PEC1 - Descripción de datos multivariantes, análisis gráfico y detección de valores atípicos

M0.517 - Análisis multivariante de datos

Juan Águila Martínez

20 de abril de 2016

Resumen

El siguiente documento recoge los procedimientos y conclusiones de la primera Prueba de Evaluación Continua (PEC) del curso **Análisis multivariante de datos**. Esta prueba desarrolla la descripción de un conjunto datos multivariante, su representación gráfica y la detección de valores atípicos mediante técnicas de análisis multivariante. El documento se estructura como sigue: en el primer apartado, se describe el objetivo del ejercicio y el entorno de software utilizado (1); a continuación, se describe el conjunto de datos utilizado en la práctica, *census.dat* (2); los siguientes apartados contienen el desarrollo del ejercicio: la descripción de las medidas de centralidad y variabilidad del conjunto de datos y su representación gráfica mediante algunas de las librerías disponibles en R (3) y el desarrollo de dos técnicas de detección de valores atípicos, junto a las conclusiones de la aplicación de dichas técnicas (4).

1. Definición de los objetivos y el entorno de trabajo

En esta PEC se desarrollan los contenidos de las primeras semanas de la asignatura **Análisis multivariante de datos**: descripción de datos multivariantes, representación gráfica de datos multivariantes y detección de valores atípicos en datos multivariantes. A este efecto, se aplican las técnicas presentadas en el libro-manual de la asignatura, Análisis de datos multivariantes (Peña, 2002), al conjunto de datos *census.dat*, que se presentará en el siguiente apartado. Dada la vocación práctica de este curso, la resolución del problema planteado se ha realizado con software de cálculo estadístico y graficado indicado a continuación.

Para la realización de esta PEC se ha utilizado el lenguaje de programación ***R*** (v3.2.0) en el entorno de desarrollo ***RStudio*** en su versión Desktop para Mac OS X 10.6+ (64bit) (v0.99). Una de las principales virtudes de *R* es la amplia variedad de paquetes de funciones desarrollados con licencia *open source* (en la mayoría de los casos *copyleft* fuertes, como *GPL*), que incorporan funcionalidades de análisis estadístico descargando el proceso de análisis de la implementación de los algoritmos de cálculo. En este sentido, para la realización de esta práctica se han utilizado los siguientes paquetes:

- ***Psych***: funciones de propósito general para la descripción de datos multivariantes (https://cran.r-project.org/web/packages/psych/)
- ***corrplot***: gráficos de visualización de matrices de correlación (https://cran.r-project.org/web/packages/corrplot/)
- ***xTable***: generación de tablas LaTex a partir de DataFrames (https://cran.r-project.org/web/packages/xtable/)
- ***xlsx***: funciones para la lectura y escritura de datos en archivos de Microsoft Office Excel (https://cran.r-project.org/web/packages/xlsx/)
- ***mvoutlier***: funciones para la detección de valores atípicos en datos multivariantes (https://cran.r-project.org/web/packages/mvoutlier/)

El redactado de la PEC se ha realizado con ***LaTex*** en el entorno web de ***Overleaf*** (https://www.overleaf.com/). *Overleaf* permite redactar documentos con *LaTex* en cualquier equipo con conexión a internet, sin necesidad de instalar una distribución de escritorio de *LaTex*.

2. El cojunto de datos: *census.dat*

census.dat es un conjunto de datos con 1.080 observaciones de 13 variables referentes a salarios, beneficios e impuestos pagados por empresas y particulares americanos durante 1995. Los datos de Census se obtuvieron en el año 2000 usando el sistema de extracción de datos del *U. S. Bureau of the Cesus*. El proceso de extracción dio lugar a un conjunto de 149.642 registros, sobre los que se aplicó el post-proceso siguiente:

1. Se descartaron 38 variables que tomaban valor nulo en más de 120.000 registros; permanecieron 16 variable
2. Se descartaron 3 variables que tomaban un rango pequeño de valores, y podían por tanto ser expresadas como variables categóricas; permanecieron 13 variables
3. Se descartaron los registros con valores nulos o 0 para 1 o más variables, permaneciendo 12.062 registros
4. Se mantuvo sólo 1 registro para cada valor de las variables FEDTAX, AFNLWGT, AGI, EMCONTTRB, PTOTVAL, TAXINC, STATETAX y POTHVAL (en este orden), de manera que las variables no tuviesen registros duplicados.

Las variables remanentes se describen a continuación:

- **AFNLWGT**: Peso final, con dos posiciones decimales implícitas
- **AGI**: Ingreso bruto ajustado
- **EMCONTRB**: Contribución del empresario al seguro de salud

- **FEDTAX**: Responsabilidad de impuestos federales sobre la renta
- **FICA**: Deducción de la nómina para la jubilación
- **INTVAL**: Cantidad de ingresos por intereses
- **PEARNVAL**: Ingresos personales
- **POTHVAL**: Ingresos de otras personas
- **PTOTVAL**: Ingresos totales
- **STATETAX**: Responsabilidad de impuestos estatales sobre la renta
- **TAXINC**: Cantidad de renta imponible
- **WSALVAL**: Salario
- **ERNVAL**: Ganancias comerciales y agrícolas

El conjunto de datos CENSUS se encuentra públicamente disponible en el siguiente enlace: http://neon.vb.cbs.nl/casc/CASCrefmicrodata.zip

3. Descripción y representación gráfica del conjunto de datos

3.1. Medidas de centralidad y variabilidad

En este apartado se presenta el análisis los estadísticos univariantes de las variables del archivo *census.dat*. Se pretende, mediante este paso, valorar las principales características de la distribución individual de cada una de las variables, a efectos de localizar medidas de forma o valores atípicos que convenga tener en cuenta en el análisis multivariante posterior. Para realizar el análisis se ha utilizado el paquete de funciones para estadística descriptiva *psych*, mientras que la exportación de los datos a LaTex se ha realizado mediante el paquete xTable (ver sección 1). Todo el código *R* utilizado se encuentra ordenado en el Anexo 1.

El Cuadro 1 presenta las medidas de distribución univariante de las 13 variables del archivo *census.dat*.

Todas las variables del archivo son variables continuas, por lo que los valores medios indican la posición del centro de gravedad de los datos. Tanto los valores medios como las desviaciones típicas de los valores muestran el diferente orden de magnitud de las variables del conjunto de datos; convendrá tener en cuenta esta consideración en algunas etapas posteriores del análisis como, por ejemplo, aquellas que involucren el cálculo de distancias. En relación al coeficiente de asimetría, se observa una distribución notablemente asimétrica de las variables POTHVAL e INTVAL, a juzgar por el elevado valor de los indicadores. Asimismo, cabe pensar en la existencia de observaciones atípicas dentro de estas dos variables, dado el elevado valor de los coeficientes de kurtosis. Las anotaciones recogidas quedan corroboradas mediante la observación de los histogramas y diagramas de dispersión de POTHVAL (Ingresos de otras

	Var.	Mean	Standard Dev.	Skewness	Kurtosis	Variation
1	AFNLWGT	196039.81	101251.42	0.94	0.96	0.52
2	AGI	56222.76	24674.84	-0.16	-1.02	0.44
3	EMCONTRB	3173.14	1401.83	0.01	-0.63	0.44
4	FEDTAX	7544.66	4905.20	0.38	-0.76	0.65
5	PTOTVAL	45230.84	21323.47	0.42	-0.44	0.47
6	STATETAX	2597.18	1826.44	1.00	1.56	0.70
7	TAXINC	39712.95	21224.16	-0.09	-0.97	0.53
8	POTHVAL	5162.23	9449.64	4.22	26.53	1.83
9	INTVAL	1421.41	3750.89	6.87	61.91	2.64
10	PEARNVAL	40068.61	20816.01	0.32	-0.52	0.52
11	FICA	2962.65	1427.23	-0.01	-0.74	0.48
12	WSALVAL	39523.38	20601.28	0.32	-0.50	0.52
13	ERNVAL	38444.56	20677.57	0.34	-0.53	0.54

Cuadro 1: Análisis descriptivo de *census.dat*

personas) e INTVAL (Cantidad de ingresos por intereses), recogidos en las Figuras 1 y 2, respectivamente. Las mismas Figuras para el resto de variables se encuentran recogidas en el Anexo 2 del documento. Todas ellas presentan una distribución más cercana a la normal, y sólo destaca la presencia de atípicos univariantes y una ligera asimetría en las variables AFNLWGT (Peso final) y STATETAX (Responsabilidad de impuestos estatales sobre la renta).

La presencia de valores atípicos pueden distorsionar las medias y desviaciones tipo de las variables POTHVAL e INTVAL, lo que podría explicar el elevado valor de sus coeficientes de variación. En este sentido, conviene analizar algunas medidas robustas de centralidad y variabilidad. El Cuadro 2 presenta las medidas de distribución robustas de las variables de *census.dat*.

Las relaciones de valores entre la media y la mediana (Mean/Med.) y entre la desviación tipo y la desviación absoluta mediana (SD/MAD) refleja, de nuevo, la presencia de observaciones atípicas en las variables POTHVAL e INTVAL.

En el tercer apartado de la práctica se realizará un estudio de valores atípicos mediante 3 técnicas multivariantes.

	Var.	Median	Mean/Med.	MAD	SD/Mad	MAD/Med.
1	AFNLWGT	180349.00	1.09	83267.26	1.22	0.46
2	AGI	58412.50	0.96	29029.31	0.85	0.50
3	EMCONTRB	3215.50	0.99	1564.88	0.90	0.49
4	FEDTAX	7068.00	1.07	5669.46	0.87	0.80
5	PTOTVAL	43278.00	1.05	23348.73	0.91	0.54
6	STATETAX	2322.00	1.12	1829.53	1.00	0.79
7	TAXINC	41155.00	0.96	24903.97	0.85	0.61
8	POTHVAL	1586.50	3.25	2095.66	4.51	1.32
9	INTVAL	353.00	4.03	466.28	8.04	1.32
10	PEARNVAL	39000.00	1.03	22239.00	0.94	0.57
11	FICA	3002.00	0.99	1786.53	0.80	0.60
12	WSALVAL	38000.00	1.04	22239.00	0.93	0.59
13	ERNVAL	36000.00	1.07	22239.00	0.93	0.62

