UNAP
UNIVERSIDAD ARTURO PRAT
DEL ESTADO DE CHILE

FIAUNAP
FACULTAD DE INGENIERÍA Y ARQUITECTURA
UNIVERSIDAD ARTURO PRAT

Informe Laboratorio

Estadística Aplicada

"*Chicago*" – BBDD

Estudiante:

Alejandro Dinamarca Cáceres

Docente:

Francisco Cartes Arenas

Fecha:

23/08/2020

Ramo:

Estadística Aplicada

Índice

Contenido

1. Entendiendo la BBDD “Chicago”

1.1. Método y fuentes.

El presente trabajo consiste en una investigación de la base de datos Chicago, bastante resumida a través del software R STUDIO, que se basa en el lenguaje de programación R con un enfoque al análisis de datos. **Si bien, habrá utilización de código en el documento, se obviará el que se considere no relevante**.

Antes de iniciar el análisis de la data recopilada, es preciso saber con qué información se trabaja, así mismo como sus fuentes y orígenes:

“*In a study of insurance availability in Chicago, the U.S. Commission on Civil Rights attempted to examine charges by several community organizations that insurance companies were redlining their neighborhoods, i.e. canceling policies or refusing to insure or renew. First the Illinois Department of Insurance provided the number of cancellations, non-renewals, new policies, and renewals of homeowners and residential fire insurance policies by ZIP code for the months of December 1977 through February 1978. The companies that provided this information account for more than 70% of the homeowners insurance policies written in the City of Chicago. The department also supplied the number of FAIR plan policies written an renewed in Chicago by zip code for the months of December 1977 through May 1978. Since most FAIR plan policyholders secure such coverage only after they have been rejected by the voluntary market, rather than as a result of a preference for that type of insurance, the distribution of FAIR plan policies is another measure of insurance availability in the voluntary market. Secondly, the Chicago Police Department provided crime data, by beat, on all thefts for the year 1975. Most Insurance companies claim to base their underwriting activities on loss data from the preceding years, i.e. a 2-3 year lag seems reasonable for analysis purposes. the Chicago Fire Department provided similar data on fires occurring during 1975. These fire and theft data were organized by zip code. Finally the US Bureau of the census supplied data on racial composition, income and age and value of residential units for each ZIP code in Chicago. To adjust for these differences in the populations size associated with different ZIP code areas, the theft data were expressed as incidents per 1,000 population and the fire and insurance data as incidents per 100 housing units.*”

Biostatistics. (s. f.). Recuperado 20 de agosto de 2020, de http://www.biostat.jhsph.edu/~iruczins/teaching/jf/ch12.pdf

La cita anterior nos permite dar cuenta que, la base de datos a trabajar representa **más del 70%** de las pólizas de seguro para propietarios de vivienda escritas en la ciudad de Chicago, con distintos datos recopilados a partir de estas pólizas, seguros, políticas de viviendas, datos demográficos, incendios y robos gracias a distintas fuentes, entre ellas, bomberos, policías, etc. De la misma ciudad. Además, se recalca que todos los datos corresponden a variables cuantitativas.

1.2. Variables.

Como es costumbre, todos estos datos se almacenan en forma de colección de forma tabulada, lo que viene a ser el "dataset", aquí una captura de las primeras cinco filas:

	race	fire	theft	age	volact	involact	income
60626	10.0	6.2	29	60.4	5.3	0.0	11744
60640	22.2	9.5	44	76.5	3.1	0.1	9323
60613	19.6	10.5	36	73.5	4.8	1.2	9948
60657	17.3	7.7	37	66.9	5.7	0.5	10656
60614	24.5	8.6	53	81.4	5.9	0.7	9730

De aquí, se aprecian los códigos ZIP al inicio de las filas y en los **encabezados de las columnas, las variables que nos son relevantes**:

- **Race**: Composición racial en porcentaje minoritario.
- **Fire**: Incendios cada 100 viviendas.
- **Theft**: Asaltos cada 1000 habitantes.
- **Age**: Porcentaje de viviendas construidas antes de 1939.
- **Volact**: Nuevas pólizas por dueños de vivienda más renovaciones, menos cancelaciones y no renovaciones cada 100 viviendas bajo vieja política de vivienda.
- **Involact**: Nuevas pólizas y renovaciones bajo nuevo plan de vivienda FAIR.
- **Income**: Ingreso medio familiar.

Una vez introducida la base de datos, se inicia la investigación con el **IDE R STUDIO.**

2. Análisis Exploratorio

Una vez teniendo en conocimiento la estructura y variables, se procede a la obtención para posterior análisis exploratorio de la colección de datos.

2.1. Medidas de resumen: obtención de estadísticos descriptivos.

En cuanto a las medidas de resumen, sean de tendencia central o de dispersión, se ocupó, en primera instancia, el comando `summary()`:

```
> summary(data)
 raza....chicago.race fuego....chicago.fire robo....chicago.theft edad....chicago.age politicavivienda1....chicago.volact politicavivienda2....chicago.involact
 Min.   : 1.00        Min.   : 2.000        Min.   : 9.00         Min.   : 7.70       Min.   : 0.50                     Min.   :0.0000
 1st Qu.: 3.75        1st Qu.: 5.650        1st Qu.:23.00         1st Qu.:48.60       1st Qu.: 3.10                     1st Qu.:0.0000
 Median :24.50        Median :10.400        Median :31.00         Median :65.00       Median : 5.90                     Median :0.4000
 Mean   :34.99        Mean   : 9.782        Mean   :29.05         Mean   :60.97       Mean   : 6.53                     Mean   :0.6149
 3rd Qu.:57.65        3rd Qu.:12.279        3rd Qu.:34.00         3rd Qu.:77.30       3rd Qu.: 9.65                     3rd Qu.:0.9000
 Max.   :99.70        Max.   :21.800        Max.   :46.00         Max.   :90.10       Max.   :14.30                     Max.   :2.2000
 ingresofamiliar....chicago.income
 Min.   : 5583
 1st Qu.: 8447
 Median :10694
 Mean   :10466
 3rd Qu.:11810
 Max.   :16250
```

A pesar de su gran utilidad, hay estadísticos descriptivos que faltan, como, por ejemplo, la desviación estándar, por lo cual, se procedió a realizar un código que nos obtiene algunos de estos datos faltantes, que fue realizado a través de un ciclo iterativo que recorre cada valor del dataset, obteniendo por cada columna, la medida deseada, repitiéndose según n columnas. Este código es el que se adjunta a continuación:

```
n <- 1
data <- data.frame(raza <- chicago$race, fuego <- chicago$fire, robo <- chicago$theft, edad <- chicago$age, politicaVivienda1 <- chicago$volact,
                   politicavivienda2 <- chicago$involact, ingresofamiliar <- chicago$income)
#summary(data)
#head(data)
dataMediaDF <- data.frame()
textoCol <- c("Composicion Racial", "Incendios por 100 Viviendas", "Robo por 1000 Habitantes", "Porcentaje Viviendas 1939", "Nueva Politica Vivienda",
              "Nueva Politica FAIR", "Ingreso Medio Familiar")
textoRow <- c("Promedio", "Mediana", "Moda unimodal", "Varianza", "Desviacion estandar", "CV", "Coef. Asimetria", "Curtosis")
for(val in data){
  dataMediaDF[1,n] <- round(mean(na.omit(data[,n])), digits=3)
  dataMediaDF[2,n] <- round(median(na.omit(data[,n])), digits=3)
  dataMediaDF[3,n] <- round(mfv(na.omit(data[,n]), na_rm=FALSE)[1], digits=3)
  dataMediaDF[4,n] <- round(var(na.omit(data[,n])), digits=3)
  dataMediaDF[5,n] <- round(sqrt(var(na.omit(data[,n]))), digits=3)
  dataMediaDF[6,n] <- round(dataMediaDF[5,n]/dataMediaDF[1,n], digits=3)
  dataMediaDF[7,n] <- round(skew((data[,n])), digits=3)
  dataMediaDF[8,n] <- round(kurtosi((data[,n])), digits=3)
  n <- n + 1
  if(n == 8){
    names(dataMediaDF) <- textoCol
    rownames(dataMediaDF) <- textoRow
  }
}
datamedidasDF <- data.frame(dataMediaDF)
datamedidasDF
```

El resultado del código anterior:

```
> datamedidasDF
                  Composicion.Racial Incendios.por.100.Viviendas Robo.por.1000.Habitantes Porcentaje.Viviendas.1939 Vieja.Politica.Vivienda
Promedio                      34.985                       12.279                    32.362                    60.328                 6.530
Mediana                       24.500                       10.400                    29.000                    65.000                 5.900
Moda unimodal                  1.000                        2.200                    27.000                    89.800                 3.100
Varianza                    1061.953                       86.532                   496.888                   509.629                15.733
Desviacion estandar           32.588                        9.302                    22.291                    22.575                 3.966
CV                             0.931                        0.758                     0.689                     0.374                 0.607
Coef. Asimetria                0.557                        1.271                     2.956                    -0.921                 0.271
Curtosis                      -1.054                        0.939                    12.552                     0.084                -1.179
                  Nueva.Politica.Vivienda Ingreso.Medio.Familiar
Promedio                            0.615                10695.830
Mediana                             0.400                10694.000
Moda unimodal                       0.000                 5583.000
Varianza                            0.402              7585606.666
Desviacion estandar                 0.634                 2754.198
CV                                  1.031                    0.258
Coef. Asimetria                     0.808                    1.155
Curtosis                           -0.433                    3.156
```

Finalmente, a través de la librería “formattable”, se customiza y ordena el dataset de los estadísticos descriptivos:

	Composicion Racial	Incendios por 100 Viviendas	Robo por 1000 Habitantes	Porcentaje Viviendas 1939	Vieja Politica Vivienda	Nueva Politica Vivienda	Ingreso Medio Familiar
Promedio	34.985	12.279	32.362	60.328	6.530	0.615	10695.830
Mediana	24.500	10.400	29.000	65.000	5.900	0.400	10694.000
Moda unimodal	1.000	2.200	27.000	89.800	3.100	0.000	5583.000
Varianza	1061.953	86.532	496.888	509.629	15.733	0.402	7585606.666
Desviacion estandar	32.588	9.302	22.291	22.575	3.966	0.634	2754.198
CV	0.931	0.758	0.689	0.374	0.607	1.031	0.258
Coef. Asimetria	0.557	1.271	2.956	-0.921	0.271	0.808	1.155
Curtosis	-1.054	0.939	12.552	0.084	-1.179	-0.433	3.156

2.2. Medidas de resumen: análisis.

Una vez obtenidos los estadísticos descriptivos faltantes es hora de someterlos a distintos análisis, los cuales se resumirán en los siguientes comentarios:

2.2.1. Composición racial (race) o `data$raza....chicago.race`.

En el caso de la composición racial es una variable que nos indica el porcentaje de extranjeros contra estadounidenses en términos de pólizas y seguros en vecindarios por los propietarios de viviendas de la ciudad de Chicago:

- Presenta una media del 34.895%. Esto quiere decir que, de los seguros, pólizas, etc. Adquiridos o renovados en vecindarios por los propietarios de viviendas, en promedio, el 65.105% era estadounidense.

- Una mediana del 24.5% que implica que, el 50% de los propietarios de viviendas en vecindarios presentaban un porcentaje menor o igual al 24.5% de extranjeros.

- Una moda del 1% que nos refleja que, en varios vecindarios se repitió una alta mayoría del 99% estadounidense.

- La desviación estándar calculada fue del 32.588%, estimando que, en promedio, la diferencia entre la media y cada uno de las pólizas y seguros de los vecindarios fue del 32.588%.

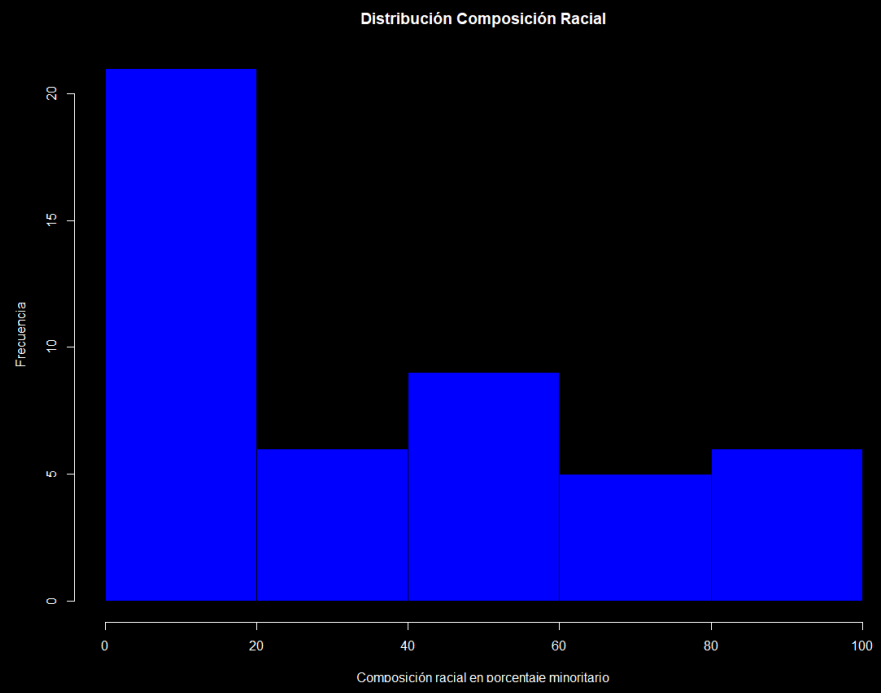
- El coeficiente de variación fue del 0.931 (93.1%), siendo superior al 50%, siendo entonces una distribución heterogénea por lo cual, los datos están dispersos, encontrándose lejanos a la media, por lo tanto, la media aritmética de la muestra (34.895%) no es representativa del conjunto de datos.

- El coeficiente de asimetría fue del 0.557, por lo cual, la distribución es asimétrica a la derecha, esto debería reflejarse en el histograma, gráfico de dispersión y diagramado de caja siendo la mayor frecuencia de datos encontrada a la izquierda de la distribución.

- La curtosis correspondió al -1.054, por lo cual es menor a 0, resultando en que la distribución es platicúrtica, resultando en que el grado de concentración de datos alrededor de la media es baja.

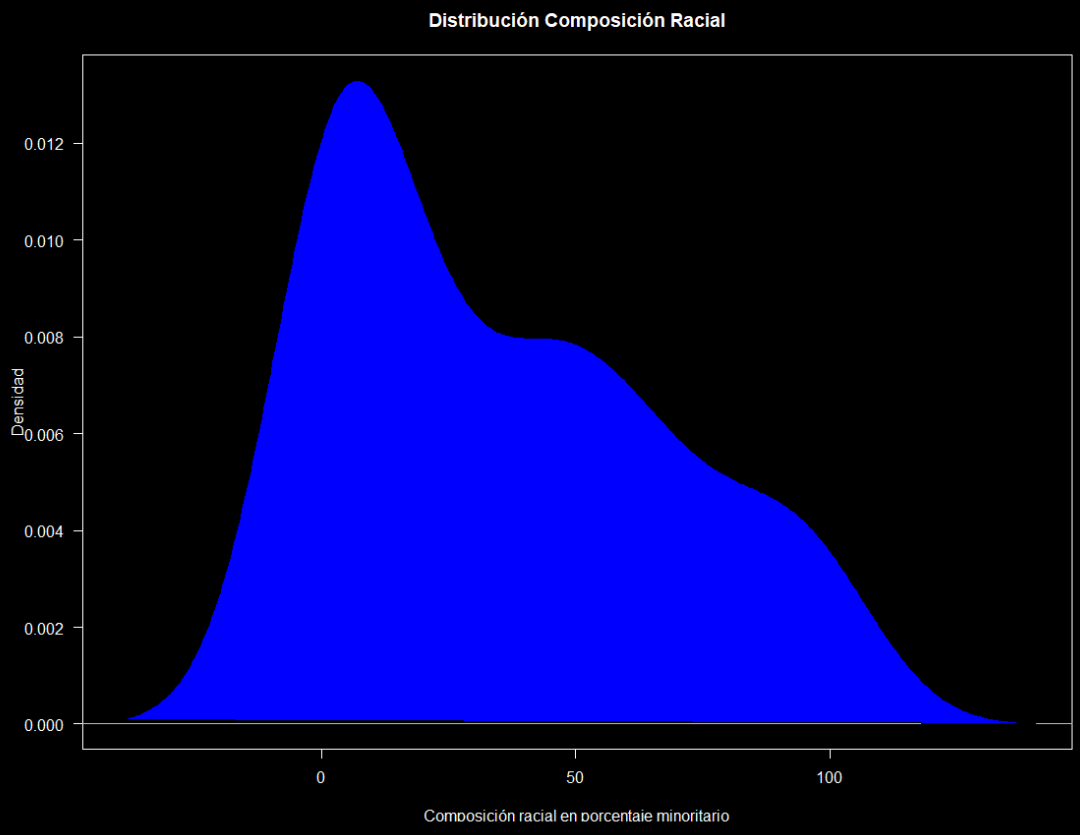
A continuación, se presentan el histograma, gráficos de densidad y caja “boxplot” para la variable.

Histograma de Composición racial (race) o `data$raza....chicago.race`



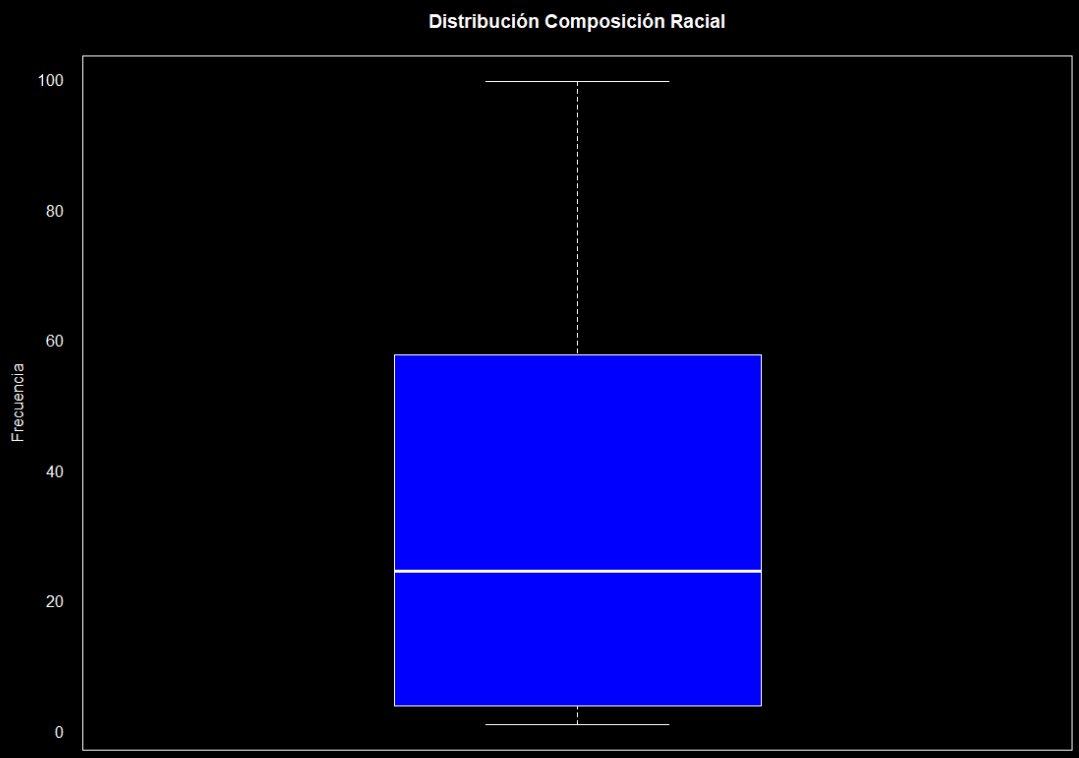
- El gráfico se acoge al análisis hecho en los estadísticos descriptivos, no se observa ninguna anomalía.

Densidad de Composición racial (race) o `data$raza....chicago.race`



- La misma situación que el gráfico anterior.

Diagramado de caja de Composición racial (race) o
`data$raza....chicago.race`



- La misma situación que el gráfico anterior.

2.2.2. Incendios por 100 viviendas (fire) o `data$fuego....chicago.fire`.

En el caso de los incendios, es una variable que nos entrega la cantidad de viviendas que sufrieron incendios, proporcionada por cada 100 viviendas:

- Presenta una media del 12.279. Esto quiere decir que, cada 100 viviendas, en promedio 12 presentaron incendios.

- Una mediana del 10.4, lo que implica que el 50% de viviendas presentaron una cantidad menor o igual a 10 incendios.

- Un valor de moda 2.2, que nos refleja que, cada 100 viviendas se repitieron en mayor frecuencia, un poco más de dos incendios.

- La desviación estándar calculada fue del 9.302 estimando que, en promedio, la diferencia entre la media y cada una de las 100 viviendas fue un poco más de 9 incendios.

