# Exploración de dataframes con R

Alejandro Zavala

## Contents

## Descripción del análisis

El siguiente análisis consiste en interactuar con el entorno y lógica de diversas funciones en R. Se simulara un dataframe con diversas variables para el análisis exploratorio.

Viendo las primeras 20 observaciones del dataframe

```
set.seed(8) # Semilla para replicar script

cantidad_datos <- 10000

edad <- sample(10:45, size = cantidad_datos, replace = TRUE)
sexo <- sample(c("Masculino","Femenino"), size = cantidad_datos, replace = TRUE)
ingresos <- ifelse(sexo == "Femenino", runif(cantidad_datos,6000,12000), rnorm(cantidad_datos,22000,350
tono_piel <- sample(c("Blanco","Negro","Moreno"), size = cantidad_datos, replace = TRUE)
auto <- sample(c("Si","No"), size = cantidad_datos, replace = TRUE)
casa <- sample(c("Si","No"), size = cantidad_datos, replace = TRUE)
licenciatura <- sample(c("Si","No", "Se desconoce"), size = cantidad_datos, replace = TRUE)

datos_ingresos <- data.frame(edad,sexo,ingresos,licenciatura,tono_piel,auto,casa)
kable(head(datos_ingresos,20),format="markdown",row.names=TRUE)
```

| | edad | sexo | ingresos | licenciatura | tono_piel | auto | casa |
|---|---|---|---|---|---|---|---|
| 1 | 41 | Femenino | 7900.708 | No | Negro | No | No |
| 2 | 43 | Masculino | 20394.662 | Se desconoce | Blanco | Si | Si |
| 3 | 24 | Femenino | 9659.721 | Si | Blanco | No | No |
| 4 | 21 | Femenino | 6040.725 | Se desconoce | Moreno | Si | Si |
| 5 | 10 | Femenino | 7654.957 | Se desconoce | Negro | Si | No |
| 6 | 38 | Femenino | 11426.962 | No | Moreno | Si | Si |
| 7 | 12 | Masculino | 18964.097 | No | Negro | No | No |
| 8 | 44 | Femenino | 6334.924 | Si | Negro | Si | No |
| 9 | 15 | Femenino | 9786.449 | No | Moreno | Si | Si |
| 10 | 31 | Masculino | 16765.669 | No | Negro | No | Si |
| 11 | 20 | Masculino | 18455.201 | Se desconoce | Negro | No | No |
| 12 | 13 | Masculino | 24190.751 | Se desconoce | Negro | No | Si |
| 13 | 18 | Femenino | 9883.722 | Si | Moreno | No | Si |
| 14 | 21 | Masculino | 18127.544 | Si | Negro | No | Si |
| 15 | 16 | Masculino | 19867.647 | Se desconoce | Negro | Si | Si |
| 16 | 28 | Femenino | 8531.917 | Se desconoce | Blanco | Si | No |
| 17 | 15 | Masculino | 26772.029 | Si | Negro | Si | No |
| 18 | 16 | Femenino | 10250.067 | Se desconoce | Moreno | No | Si |
| 19 | 11 | Masculino | 20144.154 | Se desconoce | Negro | Si | Si |
| 20 | 30 | Masculino | 19972.282 | No | Moreno | Si | Si |

Viendo las ultimas 20 observaciones del dataframe

```
kable(tail(datos_ingresos,20),format="markdown",row.names=TRUE)
```