Cuadro 2: Medidas robustas de *census.dat*

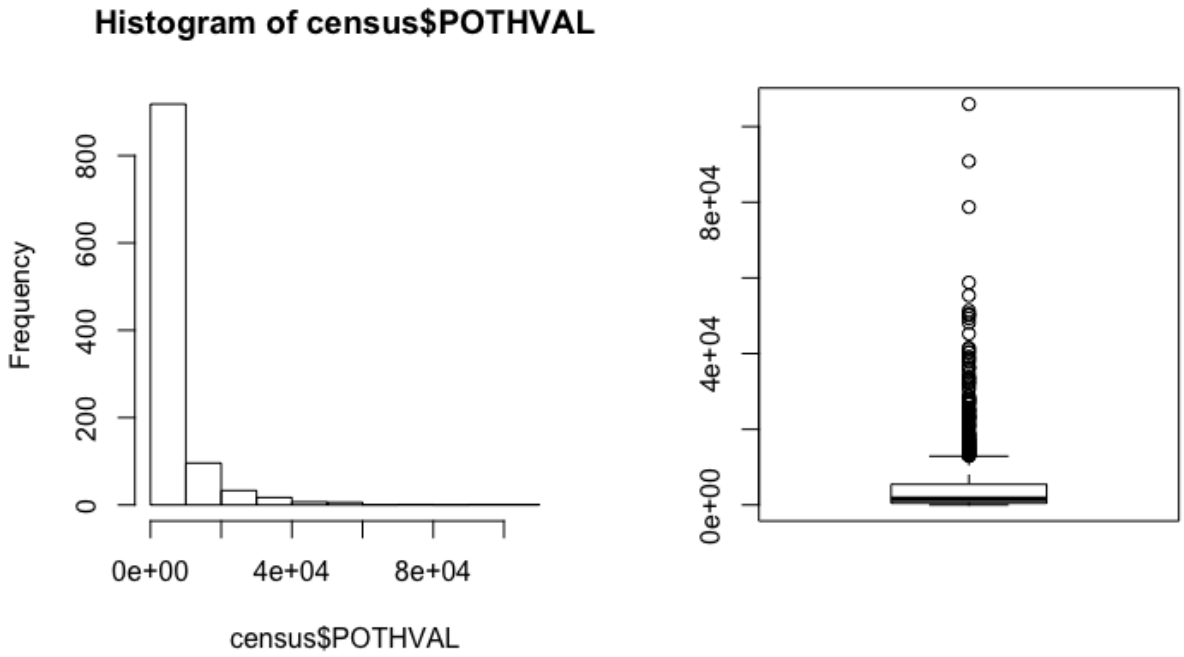


Figura 1: Histograma y diagrama de caja de la variable POTHVAL

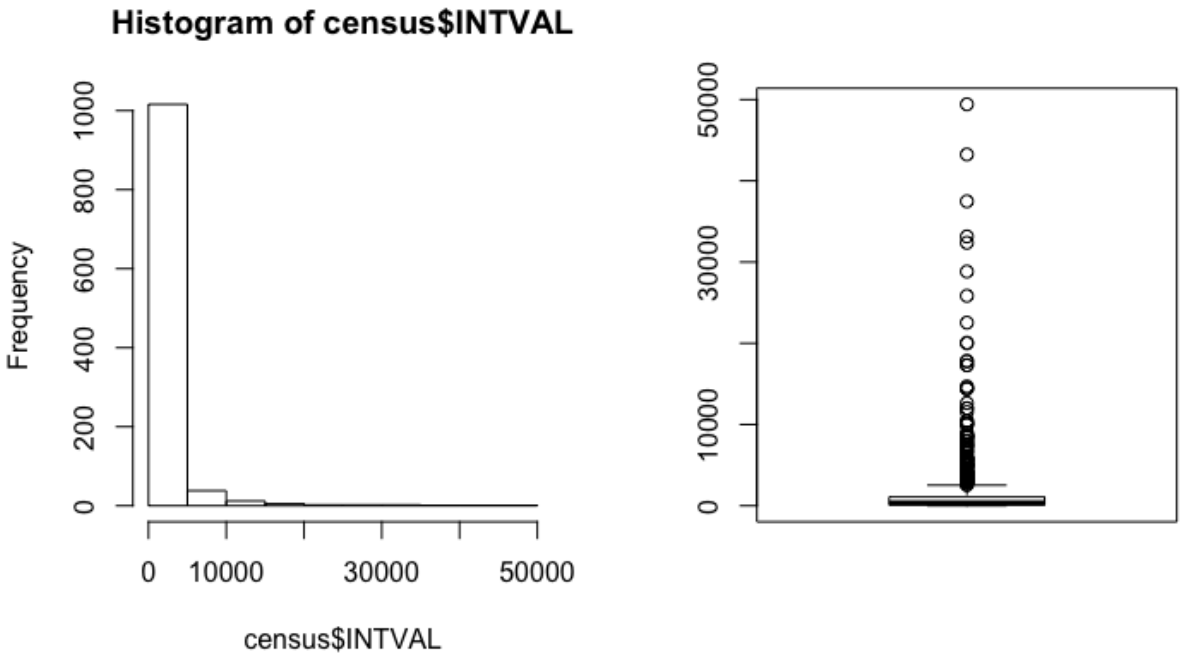


Figura 2: Histograma y diagrama de caja de la variable INTVAL

3.2. Análisis de colinealidad

Un conjunto de vectores de datos linealmente dependientes -en los que al menos un vector de datos puede escribirse como combinación lineal de otros- da lugar a una matriz singular (de determinante igual a 0) no invertible. A este fenómeno se le denomina colinealidad, y puede expresarse en dos formas:

- Cuando una variable se puede escribir, exactamente, como combinación lineal de una o más variables de la matriz de datos, la colinealidad es exacta, y el determinante de la matriz se anula, impidiendo la aplicación de algoritmos que requieran trabajar con una matriz no singular
- Alternativamente, cuando una variable se aproxima a la combinación lineal de una o más variables de la matriz de datos, la colinealidad es aproximada. En este caso el determinante se acercará a 0, dando lugar a problemas de precisión al invertir la matriz de datos por tener que dividir entre un número muy pequeño.

El siguiente apartado analiza la existencia de colinealidad entre las variables, a efectos de definir la matriz de datos reducida que, expresada como un subconjunto de la matriz de datos original que prescinde de las variables que no aportan variabilidad, permita aplicar aquellos algoritmos que requieren trabajar con matrices no singulares. La existencia de dependencias lineales entre variables de la matriz de datos original se puede analizar mediante la diagonalización de la matriz de covarianzas asociada como sigue:

1. Se calcula la matriz de covarianzas asociada a la matriz de datos original
2. Se diagonaliza la matriz de covarianzas, obteniendo los valores y vectores propios asociados
3. Se analiza el vector de autovalores: la varianza de cada componente principal es un autovalor, de manera que el número de autovalores nulos indica el número de variables que son combinación de otras. El redondeo realizado por los programas de cálculo informático puede ocultar la presencia de autovalores nulos, por lo que conviene calcular métricas adicionales que permitan definir cuando una variable no aporta variablidad adicional al problema. Una realización de este planteamiento es la presentada por Besley (Besley, 1991), que propone estimar cuando un autovalor se acerca a 0 a partir de su valor relativo con respecto al mayor. En este sentido, se define el índice de condición de una variable como la raíz cuadrada del cociente entre el valor del mayor de los autovalores y dicho autovalor, de manera que índices de condición elevados (superiores a 100) expresan una colinealidad fuerte, por relacionarse con autovalores casi-nulos.
4. Una vez determinado el número de autovalores nulos, el proceso sigue analizando la presencia de las distintas variables dentro de la relación lineal. A este efecto, se analizan los valores del autovector asociado a dicha componente principal, dividiendo los valores del autovector asociado a la componente principal nula o casi-nula entre el valor del mayor de los componentes de dicho autovector.

La aplicación mediante R de este procedimiento a la matriz de datos del archivo *census.dat* se recoge en el Anexo 1. El listado de autovalores generado y sus correspondientes índices de condición se recogen en el Cuadro 3:

	Eigenvalues	Condition
1	10254230277.01	1.00
2	2397589507.49	4.28
3	345734702.94	29.66
4	134307709.07	76.35
5	21603564.26	474.65
6	12138618.28	844.76
7	9190395.13	1115.76
8	7945312.96	1290.60
9	1399810.97	7325.44
10	1028967.02	9965.56
11	602007.13	17033.40
12	126498.72	81061.93
13	0.00	4000370927997654 40.00

Cuadro 3: Autovalores de la matriz de covarianzas

Y el valor propio asociado al autovector nulo queda representado, junto a los pesos de las variables en la combinación lineal, en el Cuadro 4:

	Eigenvector	Weight
1	0.00	0.00
2	0.00	0.00
3	0.00	0.00
4	0.00	0.00
5	-0.58	-1.00
6	-0.00	-0.00
7	-0.00	-0.00
8	0.58	1.00
9	-0.00	-0.00
10	0.58	1.00
11	-0.00	-0.00
12	-0.00	-0.00
13	-0.00	-0.00

Cuadro 4: Autovector asociado al autovalor nulo y peso relativo de cada una de las variables

Se detecta la existencia de una colinealiad muy fuerte entre 3 de las variables del conjunto de datos, que podría indicar una colinealidad exacta: la variable PTOTVAL (Ingresos totales) se puede expresar como la suma de las variables PEARNVAL (Ingresos personales) y POTHVAL (Ingresos de otras personas). Esta relación se puwsw comprobar mediante el siguiente procedimiento:

```
census$linearity_sneak =
                census$PTOTVAL - census$PEARNVAL - census$POTHVAL
hist(census$linearity_sneak)
```

Que devuelve como resultado un histograma nulo, es decir, que la variable calculada es cero para todos los registros de *census.dat*

