Introducción al Análisis Exploratorio de Datos (EDA) en R

Módulo 2

2024-03-16

Herramientas para la manipulación de datos

Paquetes en R

Los paquetes en R son colecciones de funciones, datos y documentación cuyo objetivo es extender las capacidades básicas de R. **CRAN** (The Comprehensive R Archive Network) es una red de servidores que almacenan versiones de R, así como librerías en R que cumplen las políticas del repositorio (CRAN, 2022).

Para instalar paquetes del repositorio **CRAN**:

```
install.packages("dplyr")
```

Después de instalar el paquete, se debe cargar la librería:

```
library(dplyr)
```

Para encontrar la documentación del paquete:

```
help(dplyr)
```

Tidyverse

Tidyverse es un conjunto de librerías en R diseñadas para el análisis de datos (importar, transforma, visualizar y modelar con datos) (Wickham, 2019).

Nos concentraremos en las siguientes librerías:

- dplyr
- ggplot2
- forcats*

Figure 1: Librerías en Tidyverse

Importar datos

El primer paso es definir el directorio de trabajo:

```
setwd("path")
```

Nos concentraremos en funciones para importar los siguientes formatos de datos

Formato	Formato específico	Función	Paquete
Texto o tabulares	CSV	read_csv()	readr
	Otros formatos de texto	read_delim()	readr
Formatos de otros programas	Excel	read_excel()	readxl
	SPSS	read_sav()	haven
	STATA	read_dta()	haven
	SAS	read_sas()	haven
Formatos propios de R	.rda	load()	base
	.rds	readRDS()	base

Pipe (%>%)

La tubería de comando o *pipeline* (%>%) es una herramienta utilizada para el encadenamiento de funciones. El operador nos permite escribir una secuencia de operaciones

Una secuencia en su **forma estándar** sigue la forma

```
dataset_2 <- dplyr::filter(dataset, attend > 15 & attend != 20)
```

En **forma encadenada**:

```
dataset_2 <- dataset %>% dplyr::filter(attend > 15 & attend != 20)
```

El siguiente atajo es útil: 

Dplyr

El paquete **dplyr** proporciona una sintaxis para la manipulación de datos. (El operador %>% pertenece a la sintaxis de dplyr). Nos concentraremos en las siguientes funciones:

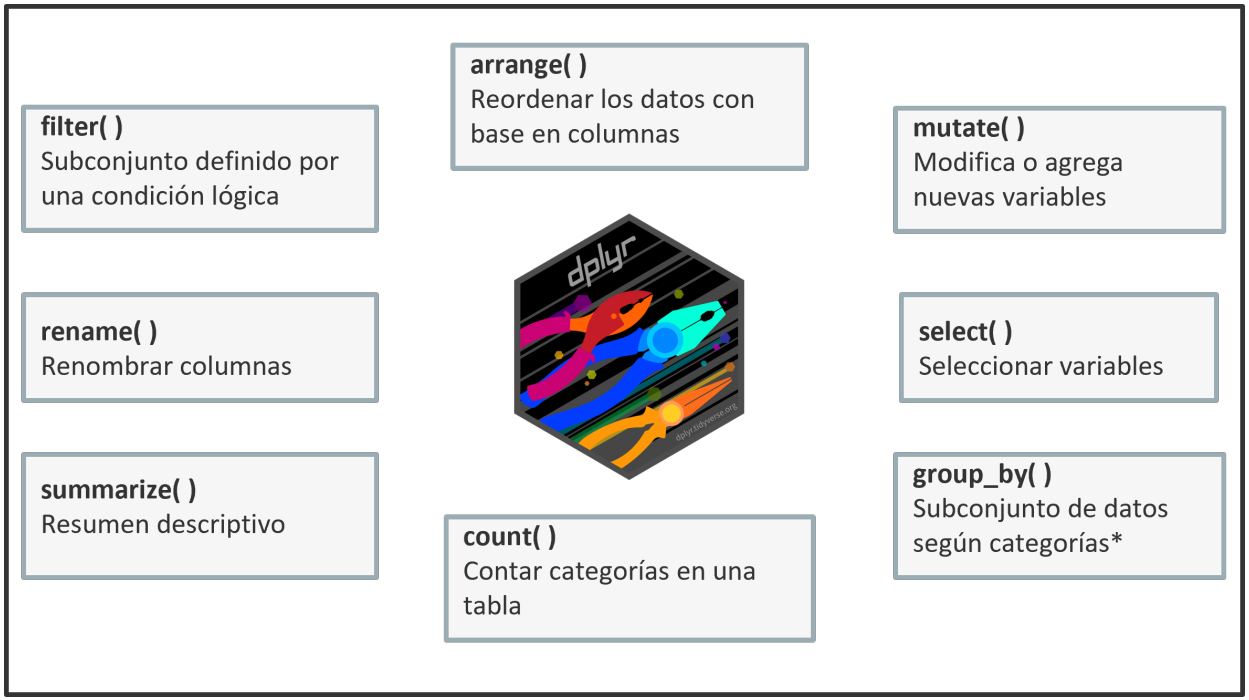


Figure 2: Algunas funciones en el paquete dplyr

Resumen por grupo

Usando las funciones `summarize()` y `group_by()`, obtenemos un resumen descriptivo de la base de datos diferenciado según una o más variables de control. Por ejemplo:

```
# Resumen general
table_1 <- new_dataset %>% filter(Int_attend == "Group 4")
%>%   summarize(MeanAttend = mean(attend),SdAttend = sd(attend))
```

```
# Resumen diferenciado
table_2 <- new_dataset %>% group_by(Int_attend) %>%
  summarize(MeanAttend = mean(attend), SdAttend = sd(attend))
```

La **Figura 10** muestra el funcionamiento de `summarize()` y `group_by()`.

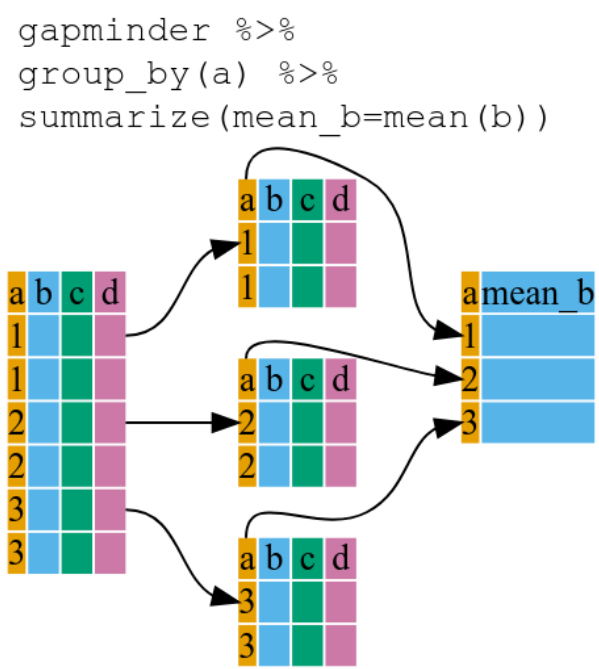


Figure 3: Caption for the picture.

ggplot2

El paquete **ggplot2** proporciona un sistema coherente para visualizar datos y crear gráficos. La versatilidad de **ggplot2** radica en el uso de la Gramática de Gráficos (*Grammar of Graphics*).

```
ggplot(dataset, aes()) + geometría + faceta + opciones
```

donde:

1. *dataset* es un data frame
2. Las características del mapa **aes()** describe los ejes (x, y), el color exterior (**color** o **colour**), el color interior (**fill**), la forma de los puntos (**shape**), el tipo de línea (**linetype**) y el tamaño (**size**)
3. Los objetos geométricos (**geometría**) determinan el tipo de gráfico:
   - Puntos (*geom_point*)
   - Líneas (*geom_lines*)
   - Histogramas (*geom_histogram*)
   - Boxplot (*geom_boxplot*)
4. La **faceta** permite dividir un gráfico en múltiples gráficos de acuerdo con grupos

Introducción al EDA

¿Qué es el EDA?

El Análisis Exploratorio de Datos (EDA, por sus siglas en inglés) proporciona una estrategia robusta para ampliar el entendimiento sobre los datos. El principio general es el siguiente:

> "Es importante comprender lo que podemos hacer antes de aprender a medir lo bien que parece que lo hemos hecho" (Tukey, 1977).

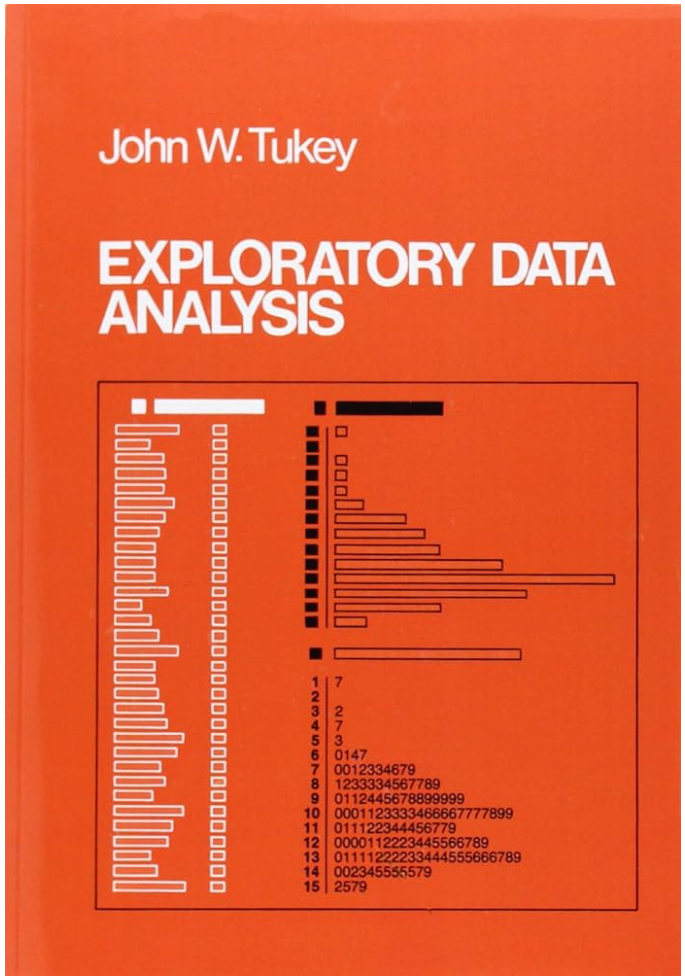


Figure 4: Tukey (1977). Exploratory Data Analysis

Utilidad

Mediante métodos gráficos y descriptivos, el EDA permite (Carranza, 2021):

- Revelar la estructura de los datos
- Determinar las variables relevantes
- Determinar valores atípicos
- Proponer hipótesis
- Proponer estrategias para la modelación

