Gráficos multidimensionales

Juan Salamanca

16 de mayo de 2017

Gráficos multidimensionales para clasificación

Asumamos que queremos clasificar cuales carros de un conjunto de varios cientos se pueden considerar como económicos en consumo de combustible. Algunos de ellos tienen cilindrajes pequeños y consumen menos en la ciudad que en autopista, pero no siempre se observa ese comportamiento. Algunos de alto cilindraje son tambien económicos. **El problema está en determinar una serie de umbrales para cada atributo (dimensión) del conjunto de datos que ayuden a clasificar los carros entre económicos y consumidores con un pequeño margen de error**

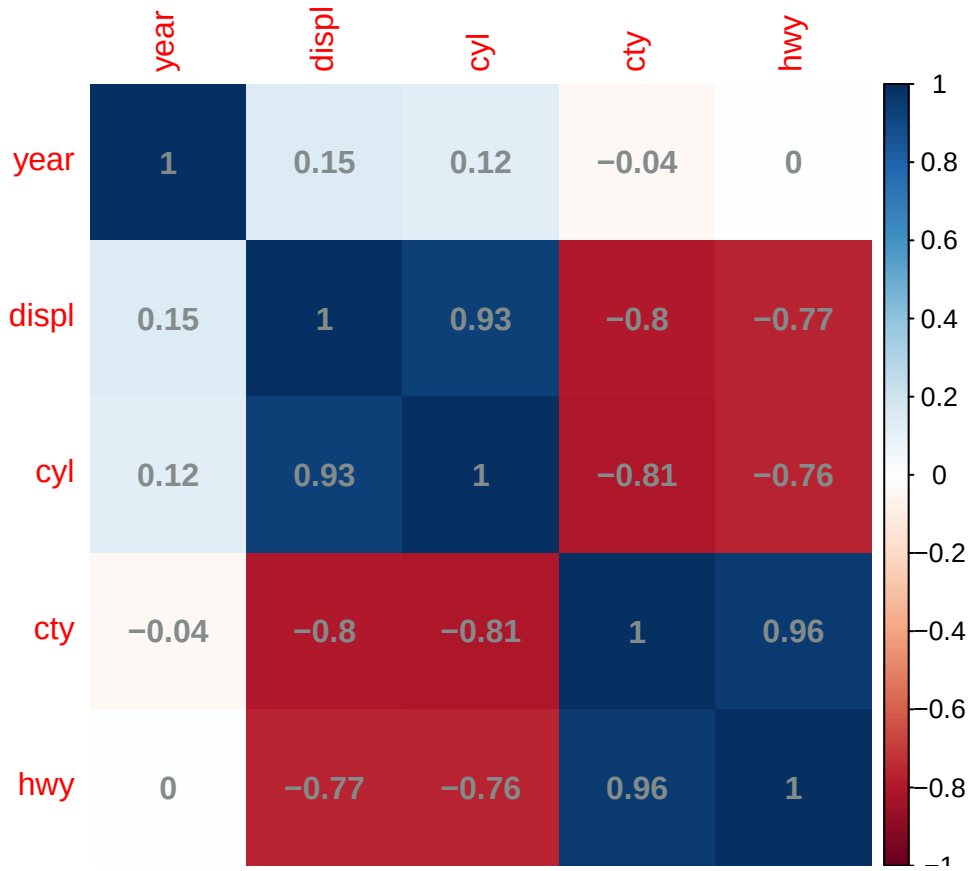
Para una referencia animada de esta tarea ver: http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

```
#Librerias a utilizar
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.2
```

```
library(corrplot)
```

En el taller de correlogramas aprendimos a hacer un mapa de temepratura con los valores de correlación entre variables numéricas y obtuvimos un resultado similar a este que representa la interacción entre variables numéricas del dataset **mpg** de **ggplot2**.