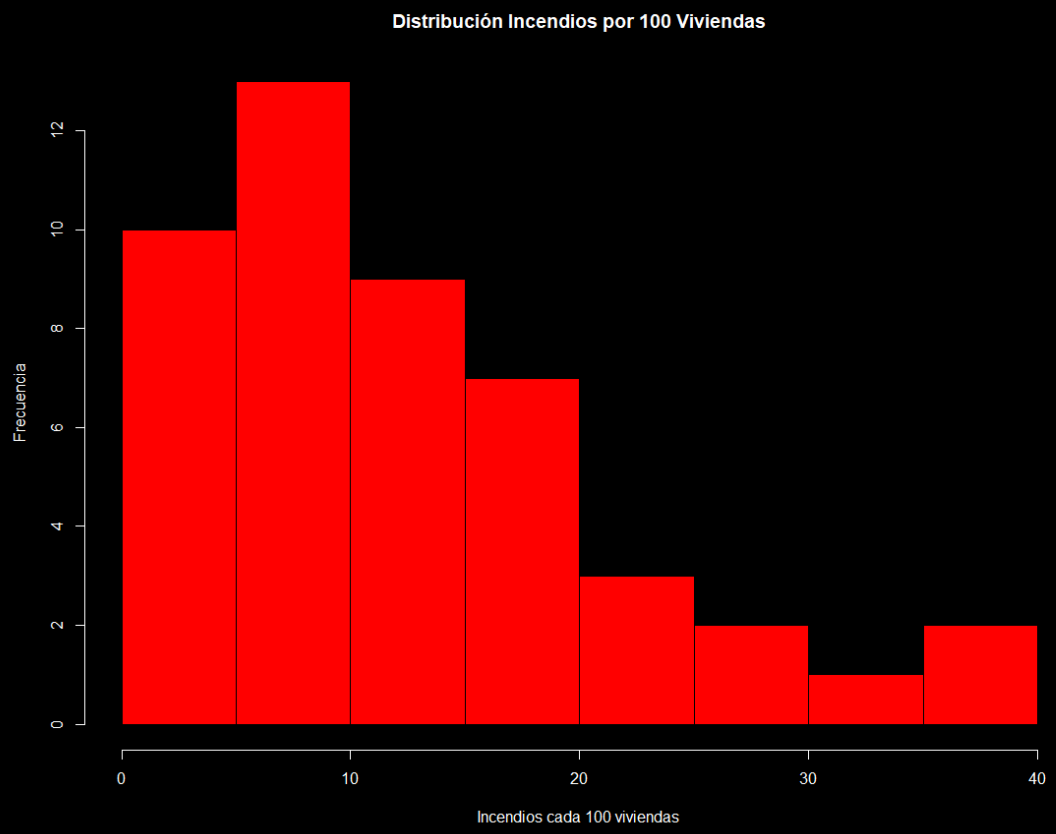
- El coeficiente de variación fue del 0.758 (75.8%), siendo superior al 50%, siendo entonces una distribución heterogénea por lo cual, los datos están dispersos, encontrándose lejanos a la media, por lo tanto, la media aritmética de la muestra (12.279) no es representativa del conjunto de datos.

- El coeficiente de asimetría fue del 1.271, siendo mayor a 0, por lo cual, la distribución es asimétrica a la derecha, esto debería reflejarse en el histograma, gráfico de dispersión y diagramado de caja siendo la mayor frecuencia de datos encontrada a la izquierda de la distribución.

- La curtosis correspondió al 0.939, por lo cual es mayor a 0, resultando en que la distribución es leptocúrtica, resultando en que el grado de concentración de datos alrededor de la media es alta.

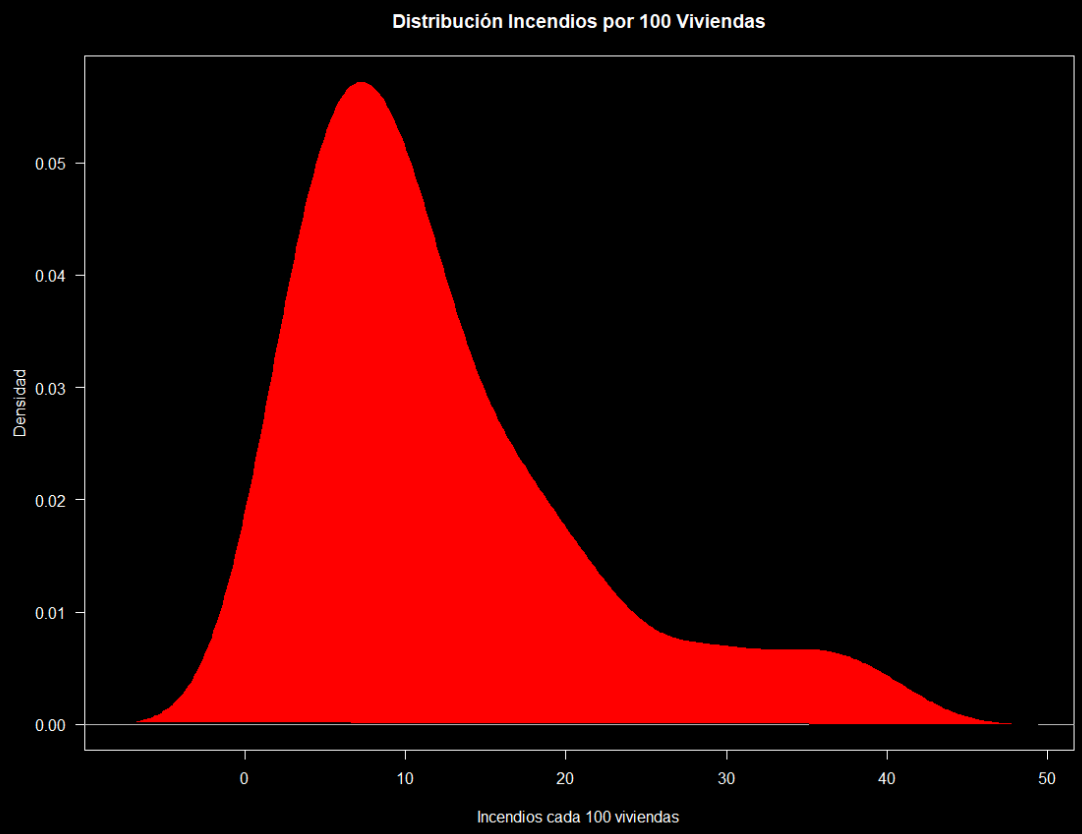
A continuación, se presentan el histograma, gráficos de densidad y caja "boxplot" para la variable.

Histograma de Incendios por 100 viviendas (fire) o
`data$fuego....chicago.fire`



- El gráfico se acoge al análisis hecho en los estadísticos descriptivos, no se observa ninguna anomalía.

Densidad de Incendios por 100 viviendas (fire) o
`data$fuego....chicago.fire`



- La misma situación que el gráfico anterior.

Diagramado de caja de Incendios por 100 viviendas (fire) o
`data$fuego....chicago.fire`

- El gráfico se acoge, por lo general, al análisis hecho en los estadísticos descriptivos, sin embargo, se observan tres datos anómalos a la distribución, es decir, tres outliers. Habrá que limpiarlos más adelante para que no interfiera en las correlaciones y por ende, en la exactitud del modelo.

2.2.4. Robo por 1000 habitantes (theft) o `data$robo....chicago.theft`.

Theft es una variable que nos entrega la cantidad de robos por cada 1000 habitantes:

- Presenta una media de 32.362. Esto quiere decir que, cada 1000 habitantes, en promedio un poco más de 32 presentaron asaltos.

- Una mediana del 29, lo que implica que el 50% de cada 1000 habitantes presentaron una cantidad menor o igual a 29 asaltos.

- Un valor de moda 27, que nos refleja que, cada 1000 habitantes se repitieron en mayor frecuencia 27 asaltos.

- La desviación estándar calculada fue del 22.291 estimando que, en promedio, la diferencia entre la media y cada uno de los 1000 habitantes fue un poco más de 22 asaltos.

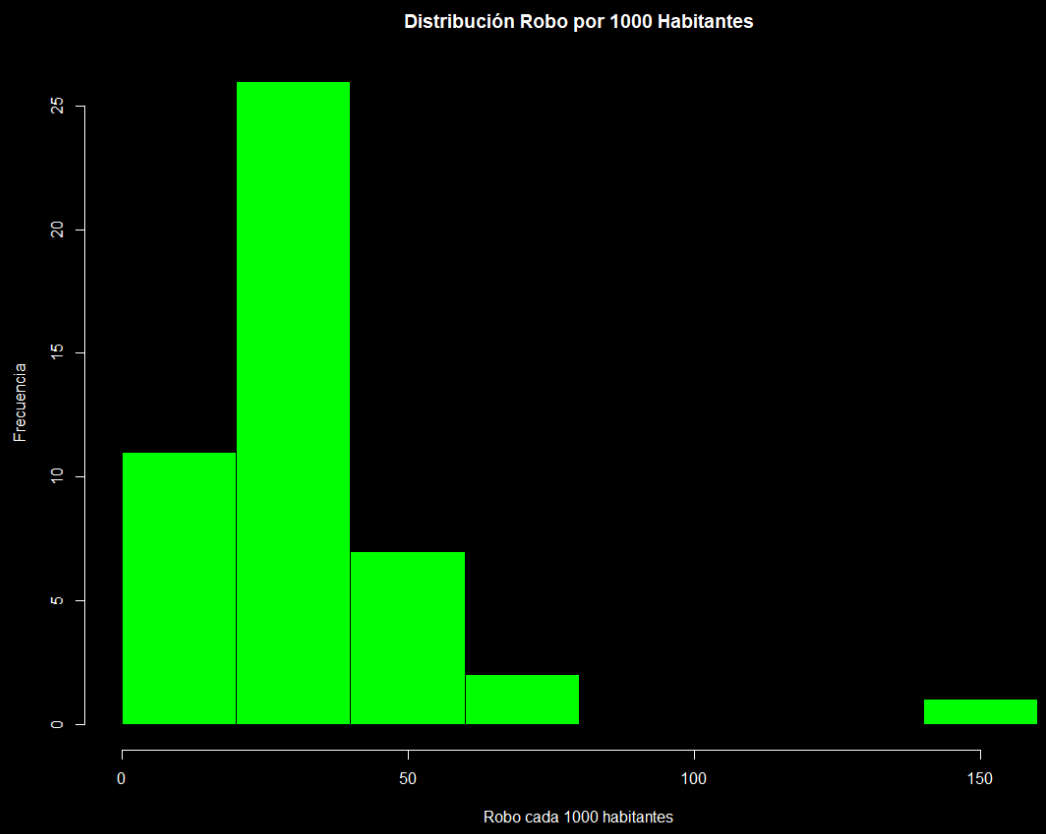
- El coeficiente de variación fue del 0.689 (68.9%), siendo superior al 50%, siendo entonces una distribución heterogénea por lo cual, los datos están dispersos, encontrándose lejanos a la media, por lo tanto, la media aritmética de la muestra (32.362) no es representativa del conjunto de datos.

- El coeficiente de asimetría fue del 2.956, siendo ampliamente mayor a 0, por lo cual, la distribución es asimétrica a la derecha, esto debería reflejarse en el histograma, gráfico de dispersión y diagramado de caja siendo la mayor frecuencia de datos encontrada pronunciadamente a la izquierda de la distribución.

- La curtosis correspondió al 12.552, por lo cual es mayor a 0, resultando en que la distribución es altamente leptocúrtica, es decir, el grado de concentración de datos alrededor de la media es muy alta.

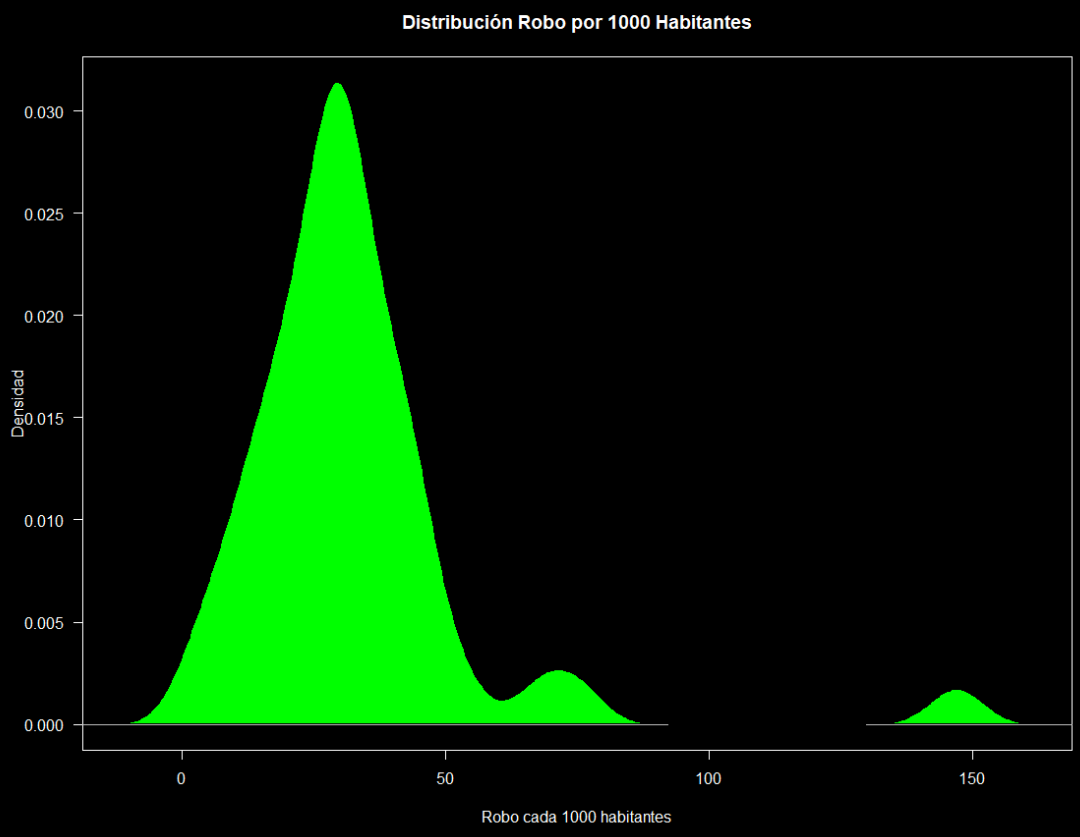
A continuación, se presentan el histograma, gráficos de densidad y caja "boxplot" para la variable.

Histograma de Robo por 1000 habitantes (theft) o
`data$robo....chicago.theft`



- El gráfico se acoge al análisis hecho en los estadísticos descriptivos, sin embargo, se aprecian eventuales outliers que se verán a mayor detalle en el diagramado de caja.

Densidad de Robo por 1000 habitantes (theft) o
`data$robo....chicago.theft`



- La misma situación que el gráfico anterior.

Diagramado de caja Robo por 1000 habitantes (theft) o
`data$robo....chicago.theft`

- En el gráfico efectivamente se observa la distribución altamente leptocúrtica, donde los datos están altamente distribuidos en torno a la media. Además, se aprecian tres outliers, uno en particular que está muy alejado de la distribución. Estos outliers deberán ser limpiados más adelante.

2.2.5. **Porcentaje de viviendas construidas antes de** 1939 (age) o `data$edad....chicago.age`.

La variable Age nos entrega el porcentaje de viviendas construidas antes de 1939:

- Presenta una media del 60.328%. Esto quiere decir que, en promedio, la mayoría de las viviendas fueron construidas en una época anterior a 1939.

- Una mediana del 65% que implica que, del 50% de las viviendas en vecindarios, el 35% fueron construidas en una época posterior o igual a 1939.

- Una moda del 89.8% que nos refleja que, en varios vecindarios se repitió una alta mayoría del 89.8% de viviendas construidas en un período anterior a 1939.

- La desviación estándar calculada fue del 22.575%, estimando que, en promedio, la diferencia entre la media y las viviendas fue del 22.575% de viviendas construidas en una época anterior a 1939.

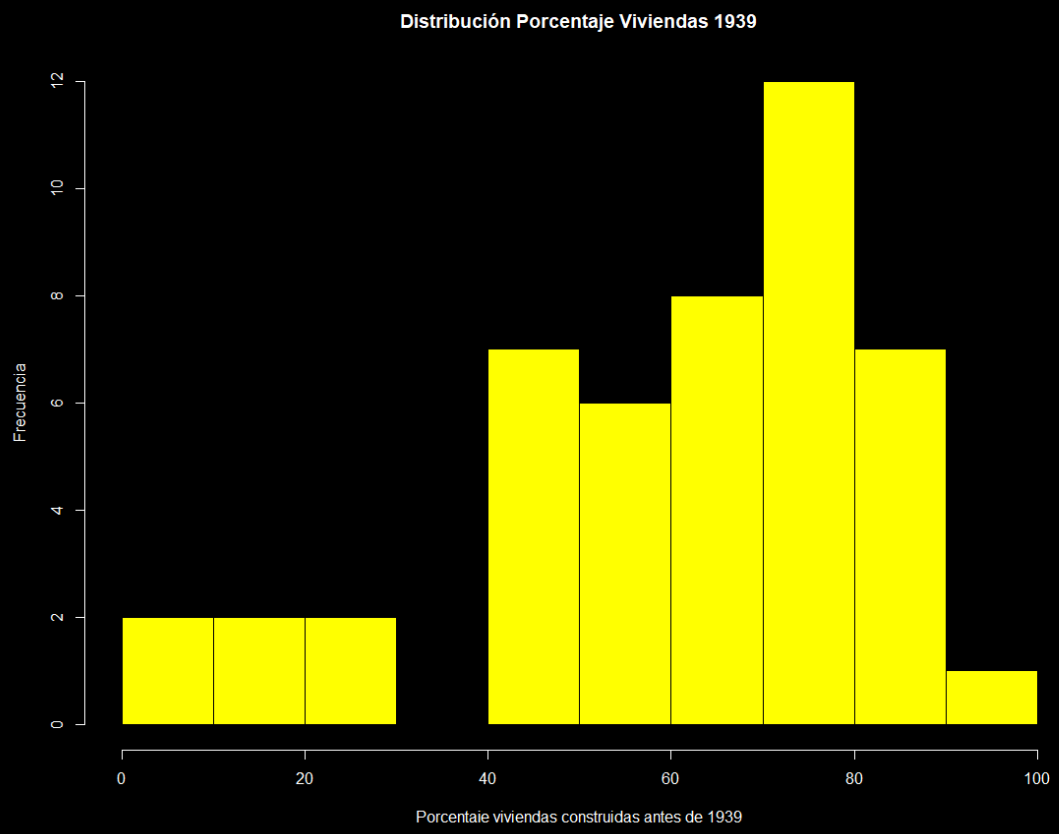
- El coeficiente de variación fue del 0.374 (37.4%), inferior al 50%, siendo entonces una distribución homogénea por lo cual, los datos no están dispersos, encontrándose cercanos a la media, por lo tanto, la media aritmética de la muestra (60.328%) es representativa del conjunto de datos.

- El coeficiente de asimetría fue del -0.921, por lo cual, la distribución es asimétrica a la izquierda, esto debería reflejarse en el histograma, gráfico de dispersión y diagramado de caja siendo la mayor frecuencia de datos encontrada a la derecha de la distribución.

- La curtosis correspondió al 0.084, por lo cual es mayor a 0, pero no por mucho, por lo cual, es un valor muy cercano al 0. A raíz, la distribución es casi mesocúrtica, resultando en que un grado de concentración de datos medio, es decir, asemejada a una distribución normal.

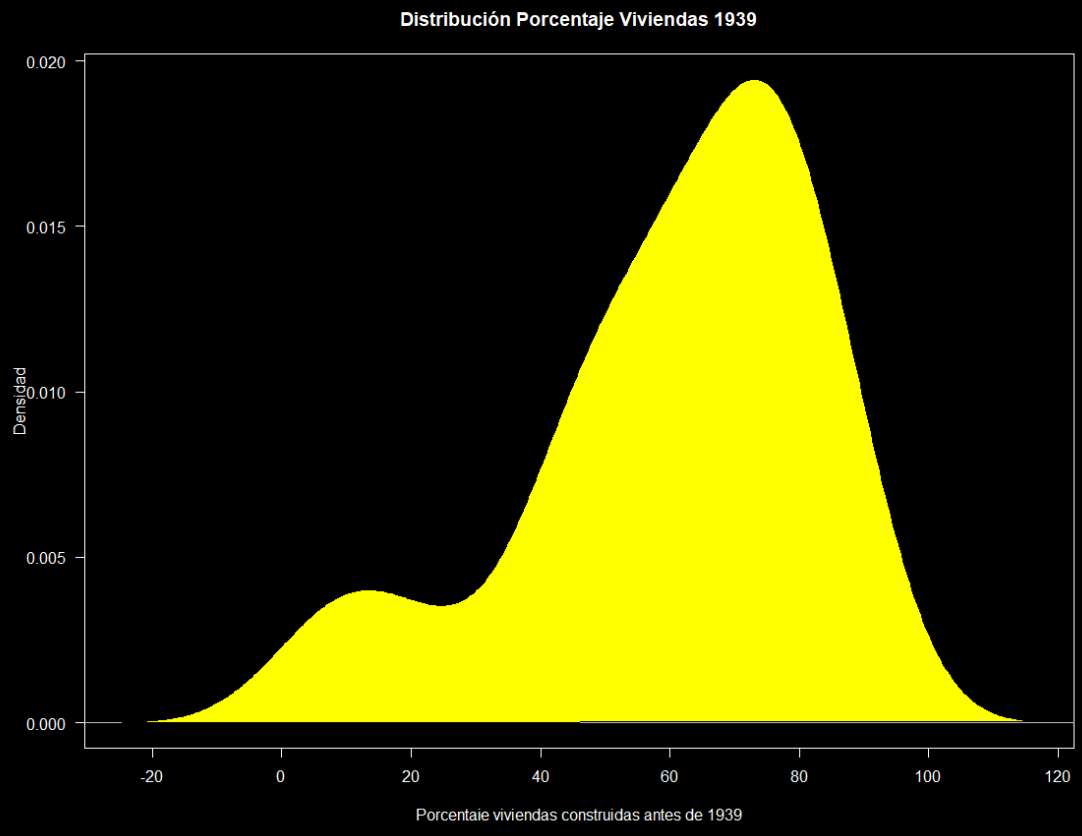
A continuación, se presentan el histograma, gráficos de densidad y caja "boxplot" para la variable.

