

Soluciones Ejercicios Sesión 5

Capacitadores R en DET

18-12-2020

Objetivo

Desarrollar visualizaciones simples y claras con el paquete `ggplot2`.

Primer ejercicio

Descargar la base de datos WDI INDICATORS LA.xlsx y cargarla en el ambiente de R.

```
library(readxl)
library(tidyverse)

latam <- read_excel("data/WDI INDICATORS LA.xlsx",
                   na="..") ## Señalamos que reconozca a ".."
```

Segundo ejercicio

¿Cuántas filas y columnas tiene la base de datos?

¿Es práctico el actual formato de la base para trabajar con `ggplot2`?

¿Que modificaciones le harías a la base de datos?

(te recomendamos aplicarle `pivot_longer`)

¿Cuántas filas y columnas tiene la nueva base de datos?

```
dim(latam)
```

```
## [1] 23040    64
```

```
names(latam)
```

```
## [1] "country"      "country_code" "serie"         "1960"         "1961"
## [6] "1962"         "1963"         "1964"         "1965"         "1966"
```

```
## [11] "1967"      "1968"      "1969"      "1970"      "1971"
## [16] "1972"      "1973"      "1974"      "1975"      "1976"
## [21] "1977"      "1978"      "1979"      "1980"      "1981"
## [26] "1982"      "1983"      "1984"      "1985"      "1986"
## [31] "1987"      "1988"      "1989"      "1990"      "1991"
## [36] "1992"      "1993"      "1994"      "1995"      "1996"
## [41] "1997"      "1998"      "1999"      "2000"      "2001"
## [46] "2002"      "2003"      "2004"      "2005"      "2006"
## [51] "2007"      "2008"      "2009"      "2010"      "2011"
## [56] "2012"      "2013"      "2014"      "2015"      "2016"
## [61] "2017"      "2018"      "2019"      "2020"
```

```
latam %>% select(country,serie,'1960','2000','2010') %>% head()
```

```
## # A tibble: 6 x 5
##   country  serie                                '1960' '2000' '2010'
##   <chr>    <chr>                                <dbl>  <dbl>  <dbl>
## 1 Argentina Access to electricity (% of population)      NA     NA    98.8
## 2 Argentina Access to clean fuels and technologies for coo~    NA    94.8   97.6
## 3 Argentina Access to electricity, rural (% of rural popul~    NA     NA    90.2
## 4 Argentina Access to electricity, urban (% of urban popul~    NA     NA    99.7
## 5 Argentina Account ownership at a financial institution o~    NA     NA     NA
## 6 Argentina Account ownership at a financial institution o~    NA     NA     NA
```

```
latam2 <- latam %>% pivot_longer(cols = -c(country,country_code,serie),
                                names_to = "anio")
```

```
dim(latam2)
```

```
## [1] 1405440      5
```

```
latam2 %>% arrange(-value) %>% head()
```

```
## # A tibble: 6 x 5
##   country country_code serie                                anio  value
##   <chr>    <chr>    <chr>                                <chr>  <dbl>
## 1 Colombia COL      Gross national expenditure (current LCU) 2019  1.13e15
## 2 Colombia COL      GDP (current LCU)                        2019  1.06e15
## 3 Colombia COL      GDP: linked series (current LCU)          2019  1.06e15
## 4 Colombia COL      Gross national expenditure (current LCU) 2018  1.03e15
## 5 Colombia COL      GNI (current LCU)                        2019  1.03e15
## 6 Colombia COL      GNI: linked series (current LCU)          2019  1.03e15
```

