Tarea 3

Daniel Camarena

2 de septiembre de 2017

Tarea 3

Objetivo

El propósito de esta tarea es que trabajen un análisis más completo usando las herramienta que hemos estudiado, en particular visualización, transformaciones y el paradigma separa-aplica-combina.

Datos

Para esta tarea pueden elegir la base de datos con la que desean trabajar, puede ser alguna de clase: billboard, flights, batting, etc. ó la base de datos de causas de mortalidad en México (tarea anterior) ó cualquier base de datos de su interés.

Una vez que elijan los datos preguntense:

- ¿Qué otras fuentes de información podrían ser útiles?
- ¿Qué quiero aprender de estos datos?
- ¿Qué datos hay disponibles?
- ¿Me puedo enfocar en un subconjunto de los datos?

Una vez que consideren los puntos anteriores pueden plantearse las preguntas que quieren responder con sus datos (tendencias en causas de mortalidad, diferencias por grupos, etc), conforme trabajen con sus datos surgirán nuevas preguntas.

Entrega

Realicen un reporte que incluya una introducción donde se explique las preguntas que abordaron, descripción de las bases de datos, gráficas (al menos 3) y tablas si es necesario.

Introducción

Para esta tarea analizaremos los datos de la base de datos de flights, para ello primero cargaremos la libreria *tidyverse* y pondremos los datos en sesión.

Preguntas a resolver:

- ¿Qué factores pueden influir en el retraso de salida de un vuelo?
- ¿Qué aerolinea tiene los aviones más veloces?
- ¿Qué Aerolinea tiene mas retrasos en salidas y llegadas?
- ¿Qué meses son los más caóticos en los aeropuertos de nueva york?
- ¿Han cambiado el número de asientos dentro de los aviones?

```
library(tidyverse)
library(nycflights13)
library(pander)
library(ggplot2)
```

```
flights <- flights
airports <- airports
planes <- planes
airlines <- airlines
weather <- weather
```

Este data set (nycflights13) fue construído por Hadley Wickham con la finalidad de poder entender qué es lo que causan los retrasos de vuelos en **NYC**. Intentemos hacer lo mismo...

Lo primero que haremos es combinar la tabla de flights con sus catálogos para obtener una visión global del *DataSet*

```
flights_comp <- flights %>% left_join(airports, by = c('dest'='faa')) %>%
                            left_join(planes, by = c('tailnum'='tailnum')) %>%
                            left_join(airlines) %>%
                            left_join(weather)
```

Luego de hacer los joins necesarios vemos las columnas que tenemos listas para nuestro análisis