EDA y CDA

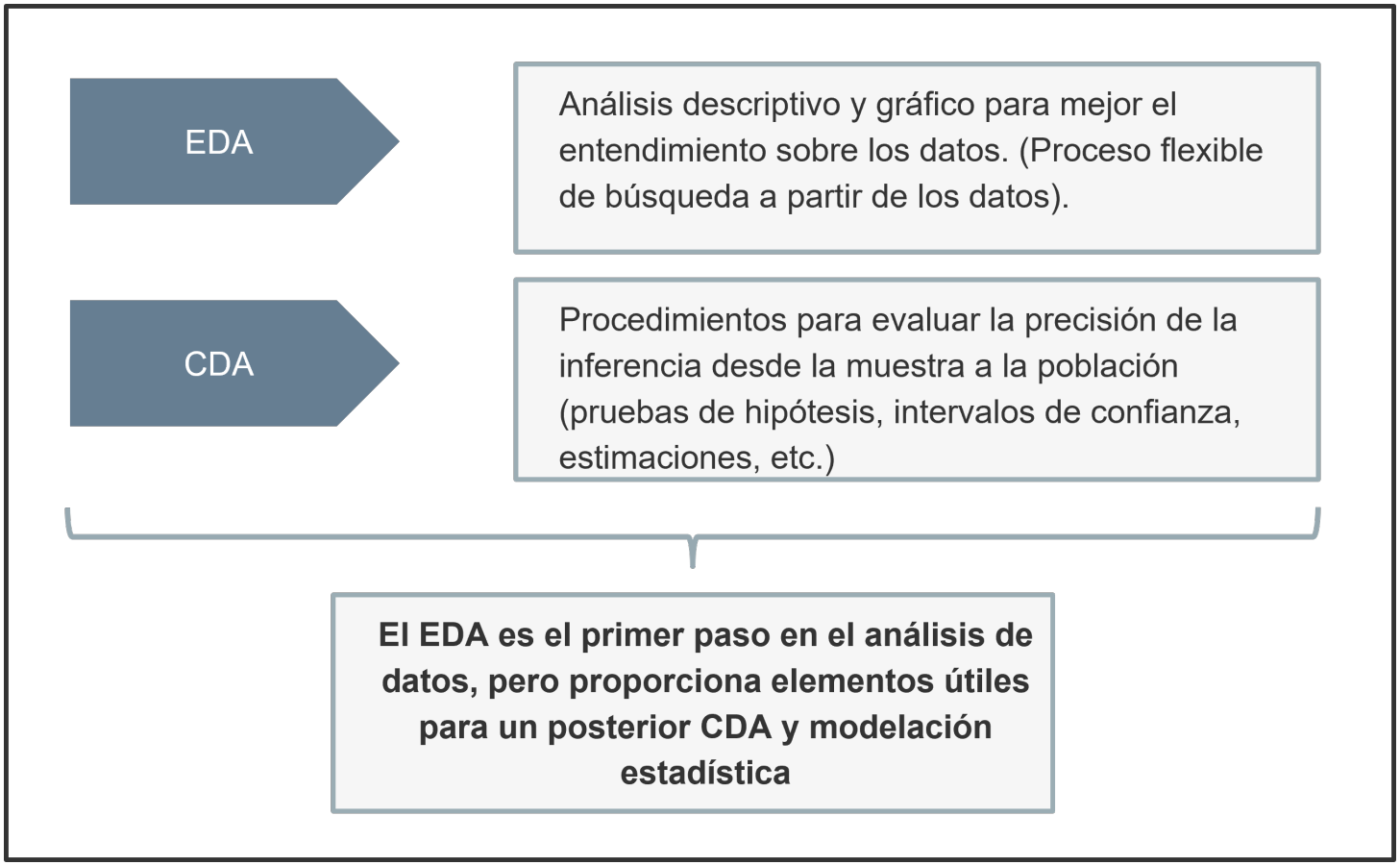


Figure 5: Análisis Exploratorio de Datos (EDA) y Análisis Confirmatorio de Datos (CDA)

Análisis Exploratorio de Datos (EDA) Univariado

Base de datos

La base de datos usada es extraída de los microdatos de la **Gran Encuesta Integrada de Hogares (GEIH)** para diciembre de 2023. El análisis considera las siguientes 13 ciudades y áreas metropolitanas:

- Medellín A.M.
- Barranquilla A.M.
- Bogotá
- Cartagena
- Manizales A.M.
- Monteria
- Villavicencio
- Pasto
- Cucuta A.M.
- Pereira A.M.
- Pereira A.M.
- Bucaramanga A.M.
- Ibague
- Cali A.M

La información es extraída de dos módulos de la GEIH:

- **Ocupados** (horas trabajadas , ingreso laboral, actividad económica, etc.)
- **Características generales, seguridad social en salud y educación** (edad, sexo, nivel de educación, etc.)

Base de datos (cont.)

Para importar la base de datos (.xlsx),

```
library(readxl)
dataset <- readxl::read_excel("Datos/Formatos/geih_dataset.xlsx")
```

La siguiente tabla muestra un resumen de la base de datos:

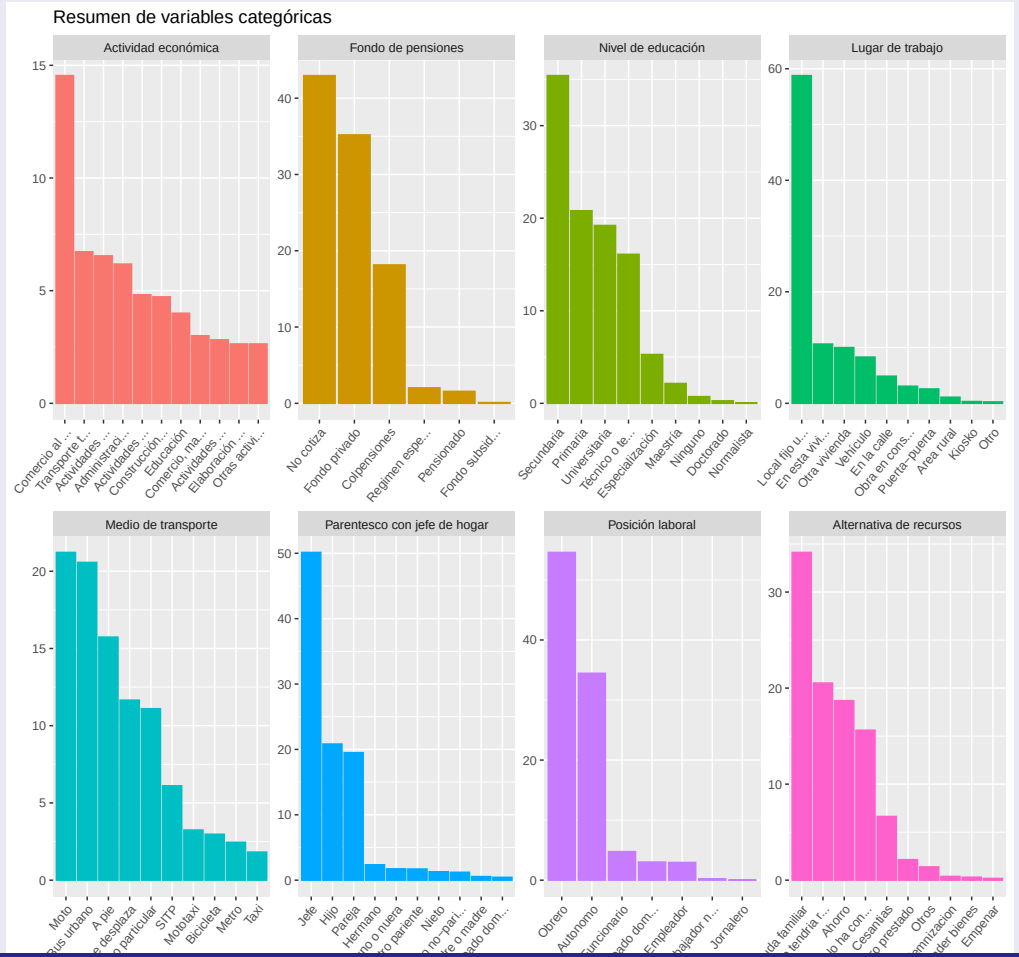
Variable	Clase	Descripción
area	Factor	Área metropolitana
dpto	Factor	Departamento
sexo	Factor	Sexo al nacer
parent	Factor	Parentesco con el jefe o jefa del hogar
edad	Numérica	Años cumplidos
edu	Factor	Mayor nivel educativo alcanzado
ingreso	Numérica	Ingreso laboral
horas_semana	Numérica	Horas trabajadas normalmente a la semana
cotiza	Factor	¿Cotiza a un fondo de pensiones?
lugar	Factor	Lugar principal de trabajo
meses	Numérica	¿Cuántos meses trabajó en los últimos 12 meses?
rama_4	Cadena	Rama de actividad CIIU REV 4 (4 dígitos)
rama_2	Cadena	Rama de actividad CIIU REV 4 (4 dígitos)
posic	Factor	Posición laboral
fondo	Factor	¿A cuál fondo cotiza?
cambiar	Factor	¿Desea cambiar su trabajo?
estable	Factor	¿Considera que su empleo es estable?
t_actual	Numérica	¿Cuánto tiempo lleva en su empleo actual?
t_viaje	Numérica	Tiempo de desplazamiento al trabajo
mas_h	Factor	¿Quiere trabajar más horas?
medio	Factor	Medio de transporte
sintrab	Factor	¿Si no tuviera trabajo, de dónde obtendría sus recursos?
n_comp	Factor	¿Cuántas personas tiene la empresa donde trabajo?
srl	Factor	¿Afiliación a ARL?
caja	Factor	¿Afiliación a caja de compensación familiar?
actividad	Factor	Actividad económica recodificada
cotiza_fondo	Factor	Fondo de pensiones recodificado
factor_exp	Numérica	Factor de expansión

Datos cualitativos:

Resumen de datos cualitativos

Considérese las siguientes variables cualitativas previamente recodificadas:

- La función `forcats::fct_lump_n()`es usada para agregar las categoría en "otros".
- Las gráficas muestran las 10 categorías más frecuentes.
- La función `ggplot2::facet_wrap` es usada para obtener los gráficos múltiples