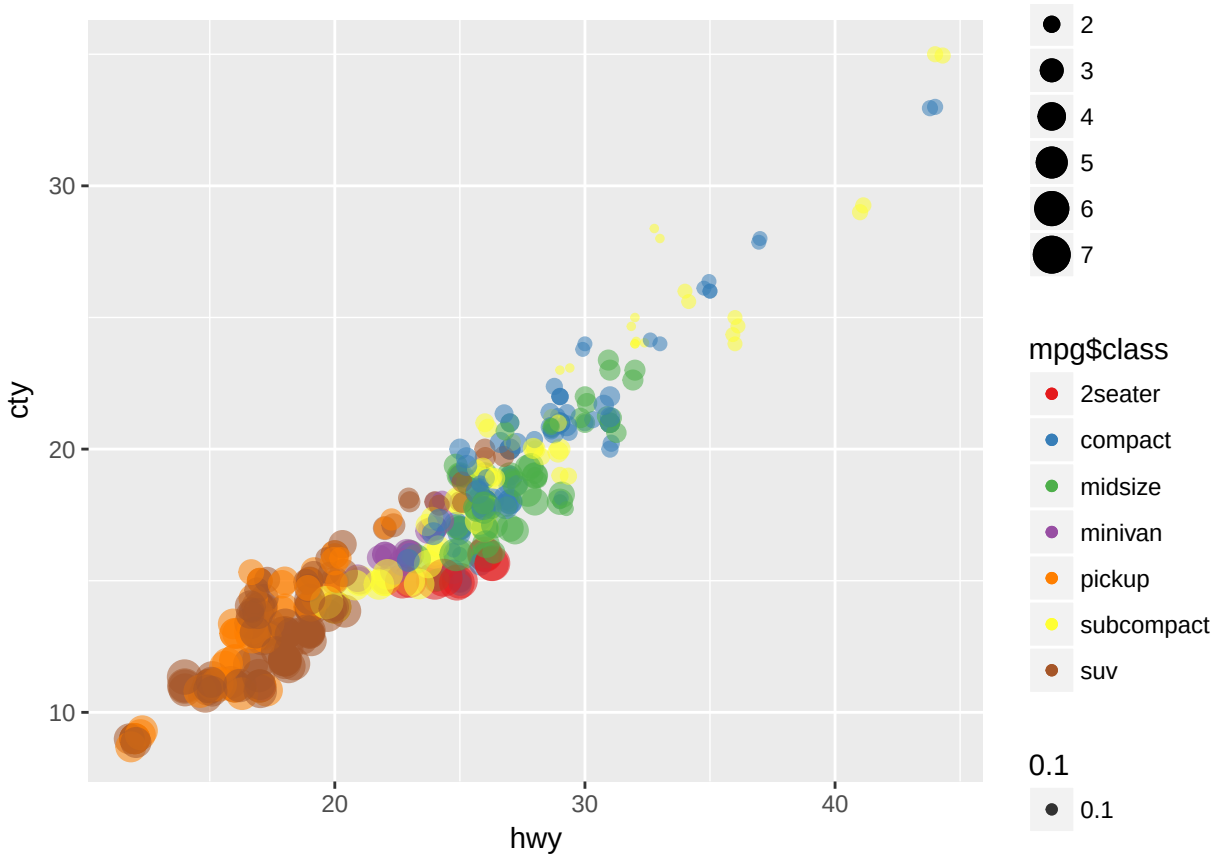


El mapa de temperatura nos indica que hay dos grupos de variables con correlaciones positivas (*disp* con *cly*, y *hwy* con *cty*). Entre esos grupos hay correlaciones negativas. Aunque el valor de correlación nos sirve para darnos cuenta de la magnitud y la direccion de la correlación, no es suficiente para darnos cuenta del efecto real (causalidad o co-ocurrencia) de una variable sobre la otra ni para **clasificar datos en categorías**. Recuerde el ejemplo de de los cuartetos de anscombe (https://en.wikipedia.org/wiki/Anscombe%27s_quartet).

Para interpretar realmente los efectos de las variables que diferencian categorias en los datos es necesario 'leer' los datos a nivel de cada registro. Para esto el método mas efectivo es un diagrama de dispersión o **scatter plot**.