Histograma de **Porcentaje de viviendas construidas antes de** 1939 (age) o `data$edad....chicago.age`



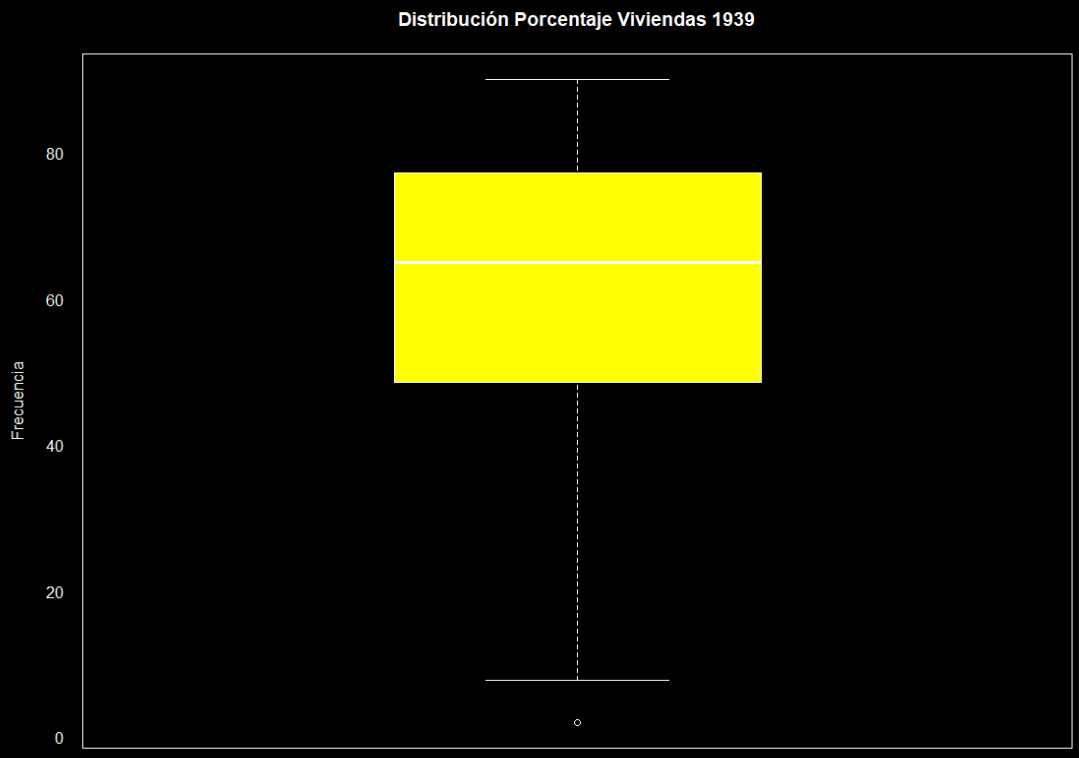
- El gráfico, por lo general, se acoge al análisis hecho en los estadísticos descriptivos, sin embargo, en]30,40[no existe frecuencia, por lo cual se podría asumir que ningún vecindario presentó un porcentaje entre el 30 y el 40% de viviendas construidas antes de 1939.

Densidad de **Porcentaje de viviendas construidas antes de** 1939 (age) o `data$edad....chicago.age`



- El gráfico se acoge al análisis hecho en los estadísticos descriptivos.

Diagramado de caja de **Porcentaje de viviendas construidas antes de** 1939 (age) o `data$edad....chicago.age`



- El gráfico, por lo general, se acoge al análisis hecho en los estadísticos descriptivos, pero, se observa la situación del histograma: no existe frecuencia en el intervalo del 30 al 40%. Además, se aprecia un outlier que habrá que limpiar.

2.2.6. **Vieja política de viviendas (volact)** o `data$politicaVivienda1....chicago.volact`.

La variable Volact nos entrega las nuevas pólizas por proporción en dueños de vivienda más renovaciones, menos cancelaciones y no renovaciones cada 100 viviendas bajo vieja política de vivienda:

- Presenta una media del 6.53%. Esto quiere decir que, por vecindarios cada 100 viviendas se solicitaron o renovaron, en promedio, más del 6% de pólizas.

- Una mediana del 5.9% implica que, por lo menos, 5.9% de pólizas se solicitaron o renovaron del 50% de 100 viviendas en vecindarios.

- Una moda del 3.1% que nos refleja que, en varios vecindarios se repitió una solicitud o renovación de un poco más del 3% de 100 viviendas en vecindarios.

- La desviación estándar calculada fue del 3.966, estimando que, en promedio, la diferencia entre la media y las pólizas renovadas o solicitadas fue más de 4.

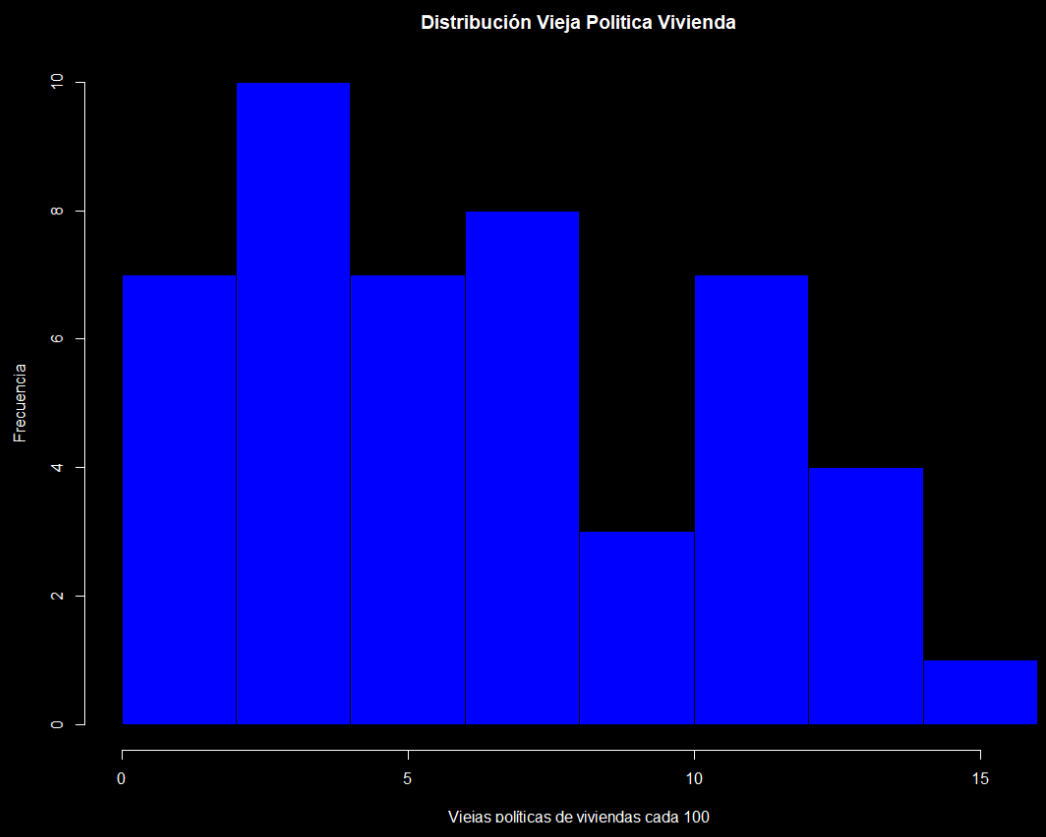
- El coeficiente de variación fue del 0.607 (60.7%), siendo superior al 50% correspondiendo a una distribución heterogénea por lo cual, los datos están dispersos, encontrándose lejanos a la media, por lo tanto, la media aritmética de la muestra (6.53%) no es representativa del conjunto de datos.

- El coeficiente de asimetría fue del 0.271, por lo cual, la distribución es asimétrica a la derecha, esto debería reflejarse en el histograma, gráfico de dispersión y diagramado de caja siendo la mayor frecuencia de datos encontrada a la izquierda de la distribución.

- La curtosis correspondió al -1.179, por lo cual es menor a 0. A raíz, la distribución es casi platicúrtica, resultando en que un grado de concentración de datos reducido, es decir, lejanos al centro.

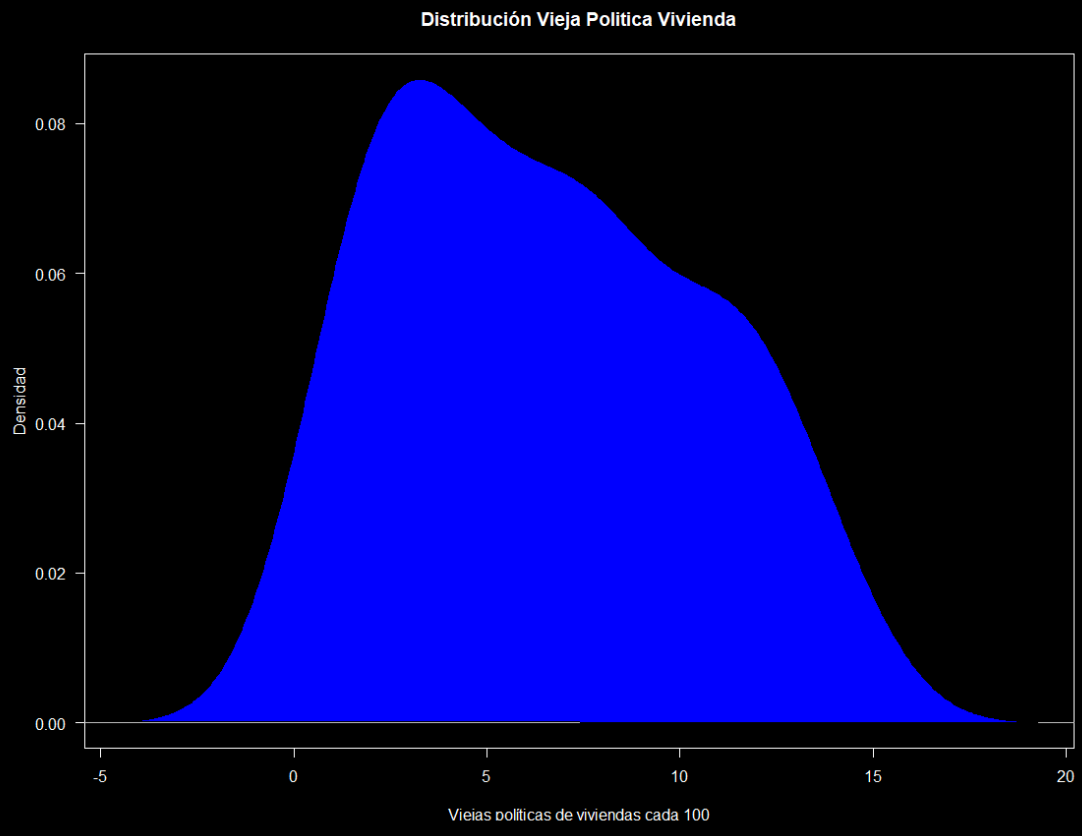
A continuación, se presentan el histograma, gráficos de densidad y caja "boxplot" para la variable.

Histograma de **Vieja política de viviendas (volact)** o `data$politicaVivienda1....chicago.volact`



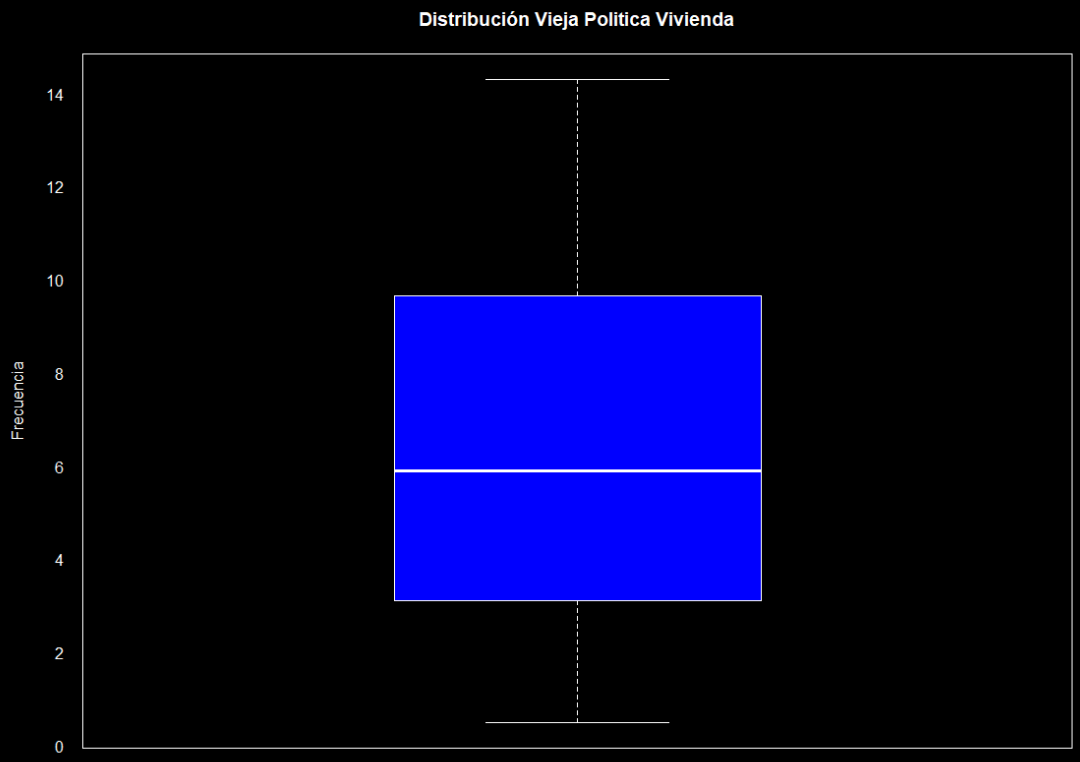
- El gráfico se acoge al análisis hecho en los estadísticos descriptivos. Se aprecia que no más del 20% de cada 100 viviendas en vecindarios, solicitaron o renovaron pólizas bajo vieja política de viviendas.

Densidad de **Vieja política de viviendas (volact)** o
`data$politicaVivienda1....chicago.volact`



- El gráfico se acoge al análisis hecho en los estadísticos descriptivos.

Diagramado de caja de **Vieja política de viviendas (volact)** o `data$politicaVivienda1....chicago.volact`



- El gráfico se acoge al análisis hecho en los estadísticos descriptivos.

2.2.7. **Nueva política de viviendas (involact)** o `data$politicaVivienda2....chicago.involact`.

La variable Involact nos entrega las nuevas pólizas en proporción por dueños de vivienda más renovaciones, menos cancelaciones y no renovaciones cada 100 viviendas bajo nueva política de vivienda FAIR:

- Presenta una media del 0.615%. Esto quiere decir que, por vecindarios cada 100 viviendas se solicitaron o renovaron, en promedio, más del 6% de pólizas.

- Una mediana del 5.9% implica que, por lo menos, 5.9% de pólizas se solicitaron o renovaron del 50% de 100 viviendas en vecindarios.

- Una moda del 3.1% que nos refleja que, en varios vecindarios se repitió una solicitud o renovación de un poco más de tres pólizas del 3% en viviendas en vecindarios.

- La desviación estándar calculada fue del 3.966%, estimando que, en promedio, la diferencia entre la media y las pólizas renovadas o solicitadas fue aproximadamente del 4%.

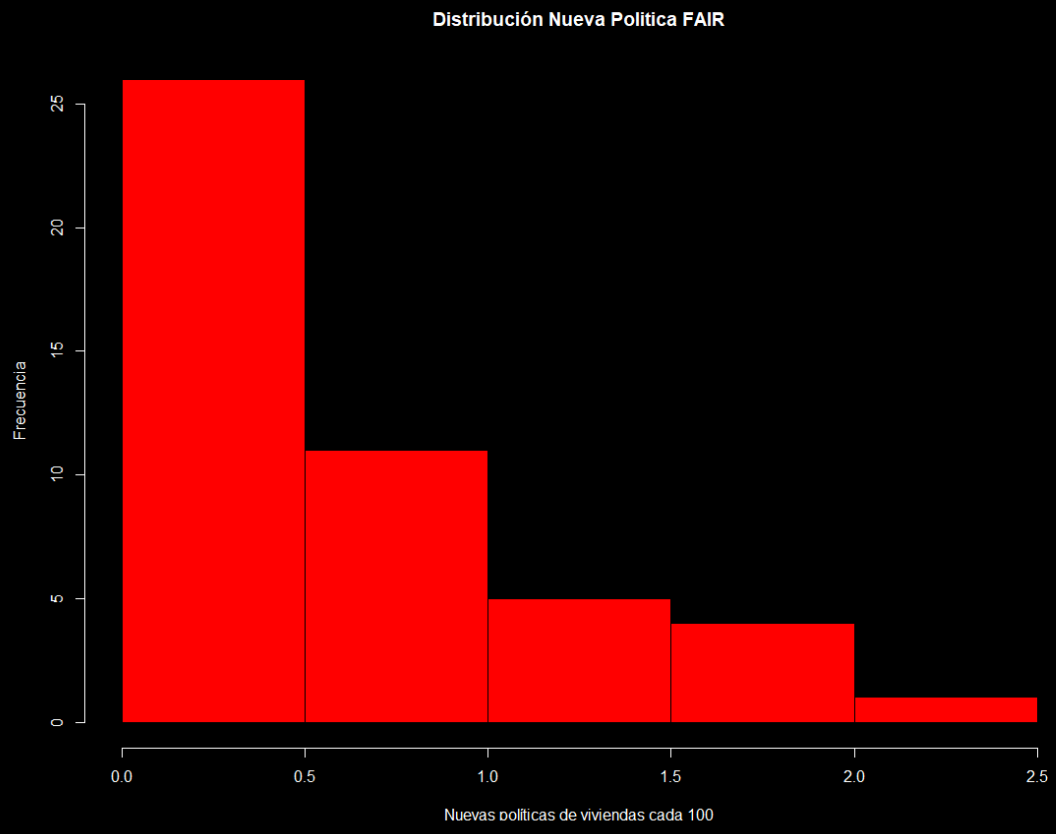
- El coeficiente de variación fue del 0.607 (60.7%), siendo superior al 50% correspondiendo a una distribución heterogénea por lo cual, los datos están dispersos, encontrándose lejanos a la media, por lo tanto, la media aritmética de la muestra (6.53%) no es representativa del conjunto de datos.

- El coeficiente de asimetría fue del 0.271, por lo cual, la distribución es asimétrica a la derecha, esto debería reflejarse en el histograma, gráfico de dispersión y diagramado de caja siendo la mayor frecuencia de datos encontrada a la izquierda de la distribución.

- La curtosis correspondió al -1.179, por lo cual es menor a 0. A raíz, la distribución es casi platicúrtica, resultando en que un grado de concentración de datos reducido, es decir, lejanos al centro.

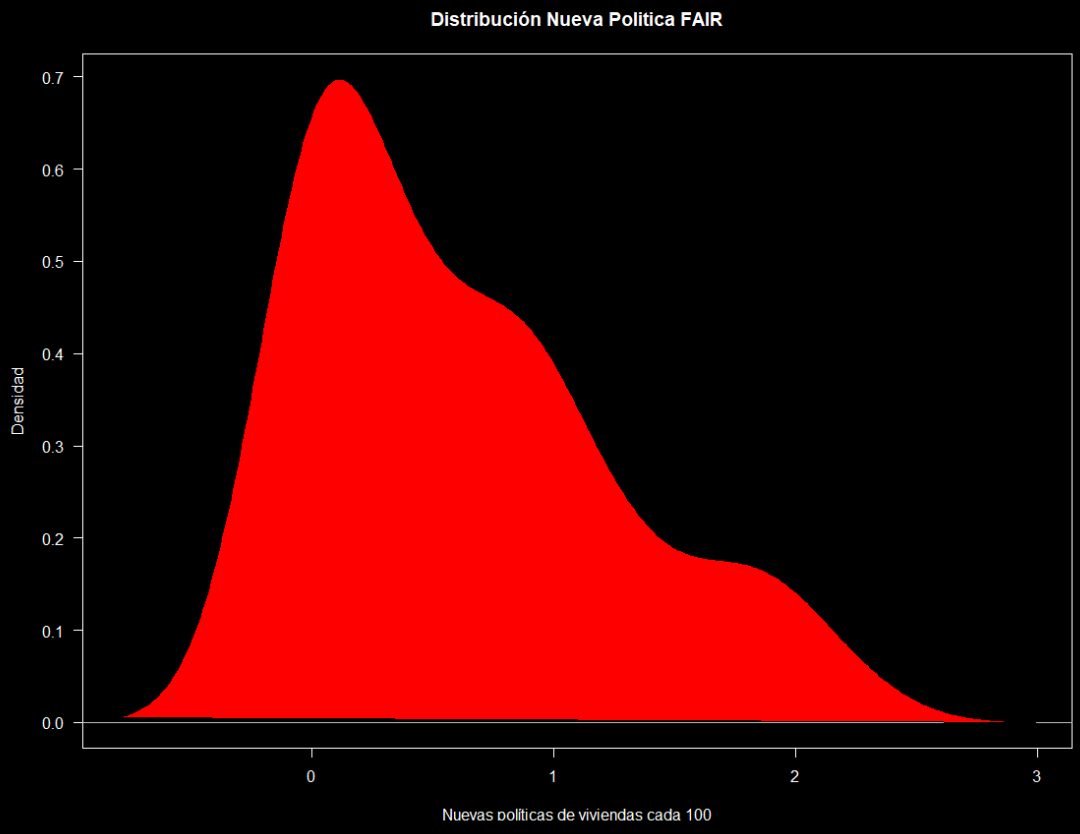
A continuación, se presentan el histograma, gráficos de densidad y caja "boxplot" para la variable.

Histograma de Nueva política de viviendas (volact) o `data$politicaVivienda2....chicago.involact`.



