Taller_Angela_Villate

Angela_Villate

2/10/2020

Exploración de datos, gráficos y análisis de MTCARS

**** Nota:

Este ejercicio tiene tres momentos: En el primero un trabajo exploratorio para entender los datos y su forma de relacionarse

Luego se realizaron filtros según la correlación de las variables

Se presentan unas gráficas para visualizar los datos

*Cada uno de los chunks se acompañó de un análisis o comentario. En muchos de ellos se indicaron los errores como por ejemplo lo impertinente de hacer cierto tipo de consulta o de usar cierto tipo de estética.

1. Instalamos el paquete de librerías para preparar nuestro entorno de trabajo

Ahora llamamos las librerías

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(visdat)
library(cowplot)
library(ggrepel)
library(mapproj)
```

```
## Loading required package: maps
```

```
##
## Attaching package: 'maps'
```

```
## The following object is masked from 'package:purrr':
##
##     map
```

```
library(ggthemes)
```

```
##
## Attaching package: 'ggthemes'
```

```
## The following object is masked from 'package:cowplot':
##
##     theme_map
```

```
library(here)
```

```
## here() starts at /cloud/project
```

```
library(extrafont)
```

```
## Registering fonts with R
```

```
library(extrafont)
library(knitr)
library(magick)
```

```
## Linking to ImageMagick 6.9.7.4
## Enabled features: fontconfig, freetype, fftw, lcms, pango, x11
## Disabled features: cairo, ghostscript, rsvg, webp
```

```
## Using 16 threads
```

```
library(ggplot2)
```

Ahora llamaremos nuestra base de datos de trabajo

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

Hacemos un head para conocer los primeros datos

```
head(mtcars)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Encontramos que este es un dataset en el que se presentan 10 características según la marca de carros: como características encontramos las siguientes: mpg: millas por galón cyl: número de cilindros Displacement: desplazamiento hp: caballos de fuerza drat: relación del eje trasero wt: peso (seguramente es el peso del motor) qsec: vs Engine (0 = V-shaped, 1 = straight): transmisión automática Vs manual gear Number of forward gears: número de engranajes número de carburadores

#Establecemos una correlación entre las variable (aún explorando los datos)

```
cor(mtcars$mpg, mtcars)
```

```
##      mpg       cyl       disp         hp      drat         wt     qsec
## [1,]   1 -0.852162 -0.8475514 -0.7761684 0.6811719 -0.8676594 0.418684
##             vs        am      gear       carb
## [1,] 0.6640389 0.5998324 0.4802848 -0.5509251
```

#Ahora empezaremos a hacer arreglos utilizando la librería tydivers

```
arrange(mtcars, mpg, desc(cyl)) %>% head()
```

```
##                      mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Cadillac Fleetwood  10.4   8  472 205 2.93 5.250 17.98  0  0    3    4
## Lincoln Continental 10.4   8  460 215 3.00 5.424 17.82  0  0    3    4
## Camaro Z28          13.3   8  350 245 3.73 3.840 15.41  0  0    3    4
## Duster 360          14.3   8  360 245 3.21 3.570 15.84  0  0    3    4
## Chrysler Imperial   14.7   8  440 230 3.23 5.345 17.42  0  0    3    4
## Maserati Bora       15.0   8  301 335 3.54 3.570 14.60  0  1    5    8
```

Este arreglo nos muestra una organización descendente de la relación entre las millas por minuto y el número de cilindros que usan cada uno de los motores de los carros.

De la representación de los datos podemos concluir que no hay una relación directa entre el número de cilindros y y las millas por hora.

Pues encontramos, por ejemplo el caso del carro Mazda RX4 que con 6 cilindros recorre 21 millas por segundo.