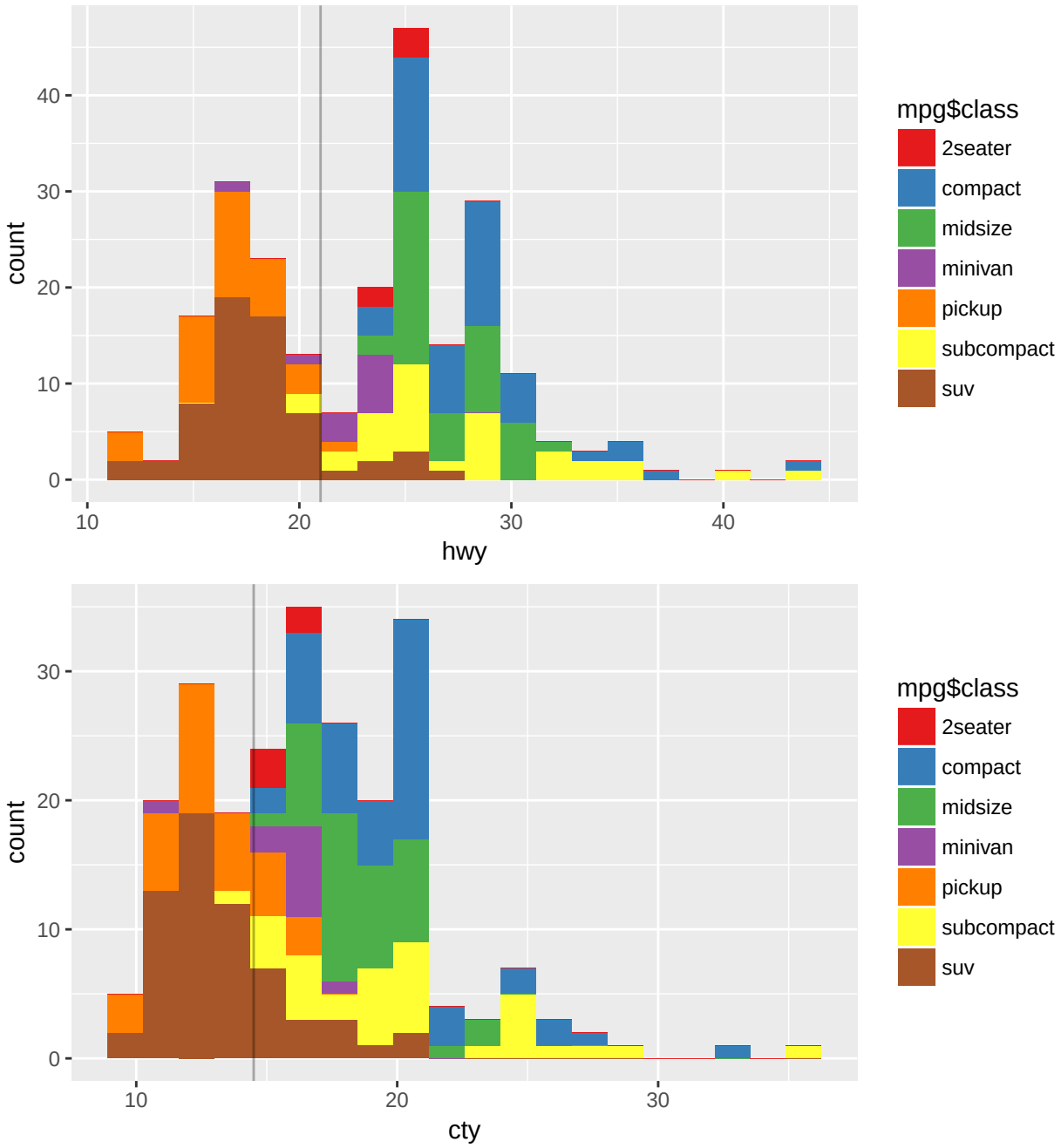
Por ejemplo, en el conjunto de vehículos, la relación entre el número de millas por litro en carretera (*hwy*), el número de millas que hace por litro de combustible (*cty*) describe una correlacion lineal. El cilindraje se representa en el diámetro de cada punto (*displ*).

```
scatterPlot <- ggplot(mpg, aes(hwy, cty, color = mpg$class, size = mpg$displ,
    alpha = 0.1))
scatterPlot <- scatterPlot + geom_point() + geom_jitter()
# scatterPlot <- scatterPlot + geom_hline(yintercept = 15, alpha = 0.3) +
# geom_vline(xintercept = 20, alpha = 0.3)
scatterPlot <- scatterPlot + scale_colour_brewer(palette = "Set1")
scatterPlot
```