3.3. Dependencias lineales por pares

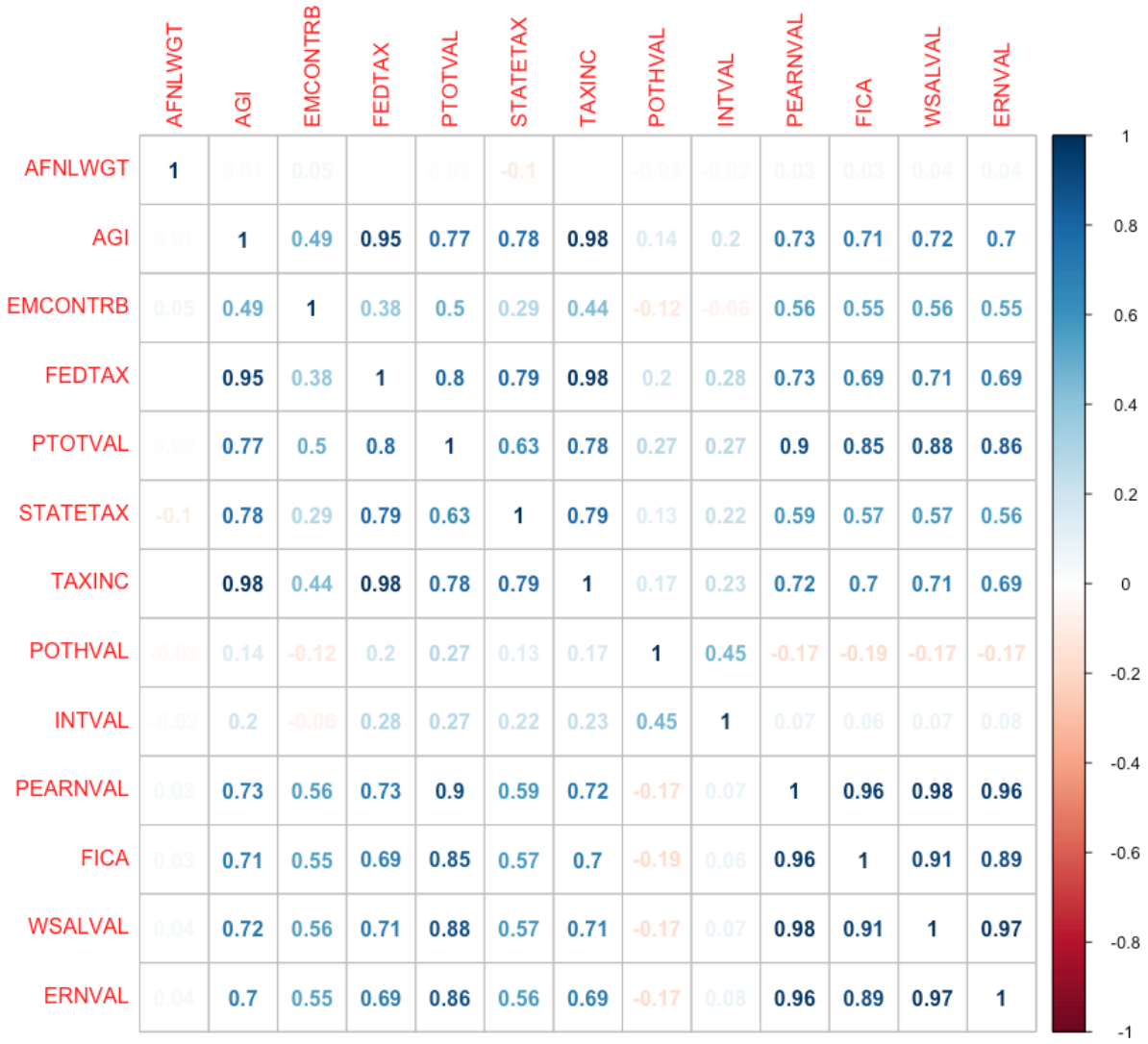
En este apartado se analizará la dependencia lineal de las variables por pares. El problema se aborda, en primer lugar, analizando la matriz de correlación de las variables. El proceso de cálculo y representación de la misma a partir de los datos de *census.dat* se encuentra recogido en el Anexo 1. A tal efecto, se ha utilizado el paquete *corrplot*, que permite realizar algunas operaciones adicionales sobre la matriz de correlación. La Figura 3 recoge la representación básica de la matriz de correlación. En la figura 4 se han reordenado las variables, clusterizando aquellas con relaciones fuertes entre sí. La Figura 4 permite comprobar que existen algunos grupos de variables bien diferenciados. Por un lado, las variables AFNLWG (Peso Final), POTHVAL (Ingresos de otras personas) e INTVAL (Ingresos por intereses) muestran correlaciones muy débiles con el resto de variables del archivo (a excepción de la correlación entre POTHVAL e INTVAL, cercana a 0,5). Por otro lado, el resto de variables presentan relaciones lineales fuertes entre sí. Se nota, además, que es posible realizar varios grupos bien identificados: en primer lugar, las variables FEDTAX (Tasas federales), AGI (Ingreso bruto) y TAXINC (Cantidad de renta imponible); en segundo lugar, las variables PTOTVAL (Ingresos totales), FICA (Deducciones por jubilación), ERNVAL (Ganancias comerciales o agrícolas), PEARNVAL (Ganancias personales), WSALVAL (Salario); por último, las variables con menor relación con el resto dentro del grupo, EMCONTRB (Contribución del empresario al seguro de salud) y STATETAX (Tasas estatales).
Después de observar la matriz de correlación entre las variables, conviene recoger los diagramas de dispersión de las variables por pares (recogidos en la Figura 5). Esta representación nos permitirá comprobar la existencia de relaciones de tipo no lineal (relaciones logarítmicas, cuadráticas u otras), así como la presencia de valores atípicos en cada una de las proyecciones.

En relación a la matriz de gráficos de dispersión, caben las siguientes consideraciones:

- Las variables AGI, FEDTAX y TAXINC presenta una relacion lineal cercana a 1 en una recta de pendiente 1, lo que indica que, salvo algunos ejemplos puntuales, representan la misma observación. La relación de estas variables con otras variables como PTOTVAL, PEARNVAL, ERNVAL o WSALVAL (el segundo grupo observado en la matriz de correlación) parece una relación heterocedástica (la variabilidad de la segunda aumenta al aumentar el valor de la variable).

- Las variables PTOTVAL, FICA, ERNVAL, PEARNVAL y WSALVAL presentan relaciones lineales fuertes entre sí, aunque menos fuertes que las anteriores (cercanas a 0,8).

- Algunas proyecciones, como la FEDTAX-STATETAX permiten visualizar valores atípicos, que pueden distorsionar el coeficiente de correlación.



	AFNLWGT	AGI	EMCONTRB	FEDTAX	PTOTVAL	STATETAX	TAXINC	POTHVAL	INTVAL	PEARNVAL	FICA	WSALVAL	ERNVAL
AFNLWGT	1	0.01	0.05	0	0.03	-0.1	0	-0.03	-0.02	0.03	0.03	0.04	0.04
AGI	0.01	1	0.49	0.95	0.77	0.78	0.98	0.14	0.2	0.73	0.71	0.72	0.7
EMCONTRB	0.05	0.49	1	0.38	0.5	0.29	0.44	-0.12	-0.06	0.56	0.55	0.56	0.55
FEDTAX	0	0.95	0.38	1	0.8	0.79	0.98	0.2	0.28	0.73	0.69	0.71	0.69
PTOTVAL	0.03	0.77	0.5	0.8	1	0.63	0.78	0.27	0.27	0.9	0.85	0.88	0.86
STATETAX	-0.1	0.78	0.29	0.79	0.63	1	0.79	0.13	0.22	0.59	0.57	0.57	0.56
TAXINC	0	0.98	0.44	0.98	0.78	0.79	1	0.17	0.23	0.72	0.7	0.71	0.69
POTHVAL	-0.03	0.14	-0.12	0.2	0.27	0.13	0.17	1	0.45	-0.17	-0.19	-0.17	-0.17
INTVAL	-0.02	0.2	-0.06	0.28	0.27	0.22	0.23	0.45	1	0.07	0.06	0.07	0.08
PEARNVAL	0.03	0.73	0.56	0.73	0.9	0.59	0.72	-0.17	0.07	1	0.96	0.98	0.96
FICA	0.03	0.71	0.55	0.69	0.85	0.57	0.7	-0.19	0.06	0.96	1	0.91	0.89
WSALVAL	0.04	0.72	0.56	0.71	0.88	0.57	0.71	-0.17	0.07	0.98	0.91	1	0.97
ERNVAL	0.04	0.7	0.55	0.69	0.86	0.56	0.69	-0.17	0.08	0.96	0.89	0.97	1

Figura 3: Matriz de correlación de las variables de *census.dat*

- Adicionalmente, se observan algunas relaciones que se alejan de la linealidad en valores altos, como la relación FICA-WSALVAL o FICA-ERNVAL. Cabe pensar que la representación de los diagramas de dispersión en escala logarítmica podría hacer aparecer correlaciones lineales muy elevadas entre los valores. En este sentifo, la Figura 6 recoge la Matriz de gráficos de dispersión en escala logarítmica de las variables de *census.dat*. Como se suponía, la relación entre FICA-WSALVAL y FICA-ERNVAL pasa a situarse ahora sobre una recta, y la correlación crece 2pp respecto a la de las variables originales. Adicionalmente, las relaciones heterocedásticas mencionadas antes se suavizan (ver WSALVAL-AGI, por ejemplo)

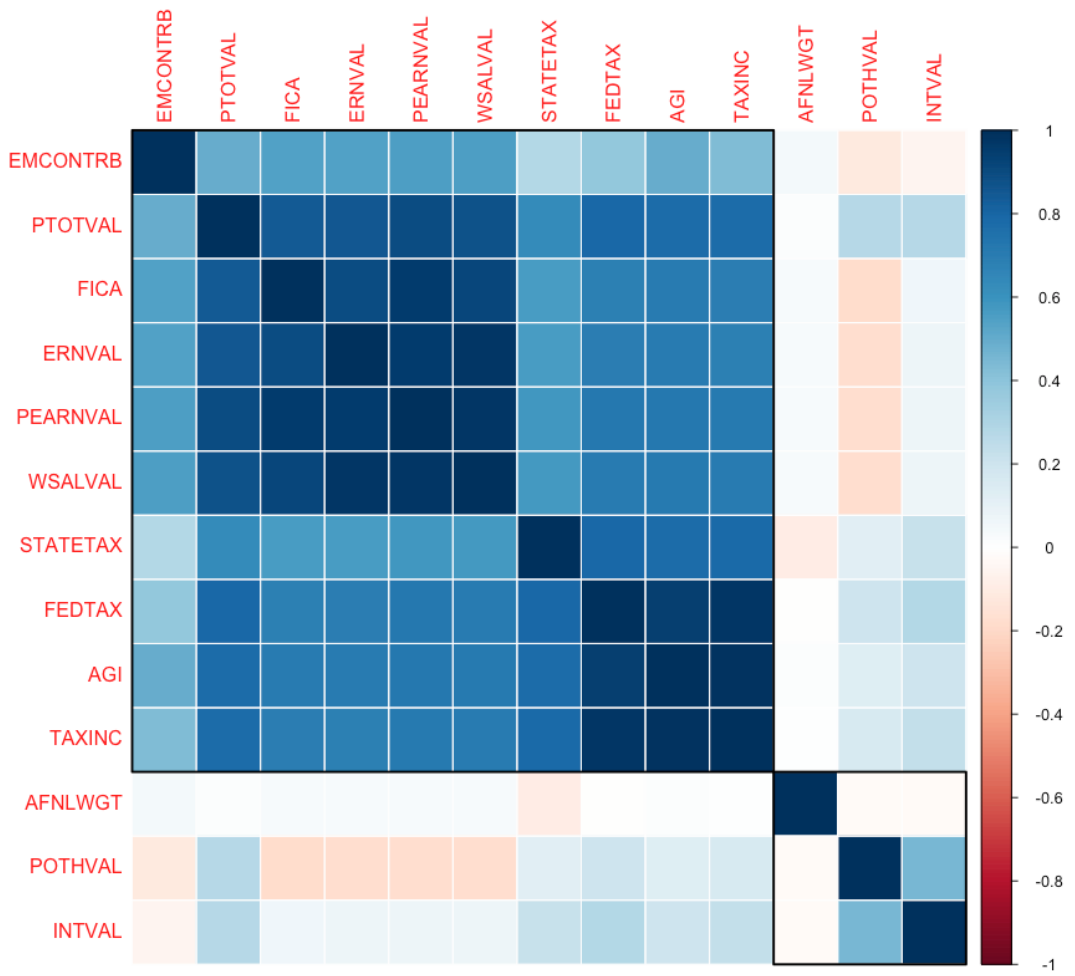


Figura 4: Agrupaciones de variables dentro de *census.dat*

3.4. Medidas globales de variabilidad

En el Cuadro 5 se recogen las medidas globales de variabilidad de *census.dat*, calculadas a partir de su matriz de covarianzas. El fragmento de código utilizado para el cálculo se encuentra en el Anexo 1. Notar que para el cálculo de las medidas de variabilidad se ha descartado la variable PTOTVAL, evitando así los problemas de precisión y singularidad indicados en la Sección 3.2. Como existe bastante correlación entre las variables estas medidas son mucho menores de los promedios de las varianzas o desviaciones tipo.

	totalVariance	avgVariance	effVariance	effStdDev
1	12731207011.46	1060933917.62	17453863.97	4177.78

