

Curso de Visualización de Datos

Sesión 1: Análisis Exploratorio de Datos

Introducción a ggplot2

Dr. Jorge de la Vega Góngora

jorge.delavegagongora@gmail.com

Google Clasroom: igyd7ee

Github: https://github.com/jvega68/IASI_Data_Viz

Curso Virtual 10 de Noviembre 2023



1

Introducción

- Análisis Exploratorio de Datos
- Visualización: Metodos gráficos para describir datos
 - Técnicas de visualización univariada
 - Técnicas de visualización multivariada
- Métodos numéricos para describir datos
 - Estadísticas univariadas
 - Medidas de tendencia central
 - Medidas de dispersión
 - Otras estadísticas relevantes
 - Estadísticas multivariadas
- Conclusión

Introducción I

- Supongan que llegan a sus manos un conjunto de datos, algo como lo que pueden obtener de [esta liga¹](#).

```
library(readxl)
temp <- tempfile(fileext = ".xlsx")
download.file(url = "https://github.com/jvega68/MIDE_DCD/raw/master/Residential-Building-Data-Set.xlsx",
              destfile = temp,
              mode = "wb",
              quiet = T)
data <- read_xlsx(temp, sheet = "Data", skip = 1)
head(data)

# A tibble: 6 x 109
  `START YEAR` `START QUARTER` `COMPLETION YEAR` `COMPLETION QUARTER` `V-1` `V-2` `V-3` `V-4` `V-5` `V-6` `V-7` `V-8` `V-11...13`
    <dbl>        <dbl>          <dbl>          <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1      81           1            85             1   1  3150   920   598.   190 1011.   16 1200   6713
2      84           1            89             4   1  7600 1140 3040   400 964.   23 2900   3152
3      78           1            81             4   1  4800   840   480   100 690.   15 630    1627
4      72           2            73             2   1  685   202  13.7   20 460.   4  140    2581.
5      87           1            90             2   1  3000   800 1230   410 632.   13 5000   6790
6      87           1            90             1   1  2500   640 1050   420 647.   12 4800   6790
# i 96 more variables: `V-12...14` <dbl>, `V-13...15` <dbl>, `V-14...16` <dbl>, `V-15...17` <dbl>, `V-16...18` <dbl>, `V-17...19` <dbl>,
# `V-18...20` <dbl>, `V-19...21` <dbl>, `V-20...22` <dbl>, `V-21...23` <dbl>, `V-22...24` <dbl>, `V-23...25` <dbl>, `V-24...26` <dbl>,
# `V-25...27` <dbl>, `V-26...28` <dbl>, `V-27...29` <dbl>, `V-28...30` <dbl>, `V-29...31` <dbl>, `V-11...32` <dbl>, `V-12...33` <dbl>,
# `V-13...34` <dbl>, `V-14...35` <dbl>, `V-15...36` <dbl>, `V-16...37` <dbl>, `V-17...38` <dbl>, `V-18...39` <dbl>, `V-19...40` <dbl>,
# `V-20...41` <dbl>, `V-21...42` <dbl>, `V-22...43` <dbl>, `V-23...44` <dbl>, `V-24...45` <dbl>, `V-25...46` <dbl>, `V-26...47` <dbl>,
# `V-27...48` <dbl>, `V-28...49` <dbl>, `V-29...50` <dbl>, `V-11...51` <dbl>, `V-12...52` <dbl>, `V-13...53` <dbl>, `V-14...54` <dbl>,
# `V-15...55` <dbl>, `V-16...56` <dbl>, `V-17...57` <dbl>, `V-18...58` <dbl>, `V-19...59` <dbl>, `V-20...60` <dbl>, ...
```

- ¿Qué harían? ¿Se pueden analizar un conjunto de datos sin saber su contexto?
Probablemente sí, pero posiblemente no se puede decir mucho de ellos.

Introducción II

- ¿En qué unidades se miden? ¿Qué variables se pueden comparar entre sí? ¿Los tipos de variables fueron codificadas correctamente? Por ejemplo, en los datos la variable con la etiqueta V-1 dice que son dobles, pero sólo toman unos cuantos valores (valores del 1 al 20):

```
unique(data$`V-1`)
[1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
```

- Sin embargo, aun así se puede hacer una ligera exploración de los datos y ganar información, tanto numérica como visualmente
 - Por el nombre del archivo, podemos ver que los datos corresponden a edificios residenciales.
 - Podemos notar que las variables START YEAR, COMPLETION YEAR corresponden a años, en el formato AA. También las variables START QUARTER y COMPLETION QUARTER indican el número del trimestre del 1 al 4.
 - Se pueden hacer algunas gráficas para ver los datos, pero siempre nos hará falta saber qué estamos viendo para poder tomar decisiones sobre qué pasos seguir. Quizá adivinaríamos el significado de algunas variables, pero difícilmente podremos llegar a conclusiones adecuadas.
 - El archivo contiene una descripción muy general de las variables, las unidades en las que se midieron que pueden ayudar con el análisis. Los datos parecen ser de 2015 o alrededor de esa fecha.

- Más allá de esta información, no hay mucho que podamos decir sobre las relaciones entre las variables, cuáles serían explicativas, cuáles respuestas. Una descripción de los datos indica que se cuenta con costos de construcción, precios de venta, 8 variables físicas de proyecto y 19 variables económicas que corresponden a apartamentos para una familia en Tehrán, Irán.
- Es muy importante contar con un **contexto de análisis**.

¹La información complementaria de los datos se encuentra en
<https://archive.ics.uci.edu/ml/datasets/Residential+Building+Data+Set>

Introducción



**“Exploratory data analysis is
detective work - numerical
detective work, counting
detective work, or graphical
detective work”**
John Tukey, 1977

Análisis Exploratorio de Datos I

- El objetivo principal del Análisis Exploratorio de Datos (EDA, por sus siglas en inglés) es descubrir la estructura de un conjunto de observaciones sin asumir hipótesis sobre la estructura de esas observaciones o variables.
- EDA es un conjunto de técnicas para tratar de entender y conocer los datos de manera efectiva. Estas técnicas nos permiten familiarizarnos con los datos e identificar sus principales características.
- Lo primero que se debe hacer con los datos es conocerlos y tratar de extraer la esencia de la información.
- EDA no es un proceso formal con reglas estrictas, es más bien una forma de ganar información acerca de posibles estructuras que permitirán generar hipótesis y preguntas de investigación y que requerirán validarse con mayor análisis o a través de modelado.

Pasos comunes del EDA

- Generar preguntas acerca de los datos (o incluso antes de obtenerlos). Preguntas usuales sobre su variabilidad, concentración, distribución. Sobre relaciones entre conjuntos de variables.
- Buscar respuestas ya sea por **visualización**, transformación y modelación de los datos.
- Se pueden considerar técnicas para visualizar datos univariados, en pares y multivariados.
- Usar lo que se aprendió en el paso anterior para refinar las preguntas y/o generar nuevas preguntas.

Sobre grandes bases de datos

- Las técnicas tradicionales estadísticas usualmente se enfocaban en conjuntos pequeños de datos, pero ahora hay bases de datos muy grandes o con cantidades enormes de información. Las técnicas de análisis y visualización se han ido adaptando a cada contexto según la necesidad.
- Es importante tomar en cuenta que cuando se tienen grandes bases de datos, no es necesario todo el tiempo tener que usar todos los datos *a la vez*. Se pueden tomar subconjuntos, siempre que el muestreo de los elementos del grupo sean elegidos al azar, y que todos los conjuntos tengan la misma probabilidad de selección.

Datos ENIGH 2020

- Consideraremos los datos de la Encuesta Nacional de Ingreso y Gasto de los Hogares (ENIGH) del año 2020 para mostrar las diferentes técnicas tanto numéricas como gráficas.
- El objetivo de la ENIGH es proporcionar un panorama estadístico del comportamiento de los ingresos y gastos de los hogares tanto en monto, procedencia y distribución. También ofrece información sobre las características ocupacionales y sociodemográficas de los integrantes del hogar.
- El tamaño de muestra de la ENIGH 2020 fue de 105,483 viviendas que se visitaron del 21 de agosto al 28 de noviembre de 2020.
- Los datos y mayor información de la encuesta se puede encontrar en [el micrositio ENIGH 2020](#).
- Se utilizará la tabla resumen con información a nivel de los hogares que se llama CONCENTRADHOGAR. La descripción de las variables se encuentra en el archivo [Descripción de Base de Datos](#).
- Nota: Para hacer un análisis preciso de estos datos es necesario tomar en cuenta el diseño muestral. Aquí sólo se utilizarán los datos para demostrar los métodos de manera general.

Preparación de los datos I

- Utilizaremos las siguientes variables como parte de nuestro análisis.
 - folioviv: folio de identificación de la vivienda. Los dos primeros dígitos corresponden a la entidad federativa
 - sexo_jefe: Distinción biológica que clasifica al jefe del hogar en hombre (1) o mujer (2)
 - clase_hogar: Diferenciación de los hogares a partir del tipo de relación consanguínea, legal, de afinidad o de costumbre entre el jefe(a) y los otros integrantes del hogar (unipersonal (1), nuclear (2), ampliado (3), compuesto (4), corresidente(5)).
 - edad_jefe: Edad del jefe del hogar.
 - tot_integ: Número de personas integrantes del hogar.
 - ing_cor: Suma de los ingresos por trabajo, los provenientes de rentas, de transferencias, de estimación del alquiler y de otros ingresos.
 - gasto_mon: Gasto monetario.
 - pago_tarje: Pago por tarjeta de crédito al banco o casa comercial.
 - deudas: Pago de deudas de los miembros del hogar a la empresa donde trabajan y/o a otras personas o instituciones.

Preparación de los datos II

- Cargamos los datos y hacemos algunas transformaciones apropiadas con las variables que no fueron convertidas correctamente.

```
library(tidyverse)
library(readr)
Entidades<-c("Aguascalientes", "Baja California", "Baja California Sur", "Campeche", "Coahuila de Zaragoza", "Colima",
           "Chiapas", "Chihuahua", "Ciudad de México", "Durango", "Guanajuato", "Guerrero", "Hidalgo", "Jalisco", "México", "Michoacán de Ocampo",
           "Morelos", "Nayarit", "Nuevo León", "Oaxaca", "Puebla", "Querétaro", "Quintana Roo", "San Luis Potosí", "Sinaloa", "Sonora", "Tabasco",
           "Tamaulipas", "Tlaxcala", "Veracruz de Ignacio de la Llave", "Yucatán", "Zacatecas")

temp <- tempfile()
download.file("https://github.com/jvega68/IASI_Data_Viz/blob/master/Datos/19_enigh2020_ns_concentradohogar_csv.zip?raw=TRUE", temp)

Conc <- read_csv(temp, show_col_types = F) # en Microdatos
```

- A continuación hacemos algunas transformaciones de los datos:

- Se toma el subconjunto de variables de interés
- Se agregan los nombres de los estados de la República Mexicana, de acuerdo a los dos primeros dígitos de la variable folioviv.
- Se convierten a factor las variables sexo y clase_hog.