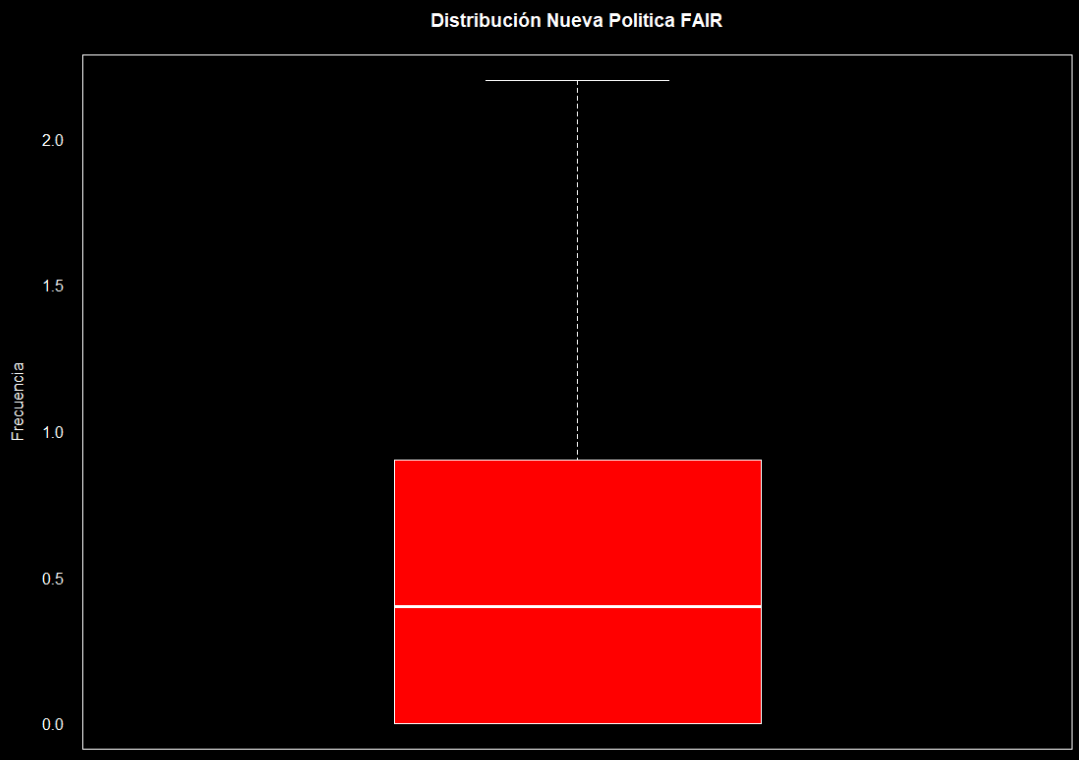
- El gráfico se acoge al análisis hecho en los estadísticos descriptivos. Se aprecia que no más del 2.5% de cada 100 viviendas en vecindarios, solicitaron o renovaron pólizas bajo nueva política de viviendas.

Densidad de Nueva política de viviendas (volact) o `data$politicaVivienda2....chicago.involact`.



- El gráfico se acoge al análisis hecho en los estadísticos descriptivos.

Diagramado de caja de **Nueva política de viviendas (volact)** o `data$politicaVivienda2....chicago.involact`.



- La misma situación que el gráfico anterior.

2.2.8. **Ingreso medio** familiar **o** `data$ingresofamiliar....chicago.income`.

La variable Income nos entrega el ingreso medio familiar:

- Presenta una media del 10695.830. Esto quiere decir que, por familia, en promedio, hay un ingreso medio del 10695.830.

- Una mediana del 10694 implica que, por lo menos, el 50% de las familias percibe un ingreso menor o igual a 10694.

- Una moda del 5583 que nos refleja que, el valor más repetido de ingreso medio familiar fue de 5583.

- La desviación estándar calculada fue del 2754.198 estimando que, en promedio, la diferencia entre la media y el ingreso medio familiar fue aproximadamente de 2754.198.

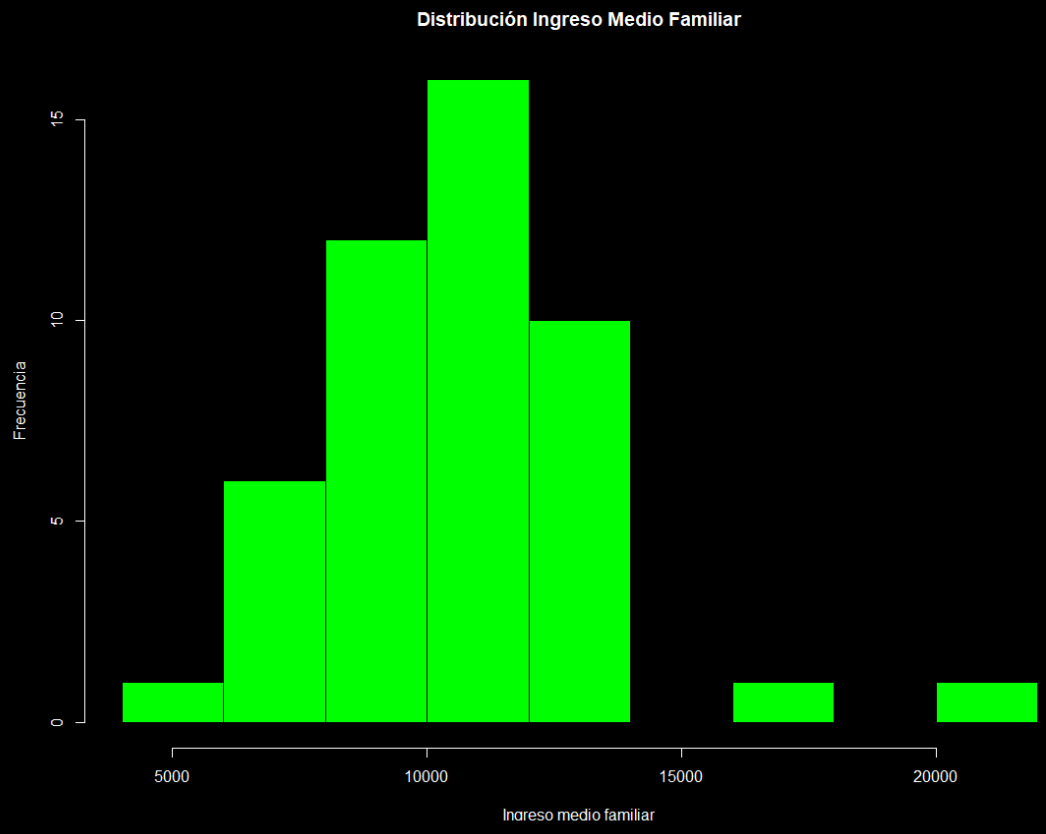
- El coeficiente de variación fue del 0.258 (25.8%), siendo inferior al 50% correspondiendo a una distribución homogénea por lo cual, los datos no están dispersos, encontrándose cercanos a la media, por lo tanto, la media aritmética de la muestra (10695.830) es representativa del conjunto de datos.

- El coeficiente de asimetría fue del 1.155, por lo cual, la distribución es asimétrica a la derecha, esto debería reflejarse en el histograma, gráfico de dispersión y diagramado de caja siendo la mayor frecuencia de datos encontrada a la izquierda de la distribución.

- La curtosis correspondió al 3.156, por lo cual es mayor a 0. A raíz, la distribución es casi leptocúrtica, resultando en que un grado de concentración de datos elevado, es decir, cercanos al centro.

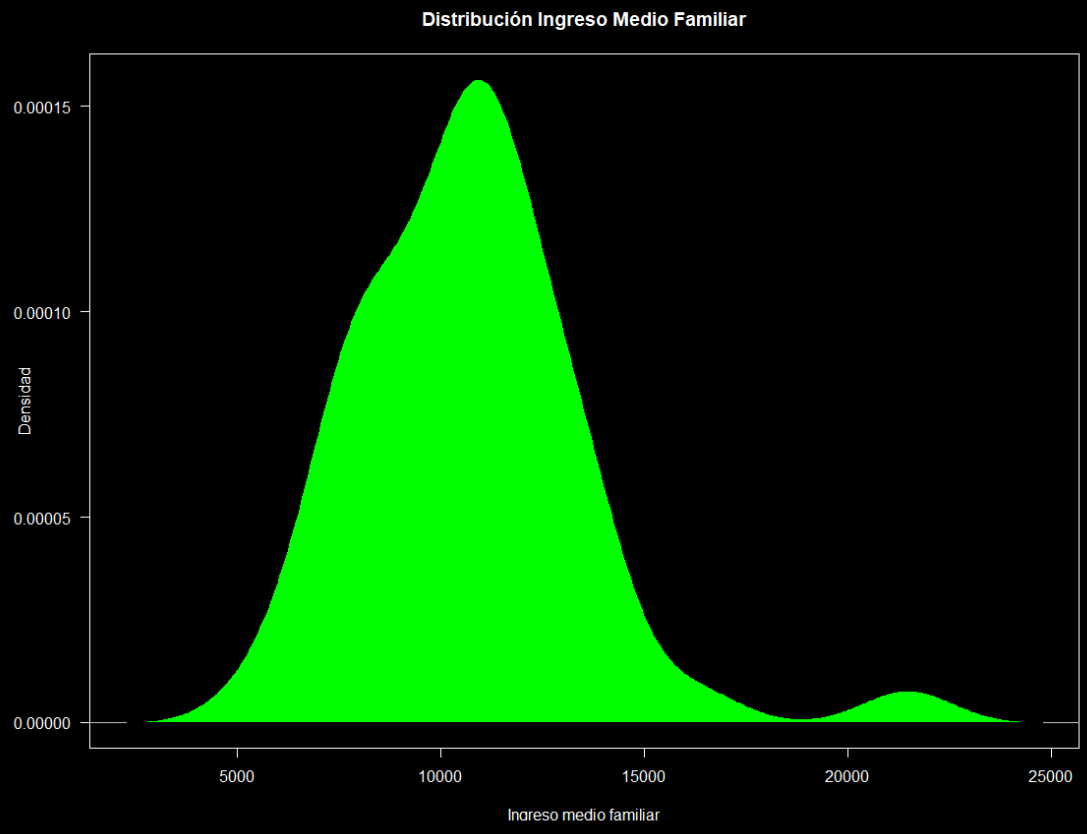
A continuación, se presentan el histograma, gráficos de densidad y caja "boxplot" para la variable.

Histograma de **Ingreso medio familiar (income)** o `data$ingresofamiliar....chicago.income`.



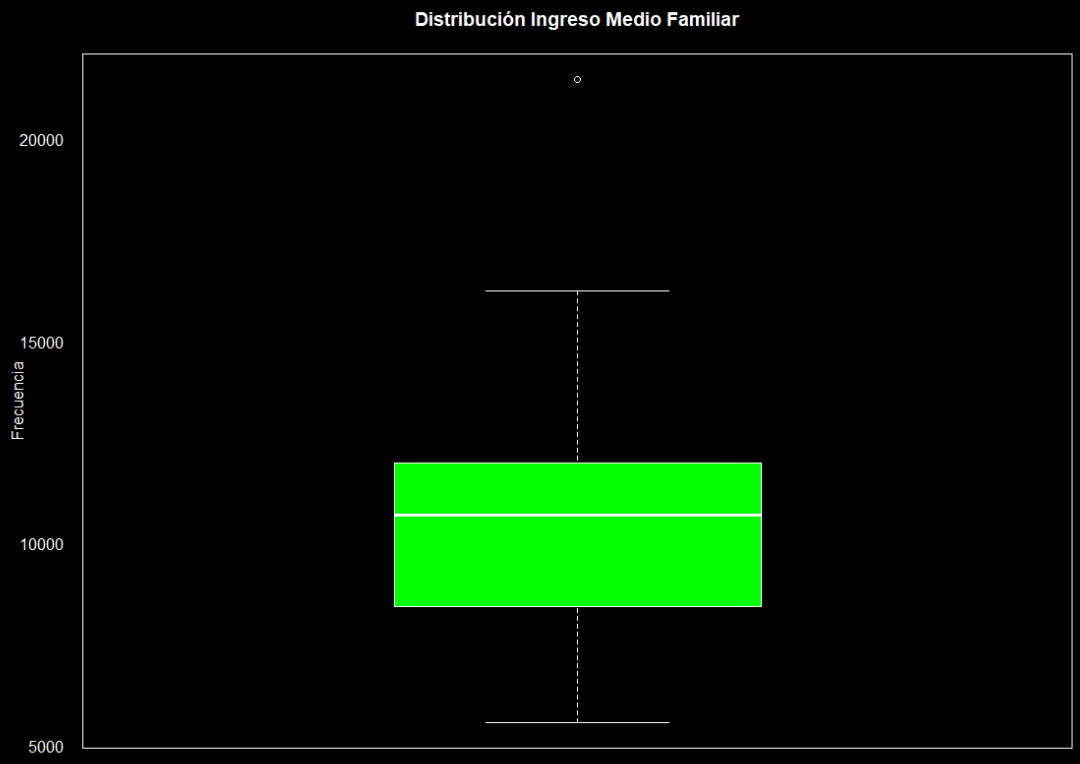
- El gráfico se acoge al análisis hecho en los estadísticos descriptivos, sin embargo, se aprecian eventuales outliers que se verán a mayor detalle en el diagramado de caja.

Densidad de **Ingreso medio familiar (income)** o `data$ingresofamiliar....chicago.income`.



- La misma situación que el gráfico anterior.

Diagramado de caja **Ingreso medio familiar (income)** o `data$ingresofamiliar....chicago.income`.



- La misma situación que el gráfico anterior, además se observa un outlier.

3. Limpieza de data y relación entre variables.

Las correlaciones nos permiten analizar la proporcionalidad o la relación lineal entre dos o más variables dependiendo del modelo que se desea investigar. Sin embargo, debido a que estas correlaciones son realizadas mediante la data disponible, estas mismas son susceptibles a extremos anómalos que a veces, son introducidos de manera errónea en la base de datos por tabulación o medición incorrecta. Es por esto por lo que, a continuación, se realizará dicho estudio del impacto de los outliers en las correlaciones, antes y después de su limpieza:

3.1. Primera correlación simple antes de limpiar la data.

Con el comando `cor()` se realiza la correlación en un dataset por fila y columna en cada variable, resultando:

```
> iniciocor
                                raza....chicago.race fuego....chicago.fire robo....chicago.theft edad....chicago.age politicavivienda1....chicago.volact
raza....chicago.race                       1.0000000             0.5927956            0.2550647          0.2505118                      -0.7594196
fuego....chicago.fire                      0.5927956             1.0000000            0.5562105          0.4122225                      -0.6864766
robo....chicago.theft                      0.2550647             0.5562105            1.0000000          0.3176308                      -0.3116183
edad....chicago.age                        0.2505118             0.4122225            0.3176308          1.0000000                      -0.6057428
politicavivienda1....chicago.volact       -0.7594196            -0.6864766           -0.3116183         -0.6057428                       1.0000000
politicavivienda2....chicago.involact      0.7137540             0.7030397            0.1496309          0.4757291                      -0.7464229
ingresofamiliar....chicago.income         -0.7037328            -0.6104481           -0.1729226         -0.5286695                       0.7509780
                                politicavivienda2....chicago.involact ingresofamiliar....chicago.income
raza....chicago.race                                       0.7137540                        -0.7037328
fuego....chicago.fire                                      0.7030397                        -0.6104481
robo....chicago.theft                                      0.1496309                        -0.1729226
edad....chicago.age                                        0.4757291                        -0.5286695
politicavivienda1....chicago.volact                       -0.7464229                         0.7509780
politicavivienda2....chicago.involact                      1.0000000                        -0.6648471
ingresofamiliar....chicago.income                         -0.6648471                         1.0000000
```

El análisis respectivo de cada correlación:

RACE

```
# La relación de RACE y FIRE es fuerte y positiva.
# La relación de RACE y THEFT es débil y positiva.
# La relación de RACE y AGE es débil y positiva.
# La relación de RACE y VOLACT es fuerte y negativa.
# La relación de RACE e INVOLACT es fuerte y positiva.
# La relación de RACE e INCOME es fuerte y negativa.
```

FIRE

```
# La relación de FIRE y THEFT es fuerte y positiva.
# La relación de FIRE y AGE es moderada y positiva.
# La relación de FIRE y VOLACT es fuerte y negativa.
# La relación de FIRE e INVOLACT es fuerte y positiva.
# La relación de FIRE e INCOME es fuerte y negativa.
```

THEFT

```
# La relación de THEFT y AGE es moderada y positiva.
# La realción de THEFT y VOLACT es fuerte y negativa.
# La relación de THEFT e INVOLACT es moderada y positiva.
# La relación de THEFT e INCOME es fuerte y negativa.
```

AGE

```
# La relación de AGE y VOLACT es fuerte y negativa.
# La relación de AGE e INVOLACT es moderada y positiva.
# La relación de AGE e INCOME es fuerte y negativa.
```

VOLACT

```
# La relación de VOLACT e INVOLACT es fuerte y negativa.
# La relación de VOLACT e INCOME es fuerte y positiva.
```

INVOLACT

```
# La relación de INVOLACT e INCOME es fuerte y negativa.
```

3.2. Limpieza de data.

En una primera instancia se pensó en eliminar la fila completa que contuviera el outlier (dado que por código era más simple), sin embargo, dicho método no considera la enorme pérdida de datos al eliminar una fila, puesto no solo se elimina el outlier, sino también los datos de la fila completa.

A razón de lo anterior, y gracias a la librería Tidyverse más la utilidad de obtener aquellos outliers a través del diagramado de caja, se utilizó el método de reemplazar dichos outliers por la media de cada variable, esto a través de la obtención manual de la fila y columna que contenía dichos outliers con el comando `boxplot(data[, n], plot=FALSE)$out` donde n corresponde a la columna deseada, y luego realizando la búsqueda a través de `data[which(data %in%), boxplot(data[, n], plot=FALSE)$out]` arrojando las filas que contuvieran los outliers presentes en los diagramas de caja adjuntados en la parte superior. Luego, una vez obtenida la fila y columna se procedió a reemplazar de forma manual dichos outliers por la media. Cabe recalcar que dicho proceso fue realizado sin un ciclo iterativo puesto no era necesario, sin embargo, si se tuviera que eliminar una gran cantidad de outliers de una base de datos más grande, se requeriría utilizar un código más automatizado.

En total, se realizaron dos ciclos de limpieza hasta que en el diagramado de caja no apareció ningún outlier:

```
# Termino de limpieza de outliers

#(total limpiados: 5 filas de datos)

#--> Tras repensar el método de limpieza y la pérdida de datos si se
eliminan filas, se optó por

# Corregir dichos outlier colocando la media en las celdas
correspondientes. Por lo tanto

# Se corregieron 6 outlier de la variable FIRE

# Se corregieron 6 outlier de la variable THEFT

# Se corregieron 1 outlier de la variable AGE

# Se corregieron 1 outlier de la variable INCOME
```

Por lo tanto, a partir del comentario extraído del código, se corrigieron, en total, 18 outliers de las variables FIRE, THEFT, AGE e INCOME, por lo cual, dichas variables sufrirán cambios en sus correlaciones simples.

3.3. Segunda correlación simple tras limpiar la data.

Nuevamente con la utilización del comando `cor()` se realiza correlación, pero esta vez, con la data sin outliers:

```
> cor
                                      raza....chicago.race fuego....chicago.fire robo....chicago.theft edad....chicago.age politicavivienda1....chicago.volact
raza....chicago.race                             1.0000000             0.7291560             0.5912335          0.2339431                        -0.7594196
fuego....chicago.fire                            0.7291560             1.0000000             0.5116007          0.4764259                        -0.8082098
robo....chicago.theft                            0.5912335             0.5116007             1.0000000          0.4276477                        -0.6728421
edad....chicago.age                              0.2339431             0.4764259             0.4276477          1.0000000                        -0.5931793
politicavivienda1....chicago.volact             -0.7594196            -0.8082098            -0.6728421         -0.5931793                         1.0000000
politicavivienda2....chicago.involact            0.7137540             0.7816447             0.3945060          0.4748992                        -0.7464229
ingresofamiliar....chicago.income               -0.7697291            -0.7605742            -0.6138204         -0.5567949                         0.8884537
                                      politicavivienda2....chicago.involact ingresofamiliar....chicago.income
raza....chicago.race                                             0.7137540                        -0.7697291
fuego....chicago.fire                                            0.7816447                        -0.7605742
robo....chicago.theft                                            0.3945060                        -0.6138204
edad....chicago.age                                              0.4748992                        -0.5567949
politicavivienda1....chicago.volact                             -0.7464229                         0.8884537
politicavivienda2....chicago.involact                            1.0000000                        -0.7169316
ingresofamiliar....chicago.income                               -0.7169316                         1.0000000
```

El análisis respectivo del cambio de cada correlación respecto a la anterior, ahora con la data limpia y con menos ruido:

RACE

```
# La relación de RACE y FIRE es más fuerte y positiva.
# La relación de RACE y THEFT no es débil, sino fuerte y positiva.
# La relación de RACE y AGE es un poco mayor pero débil y positiva.
# La relación de RACE y VOLACT no sufrió cambios.
# La relación de RACE e INVOLACT no sufrió cambios.
# La relación de RACE e INCOME es más fuerte y negativa.
```

FIRE

```
# La relación de FIRE y THEFT es menos fuerte y positiva.

# La relación de FIRE y AGE es mayor pero moderada y positiva.

# La relación de FIRE y VOLACT es más fuerte y negativa.

# La relación de FIRE e INVOLACT es más fuerte y positiva.

# La relación de FIRE e INCOME es más fuerte y negativa.
```

THEFT

```
# La relación de THEFT y AGE es mayor pero moderada y positiva.

# La relación de THEFT y VOLACT es más fuerte y negativa.

# La relación de THEFT e INVOLACT no es moderada. Es débil y
positiva.

# La relación de THEFT e INCOME es no es fuerte. Es débil y
negativa.
```

AGE

```
# La relación de AGE y VOLACT es un poco más fuerte y negativa.

# La relación de AGE e INVOLACT es más moderada y positiva. Casi
fuerte.

# La relación de AGE e INCOME es más fuerte y negativa.
```

VOLACT

```
# La relación de VOLACT e INVOLACT no sufrió cambios.

# La relación de VOLACT e INCOME es más fuerte y positiva.
```