| | edad | sexo | ingresos | licenciatura | tono_piel | auto | casa |
|---|---|---|---|---|---|---|---|
| 9981 | 36 | Masculino | 19897.309 | No | Blanco | No | Si |
| 9982 | 13 | Femenino | 7846.756 | Si | Moreno | Si | Si |
| 9983 | 26 | Masculino | 17841.754 | No | Blanco | No | No |
| 9984 | 44 | Masculino | 18921.509 | Si | Moreno | No | Si |
| 9985 | 11 | Femenino | 7847.940 | No | Negro | Si | Si |
| 9986 | 16 | Masculino | 13518.391 | Si | Blanco | Si | Si |
| 9987 | 41 | Masculino | 17185.328 | Se desconoce | Moreno | Si | Si |
| 9988 | 23 | Femenino | 11172.003 | Se desconoce | Negro | Si | Si |
| 9989 | 32 | Femenino | 10646.881 | Si | Blanco | Si | No |
| 9990 | 28 | Femenino | 7448.916 | Se desconoce | Blanco | Si | No |
| 9991 | 23 | Femenino | 8201.323 | No | Blanco | Si | No |
| 9992 | 38 | Masculino | 17530.083 | Se desconoce | Blanco | Si | No |
| 9993 | 31 | Femenino | 11764.478 | Se desconoce | Moreno | No | No |
| 9994 | 19 | Femenino | 8917.782 | Si | Moreno | Si | No |
| 9995 | 25 | Masculino | 18372.749 | Si | Negro | No | Si |
| 9996 | 38 | Masculino | 22638.826 | Se desconoce | Negro | No | Si |
| 9997 | 28 | Femenino | 8368.195 | Se desconoce | Moreno | No | Si |
| 9998 | 32 | Femenino | 8800.436 | Si | Negro | No | No |
| 9999 | 42 | Femenino | 11141.950 | Si | Negro | No | No |
| 10000 | 39 | Masculino | 19697.124 | Se desconoce | Blanco | No | Si |

Viendo el nombre de columna del dataframe

```
names(datos_ingresos)
```

```
## [1] "edad"         "sexo"         "ingresos"     "licenciatura" "tono_piel"
## [6] "auto"         "casa"
```

Donde:

1. edad: representa la edad del individuo
2. Sexo: representa el sexo del individuo
3. ingresos: Ingresos mensuales del individuo mensual (en moneda mexicana)
4. licenciatura: si el individuo cuenta con licenciatura
5. tono_piel: raza del individuo
6. auto: si el individuo tiene auto propio
7. casa: si el individuo tiene casa propio

## Explorando el dataframe

Para conocer la estructura del dataframe, usamos la función **str**

```
str(datos_ingresos)
```

```
## 'data.frame':    10000 obs. of  7 variables:
##  $ edad        : int  41 43 24 21 10 38 12 44 15 31 ...
##  $ sexo        : chr  "Femenino" "Masculino" "Femenino" "Femenino" ...
##  $ ingresos    : num  7901 20395 9660 6041 7655 ...
##  $ licenciatura: chr  "No" "Se desconoce" "Si" "Se desconoce" ...
##  $ tono_piel   : chr  "Negro" "Blanco" "Blanco" "Moreno" ...
##  $ auto        : chr  "No" "Si" "No" "Si" ...
##  $ casa        : chr  "No" "Si" "No" "Si" ...
```

## Selección de variables

Si deseamos mostrar las primeras 10 filas de nuestro dataframe

```
kable(datos_ingresos[1:10,])
```

| edad | sexo | ingresos | licenciatura | tono_piel | auto | casa |
|---|---|---|---|---|---|---|
| 41 | Femenino | 7900.708 | No | Negro | No | No |
| 43 | Masculino | 20394.662 | Se desconoce | Blanco | Si | Si |
| 24 | Femenino | 9659.721 | Si | Blanco | No | No |
| 21 | Femenino | 6040.725 | Se desconoce | Moreno | Si | Si |
| 10 | Femenino | 7654.957 | Se desconoce | Negro | Si | No |
| 38 | Femenino | 11426.962 | No | Moreno | Si | Si |
| 12 | Masculino | 18964.097 | No | Negro | No | No |
| 44 | Femenino | 6334.924 | Si | Negro | Si | No |
| 15 | Femenino | 9786.449 | No | Moreno | Si | Si |
| 31 | Masculino | 16765.669 | No | Negro | No | Si |

Si deseamos ver las primeras 10 observaciones (renglones) de las primeras 2 columnas

```
kable(datos_ingresos[1:10,1:2],format="markdown",row.names=TRUE)
```

| | edad | sexo |
|---|---|---|
| 1 | 41 | Femenino |
| 2 | 43 | Masculino |
| 3 | 24 | Femenino |
| 4 | 21 | Femenino |
| 5 | 10 | Femenino |
| 6 | 38 | Femenino |
| 7 | 12 | Masculino |
| 8 | 44 | Femenino |
| 9 | 15 | Femenino |
| 10 | 31 | Masculino |

