Ejercicio en clase: Análisis Exploratorio de los Datos - Visualización de Datos

Introducción

En este documento se tratan dos conceptos muy importantes a considerar un proyecto de Ciencia de Datos: *Visualización de Datos* y *Análisis Exploratorio de los Datos (EDA).*

Análisis Exploratorio de los Datos

El Análisis Exploratorio de Datos tiene como objetivo, *examinar los datos* para conseguir un *entendimiento básico* con respecto al comportamiento de los datos y algunas *relaciones* existentes entre ellos. Además, este análisis nos permite encontrar problemas de calidad no detectados con anterioridad, como datos ausentes, casos atípicos, etc. [1]

Para enriquecer el Análisis Exploratorio de los Datos, se utilizan medidas estadísticas básicas, tablas y gráficas, sin embargo, su cálculo no es suficiente; es necesario plasmar los resultados y la información construida, lo más claro y sencillo posible, es por eso, que se siguen los principios básicos del concepto de *visualización de datos*.

Visualización de los Datos

Personas importantes:

John Wilder Tukey: Fundador del Análisis Exploratorio de Datos o EDA (Exploratory Data Analysis). Su libro Exploratory Data Analysis (1977) es el clásico sobre este tema. EDA es una filosofía básicamente gráfica de exploración de datos estadísticos. [2]

Edward Tufte: (Estadístico y Artista) De acuerdo al New York Times, Edward Tufte es el Leonardo da Vinci de los datos". [3]

Charles Joseph Minard: Ingeniero civil francés reconocido por su importante aportación en el terreno de los gráficos. [4]

Principios del Diseño Analítico - Edward Tufte [5]

1. Muestra comparaciones.
2. Muestra causalidad.
3. Utiliza datos multivariados.
4. Modos de Integración completos (palabras, números, imágenes y diagramas).
5. Establecer credibilidad.
6. Se centra en el contenido.

Chartjunk

El *chartjunk* son aquellos elementos gráficos que no corresponden a variación de datos, o que entorpecen la interpretación de una gráfica.[6] * Todo lo que quita atención de los datos.

Charles Joseph Minard - Marcha de Napoleón sobre Moscú [6]

"Bien podría ser el mejor gráfico estadístico jamás dibujado."

"Cuenta una historia rica y coherente con sus datos multivariados, mucho más esclarecedora que un solo número que rebota en el tiempo."

-Edward Tufte.

Variables: tropas de Napoleón, distancia, temperatura, latitud y longitud, dirección en que viajaban las tropas y la localización relativa a fechas específicas.

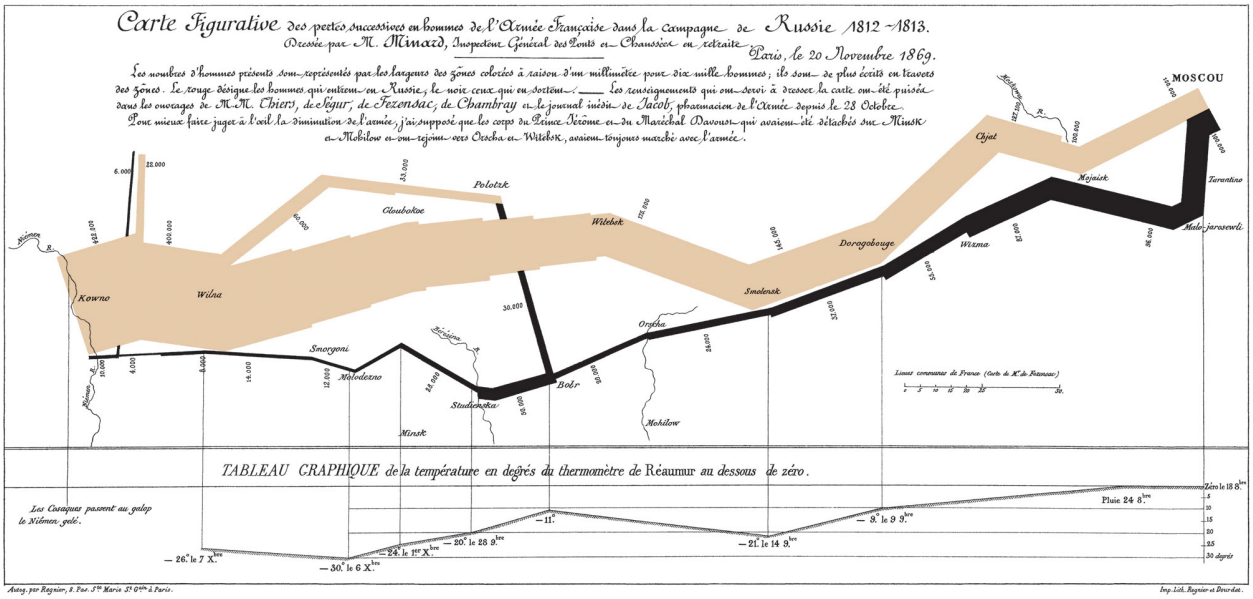


Figure 1: Charles Joseph Minard - Marcha de Napoleón sobre Moscú.

Ejercicio en Clase

```
#Librería base
library(tidyverse)
```

```
#install.packages("ggplot2")
library(ggplot2)
```

ggplot2

Tips para manejar ggplot2: https://rstudio.com/wp-content/uploads/2015/04/ggplot2-spanish.pdf

Notación básica:

ggplot(dataframe, aes(x=, y=, fill=)+

geom_bar()

geom_point()

geom_boxplot()

geom_line()

MPG - MILLES PER GALLON

```
ejercicio_2 <- mpg
head(ejercicio_2)
```

```
## # A tibble: 6 x 11
##   manufacturer model displ  year   cyl trans  drv     cty   hwy fl    class
##   <chr>        <chr> <dbl> <int> <int> <chr>  <chr> <int> <int> <chr> <chr>
## 1 audi         a4      1.8  1999     4 auto(~ f        18    29 p     comp~
## 2 audi         a4      1.8  1999     4 manua~ f        21    29 p     comp~
## 3 audi         a4      2    2008     4 manua~ f        20    31 p     comp~
## 4 audi         a4      2    2008     4 auto(~ f        21    30 p     comp~
## 5 audi         a4      2.8  1999     6 auto(~ f        16    26 p     comp~
## 6 audi         a4      2.8  1999     6 manua~ f        18    26 p     comp~
```

```
#?mpg
```

Variables

Manufacturer: Empresa manufacturera.

