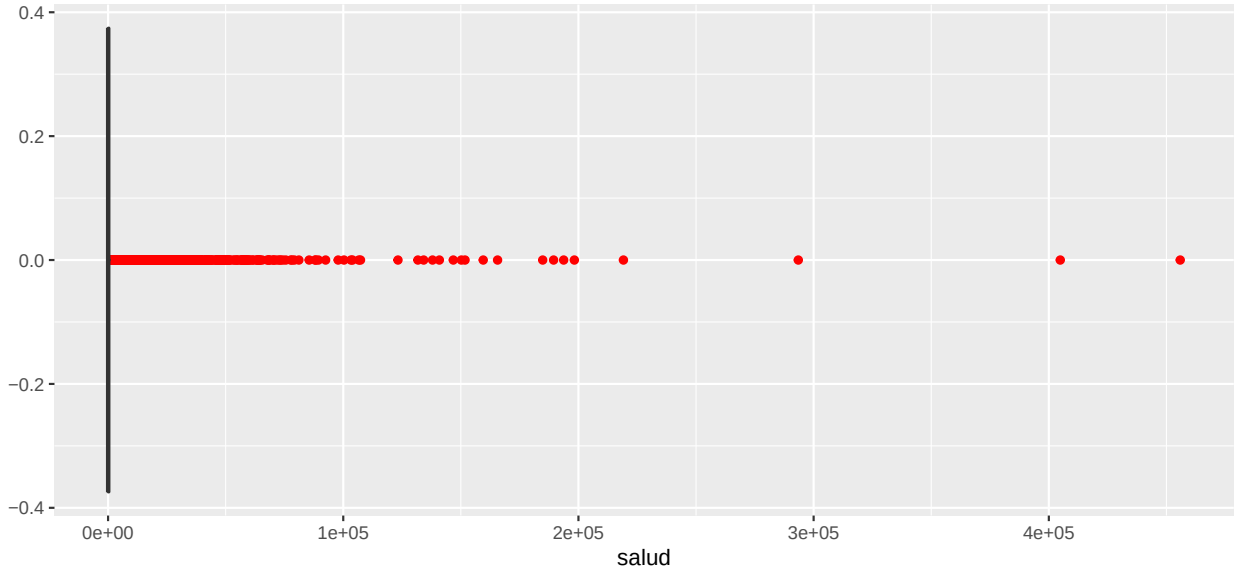

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0   12652   21164   28056   34264 1447746
```

El promedio para el gasto mensual es de 28056, y que la mayor frecuencia de estos datos esta por debajo del 3er quantil que equivale a 34264

Haciendo estadistica descriptiva para la variable **salud**

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```