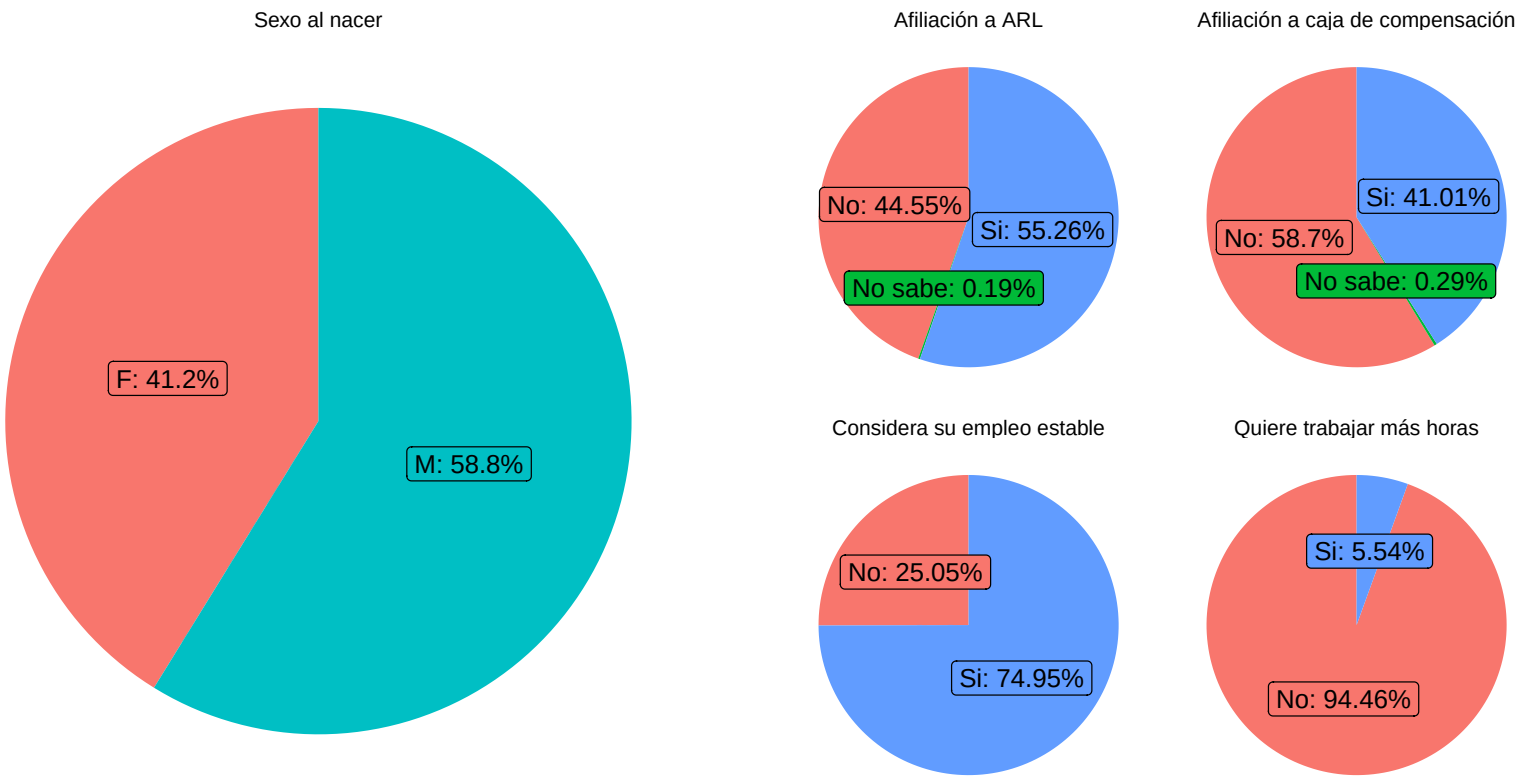
Resumen de datos cualitativos (cont.):

Idéntica información puede ser representada mediante el siguiente resumen:

	N	Proporción (%)
Actividad económica		
Actividades de atención de la salud humana	53	0.60
Actividades de los hogares individuales como empleadores de personal doméstico	31	0.35
Actividades de servicios de comidas y bebidas	72	0.82
Administración pública y defensa; planes de seguridad social de afiliación obligatoria	68	0.77
Comercio al por menor (incluso el comercio al por menor de combustibles), excepto el de vehículos automotores y motocicletas	160	1.82
Comercio, mantenimiento y reparación de vehículos automotores y motocicletas, sus partes, piezas y accesorios	33	0.38
Construcción de edificios	52	0.59
Educación	44	0.50
Elaboración de productos alimenticios	29	0.33
Other	455	5.17
Otras actividades de servicios personales	29	0.33
Transporte terrestre; transporte por tuberías	74	0.84
Fondo de pensiones		
Colpensiones	227	2.58
Fondo privado	360	4.09
No cotiza	464	5.27
Pensionado	23	0.26
Regimen especial	26	0.30
Educación		
Doctorado	1	0.01
Especialización	62	0.70
Maestría	24	0.27
Ninguno	9	0.10
Primaria	224	2.55
Secundaria	391	4.44
Técnico o tecnológico	177	2.01
Universitaria	212	2.41

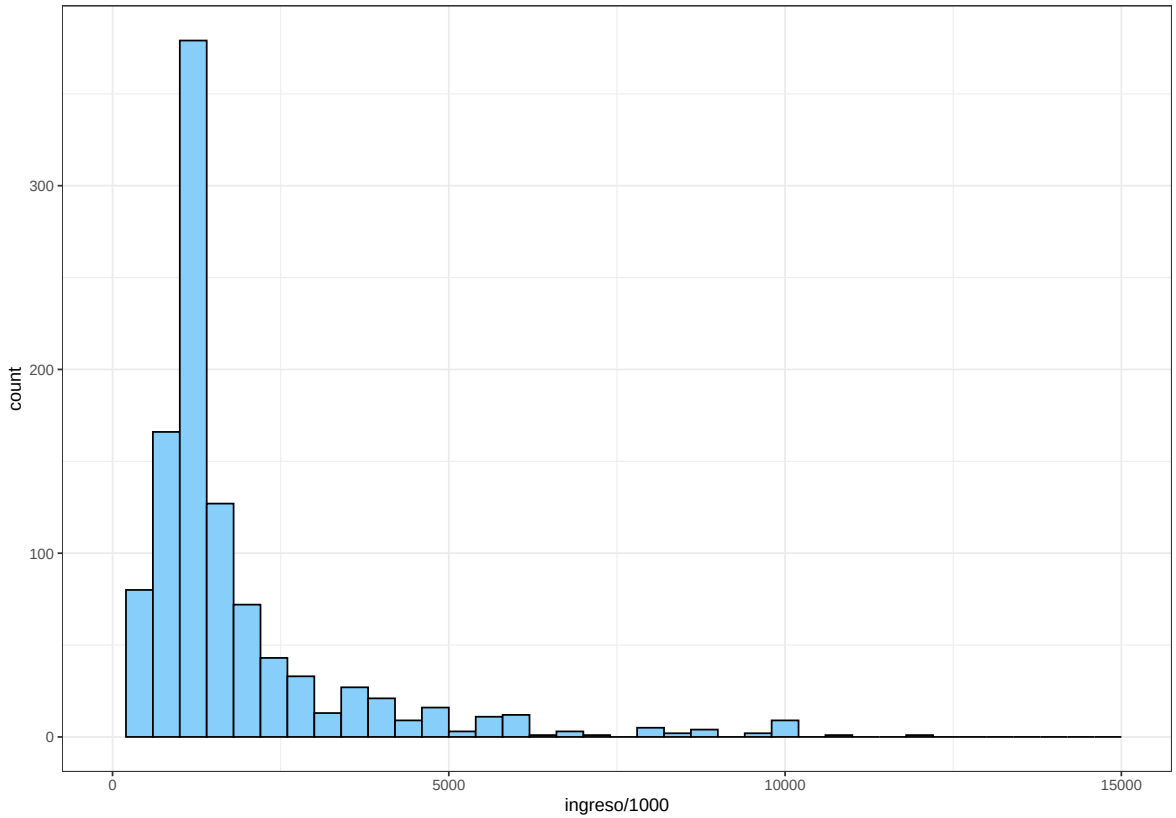
Resumen de datos cualitativos (cont.):

Un panel de **gráficos circulares** es útil para presentar un resumen sobre las variables categóricas con un número menor de niveles. Considérese las siguientes variables



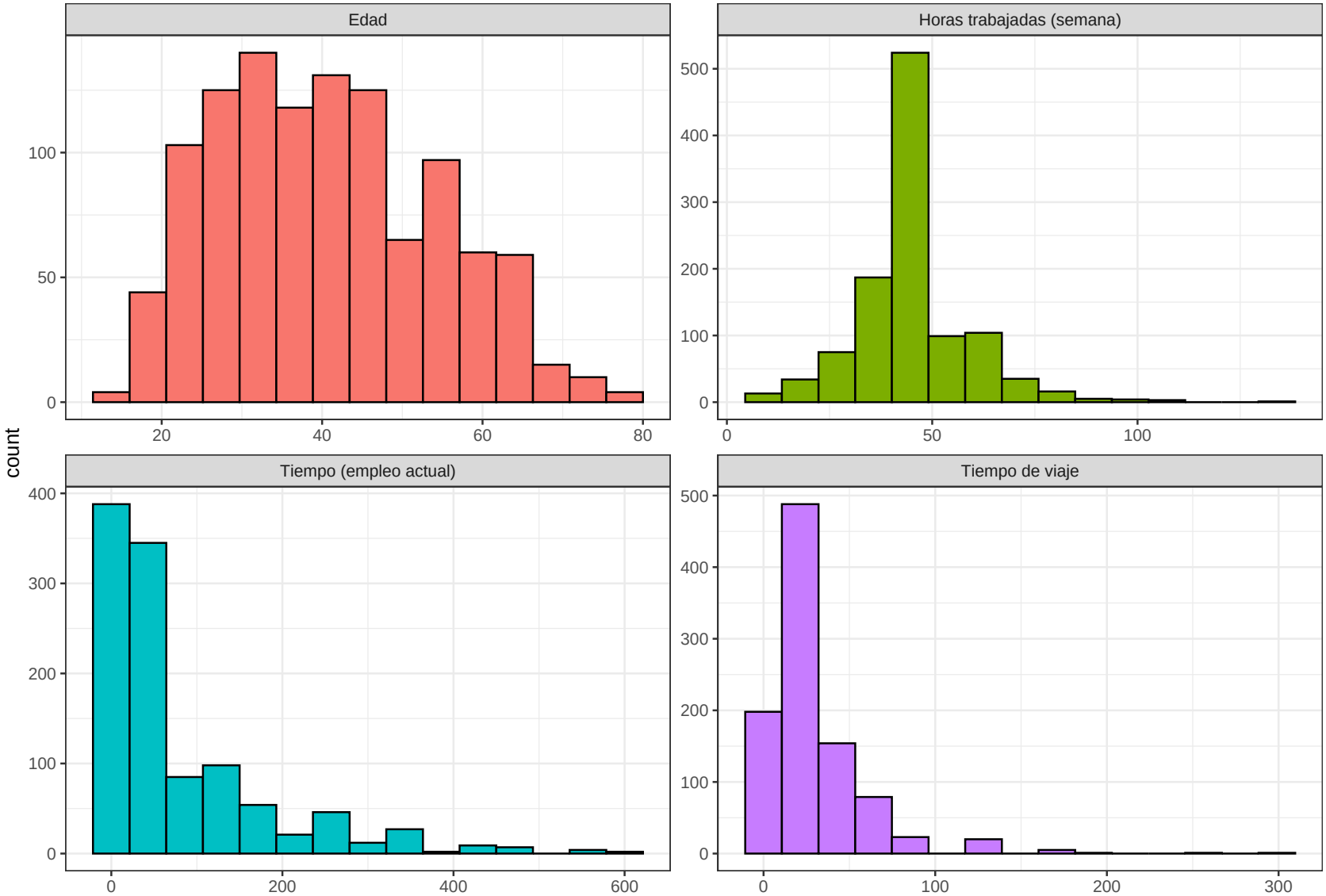
Datos cuantitativos

El **Histograma** es una representación gráfica de los datos que muestra la frecuencia de los casos (valores) en categorías de datos (véase la tabla inferior).