Model: Modelo

Displ: Desplazamiento del motor (en litros)

Year: Año de creación

Cyl: Número de cilindros

Trans: Tipo de transmisión

Drv: f = front-wheel drive, r = rear wheel drive, 4 = 4wd

City: City miles per gallon (Millas en ciudad por galón).

Hwy: Highway miles per gallon (Millas en carretera por galón).

Fl: Tipo de combustible

Class: Tipo de coche

```
summary(ejercicio_2)
```

```
##  manufacturer          model               displ            year
##  Length:234         Length:234         Min.   :1.600   Min.   :1999
##  Class :character   Class :character   1st Qu.:2.400   1st Qu.:1999
##  Mode  :character   Mode  :character   Median :3.300   Median :2004
##                                        Mean   :3.472   Mean   :2004
##                                        3rd Qu.:4.600   3rd Qu.:2008
##                                        Max.   :7.000   Max.   :2008
##       cyl           trans               drv                 cty
```

```
##  Min.   :4.000   Length:234         Length:234         Min.   : 9.00
##  1st Qu.:4.000   Class :character   Class :character   1st Qu.:14.00
##  Median :6.000   Mode  :character   Mode  :character   Median :17.00
##  Mean   :5.889                                         Mean   :16.86
##  3rd Qu.:8.000                                         3rd Qu.:19.00
##  Max.   :8.000                                         Max.   :35.00
##       hwy             fl                class
##  Min.   :12.00   Length:234         Length:234
##  1st Qu.:18.00   Class :character   Class :character
##  Median :24.00   Mode  :character   Mode  :character
##  Mean   :23.44
##  3rd Qu.:27.00
##  Max.   :44.00
```

Preguntas

1. ¿Cuántos años tenemos de registros?

```
pregunta_1 <- ejercicio_2 %>%
  select(year) %>%
  arrange(year) %>%
  unique()

pregunta_1
```

```
## # A tibble: 2 x 1
##    year
##   <int>
## 1  1999
## 2  2008
```

2. ¿Cuántas marcas tenemos?

```
pregunta_2 <- ejercicio_2 %>%
  select(manufacturer) %>%
  arrange(manufacturer) %>%
  unique()

pregunta_2
```

```
## # A tibble: 15 x 1
##    manufacturer
##    <chr>
##  1 audi
##  2 chevrolet
##  3 dodge
##  4 ford
##  5 honda
##  6 hyundai
##  7 jeep
##  8 land rover
```

```
##  9 lincoln
## 10 mercury
## 11 nissan
## 12 pontiac
## 13 subaru
## 14 toyota
## 15 volkswagen
```

3. ¿Cuántos modelos distintos hay?

```
pregunta_3 <- ejercicio_2 %>%
  select(model) %>%
  arrange(model) %>%
  unique()

pregunta_3
```

```
## # A tibble: 38 x 1
##    model
##    <chr>
##  1 4runner 4wd
##  2 a4
##  3 a4 quattro
##  4 a6 quattro
##  5 altima
##  6 c1500 suburban 2wd
##  7 camry
##  8 camry solara
##  9 caravan 2wd
## 10 civic
## # ... with 28 more rows
```

4. ¿Cuántos tipos de transmición hay?

```
pregunta_4 <- ejercicio_2 %>%
  select(trans) %>%
  arrange(trans) %>%
  unique()

pregunta_4
```

```
## # A tibble: 10 x 1
##    trans
##    <chr>
##  1 auto(av)
##  2 auto(l3)
##  3 auto(l4)
##  4 auto(l5)
##  5 auto(l6)
##  6 auto(s4)
##  7 auto(s5)
##  8 auto(s6)
##  9 manual(m5)
## 10 manual(m6)
```

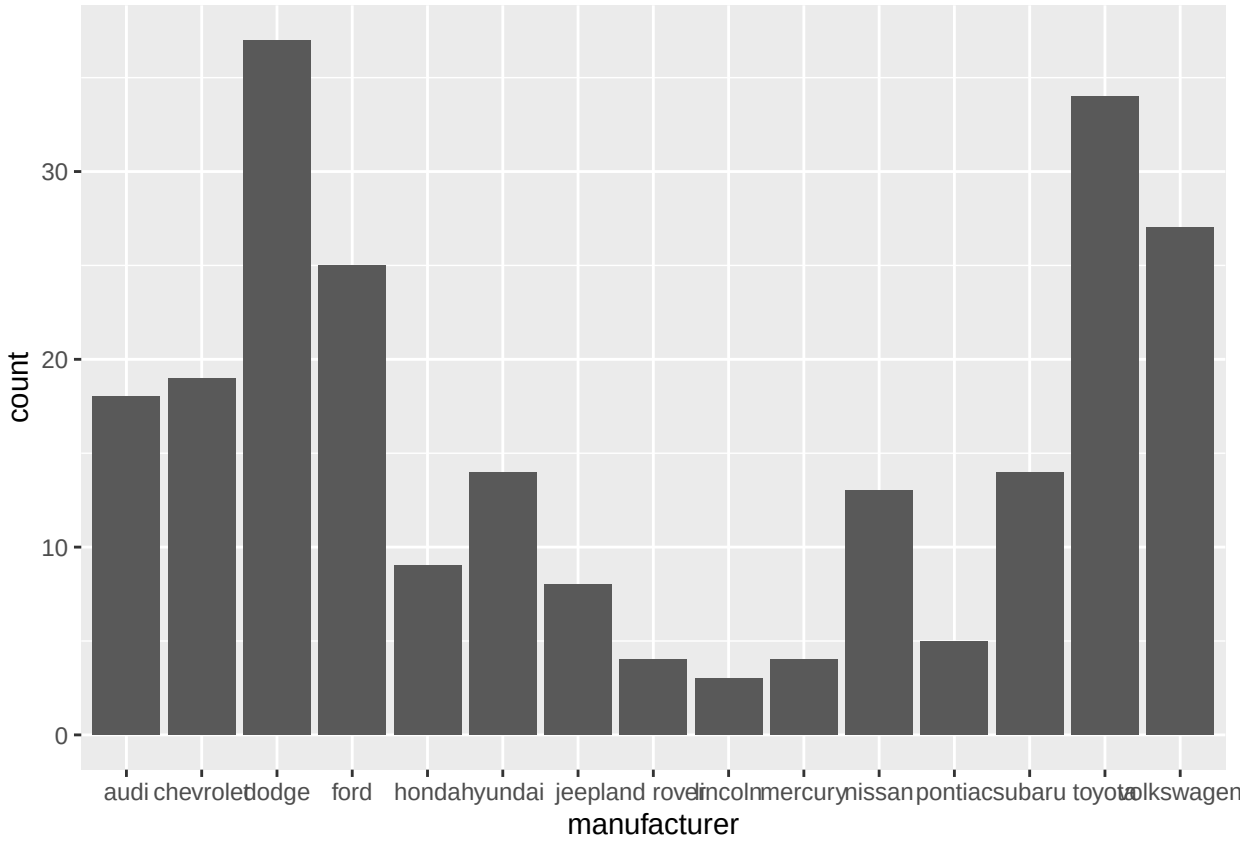
5. ¿Cuántos tipos de combustuble hay?

```
pregunta_5 <- ejercicio_2 %>%
  select(fl) %>%
  arrange(fl) %>%
  unique()

pregunta_5
```