En el diagrama de dispersión se aprecia a simple vista que hay clases de vehículos que consistentemente consumen mas combustible y que tienen un volumen de cilindraje superior a la mayoría del conjunto. Pero es dificil determinar en qué punto se pude determinar con 0% de error los vehículos que consideraríamos económicos de los que no. Los siguientes histogramas nos permiten mayor pecisión y nos muestran que para la variable *hwy* el valor de corte con menor error esta cerca a 21 millas/litro y para la variable *cty* esta entre 14 y 15 millas/litro.

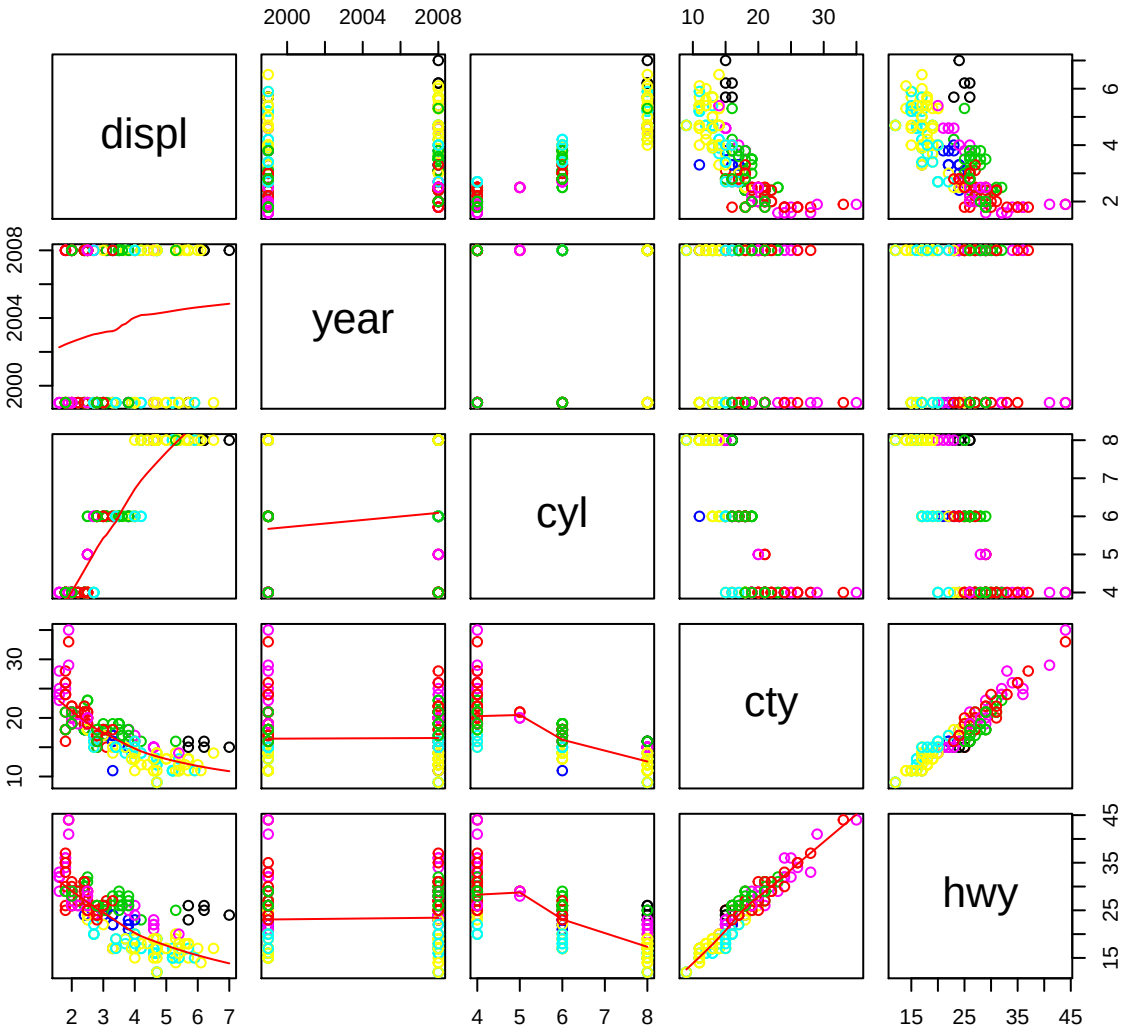
```
# HWY
histogram <- ggplot(mpg, aes(hwy, fill = mpg$class))
# Cambiel la cantidad de bins para tener mayor resoluci<U+00F3>n
histogram <- histogram + geom_histogram(bins = 20) + geom_vline(xintercept = 21,
    alpha = 0.3)
histogram <- histogram + scale_fill_brewer(palette = "Set1")
histogram
```