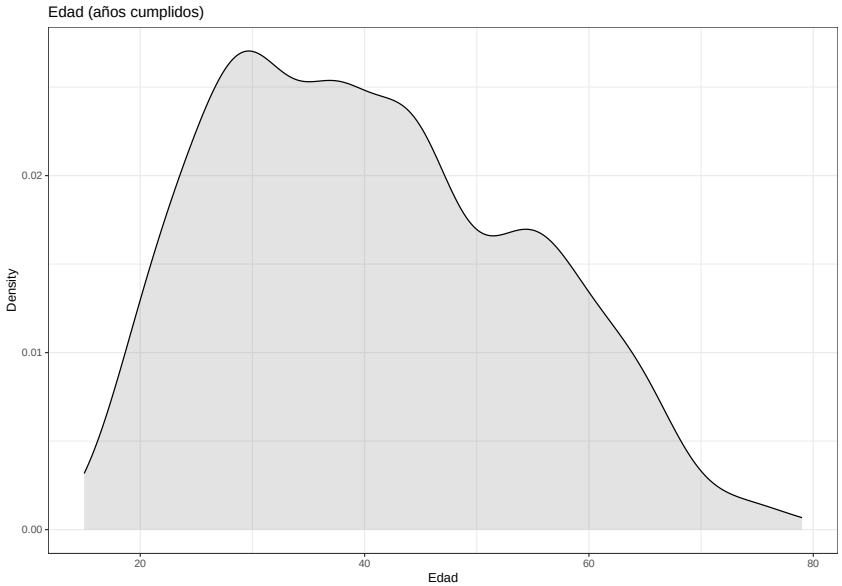
[0,1000]	(1000,2000]	(2000,3000]	(3000,4000]	(4000,5000]	(5000,6000]	(6000,7000]	(7000,8000]
246	556	98	57	29	25	5	6

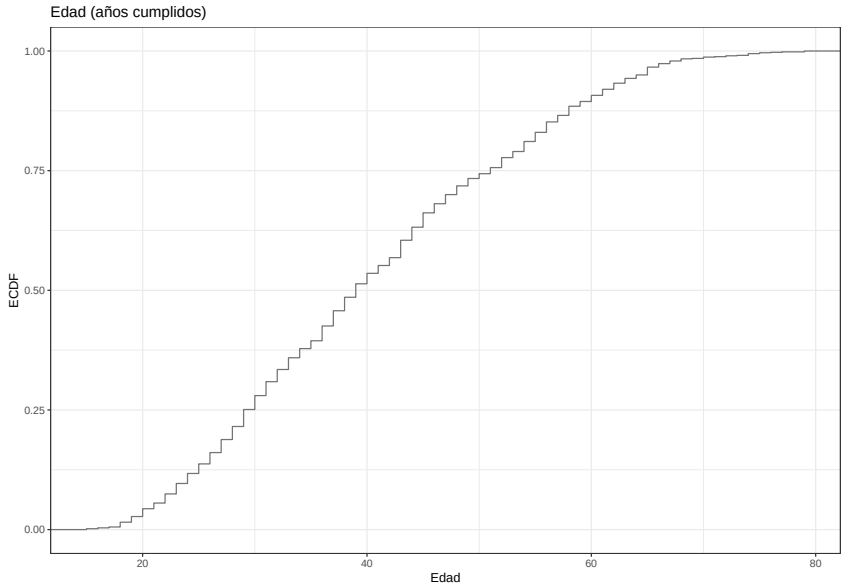
Histograma


Edad
Horas trabajadas (semana)
Tiempo (empleo actual)
Tiempo de viaje
count
0
50
100
20
40
60
80
200
400
300
500
600

PDF y ECDF

La **función de densidad empírica** (PDF) $f(x)$ y la **función de distribución acumulada empírica** (ECDF) $F(x)$ son obtenidas mediante las funciones `density()` y `ecdf()` del paquete `Stats`.





La función de densidad $f(x)$ satisface que $\int_a^b f(x)dx = P[a \leq X \leq b]$, donde $P[a \leq X \leq b]$ significa la probabilidad de que X se encuentre en el intervalo a a b. Por definición, $F(x) = P(X \leq x)$, es decir, expresa la probabilidad de que X no sea mayor que el valor de x.

Función de Distribución Acumulada Empírica

Un ejercicio análogo es implementado para las variables cuantitativas restantes.

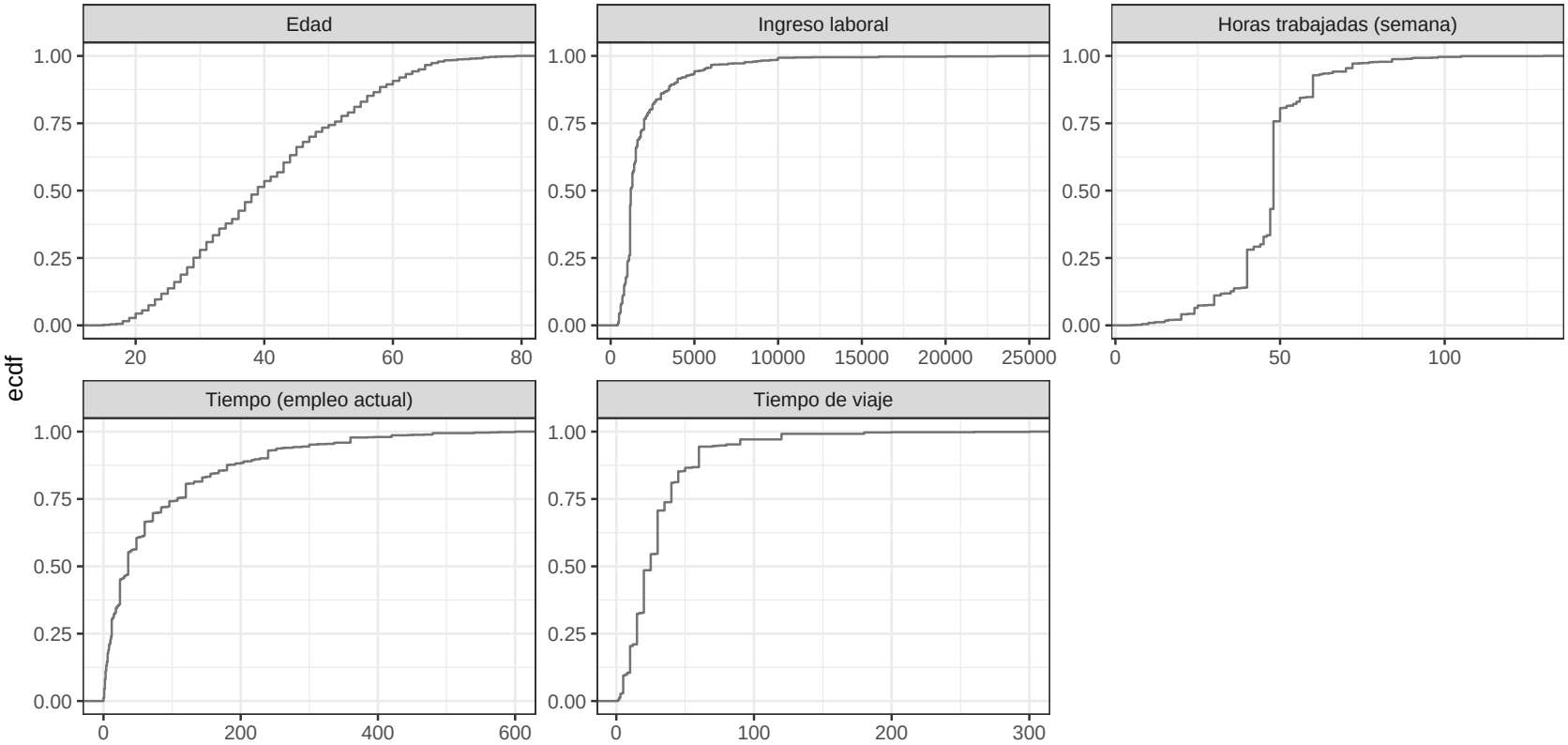


Diagrama de caja

Un **diagrama de caja** es un resumen gráfico de los datos con base en la los cuartiles Q1 y Q3, la mediana y el rango intercuartílico (RIC). El siguiente diagrama muestra su elaboración:

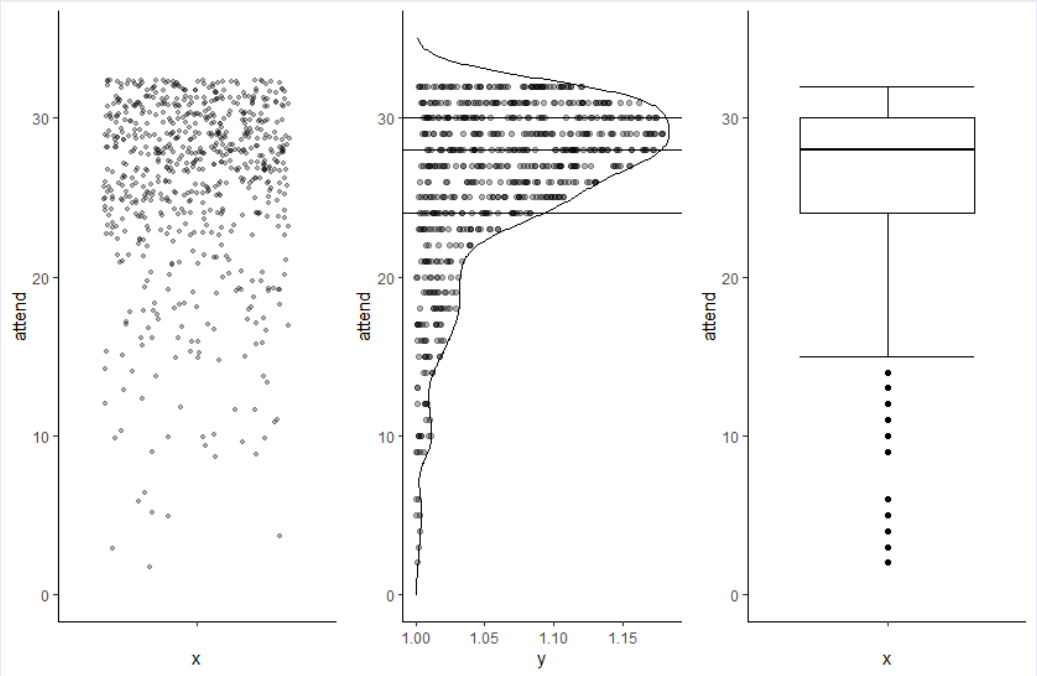


Figure 6: Construcción de un diagrama de caja

Un resumen de datos cuantitativos usando StatDA()