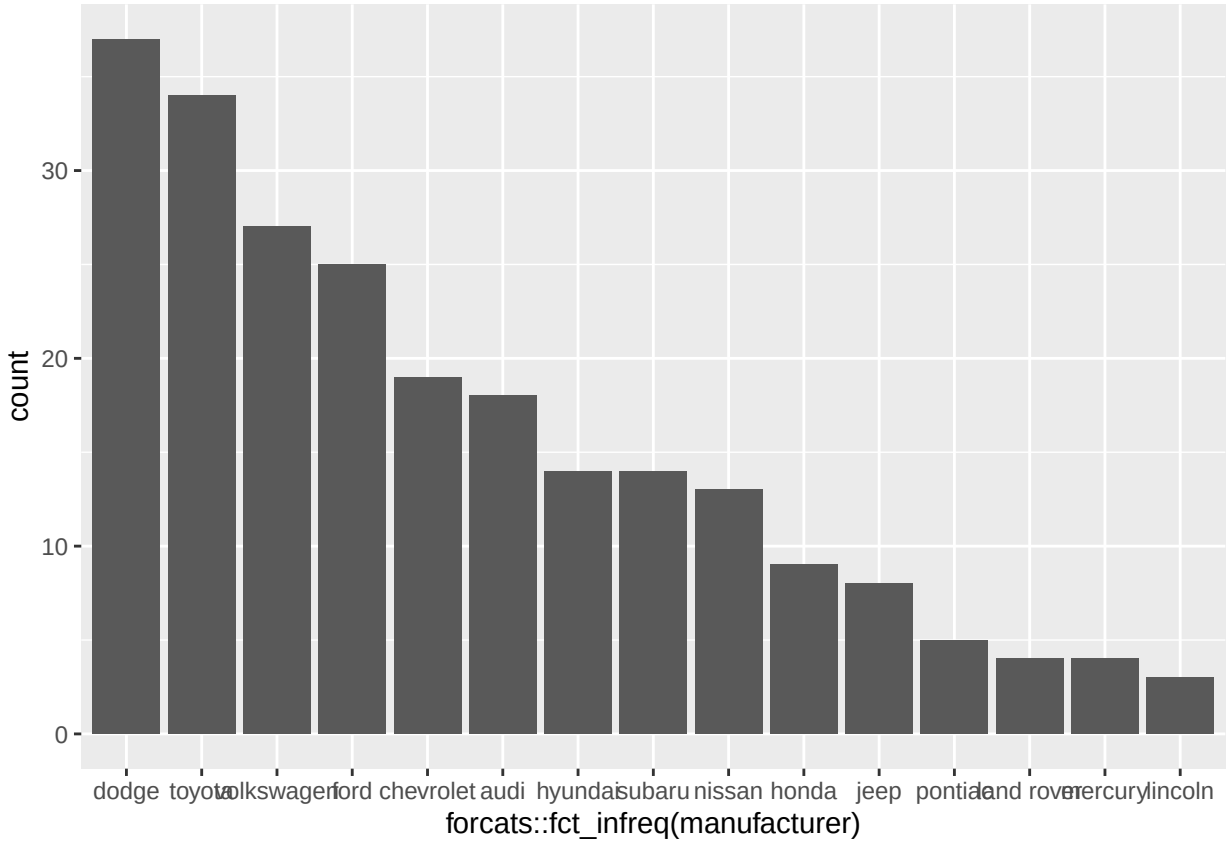
```
## # A tibble: 5 x 1
##   fl
##   <chr>
## 1 c
## 2 d
## 3 e
## 4 p
## 5 r
```

6. Con relación a las marcas manufactureras, ¿Cómo se comportan los registros? - ¿Qué marca manufacturera tiene más registros?

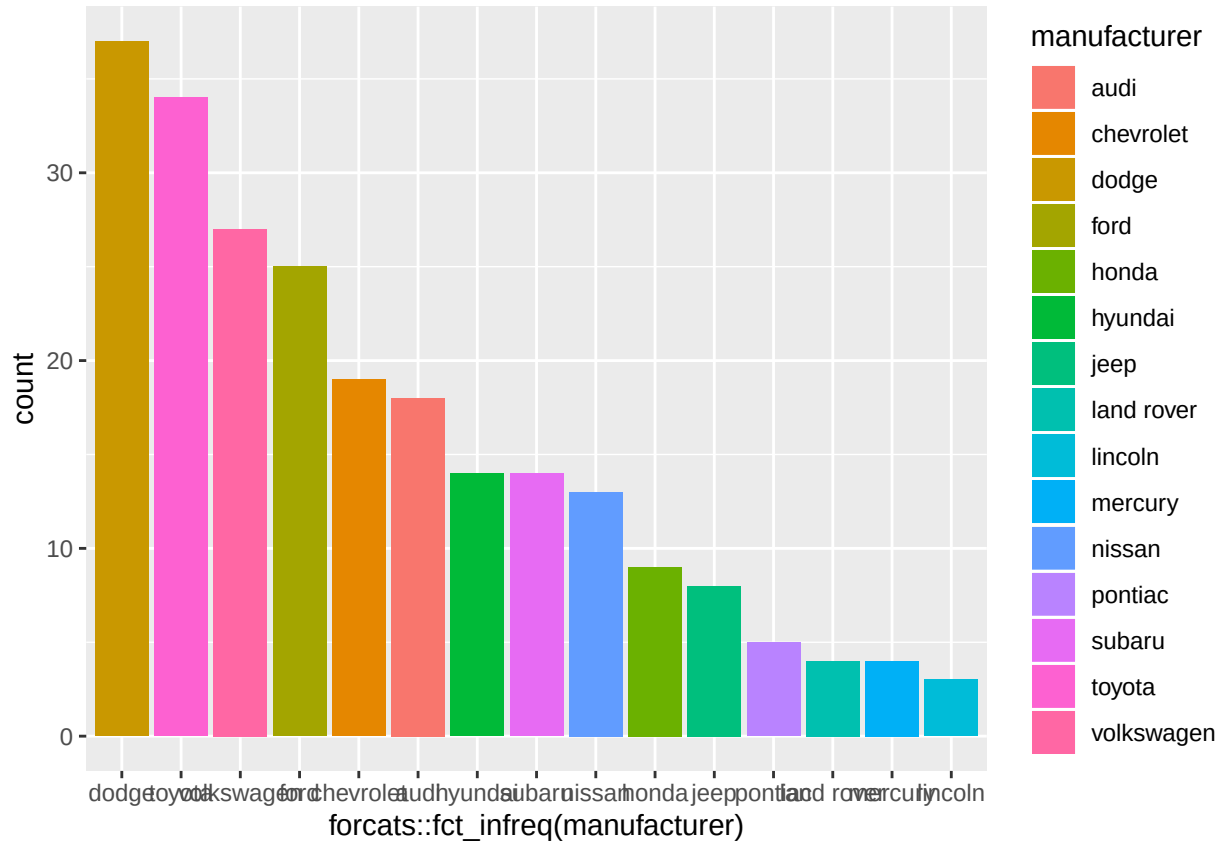
```
im_pg6 <- ggplot(ejercicio_2, aes(x = manufacturer)) +
  geom_bar()
im_pg6
```