Cuadro 5: Medidas de variabilidad global

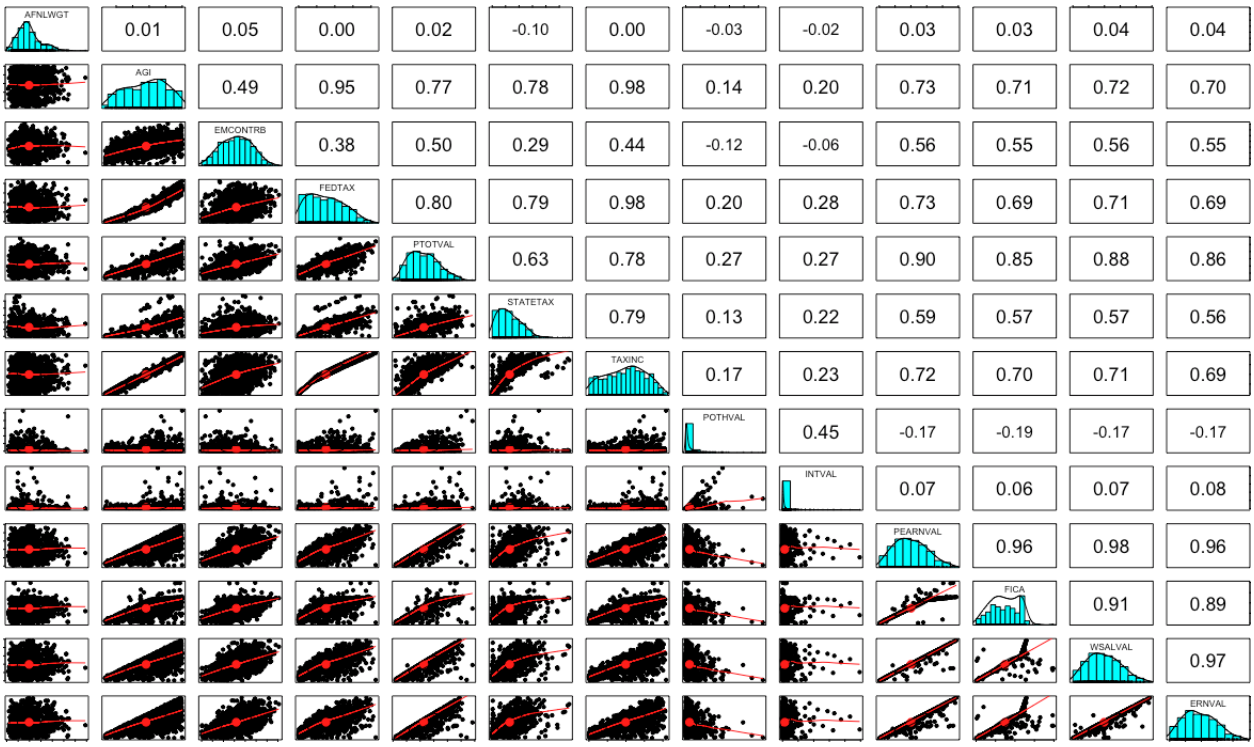


Figura 5: Matriz de gráficos de dispersión

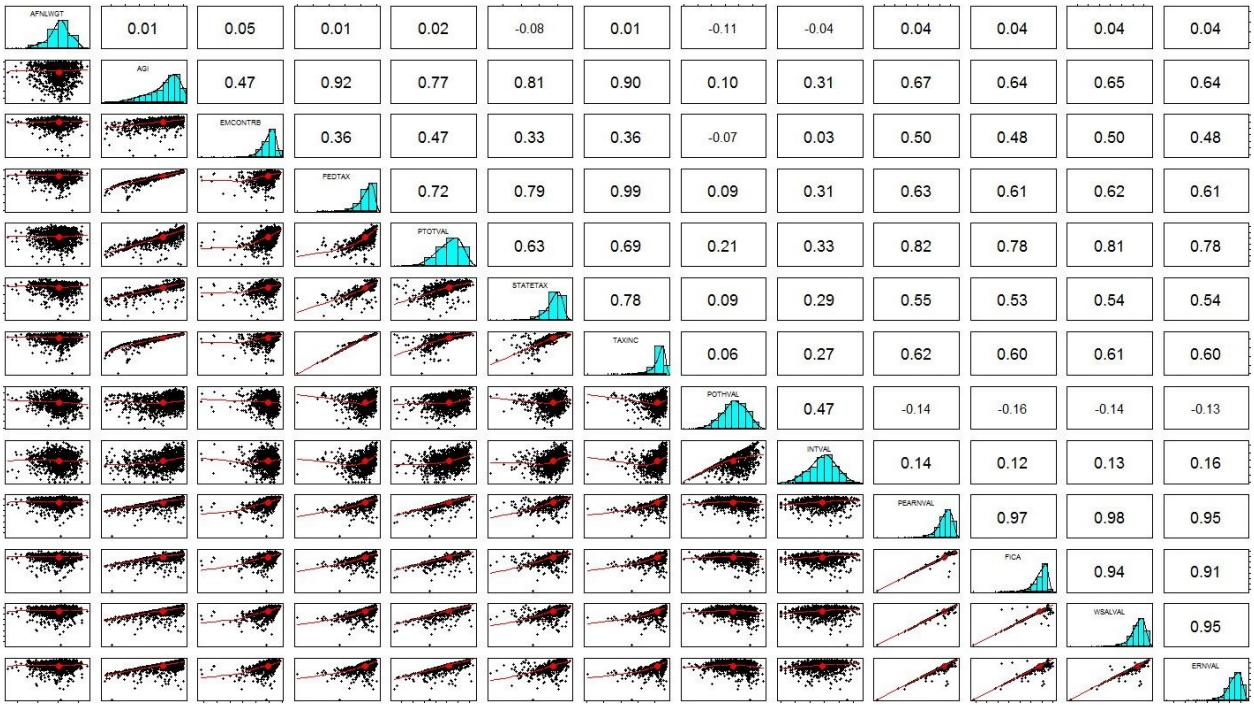


Figura 6: Matriz de gráficos de dispersión en escala logarítmica

4. Datos atípicos

Se analizan a continuación dos algoritmos para la detección de valores atípicos multivariantes dentro del conjunto *census.dat*.

4.1. Local Outlier Factor (LOF)

El algoritmo LOF, propuesto por Marus M.Breunig (Breunig et al, 2000), realiza detección de valores atípicos midiendo la concentración o dispersión de puntos alrededor de un punto y comparándola con la sus vecinos. La base del algoritmo es el concepto de densidad local, dónde la localidad viene marcada por los k puntos más próximos, y k es una constante que interviene como argumento de la función. El algoritmo permite identificar regiones de densidad similar, y dentro de éstas puntos con una densidad substancialmente inferior que sus vecinos. Dichos puntos se considerarán puntos atípicos. La densidad local de un punto se calcula como

$$LOF(p) = \frac{\sum \frac{lrd(O)}{lrd(p)}}{abs(Nmin)}$$

Dónde *lrd* es la densidad local de accesibilidad, calculada como la distancia media entre P y los objetos de su k-vecindad. Para la implementación del algoritmo LOF en R se ha utilizado el paquete *DMwR* (https://cran.r-project.org/web/packages/DMwR/). A efectos de evitar que las diferentes magnitudes de las variables afecte al cálculo de distancias dentro del algoritmo, se ha realizado un escalado lineal del espacio de variables, que consiste en transformar el valor x en $fracx - MINMAX - MIN$. El código de la implementación de LOF en R se encuentra en el Anexo 1. Las Figuras 7 y 8 representan los resultados de la ejecución del algoritmo: en primer lugar, la distribución de densidad local de los 1.080 puntos de la matriz de datos; en segundo lugar, la representación sobre la matriz de gráficos de dispersión de los valores considerados como atípicos (los 10 con menor densidad local). La principal ventaja de LOF es la posibilidad de localizar datos atípicos en un área del conjunto de datos que no lo serían en otra área del conjunto: por ejemplo, un punto a una distancia pequeña de un grupo muy denso será un atípico, mientras que un punto más distante de una nube dispersa. Por contra, el principal inconveniente del algoritmo es la dificultad para interpretar los resultados del cociente, dado que no hay una regla clara para determinar cuándo un punto es un dato atípico. Por contra, la definición de un valor como atípico depende del problema concreto y del valor de la frontera que el usuario defina.