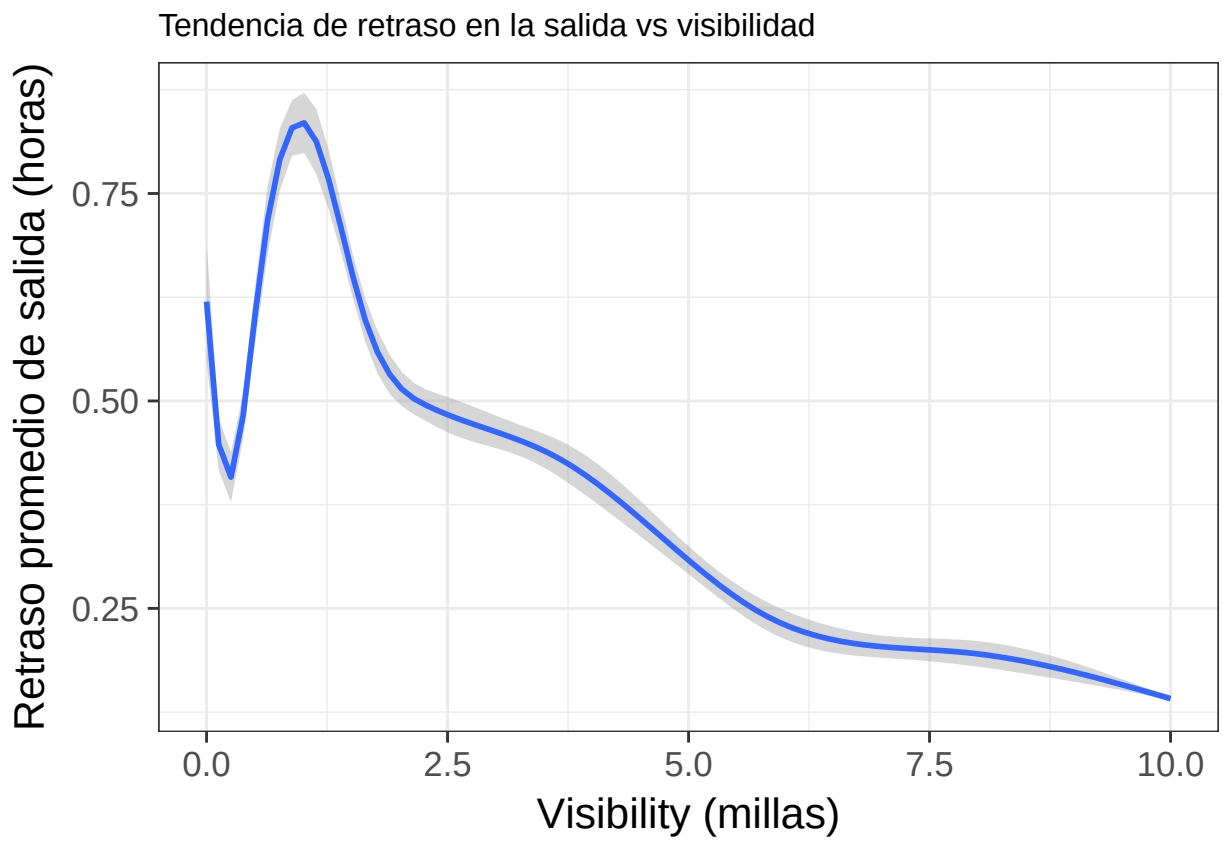
```
##  [1] "year.x"          "month"           "day"             "dep_time"
##  [5] "sched_dep_time"  "dep_delay"       "arr_time"        "sched_arr_time"
##  [9] "arr_delay"       "carrier"         "flight"          "tailnum"
## [13] "origin"          "dest"            "air_time"        "distance"
## [17] "hour"            "minute"          "time_hour"       "name"
## [21] "lat"             "lon"             "alt"             "tz"
## [25] "dst"             "tzone"           "year.y"          "type"
## [29] "manufacturer"    "model"           "engines"         "seats"
## [33] "speed"           "engine"          "year"            "temp"
## [37] "dewp"            "humid"           "wind_dir"        "wind_speed"
## [41] "wind_gust"       "precip"          "pressure"        "visib"
```

Ya con nuestro data set listo vamos a hacer análisis... a ver que sale...

Centremos nuestra atención en las siguientes variables: *dep_delay*, *visibility*. Es muy probable que una baja visibilidad dentro del aeropuerto cause un retraso en la salida de los vuelos.