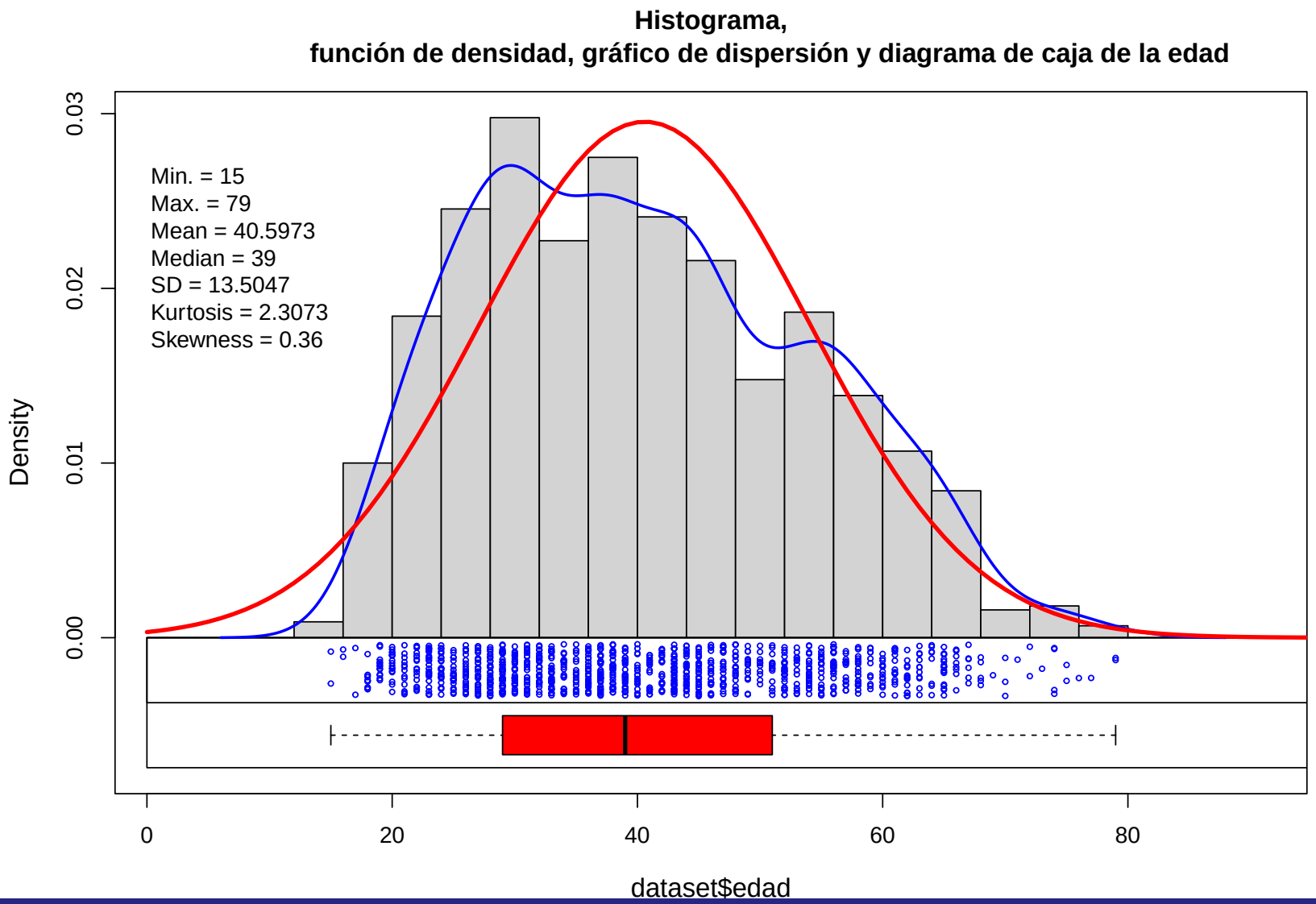
La librería `StatDA()` proporciona una utilidad para representar la distribución y los principales elementos descriptivos de las variables continuas.

```
library(StatDA)
library(moments)

me = mean(dataset$edad)
sd = sd(dataset$edad)

StatDA::edaplot(dataset$edad, scatter=TRUE, H.freq=FALSE, box=TRUE,
                H.breaks=seq(0,100, by = 4),H.col="lightgray", H.border=TRUE, H.labels=FALSE,
                S.pch=1, S.col="blue", S.cex=0.5, D.lwd=2, D.lty=1, D.plot=FALSE,
                P.xlim=c(1, 91), P.cex.lab =1.2, P.log=FALSE, P.main="Histograma,
                función de densidad, gráfico de dispersión y diagrama de caja de la edad",
                P.xlab="Edad (años)", P.plot=TRUE,
                P.ylab="Densidad",
                B.pch=1,B.cex=0.5, B.col="red")
lines(density(dataset$edad), lwd=2, col='blue')
curve(dnorm(x, mean=me, sd=sd), from=0, to=100, add=T,
      col='red', lwd=3)
leg.txt <- c(paste0("Min. = ", round(min(dataset$edad),4)),
             paste0("Max. = ", round(max(dataset$edad),4)),
             paste0("Mean = ", round(mean(dataset$edad),4)),
             paste0("Mediana = ", round(median(dataset$edad),4)),
             paste0("SD = ", round(sd(dataset$edad),4)),
             paste0("Kurtosis = ", round(kurtosis(dataset$edad),4)),
             paste0("Skewness = ", round(skewness(dataset$edad),4)))
legend (x=-3, y=0.028, bty="n", leg.txt)
```

Resumen de datos cuantitativos (cont.)



Valores atípicos

La perspectiva univariante selecciona como **valores atípicos** u **outliers** aquellas observaciones que caen fuera de los rangos de la distribución. Un valor atípico se puede producir por alguna de las siguientes cuatro causas (Aldás y Uriel, 2017):

- **Errores en los datos**: errores en la recolección o introducción de los datos.
- **Erorres voluntarios**: errores intencionados en la respuesta del entrevistado.
- **Errores de muestreo**: errores que son el resultado de introducir en la muestra a individuos pertenecientes a una población distinta a la **población objetivo**.
- **Outliers legítimos**: caso de la población objetivo que, por la variabilidad de las muestras, difiere del resto de casos.

Detección univariada de valores atípicos

Considérese las siguientes alternativas para la detección univariante de *outliers*:

Criterio intercuartílico

$$x^* \in [q_{0.25} - 1.5IQR, q_{0.75} + 1.5IQR]$$

Criterio de valores estandarizados (Hair et al., 2014)

- Para muestras pequeñas ($n < 80$), x^* tiene valores estándar de 2.5 o superiores.
- Para muestras mayores ($n \geq 80$), x^* tiene valores estándar de 3-4 o superiores.

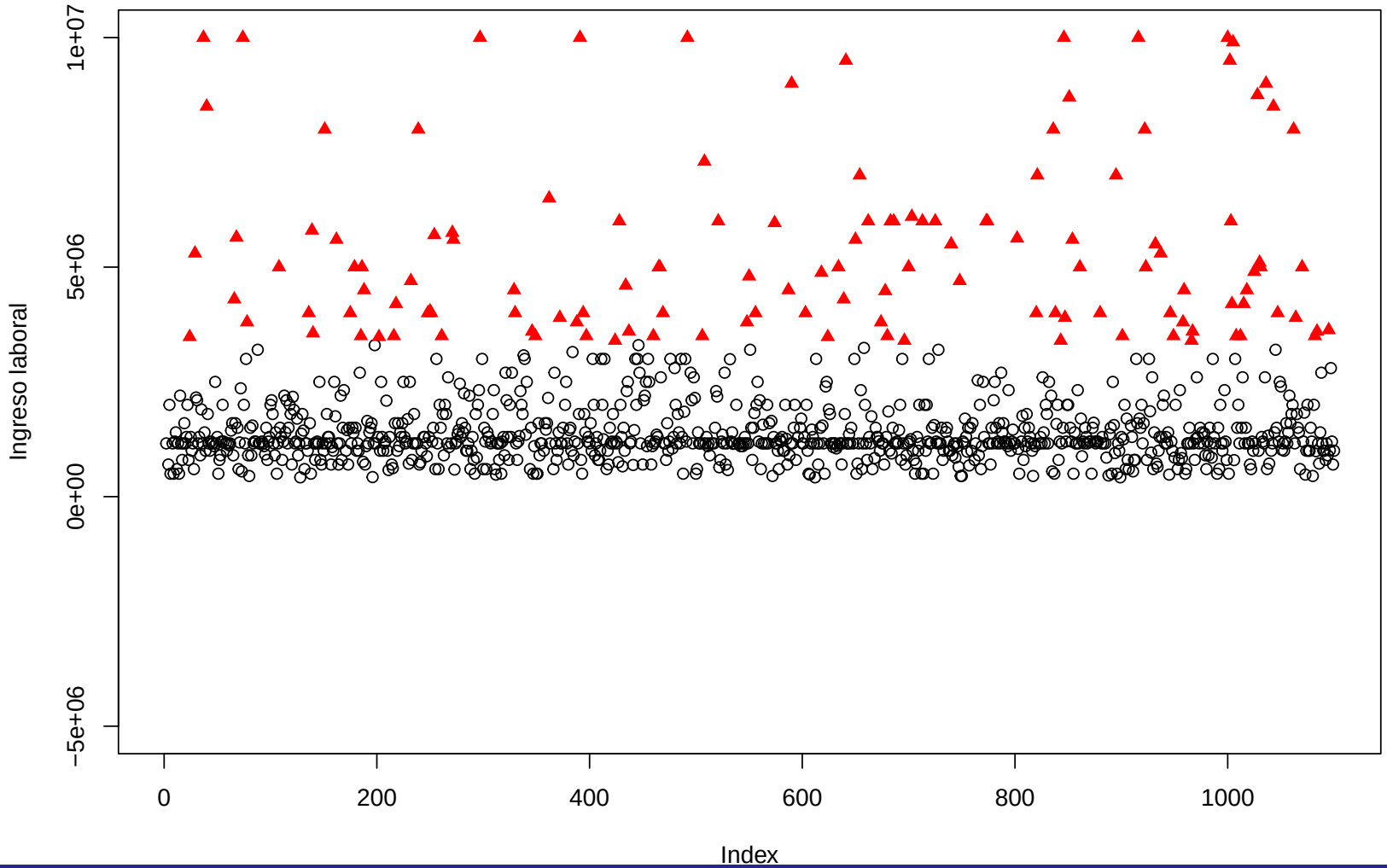
Test de Grubbs

El **Test de Grubbs** supone la normalidad de la distribución (Grubbs, 1969; Stefansky, 1971). La hipótesis nula (no hay *outliers*) se rechaza si

$$G > \frac{n-1}{n}\sqrt{\frac{t^2_{(\alpha/2n,n-2)}}{n-2+t^2_{(\alpha/2n,n-2)}}}$$