```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##      0.0      0.0     23.5    797.8    450.0 455869.6
```

Podemos notar que en promedio se gasta en cosas de salud 797.8 pesos, habiendo casos donde no se gasta en salud y otros graves donde el costo de salud se incrementa por tener una enfermedad hasta en casi 455869.6 pesos

4. Análisis de variables de "ConcentradoHogarENIGH2018.xlsx"

Listando las ultimas 20 observaciones de dos variables: "mayores" y "menores"

Tomaremos la variable "mayores" y "menores" para mostrar las ultimas 20 observaciones

	folioviv	mayores	menores
74628	3260798022	2	0
74629	3260798023	3	0
74630	3260798024	2	0
74631	3260798413	5	0
74632	3260798414	2	2
74633	3260798415	4	2
74634	3260798416	2	0
74635	3260798418	4	0
74636	3260798807	4	2
74637	3260798808	2	0
74638	3260798809	4	0
74639	3260798810	4	1
74640	3260798811	2	0
74641	3260798812	1	0
74642	3260798901	3	2
74643	3260798902	2	0
74644	3260798903	2	0
74645	3260798904	5	2
74646	3260798905	1	0
74647	3260798906	2	0

Listando las 10 observaciones mas altas para sueldos,ing_trab,gasto_mon, mayores y menores

Listando las 10 observaciones mas altas para **sueldos**

	folioviv	sueldos
43214	1901398801	959016.4
22912	919519304	733695.7
21556	902658201	713038.7
5131	260301506	686739.1
10483	501910002	675000.0
23114	960004920	672652.2
54113	2302287805	649180.3
22911	919519302	645652.2
22717	917067801	590163.9
22718	917067802	590163.9

Listando las 10 observaciones mas altas para **ing_trab**

	folioviv	ing_trab
43214	1901398801	2797131
64572	2804081804	2081299
7990	360196911	1936467
23114	960004920	1778185
22912	919519304	1589674
5131	260301506	1492092
21556	902658201	1483757
22911	919519302	1467391
22718	917067802	1465533
2797	202988905	1447552

Listando las 10 observaciones mas altas para **gasto_mon**

	folioviv	gasto_mon
43210	1901367905	1447746.1
41623	1800749406	788680.7
20423	860308809	728120.8
22717	917067801	702380.1
22753	917215606	681451.1
50568	2201087602	667589.8
43214	1901398801	629546.6
22705	916965302	610841.7
7960	360185906	601511.5
54113	2302287805	559055.0

Podemos notar que las viviendas con folio:

1. 1901398801
2. 917067802
3. 919519302
4. 2302287805

Siendo las principales apareciendo en los sueldos, ing_trab y gasto_mon mas altos

Listando otras 10 observaciones altas de 2 variables

Listando las 10 observaciones mas altas para **becas**

	folioviv	becas
16128	702239702	97377.03
65003	2806712602	82622.94
5629	300001001	71902.16
55294	2401150906	58695.65
31400	1302752809	49180.32
58231	2502801903	48423.90
13451	600480503	47543.46
18240	802870806	44262.29
845	101430001	44021.73
12006	506020403	44021.73

Listando las 10 observaciones mas altas para **donativos**

	folioviv	sueldos
43214	1901398801	959016.4
22912	919519304	733695.7
21556	902658201	713038.7
5131	260301506	686739.1
10483	501910002	675000.0
23114	960004920	672652.2
54113	2302287805	649180.3
22911	919519302	645652.2
22717	917067801	590163.9
22718	917067802	590163.9

5. Modelo lineal

Se propone el modelo ajustado:

$$\hat{gastomon} = \hat{\alpha}_0 + \hat{\alpha}_1 trabajo_i + \hat{\alpha}_2 jubilacion_i + \hat{\alpha}_3 negocio_i + \hat{\alpha}_4 transfhog_i$$
$$\hat{\alpha}_i > 0 \quad \text{i=1,2,3,4}$$

Estimando el modelo con MCO

```
##  (Intercept)      trabajo   jubilacion      negocio   transf_hog
## 1.362855e+04 4.884662e-01 2.069214e-01 2.565495e-01 1.580828e-01
```

El modelo queda finalmente:

$$\hat{gastomon} = 13628.55 + 0.48846 trabajo_i + 0.20692 jubilacion_i + .25654 negocio_i + 0.15808 transfhog_i$$

Inferencia de los coeficientes

```
##
## Call:
## lm(formula = gasto_mon ~ trabajo + jubilacion + negocio + transf_hog,
##     data = enigh_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -865805   -9880   -4133    4572 1434118
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.363e+04  1.106e+02 123.256  < 2e-16 ***
## trabajo     4.885e-01  2.412e-03 202.548  < 2e-16 ***
## jubilacion  2.069e-01  3.926e-03  52.708  < 2e-16 ***
## negocio     2.565e-01  5.459e-03  46.996  < 2e-16 ***
## transf_hog  1.581e-01  2.152e-02   7.347 2.05e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22350 on 74642 degrees of freedom
## Multiple R-squared:  0.3736, Adjusted R-squared:  0.3736
## F-statistic: 1.113e+04 on 4 and 74642 DF,  p-value: < 2.2e-16
```

La region de rechazo (tomando por ejemplo trabajo=0) de los parametros (t value) se rechaza si $|t| \geq 1.96$ Notemos que el valor 1.96 corresponde a la probabilidad de la tabla T (Student) un nivel de confianza del 95% con n-2=74645 grados de libertad.

Como todos los **t value** son mayores a 1.96 se rechaza la hipotesis nula; es decir son significantes para el modelo

Coeficiente de determinación