##Ahora, para relaizar otro tipo de consulta haremos un fitro específico por cilindro

```
mtcars %>% filter(mpg >= 18) %>% head()
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Este parámetro se pensó según una hipótesis (no comprobada) de que aquellos vehículos que recorren más de 18 mpg pueden ser los más competitivos en el mercado, en consecuencia podrías concluri que atuos como Mazda RX4, Datsun, Hornet 4 Drive y Hornet Sportabout son algunos de los más competititvos según las mpg recorridas.

##Ahora seguimos realizando nuevas consultas en el dataset

```
mtcars %>% filter(mpg %in% c(18 ,21)) %>% head()
```

```
##               mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4      21   6  160 110  3.9 2.620 16.46  0  1    4    4
## Mazda RX4 Wag  21   6  160 110  3.9 2.875 17.02  0  1    4    4
```

*** Nos damos cuenta que la consulta anterior no tiene sentido debido a que no existe un valor en mpg en 18, pues existen intervalos entre 18 y 21. Por esta razón la consulta solo nos trajo dos valores.

Por lo tanto, haremos nuevamente la consulta pero con los cilindros, esta puede arrojarnos más información valiosa.

```
mtcars %>% filter(cyl %in% c(6 ,8)) %>% head()
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
```

```
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
## Duster 360        14.3   8  360 245 3.21 3.570 15.84  0  0    3    4
```

Esta consulta tiene un poco más sentido que la anterior, porque nos muestra los datos que se registran en los cilindros 6 y 8

```
mtcars %>%

  filter(mpg >= 18, gear==4) %>%

  dplyr::select(mpg, cyl,carb, gear) %>%

  head(n=20)
```

```
##                 mpg cyl carb gear
## Mazda RX4      21.0   6    4    4
## Mazda RX4 Wag  21.0   6    4    4
## Datsun 710     22.8   4    1    4
## Merc 240D      24.4   4    2    4
## Merc 230       22.8   4    2    4
## Merc 280       19.2   6    4    4
## Fiat 128       32.4   4    1    4
## Honda Civic    30.4   4    2    4
## Toyota Corolla 33.9   4    1    4
## Fiat X1-9      27.3   4    1    4
## Volvo 142E     21.4   4    2    4
```

Esta última consulta nos ayuda a enteder un poco más la lógica de la data (desconozco de carros). De aquí podemos comprender que, por ejemplo, si queremos comprender cuáles son los carros más comerciales, es decir, aquellos que las familias prefieren con más frecuencia se deben tener en cuenta variables como: mpg, cyl, ho, gear, carb.

Las calidades que buscamos de los autos son aquellos que recorran más millas, tengan más cilindros (pues estos le dan potencia al motor), tengan más hp (fuerza en los engranajes), revisar la caja de velocidades (gear), pero que no tenga tantos carburadores; estos últimos aumentarían el consumo de gasolina.

Así que haremos un nuevo filtro para tener una relación de estas variables seleccionadas.

```
mtcars %>%

  filter(mpg >= 18, carb <= 4) %>%

  dplyr::select(mpg, cyl,carb, gear, hp) %>%

  head(n=20)
```

```
##                   mpg cyl carb gear  hp
## Mazda RX4         21.0   6    4    4 110
## Mazda RX4 Wag     21.0   6    4    4 110
## Datsun 710        22.8   4    1    4  93
## Hornet 4 Drive    21.4   6    1    3 110
## Hornet Sportabout 18.7   8    2    3 175
## Valiant           18.1   6    1    3 105
## Merc 240D         24.4   4    2    4  62
```

```
## Merc 230          22.8   4    2    4  95
## Merc 280          19.2   6    4    4 123
## Fiat 128          32.4   4    1    4  66
## Honda Civic       30.4   4    2    4  52
## Toyota Corolla    33.9   4    1    4  65
## Toyota Corona     21.5   4    1    3  97
## Pontiac Firebird  19.2   8    2    3 175
## Fiat X1-9         27.3   4    1    4  66
## Porsche 914-2     26.0   4    2    5  91
## Lotus Europa      30.4   4    2    5 113
## Volvo 142E        21.4   4    2    4 109
```

Esta última consulta nos permite tener una idea de los carros según las características que consideramos óptimas para establecer si son comerciales.

De los datos podemos concluir que, automóviles como el mazda rx4, rx4 wag, el datsum, el hornet, el fiat 128 son las marcas de automóviles que podríamos decir que son comercialmente "atractivos", tienen buen rendimiento y bajo consumo de gasolina (bajo número de carburadores).