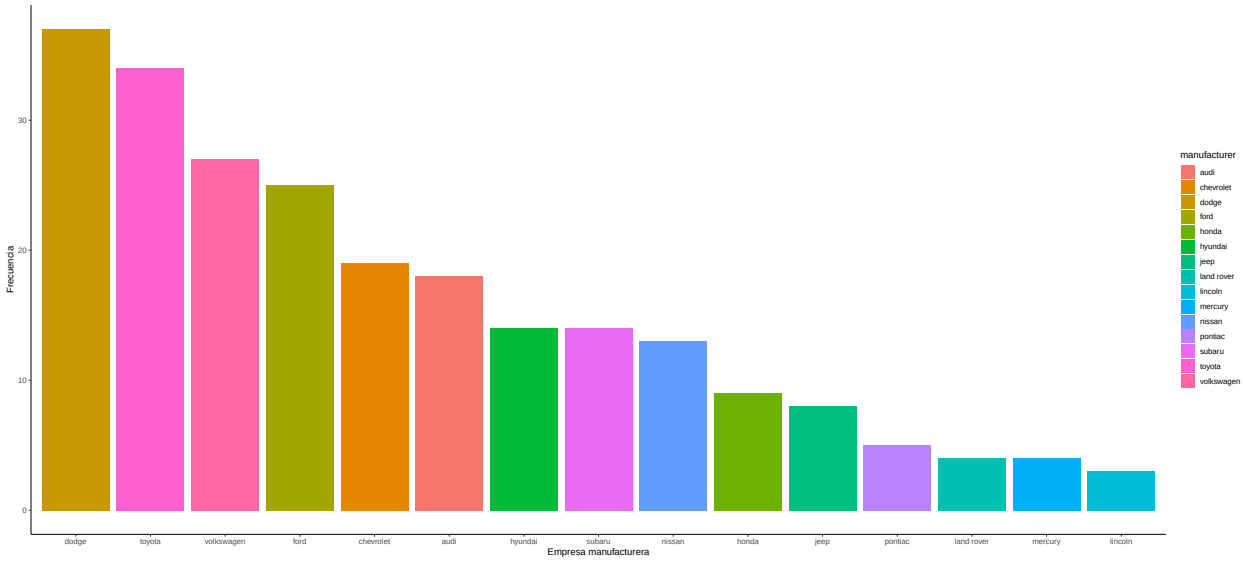
```
# fct_infreq: Reorder factors levels by first appearance, frequency, or numeric order.

im_pg6 <- ggplot(ejercicio_2, aes(x = forcats::fct_infreq(manufacturer))) +
  geom_bar()
im_pg6
```




```
im_pg6 <- ggplot(ejercicio_2, aes(x = forcats::fct_infreq(manufacturer), fill = manufacturer)) +
  geom_bar()
im_pg6
```

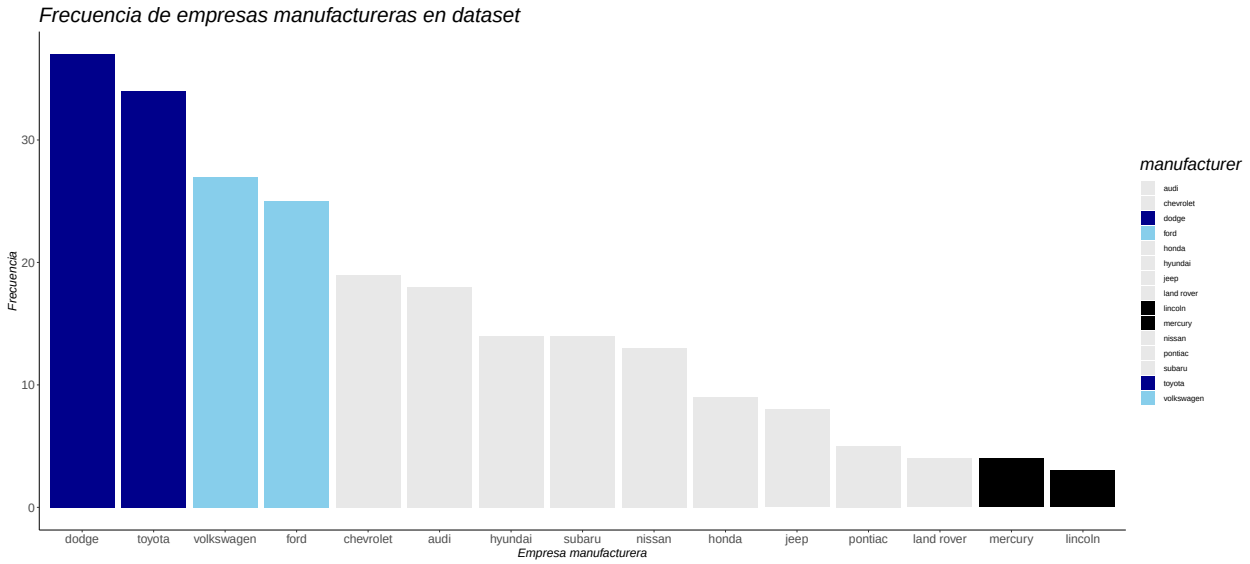



```
im_pg6 <- ggplot(ejercicio_2, aes(x = forcats::fct_infreq(manufacturer), fill = manufacturer)) +
  geom_bar() +
  theme_classic()+
  xlab("Empresa manufacturera") +
  ylab("Frecuencia")

im_pg6
```