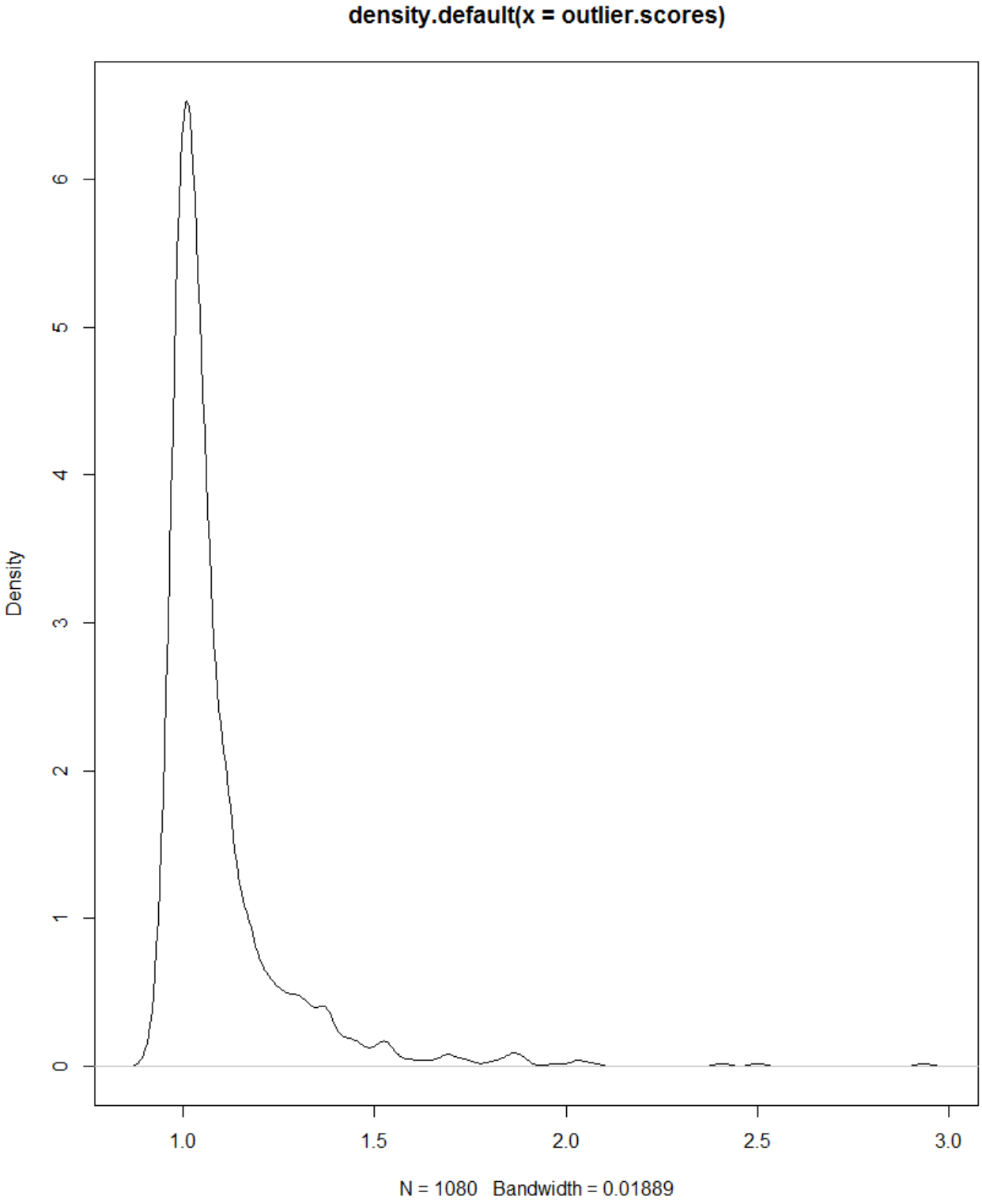


Figura 7: Gráfico de densidad local de los puntos de *census.dat*

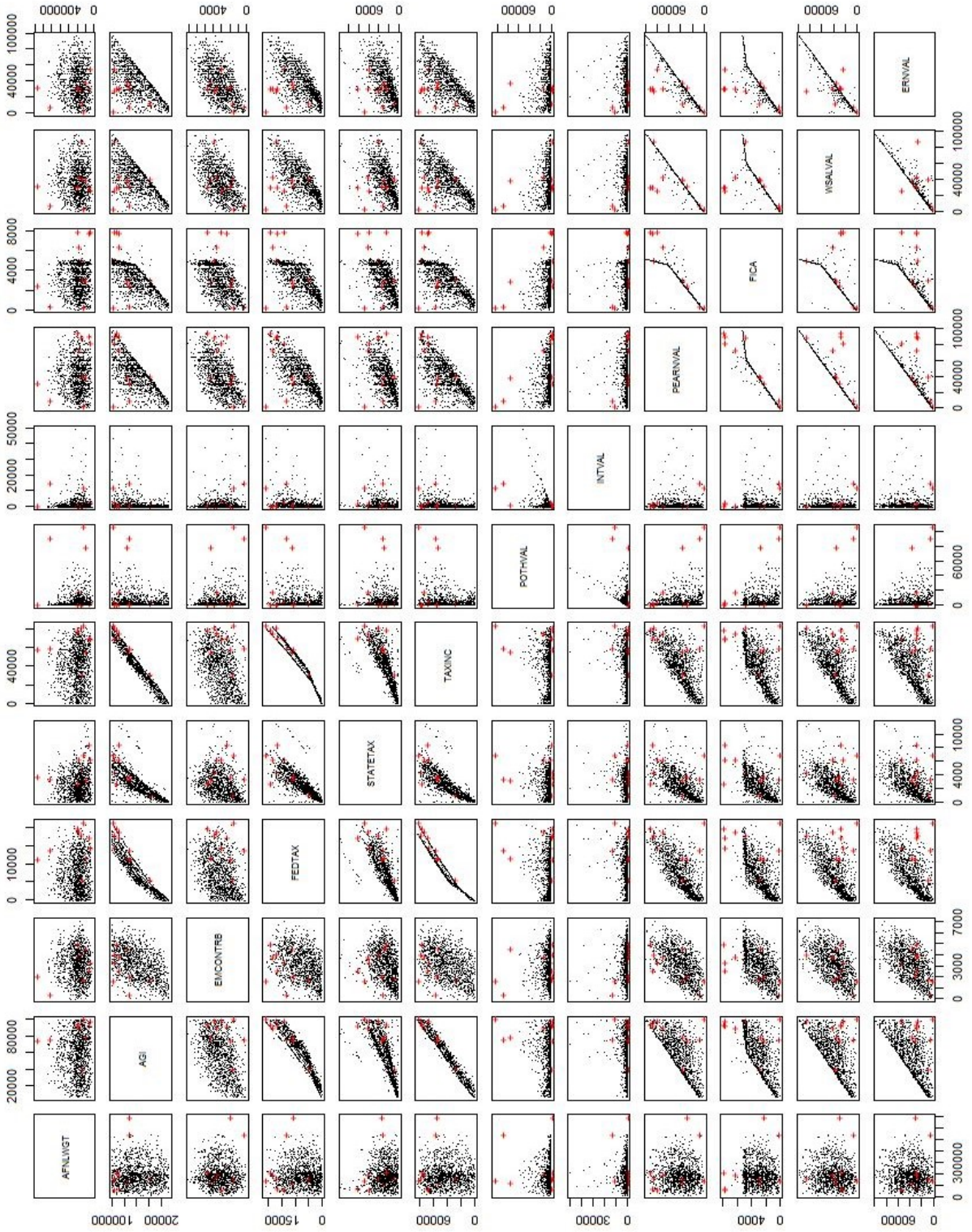


Figura 8: Representación de los 10 puntos con menor densidad local sobre la matriz de gráficos de dispersión

4.2. Distancia de Mahalanobis

Otra posibilidad para localizar grupos de datos atípicos es la utilización de la distancia de cada punto al centro geométrico de los datos, de manera que los puntos que más se alejen del mismo podrán etiquetarse como atípicos. En este sentido, es posible utilizar una métrica normalizada como la distancia de Mahalanobis, que tiene en cuenta la ponderación por correlación entre las variables y la diferencia de escala entre éstas. Para implementar el cálculo de distancias en R utilizaremos el paquete *chemometrics*, que incorpora los cálculos de la distancia de Mahaloanobis estándar y la distancia de Mahalanobis robusta. Como en las ocasiones anteriores, el Anexo 1 recoge el código utilizado en este punto. La Figura 9 recoge el histograma y el diagrama de caja de las distancias de Mahalanobis de los datos de *census.dat*. La aplicación del criterio de Mahalanobis con percentil 99 deja 86 valores identificados como atípicos dentro de la matriz de datos.