```
## [1] 0.3736123
```

Como el cooeficiente de determinacion R es de : 0.3736 ajusta el modelo en un 37.36%

Coeficiente de determinación ajustado

```
## [1] 0.3735787
```

Residuales y valores ajustados

Se mostrarán los primeros 5 valores **residuales**

```
##           1           2           3           4           5
## -23660.3552  35942.5034  17536.5527    826.8109 -14019.7625
```

Se mostrarán los primeros 5 valores **ajustados**

```
##        1        2        3        4        5
## 41466.13 17635.20 82934.68 18513.25 26789.09
```

A manera de resumen se muestran las primeras 5 observaciones reales de la variable **gasto_mon**

```
## [1]  17805.77  53577.70 100471.23  19340.06  12769.33
```

6. Modelo logaritmico

Valores de las constantes

Ahora se propone el modelo:

El modelo queda finalmente:

$$\log(\hat{gastomon}) = \hat{\beta}_0 + \hat{\beta}_1 \log(trabajo) + \hat{\beta}_2 \log(jubilacion) + \hat{\beta}_3 \log(negocio) + \hat{\beta}_4 log(transfhog)$$
$$\hat{\beta}_j > 0 \quad \text{j=1,2,3,4}$$

Para poder aplicar este modelo debemos tener en cuenta que:

1. El $log(x) > 0$ .Es decir no existen logaritmos de valores menores o iguales a 0
2. Se aplica un filtro que elimina valores nulos (que son renglones con valores a cero)
3. El tamaño del filtro disminuye el del original (por suprimir valores nulos)

El nuevo tamaño de datos por columna es: 443

Estimacion del modelo por MCO

```
##     (Intercept)    log(trabajo) log(jubilacion)    log(negocio) log(transf_hog)
##    5.7491984041    0.1717730990    0.2479575610    0.0785542168   -0.0007756742
```

El modelo corresponde a:

$$\log(\hat{gastomon}) = 5.74919 + 0.17177 \log(trabajo_i) + 0.24795 \log(jubilacion_i) + 0.07855 \log(negocio_i) - 0.0007 log(transfhog_i)$$

Inferencia a los coeficientes

```
summary(mco_log)
```

```
##
## Call:
## lm(formula = log(gasto_mon) ~ log(trabajo) + log(jubilacion) +
##     log(negocio) + log(transf_hog), data = datalog)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.02017 -0.39547 -0.00785  0.36808  2.99190
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.7491984  0.3496407  16.443  < 2e-16 ***
## log(trabajo)     0.1717731  0.0190610   9.012  < 2e-16 ***
## log(jubilacion)  0.2479576  0.0292666   8.472 3.69e-16 ***
## log(negocio)     0.0785542  0.0177693   4.421 1.24e-05 ***
## log(transf_hog) -0.0007757  0.0236762  -0.033    0.974
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5972 on 438 degrees of freedom
## Multiple R-squared:  0.3152, Adjusted R-squared:  0.309
## F-statistic: 50.41 on 4 and 438 DF,  p-value: < 2.2e-16
```

La region de rechazo (tomando por ejemplo log(trabajo)=0) de los parametros (t value) se rechaza si $|t| \geq 1.96$ Notemos que el valor 1.96 corresponde a la probabilidad de la tabla T (Student) un nivel de confianza del 95% con n-2=441 grados de libertad.

Como la mayoria de los **t value** son mayores a 1.96 se rechaza la hipotesis nula; es decir son significantes para el modelo.

El unico valor **t value** con valor -0.033 que corresponde al estadistico t de la variable log(transf_hog) no es mayor a 1.96 no se rechaza la hipotesis nula es decir no es significativo para el modelo el cual podria ser omitido al modelo (Claro que depende de otros factores y análisis):

$$\log(gastomon) = 5.7491984041 + 0.1717730990 \log(trabajo) + 0.247957561 \log(jubilacion) + 0.0785542168 \log(negocio)$$

Coeficiente de determinación

```
## [1] 0.3152361
```

Como el cooeficiente de determinación R^2 es de : 0.3152 ajusta el modelo en un 31.52%

Coeficiente de determinacion ajustado