Preparación de los datos III

```
Conc <- Conc %>%
  # El folio de la vivienda contiene el estado y el municipio
  select(folioviv, sexo_jefe, clase_hog, edad_jefe, tot_integ, ing_cor, gasto_mon, pago_tarje, deudas) %>%
  mutate(entidad = Entidades[as.numeric(substr(folioviv,1,2))],
         sexo_jefe = factor(sexo_jefe, levels = c(1,2), labels = c("H","M")),
         clase_hog = as.factor(clase_hog))
head(Conc) # también se puede usar tail(Conc) para ver los últimos datos

# A tibble: 6 x 10
  folioviv sexo_jefe clase_hog edad_jefe tot_integ ing_cor gasto_mon pago_tarje deudas entidad
  <chr>     <fct>    <fct>      <dbl>     <dbl>     <dbl>     <dbl>     <dbl> <chr>
1 0100013605 M        2          48        3   16229.   24626.       0     0 Aguascalientes
2 0100013606 H        2          46        4   31426.   20397.       0   2361. Aguascalientes
3 0100017801 H        2          26        2   33979.   44956.       0     0 Aguascalientes
4 0100017802 H        2          29        2   71557.   82950.       0     0 Aguascalientes
5 0100017803 H        2          63        2   90703.   30141.       0     0 Aguascalientes
6 0100017804 H        2          33        4   30369.   39992.       0     0 Aguascalientes
```

Descripción general del conjunto de datos I

- Como resumen general podemos usar la función `summary` como resumen de cada variable y algunas estadísticas básicas, o `str` para conocer la estructura del archivo de datos.

```
summary(Conc) # resumen más básico de datos

  folioviv      sexo_jefe clase_hog   edad_jefe      tot_integ      ing_cor      gasto_mon      pago_tarje
Length:89006    H:63230   1:10842   Min.   : 14.00   Min.   : 1.000   Min.   :     0   Min.   :     0   Min.   :     0
Class :character M:25776   2:55339   1st Qu.: 39.00   1st Qu.: 2.000   1st Qu.: 21392   1st Qu.: 13875   1st Qu.:     0
Mode  :character            3:21819   Median : 50.00   Median : 3.000   Median : 35172   Median : 22106   Median :     0
                           4: 717    Mean   : 51.09   Mean   : 3.546   Mean   : 47838   Mean   : 28229   Mean   : 754
                           5: 289    3rd Qu.: 62.00   3rd Qu.: 5.000   3rd Qu.: 57640   3rd Qu.: 34466   3rd Qu.:     0
                           Max.   :107.00   Max.   :25.000   Max.   :10702107  Max.   :1007112   Max.   :440217

  deudas      entidad
Min.   : 0.0 Length:89006
1st Qu.: 0.0 Class :character
Median : 0.0 Mode  :character
Mean   : 507.6
3rd Qu.: 0.0
Max.   :377213.1

str(Conc)      # muestra la estructura de los datos

#> #> #> tibble [89,006 x 10] (S3:tbl_df/tbl/data.frame)
#> #> #> $ folioviv : chr [1:89006] "0100013605" "0100013606" "0100017801" "0100017802" ...
#> #> #> $ sexo_jefe : Factor w/ 2 levels "H","M": 2 1 1 1 1 1 1 1 1 ...
#> #> #> $ clase_hog : Factor w/ 5 levels "1","2","3","4",...: 2 2 2 2 2 2 2 2 2 ...
#> #> #> $ edad_jefe : num [1:89006] 48 46 26 29 63 33 60 76 74 37 ...
#> #> #> $ tot_integ : num [1:89006] 3 4 2 2 2 4 3 2 2 6 ...
#> #> #> $ ing_cor   : num [1:89006] 16229 31426 33979 71557 90703 ...
#> #> #> $ gasto_mon: num [1:89006] 24626 20397 44956 82950 30141 ...
#> #> #> $ pago_tarje: num [1:89006] 0 0 0 0 0 0 0 0 0 ...
#> #> #> $ deudas    : num [1:89006] 0 2361 0 0 0 ...
#> #> #> $ entidad   : chr [1:89006] "Aguascalientes" "Aguascalientes" "Aguascalientes" ...
```

Descripción general del conjunto de datos II

- También podemos usar skimr

```
library(skimr)
kableExtra::kable(kable(format = "latex") %>%
kableExtra::kable_styling(latex_options = "scale_down"))
```

- Más adelante veremos otras funciones para obtener estadísticas sumarias de los datos.

Preguntas guía para el análisis exploratorio

- ¿Hay más jefas o más jefes de familia? ¿Cuál es la proporción de mujeres?
- ¿En qué entidad se da el mayor/menor ingreso corriente? ¿el mayor/menor gasto?
- ¿Cuál es el ingreso promedio a nivel nacional? ¿A nivel estatal? ¿Cuál es su variación?
- ¿Cuál es la moda de categorías del hogar?
- ¿Cómo se relacionan el gasto corriente y el ingreso corriente a nivel estatal?
- ¿Hay alguna relación entre el gasto corriente y la edad de los jefes de familia? ¿el sexo del jefe de familia?

Importancia de la visualización

The greatest value of a picture is when it forces us to notice what we never expected to see

John Tukey (1977)

Graphs provide powerful tools both for analyzing scientific data and for communicating quantitative information

William Cleveland (1985)

```
data(anscombe)
coef(lm(y1~x1, data = anscombe))

(Intercept)      x1
 3.0000909    0.50000909

coef(lm(y2~x2, data = anscombe))

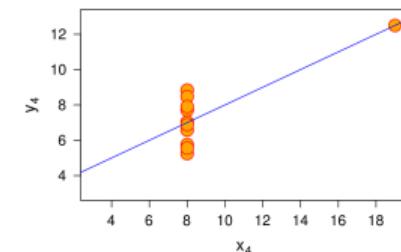
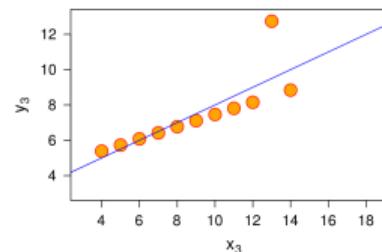
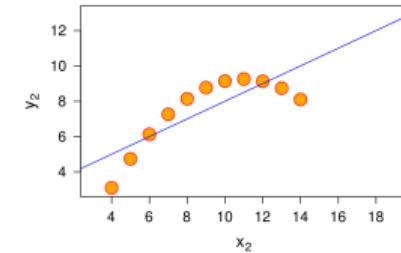
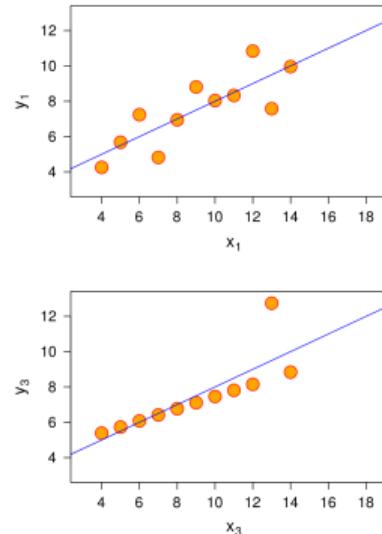
(Intercept)      x2
 3.000909     0.5000000

coef(lm(y3~x3, data = anscombe))

(Intercept)      x3
 3.0024545    0.4997273

coef(lm(y4~x4, data = anscombe))

(Intercept)      x4
 3.0017273    0.4999091
```



- ¿Cómo se usan las gráficas en el análisis de datos?
 - Comprender propiedades de los datos
 - Encontrar patrones en los datos
 - Sugerir estrategias de modelado
 - Validar el análisis
 - Comunicar resultados
- Las gráficas pueden ser de dos tipos: exploratorias o analíticas y de presentación o comunicación. Las características de las gráficas exploratorias incluyen:
 - Se hacen de manera rápida y/o interactiva
 - Se tiende a hacer muchas para comprender diversos aspectos de los datos.
 - El objetivo es lograr un entendimiento personal de la información: cuáles son los datos, cómo se ven, qué problemas pueden tener. Aquí conviene tener preguntas para guiar el análisis.
 - Diferir el uso de leyendas y títulos (no son para presentación)
 - El color y el tamaño, así como otros atributos, se usan para incorporar información y no por decoración.

Principios de gráficas analíticas (Roger Peng/Edward Tufte)

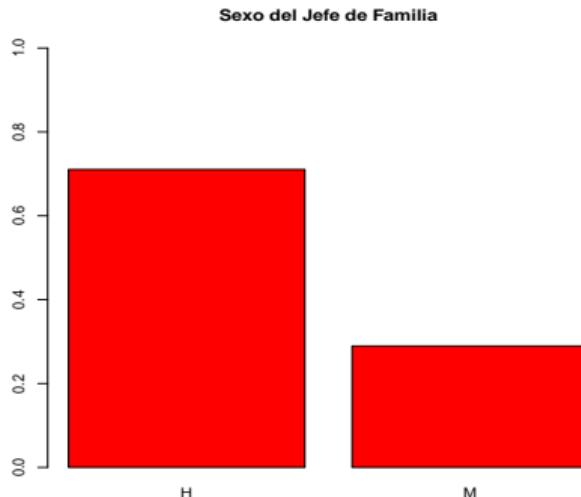
- ① Mostrar comparaciones entre grupos
 - Evidencia para una hipótesis es siempre relativa a otra hipótesis competitiva.
 - Siempre hay que preguntar ¿comparado a qué?
- ② Mostrar causalidad, mecanismo, explicación, estructura sistemática. ¿Cuál es el marco causal para pensar acerca del problema?
- ③ Mostrar datos multivariados: tratar de incorporar varias dimensiones al problema.
- ④ Integrar múltiples modos de evidencia
- ⑤ Describir y documentar la evidencia. Una gráfica debe decir una historia lo más completa posible y que sea creíble.
- ⑥ El contenido es rey: ¿Cuál es la historia que se quiere contar? Si no hay historia, la gráfica no sirve.
 - Las presentaciones analíticas dependen de su calidad, relevancia e integridad del contenido.