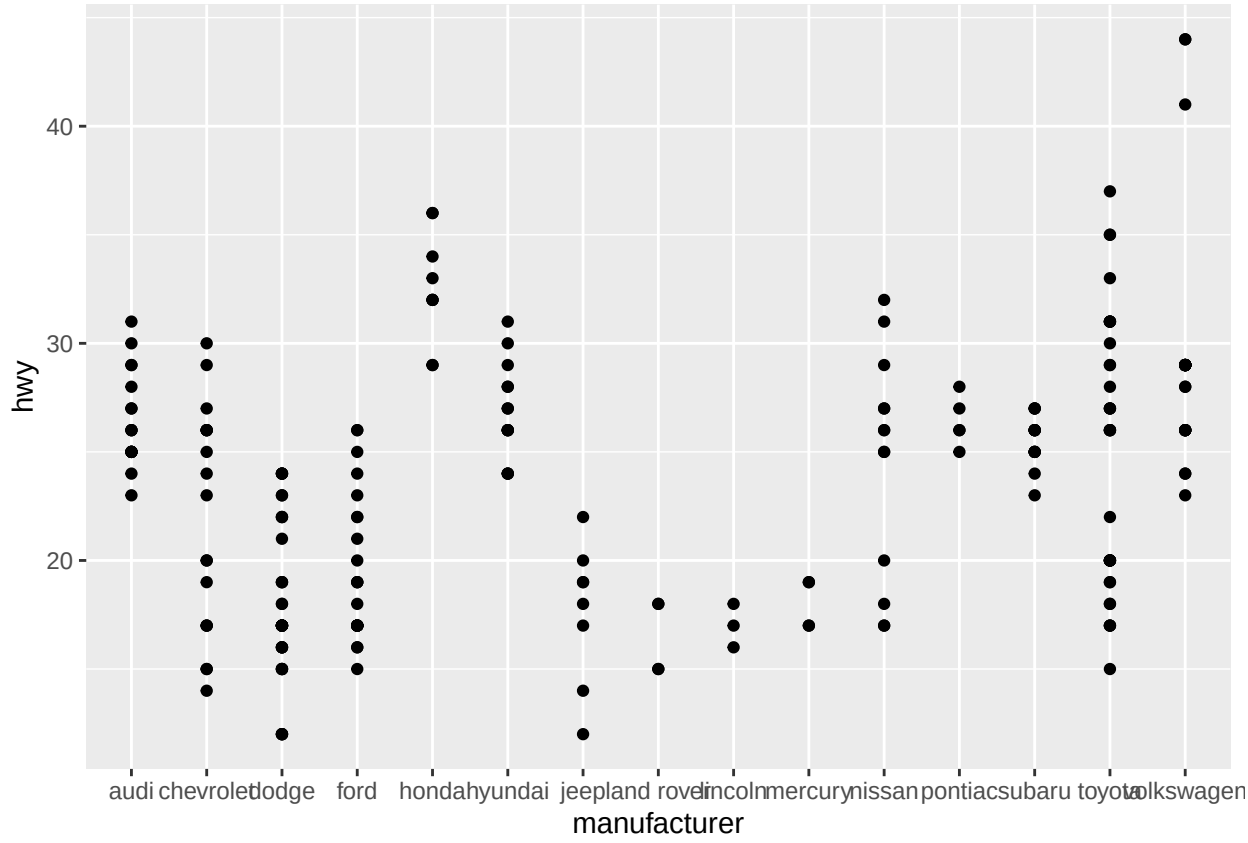
Colores: http://applied-r.com/wp-content/uploads/2019/01/rcolors_byname.png

```
im_pg6 <- ggplot(ejercicio_2, aes(x = forcats::fct_infreq(manufacturer), fill = manufacturer)) +
  geom_bar() +
  theme_classic()+
  xlab("Empresa manufacturera") +
  ylab("Frecuencia")+
  scale_fill_manual(values=c("gray91","gray91","darkblue","skyblue","gray91","gray91","gray91","gray91"
  theme(axis.text=element_text(size=14),
        axis.title=element_text(size=14,face="italic"),
        title = element_text(size=20,face="italic"))+
  labs(title="Frecuencia de empresas manufactureras en dataset")

im_pg6
```

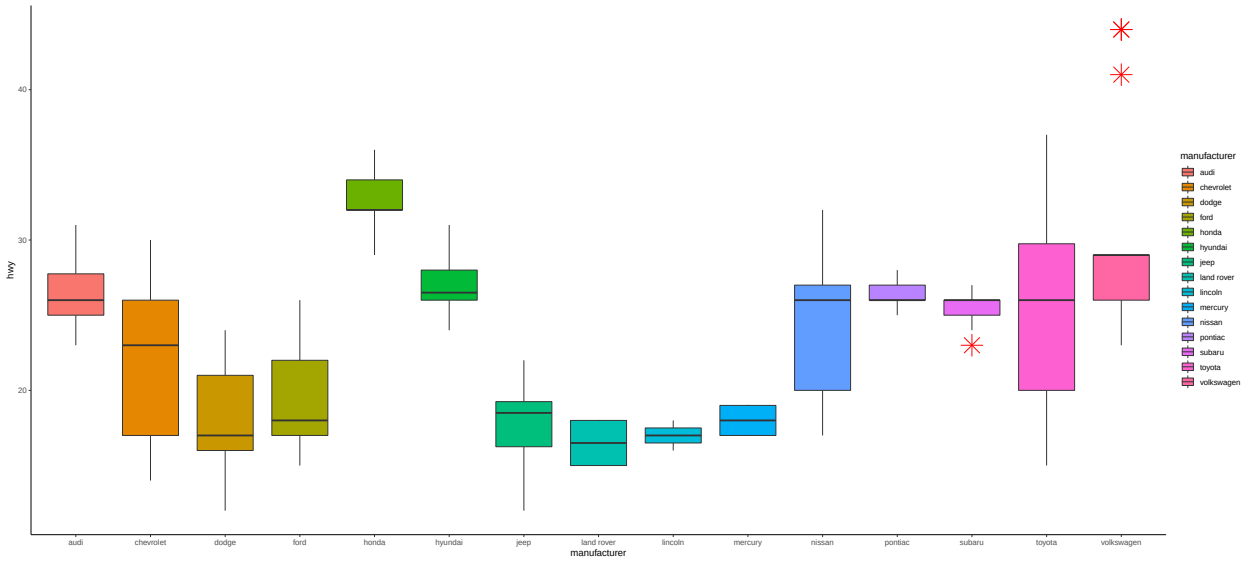


7. ¿Cómo se comportan las diferentes empresas manunfactureras con respecto a las millas por galón en autopista?

```
im_pg7 <- ggplot(ejercicio_2, aes(x = manufacturer, y = hwy)) +
  geom_point()

im_pg7
```