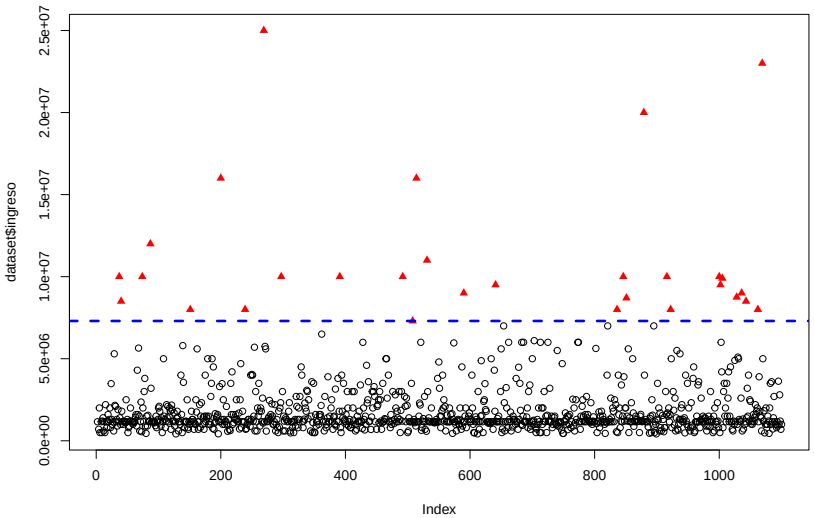
donde $G = max|x_i - \bar{x}|/\sigma$

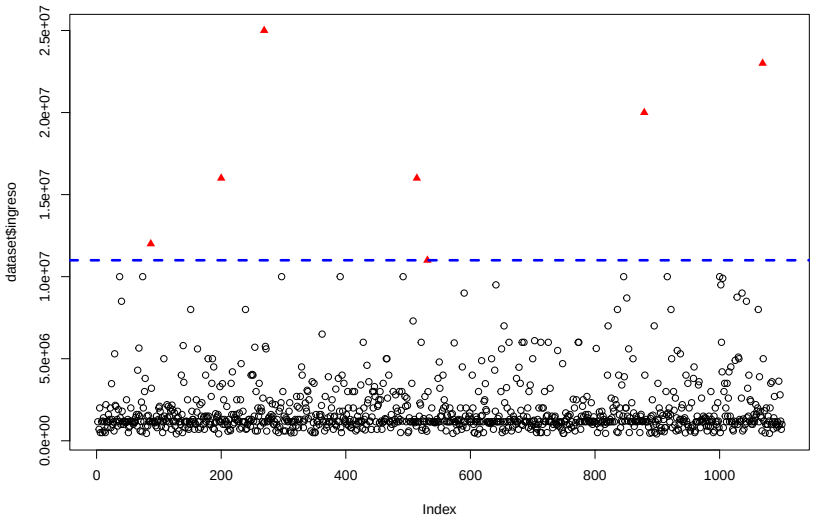
Criterio intercuartílico



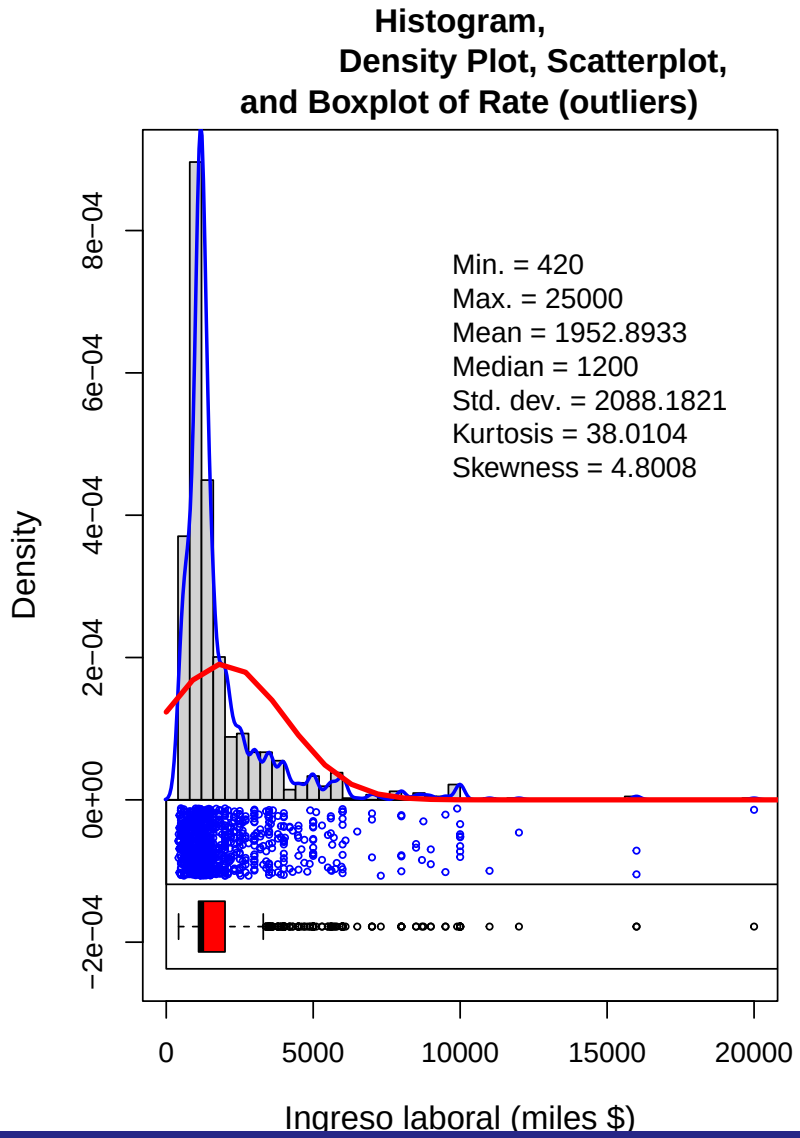
Criterio de valores estandarizados

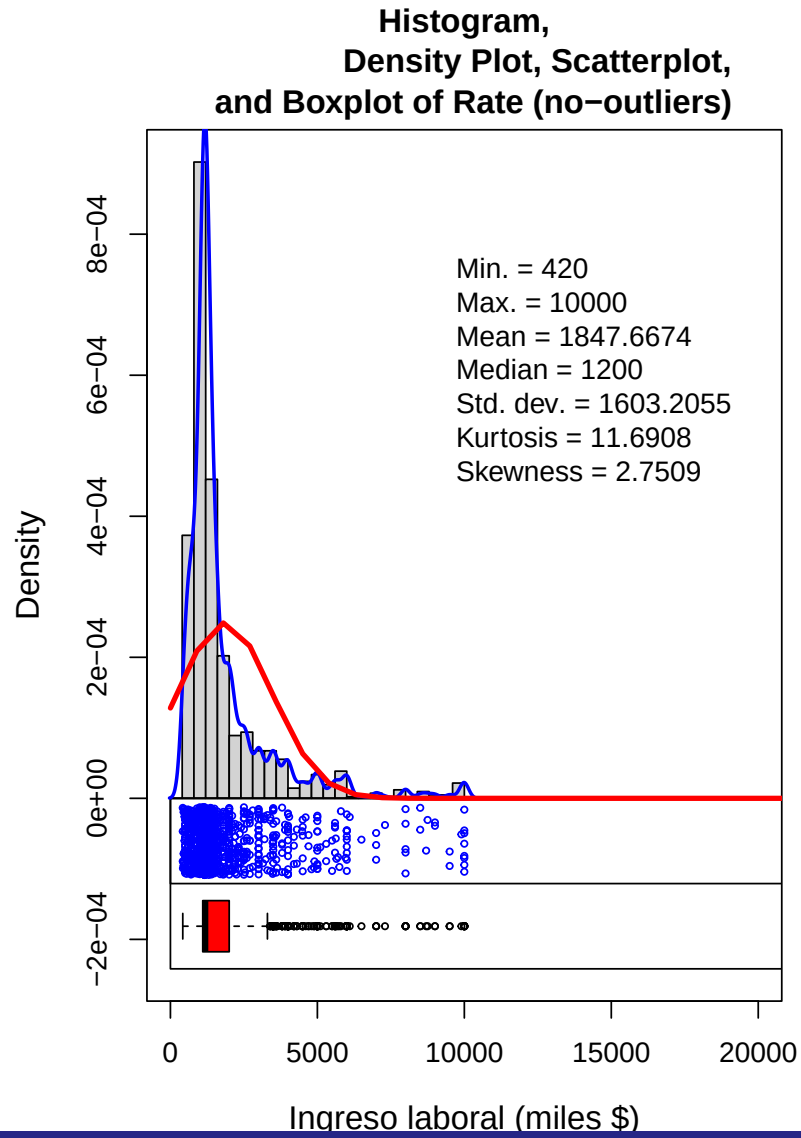
Considérese la detección de valores atípicos mediante los siguientes valores estandarizados: x^* es un valor atípico cuando $z^* \geq 2.5$ (**figura A**); y x^* es un valor atípico cuando $z^* \geq 4$ (**figura B**).





Resumen de la variable continua con y sin outliers





Alternativas

Algunas alternativas para el tratamiento de valores atípicos son las siguientes:

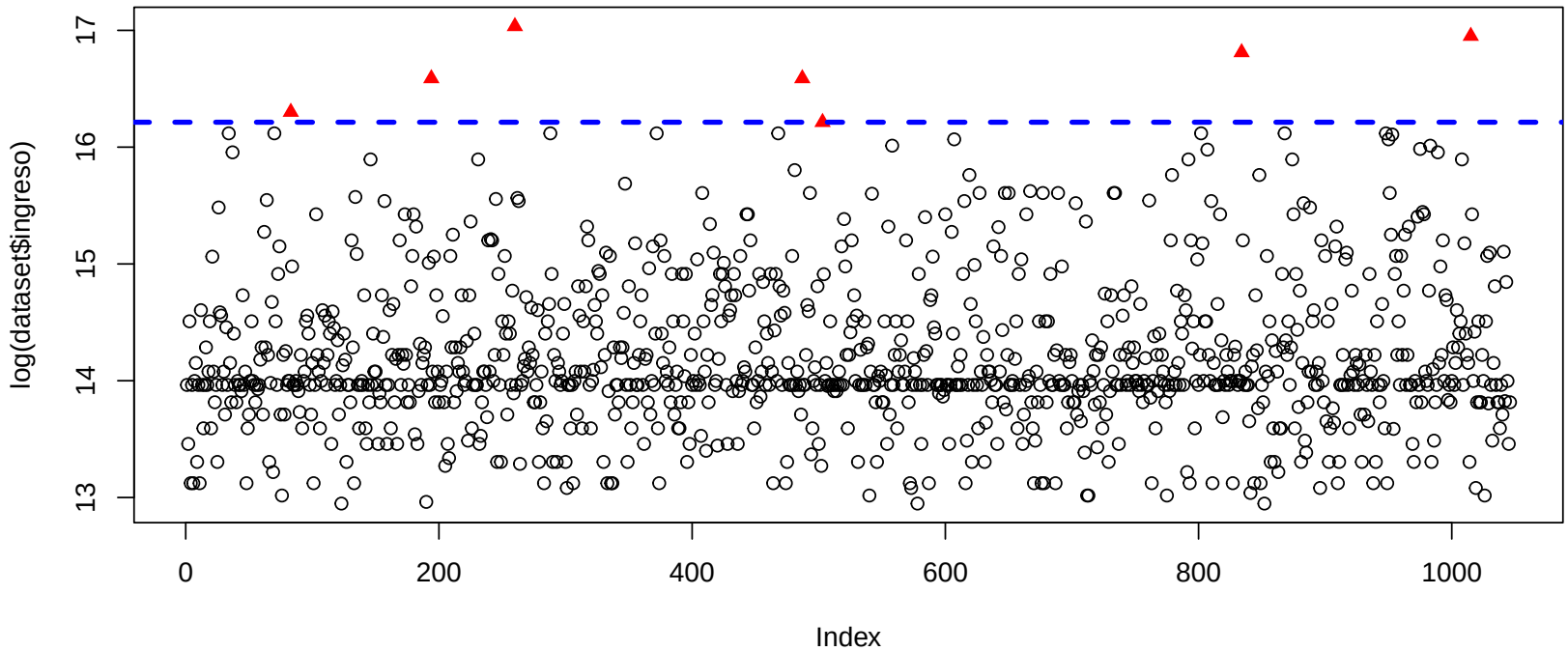
1. Eliminación de los valores atípicos para garantizar estimaciones correctas sobre la mayoría de la población (Judd et al., 2009).
2. Suavizar la influencia de los valores atípicos mediante el uso de transformaciones (raíces o logaritmos) para reducir su rango (Hamilton, 1992)
3. Análisis estadístico **robusto**.