Tercer ejercicio

¿Cuántos países existen en la base de datos?

```
## Opción más simple: tabla y contar a mano.
```

```
table(latam2$country)
```

```
##
##      Argentina      Bolivia      Brazil      Chile
##      87840          87840      87840      87840
##      Colombia      Costa Rica      Cuba Dominican Republic
##      87840          87840      87840      87840
##      Ecuador      El Salvador      Honduras      Panama
##      87840          87840      87840      87840
##      Paraguay      Peru      Uruguay      Venezuela, RB
##      87840          87840      87840      87840
```

```
## Otra opción
```

```
latam2 %>% group_by(country) %>% tally() %>% dim()
```

```
## [1] 16 2
```

Crea una nueva base de datos que solamente contenga datos para un país (el que tú quieras). Dale el nombre del país seleccionado a esta nueva base de datos.

```
cuba<-latam2 %>% filter(country=="Cuba")
```

```
cuba %>% arrange(-value) %>% head()
```

```
## # A tibble: 6 x 5
##   country country_code serie          anio      value
##   <chr>    <chr>      <chr>          <chr>    <dbl>
## 1 Cuba    CUB        GDP (current LCU)      2018    1.00e11
## 2 Cuba    CUB        GDP (current US$)      2018    1.00e11
## 3 Cuba    CUB        GDP: linked series (current LCU) 2018    1.00e11
## 4 Cuba    CUB        Gross value added at basic prices (GVA)~ 2018    9.89e10
## 5 Cuba    CUB        Gross value added at basic prices (GVA)~ 2018    9.89e10
## 6 Cuba    CUB        Gross national expenditure (current LCU) 2018    9.81e10
```

Filtra la base del país, dejando solamente los datos de una variable (*serie*), la que te parezca más interesante y que ojalá no tenga muchos valores perdidos (NA o ..).

```
## Con esta línea vemos cuáles variables tienen menos NA
cuba %>% filter(is.na(value)) %>% group_by(serie) %>% tally() %>% arrange(n) %>% head()
```

```
## # A tibble: 6 x 2
##   serie                                n
##   <chr>                             <int>
## 1 Age dependency ratio (% of working-age population)      1
## 2 Age dependency ratio, old (% of working-age population)  1
## 3 Age dependency ratio, young (% of working-age population) 1
## 4 Fixed telephone subscriptions                          1
## 5 Fixed telephone subscriptions (per 100 people)           1
## 6 Merchandise exports (current US$)                       1
```

```
## Resulta interesante la mortalidad infantil.
cuba_mortalidad<-cuba %>% filter(serie=="Mortality rate, infant (per 1,000 live births)")
cuba_mortalidad %>% head()
```

```
## # A tibble: 6 x 5
##   country country_code serie                                anio  value
##   <chr>    <chr>        <chr>                                <chr> <dbl>
## 1 Cuba    CUB          Mortality rate, infant (per 1,000 live birth~ 1960  47.1
## 2 Cuba    CUB          Mortality rate, infant (per 1,000 live birth~ 1961  45.3
## 3 Cuba    CUB          Mortality rate, infant (per 1,000 live birth~ 1962  43.6
## 4 Cuba    CUB          Mortality rate, infant (per 1,000 live birth~ 1963  41.9
## 5 Cuba    CUB          Mortality rate, infant (per 1,000 live birth~ 1964  40.3
## 6 Cuba    CUB          Mortality rate, infant (per 1,000 live birth~ 1965  38.7
```