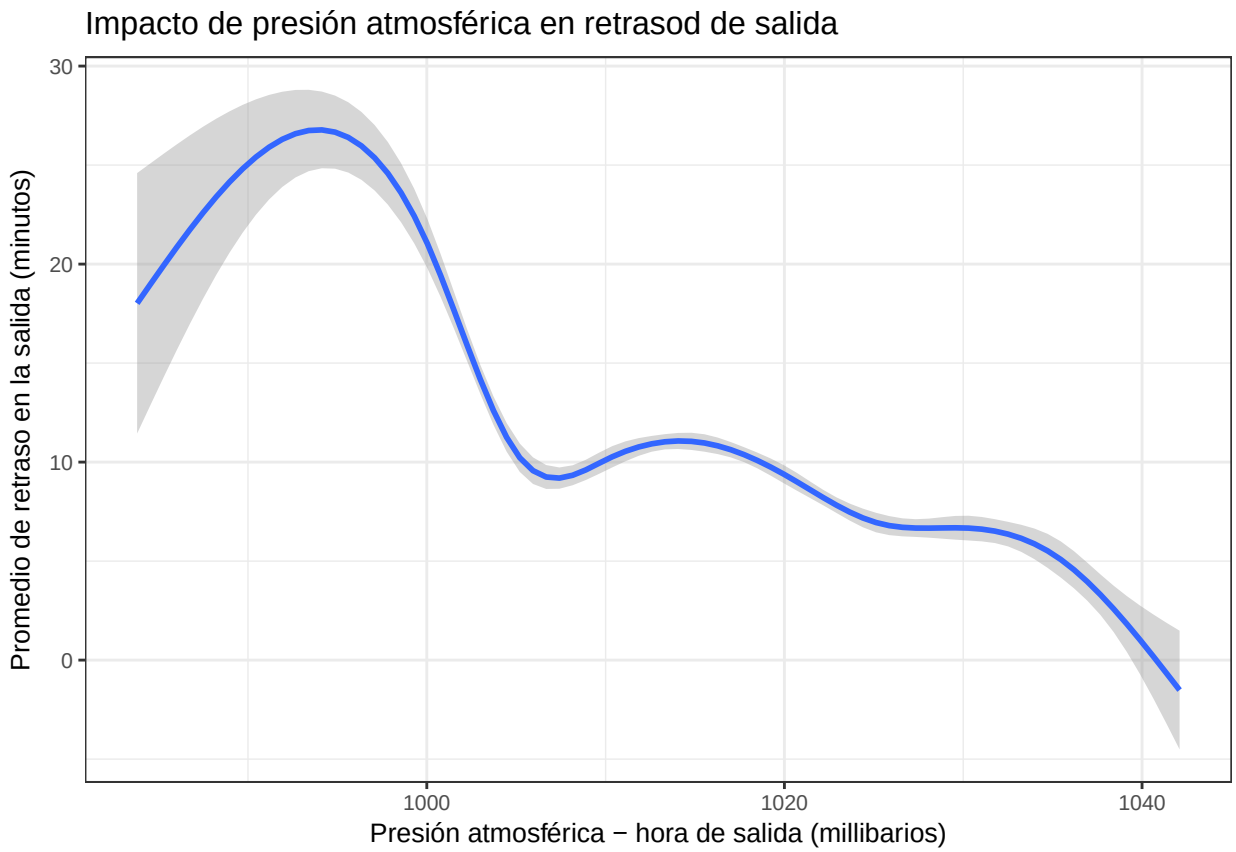
También vamos a hacer una variable que nos dé el tiempo de retraso en horas

```
flights_comp %>% mutate(dep_delay_hr = dep_delay/60) %>%
              select(origin, year, month, day, hour, dep_delay_hr, visib) %>%
              filter(!is.na(dep_delay_hr) & !is.na(visib)) %>%
              ggplot(aes(x = visib, y = dep_delay_hr)) +
              geom_smooth() +
              theme_bw(base_size = 16) +
              xlab("Visibility (millas)") +
              ylab("Retraso promedio de salida (horas)") +
              ggtitle("Tendencia de retraso en la salida vs visibilidad") +
              theme(plot.title=element_text(size=12))
```