Desventajas

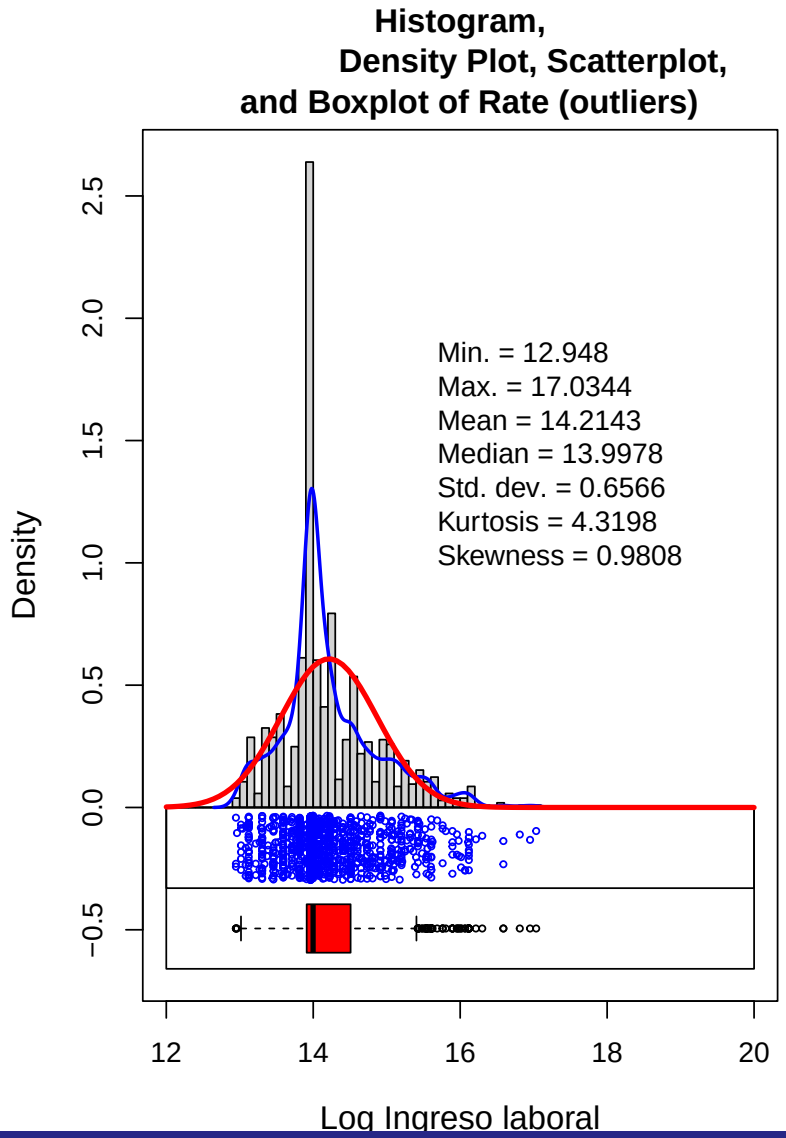
- Perdida de información
- No todas las transformaciones conservan el sentido teórico de la escala original.

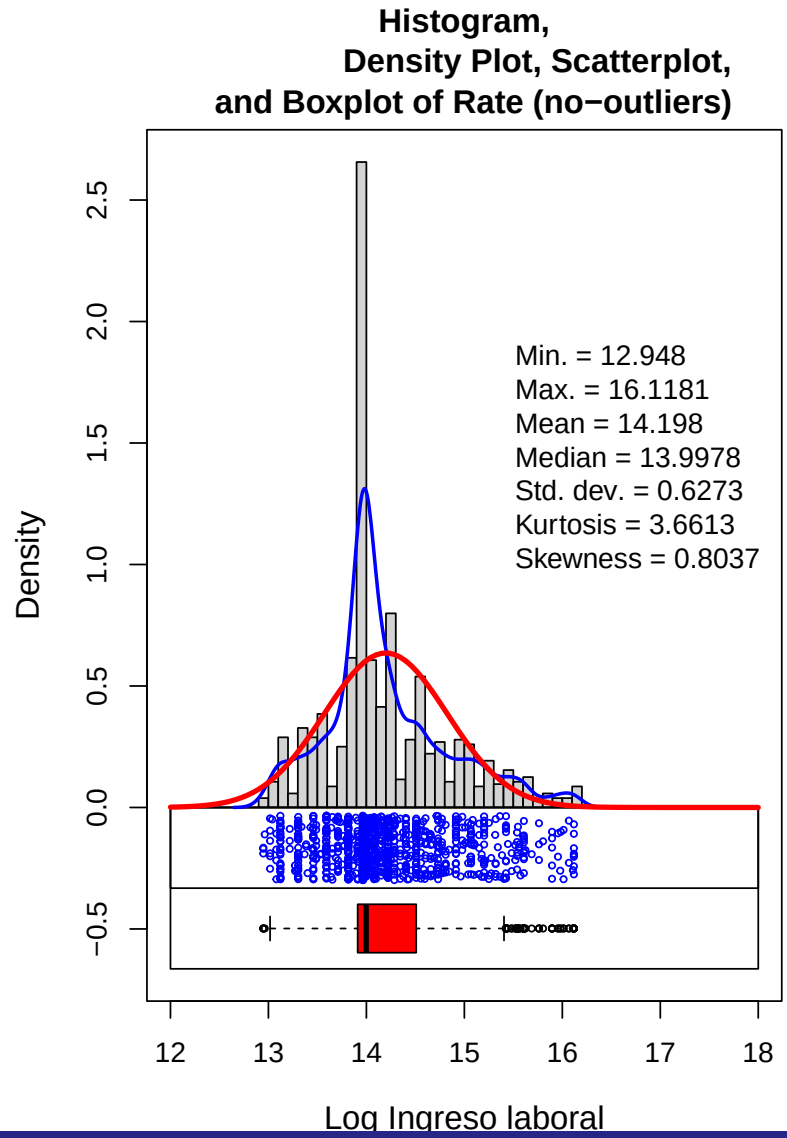
Transformaciones para reducir su rango

En lo sucesivo, consideramos el efecto de una transformación logarítmica. Nótese que la transformación logra **reducir el rango** y suavizar, en consecuencia, la influencia de los valores atípicos. (La línea azul indica el umbral a partir del cual un valor es considerado atípico según el **criterio de valores estandarizados**).



Transformaciones





Recursos alternativos

Recursos alternativos

- La librería `swirl` proporciona un tutorial sobre elementos básicos en R

```
install.packages("swirl")
library (swirl)
swirl()
```

- Data wrangling with dplyr and tidyr (Cheat Sheet): Recurso 1.2
- Visualización de datos usando ggplot2 (Guía Rápida): Recurso 1.3
- Factors with forcats (Cheat Sheet): Recurso 1.4
  \end{itemize}

Bibliografía de consulta

Bibliografía de consulta

- Wickham, H. (2016) GGplot2. Elegant Graphics for Data Analysis. Springer
- Grolemund, G. (2014). Hands-On Programming with R. O'Reilly Media: Sebastopol, CA.
- Schutt, R. & O'Neil, C. (2014). Doing Data Science. O'Reilly Media: Sebastopol, CA.
- Wickham & Grolemund, G. (2016). R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. O'Reilly Media: Sebastopol, CA.

Pruebas diagnóstico

Distribución normal univariada

Una variable aleatoria continua X está normalmente distribuida si su **función de densidad** sigue la forma:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left\{\frac{-1}{2\sigma^2}(x-\mu)^2\right\}$$

donde

$$f(x) \geq 0$$

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

$$\int_{a}^{b} f(x)dx = P(a \leq x \leq b)$$

Parámetros de la distribución: media(μ) y varianza (σ^2) de la distribución.