Visualizando datos univariados I

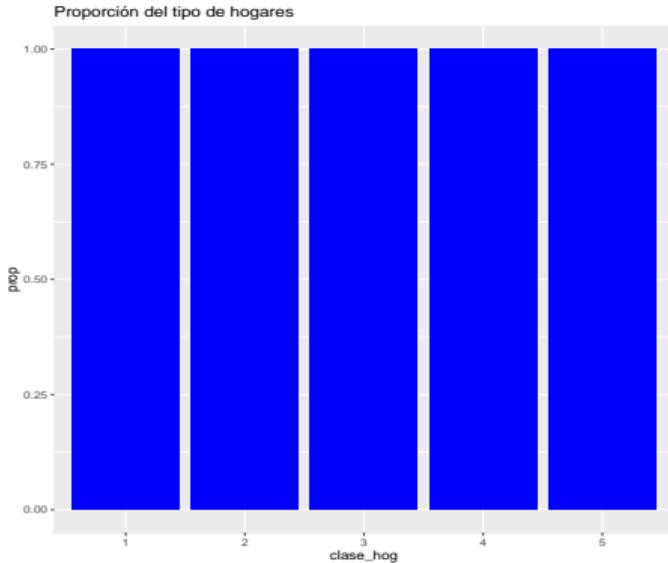
- Hay distintas visualizaciones que nos ayudan a entender el comportamiento para distintos tipos de variables. Cada tipo de gráfica enfatiza diferentes características de la información:
 - Para datos cuantitativos continuos o discretos: boxplots, histogramas, densidades, dotplots
 - Para datos cuantitativos indexados por tiempo (o algún índice): series de tiempo
 - Para datos cualitativos o categóricos: gráficas de barras (frecuencias)

Datos categóricos: Gráficas de barras

```
# Versión base plot  
barplot(height = table(Conc$sexo_jefe)/nrow(Conc), ylim = c(0,1), col = "red",  
        main = "Sexo del Jefe de Familia")
```



```
# Versión ggplot  
Conc %>% ggplot(aes(x = clase_hog, y = after_stat(prop))) +  
  geom_bar(fill = "blue") +  
  labs(title = "Proporción del tipo de hogares")
```



```
table(Conc$sexo_jefe)
```

Sexo	Cantidad
H	63230
M	25776

```
table(Conc$clase_hog)
```

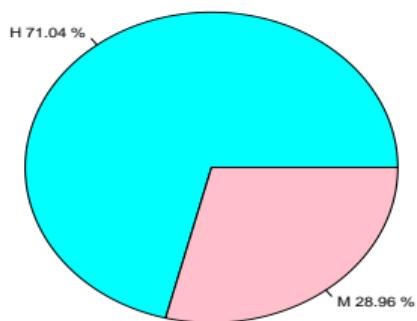
Tipo de Hogar	Cantidad
1	10842
2	55339
3	21819
4	717
5	289

Datos categóricos: Pies (NO recomendado) I

- Hay varias razones por las que no se recomienda hacer gráficas de pie, a menos que se tengan muy pocas categorías.

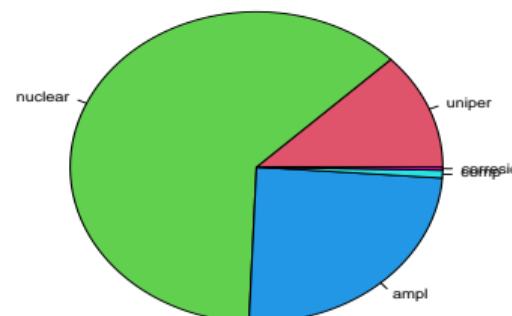
```
pct <- round(100*table(Conc$sexo_jefe)/length(Conc$sexo_jefe),2)
pie(x = table(Conc$sexo_jefe),
  labels = paste(c("H","M"),pct,"%"),
  main = "Gráfica de pie (no recomendada)",
  col = c("cyan","pink"))
```

Gráfica de pie (no recomendada)



```
pie(x = table(Conc$clase_hog),
  labels = c("uniper","nuclear","ampl","comp","corresid"),
  main = "Distribución de los tipos de hogar",
  col = 2:6)
```

Distribución de los tipos de hogar



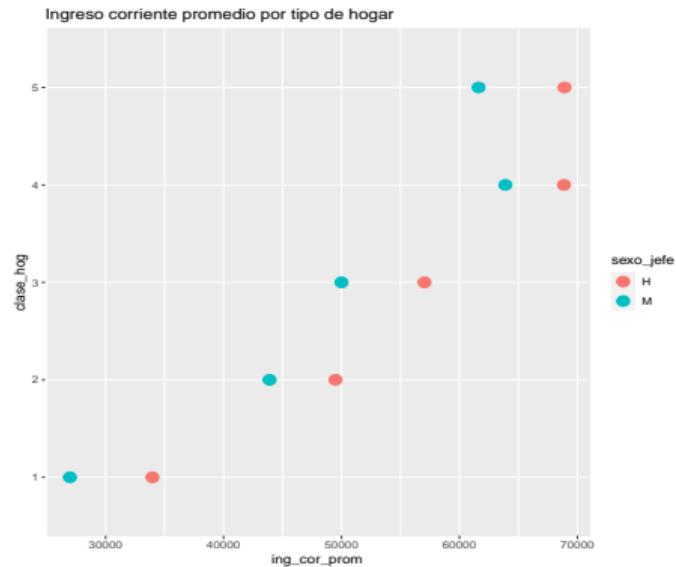
Datos categóricos: Pies (NO recomendado) II

- Las gráficas de pie son malas porque los humanos somos malos para interpretar ángulos correctamente.
- En lugar de gráficas de pie, se recomienda considerar gráficas de barras o gráficas de puntos como la siguiente:

```
A <- Conc %>%
  group_by(clase_hog, sexo_jefe) %>%
  summarise(ing_cor_prom = mean(ing_cor), .groups = "drop")

A %>% ggplot(aes(x = clase_hog, y = ing_cor_prom, color = sexo_jefe)) +
  geom_point(size = 4) +
  coord_flip() +
  labs(title = "Ingreso corriente promedio por tipo de hogar")
```

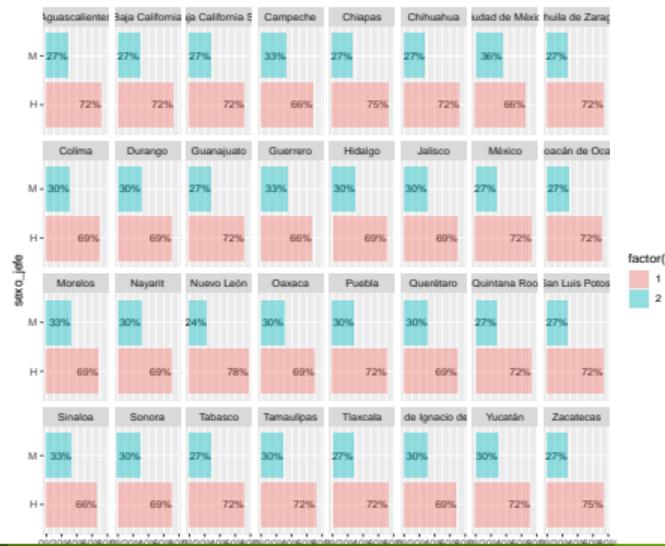
Datos categóricos: Pies (NO recomendado) III



Gráficas de barras: comparando resultados por entidad

Conc %>%

```
group_by(entidad) %>%
  ggplot(aes(x = sexo_jefe,
             y = ..prop..,
             fill = factor(..x..),
             group = entidad), stat = "count") +
  geom_text(aes(label = scales::percent(..prop..,accuracy = 3),
                y= ..prop.. ), stat= "count", hjust = 1, size = 3) +
  geom_bar(alpha = 0.4) +
  facet_wrap(~ entidad, nrow = 4) +
  coord_flip() +
  scale_y_continuous(labels = scales::percent) +
  ylab("Frecuencia Relativa")
```



- Supongamos que X denota una variable aleatoria. Para un número $0 \leq p \leq 1$, el p -ésimo cuantil de X es el valor q_p tal que $P(X \leq q_p) = p$. Es decir, q_p es el valor tal que la cantidad de área bajo la curva de densidad (de X) a la izquierda de q_p es p .
- Un término relacionado es *percentil*, que divide las porciones de área en 100 partes. El cuantil 0.3 es el percentil 30 %. Los *cuartiles* dividen el área en 4 partes proporcionales (25 %, 50 %, 75 %). Los *deciles* dividen el área en 10 partes iguales, etc.
- Por ejemplo, los cuartiles de la distribución normal estándar son

```
qnorm(c(0.25,0.5,0.75))  
[1] -0.6744898  0.0000000  0.6744898
```

Boxplots I

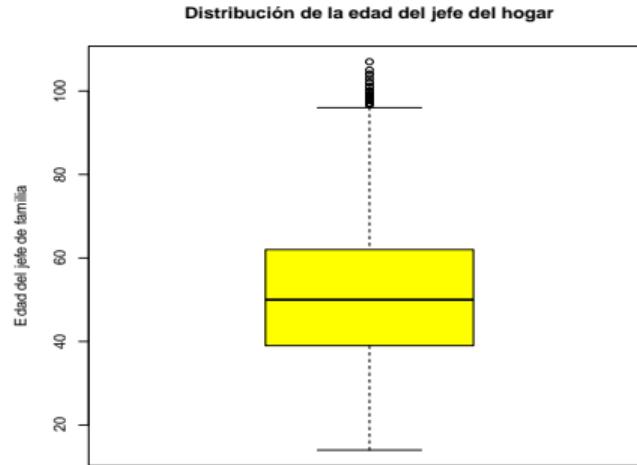
- Un boxplot o gráfica de caja y brazos es una versión estilizada de una densidad en donde se enfatizan 5 medidas resumen de los datos:
 - 1 valor mínimo
 - 2 los cuartiles $q_1 = 25\%$, $q_2 = 50\%$ (mediana), $q_3 = 75\%$
 - 3 valor máximo
- Adicionalmente representa los valores extremos ($q_1 - 1.5IQR$ y $q_3 + 1.5IQR$ donde IQR es el rango intercuartil, que se define como $IQR = q_3 - q_1$)

Valores atípicos

Una observación se considera un valor extremo o atípico (*outlier*) si está más lejos que $1.5IQR$ del cuartil más cercano. Decimos que el valor atípico es extremo si está a más de $3IQR$ del cuartil más cercano.

Boxplots II

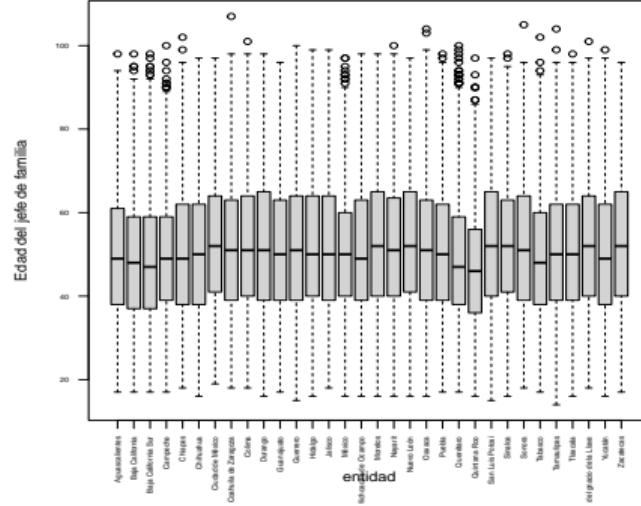
```
boxplot(Conc$edad_jefe,  
       col = "yellow",  
       main = "Distribución de la edad del jefe del hogar",  
       ylab = "Edad del jefe de familia")
```



Boxplots III

- Los boxplots son particularmente útiles para comparar diferentes poblaciones de manera muy general.