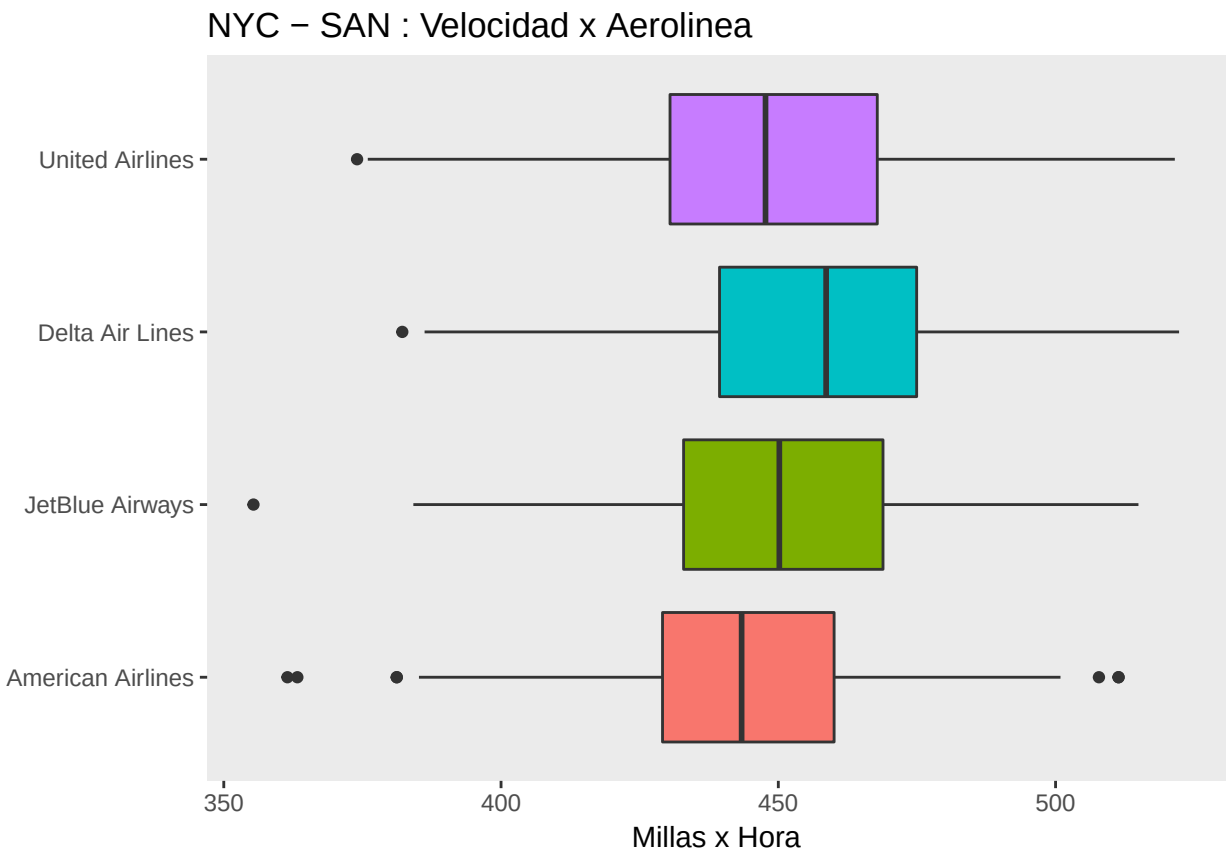
Otra causa de los retrasos en la salida de los vuelos puede ser un mal clima. Para poder detectar un mal clima podemos fijarnos en la presión atmósferica registrada. Una baja presión atmosférica puede ser señal de un mal clima (vientos fuertes, tormentas, ciclones, etc), y una alta presión atmosférica puede indicar un mejor clima.

```
flights_comp %>%
    select(dep_delay, pressure) %>%
    filter(!is.na(dep_delay) & !is.na(pressure)) %>%
    ggplot(aes(x = pressure, y = dep_delay)) +
    geom_smooth() +
    theme_bw(base_size = 10) +
    xlab("Presión atmosférica - hora de salida (millibarios)") +
    ylab("Promedio de retraso en la salida (minutos)") +
    ggtitle("Impacto de presión atmosférica en retrasod de salida")
```



Podemos suponer que San Diego es el punto mas alejado de Nueva York dentro de estados Unidos. Veamos que aerolinea tiene los aviones que cruzan más rápido todo el territorio de Estados Unidos

```
flights_comp %>%
    filter(dest == 'SAN') %>%
    mutate(millasxminuto = distance / air_time,
           millasxhora = millasxminuto * 60) %>%
    ggplot(aes(x = carrier, y = millasxhora, fill = carrier)) +
    geom_boxplot() +
    guides(fill=FALSE) +
    ggtitle("NYC - SAN : Velocidad x Aerolinea") +
    ylab("Millas x Hora") +
    coord_flip() +
    theme(axis.title.y = element_blank(),
          panel.grid.minor=element_blank(),
          panel.grid.major=element_blank()) +
    scale_x_discrete(breaks=c("AA", "B6", "DL", "UA"),
                     labels=c("American Airlines", "JetBlue Airways",
                              "Delta Air Lines", "United Airlines"))
```



Veamos qué aerolinea tiene mas retrasos en salidas y llegadas:

```
flights %>%
    rename(destination = dest,
           departure_delay = dep_delay,
           arrival_delay = arr_delay,
           time_in_air = air_time) %>%
    select(-year, -dep_time, -arr_time, -tailnum,
           -flight, -hour, -minute, -day) %>%
    mutate(ind_delayed_dep = ifelse(departure_delay > 0, 1, 0),
           ind_delayed_arr = ifelse(arrival_delay > 0, 1, 0)) %>%
    left_join(airlines, by = "carrier") %>%
    group_by(origin, name) %>%
    summarize(n_obs = n(),
              per_delayed_dep = round(sum(ind_delayed_dep, na.rm=TRUE) / n(),2),
              per_delayed_arr = round(sum(ind_delayed_arr, na.rm=TRUE) / n(),2)) %>%
    rename(Airport = origin,
           `Aerolinea` = name,
           `Numero de Vuelos` = n_obs,
           `Proporcion de salidas retrasadas` = per_delayed_dep,
           `Proporcion de llegadas retrasadas` = per_delayed_arr) %>%
    pander(style = "rmarkdown", split.tables = 200)
```