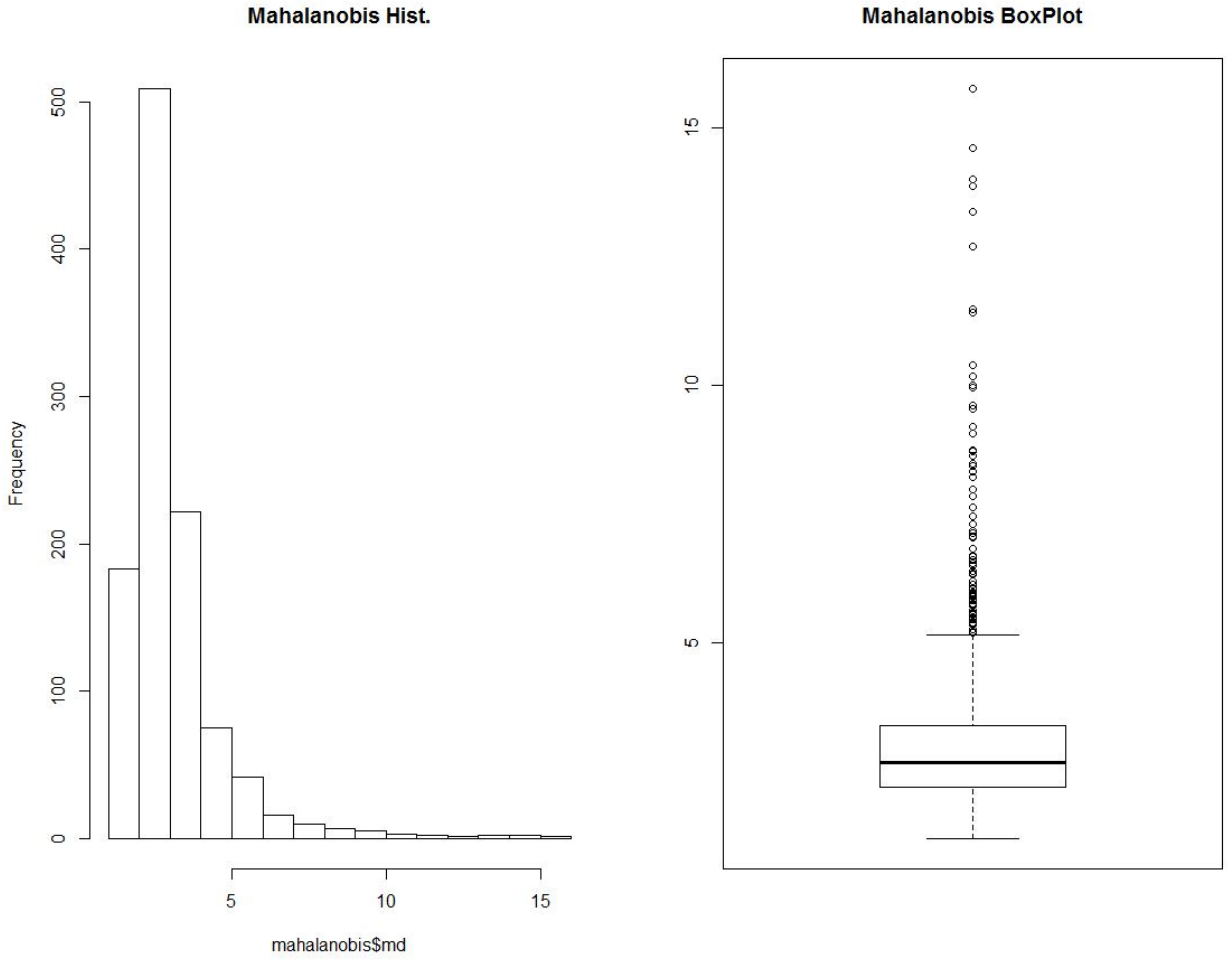


Figura 9: Histograma y diagrama de caja de las distancias de Mahalanobis

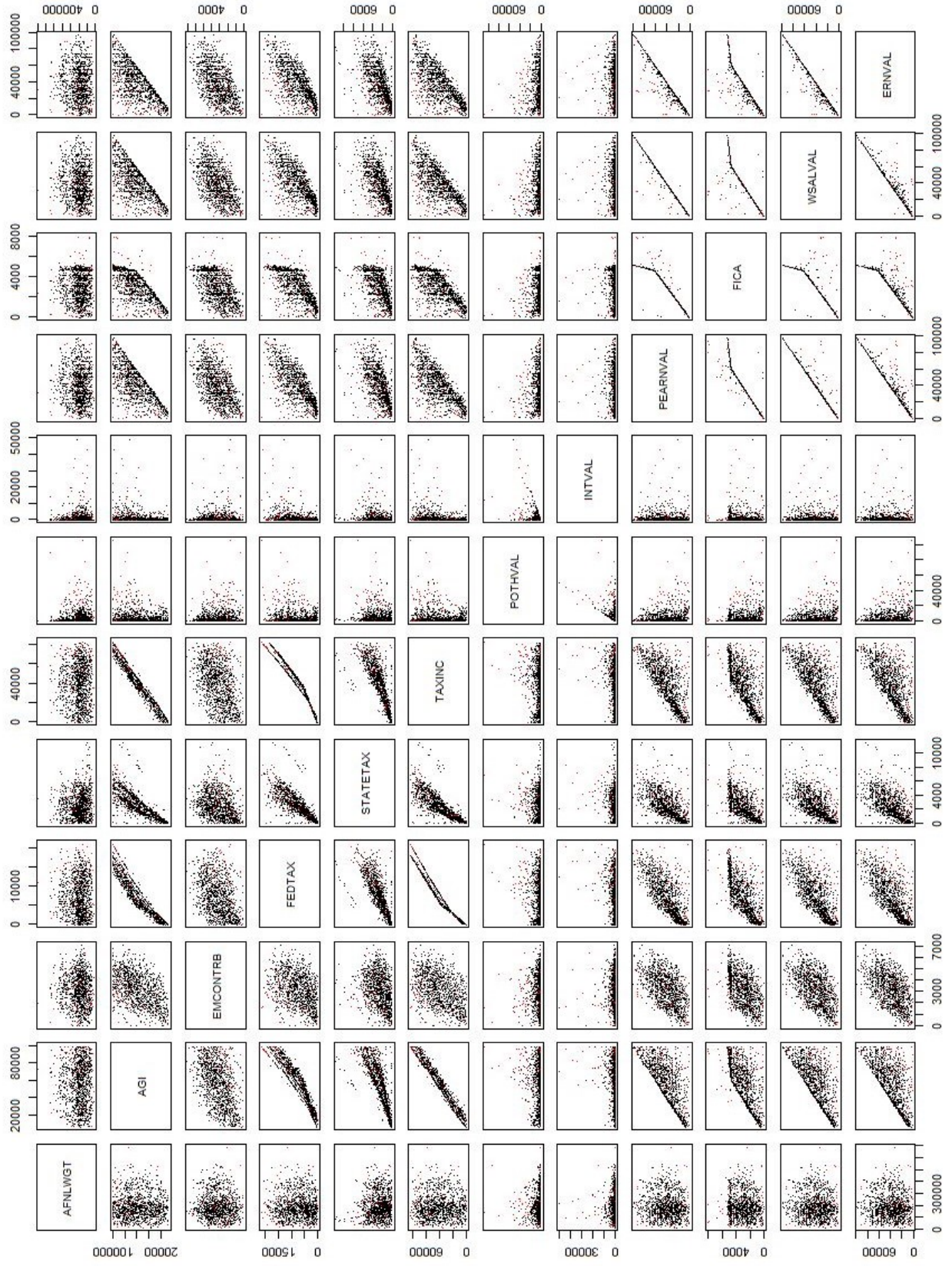


Figura 10: Representación de los datos atípicos para percentil 99 de Mahalanobis

5. Anexo 1: Resumen de código utilizado

Sección 3: Análisis de las distribuciones univariantes de los datos

```
setwd("./7. AnAlisis de datos multivariantes/")
census = read.csv(file="CASCrefmicrodata.csv")
stats = describe(census)
stats$variation = stats$sd / stats$mean
stats$m_m = stats$mad / stats$median
stats$centrality_dis = stats$mean / stats$median
stats$variability_dis = stats$sd / stats$mad
latexCentralityStats
                = data.frame(stats$vars, stats$mean, stats$sd,
                stats$skew, stats$kurtosis, stats$variation)
latexCentralityStats[,1] = colnames(census)
colnames(latexCentralityStats)
                = c("Var.", "Mean", "Standard Dev.", "Skewness",
                        "Kurtosis", "Variation")
<<results=tex>>
  xtable(latexCentralityStats)
write.xlsx(stats, file = "stats.xlsx")
latexRobustStats
                = data.frame(stats$vars, stats$median, stats$centrality_dis,
                stats$mad, stats$variability_dis, stats$m_m)
latexRobustStats[,1] = colnames(census)
colnames(latexRobustStats)
                = c("Var.", "Median", "Mean/Med.", "MAD",
                "SD/Mad", "MAD/Med.")
<<results=tex>>
  xtable(latexRobustStats)
write.xlsx(stats, file = "robust_stats.xlsx")
```

Sección 3: Localización de autovalores nulos

```
covariance = cov(census)
write.xlsx(covariance, file = "covariance.xlsx")
logCensus = log(census)
logCovariance = cov(logCensus)
write.xlsx(logCovariance, file = "logCovariance.xlsx")
eigen = eigen(covariance)
write.xlsx(eigen$values, file = "eigenvalues.xlsx")
besley = max(eigen$values) / eigen$values
write.xlsx(besley, file = "besley.xlsx")
weights = eigen$vectors[,13] / max(eigen$vectors[,13])
write.xlsx(weights, file = "weights.xlsx")
```

Sección 3: Cálculo de la matriz de correlación

```
correlations = cor(census)
par(mfrow=c(1,1))
corrplot(correlations, method = "number")
corrplot(correlations, order = "hclust", addrect = 2, method="number")
```

Sección 3: Cálculo de medidas globales de variabilidad

```
covariance = cov(census)
totalVariance = sum(diag(covariance))
avgVariance = totalVariance / length(diag(covariance))
genVariance = det(covariance)
effVariance = genVariance**(1/length(diag(covariance)))
effStdDev = sqrt(effVariance)
variances = data.frame()
variances$totalVariance = totalVariance
variances$avgVariance = avgVariance
variances$genVariance = genVariance
variances$effVariance = effVariance
variances$effStDev = effStdDev
```

Sección 4: Algoritmo LOF

```
library(DMwR)
outlier.scores = lofactor(LinearScaling(census), k = 5)
plot(density(outlier.scores))
outliers = order(outlier.scores, decreasing = TRUE)[1:10]
n <- nrow(census)
pch <- rep(".", n)
pch[outliers] <- "+"
col <- rep("black", n)
col[outliers] <- "red"
pairs(census, pch=pch, col=col)
```

Sección 4: Mahalanobis

```
library(chemometrics)
mahalanobis = Moutlier(census, quantile = 0.99, plot = TRUE)
summary(mahalanobis$md)
hist(mahalanobis$md, main = "Mahalanobis Hist.")
boxplot(mahalanobis$md, main = "Mahalanobis BoxPlot")
mahalanobis$ind = (mahalanobis$md > mahalanobis$cutoff)
census$outlier = mahalanobis$ind
summary(census$outlier)
```