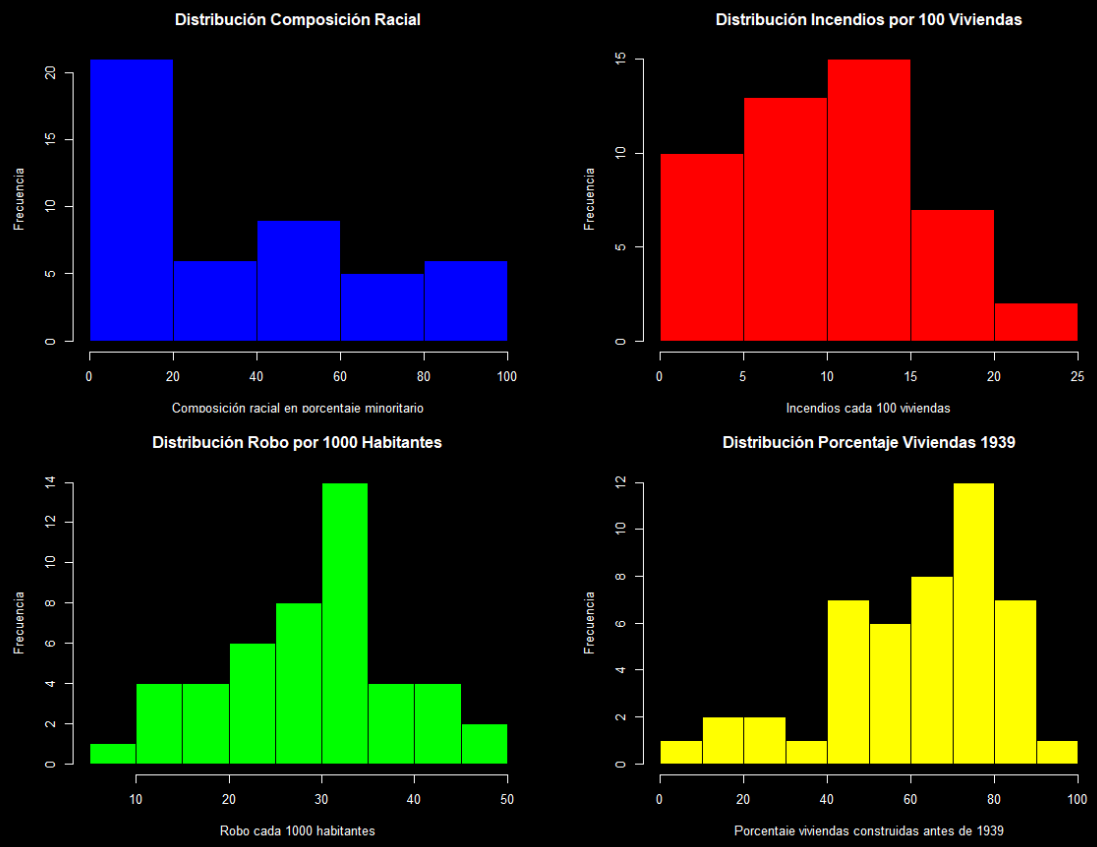
INVOLACT

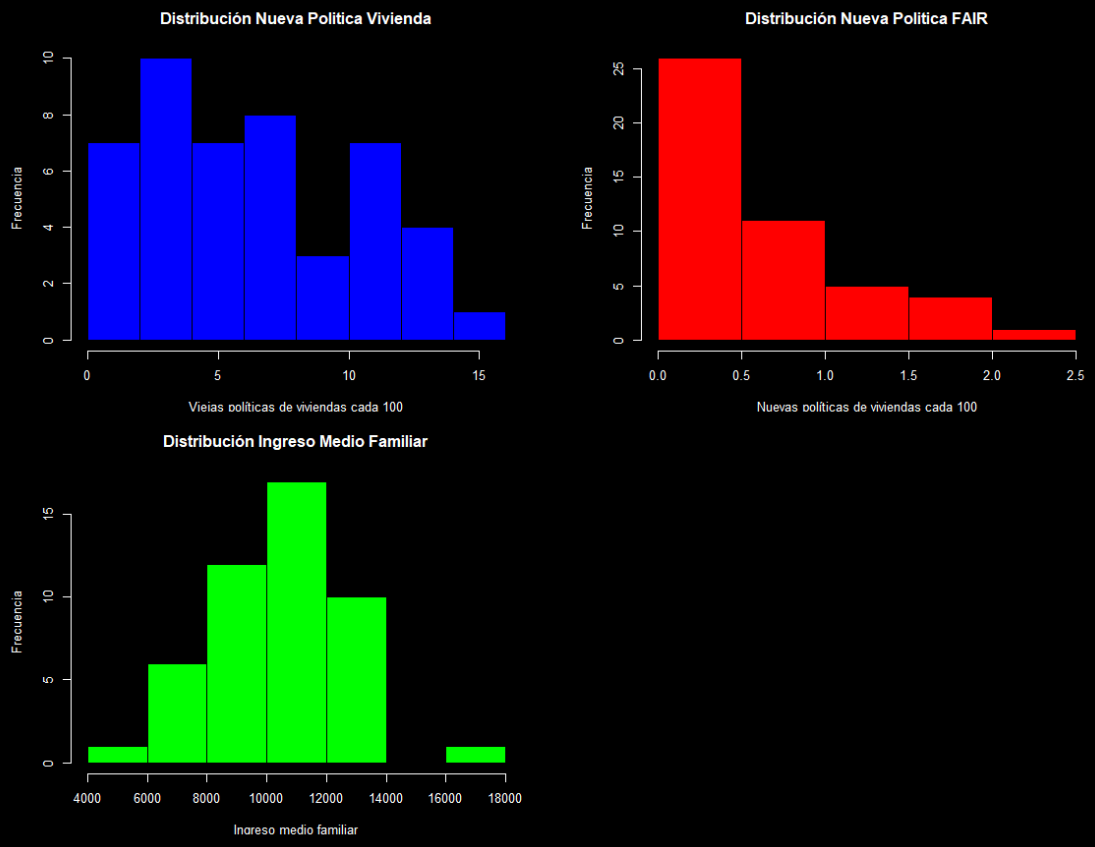
```
# La relación de INVOLACT e INCOME es más fuerte y negativa.
```

3.4. Cambio en gráficos.

A modo de resumen, se adjuntan los gráficos superiores por cada variable en una distribución de cuatro gráficos por hoja, pero esta vez con la data limpia para observar cómo variaron las distribuciones. Cabe recalcar que estos cambios nos servirán para realizar un mejor ajuste del modelo más adelante.

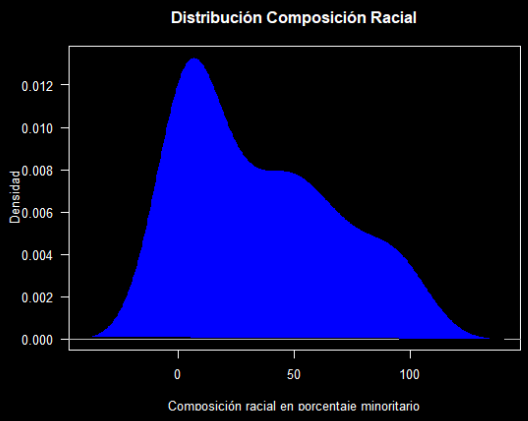
Histogramas

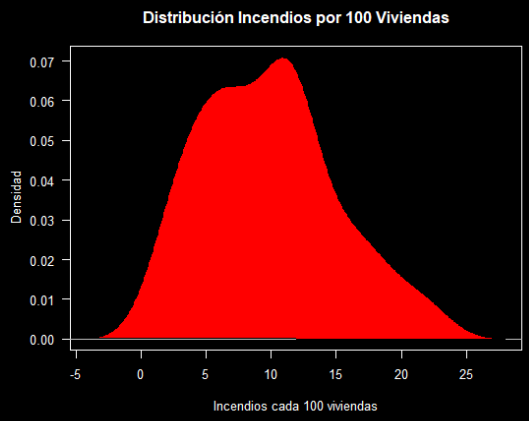


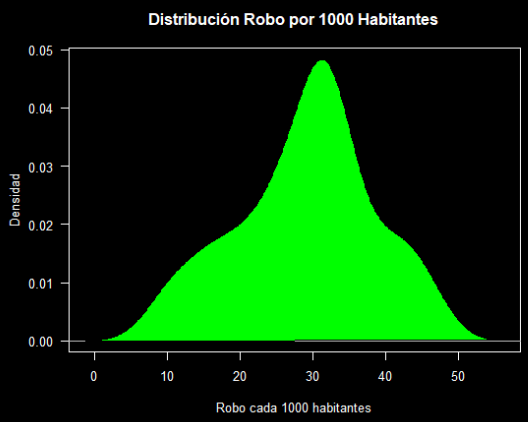


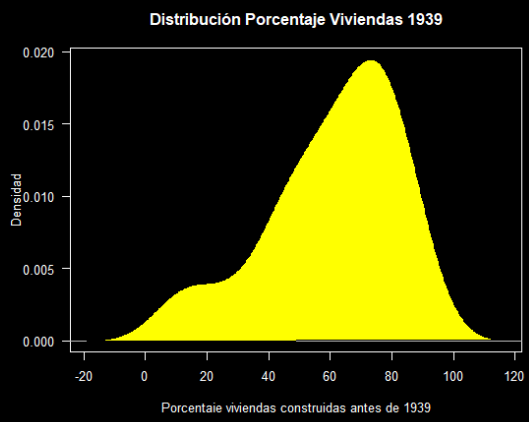
- Los histogramas sufrieron grandes cambios. A modo general tuvieron un impacto positivo en las distribuciones, dado que en todas el nivel de dispersión disminuyó, permitiendo observar gráficas menos irregulares.

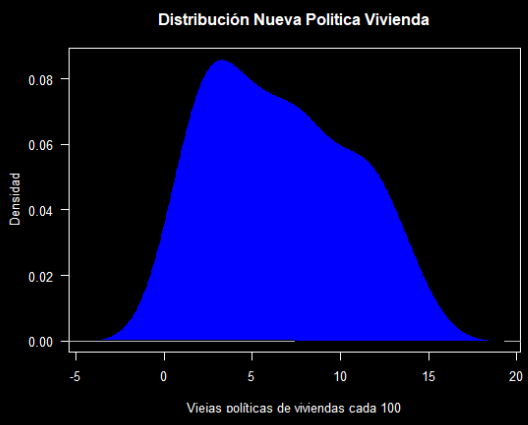
Densidad

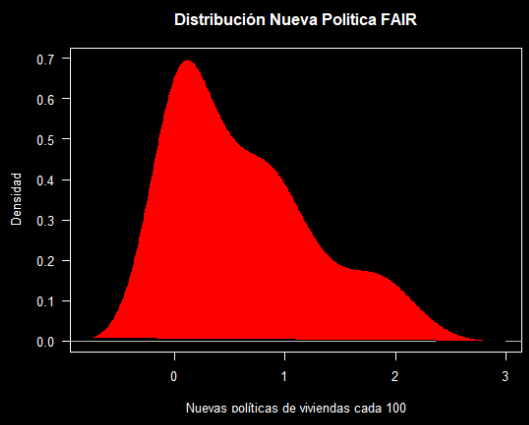

Distribución Composición Racial
Densidad
0.012
0.010
0.008
0.006
0.004
0.002
0.000
0
50
100
Composición racial en porcentaje minoritario

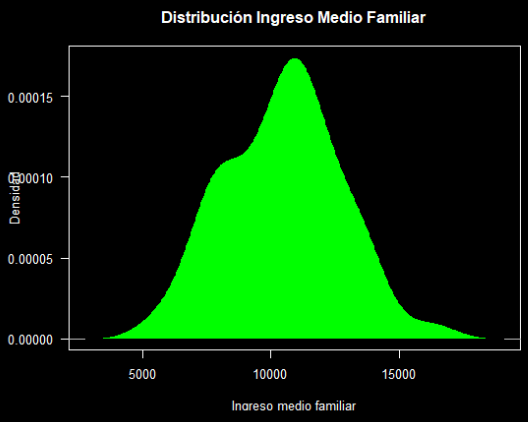

Distribución Incendios por 100 Viviendas
Densidad
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
-5
0
5
10
15
20
25
Incendios cada 100 viviendas


Distribución Robo por 1000 Habitantes
Densidad
0.05
0.04
0.03
0.02
0.01
0.00
0
10
20
30
40
50
Robo cada 1000 habitantes


Distribución Porcentaje Viviendas 1939
Densidad
0.020
0.015
0.010
0.005
0.000
-20
0
20
40
60
80
100
120
Porcentaje viviendas construidas antes de 1939

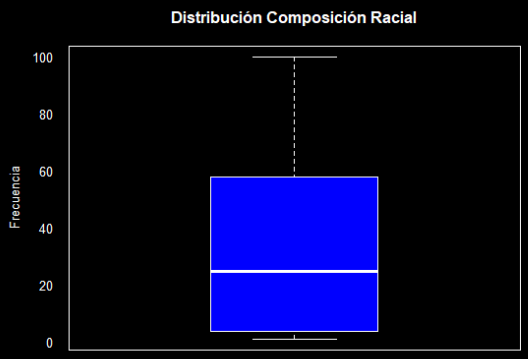


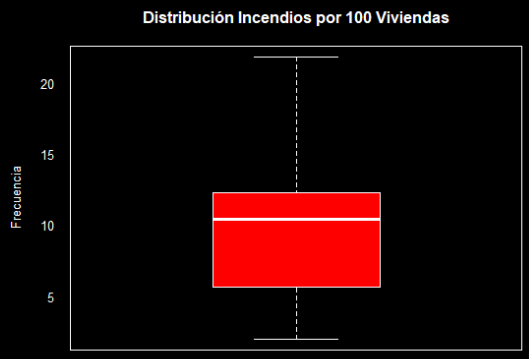


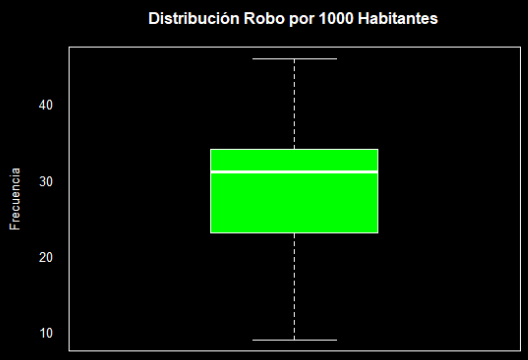


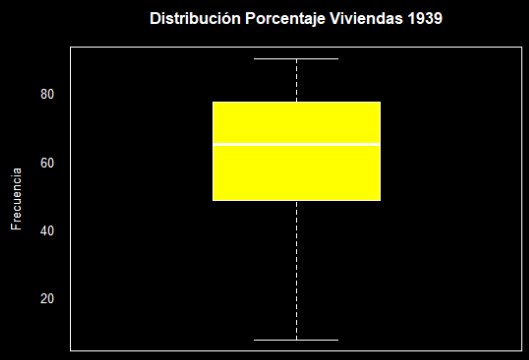
- Los diagramas de densidad sufrieron cambios similares a los histogramas.

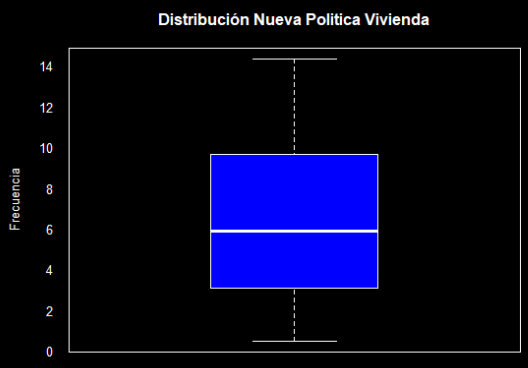
Diagramado de Caja (boxplot)

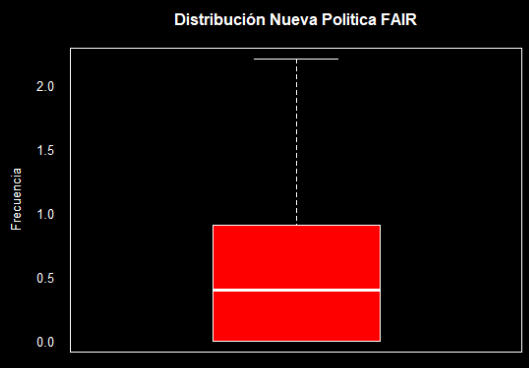


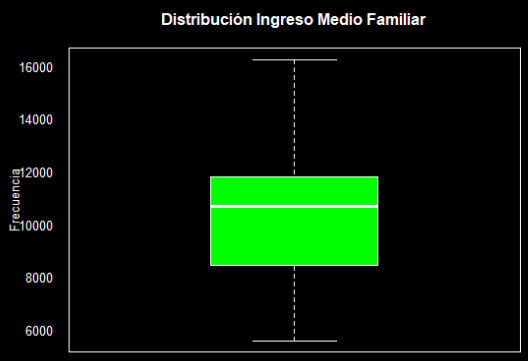












- En los diagramados de caja se observan los mayores cambios de los tres gráficos. Comparando los gráficos antes y después de la limpieza, se logra ver que la limpieza de data fue efectiva, puesto todos los outliers fueron eliminados, logrando que la distribución se aprecie de mejor manera, con sus características y descripciones previstas a través de los estadísticos descriptivos.

3.5. Correlaciones parciales.

Si bien las correlaciones simples nos pueden dar una idea bastante intuitiva del comportamiento de las variables entre sí, dichas correlaciones pueden ser engañosas al no restar la intersección de varianzas de, en este caso, las siete series de datos, puesto su intersección de varianzas añade una variable adicional C. Esta variable adicional C añade ruido y modifica dichas correlaciones.

Es por lo anterior que si se elimina el ruido de esta variable C, que afecta las correlaciones simples, se obtiene la correlación parcial, que es una correlación más segura que la simple.

En la práctica, son las mismas correlaciones simples, pero sin el ruido interfiriendo entre variables.

"*Eliminando la varianza compartida por las variables de interés con la o las variables auxiliares, obtenemos una medida de r que refleja los efectos de las variables de interés primario*."

Centro de Ciencias Genómicas de la UNAM Campus Morelos. (s. f.). Recuperado 22 de agosto de 2020, de https://www.ccg.unam.mx/~vinuesa/R4biosciences/docs/Tema8_correlacion.html

Por lo cual, ejecutando el comando `pcor()` en R STUDIO, se nos brinda las correlaciones parciales:

```
> corr <- pcor(data)$estimate
> corr
                                    raza....chicago.race fuego....chicago.fire robo....chicago.theft edad....chicago.age politicaviviendа1....chicago.volact
raza....chicago.race                          1.0000000           0.15699350          0.32123477        -0.53475154                        -0.1133044
fuego....chicago.fire                         0.1569935           1.00000000         -0.02755743         0.04581949                        -0.2778653
robo....chicago.theft                         0.3212348          -0.02755743          1.00000000         0.21795983                        -0.3040498
edad....chicago.age                          -0.5347515           0.04581949          0.21795983         1.00000000                        -0.2233900
politicaviviendа1....chicago.volact          -0.1133044          -0.27786532         -0.30404977        -0.22339005                         1.0000000
politicaviviendа2....chicago.involact         0.3494078           0.34821047         -0.30816066         0.24428415                        -0.1411004
ingresofamiliar....chicago.income            -0.3360462          -0.02973155          0.03423613        -0.23525716                         0.4979253
                                    politicaviviendа2....chicago.involact ingresofamiliar....chicago.income
raza....chicago.race                                        0.349407779                       -0.336046226
fuego....chicago.fire                                       0.348210467                       -0.029731554
robo....chicago.theft                                      -0.308160661                        0.034236129
edad....chicago.age                                         0.244284153                       -0.235257165
politicaviviendа1....chicago.volact                        -0.141100394                        0.497925266
politicaviviendа2....chicago.involact                       1.000000000                        0.003857597
ingresofamiliar....chicago.income                           0.003857597                        1.000000000
```

Y esto nos trae a su vez, una redefinición de las correlaciones entre variables. Se pueden apreciar los cambios comparando con los análisis hechos en Segunda correlación simple tras limpiar la data.:

RACE

```
# La relación de RACE y FIRE es débil y positiva.
# La relación de RACE y THEFT es moderada y positiva.
# La relación de RACE y AGE es fuerte y negativa.
# La relación de RACE y VOLACT es débil y negativa.
# La relación de RACE e INVOLACT es moderada y positiva.
# La relación de RACE e INCOME es moderada y negativa.
```

FIRE

```
# La relación de FIRE y THEFT es casi nula y negativa.
# La relación de FIRE y AGE es casi nula y positiva.
# La relación de FIRE y VOLACT es débil y negativa.
# La relación de FIRE e INVOLACT es moderada y positiva.
# La relación de FIRE e INCOME es casi nula y negativa.
```

THEFT

```
# La relación de THEFT y AGE es débil y positiva.
# La relación de THEFT y VOLACT es casi débil y negativa.
# La relación de THEFT e INVOLACT es casi débil y negativa.
# La relación de THEFT e INCOME es casi nula y positiva.
```

AGE

```
# La relación de AGE y VOLACT es débil y negativa.
# La relación de AGE e INVOLACT es débil y positiva.
# La relación de AGE e INCOME es débil y negativa.
```

VOLACT

```
# La relación de VOLACT e INVOLACT es débil y negativa.
# La relación de VOLACT e INCOME es casi fuerte y positiva.
```

INVOLACT

```
# La relación de INVOLACT e INCOME es casi nula y positiva.
```

- Se pueden apreciar cambios considerables respecto a las correlaciones simples, pero una vez más, el ruido de la intersección de varianzas interfiere en dichos valores produciendo que, al restar dicha intersección de varianzas, las correlaciones parciales sean más precisas.

3.6. P valor.

El P valor nos entrega aquellas relaciones **estadísticamente significativas** a partir de una hipótesis. En concreto, R STUDIO comprueba cada una de las variables con la hipótesis nula de que el coeficiente es igual a cero, o que, entre sí, los cambios a los que responden las relaciones entre variables son significativas para un modelo de regresión. En otras palabras, nos entrega la significancia, en este caso, de las relaciones entre variables.