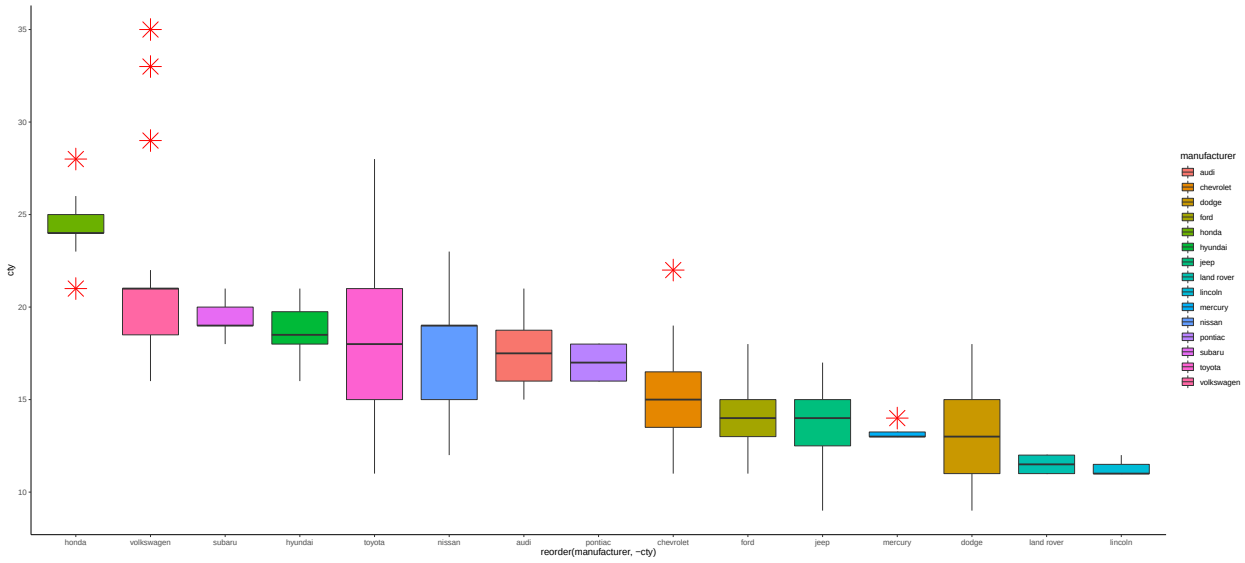

```
im_pg7 <- ggplot(ejercicio_2, aes(x = manufacturer, y = hwy, fill=manufacturer)) +
  geom_boxplot(outlier.colour="red", outlier.shape=8, outlier.size=8)+
  theme_classic()

im_pg7
```



8. ¿Cómo se comportan las diferentes empresas manunfactureras con respecto a las millas por galón en

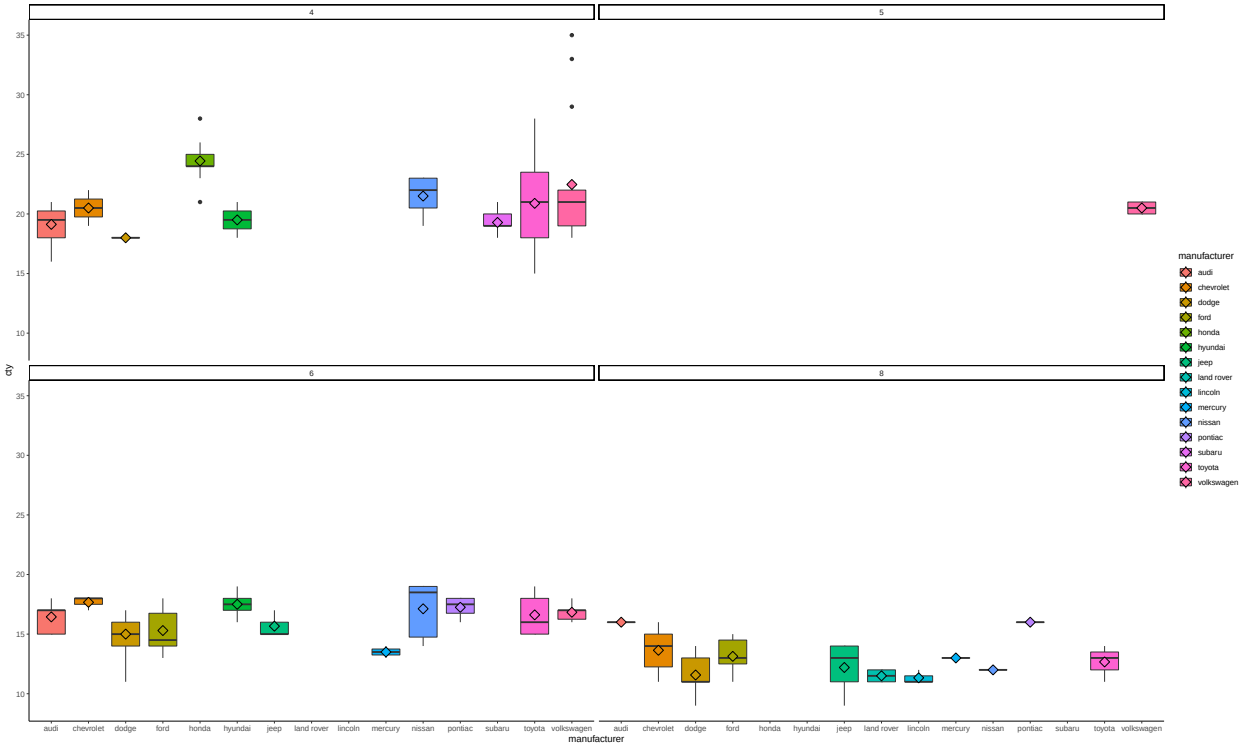
ciudad?

```
im_pg8<- ggplot(ejercicio_2, aes(x = reorder(manufacturer,-cty), y = cty, fill=manufacturer)) +
  geom_boxplot(outlier.colour="red", outlier.shape=8, outlier.size=8)+
  theme_classic()

im_pg8
```