```
## [1] 0.3152361
```

Residuales y valores ajustados

Se mostrarán los primeros 5 valores **residuales**

```
##         10         62        171        199        206
##  0.3384478  1.2863579  0.5029658  0.3538897 -0.3219813
```

Se mostrarán los primeros 5 valores **ajustados**

```
##        10        62       171       199       206
## 10.519073 10.283676 10.564504 10.809007  9.639279
```

A manera de resumen se muestran las primeras 5 observaciones reales de la variable **gasto_mon**

```
## [1] 10.857521 11.570034 11.067469 11.162897  9.317298
```

Comparando modelo lineal y logaritmico

Para el modelo lineal se tiene que los coeficientes son:

	x
(Intercept)	1.362855e+04
trabajo	4.884662e-01
jubilacion	2.069214e-01
negocio	2.565495e-01
transf_hog	1.580828e-01

Para el modelo logaritmico se tiene que los coeficientes son:

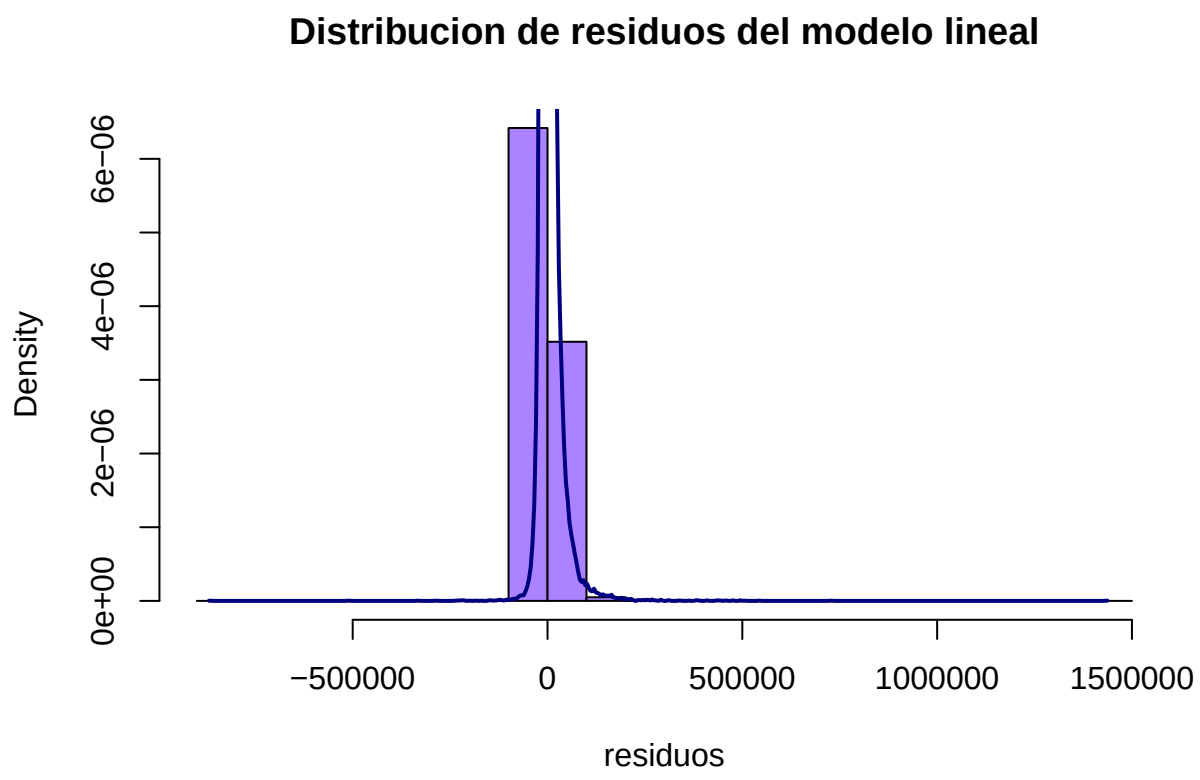
	x
(Intercept)	5.7491984
log(trabajo)	0.1717731
log(jubilacion)	0.2479576
log(negocio)	0.0785542
log(transf_hog)	-0.0007757

1. Podemos notar que los valores obtenidos difieren entre si a excepcion del valor del coeficiente en jubilacion que son algo parecidos.
2. Ademas de que en el modelo logaritmico podria ser omitida la variable **transf_hog**
3. El coeficiente de determinacion mas alto es el del modelo lineal; es decir ajusta mejor los datos con una diferencia aproximada de 5.837%

Normalidad

Mostrando el histgrama de los residuos del modelo lineal