6. Anexo 2: Histogramas y diagramas de caja para las variables no destacadas

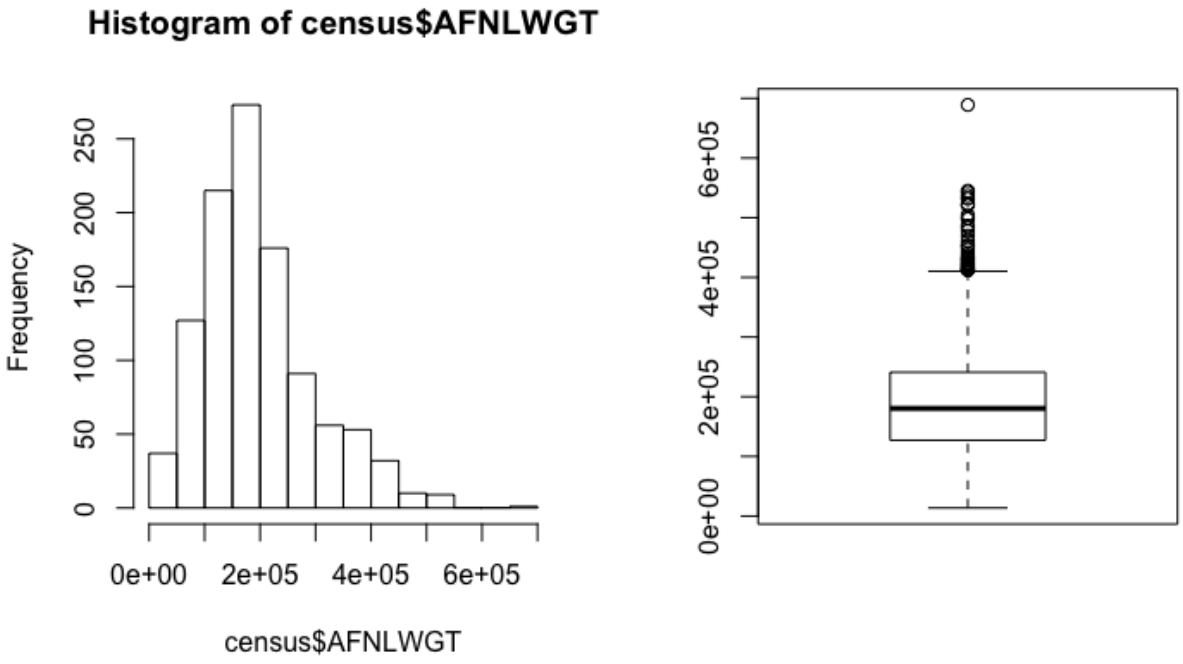


Figura 11: Histograma y diagrama de caja de la variable AFNLWGT

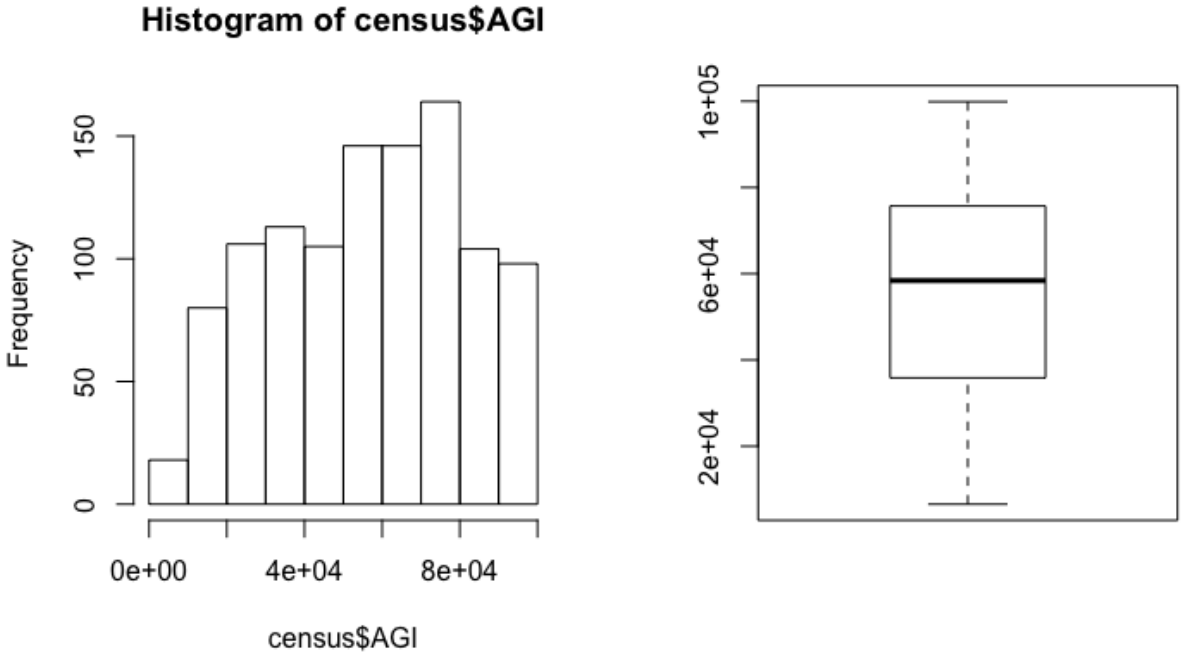


Figura 12: Histograma y diagrama de caja de la variable AGI

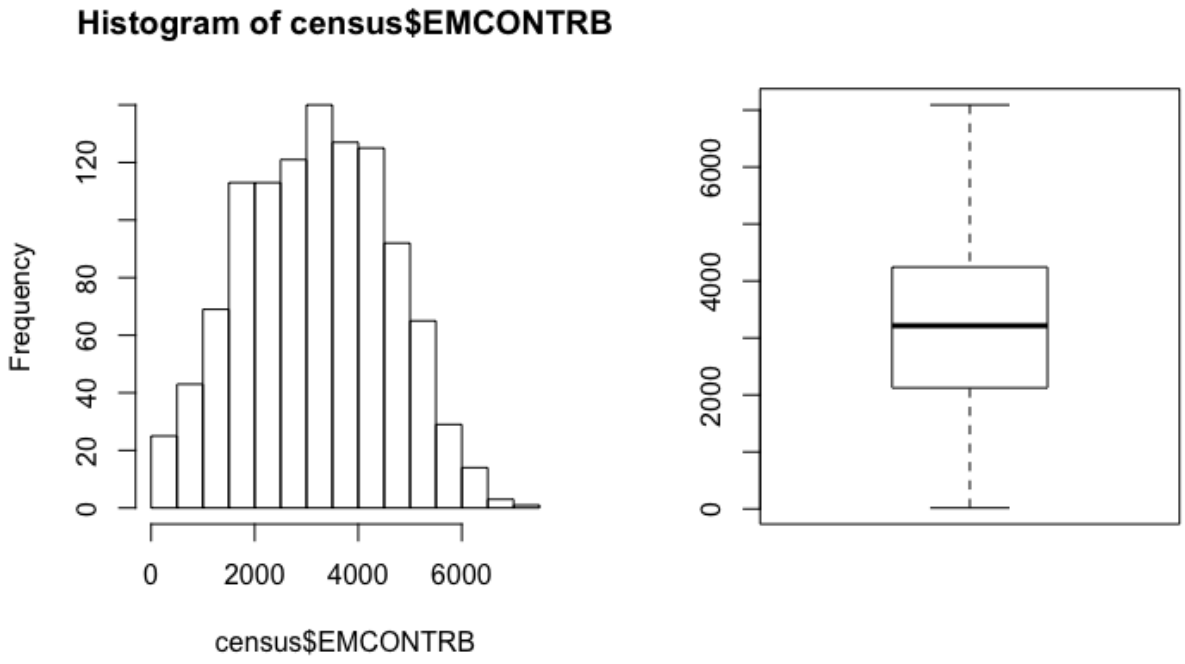


Figura 13: Histograma y diagrama de caja de la variable EMCONTRB

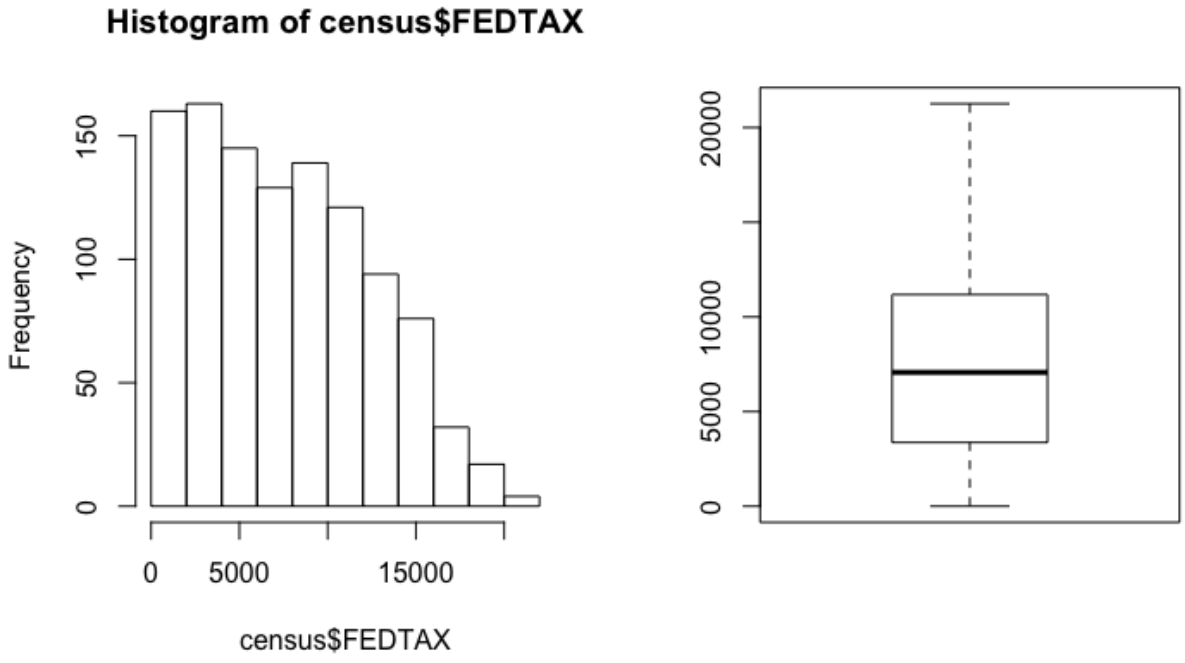


Figura 14: Histograma y diagrama de caja de la variable FEDTAX

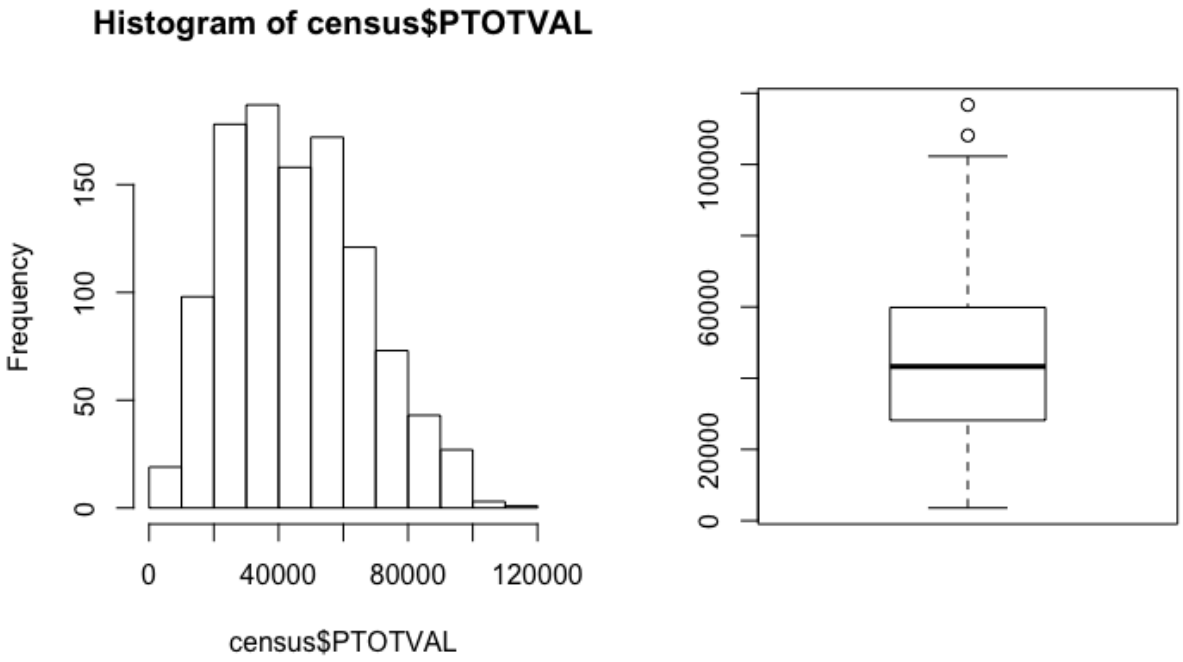


Figura 15: Histograma y diagrama de caja de la variable PTOTVAL