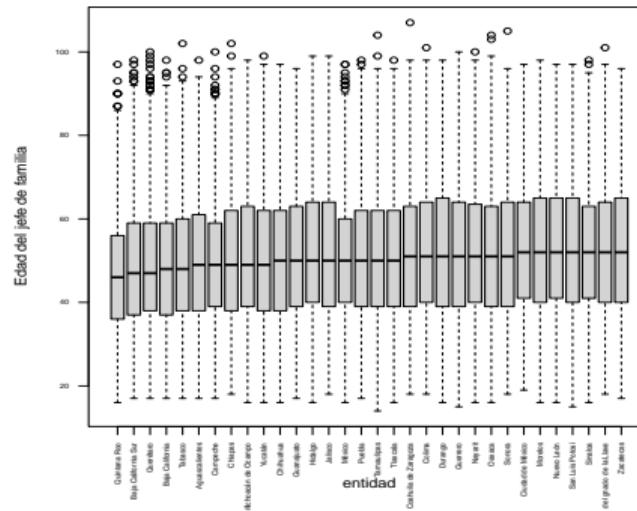
```
boxplot(edad_jefe ~ entidad,
        data = Conc,
        las = 2,
        cex.axis = 0.5,
        ylab = "Edad del jefe de familia")
```



Boxplots IV

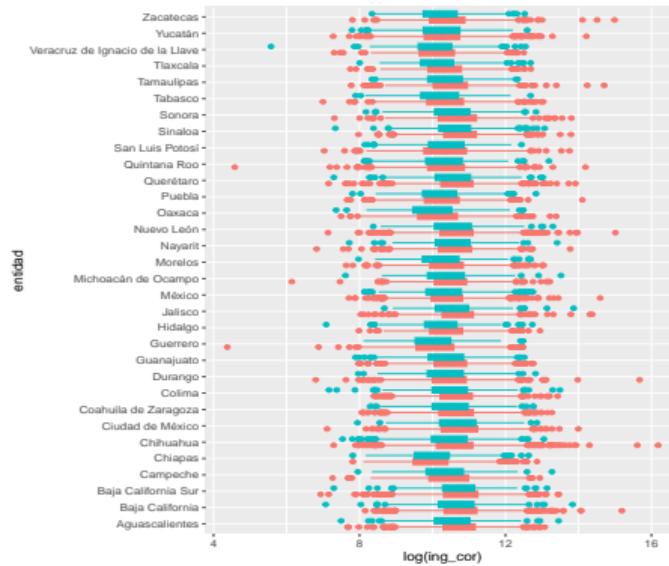
- A veces es mejor ordenar las categorías para tener una mejor apreciación de los datos.

```
orden <- with(Conc, reorder(entidad,edad_jefe,median))
boxplot(Conc$edad_jefe ~ orden, las = 2, cex.axis = 0.5, xlab = "entidad", ylab = "Edad del jefe de familia")
```



Comparación de subpoblaciones de hombres y mujeres por estado I

```
Conc %>%
  group_by(entidad) %>%
  ggplot(aes(y = entidad, x = log(ing_cor))) +
  geom_boxplot(aes(fill = sexo_jefe, color = sexo_jefe), position = "dodge") +
  theme(legend.position = "none")
```



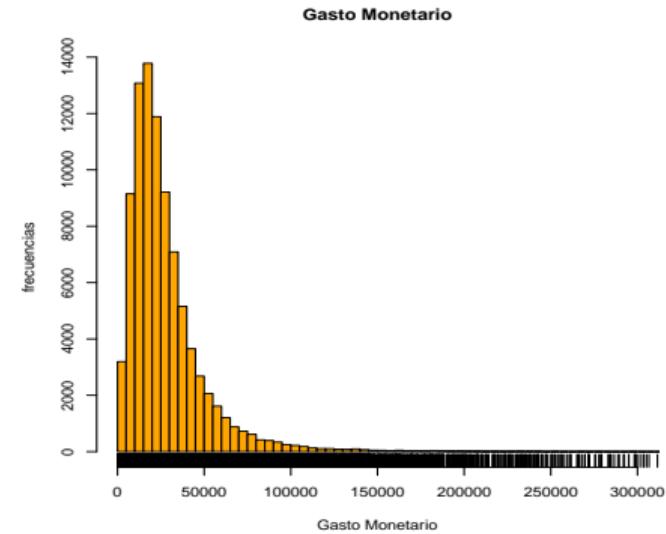
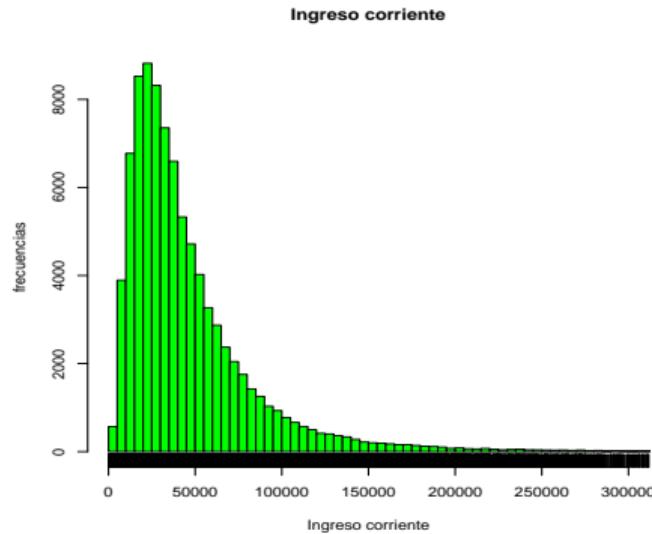
Histogramas I

- Un histograma es una gráfica de la distribución de la frecuencia o frecuencia relativa. Cada frecuencia (relativa) es representada por un rectángulo sobre el correspondiente valor (o rango de valores) y el área del rectángulo es proporcional a la frecuencia (relativa) correspondiente.
- El histograma es útil para datos numéricos.
- Características que resaltan en un histograma:
 - Centro o valor típico.
 - Extensión o variabilidad de los datos
 - Forma general
 - Localización y número de picos (modas)
 - Presencia de huecos y posibles valores atípicos.

Histogramas II

```
hist(Conc$ing_cor, col = "green", breaks = 2000,
  main = "Ingreso corriente",
  xlab = "Ingreso corriente",
  ylab = "frecuencias",
  xlim = c(0,300000))
rug(Conc$ing_cor)
```

```
hist(Conc$gasto_mon, col = "orange", breaks = 300,
  main = "Gasto Monetario",
  xlab = "Gasto Monetario",
  ylab = "frecuencias",
  xlim = c(0,300000))
rug(Conc$gasto_mon)
```

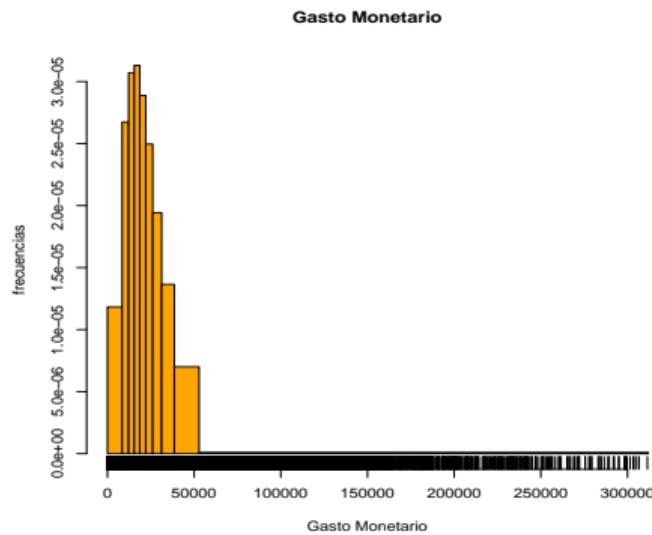


- La frecuencia relativa también se conoce como *densidad*. Es común usar las densidades cuando los intervalos del histograma no son iguales.
- Algunas características adicionales de los histogramas:
 - Unimodal (un pico), bimodal (dos picos), multimodal (varios picos)
 - Simétrico, sesgado positivamente o sesgado a la derecha, si tiene cola positiva larga. Sesgado negativamente o a la izquierda si tiene cola negativa larga.
 - Colas anchas o colas delgadas.

Histogramas con diferentes longitudes de clase

- Es posible tener histogramas con diferentes longitudes de clase, cuando los grupos en los que se dividen los datos no son uniformes. Por ejemplo, el siguiente histograma usa los deciles de la distribución

```
hist(Conc$gasto_mon, col = "orange", breaks = quantile(Conc$gasto_mon,seq(0,1,by=0.1)), prob = T,
     main = "Gasto Monetario",
     xlab = "Gasto Monetario",
     ylab = "frecuencias",
     xlim = c(0,300000))
rug(Conc$gasto_mon)
```

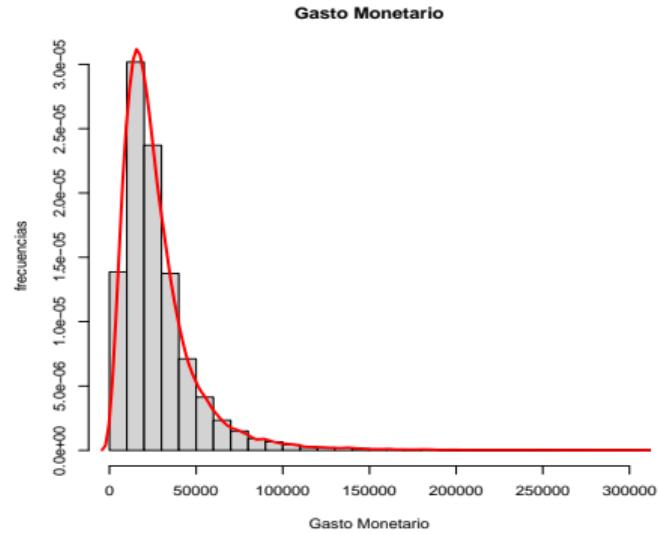


Densidades I

- Para el caso de datos cuantitativos en escala continua, una distribución de probabilidad se especifica por una curva que se llama la *curva de densidad*. La función $f(x)$ que define esa curva se llama función de densidad.
- La función de densidad puede considerarse como el límite de un histograma, cuando los rectángulos se hacen más y más pequeños.
- Propiedades de la función de densidad:
 - $f(x) \geq 0$ para todos los valores de la variable aleatoria X .
 - El área total bajo la densidad es igual a 1: $\int f(x)dx = 1$

```
hist(Conc$gasto_mon,prob=T,breaks = 100,
  main = "Gasto Monetario",
  xlab = "Gasto Monetario",
  ylab = "frecuencias",
  xlim = c(0,300000))
lines(density(Conc$gasto_mon), lwd = 3, col ="red")
```

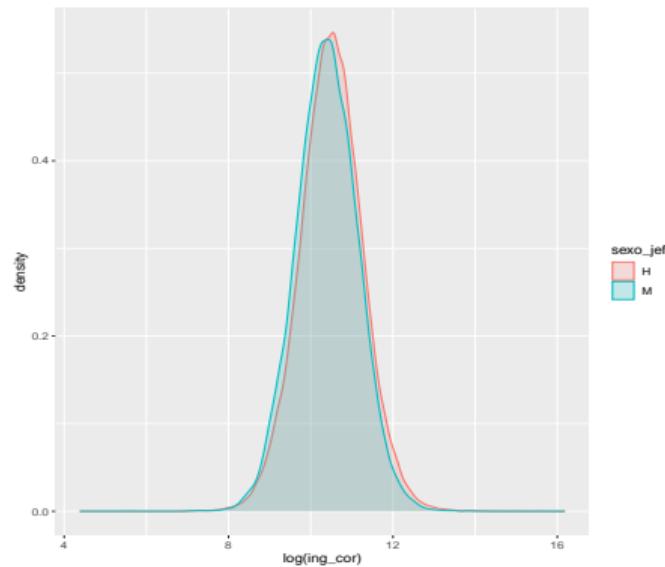
Densidades II



Comparación de Subpoblaciones con Densidades I

- Un ejercicio interesante es comparar las densidades de la población de hombres con la de mujeres

```
Conc %>%
  group_by(sexo_jefe) %>%
  ggplot(aes(x = log(ing_cor), col = sexo_jefe, fill = sexo_jefe)) +
  geom_density(alpha = 0.2)
```