```
mtcars %>%

  group_by(cyl) %>%

  summarize("min" = min(mpg, na.rm = TRUE),

            "Q1" = quantile(mpg,  probs = 0.25, na.rm = TRUE),

            "median (Q2)" = median(mpg, na.rm = TRUE),

            "mean" = mean(mpg, na.rm = TRUE),

            Q3 = quantile(mpg,  probs = 0.75, na.rm = TRUE),

            "max Q4" = max(mpg, na.rm = TRUE)

  ) %>% head()
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 3 x 7
##     cyl   min    Q1 `median (Q2)`  mean    Q3 `max Q4`
##   <dbl> <dbl> <dbl>         <dbl> <dbl> <dbl>    <dbl>
## 1     4  21.4  22.8          26    26.7  30.4     33.9
## 2     6  17.8  18.6          19.7  19.7  21       21.4
## 3     8  10.4  14.4          15.2  15.1  16.2     19.2
```

Esta última consulta nos permitió agrupar los datos por cilindros (entendiendo que estos son un indicador de la potencia del motor) y calculamos los estadísticos de las mpg para establecer la correlación entre la distancia recorrida y la potencia.

Tenemos que a mayor número de cilindros el promedio de mpg disminuye.

Ahora realizaremos la misma consulta, pero el criterio de agrupación serán los carburadores.

```
mtcars %>%

  group_by(gear) %>%
```

```
  summarize("min" = min(mpg, na.rm = TRUE),

            "Q1" = quantile(mpg,  probs = 0.25, na.rm = TRUE),

            "median (Q2)" = median(mpg, na.rm = TRUE),

            "mean" = mean(mpg, na.rm = TRUE),

            Q3 = quantile(mpg,  probs = 0.75, na.rm = TRUE),

            "max Q4" = max(mpg, na.rm = TRUE)

  ) %>% head()
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 3 x 7
##    gear   min    Q1 `median (Q2)`  mean    Q3 `max Q4`
##   <dbl> <dbl> <dbl>         <dbl> <dbl> <dbl>    <dbl>
## 1     3  10.4  14.5          15.5  16.1  18.4     21.5
## 2     4  17.8  21            22.8  24.5  28.1     33.9
## 3     5  15    15.8          19.7  21.4  26       30.4
```

Estos datos nos muestran la relación entre las velocidades de un motor y el promedio de mpg tenemos, entonces que, un motor con una caja de 5 velocidades puede recorrer en promedio 21, 4 mpg mientras que uno de 3 hará un recorrido promedio de 18,4 mpg.

Ahora vamos a reescalar los displacements que está en milímetros y la pasamos a metros.

```
mtcars %>%

  mutate(disp = disp/1000) %>%

  head()
```

```
##    mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## 1 21.0   6 0.160 110 3.90 2.620 16.46  0  1    4    4
## 2 21.0   6 0.160 110 3.90 2.875 17.02  0  1    4    4
## 3 22.8   4 0.108  93 3.85 2.320 18.61  1  1    4    1
## 4 21.4   6 0.258 110 3.08 3.215 19.44  1  0    3    1
## 5 18.7   8 0.360 175 3.15 3.440 17.02  0  0    3    2
## 6 18.1   6 0.225 105 2.76 3.460 20.22  1  0    3    1
```

Quizas porque estamos acostumbrados al sistema métrico esta sea una mejor opción de presentar este dato.

Ahora realizaremos un MAS de la data

```
#Muestreo aleatorio simple
sample_n(mtcars, 9)
```

```
##                   mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Ford Pantera L   15.8   8 351.0 264 4.22 3.170 14.50  0  1    5    4
## Camaro Z28       13.3   8 350.0 245 3.73 3.840 15.41  0  0    3    4
## Datsun 710       22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
## Merc 230         22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
## Fiat 128         32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
## Merc 240D        24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
## Mazda RX4        21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
```

```
## Dodge Challenger 15.5   8 318.0 150 2.76 3.520 16.87  0  0    3    2
## Toyota Corona    21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
```