count
hwy
mpg$class
2seater
compact
midsize
minivan
pickup
subcompact
suv
count
cty
mpg$class
2seater
compact
midsize
minivan
pickup
subcompact
suv

Matriz de diagramas de dispersión

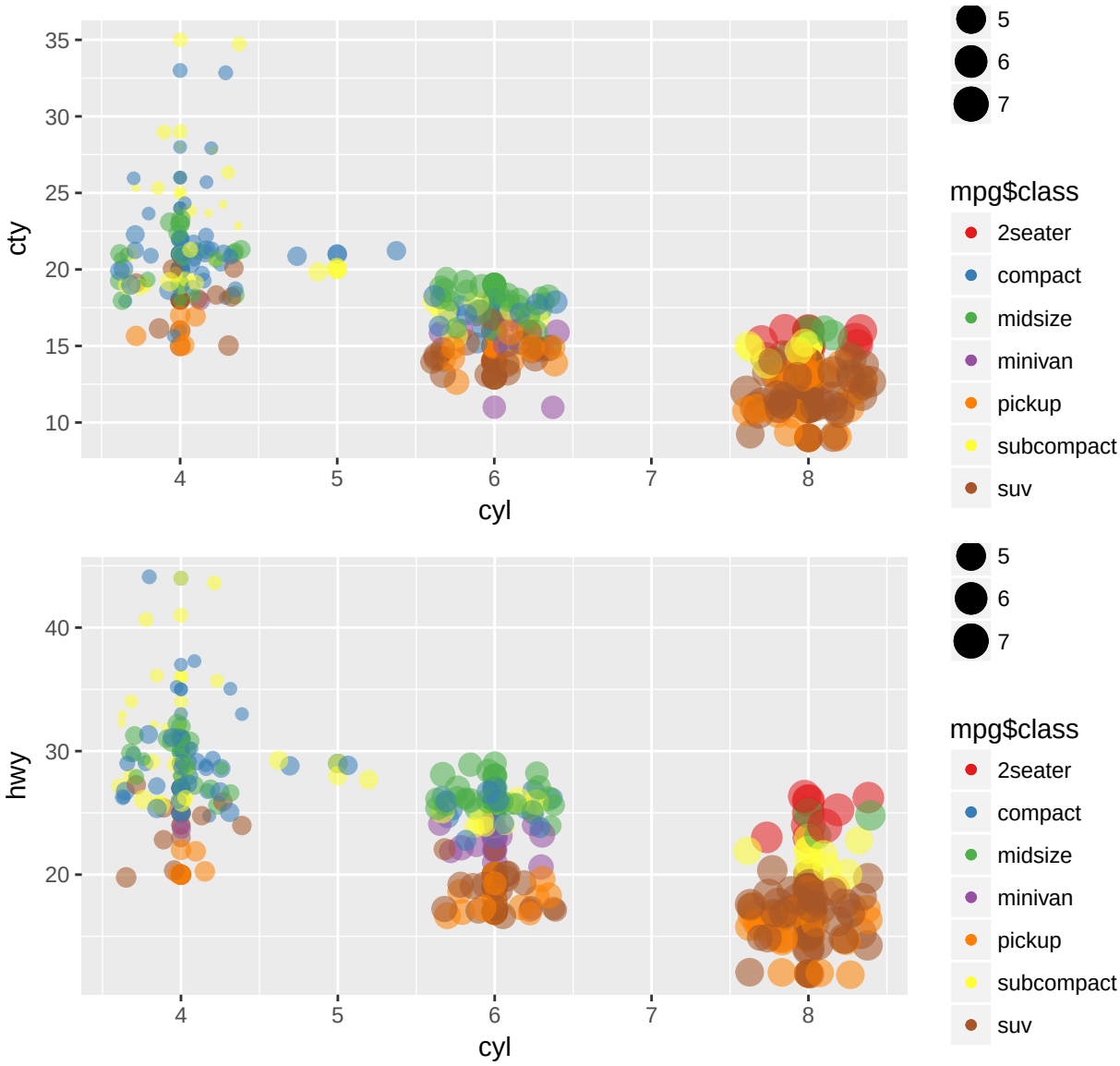
Para efectuar este análisis sobre todas las combinaciones de dimensiones utilizamos una **matriz de diagramas de dispersión**. Para esto seleccionamos las dimensiones numéricas mas relevantes y utilizamos el método de graficación *pairs* nativo de R.

```
# Selecciona las columnas relevantes
relevant <- mpg[, c(3, 4, 5, 8, 9)]
pairs(relevant[, 1:5], col = as.factor(mpg$class), lower.panel = panel.smooth)
```