Si deseamos ver los elementos únicos de una columna, tomando la de sexo

```
unique_sex <- unique(datos_ingresos["sexo"]);kable(unique_sex)
```

| sexo |
|---|
| Femenino |
| Masculino |

Si deseamos filtrar nuestro dataframe por individuos que sean del sexo “Femenino”. Mostraremos las primeras 15 observaciones

```
data_femenino <- datos_ingresos[datos_ingresos["sexo"]=="Femenino",]
kable(data_femenino[1:15,],format="markdown",row.names=TRUE)
```

| | edad | sexo | ingresos | licenciatura | tono_piel | auto | casa |
|---|---|---|---|---|---|---|---|
| 1 | 41 | Femenino | 7900.708 | No | Negro | No | No |
| 3 | 24 | Femenino | 9659.721 | Si | Blanco | No | No |
| 4 | 21 | Femenino | 6040.725 | Se desconoce | Moreno | Si | Si |
| 5 | 10 | Femenino | 7654.957 | Se desconoce | Negro | Si | No |
| 6 | 38 | Femenino | 11426.962 | No | Moreno | Si | Si |
| 8 | 44 | Femenino | 6334.924 | Si | Negro | Si | No |
| 9 | 15 | Femenino | 9786.449 | No | Moreno | Si | Si |
| 13 | 18 | Femenino | 9883.722 | Si | Moreno | No | Si |
| 16 | 28 | Femenino | 8531.917 | Se desconoce | Blanco | Si | No |
| 18 | 16 | Femenino | 10250.067 | Se desconoce | Moreno | No | Si |
| 22 | 24 | Femenino | 11221.995 | No | Blanco | Si | Si |
| 23 | 19 | Femenino | 10886.229 | Si | Moreno | Si | No |
| 24 | 17 | Femenino | 6457.714 | Se desconoce | Blanco | Si | Si |
| 27 | 41 | Femenino | 11601.926 | No | Moreno | No | Si |
| 28 | 15 | Femenino | 8091.790 | No | Blanco | No | No |

Si queremos extraer a todos los individuos de sexo “Masculino” y que sean de tono de piel “moreno”. Mostrando las primeras 10 observaciones

```
data_masculino_moreno <- datos_ingresos[datos_ingresos["sexo"]== "Masculino" & datos_ingresos["tono_piel
kable(data_masculino_moreno[1:10,],format="markdown",row.names=TRUE)
```

| | edad | sexo | ingresos | licenciatura | tono_piel | auto | casa |
|---|---|---|---|---|---|---|---|
| 20 | 30 | Masculino | 19972.28 | No | Moreno | Si | Si |
| 32 | 40 | Masculino | 20157.02 | Si | Moreno | Si | No |
| 40 | 16 | Masculino | 24480.01 | Si | Moreno | No | No |
| 45 | 21 | Masculino | 22229.74 | Se desconoce | Moreno | No | Si |
| 48 | 21 | Masculino | 20708.11 | No | Moreno | No | No |
| 58 | 14 | Masculino | 20593.65 | Si | Moreno | Si | Si |
| 61 | 34 | Masculino | 26534.90 | Se desconoce | Moreno | Si | No |
| 68 | 12 | Masculino | 19300.19 | Si | Moreno | Si | No |
| 71 | 22 | Masculino | 29756.07 | No | Moreno | No | Si |
| 90 | 13 | Masculino | 26358.30 | No | Moreno | Si | No |

Si deseamos filtrar por individuos de sexo femenino y con ingresos mayores iguales a 1000 dolares. Mostrando las primeras 10 observaciones

```
data_femenino_1000 <- datos_ingresos[datos_ingresos["sexo"]== "Femenino" & datos_ingresos["ingresos"]>=

kable(data_femenino_1000[1:10,],format="markdown",row.names=TRUE)
```

| | edad | sexo | ingresos | licenciatura | tono_piel | auto | casa |
|---|---|---|---|---|---|---|---|
| 1 | 41 | Femenino | 7900.708 | No | Negro | No | No |
| 3 | 24 | Femenino | 9659.721 | Si | Blanco | No | No |
| 4 | 21 | Femenino | 6040.725 | Se desconoce | Moreno | Si | Si |
| 5 | 10 | Femenino | 7654.957 | Se desconoce | Negro | Si | No |
| 6 | 38 | Femenino | 11426.962 | No | Moreno | Si | Si |
| 8 | 44 | Femenino | 6334.924 | Si | Negro | Si | No |
| 9 | 15 | Femenino | 9786.449 | No | Moreno | Si | Si |
| 13 | 18 | Femenino | 9883.722 | Si | Moreno | No | Si |
| 16 | 28 | Femenino | 8531.917 | Se desconoce | Blanco | Si | No |
| 18 | 16 | Femenino | 10250.067 | Se desconoce | Moreno | No | Si |