```
sample_frac(mtcars, 0.2) %>% head()
```

```
##                    mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Chrysler Imperial  14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
## Porsche 914-2      26.0   4 120.3  91 4.43 2.140 16.70  0  1    5    2
## Merc 240D          24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
## Merc 450SE         16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
## Cadillac Fleetwood 10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
## Hornet 4 Drive     21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
```

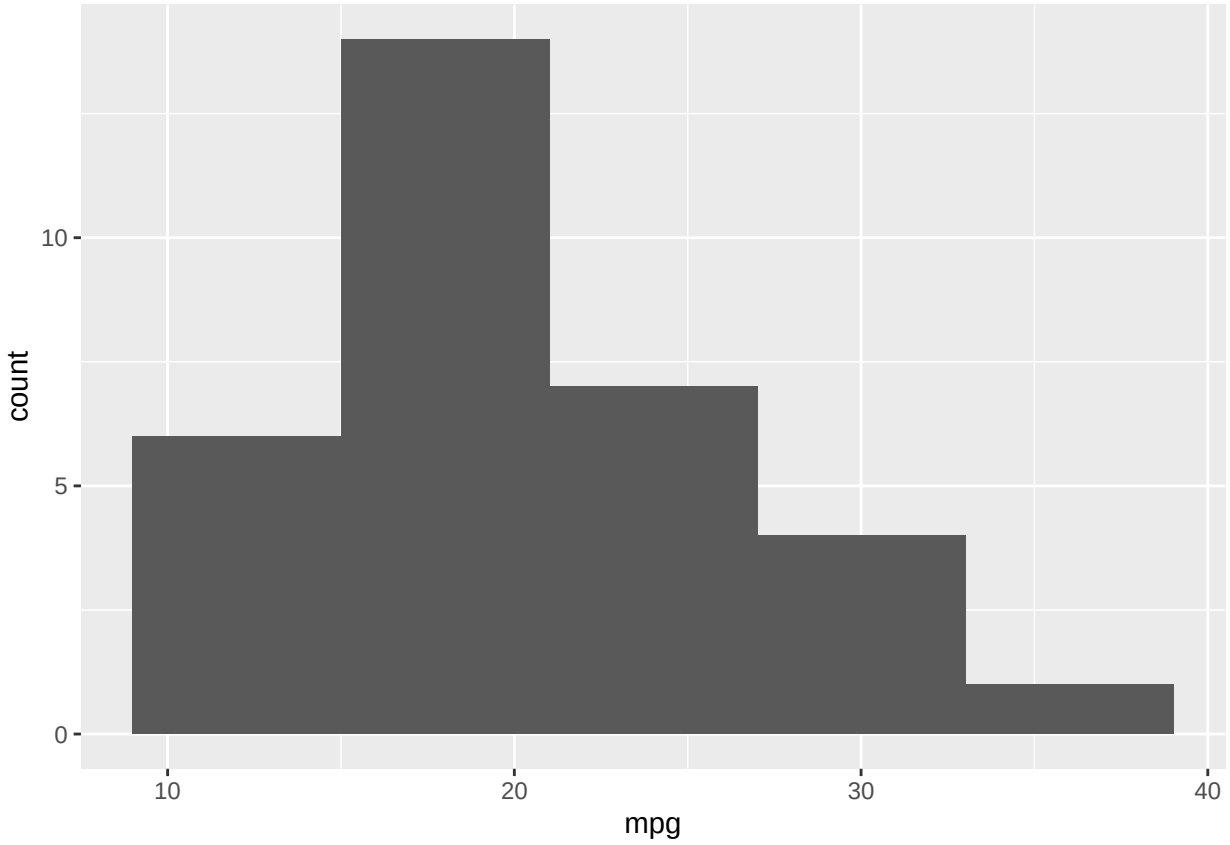
2. Gráficas de la data

Para eso utilizaremos la libreria ggplot2