En el heatmap vimos que la dimensión *year* no tiene mayor correlación con las demás dimensiones. En esta matriz confirmamos que el año no hace diferencia en el consumo de combustible. Sin embargo se alcanza a apreciar que en el año 2000 se produjeron un par de vehículos con una economía superior al de todos los que se produjeron en 2008. Aún con esta evidencia, la dimensión *year* no es un bue clasificador porque no muestra grupos diferenciados.

En general es posible ver que la cantidad de cilindros (*cly*) tambien permite diferenciar los vehículos de acuerdo a su consumo. Los diagramas de dispersión correspondientes a cada dimensión de consumo de gasolina son:



En este caso vemos que no hay un umbral único porque la dimensión *cyl* es numérica y discreta (números enteros). En esta situación lo ideal es clasificar primero por el consumo en ciudad y en autopista y posteriormente verificar que ninguno de los escogidos tiene mas de 5 cilindros.

Taller

Utilice el dataset de sobrevivientes del Titanic y determine 2 dimensiones que le permitan confirmar si las mujeres y los niños tuvieron prioridad a la hora de la evacuación o si hubo algún privilegio por haber pagado un tarifa alta, la clase en la que viajaban, o cualquiera de los demás atributos.

Dataset disponible en: https://www.kaggle.com/c/titanic/data o en la carpeta del curso en Moodle.

Gráfico de coordenadas paralelas

Reference: https://datascience.blog.wzb.eu/2016/09/27/parallel-coordinate-plots-for-discrete-and-categorical-data-in-r-a-com

En esta técnica, cada dimensión corresponde a un eje con la misma longitud de uno de los dos ejes del sistema de coordenadas. Los ejes que representan las dimensiones seleccionadas por el usuario están dispuestos de manera equidistante y paralela a lo largo del otro eje de visualización. Por último, se dibujan las líneas que conectan los puntos de datos a través de las dimensiones.