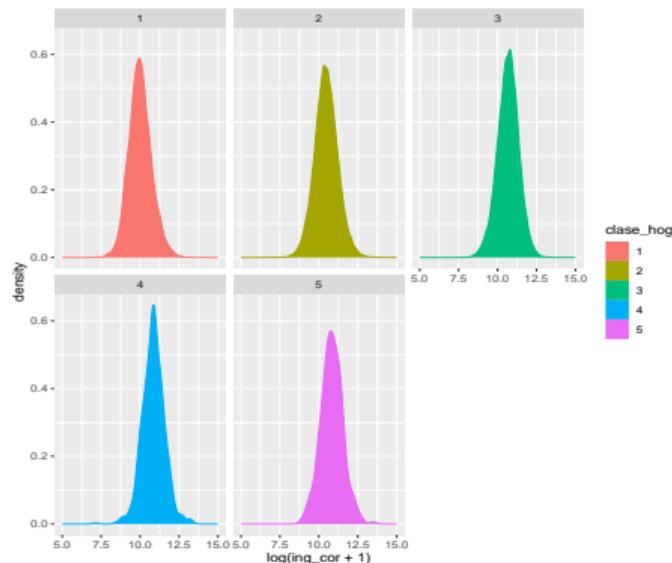


Comparación de Subpoblaciones con Densidades II

- Otro ejercicio usando el tipo de hogar

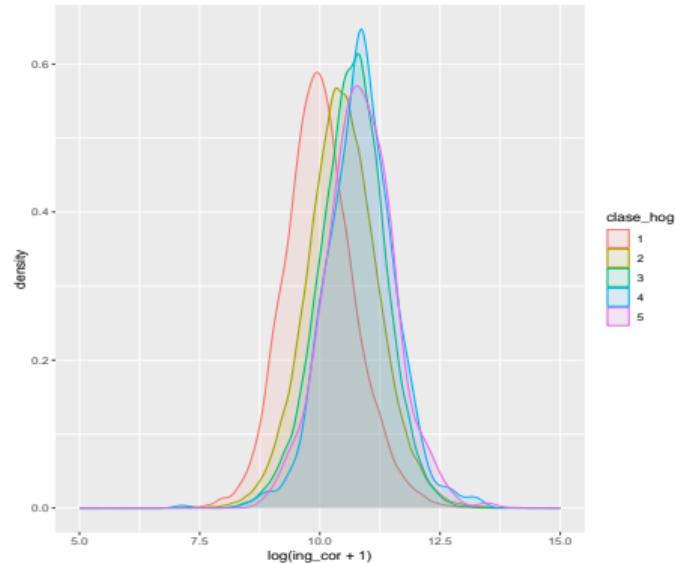
Conc %>%

```
group_by(clase_hog) %>%
  ggplot(aes(x = log(ing_cor + 1), col = clase_hog, fill = clase_hog)) +
  geom_density() +
  lims(x = c(5,15)) +
  facet_wrap(~ clase_hog)
```



Conc %>%

```
group_by(clase_hog) %>%
  ggplot(aes(x = log(ing_cor + 1), col = clase_hog, fill = clase_hog)) +
  geom_density(alpha = 0.1) +
  lims(x = c(5,15))
```

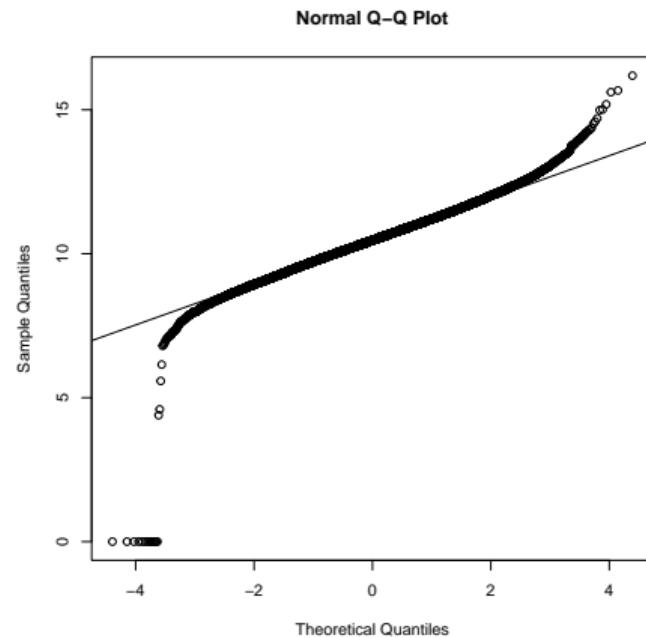


Cuantiles y qq-plots I

- Una gráfica que nos permite comparar los datos de dos distribuciones a través de sus cuantiles se llama qq-plot o gráfica cuantil-cuantil.
- Usamos los cuantiles para crear una gráfica que es útil para juzgar si los datos siguen una distribución normal. Una gráfica qq-plot es una gráfica con puntos $(x_{(1)}, q_1), (x_{(2)}, q_2), \dots, (x_{(n)}, q_n)$ donde q_k es el cuantil $k/(n + 1)$ de la distribución normal estándar y $x_{(1)} \leq \dots \leq x_{(n)}$ son los valores ordenados de la muestra.
- Si los puntos caen cercanos a una línea recta, podemos decir que los datos provienen de una distribución normal.

```
qqnorm(log(Conc$ing_cor + 1))
qqline(log(Conc$ing_cor + 1)) # agrega una línea entre el primer y tercer cuartil de los datos.
```

Cuantiles y qq-plots II



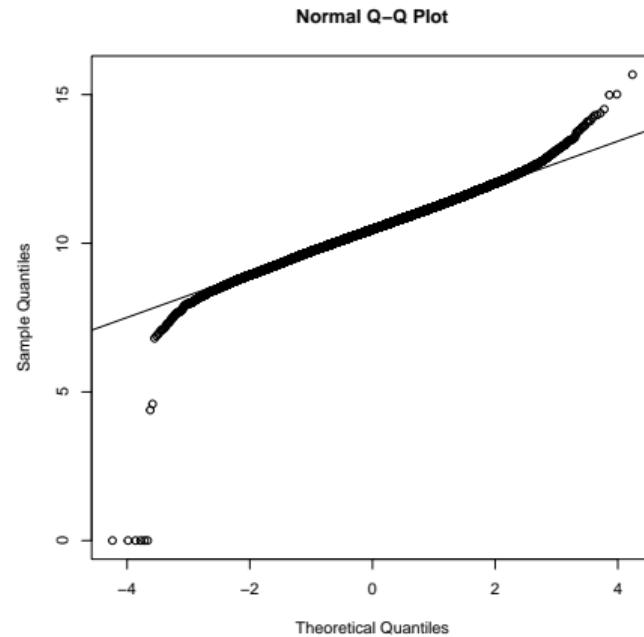
Vemos que los datos de las colas no caen sobre la línea recta, por lo tanto, los datos no se apegan a una distribución normal.

Cuantiles y qq-plots III

- Tomamos una submuestra de 50 % de los datos y repetimos la gráfica para ver si se conserva la conclusión

```
ind <- sample(1:nrow(Conc), round(0.50*nrow(Conc),0),replace=F)
qqnorm(log(Conc$ing_cor[ind] + 1))
qqline(log(Conc$ing_cor[ind] + 1)) # agrega una línea entre el primer y tercer cuartil de los datos.
```

Cuantiles y qq-plots IV



- Los qq-plots pueden ayudar a ver mejor si las colas de una distribución son más anchas o planas que las de la distribución normal.

Distribuciones acumulativas I

- Con datos numéricos es común tratar de entender cómo se ve la proporción de valores que están por encima o por abajo de un valor dado x . Estos valores se pueden representar como

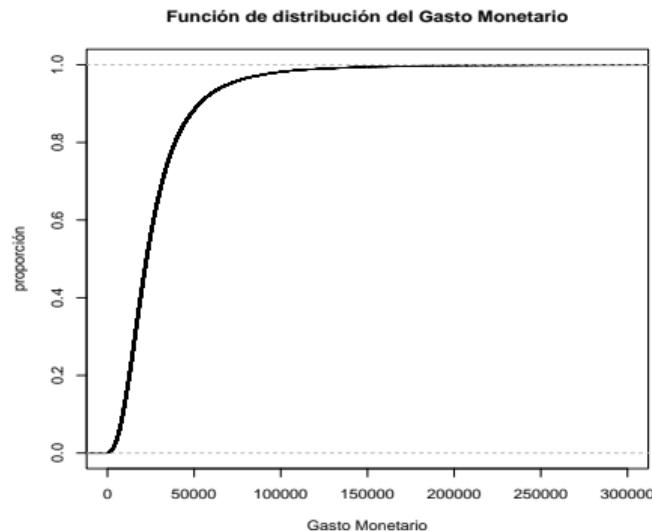
$$F(x) = P(X \leq x) \text{ de tal manera que } P(X > x) = 1 - F(x)$$

A $F(x)$ se le conoce como función de distribución acumulativa.

- Usualmente la función de distribución que parte de los datos, se le llama *función de distribución empírica*, y a la versión que se determina matemáticamente se le llama simplemente función de distribución.

```
plot(ecdf(Conc$gasto_mon),
      main = "Función de distribución del Gasto Monetario",
      xlab = "Gasto Monetario",
      ylab = "proporción",
      xlim = c(0,300000))
```

Distribuciones acumulativas II



- La función de distribución empírica (ecdf) es un estimador de la función de distribución acumulativa para una muestra. La función de distribución empírica, denotada por \hat{F} es una función de la forma:

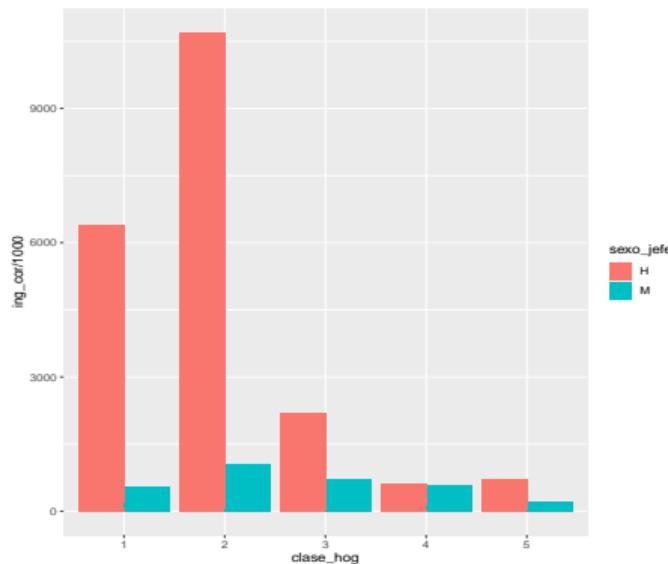
$$F_n(x) = \frac{\#(X_i \leq x)}{n} = \frac{\sum_{i=1}^n I_{(-\infty, x]}(X_i)}{n}$$

- Las gráficas anteriores ya son de alguna forma gráficas multivariadas: podemos ver que mezclando una variable continua con variables categóricas, podemos representar los datos *condicionales* a los niveles de esa variable categórica.
- En términos generales, una variable categórica define grupos de datos. En nuestra base de datos, las variables categóricas que tenemos son:
 - folioviv: esta variable es de tipo identificadora, y forma grupos de un sólo dato
 - sexo_jefe: esta variable es dicotómica, sólo toma dos valores
 - clase_hog: Esta variable, aunque usa números es categórica, y divide los datos en 5 grupos, del 1 al 5.