9. Comportamiento anterior, haciendo un análisis por número de cilindros

```
im_pg9 <- ggplot(ejercicio_2, aes(x = manufacturer, y = cty, fill=manufacturer)) +
  geom_boxplot() +
  facet_wrap(~ cyl) +
  theme_classic() +
  stat_summary(fun.y=mean, geom="point", shape=23, size=4)

im_pg9
```



BABY NAMES

```
#install.packages("babynames")
library(babynames)
```

```
ejercicio_3 <- babynames
glimpse(ejercicio_3)
```

```
## Observations: 1,924,665
## Variables: 5
## $ year <dbl> 1880, 1880, 1880, 1880, 1880, 1880, 1880, 1880, 1880, 188...
## $ sex  <chr> "F", "F", "F", "F", "F", "F", "F", "F", "F", "F", "F", "F...
## $ name <chr> "Mary", "Anna", "Emma", "Elizabeth", "Minnie", "Margaret"...
## $ n    <int> 7065, 2604, 2003, 1939, 1746, 1578, 1472, 1414, 1320, 128...
## $ prop <dbl> 0.07238359, 0.02667896, 0.02052149, 0.01986579, 0.0178884...
```

```
head(ejercicio_3)
```

```
## # A tibble: 6 x 5
##    year sex   name          n   prop
##   <dbl> <chr> <chr>     <int>  <dbl>
## 1  1880 F     Mary       7065 0.0724
## 2  1880 F     Anna       2604 0.0267
## 3  1880 F     Emma       2003 0.0205
## 4  1880 F     Elizabeth  1939 0.0199
## 5  1880 F     Minnie     1746 0.0179
## 6  1880 F     Margaret   1578 0.0162
```

1. ¿Cuántos años hay de registro?

Registros desde 1880 hasta 2017.

```
ejercicio_3 %>% select(year)%>%unique()
```

```
## # A tibble: 138 x 1
##     year
##    <dbl>
##  1  1880
##  2  1881
##  3  1882
##  4  1883
##  5  1884
##  6  1885
##  7  1886
##  8  1887
##  9  1888
## 10  1889
## # ... with 128 more rows
```

2. ¿Cuántos nombres distintos hay?

```
ejercicio_3 %>% select(name)%>%unique()
```

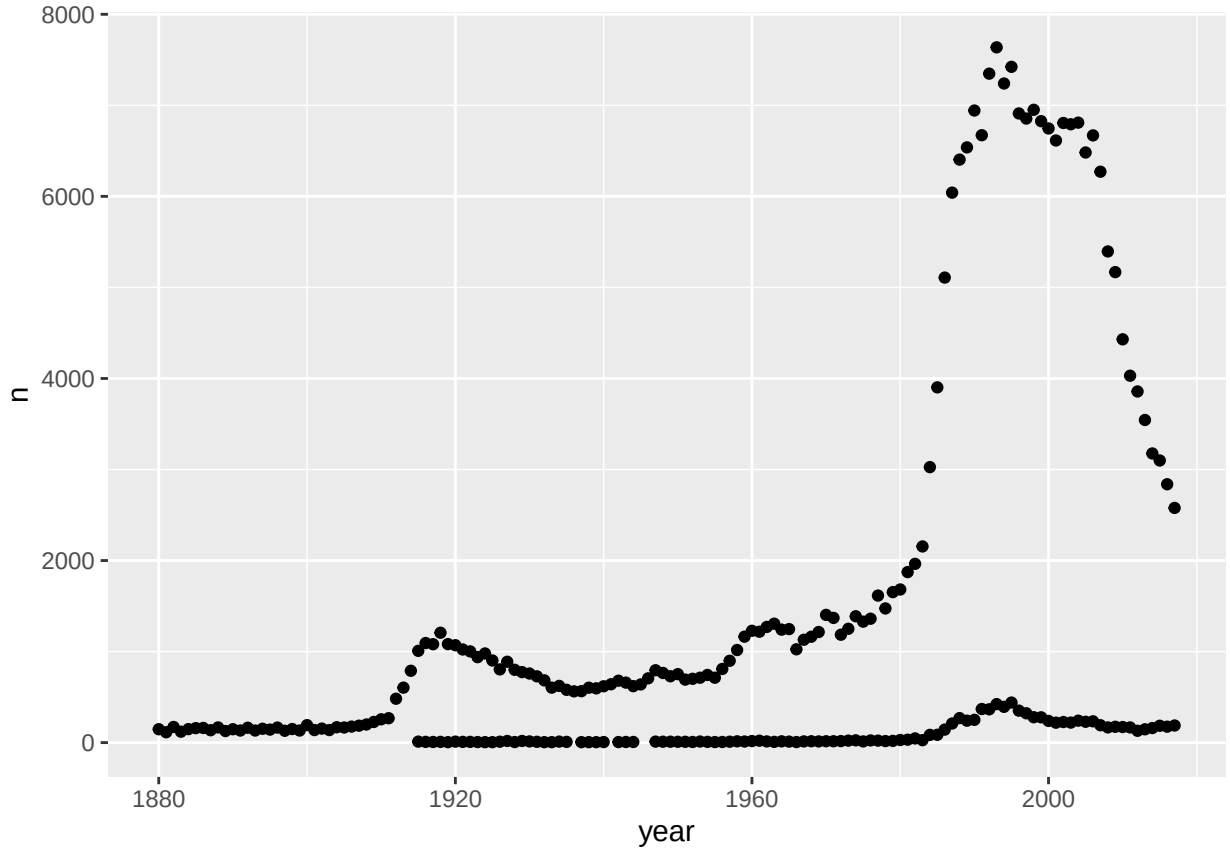
```
## # A tibble: 97,310 x 1
##    name
##    <chr>
##  1 Mary
##  2 Anna
##  3 Emma
##  4 Elizabeth
##  5 Minnie
##  6 Margaret
##  7 Ida
##  8 Alice
##  9 Bertha
## 10 Sarah
## # ... with 97,300 more rows
```