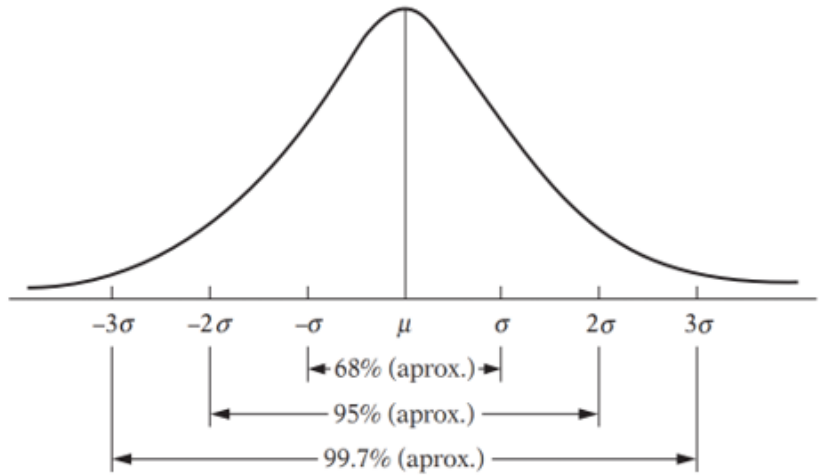
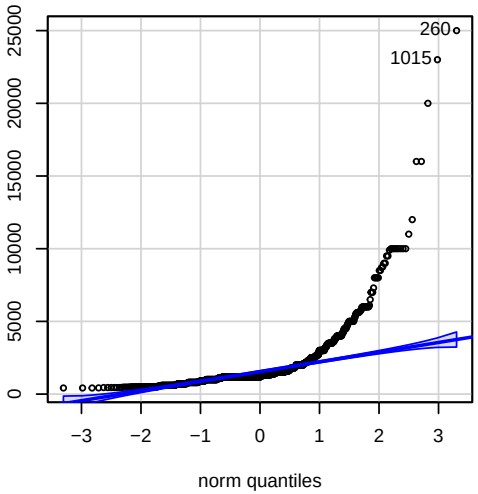


Figure 7: Áreas bajo la curva normal

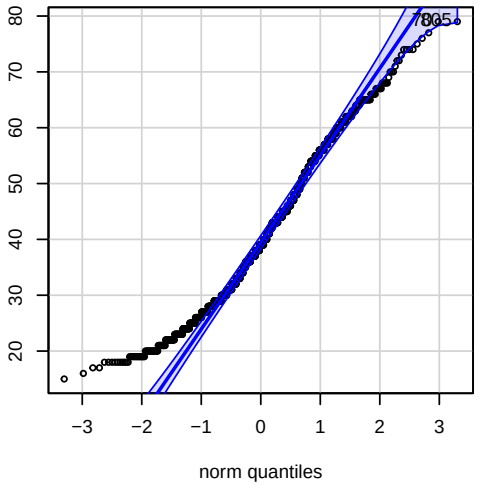
Normalidad univariada

El gráfico **cuantil-cuantil** (gráfico Q-Q) compara dos distribuciones de probabilidad usando sus cuantiles. Usando la librería `car`, comparamos la distribución de probabilidad de una muestra aleatoria con la distribución normal.

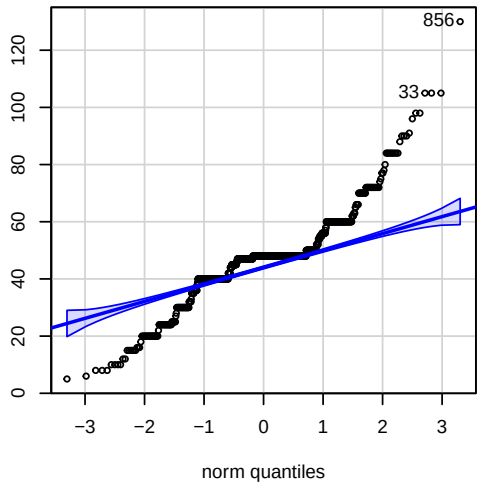
Ingreso laboral (miles $)



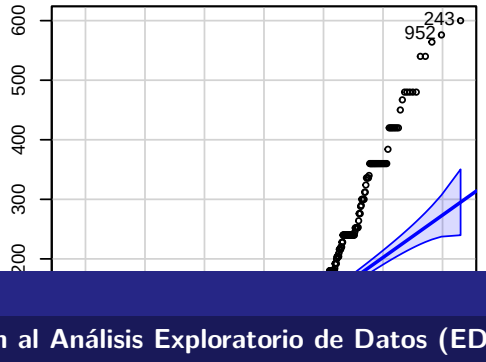
Edad (años)



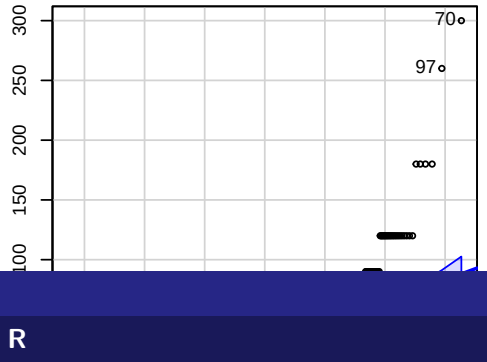
Horas trabajadas (semana)



Tiempo en el trabajo actual



Tiempo de desplazamiento



Normalidad univariada (cont.)

La intuición del método gráfico es verificada mediante pruebas formales. Usando la librería `nortest`:

variable	Anderson-Darling	Lilliefors (Kolmogorov-Smirnov)	Pearson chi-square
ingreso	0	0	0
edad	0	0	0
horas_semana	0	0	0
t_actual	0	0	0
t_viaje	0	0	0

Se examinan las siguientes transformaciones

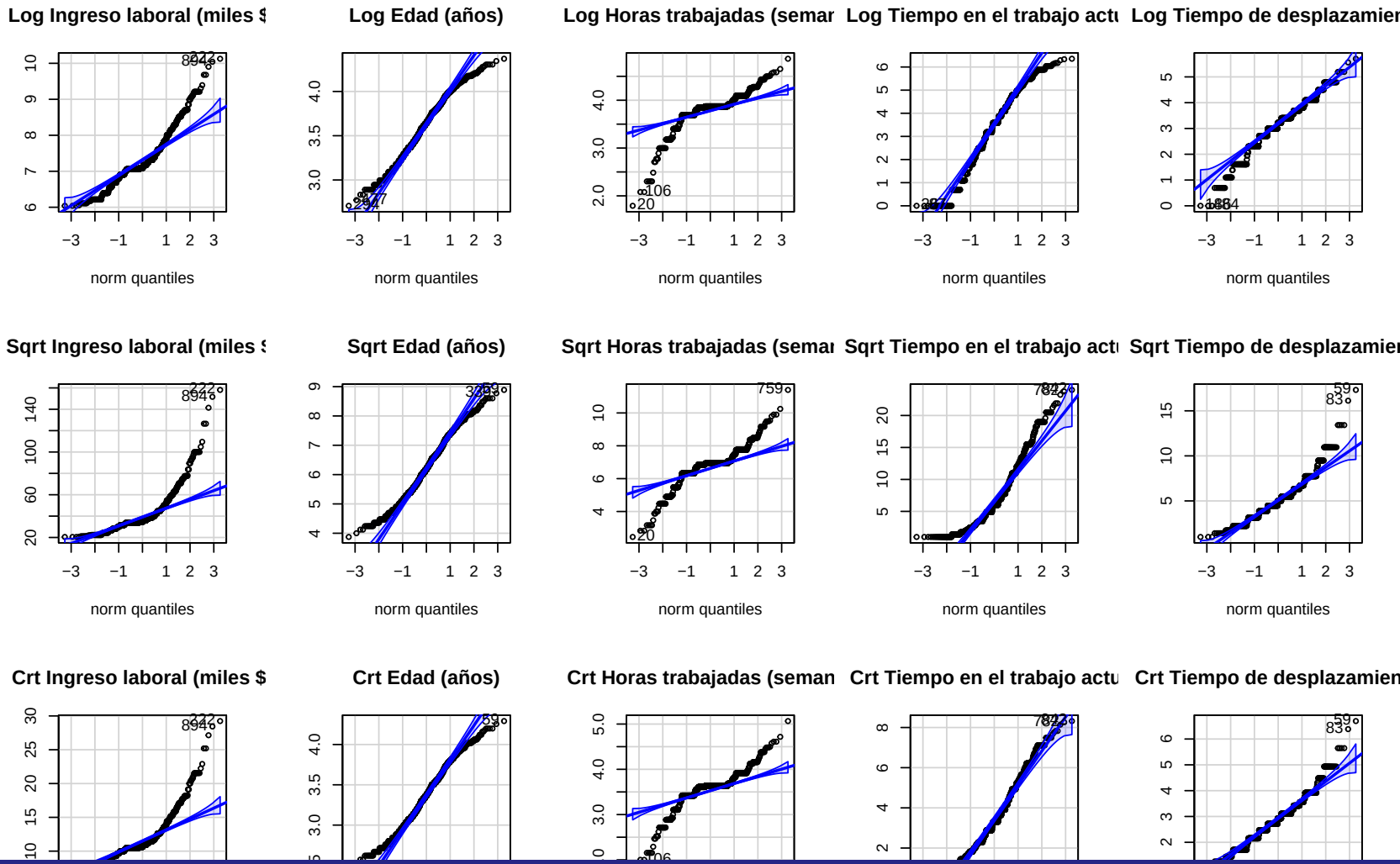
log(x)	A-D	K-S)	P
ingreso	0	0	0
edad	0	0	0
horas_semana	0	0	0
t_actual	0	0	0
t_viaje	0	0	0

sqrt(x)	A-D	K-S)	P
ingreso	0	0	0
edad	0	0	0
horas_semana	0	0	0
t_actual	0	0	0
t_viaje	0	0	0

cuberoot(x)	A-D	K-S)	P
ingreso	0	0	0
edad	0	0	0
horas_semana	0	0	0
t_actual	0	0	0
t_viaje	0	0	0

Normalidad univariada (cont.)

El siguiente panel corresponde a los gráficos Q-Q para las variables transformadas:



Resumen descriptivo (mediana e IQR)

Área	Ingreso	Edad	Horas (semana)	Tiempo actual	Tiempo de viaje
Total	1200000 (1100000 - 2e+06)	39 (29 - 50)	48 (40 - 48)	36 (12 - 108)	25 (15 - 40)
Barranquilla	1200000 (1160000 - 1875000)	36 (29 - 47.5)	48 (47.25 - 48)	33 (11.25 - 81)	30 (20 - 45)
Bogota	1360000 (1160000 - 2337500)	42 (27.5 - 49.5)	48 (41 - 56)	30 (10.5 - 84)	40 (20 - 60)
Bucaramanga	1250000 (1160000 - 2e+06)	38.5 (30 - 50.25)	47 (40 - 48)	48 (15 - 120)	28 (15 - 30)
Cali	1160000 (1112500 - 1675000)	39.5 (31 - 53)	47 (40 - 48)	36 (12 - 96)	30 (20 - 45)
Cartagena	1200000 (9e+05 - 2e+06)	37.5 (29 - 47.5)	48 (40 - 48)	60 (24 - 120)	20 (15 - 30)
Cucuta	1160000 (8e+05 - 1487500)	38 (31 - 47.75)	48 (40 - 55.75)	24 (7.25 - 82.5)	20 (15 - 30)
Ibague	1450000 (1160000 - 3400000)	40 (27 - 46)	48 (40 - 48)	28 (11 - 120)	20 (10 - 30)
Manizales	1200000 (1160000 - 2e+06)	41 (31 - 52)	47 (41 - 48)	36 (11 - 108)	20 (15 - 30)
Medellin	1303000 (1160000 - 2e+06)	36 (27.25 - 47.75)	48 (47 - 48)	24 (9.25 - 105)	35 (20 - 46.25)
Monteria	1160000 (780000 - 2050000)	37.5 (30.25 - 48)	47 (40 - 48)	40 (20.25 - 75.75)	15 (12 - 20)
Pasto	1200000 (9e+05 - 1820000)	44 (34 - 54)	48 (40 - 50)	36 (12 - 144)	20 (10 - 30)
Pereira	1275000 (1160000 - 2e+06)	38 (30 - 46.5)	48 (40 - 48)	36 (12 - 72)	20 (10 - 40)
Villavicencio	1225500 (1160000 - 2e+06)	42 (28 - 53.25)	48 (47 - 52.25)	20 (4 - 120)	25 (15 - 30)