Gráficas de variables categóricas I

- Podemos usar dos variables categóricas y una variable continua con gráficas de barras.

```
Conc %>%
  ggplot(aes(x=clase_hog, y = ing_cor/1e3, fill = sexo_jefe)) +
  geom_bar(stat = "identity", position = position_dodge())
```



Gráficas de variables categóricas II

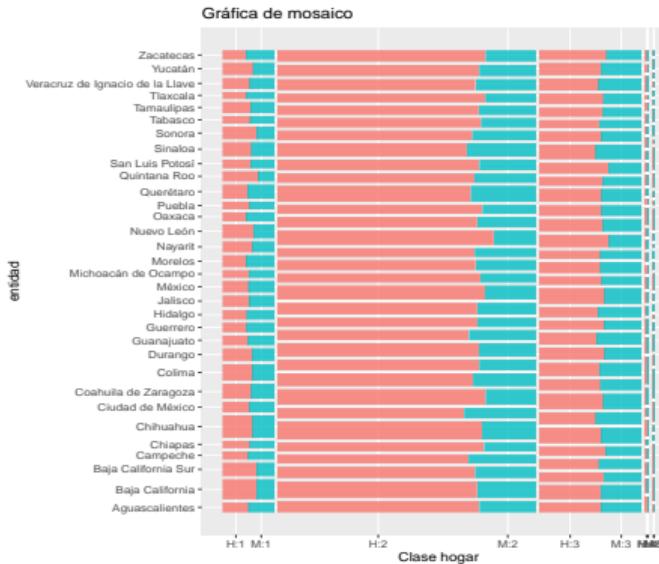
- Podemos graficar las frecuencias cruzadas de variables categóricas generando *tablas de contingencia*

```
A <- table(Conc$entidad,Conc$clase_hog) # tabla de contingencia  
head(A,20)
```

	1	2	3	4	5
Aguascalientes	293	1750	587	33	6
Baja California	622	2611	859	30	20
Baja California Sur	408	1673	585	25	26
Campeche	239	1341	568	22	4
Chiapas	198	1346	564	10	5
Chihuahua	702	2862	958	33	17
Ciudad de México	299	1472	758	16	25
Coahuila de Zaragoza	456	2470	951	36	9
Colima	526	1997	686	59	14
Durango	370	1565	771	33	7
Guanajuato	292	2017	740	31	3
Guerrero	273	1494	700	18	5
Hidalgo	287	1345	554	21	6
Jalisco	326	1805	610	24	14
México	345	2237	964	16	6
Michoacán de Ocampo	208	1321	493	19	6
Morelos	353	1533	658	18	2
Nayarit	314	1287	480	16	6
Nuevo León	431	2228	789	36	18
Oaxaca	253	1590	741	9	3

Gráficas de variables categóricas III

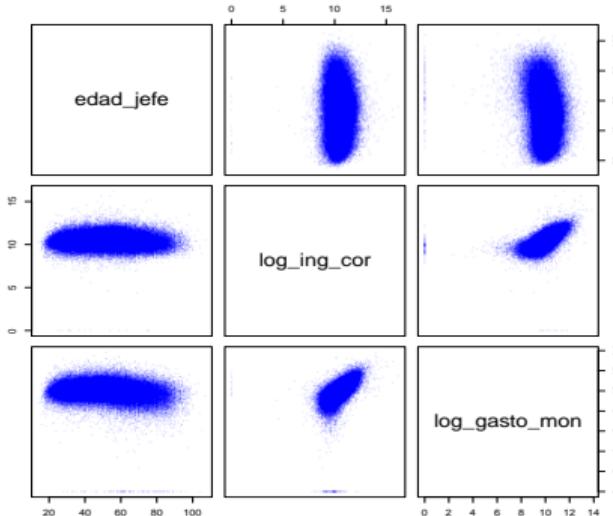
```
library(ggmosaic)
Conc %>%
  ggplot() +
  geom_mosaic(aes(x = product(entidad,clase_hog), fill = sexo_jefe)) +
  labs(x = "Clase hogar", y = "entidad", title = "Gráfica de mosaico") +
  theme(legend.position = "none")
```



Relaciones bivariadas (scatterplots) I

- Podemos graficar las variables por pares, para identificar patrones y tendencias generales

```
Conc %>%
  select(ing_cor,gasto_mon, edad_jefe) %>%
  mutate(log_ing_cor = log(ing_cor + 1),
    log_gasto_mon = log(gasto_mon+1)) %>%
  select(-ing_cor, -gasto_mon) %>%
  pairs(pch = 16, cex = 0.4, col = alpha("blue",0.1))
```

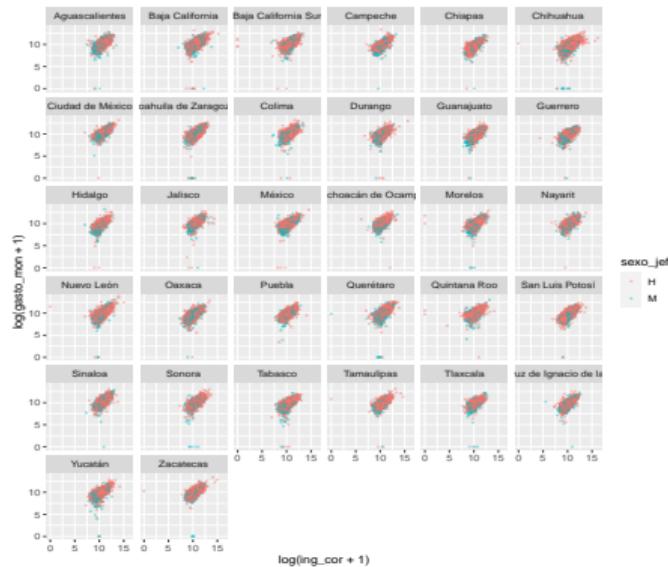


Relaciones bivariadas (scatterplots) II

- Podemos condicionar con respecto a una de las variables categóricas, poniendo todas las observaciones en una misma gráfica o separándolas, según se facilite la interpretación.

Conc %>%

```
group_by(entidad) %>%
  ggplot(aes(x = log(ing_cor + 1), y = log(gasto_mon+1), col = sexo_jefe)) +
  geom_point(size = 0.2, alpha = 0.4) +
  facet_wrap(~ entidad)
```



Caras de Chernoff I

- Las caras de Chernoff asocian a una dimensión de un dibujo una variable. Su principal intención es identificar patrones, y agrupar las variables, aunque puede ser muy difícil de decodificar.
- No funciona para muchas observaciones, usualmente el conjunto debe acotarse a un número pequeño de observaciones. Por ejemplo, podemos considerar este tipo de gráfica para mediciones de los diferentes estados

```
X <- Conc %>%
  select(-folioviv, sexo_jefe, clase_hog) %>%
  group_by(entidad) %>%
  summarize(ingcor = mean(ing_cor),
            gascor = mean(gasto_mon),
            integ_prom = mean(tot_integ),
            ppromtc = mean(pago_tarje),
            deudaprom = mean(deudas),
            edadprom = mean(edad_jefe)) %>%
  as.data.frame()
```

Caras de Chernoff II

- Los datos ahora se pueden ver en una tabla resumen

```
head(X)

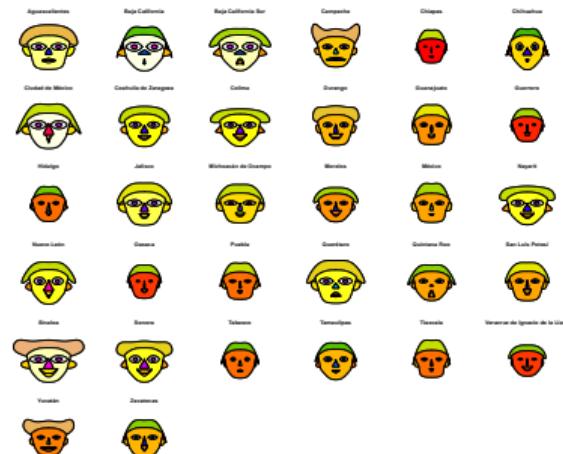
  entidad    ingcor    gascor integ_prom    ppromtc deudaprom edadprom
1 Aguascalientes 55597.85 33592.56 3.763582 1007.3213 655.3173 49.94268
2 Baja California 62025.77 34208.02 3.361420 775.4446 146.6195 48.58112
3 Baja California Sur 61035.48 33859.03 3.266470 1188.5807 550.6374 48.39750
4 Campeche 46077.36 28073.06 3.579117 786.2206 1196.2183 49.67157
5 Chiapas 29010.53 19428.12 3.897786 165.0607 326.7976 50.26189
6 Chihuahua 57482.27 25271.96 3.282808 506.6191 190.6797 50.61899

rownames(X) <- X$entidad # nombra a los renglones con los nombres de las entidades
X$entidad <- NULL #quitamos la variable con los nombres de las entidades
```

- Se ilustran las caras de Chernoff con los promedios de las entidades

```
suppressMessages(library(aplpack))
par(mar = c(1,1,1,1))
faces(X, face.type = 1, cex = 0.7)
```

Caras de Chernoff III

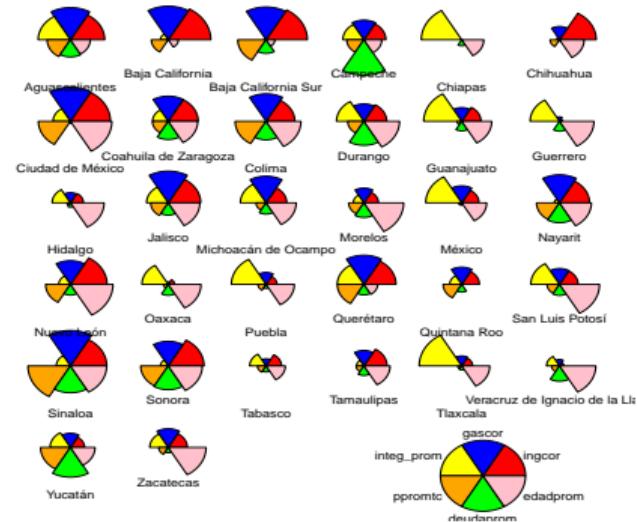


Caras de Chernoff IV

```
effect of variables:  
modified item      Var  
"height of face   " "ingcor"  
"width of face    " "gascor"  
"structure of face" "integ_prom"  
"height of mouth   " "ppromtc"  
"width of mouth    " "deudaprom"  
"smiling           " "edadprom"  
"height of eyes    " "ingcor"  
"width of eyes     " "gascor"  
"height of hair    " "integ_prom"  
"width of hair     " "ppromtc"  
"style of hair     " "deudaprom"  
"height of nose    " "edadprom"  
"width of nose     " "ingcor"  
"width of ear       " "gascor"  
"height of ear     " "integ_prom"
```