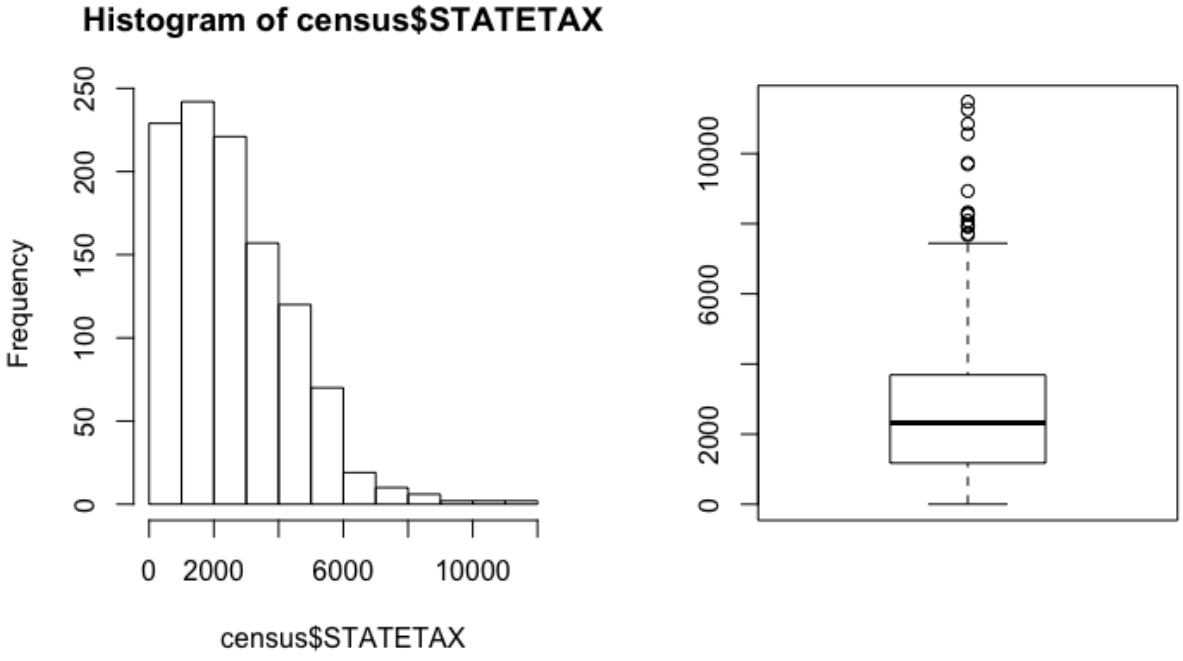


Figura 16: Histograma y diagrama de caja de la variable STATETAX

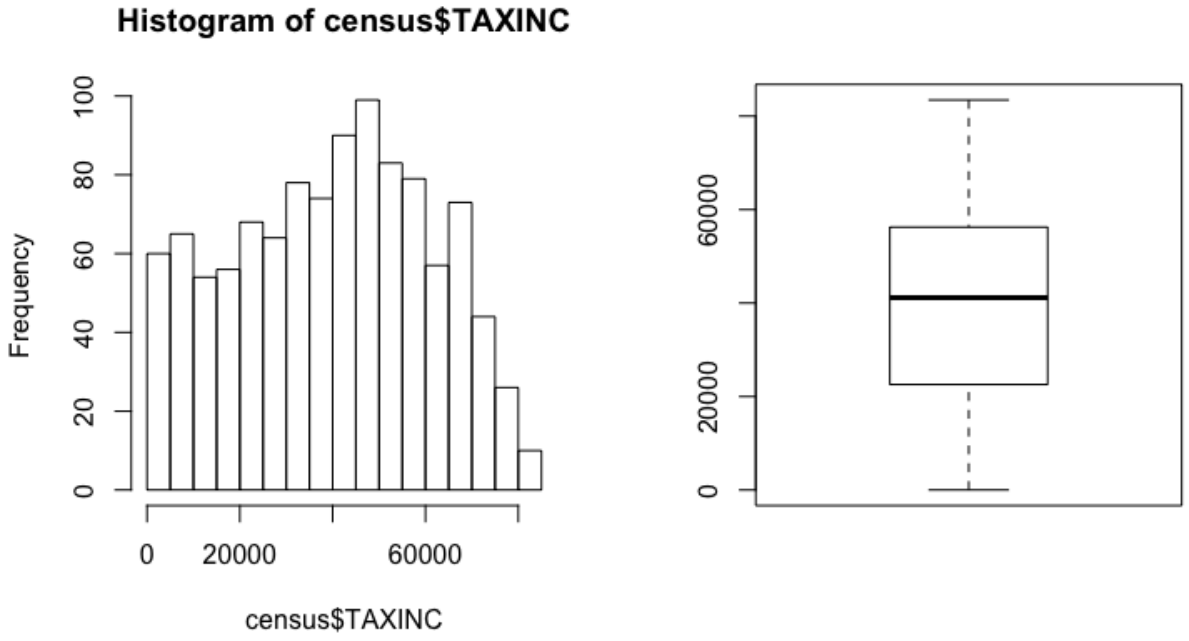


Figura 17: Histograma y diagrama de caja de la variable TAXINC

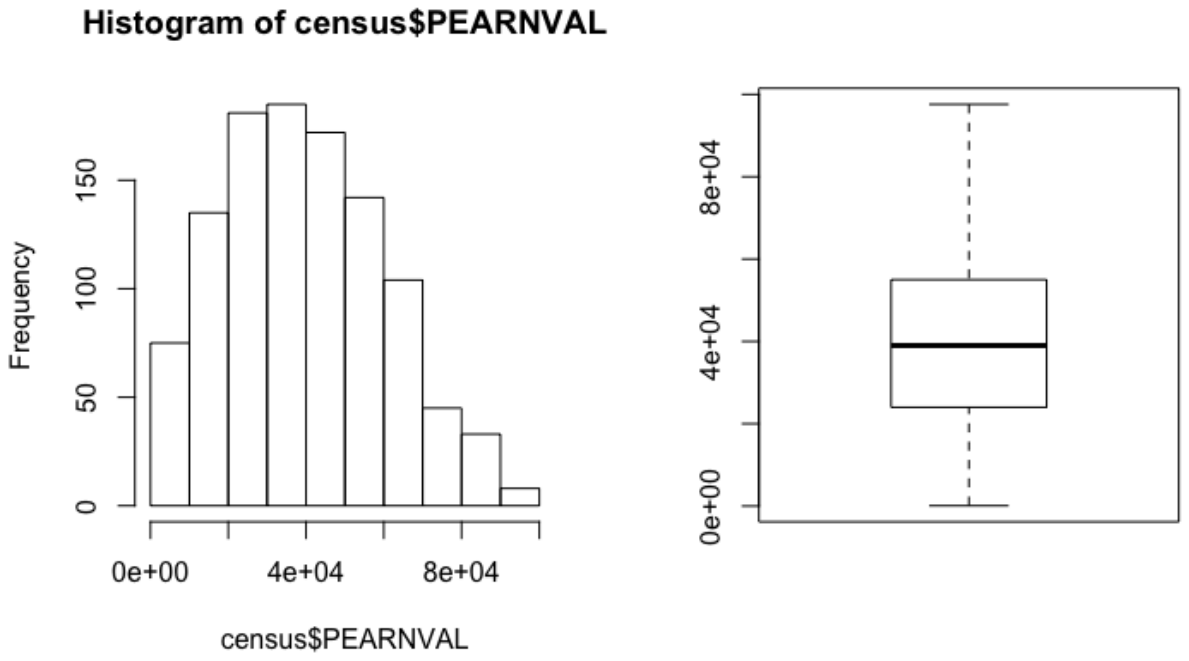


Figura 18: Histograma y diagrama de caja de la variable PEARNVAL

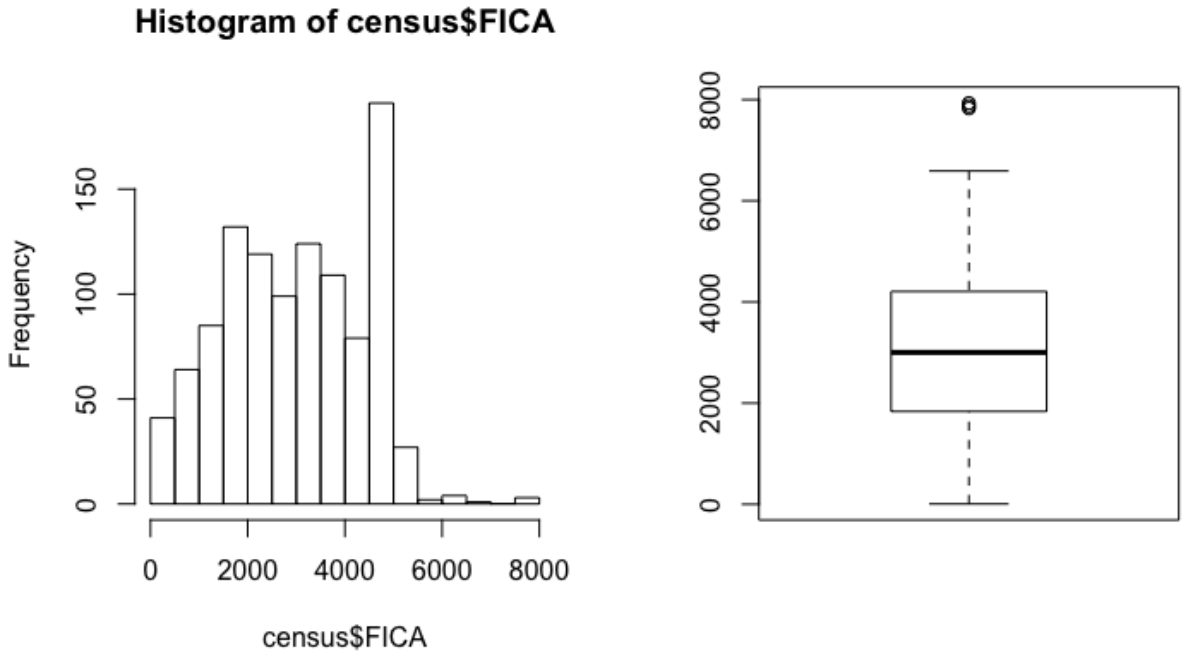


Figura 19: Histograma y diagrama de caja de la variable FICA

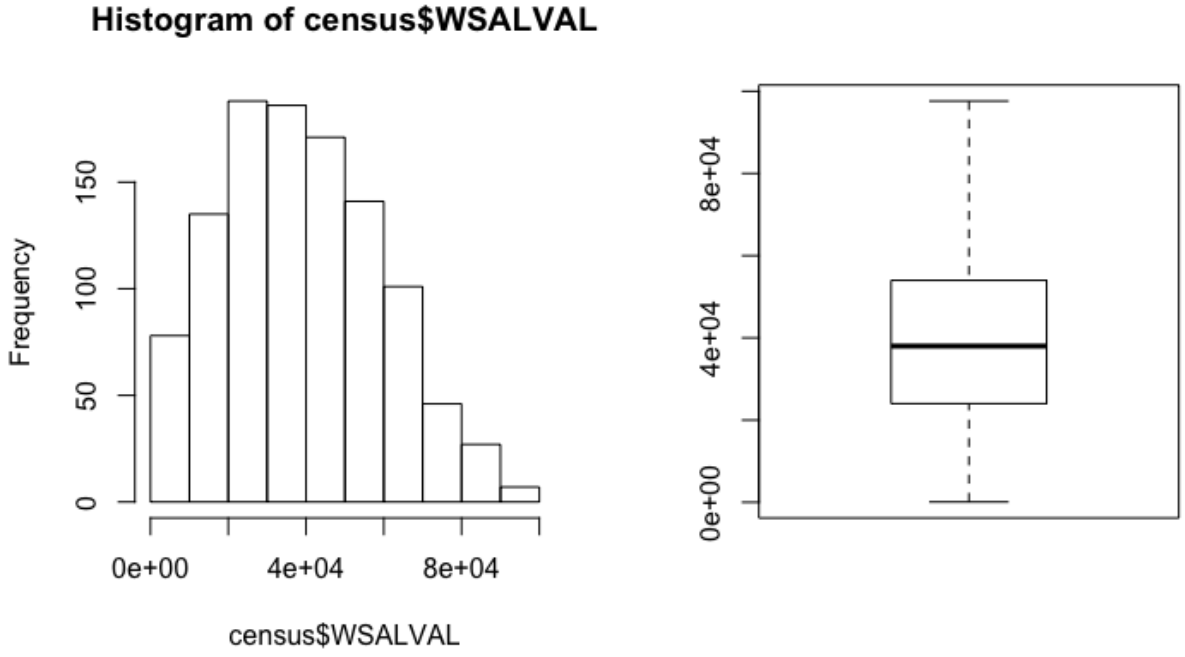


Figura 20: Histograma y diagrama de caja de la variable WSALVAL

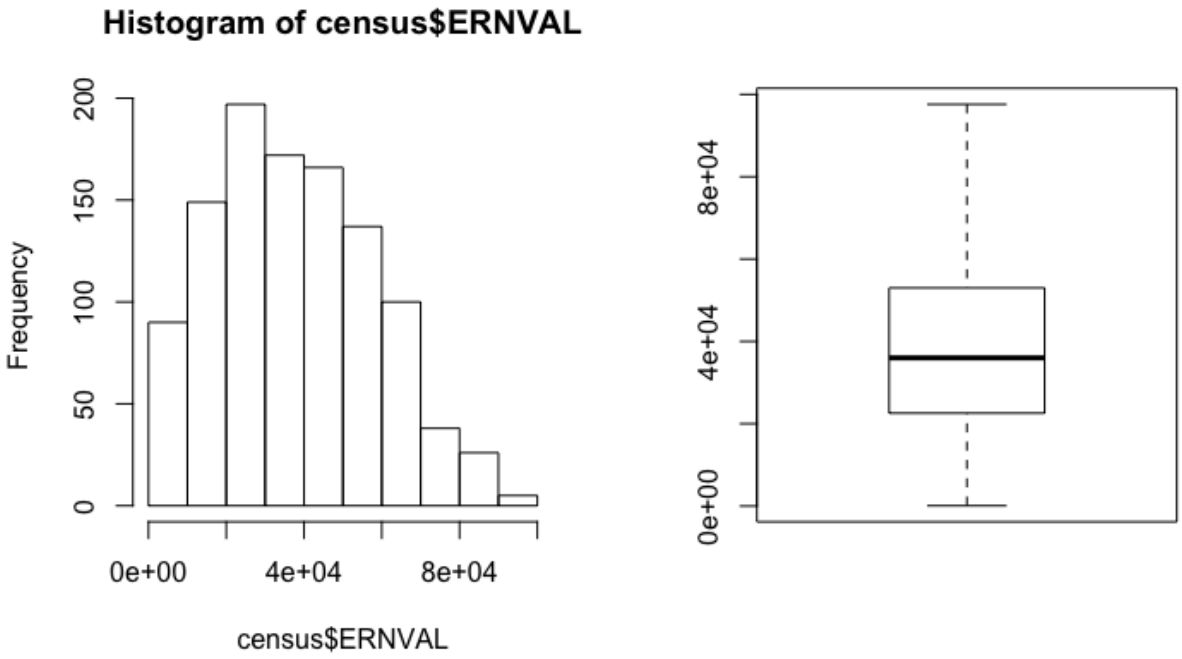


Figura 21: Histograma y diagrama de caja de la variable ERNVAL