El párrafo anterior se basó en el vínculo a continuación.

Antonio, J. (2016, 19 diciembre). Cómo interpretar los resultados del análisis de regresión: p-valores y coeficientes. Recuperado 22 de agosto de 2020, de este vínculo.

Por lo cual, a partir del comando `pcor(data)$p.value` se obtienen dichos p valores:

```
> pvalor
                                      raza....chicago.race fuego....chicago.fire robo....chicago.theft edad....chicago.age politicaviviendal....chicago.volact
raza....chicago.race                          0.0000000000           0.32075677          0.03804718       0.0002639966                        0.474953805
fuego....chicago.fire                         0.3207567693           0.00000000          0.86246618       0.7732440335                        0.074798256
robo....chicago.theft                         0.0380471847           0.86246618          0.00000000       0.1655487642                        0.050273979
edad....chicago.age                           0.0002639966           0.77324403          0.16554876       0.0000000000                        0.155001345
politicaviviendal....chicago.volact           0.4749538046           0.07479826          0.05027398       0.1550013450                        0.000000000
politicaviviendal2....chicago.involact        0.0233232015           0.02383369          0.04709424       0.1189747777                        0.372763144
ingresofamiliar....chicago.income             0.0295688379           0.85173214          0.82957819       0.1336743129                        0.000792452
                                      politicaviviendal2....chicago.involact ingresofamiliar....chicago.income
raza....chicago.race                                          0.02332320                          0.029568838
fuego....chicago.fire                                         0.02383369                          0.851732137
robo....chicago.theft                                         0.04709424                          0.829578195
edad....chicago.age                                           0.11897478                          0.133674313
politicaviviendal....chicago.volact                           0.37276314                          0.000792452
politicaviviendal2....chicago.involact                        0.00000000                          0.980656638
ingresofamiliar....chicago.income                             0.98065664                          0.000000000
```

Del resultado anterior se pueden obtener las siguientes conclusiones. Estas mismas resultarán relevantes en la realización del modelo de regresión:

RACE

```
# RACE y THEFT.

# RACE y AGE.

# RACE e INVOLACT.

# RACE e INCOME.
```

FIRE

```
# FIRE y VOLACT.

# FIRE e INVOLACT.
```

THEFT

```
# THEFT y FIRE.

# THEFT y VOLACT.

# THEFT e INVOLACT.
```

VOLACT

```
# VOLACT e INCOME.
```

INVOLACT

```
# INVOLACT y FIRE.
```

Dado que estas relaciones son aquellas que sí son relevantes (luego de calcular correlación simple, parcial y p valor), se realizan dos comparaciones a modo de ejemplo entre volact e involact:

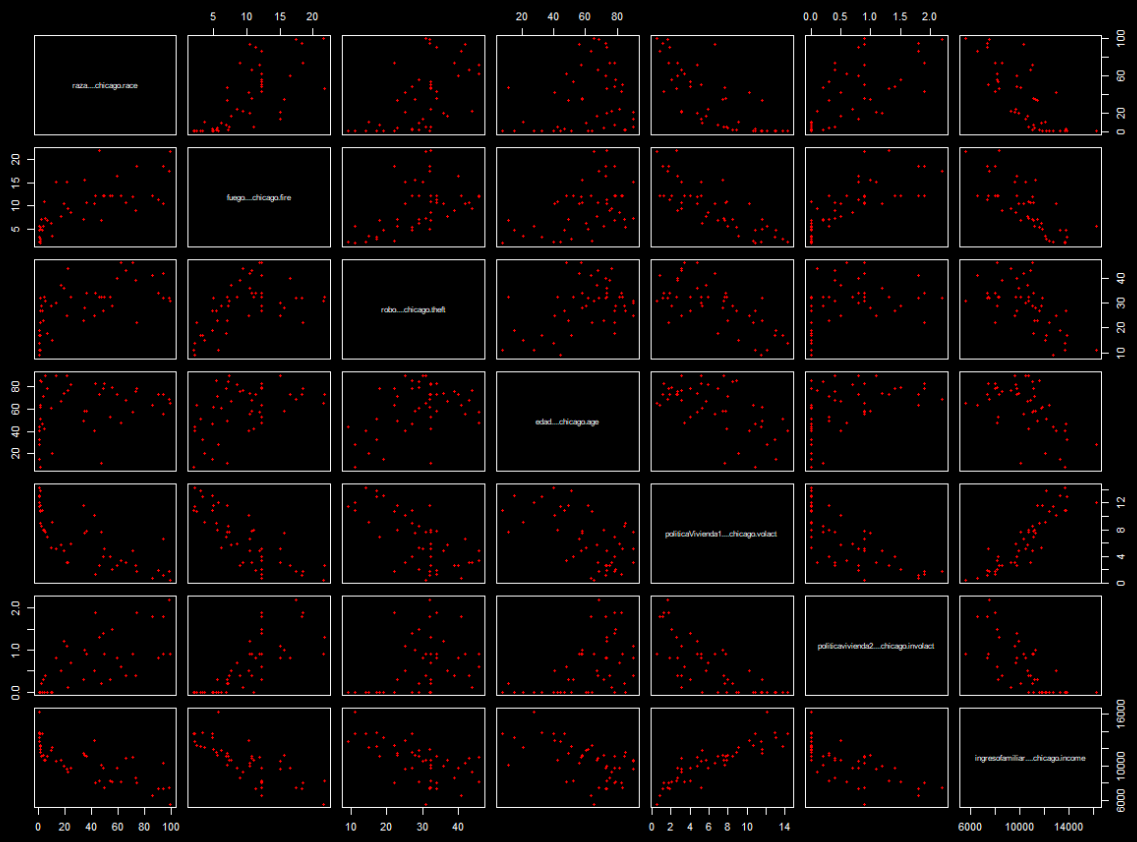
- Tal como específica en el nombre, el nuevo plan de vivienda "FAIR" hace alusión a un plan más justo, o más equitativo. El viejo plan de viviendas con la variable income tenía un p valor del 0.00079, el cual implica que ambas variables tienen una alta asociación. Se podría asumir que los ingresos medios era una variable relevante a la hora de acceder a estas viviendas.
  Sin embargo, con el nuevo plan de viviendas e income se observa un p valor igual a 0.98, es decir, los ingresos medios dejaron de tener una relación significativa entre sí, pudiendo asumir que efectivamente el nuevo plan de viviendas fue equitativo a la hora de ser más accesible a todo tipo de familias, independientemente de los ingresos medios que estas tuvieran.

- La variable race y volact tienen un p valor igual a 0.4749, por lo cual se podría asumir que extranjeros no renovaban o solicitaban pólizas con la vieja política de viviendas, lo cual tiene lógica dado que es evidente que viviendas antiguas fueran pertenecientes a dueños estadounidenses.
  La relación entre race e involact tiene un p valor del 0.0223, es decir, coincide con que extranjeros son los que más solicitan o renuevan pólizas bajo la nueva política de viviendas.

Como spoiler, estas dos variables (volact e involact) serán utilizadas como base de modelo de regresión múltiple y sometidas a comparación más adelante.

3.7. Correlaciones en forma gráfica.

A través del comando `pairs(data)` en R STUDIO es posible realizar una gráfica en forma matricial, que entrega un variable de dispersión por cada relación entre dos variables. Este comando es sumamente intuitivo para darse una idea inmediatamente del tipo de correlaciones en una base de datos. A continuación, el resultado:



A modo de ejemplo, se adjuntan cuatro comentarios acerca del gráfico pairs:

- La relación entre volact e income es una correlación positiva y fuerte. Es la más evidente de todas y con menor dispersión entre sí. Esto nos entrega la posibilidad de analizar que, en viviendas con solicitudes o renovaciones de pólizas, bajo la vieja política de viviendas, está asociada a familias con ingresos medios tanto altos como bajos, donde la mayoría de estas viviendas presentan familias que poseen un nivel bajo-medio de ingreso, con una minoría con alto nivel de ingresos.

- La relación entre involact e income es una correlación negativa moderada. A mayor cantidad de pólizas bajo el nuevo plan de vivienda FAIR, menor es el ingreso medio de las familias.

- En comparación con la relación entre volact e income, se observa que, mayor cantidad de familias con bajos ingresos acceden a este tipo de viviendas.

- La relación entre fire e income es una correlación negativa moderada. A mayor ingresos medios menos viviendas sufren incendios.

4. Modelos de regresión lineal simple.

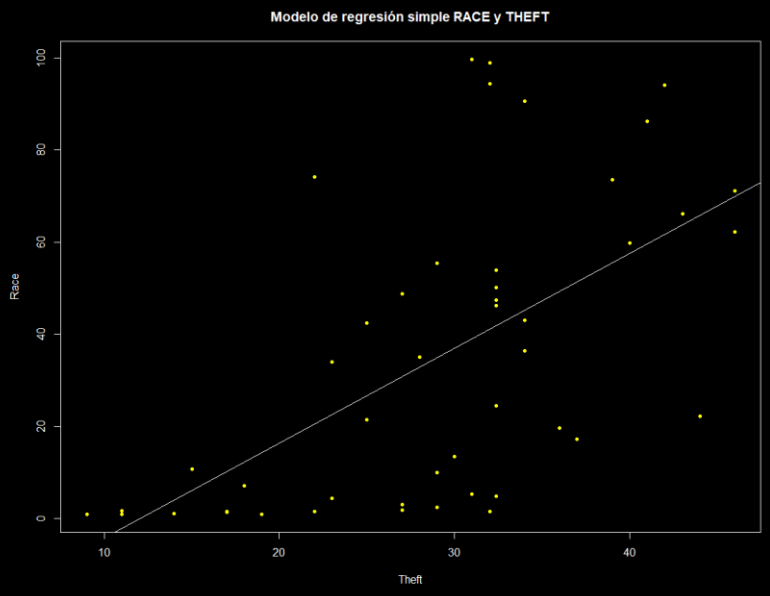
A partir de las relaciones significativas vistas en el punto 3.6 se proceden a realizar las regresiones lineales simples de dichas relaciones.

4.1. RACE y THEFT.

A continuación, un modelo simple entre RACE y THEFT.

```
rel<-lm(data$raza....chicago.race~+data$robo....chicago.theft)

summary(rel)

x = data$robo....chicago.theft

y = data$raza....chicago.race

plot(x=x, y=y , xlab="Theft", ylab="Race", col.main="white",
col.sub="white", col.lab="white",col.axis="white",
fg="white",col="yellow", pch=20, main="Modelo de regresión simple
RACE y THEFT")

abline(lm(y~x))
```

La gráfica:



El resumen del modelo:

```
Call:
lm(formula = data$raza....chicago.race ~ +data$robo....chicago.theft)

Residuals:
    Min      1Q  Median      3Q     Max
-43.614 -18.490   1.263  11.829  60.687

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                -24.8961    12.7787  -1.948   0.0576 .
data$robo....chicago.theft   2.0616     0.4192   4.918 1.21e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 '**' 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26.57 on 45 degrees of freedom
Multiple R-squared:  0.3496,    Adjusted R-squared:  0.3351
F-statistic: 24.18 on 1 and 45 DF,  p-value: 1.209e-05
```

- Es decir, el modelo realizado logra explicar solo el 33.51% de lo datos, no siendo del todo certero al tener un gran porcentaje de error.

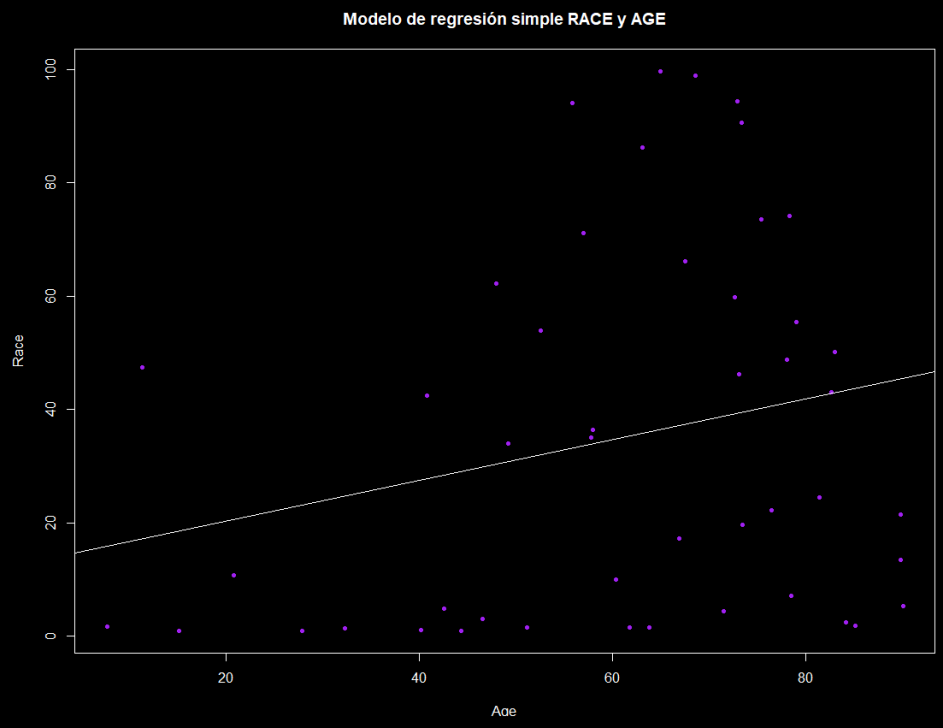
- Se observa que, el gráfico no posee un patrón mediante el cual se pueda establecer una linealidad fuerte entre ambas variables, puesto los valores están sumamente dispersos unos del otro, resultando en un R cuadrado bajo.

4.2. RACE y AGE.

A continuación, un modelo simple entre RACE y AGE.

```
rel<-lm(data$raza....chicago.race~+data$edad....chicago.age)

rel

summary(rel)

x = data$edad....chicago.age

y = data$raza....chicago.race

plot(x=x, y=y , xlab="Age", ylab="Race", col.main="white",
col.sub="white", col.lab="white",col.axis="white",
fg="white",col="purple", pch=20, main="Modelo de regresión simple
RACE s/ y AGE")

abline(lm(y~x))
```

La gráfica:



El resumen del modelo:

```
Call:
lm(formula = data$raza....chicago.race ~ +data$edad....chicago.age)

Residuals:
    Min      1Q  Median      3Q     Max
-41.834 -25.609  -9.883  25.396  63.271

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               13.1267    14.3254   0.916    0.364
data$edad....chicago.age   0.3585     0.2221   1.614    0.113

Residual standard error: 32.03 on 45 degrees of freedom
Multiple R-squared:  0.05473,   Adjusted R-squared:  0.03372
F-statistic: 2.605 on 1 and 45 DF,  p-value: 0.1135
```

- Es decir, el modelo realizado logra explicar solo el 5.473% de lo datos. No es una relación relevante.

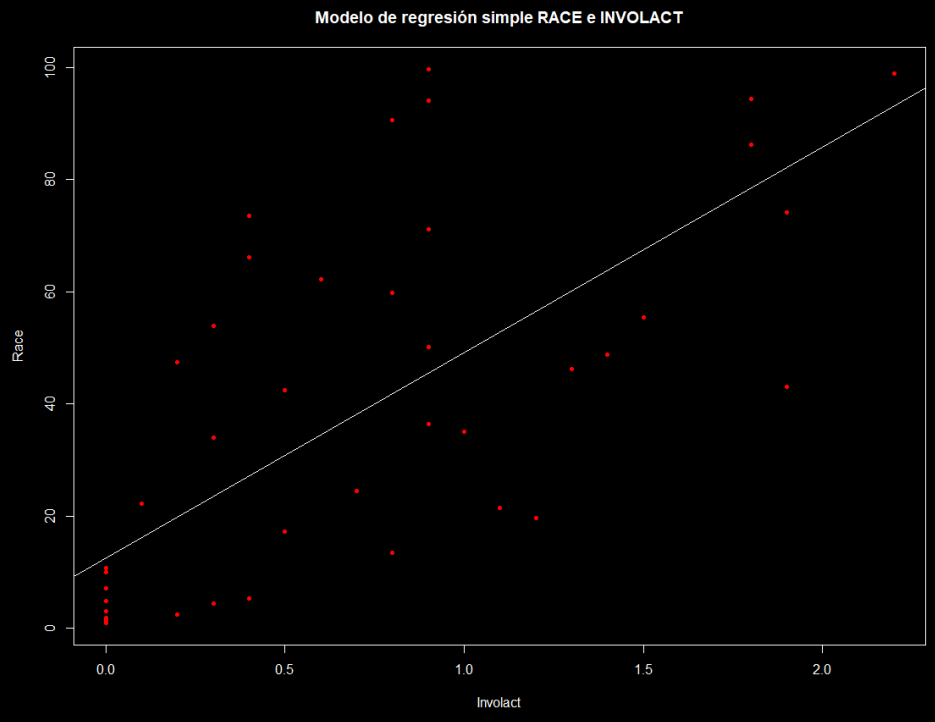
- La situación se repite respecto al gráfico anterior, pero con aún más dispersión, causando que el modelo se descarte, es decir, no es una relación relevante.

4.3. RACE e INVOLACT.

A continuación, un modelo simple entre RACE e INVOLACT.

```
rel<-
lm(data$raza....chicago.race~+data$politicavivienda2....chicago.i
nvolact)

summary(rel)

x = data$politicavivienda2....chicago.involact

y = data$raza....chicago.race

plot(x=x, y=y , xlab="Involact", ylab="Race", col.main="white",
col.sub="white", col.lab="white",col.axis="white",
fg="white",col="red", pch=20, main="Modelo de regresión simple
RACE s/ e INVOLACT")

abline(lm(y~x))
```

La gráfica:



El resumen del modelo:

```
Call:
lm(formula = data$raza....chicago.race ~ +data$politicavivienda2....chicago.involact)

Residuals:
   Min     1Q Median     3Q    Max
-39.05 -12.72  -9.32  11.15  54.25

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                                 12.420      4.714   2.635   0.0115 *
data$politicavivienda2....chicago.involact  36.697      5.368   6.836 1.78e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 '**' 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23.08 on 45 degrees of freedom
Multiple R-squared:  0.5094,    Adjusted R-squared:  0.4985
F-statistic: 46.73 on 1 and 45 DF,  p-value: 1.784e-08
```

- Es decir, el modelo realizado logra explicar el 49.85% de los datos.

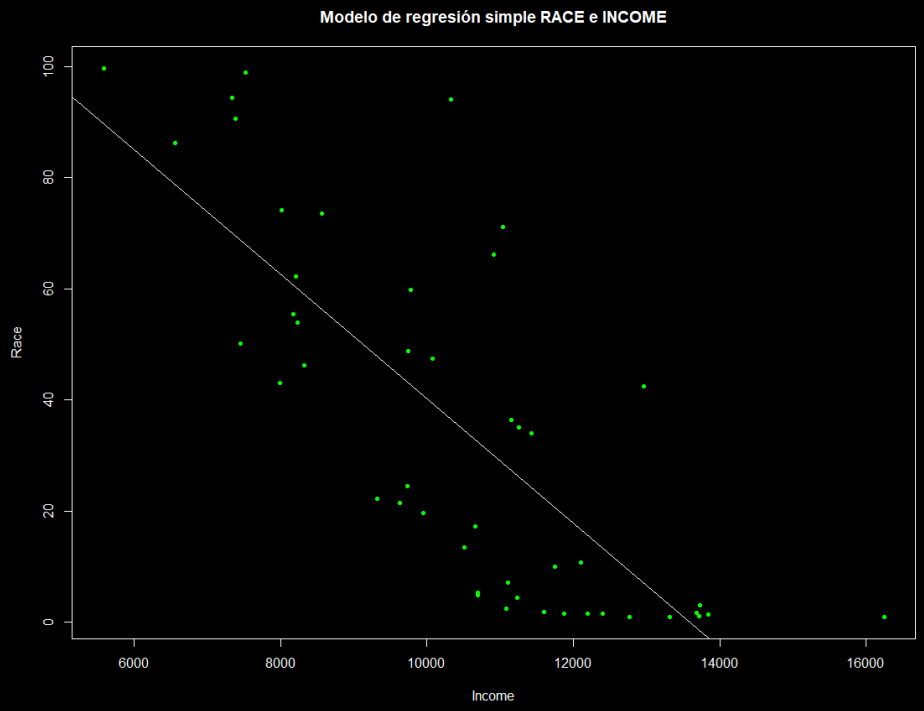
- Esta relación posee una correlación positiva algo más fuerte que las anteriores, prácticamente del 50%. Aquí se podría asumir que, si bien existe una correlación entre la composición racial y la cantidad de pólizas solicitadas o renovadas, no es lo suficientemente fuerte como para llegar a una conclusión significativa.

4.4. RACE e INCOME.

A continuación, un modelo simple entre RACE e INCOME.

```
rel<-
lm(data$raza....chicago.race~+data$ingresofamiliar....chicago.income)

summary(rel)

x = data$ingresofamiliar....chicago.income

y = data$raza....chicago.race

plot(x=x, y=y , xlab="Income", ylab="Race", col.main="white", col.sub="white", col.lab="white",col.axis="white", fg="white",col="purple", pch=20, main="Modelo de regresión simple RACE s/ e INCOME")

abline(lm(y~x))
```

La gráfica:



El resumen del modelo:

```
Call:
lm(formula = data$raza....chicago.race ~ +data$ingresofamiliar....chicago.income)

Residuals:
    Min      1Q  Median      3Q     Max
-27.512 -18.626  -1.952   9.876  57.608

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                       152.352127  14.831037  10.273 2.22e-13 ***
data$ingresofamiliar....chicago.income  -0.011214   0.001386  -8.089 2.56e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 '**' 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ' 1

Residual standard error: 21.03 on 45 degrees of freedom
Multiple R-squared:  0.5925,    Adjusted R-squared:  0.5834
F-statistic: 65.42 on 1 and 45 DF,  p-value: 2.563e-10
```

- Es decir, el modelo realizado logra explicar el 58.34%, lo cual es un valor considerablemente alto.