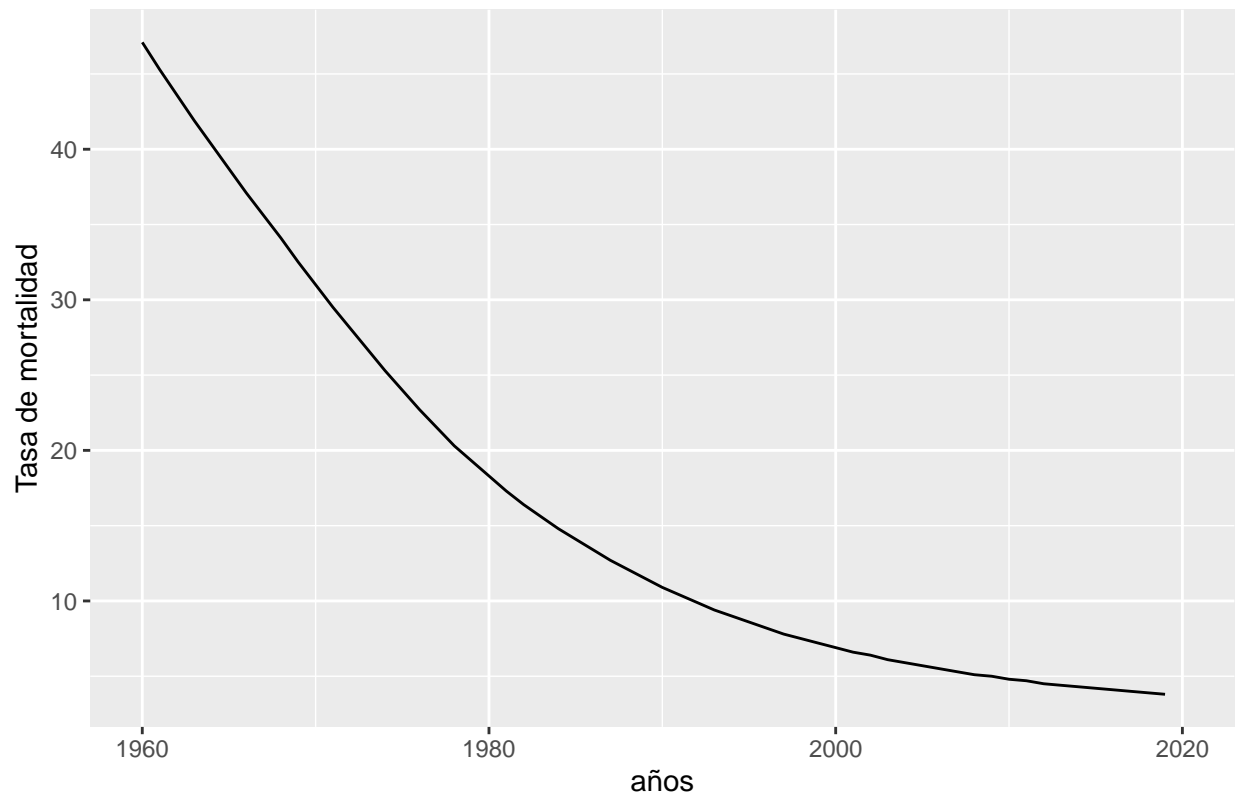
Grafica la evolución en el tiempo de la variable que seleccionaste.

```
library(ggplot2)

## Ver formato de variable.
cuba_mortalidad$anio<-as.numeric(cuba_mortalidad$anio)

cuba_mortalidad %>%
  ggplot(aes(x=anio,y=value)) +
  geom_line() +
  labs(x="años",y="Tasa de mortalidad",
       title = "Evolución de la tasa de mortalidad infantil en Cuba, desde 1960 a 2019")
```

Evolución de la tasa de mortalidad infantil en Cuba, desde 1960 a 2019



Cuarto ejercicio

Considerando la misma variable, u otra, compara la evolución de esta en el tiempo entre 3 o más países.