```
library('ggplot2')
```

Histograma

```
ggplot(mtcars,aes(x=mpg)) + geom_histogram(binwidth=6)
```

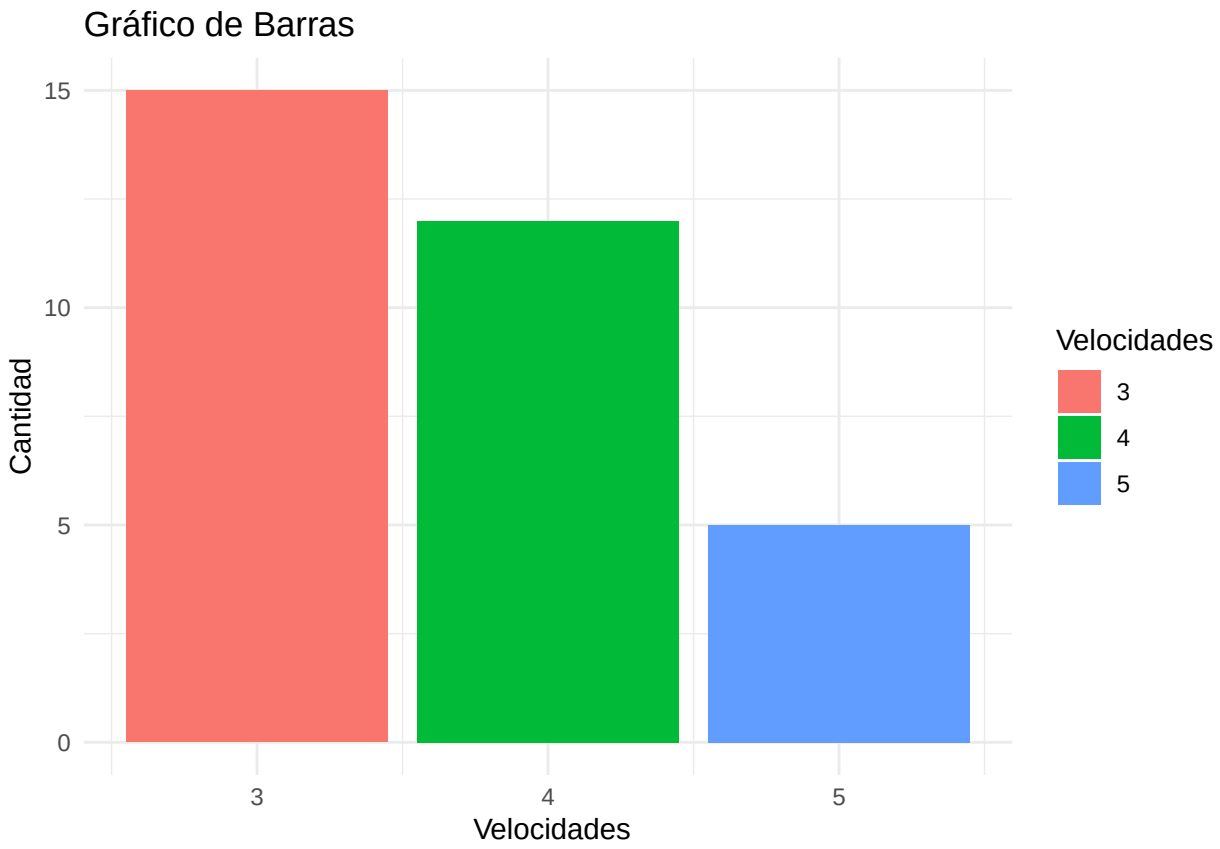


Este histograma nos permite ver cómo están distribuidos los datos según las mpgo las millas por galón. Lo que podemos observar es que la mayoría de los datos