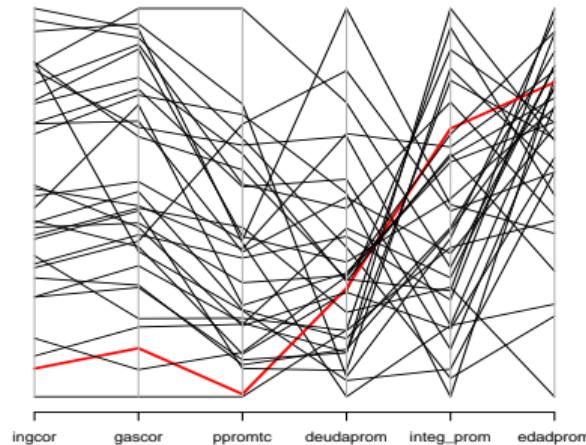
Gráficas de estrella

```
stars(X,key.loc=c(12,1.5),scale = T,  
      col.segments = c("red","blue","yellow","orange","green","pink"),  
      draw.segments =T)
```



Gráficas de coordenadas paralelas

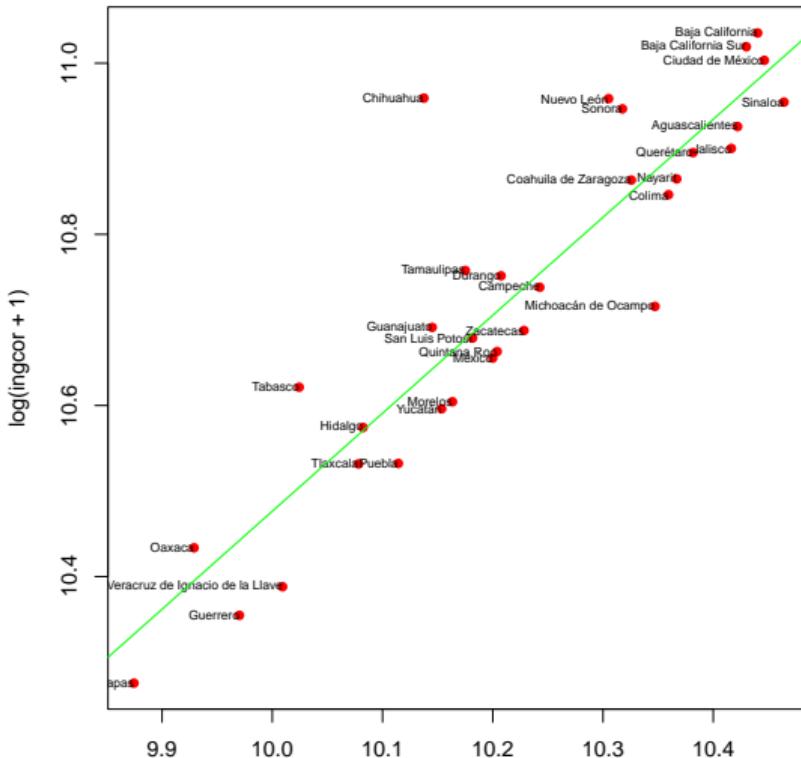
```
library(MASS)
parcoord(X[,c(1,2,4,5,3,6)], # Podemos cambiar el orden de las variables
          col = c(rep(1,11),"red",rep(1,21)),
          lwd = c(rep(1,11),3,rep(1,21)))
```



Relación entre ingreso y gasto I

```
par(mar = c(4,4,4,4))
plot(log(ingcor+1) ~ log(gascor+1), data = X, pch =16, col = "red")
text(log(X$gascor+1), log(X$ingcor+1), labels = rownames(X), cex = 0.6, adj = 1)
abline(lm(log(ingcor+1) ~ log(gascor+1), data =X)$coef, col = "green")
```

Relación entre ingreso y gasto II



Mapas

- Hay muchas maneras de hacer mapas en R. Una de las más fáciles es usar el paquete `mxmaps`.

```
library(mxmaps) # para mapas tipo cloropetas. Ver https://www.diegovalle.net/mxmaps/
X$state_name_official <- rownames(X)
X_map <- full_join(df_mxstate_2020, X, by = "state_name_official") %>%
  dplyr::select(region, state_name_official, deudaprom)
#Correcciones
X_map <- X_map[complete.cases(X_map),] # Considera sólo los casos completos
X_map$value <- X_map$deudaprom

mxhexbin_choropleth(X_map, num_colors = 1, label_size = 3, title = "Deuda promedio") mxstate_choropleth(X_map, num_colors = 1, title = "Deuda promedio") +
  scale_fill_gradient(low = "green", high = "red", guide = "colourbar") +
  scale_fill_gradient(low = "white", high = "blue", guide = "colourbar")
```

Deuda promedio



Deuda promedio



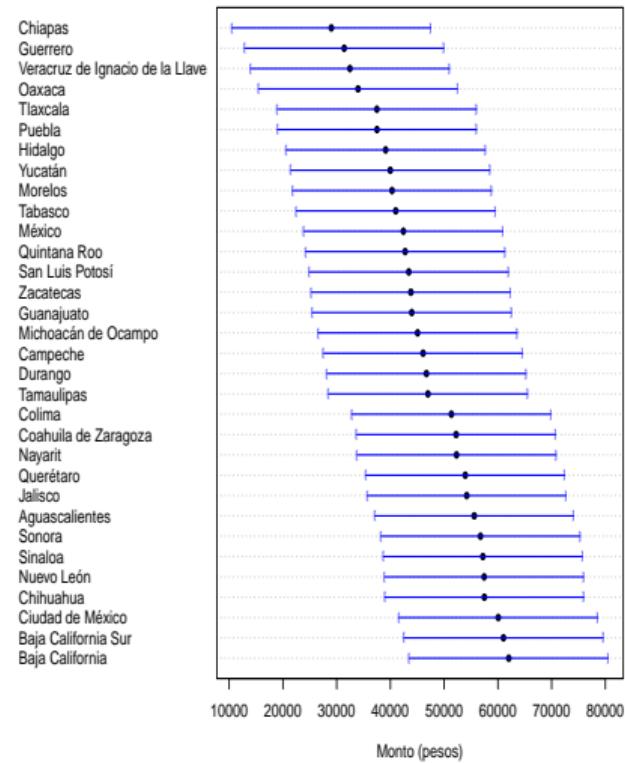
Gráficas de puntos I

```
a <- match(sort(X$ingcor, decreasing = T), X$ingcor)
B <- X[a,]
r <- nrow(X)
lim_inf <- B$ingcor - 2*sd(B$ingcor)
lim_sup <- B$ingcor + 2*sd(B$ingcor)

dotchart(B$ingcor, labels = rownames(B), pch = 16, cex = 0.8,
         main = "Ingreso corriente por entidad",
         xlab = "Monto (pesos)",
         xlim = c(min(lim_inf),max(lim_sup)))
points(lim_inf, 1:r, col = "blue", pch = "[", cex=0.5)
points(lim_sup, 1:r, col = "blue", pch = "]", cex=0.5)
segments(lim_inf, 1:r, lim_sup, 1:r, col = "blue")
```

Gráficas de puntos II

Ingreso corriente por entidad



Matriz de gráficas de dispersión I

- Vamos a crear un conjunto de datos que separe a hombres y mujeres en cada una de las variables resumen

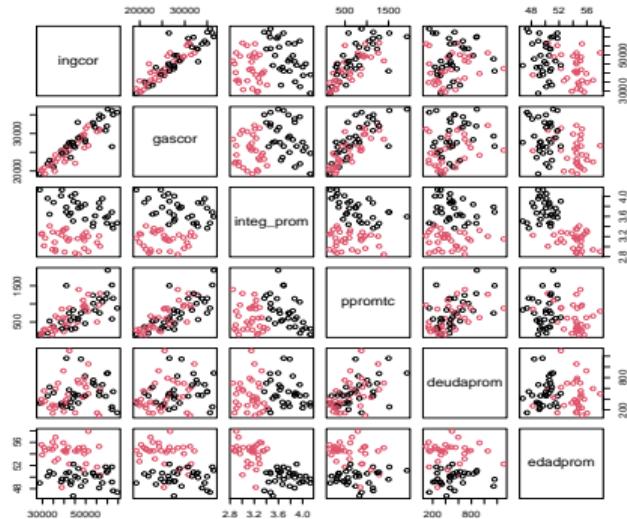
```
X <- Conc %>%
  group_by(entidad,sexo_jefe) %>%
  dplyr::select(-folioviv) %>%
  summarize(ingcor = mean(ing_cor),
            gascor = mean(gasto_mon),
            integ_prom = mean(tot_integ),
            ppromtc = mean(pago_tarje),
            deudaprom = mean(deudas),
            edadprom = mean(edad_jefe), .groups = "drop") %>%
  as.data.frame()
head(X)

      entidad sexo_jefe    ingcor     gascor integ_prom ppromtc deudaprom edadprom
1 Aguascalientes      H 57751.64 34452.86  3.952356 1050.6173  669.0779 49.00890
2 Aguascalientes      M 50177.90 31427.65  3.288538  898.3683  620.6893 52.29249
3 Baja California      H 64602.58 35624.14  3.468291  881.1544  147.0960 47.37283
4 Baja California      M 55289.17 30505.84  3.082024  499.0864  145.3739 51.73997
5 Baja California Sur      H 62710.20 35004.97  3.366276 1151.5375  539.4169 47.63321
6 Baja California Sur      M 56675.42 30875.62  3.006631 1285.0206  579.8494 50.38727
```

- La siguiente gráfica muestra los puntos identificando los valores para cada estado y para hombres y mujeres.

```
pairs(X[,-c(1:2)], col = factor(X$sexo_jefe))
```

Matriz de gráficas de dispersión II



- Los resúmenes de información involucran usualmente estadísticas sumarias tales como la media, percentiles, mediana, desviación estándar y correlación.
- En ocasiones se pueden incluir estadísticas más avanzadas, como componentes principales, que buscan reducir la dimensión de los datos.

Población y muestra I

- La colección entera de individuos u objetos de interés es conocida como la **población**. Una **muestra** es un subconjunto de la población.
- Usualmente, en el proceso estadístico es muy importante la manera en que la muestra es elegida de la población. El esquema de muestreo es muy relevante para el estudio y las conclusiones. Usualmente las características de la población se denotan con letras griegas, como μ , Σ , y las características de la muestra se ven con letras latinas, como m o \mathbf{S} .
- Usualmente se supone que la muestra de n elementos es *al azar*, lo que supone que la muestra fue seleccionada con la misma posibilidad de ser seleccionada que cualquier otras.
- Una muestra aleatoria de una característica o medida X se representa como $\{X_1, X_2, \dots, X_n\}$ que se supone proviene de la misma población.
- El proceso de inferencia estadística busca obtener de la muestra información sobre la población.