Airport	Aerolinea	Numero de Vuelos	Proporcion de salidas retrasadas	Proporcion de llegadas retrasadas
EWR	Alaska Airlines Inc.	714	0.32	0.26

Airport	Aerolinea	Numero de Vuelos	Proporcion de salidas retrasadas	Proporcion de llegadas retrasadas
EWR	American Airlines Inc.	3487	0.28	0.31
EWR	Delta Air Lines Inc.	4342	0.3	0.4
EWR	Endeavor Air Inc.	1268	0.21	0.29
EWR	Envoy Air	2276	0.36	0.45
EWR	ExpressJet Airlines Inc.	43939	0.44	0.47
EWR	JetBlue Airways	6557	0.35	0.39
EWR	SkyWest Airlines Inc.	6	0.5	0.33
EWR	Southwest Airlines Co.	6188	0.53	0.44
EWR	United Air Lines Inc.	46087	0.49	0.38
EWR	US Airways Inc.	4405	0.23	0.36
EWR	Virgin America	1566	0.35	0.31
JFK	American Airlines Inc.	13783	0.36	0.35
JFK	Delta Air Lines Inc.	20701	0.34	0.31
JFK	Endeavor Air Inc.	14651	0.41	0.37
JFK	Envoy Air	7193	0.34	0.44
JFK	ExpressJet Airlines Inc.	1408	0.37	0.44
JFK	Hawaiian Airlines Inc.	342	0.2	0.28
JFK	JetBlue Airways	42076	0.4	0.43
JFK	United Air Lines Inc.	4534	0.33	0.37
JFK	US Airways Inc.	2995	0.32	0.39
JFK	Virgin America	3596	0.47	0.35
LGA	AirTran Airways Corporation	3260	0.51	0.58
LGA	American Airlines Inc.	15459	0.27	0.31
LGA	Delta Air Lines Inc.	23067	0.3	0.36
LGA	Endeavor Air Inc.	2541	0.28	0.33
LGA	Envoy Air	16928	0.28	0.44
LGA	ExpressJet Airlines Inc.	8826	0.38	0.36
LGA	Frontier Airlines Inc.	685	0.5	0.57
LGA	JetBlue Airways	6002	0.38	0.47
LGA	Mesa Airlines Inc.	601	0.39	0.43
LGA	SkyWest Airlines Inc.	26	0.23	0.31
LGA	Southwest Airlines Co.	6087	0.54	0.42
LGA	United Air Lines Inc.	8044	0.38	0.36
LGA	US Airways Inc.	13136	0.22	0.35

Podemos ver que saliendo desde el aeropuerto **EWR** las peores opciones son:

- ExpressJet Airlines Inc.
- SkyWest Airlines Inc.
- Southwest Airlines Co.
- United Air Lines Inc

Saliendo desde **JFK** las peores opciones son:

- Virgin America
- Endeavor Air Inc.
- JetBlue Airways

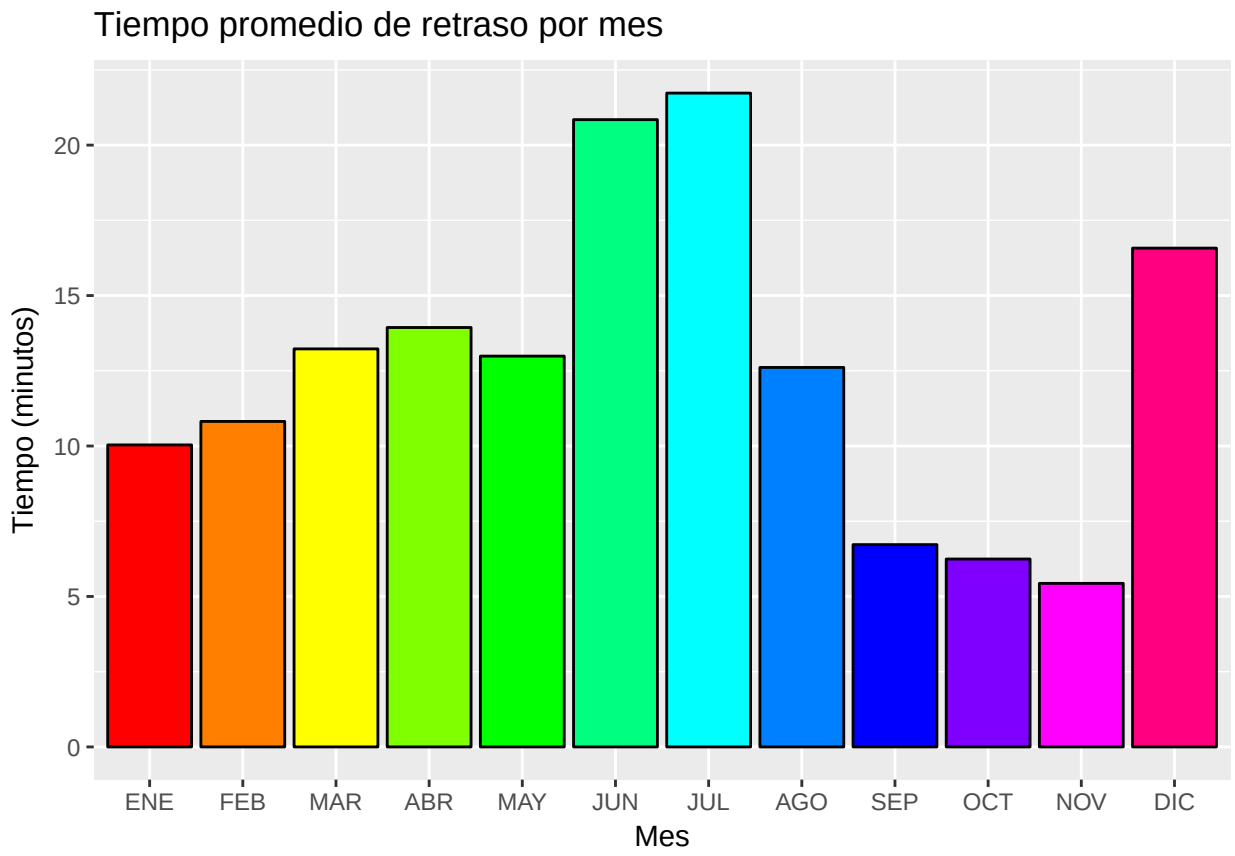
Saliendo desde **LGA** las peores opciones son:

- Southwest Airlines Co.
- AirTran Airways Corporation
- Frontier Airlines Inc.

Ahora estudiemos qué meses son los más caóticos en los aeropuertos de *NYC*

```
months_Str <- c("ENE","FEB","MAR","ABR","MAY","JUN",
                "JUL","AGO","SEP","OCT","NOV","DIC")

flights_comp %>%
    filter(!is.na(dep_delay) & !is.na(month)) %>%
    group_by(month) %>%
    summarize(mean_dep_delay = mean(dep_delay)) %>%
    ggplot(aes(x = factor(month, labels = months_Str),
               y = mean_dep_delay)) +
    geom_bar(stat = "identity", fill = rainbow(12), color = "black") +
    xlab("Mes") +
    ylab("Tiempo (minutos)") +
    ggtitle("Tiempo promedio de retraso por mes")
```



Claramente los meses mas caóticos son:

- Junio
- Julio
- Diciembre

Por último estudiemos cómo han cambiado el número de asientos a lo largo del tiempo

```
marca_avion <- flights_comp %>%
    select(carrier, tailnum, year.y,manufacturer,seats) %>%
    mutate(marca = ifelse((manufacturer == "AIRBUS" | manufacturer == "AIRBUS INDUSTRIE"), "AIRBUS", as

list <- c("AIRBUS",
          "BOEING",
          "EMBRAER")

ggplot(data = marca_avion %>%
           filter(marca %in% list) %>%
           select(marca, year.y, seats) %>%
           distinct(marca, year.y, seats),
       aes(x = year.y, y=seats, color = marca)) +
  geom_point() +
  theme_light() + xlab("años") + ylab("numero de asientos") +
  ggtitle("¿Cómo ha cambiado el numero de asientos a lo largo del tiempo?")
```