se encuentran en los intervalos de 16 a 21 mpg.

De esta manera podemos tener una lectura más rápida de algunas de las conclusiones que habíamos identificado con los filtros.

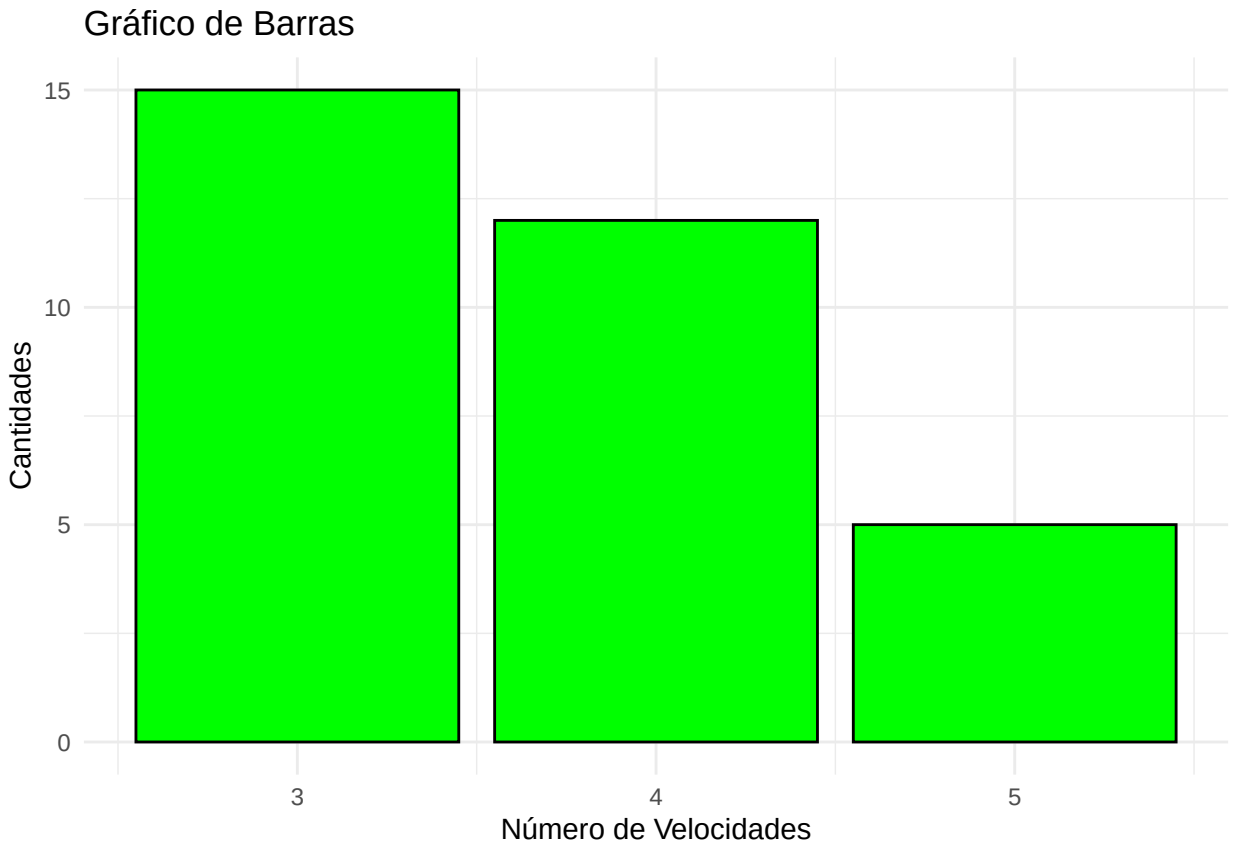
```
ggplot(data = mtcars, aes(x = gear, fill = as.factor(gear))) +
  geom_bar() +
```