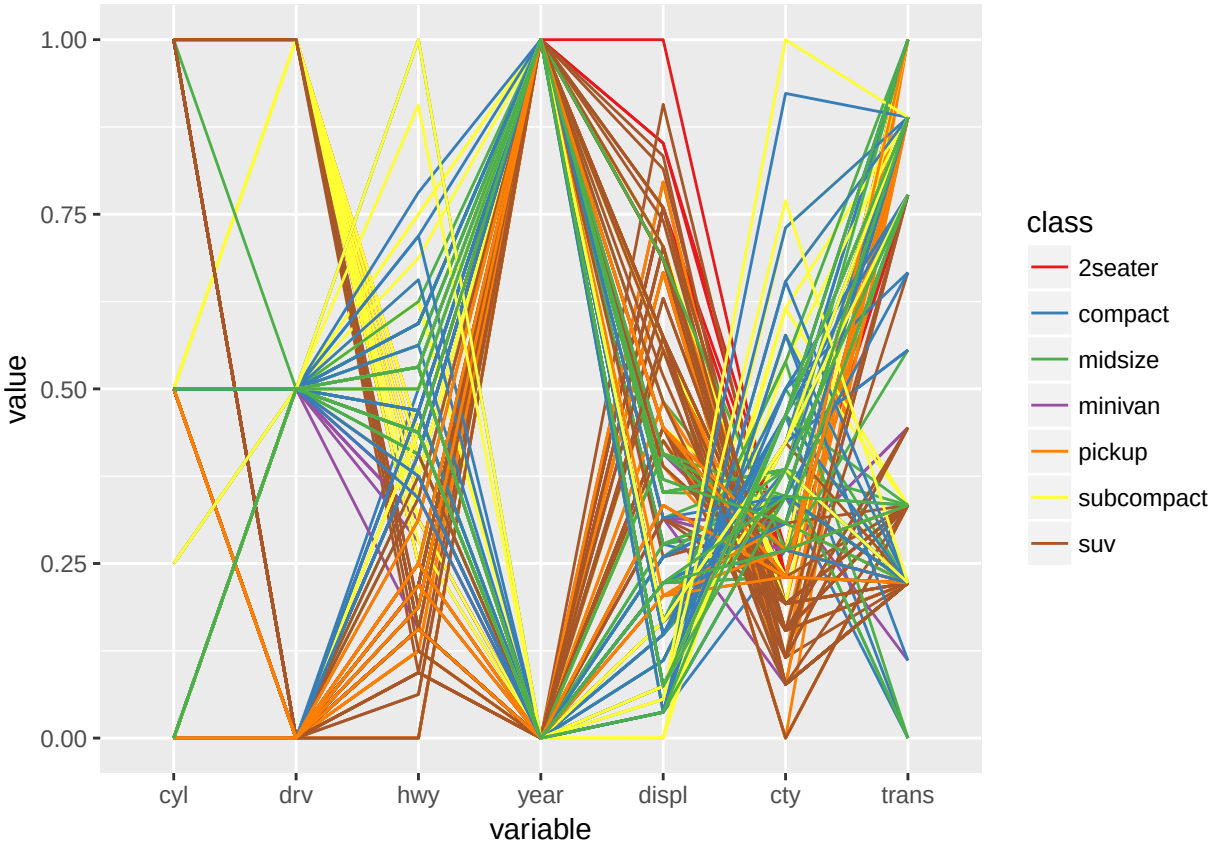
Usando GGPLOT

ggplot2 no tiene una función para dibujar directamente un sistema de coordenadas paralelas. Utilizaremos el método *ggparcoord* del paquete *GGally*.

```
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 3.3.2
```

```
parallelPlot <- ggparcoord(mpg, columns = 3:9, groupColumn = "class", scale = "uniminmax",
    order = "allClass")
parallelPlot <- parallelPlot + scale_colour_brewer(palette = "Set1")
parallelPlot
```



Para que la gráfica sea mas inteligible es necesario ordenar los ejes. Se recomienda poner contiguos los que tienen correlación en la misma dirección y magnitud.

```
parallelPlot <- ggparcoord(mpg, columns = c(3, 5, 6, 7, 8, 9), groupColumn = "class",
    scale = "uniminmax", order = c(7, 5, 3, 6, 8, 9))
parallelPlot <- parallelPlot + scale_colour_brewer(palette = "Set1")
parallelPlot <- parallelPlot + facet_grid(year ~ .)
parallelPlot
```