- La *media* o *promedio* corresponde al punto de equilibrio o centro de masa de los datos. se calcula como

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

- la media poblacional es el promedio de todas las observaciones en una población. Usualmente lo denotamos como $\mu = E(X)$.
- Por ejemplo, podemos obtener el ingreso promedio de la muestra de hogares que tenemos

```
(n <- nrow(Conc))  
[1] 89006  
(sum(Conc$ing_cor)/n)  
[1] 47838.49  
mean(Conc$ing_cor)  
[1] 47838.49
```

- Algunas características de la media:

- No es una estadística robusta: es sensible a valores atípicos.
- Puede ser diferente a cualquiera de los valores de la muestra.

Mediana

- La *mediana* o *percentil 50 %* se obtiene ordenando las observaciones de menor a mayor (incluyendo observaciones repetidas). Se representan estos valores con las *estadísticas de orden* $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$. Entonces la mediana corresponde a:

$$\text{mediana} = \text{med} = \begin{cases} X_{((n+1)/2)} & \text{si } n \text{ es impar} \\ \frac{X_{(n/2)} + X_{(n/2+1)}}{2} & \text{si } n \text{ es par} \end{cases}$$

```
median(Conc$ing_cor)
[1] 35172.01
```

- Algunas características de la mediana:

- Es una estadística robusta: es poco sensible a valores atípicos.
- La mediana y la media pueden diferir considerablemente. Cuando coinciden, la distribución de los datos es simétrica

Medias recortadas

- Una medida de tendencia central que está entre la media y la mediana son las *medias recortadas*. Una media recortada de 25 %, por ejemplo, ordena los datos y elimina 25 % de los datos a la izquierda y 25 % a la derecha, y de los datos que quedan, calcula la media.

```
mean(Conc$ing_cor, trim = 0.10) # elimina 20% de los datos  
[1] 39495.62
```

- Es importante notar que los programas pueden decidir redondear de diferente manera el porcentaje de los datos. Por ejemplo, R con $n = 15$ y $\alpha = 0.25$, se tiene que $0.25n = 3.75$ entonces se quita el 0.75, omitiendo tres observaciones de cada lado.

Medidas de dispersión

- Para medidas de dispersión, tres elecciones comunes son el rango, el rango intercuartil y la desviación estándar.
 - El *rango* corresponde a la diferencia entre el mínimo y el máximo de los valores.
 - El *rango intercuartil* es la diferencia entre el tercer y el primer cuartil de datos, como ya vimos con el boxplot.
 - La *desviación estándar*:

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

o la *varianza* que es simplemente s^2 .

```
range(Conc$ing_cor) # rango  
[1] 0 10702107  
  
IQR(Conc$ing_cor) # rango intercuartil  
[1] 36248.09  
  
sd(Conc$ing_cor) # desviación estándar  
[1] 71276.03
```

- Otras medidas de dispersión incluyen

- La desviación media absoluta (MAD): $\frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$

```
mad(Conc$ing_cor)  
[1] 24159.02
```

Medidas de forma I

- Para describir la forma de los conjuntos de datos, se puede utilizar la oblicuidad o asimetría (para medir asimetría) y la curtosis (para medir qué tan picuda es la densidad de los datos):

- La oblicuidad o asimetría se define como: $\gamma_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{s^3}$

- La curtosis es: $\gamma_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{s^4}$ y a veces se considera el exceso de curtosis como $\gamma_2 - 3$.

```
library(moments)
skewness(Conc$ing_cor)
[1] 59.5183
kurtosis(Conc$ing_cor)
[1] 7136.943
```

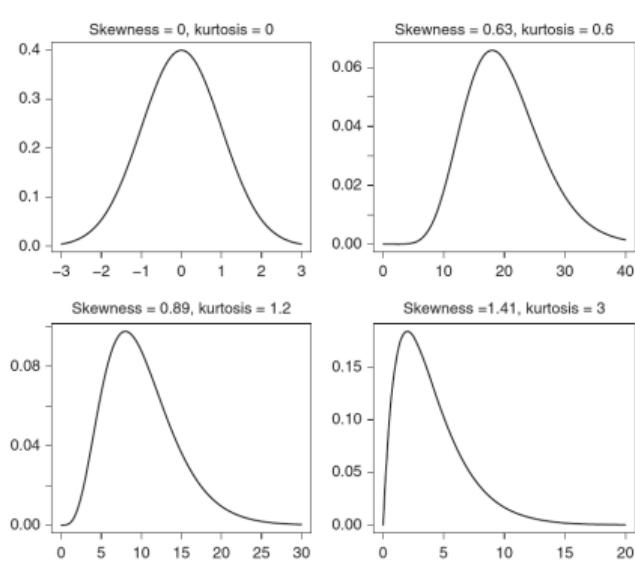
- Podemos usar el paquete psych que contiene buenas funciones para resumir los datos

```
library(psych)
describe(Conc)

   vars      n     mean       sd    median   trimmed      mad      min       max     range skew kurtosis      se
folioiv*    1 89006 43905.36 25316.07 43938.50 43910.52 32488.21 1   87754.0 87753.0  0.00  -1.20 84.86
sexo_jefe*  2 89006    1.29     0.45     1.00    1.24    0.00    1      2.0     1.0  0.93  -1.14  0.00
clase_hog*  3 89006    2.15     0.64     2.00    2.17    0.00    1      5.0     4.0  0.35  0.94  0.00
edad_jefe  4 89006   51.09    15.99    50.00   50.56   17.79   14     107.0    93.0  0.28  -0.56  0.05
tot_integ  5 89006    3.55     1.81     3.00    3.42    1.48    1     25.0     24.0  0.92  1.96  0.01
ing_cor    6 89006 47838.49 71276.03 35172.01 39495.62 24159.02 0 10702107.4 10702107.4 59.52 7133.78 238.91
gasto_mon  7 89006 28228.96 25610.94 22106.02 24190.73 14257.77 0 1007112.5 1007112.5  5.39  76.55 85.85
pago_tarje 8 89006   754.01   5436.46    0.00    0.00    0.00    0   440217.4 440217.4 24.90 1138.87 18.22
deudas     9 89006   507.57   3342.30    0.00    0.00    0.00    0 377213.1 377213.1 35.37 2855.95 11.20
entidad*   10 89006   15.87     9.24    15.00   15.74   11.86    1     32.0     31.0  0.09  -1.20  0.03
```

Medidas de forma II

- Ejemplos de las medidas y la forma de las densidades:



Varianzas, covarianzas y correlaciones I

- La matriz de varianzas y covarianzas contiene las mediciones de variables entre sí.

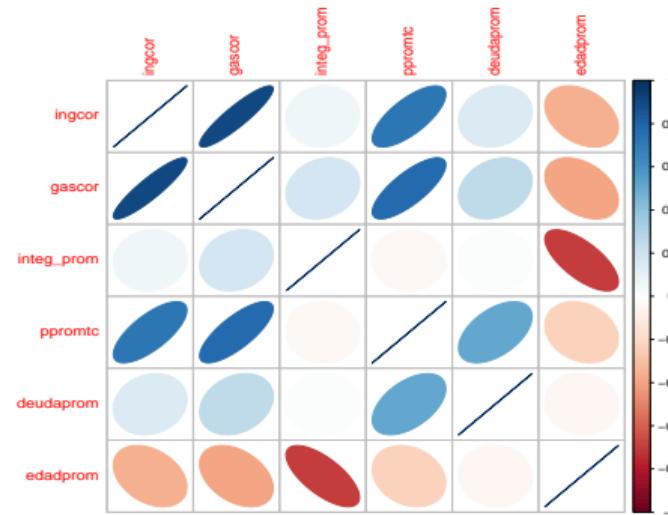
```
X$state_name_official <- NULL # Quita la variable que se agregó en los mapas.  
X$entidad <- NULL  
X$sexo_jefe <- NULL  
var(X)  
  
      ingcor      gascor  integ_prom     ppromtc   deudaprom    edadprom  
ingcor  89746060.0198 39942597.0330 216.0719055 2727451.2908 396372.897043 -9068.8107460  
gascor  39942597.0330 21724146.6498 307.5106707 1432167.4838 317181.687645 -5019.9921472  
integ_prom  216.0719    307.5107  0.1273491    -4.4068    1.726714    -0.6725588  
ppromtc  2727451.2908 1432167.4838 -4.4067995  158622.5262  54516.622913 -240.5373857  
deudaprom 396372.8970  317181.6876  1.7267136  54516.6229  71449.615996 -35.1584110  
edadprom -9068.8107   -5019.9921  -0.6725588   -240.5374  -35.158411    7.3511239
```

```
cor(X)  
  
      ingcor      gascor  integ_prom     ppromtc   deudaprom    edadprom  
ingcor  1.00000000  0.9046014  0.06391351  0.7228811  0.15652954 -0.35307403  
gascor  0.90460137  1.0000000  0.18488039  0.7715073  0.25458733 -0.39724167  
integ_prom  0.06391351  0.1848804  1.00000000 -0.0310058  0.01810183 -0.69511297  
ppromtc  0.72288111  0.7715073 -0.03100580  1.0000000  0.51209032 -0.22275286  
deudaprom 0.15652954  0.2545873  0.01810183  0.5120903  1.00000000 -0.04851236  
edadprom -0.35307403 -0.3972417 -0.69511297 -0.2227529 -0.04851236  1.00000000
```

Varianzas, covarianzas y correlaciones II

- Podemos visualizar las correlaciones entre los datos

```
library(corrplot)
corrplot(cor(X), method = "ellipse", cex.lab = 0.5)
```



Respuestas a las preguntas exploratorias I

- ¿Hay más jefas o más jefes de familia? ¿Cuál es la proporción de mujeres?

Los datos muestran que hay más jefes de familia, en una proporción de 3 a 1, lo cual se ve muy consistente a nivel nacional y a nivel estatal. La proporción de mujeres jefas de hogar es de 29 % a nivel nacional.

- ¿En qué entidad se da el mayor/menor ingreso corriente? ¿el mayor/menor gasto?

De acuerdo a la gráfica de regresión, El estado con menor ingreso y gasto es Chiapas, y el estado con mayor ingreso es Baja California y el que tiene mayor gasto es Sinaloa.

- ¿Cuál es el ingreso promedio a nivel nacional? ¿A nivel estatal? ¿Cuál es su variación?

El ingreso corriente promedio a nivel nacional es de 47,838.49, con una desviación estándar de 71,276. A nivel estatal, podemos ver la gráfica de puntos que generamos previamente y que incluye un intervalo de 95 % de confianza aprox.

- ¿Cuál es la moda de categorías del hogar?

Es la clase de hogar 2, que es de tipo nuclear y que representa el 63 % de los hogares del país.

Respuestas a las preguntas exploratorias II

- ¿Cómo se relacionan el gasto corriente y el ingreso corriente a nivel estatal?
Hicimos el análisis de regresión y calculamos la correlación de las variables
- ¿Hay alguna relación entre el gasto corriente y la edad de los jefes de familia? ¿el sexo del jefe de familia?
No parece haber una relación entre el gasto corriente y la edad de los jefes de familia, pero el sexo del jefe de familia si muestra estar relacionado con el gasto corriente. Es claro de los resultados que los hombres tienen un mayor ingreso y gasto.