```
  xlab(" Velocidades") +
  ylab("Cantidad") +
  ggtitle("Gráfico de Barras") +
  labs(fill = "Velocidades") +
  theme_minimal()
```



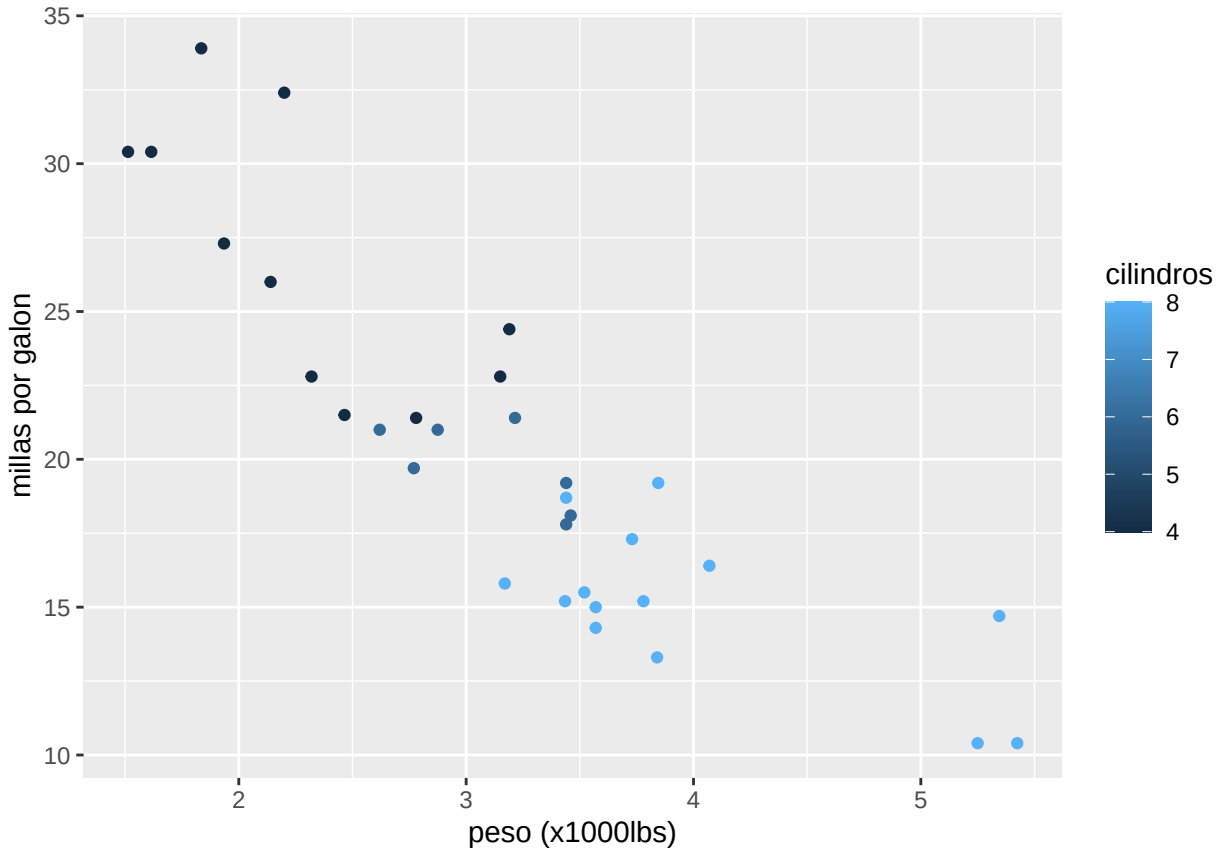
En esta gráifa de barras representamos los vehículos acumulados por velocidades. Identificamos entonces que, la mayor cantidad de vehículos cuenta con 3 velocidades deguidos de los de 4 y 5. En conclusión podríamos decir que las personas tienden a adquirir vehículos de 3 velocidades y que estos tienen un recorrido promedio de 16 a 21 mpg.