## Agregando nuevas variables

Si al dataframe inicial queremos convertir la cantidad a dólares americanos (considerando este a un intercambio de 1 dólar a 20 pesos mexicanos). Listando las primeras 10 observaciones

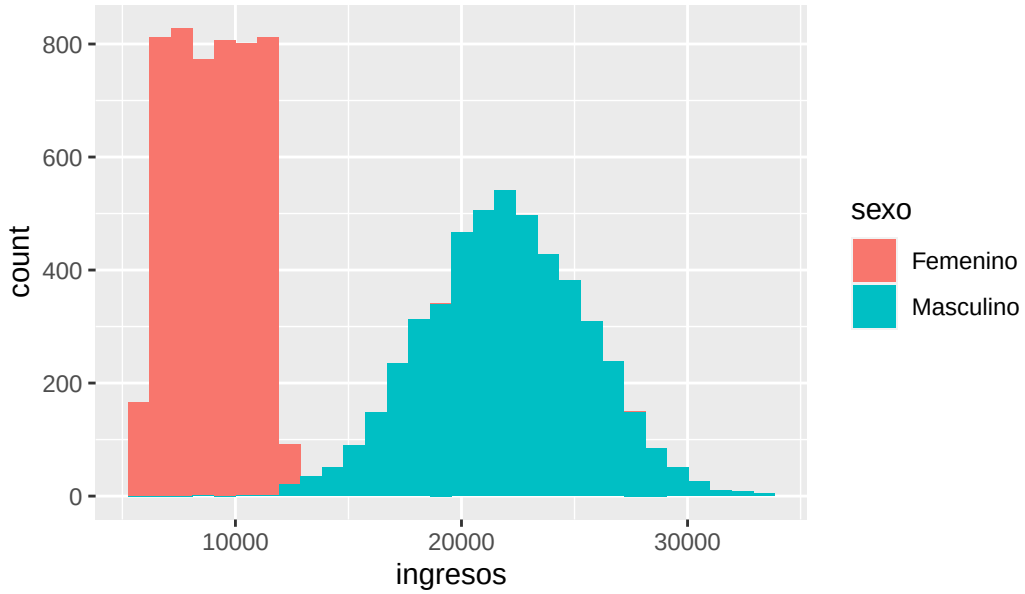
```
datos_ingresos["ingresos_us"] <- datos_ingresos["ingresos"]/20

kable(datos_ingresos[1:10,],format="markdown",row.names=TRUE)
```

| | edad | sexo | ingresos | licenciatura | tono_piel | auto | casa | ingresos_us |
|---|---|---|---|---|---|---|---|---|
| 1 | 41 | Femenino | 7900.708 | No | Negro | No | No | 395.0354 |
| 2 | 43 | Masculino | 20394.662 | Se desconoce | Blanco | Si | Si | 1019.7331 |
| 3 | 24 | Femenino | 9659.721 | Si | Blanco | No | No | 482.9861 |
| 4 | 21 | Femenino | 6040.725 | Se desconoce | Moreno | Si | Si | 302.0362 |
| 5 | 10 | Femenino | 7654.957 | Se desconoce | Negro | Si | No | 382.7479 |
| 6 | 38 | Femenino | 11426.962 | No | Moreno | Si | Si | 571.3481 |
| 7 | 12 | Masculino | 18964.097 | No | Negro | No | No | 948.2048 |
| 8 | 44 | Femenino | 6334.924 | Si | Negro | Si | No | 316.7462 |
| 9 | 15 | Femenino | 9786.449 | No | Moreno | Si | Si | 489.3225 |
| 10 | 31 | Masculino | 16765.669 | No | Negro | No | Si | 838.2834 |

## Distribuciones de los datos

Veamos la distribución de los ingresos por sexo con un histograma y en un diagrama de caja

```
hist_ingresos_sex <- ggplot( datos_ingresos,aes(x=ingresos, fill=sexo)) +
  geom_histogram(bins=30)
boxplot_ingresos_sexo <- ggplot( datos_ingresos,aes(y=ingresos, x = sexo)) +
  geom_boxplot()
hist_ingresos_sex
```




```
boxplot_ingresos_sexo
```