3. Comportamiento de nombre_____ a lo largo de los años

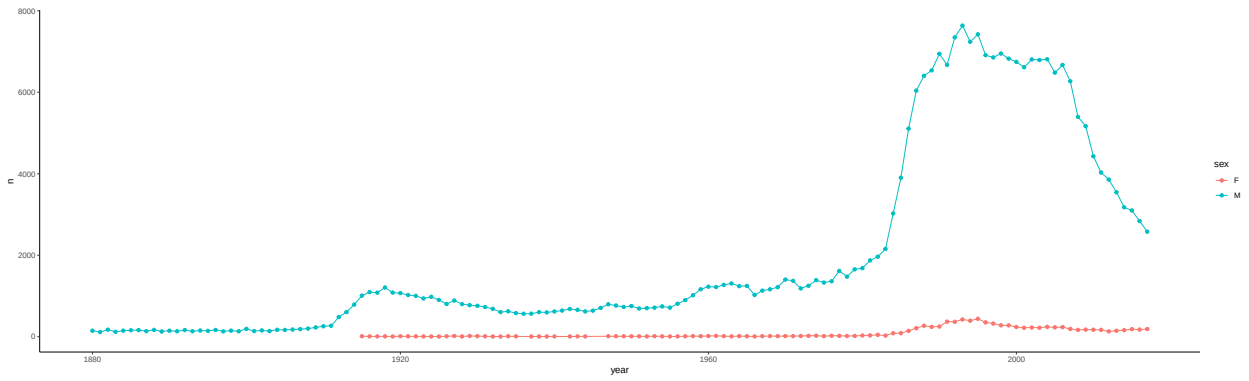
Caso 1: Alex

```
caso1 <- ejercicio_3%>%filter(name=="Alex")
```

```
ggplot(caso1, aes(x = year, y = n)) +
  geom_point()
```




```
ggplot(caso1, aes(x = year, y = n, color=sex)) +
  geom_point()+
  geom_line() +
  theme_classic()
```

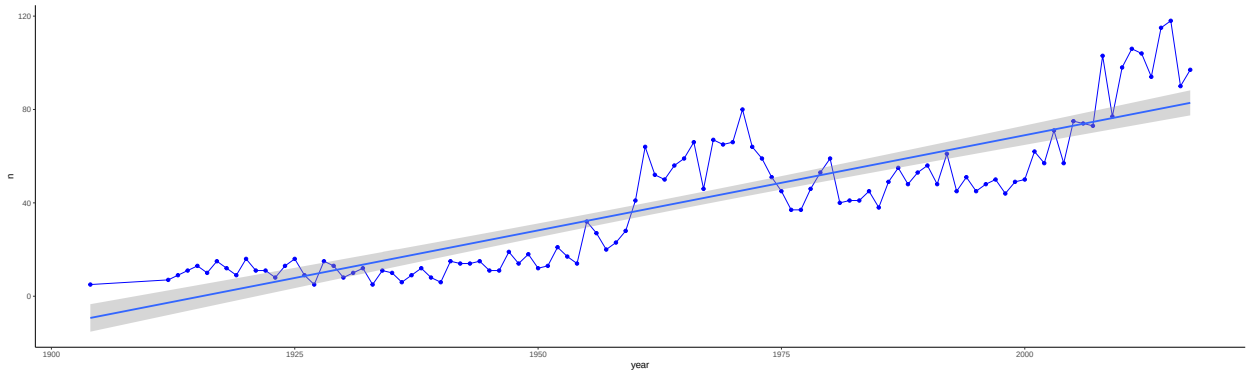


Caso 2: Thor

```
caso2 <- ejercicio_3%>%filter(name=="Thor")
```

```
ggplot(caso2, aes(x = year, y = n)) +
  geom_point(color="blue")+
  geom_line(color="blue") +
```

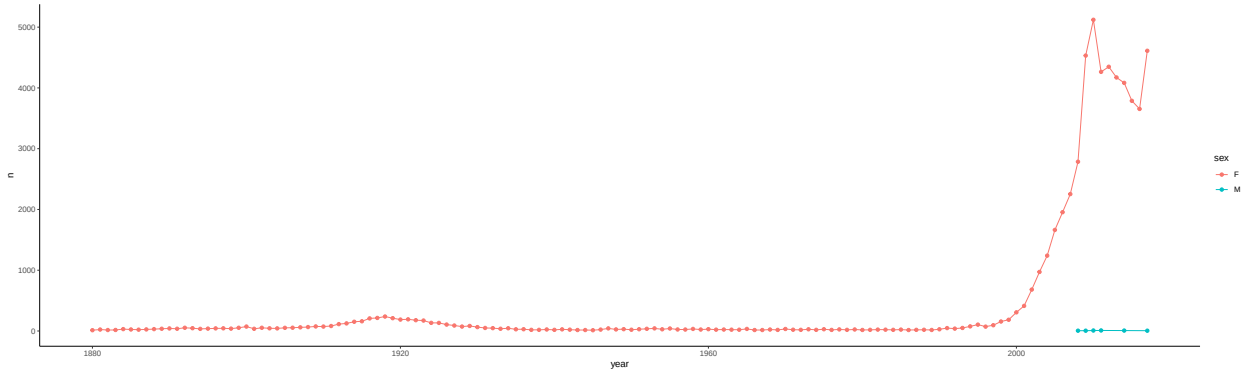
```
  theme_classic() +
  geom_smooth(method = "lm")
```



Caso 3: Bella

```
caso3 <- ejercicio_3 %>% filter(name=="Bella")
```

```
ggplot(caso3, aes(x = year, y = n, color=sex)) +
  geom_point()+
  geom_line() +
  theme_classic()
```



Referencias

[1] Salvador Figueras, M y Gargallo, P. (2003). "Análisis Exploratorio de Datos". 03 de septiembre de 2020, de 5campus.com Sitio web: https://ciberconta.unizar.es/leccion/aed/ead.pdf

[2] Smyers, K. "John Wilder Tukey: The Pioneer of Big Data and Visualization". (2013). 03 de septiembre de 2020, de Control Trends Sitio web: https://controltrends.org/controltalk-now-2/control-talk/05/john-wilder-tukey-the-pioneer-of-big-data-and-visualization/

[3] Graphics Press. "The work of Edward Tufte and Graphics Press". (s.f). 03 de septiembre de 2020, de EdwardTufte Sitio web: https://www.edwardtufte.com/tufte/

[4] INE. "Tercera etapa: 1851-1900 / Charles Joseph Minard (1781-1870)". (s.f). 03 de septiembre de 2020, de INE Sitio web: https://www.ine.es/expo_graficos2010/expogra_autor3.htm

[5] Sites Google. "Tufte on Design and Data". (s.f). 03 de septiembre de 2020, de Sites Google Sitio web: https://sites.google.com/site/tufteondesign/home/six-fundamental-principles-of-design#:~:text=Tufte%20suggests%20six%20fundamental%20principles,credibility%2C%20and%20focus%20on%20content.&text=For%20each%20principle%2C%20we%20outline,it%20to%20improve%20your%20visualizations.

[6] Ortiz, T. "Estadística Computacional". (2018). 03 de septiembre de 2020, de Github Sitio web: https://github.com/tereom/est-computacional-2018