```
ggplot(data = mtcars, aes(x = gear, fill = as.factor(gear))) +
  geom_bar(color = 'black', fill = 'green') +
  xlab("Número de Velocidades") +
  ylab("Cantidades") +
  ggtitle("Gráfico de Barras") +
  labs(fill = "Velocidades") +
  theme_minimal()
```



Esta es la misma gráfica anterior pero con otra estética, que la verdad no es muy favorable porque los colores nos permiten identificar con mayor claridad una de las variables que estamos representando y es la agrupación de los datos porvelocidades.

```
library(ggthemes)
my_scatplot <- ggplot(mtcars,aes(x=wt,y=mpg,col=cyl)) + geom_point()
my_scatplot + labs(x='peso (x1000lbs)',y='millas por galon',colour='cilindros')
```



Esta es una gráfica multidimensional en la que podemos representar tres variables y su densidad: millas por galón, peso del motor y número de cilindros.

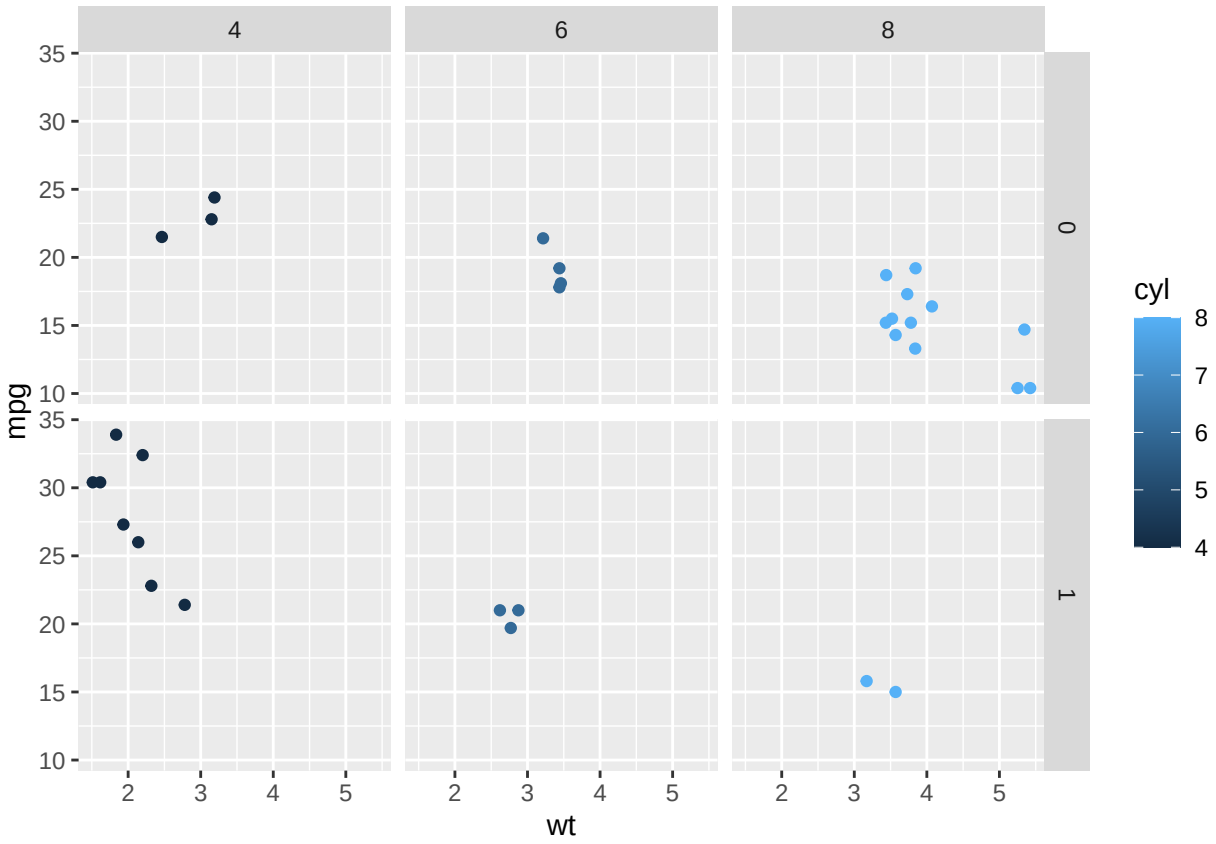
Recordemos que los cilindros indican la potencia del motor, en ese sentido los puntos más oscuros representarán aquellos vehículos con menos cilindros y a más claro la gráfica nos indicará que los vehículos tienen más cilindraje.

Podemos concluir de la gráfica que:

El peso y el número de cilindros está relacionado direcatamente, así a menor número de cilindros el peso es menor.

Para establecer nuestro producto óptimo tenemos que es aquel que recorre más millas, tiene menor cantidad de cilindros y su peso es menor.

```
my_scatplot <- ggplot(mtcars,aes(x=wt,y=mpg,col=cyl)) + geom_point()
my_scatplot + facet_grid(am~cyl)
```



En esta gráfica tenemos los mismos datos que en la anterior, pero se han agrupado por cilindros. Así se puede hacer un análisis más detallado por cada uno de los grupos de interés.