```
hist(residuos, freq=FALSE ,main="Distribucion de residuos del modelo lineal",
     col="mediumpurple1")
lines(density(residuos),col="navyblue",lwd=2)
```



Calculando su media y varianza se tiene que:

```
mean(residuos)
```

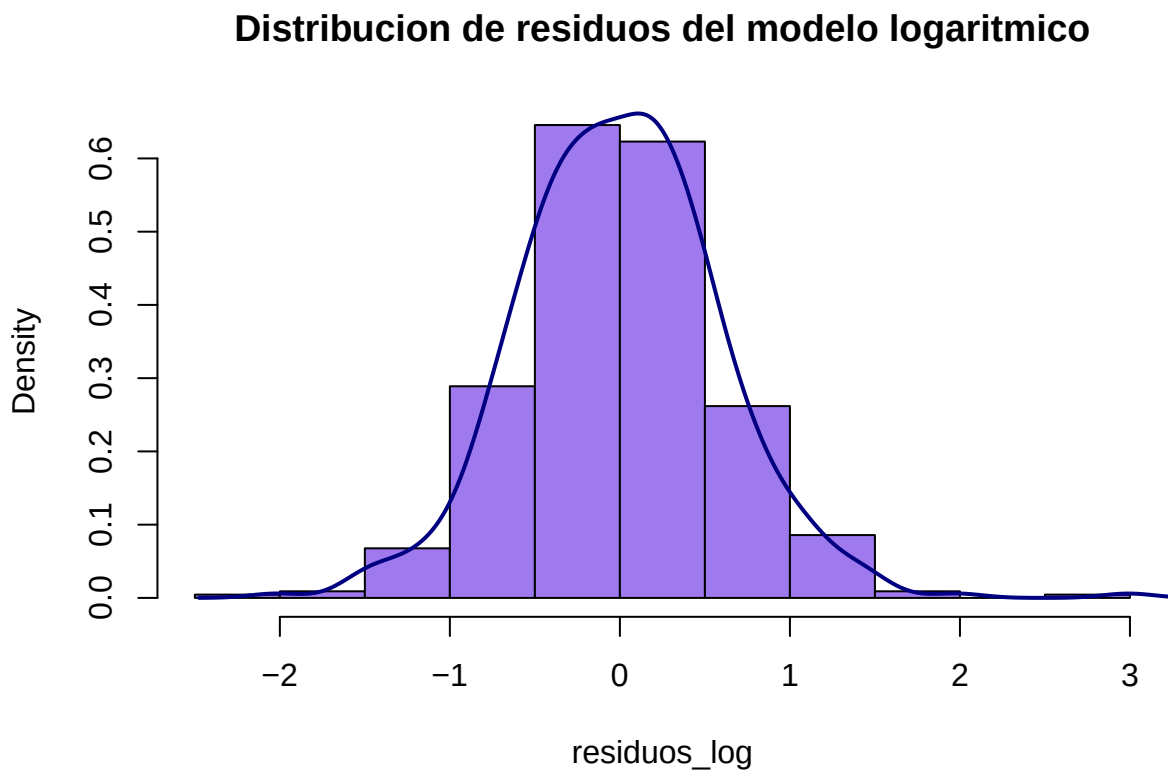
```
## [1] -3.302641e-12
```

```
sd(residuos)
```

```
## [1] 22354
```

Mostrando el histgrama de los residuos del modelo logaritmico

```
hist(residuos_log, freq=FALSE,main="Distribucion de residuos del modelo logaritmico",
     col="mediumpurple2")
lines(density(residuos_log),col="navyblue",lwd=2)
```



Calculando su media y varianza se tiene que:

```
mean(residuos_log)
```

```
## [1] 1.266378e-17
```

```
sd(residuos_log)
```

```
## [1] 0.5944836
```