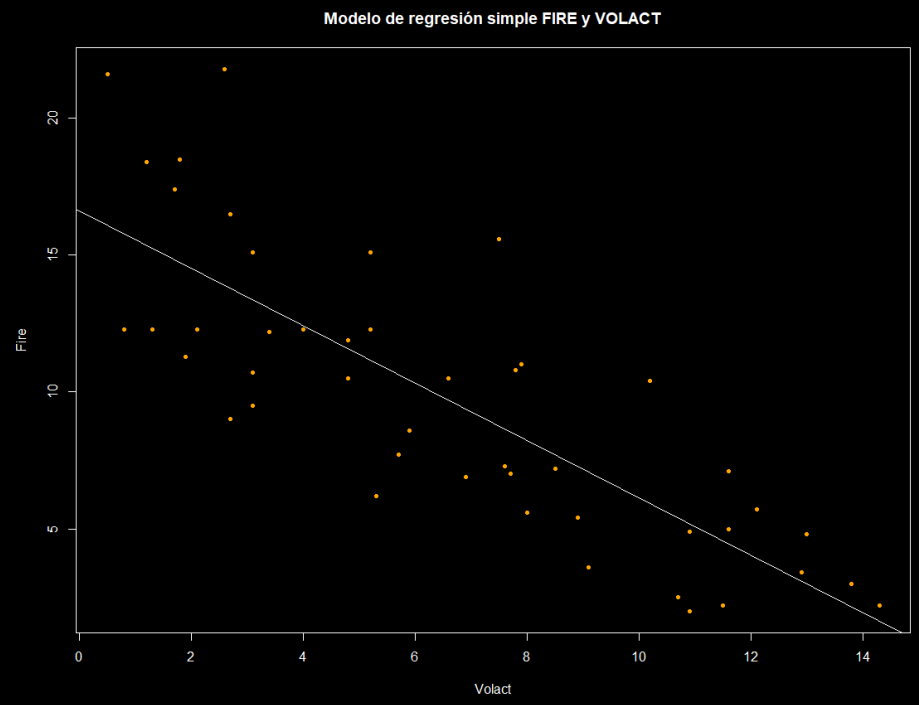
- El modelo explica casi al 60% los valores reales. Existe una correlación negativa entre moderada y fuerte entre el ingreso medio y la composición racial de las familias de las viviendas. Se observa que, aquellas viviendas que presentaron una composición racial superior son aquellas que menos ingresos medios perciben, mientras, aquellas viviendas con familias estadounidenses perciben un ingreso medio alto. A mayor cantidad de extranjeros dueños de viviendas que solicitan o renuevan pólizas, menor es el ingreso medio que perciben dichas familias.

4.5. FIRE y VOLACT.

A continuación, un modelo simple entre FIRE y VOLACT.

```
rel<-
lm(data$fuego....chicago.fire~+data$politicaVivienda1....chicago.
volact)

summary(rel)

x = data$politicaVivienda1....chicago.volact

y = data$fuego....chicago.fire

plot(x=x, y=y , xlab="Volact", ylab="Fire", col.main="white",
col.sub="white", col.lab="white", col.axis="white",
fg="white",col="orange", pch=20, main="Modelo de regresión simple
FIRE y VOLACT")

abline(lm(y~x))
```

La gráfica:



El resumen del modelo:

```
Call:
lm(formula = data$fuego....chicago.fire ~ +data$politicaVivienda1....chicago.volact)

Residuals:
    Min      1Q  Median      3Q     Max
-4.8710 -2.5682 -0.1542  2.0729  7.8998

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                          16.6245     0.8671  19.174  < 2e-16 ***
data$politicaVivienda1....chicago.volact  -1.0478     0.1138  -9.206 6.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 '**' 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.062 on 45 degrees of freedom
Multiple R-squared:  0.6532,    Adjusted R-squared:  0.6455
F-statistic: 84.76 on 1 and 45 DF,  p-value: 6.491e-12
```

- Es decir, el modelo realizado logra explicar el 64.55%, lo cual es un valor considerablemente alto.

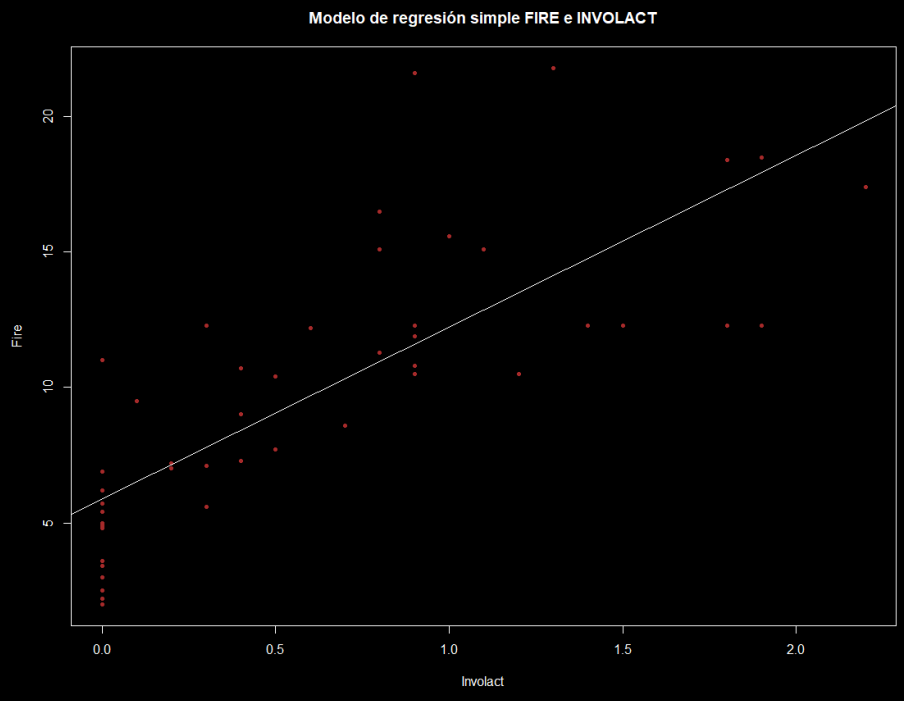
- El gráfico nos permite observar que existe una correlación negativa entre ambas variables, sin embargo, su lectura se hace difícil puesto existe una dispersión considerable en el mismo.

4.6. FIRE e INVOLACT.

A continuación, un modelo simple entre FIRE e INVOLACT.

```
rel<-
lm(data$fuego....chicago.fire~+data$politicavivienda2....chicago.
involact)

summary(rel)

x = data$politicavivienda2....chicago.involact

y = data$fuego....chicago.fire

plot(x=x, y=y , xlab="Involact", ylab="Fire", col.main="white",
col.sub="white", col.lab="white",col.axis="white",
fg="white",col="brown", pch=20, main="Modelo de regresión simple
FIRE e INVOLACT")

abline(lm(y~x))
```

La gráfica:



El resumen del modelo:

```
Call:
lm(formula = data$fuego....chicago.fire ~ +data$politicavivienda2....chicago.involact)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6533 -2.3589 -0.4829  1.2240 10.0095

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                               5.8829     0.6625   8.880 1.88e-11 ***
data$politicavivienda2....chicago.involact   6.3418     0.7544   8.407 8.89e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 '**' 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.243 on 45 degrees of freedom
Multiple R-squared:  0.611,     Adjusted R-squared:  0.6023
F-statistic: 70.67 on 1 and 45 DF,  p-value: 8.889e-11
```

- Es decir, el modelo realizado logra explicar el 60.23%, lo cual es un valor considerablemente alto.

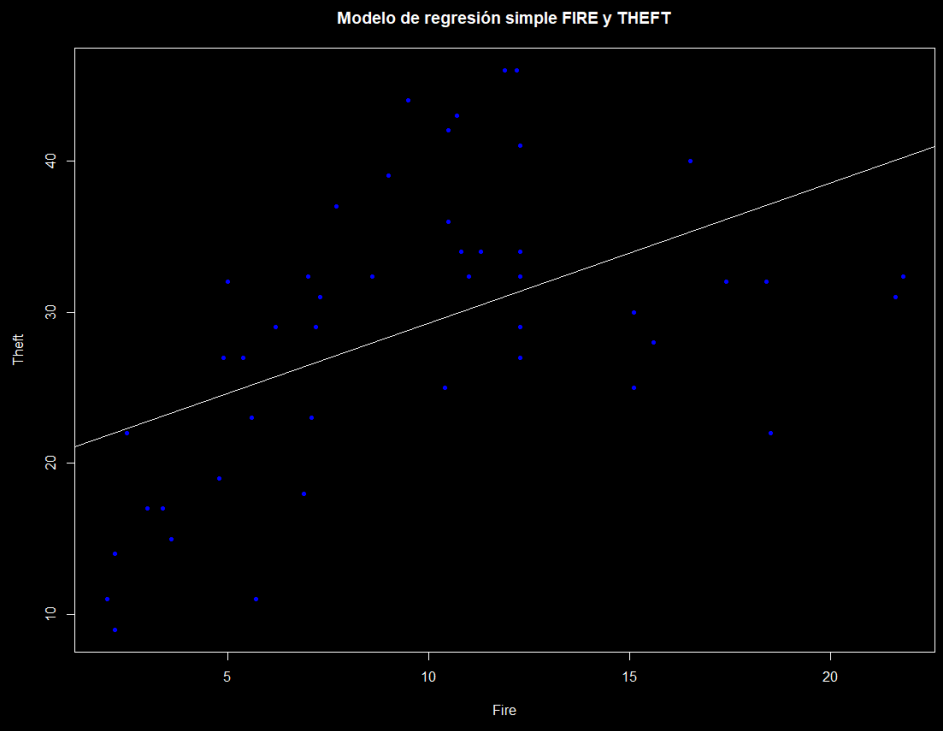
- Existe una correlación positiva entre ambas variables. Se observa que, de aquellas viviendas que presentaron incendios, gran parte no solicitaron o renovaron bajo la nueva política de viviendas FAIR.

4.7. FIRE y THEFT.

A continuación, un modelo simple entre FIRE e THEFT.

```
rel<-lm(data$robo....chicago.theft~+data$fuego....chicago.fire)

summary(rel)

x = data$fuego....chicago.fire

y = data$robo....chicago.theft

plot(x=x, y=y , col.main="white", col.sub="white",
col.lab="white",col.axis="white", fg="white",col="cyan", pch=20,
ylab="Theft", xlab="Fire", main="Modelo de regresión simple FIRE
y THEFT")

abline(lm(y~x))
```

La gráfica:



El resumen del modelo:

```
Call:
lm(formula = data$robo....chicago.theft ~ +data$fuego....chicago.fire)

Residuals:
     Min       1Q   Median       3Q      Max
-15.1515  -5.9261   0.9946   4.5616  15.2164

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 19.9509     2.5667   7.773 7.39e-10 ***
data$fuego....chicago.fire   0.9298     0.2328   3.994 0.000238 ***
---
Signif. codes:  0 ‘***’ 0.001 '**' 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.119 on 45 degrees of freedom
Multiple R-squared:  0.2617,    Adjusted R-squared:  0.2453
F-statistic: 15.95 on 1 and 45 DF,  p-value: 0.0002375
```

- Es decir, el modelo realizado logra explicar el 24.53% del total de datos reales, lo cual es un valor bajo.

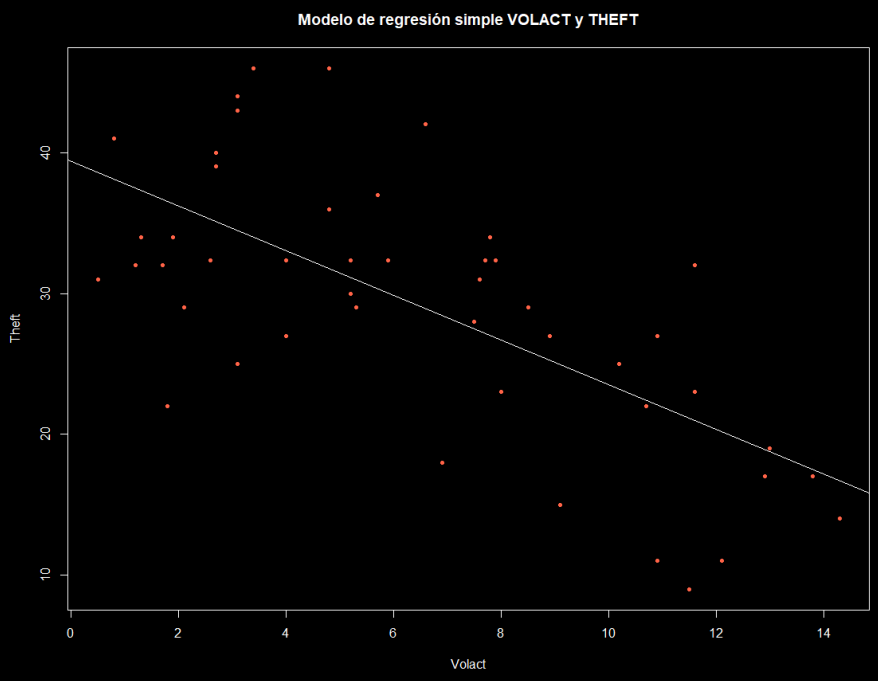
- El valor de la correlación es muy bajo dado que existe una gran dispersión entre ambas variables, resultando en un R cuadrado bajo.

4.8. THEFT y VOLACT.

A continuación, un modelo simple entre THEFT y VOLACT.

```
rel<-
lm(data$robo....chicago.theft~+data$politicaVivienda1....chicago.
volact)

summary(rel)

x = data$politicaVivienda1....chicago.volact

y = data$robo....chicago.theft

plot(x=x, y=y , col.main="white", col.sub="white",
col.lab="white",col.axis="white", fg="white",col="tomato",
pch=20, ylab="Theft", xlab="Volact", main="Modelo de regresión
simple VOLACT y THEFT")

abline(lm(y~x))
```

La gráfica:



El resumen del modelo:

```
Call:
lm(formula = data$robo....chicago.theft ~ +data$politicaVivienda1....chicago.volact)

Residuals:
     Min       1Q   Median       3Q      Max
-14.5445  -4.2092   0.2112   4.5468  14.2115

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                          39.3980     1.9795  19.903  < 2e-16 ***
data$politicaVivienda1....chicago.volact  -1.5853     0.2598  -6.101 2.21e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 '**' 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ' 1

Residual standard error: 6.99 on 45 degrees of freedom
Multiple R-squared:  0.4527,    Adjusted R-squared:  0.4406
F-statistic: 37.22 on 1 and 45 DF,  p-value: 2.209e-07
```

- El modelo realizado logra explicar el 44.06% del total de datos reales, por lo que la mitad y un poco más de los datos quedan sin explicar.

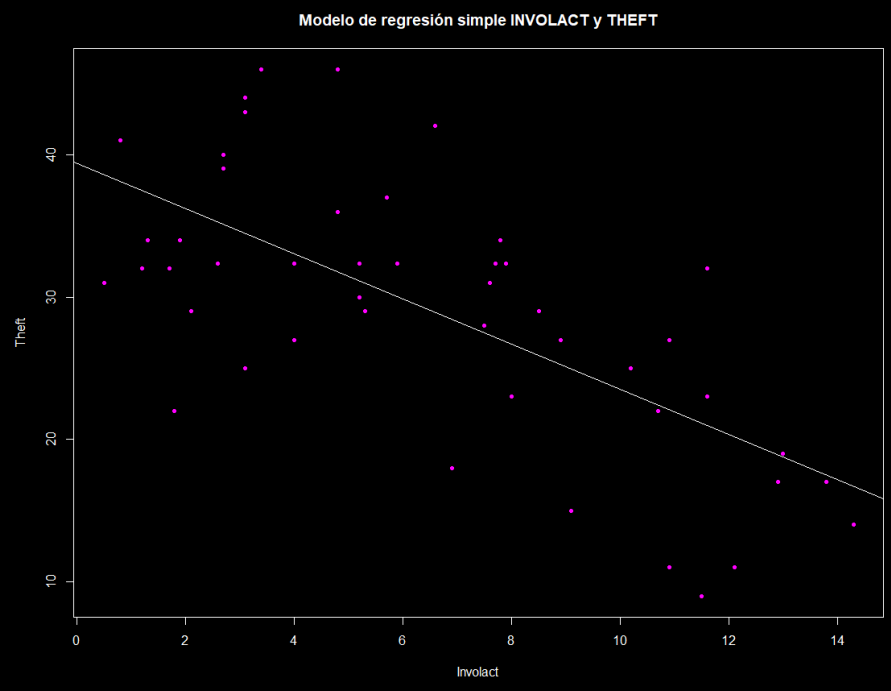
- Si bien, es una correlación significativa, no lo es del todo para obtener conclusiones relevantes a partir del modelo.

4.9. THEFT y VOLACT.

A continuación, un modelo simple entre THEFT y VOLACT.

```
rel<-
lm(data$robo....chicago.theft~+data$politicavivienda2....chicago.
involact)

summary(rel)

x = data$politicavivienda2....chicago.involact

y = data$robo....chicago.theft

plot(x=x, y=y , col.main="white", col.sub="white",
col.lab="white",col.axis="white", fg="white",col="magenta",
pch=20, ylab="Theft", xlab="Involact", main="Modelo de regresión
simple INVOLACT y THEFT")

abline(lm(y~x))
```

La gráfica:



El resumen del modelo:

```
Call:
lm(formula = data$robo....chicago.theft ~ +data$politicavivienda2....chicago.involact)

Residuals:
     Min       1Q   Median       3Q      Max
-16.4694  -6.3681   0.2953   5.1038  17.9489

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                               25.469      1.774   14.36  < 2e-16 ***
data$politicavivienda2....chicago.involact    5.817      2.020    2.88  0.00607 **
---
Signif. codes:  0 ‘***’ 0.001 '**' 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.683 on 45 degrees of freedom
Multiple R-squared:  0.1556,    Adjusted R-squared:  0.1369
F-statistic: 8.294 on 1 and 45 DF,  p-value: 0.006069
```

- El modelo realizado logra explicar el 13.69% del total de datos reales, siendo un valor bajo.

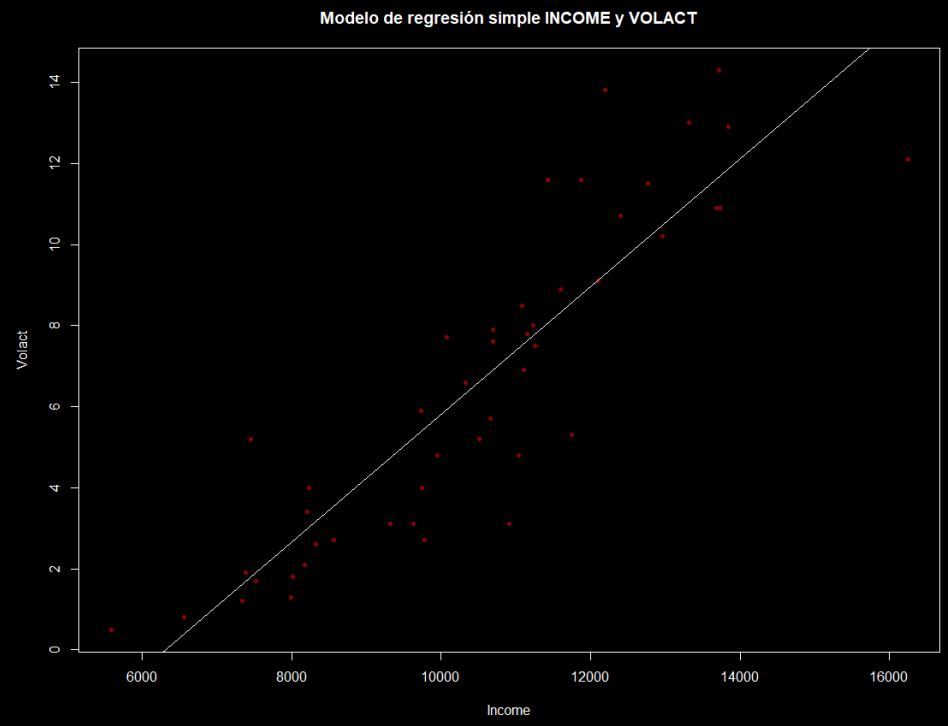
- Definitivamente, la correlación no es significativa a causa de la enorme dispersión entre variables.

4.10. INCOME y VOLACT.

A continuación, un modelo simple entre INCOME y VOLACT.

```
rel<-
lm(data$politicaVivienda1....chicago.volact~+data$ingresofamiliar
....chicago.income)

summary(rel)

x = data$ingresofamiliar....chicago.income

y = data$politicaVivienda1....chicago.volact

plot(x=x, y=y , col.main="white", col.sub="white",
col.lab="white",col.axis="white", fg="white",col="dark red",
pch=20, ylab="Volact", xlab="Income", main="Modelo de regresión
simple INCOME y VOLACT")

abline(lm(y~x))
```

La gráfica:



El resumen del modelo:

```
Call:
lm(formula = data$politicaVivienda1....chicago.volact ~ +data$ingresofamiliar....chicago.income)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1255 -0.8897 -0.0066  1.0030  4.5422

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                         -9.9592860  1.2978746  -7.674 1.03e-09 ***
data$ingresofamiliar....chicago.income  0.0015754  0.0001213  12.986  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 '**' 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.841 on 45 degrees of freedom
Multiple R-squared:  0.7893,    Adjusted R-squared:  0.7847
F-statistic: 168.6 on 1 and 45 DF,  p-value: < 2.2e-16
```

- El modelo realizado logra explicar el 78.47% del total de datos reales, siendo un valor alto.

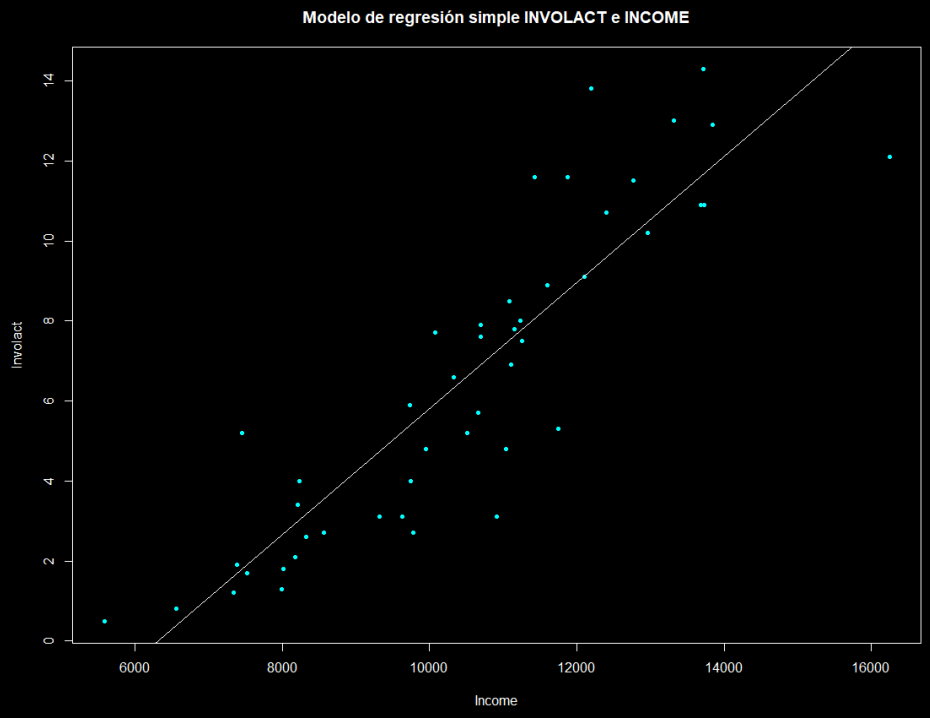
- En el gráfico se observa una correlación positiva fuerte. Se podría asumir que a mayor cantidad de ingresos, mayor es la cantidad de solicitudes o renovaciones de pólizas bajo el nuevo plan de viviendas FAIR. Sin embargo, gran parte gran parte de estas solicitudes o renovaciones las realizan viviendas con familias de un nivel de ingreso medio entre bajo y moderado.

4.11. THEFT e INVOLACT.

A continuación, un modelo simple entre THEFT y INVOLACT.

```
rel<-
lm(data$politicavivienda2....chicago.involact~+data$ingresofamili
ar....chicago.income)

summary(rel)

x = data$ingresofamiliar....chicago.income

y = data$politicavivienda2....chicago.involact

plot(x=x, y=y , col.main="white", col.sub="white",
col.lab="white",col.axis="white", fg="white",col="cyan", pch=20,
ylab="Involact", xlab="Income", main="Modelo de regresión simple
INVOLACT e INCOME")

abline(lm(y~x))
```

La gráfica:



El resumen del modelo:

```
Call:
lm(formula = data$politicavivienda2....chicago.involact ~ +data$ingresofamiliar....chicago.income)

Residuals:
     Min       1Q   Median       3Q      Max
-0.76900 -0.32718 -0.06448  0.39212  0.98657

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                        2.741e+00  3.150e-01   8.701 3.36e-11 ***
data$ingresofamiliar....chicago.income -2.031e-04  2.945e-05  -6.899 1.44e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 '**' 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4467 on 45 degrees of freedom
Multiple R-squared:  0.514,     Adjusted R-squared:  0.5032
F-statistic: 47.59 on 1 and 45 DF,  p-value: 1.441e-08
```

- El modelo realizado logra explicar el 50.32% del total de datos reales, siendo un valor moderado.

- Si bien existe una correlación positiva entre moderada y fuerte, se logra comparar que entre la vieja política de viviendas y la nueva, se logró bajar la significancia de la variable INCOME en la solicitud o renovación de pólizas a viviendas, por lo cual, la accesibilidad a dichas viviendas bajo el plan FAIR fue efectiva, reduciendo la implicancia del ingreso medio de las familais.

5. Modelos de regresión lineal múltiple.

Cabe recalcar que este tipo de modelos son sumamente más complejos, y en la mayoría de las veces, es imposible hallar una relación lineal con más de dos variables, puesto a medida cada vez se añaden más variables, más compleja debería ser la función que intentará explicar el fenómeno.

5.1. Metodología de modelo.

No todos los problemas pueden ser solucionados con el mismo algoritmo, sin embargo, se asume linealidad entre variables y es la metodología que se utilizará para la presente sección.

Un modelo es, netamente, una relación entre dos o más variables que intenta explicar un fenómeno. En este caso, se intentará explicar cómo variables demográficas y económicas inciden en las pólizas solicitadas por dueños de viviendas según política antigua o renovada en la ciudad de Chicago, para así poder llegar a ciertas conclusiones comparativas y características de cada política de vivienda, como, por ejemplo, el caso de la seguridad ante robos o incendios con las variables theft y fire, viendo como estas inciden en ambos tipos de políticas.

A partir del comando `lm(y~x)` en R STUDIO, podemos construir un modelo de regresión lineal, donde Y vendría a ser nuestra variable dependiente, que es la respuesta de las variables independientes y predictoras X. En este caso, se intenta responder el número de pólizas solicitadas o renovadas (volact e involact en forma separada) a partir de las variables predictoras (race, fire, theft, age, income), por lo cual se requerirá de la extensión del modelo de regresión lineal, que es el múltiple.

La naturaleza de un modelo es el error, más aún cuando se intenta corresponder a un fenómeno de la vida real, donde el cambio es una constante y las variables son infinitas.

Es por ello por lo que, se intentará minimizar lo más posible el error, **pero nunca estaremos libres de él**. Por lo que, durante el proceso, es posible que algunas de estas variables predictoras no sean significativas para el modelo, y a veces, pueden entorpecer en el mismo, aumentando el error. En tales instancias, se procederá a eliminar dichas variables hasta lograr el modelo con mayor exactitud posible.

5.1. Primera regresión lineal múltiple: involact.

Como se especificó en la sección anterior, debemos recurrir a la **regresión lineal múltiple,** cuya fórmula puede ser empleada a partir de `lm(data$politicavivienda2....chicago.involact~data$raza....chicago.race+data$fuego....chicago.fire+data$robo....chicago.theft+data$edad....chicago.age+data$ingresofamiliar....chicago.income).`

Aquí, es importante observar el símbolo "~" continuado de la suma de variables. El símbolo "~" asume una función lineal, donde la variable de la izquierda será nuestra variable a explicar o predecir y la o las variables de la derecha nuestras predictoras.

El resultado es el siguiente:

```
Call:
lm(formula = regresion)

Coefficients:
                        (Intercept)              data$raza....chicago.race              data$fuego....chicago.fire              data$robo....chicago.theft
                          1.026e-01                              8.191e-03                              5.075e-02                             -1.449e-02
              data$edad....chicago.age  data$ingresofamiliar....chicago.income
                          6.663e-03                             -2.446e-05
```

Es decir, nuestra función sería:

$$1.026 * 10^{-1} + (8.191 * 10^{-3})x_1 + (5.075 * 10^{-2})x_2 + (-1.449 * 10^{-2})x_3 + (6.663 * 10^{-3})x_4 - (1.449 * 10^{-2})x_5 = y$$

A continuación, se adjunta el resumen del modelo a partir de `summary()`:

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.96266 -0.15721 -0.01726  0.14679  0.95854

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                        1.026e-01  7.705e-01   0.133   0.8947
data$raza....chicago.race          8.191e-03  3.197e-03   2.562   0.0142 *
data$fuego....chicago.fire         5.075e-02  1.777e-02   2.856   0.0067 **
data$robo....chicago.theft        -1.449e-02  7.724e-03  -1.877   0.0677 .
data$edad....chicago.age           6.663e-03  3.489e-03   1.910   0.0632 .
data$ingresofamiliar....chicago.income -2.446e-05  4.924e-05  -0.497   0.6220
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3639 on 41 degrees of freedom
Multiple R-squared:  0.7062,    Adjusted R-squared:  0.6703
F-statistic: 19.71 on 5 and 41 DF,  p-value: 5.814e-10
```

De aquí es posible resumir:

- El modelo tiene una media de error de -0.01726.

- El R^2 nos entrega cuánto porcentaje explica el modelo de la variable, es decir, el 67.03%. El modelo no logra explicar un 32.97% de los datos.

- Se observa que aquellas variables significativas van en el orden del "***", "**", "*", "." Y "". El intercepto y la variable income no son significativas, habrá que removerlas para lograr que el modelo explique en un porcentaje más elevado los datos reales.

5.2. Segunda regresión lineal múltiple: mejoramiento del modelo de involact y gráficas.

Siguiendo el procedimiento anterior, se procede a eliminar las variables no significativas para así lograr que el modelo explique mayor número de datos.

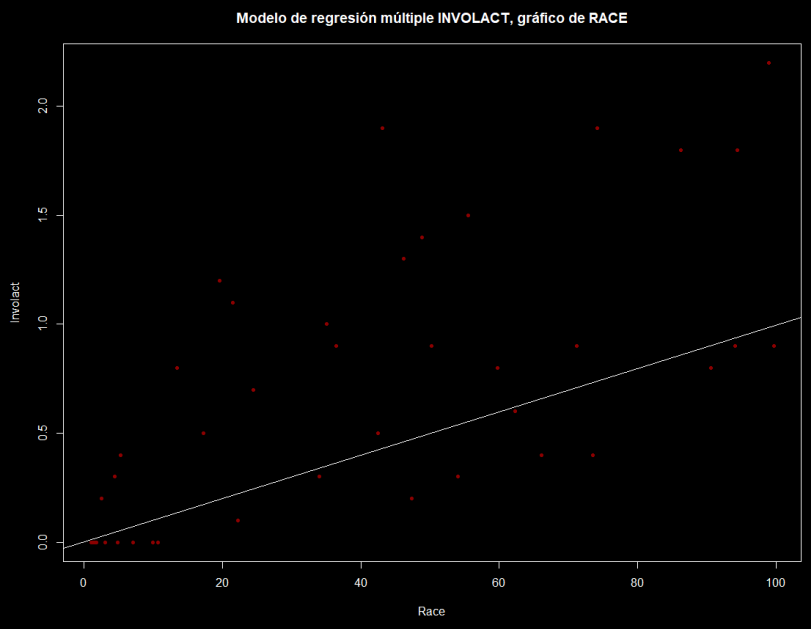
Resultando:

```
Call:
lm(formula = regresionfix)

Coefficients:
 data$raza....chicago.race  data$fuego....chicago.fire  data$robo....chicago.theft     data$edad....chicago.age
                  0.009963                     0.050014                    -0.019856                     0.006095
```

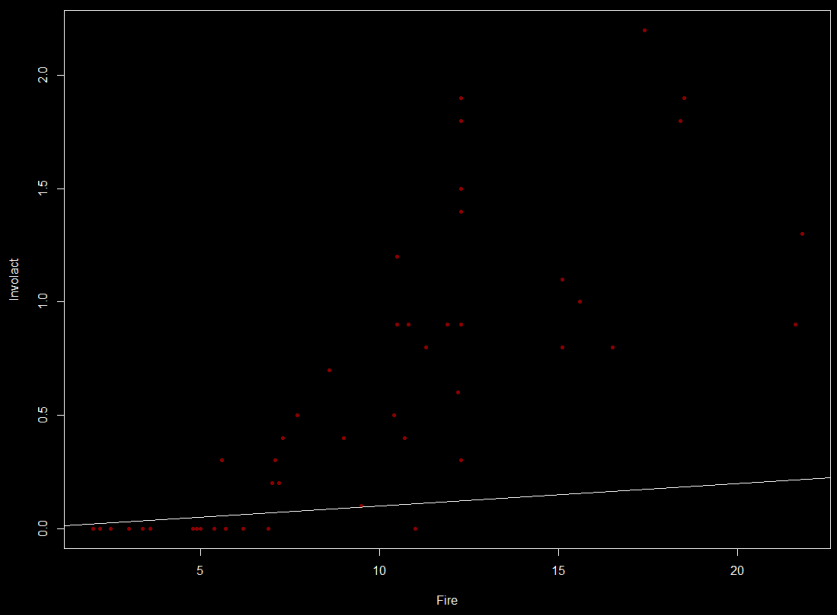
Es decir, nuestra función sería:

$$0.009963x_1 + 0.050014x_2 - 0.019856x_3 + 0.006095x_4 = y$$

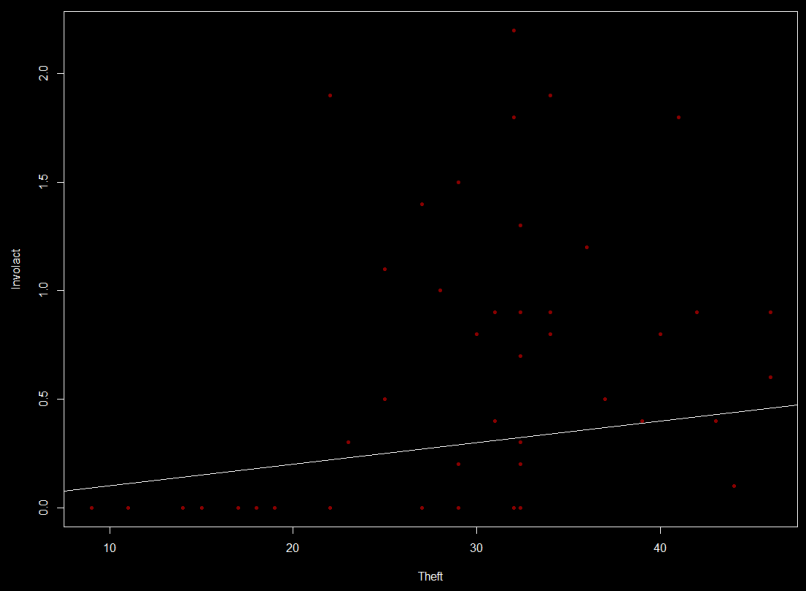
Las gráficas 2D para cada variable (hechas con `plot()` y `abline()`):

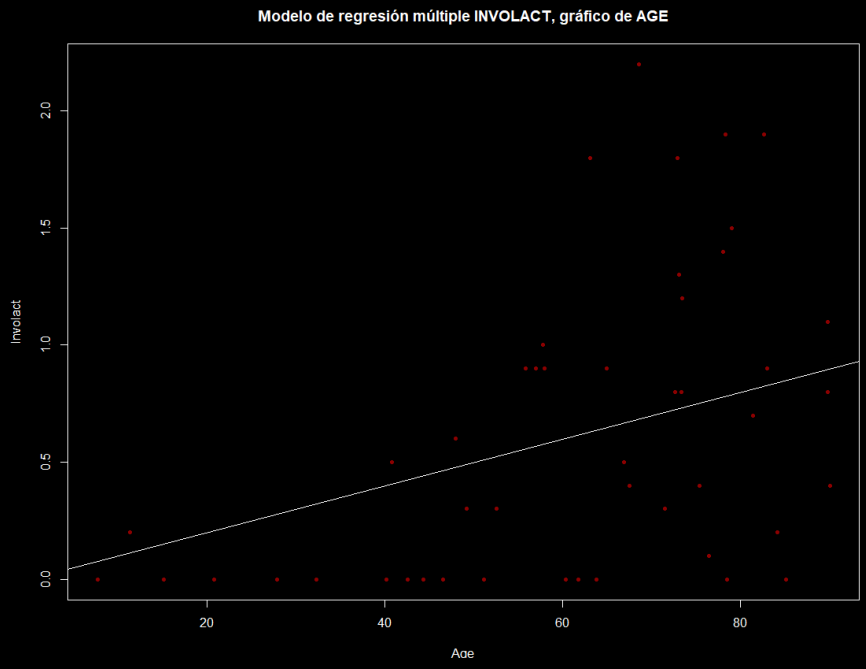


Modelo de regresión múltiple INVOLACT, gráfico de FIRE



Modelo de regresión múltiple INVOLACT, gráfico de THEFT





A continuación, se adjunta el resumen del modelo a partir de `summary()`:

```
Call:
lm(formula = regresionfix)

Residuals:
     Min       1Q   Median       3Q      Max
-0.95430 -0.20698 -0.04927  0.10495  1.02749

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
data$raza....chicago.race   0.009963   0.002625   3.796 0.000456 ***
data$fuego....chicago.fire  0.050014   0.017006   2.941 0.005250 **
data$robo....chicago.theft -0.019856   0.006197  -3.204 0.002553 **
data$edad....chicago.age    0.006095   0.002930   2.080 0.043514 *
---
Signif. codes:  0 ‘***’ 0.001 '**' 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3636 on 43 degrees of freedom
Multiple R-squared:  0.8431,    Adjusted R-squared:  0.8286
F-statistic: 57.79 on 4 and 43 DF,  p-value: < 2.2e-16
```

De aquí es posible resumir:

- El modelo tiene una media de error de -0.04927.

- El R^2 es 84.31%. El modelo no logra explicar un 15.69% de los datos. En general, el modelo tuvo una mejora, sin embargo, la media del error aumento, pero por fortuna, parece ser no representativa.

5.3. Tercera regresión lineal múltiple: volact.

Siguiendo el procedimiento para involact, ahora se realiza la regresión lineal múltiple para volact.

Resultando:

```
Call:
lm(formula = regresion)

Coefficients:
                        (Intercept)                data$raza....chicago.race               data$fuego....chicago.fire               data$robo....chicago.theft
                          4.0608144                           -0.0155861                           -0.1831528                           -0.0608794
data$edad....chicago.age  data$ingresofamiliar....chicago.income
                         -0.0266341                            0.0007833
```

Es decir, nuestra función sería:

$$4.06088144 - 0.0155861x_1 - 0.1831528x_2 - 0.0608794x_3 - 0.0266341x_4 + 0.0007833x_5 = y$$

A continuación, se adjunta el resumen del modelo a partir de `summary()`:

```
lm(formula = regresion)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2942 -0.9254  0.0383  0.9145  3.1560

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                            4.0608144  3.2945238   1.233 0.224748
data$raza....chicago.race             -0.0155861  0.0136707  -1.140 0.260857
data$fuego....chicago.fire            -0.1831528  0.0759725  -2.411 0.020486 *
data$robo....chicago.theft            -0.0608794  0.0330241  -1.843 0.072495 .
data$edad....chicago.age              -0.0266341  0.0149168  -1.786 0.081578 .
data$ingresofamiliar....chicago.income 0.0007833  0.0002105   3.721 0.000597 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.556 on 41 degrees of freedom
Multiple R-squared:  0.8628,    Adjusted R-squared:  0.8461
F-statistic: 51.58 on 5 and 41 DF,  p-value: < 2.2e-16
```

De aquí es posible resumir:

- El modelo tiene una media de error de -0.0383.
- El R^2 fue 84.61%. El modelo no logra explicar un 15.39% de los datos.
- El intercepto y race no son significativos, se removerán del modelo.

5.4. Cuarta regresión lineal múltiple: mejoramiento del modelo de volact.

Siguiendo el procedimiento de mejoramiento para involact, se eliminarán variables no significativas, es decir, el intercepto y race.

Resultando:

```
call:
lm(formula = regresionfix)

Coefficients:
           data$robo....chicago.theft              data$edad....chicago.age            data$fuego....chicago.fire  data$ingresofamiliar....chicago.income
                             -0.060057                             -0.014481                             -0.186881                              0.001048
```

Es decir, nuestra función sería:

$$-0.060057x_1 - 0.014481x_2 - 0.186881x_3 + 0.001048x_4 = y$$

Las gráficas 2D para cada variable se omiten dado que, el modelo pareciera no ajustarse a la dispersión de ninguna variable, puesto en las gráficas no aparece la línea ajustada. Habría que analizar el por qué de dicho detalle, pero para efectos del presente documento se omite.

A continuación, se adjunta el resumen del modelo a partir de `summary()`:

```
Call:
lm(formula = regresionfix)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2354 -1.0821 -0.0157  0.9752  3.3367

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
data$robo....chicago.theft              -6.006e-02  2.755e-02  -2.180  0.03479 *
data$edad....chicago.age                -1.448e-02  1.212e-02  -1.195  0.23872
data$fuego....chicago.fire              -1.869e-01  5.695e-02  -3.281  0.00205 **
data$ingresofamiliar....chicago.income   1.048e-03  5.136e-05  20.411  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 '**' 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.555 on 43 degrees of freedom
Multiple R-squared:  0.9619,    Adjusted R-squared:  0.9583
F-statistic: 271.2 on 4 and 43 DF,  p-value: < 2.2e-16
```

De aquí es posible resumir:

- El modelo tiene una media de error de -0.0157.
- El R^2 fue 95.38%. El modelo no logra explicar un 4.62% de los datos. El modelo tuvo una mejora.

6. Conclusiones.

A partir de la investigación realizada de la base de datos Chicago, se puede concluir que no todos los fenómenos se pueden solucionar a partir de un único algoritmo. Claro ejemplo de ello es el cómo se utilizó la regresión lineal simple y posteriormente la múltiple, donde ambos resultados fueron ampliamente distintos.

Por un lado, la regresión lineal simple es una excelente herramienta para hallar relaciones de linealidad entre dos variables, siendo gran parte de las veces, muy útil para este tipo de relaciones.

Sin embargo, si dicho modelo lo extrapolamos a la regresión lineal múltiple es poco probable que logre explicar un fenómeno, puesto al involucrar más de dos variables los modelos tienden a complejizarse, no pudiendo ser explicados de manera lineal, sino por métodos menos convencionales, como exponenciales o cuadráticos.

Cabe recalcar que la presente investigación se realizó bajo un alero de aprendizaje, y en definitiva, consistió en la continua experimentación, pudiendo haber errores, por lo cual se pudo no haber sido consistente en algunas conclusiones.

Finalmente, el script utilizado:

O en el siguiente vínculo.