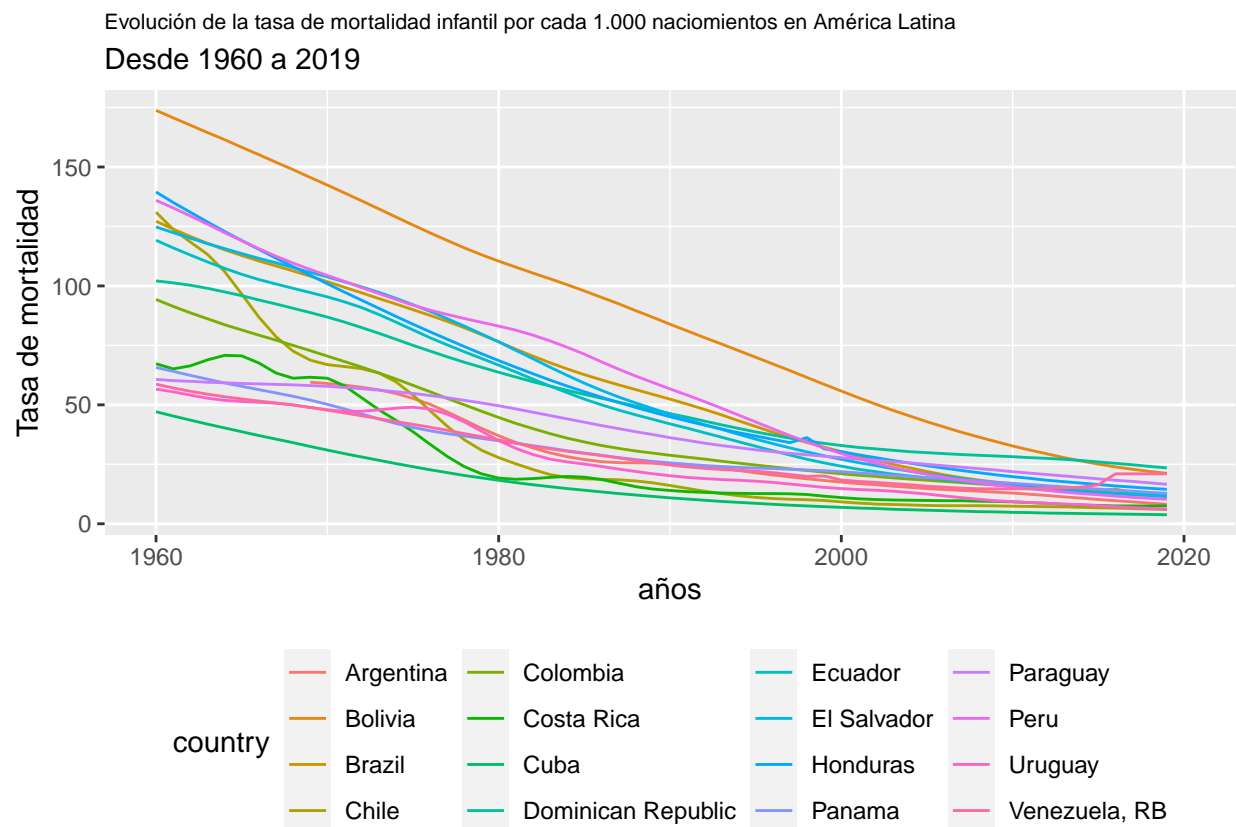
```
latam2_mortalidad<-latam2 %>% filter(serie=="Mortality rate, infant (per 1,000 live birt
table(latam2_mortalidad$country)
```

```
##
##      Argentina      Bolivia      Brazil      Chile
##      61            61            61            61
##      Colombia      Costa Rica      Cuba Dominican Republic
##      61            61            61            61
##      Ecuador      El Salvador      Honduras      Panama
##      61            61            61            61
##      Paraguay      Peru      Uruguay      Venezuela, RB
##      61            61            61            61
```

```
latam2_mortalidad$anio<-as.numeric(latam2_mortalidad$anio)

## Todos los países de AL
latam2_mortalidad %>%
  ggplot(aes(x=anio,y=value,color=country)) +
  geom_line() +
  labs(x="años",y="Tasa de mortalidad",
       title = "Evolución de la tasa de mortalidad infantil por cada 1.000 nacimientos e",
       subtitle = "Desde 1960 a 2019") +
  theme(plot.title = element_text(size=8),
        legend.position = "bottom")
```

```
## Warning: Removed 25 row(s) containing missing values (geom_path).
```



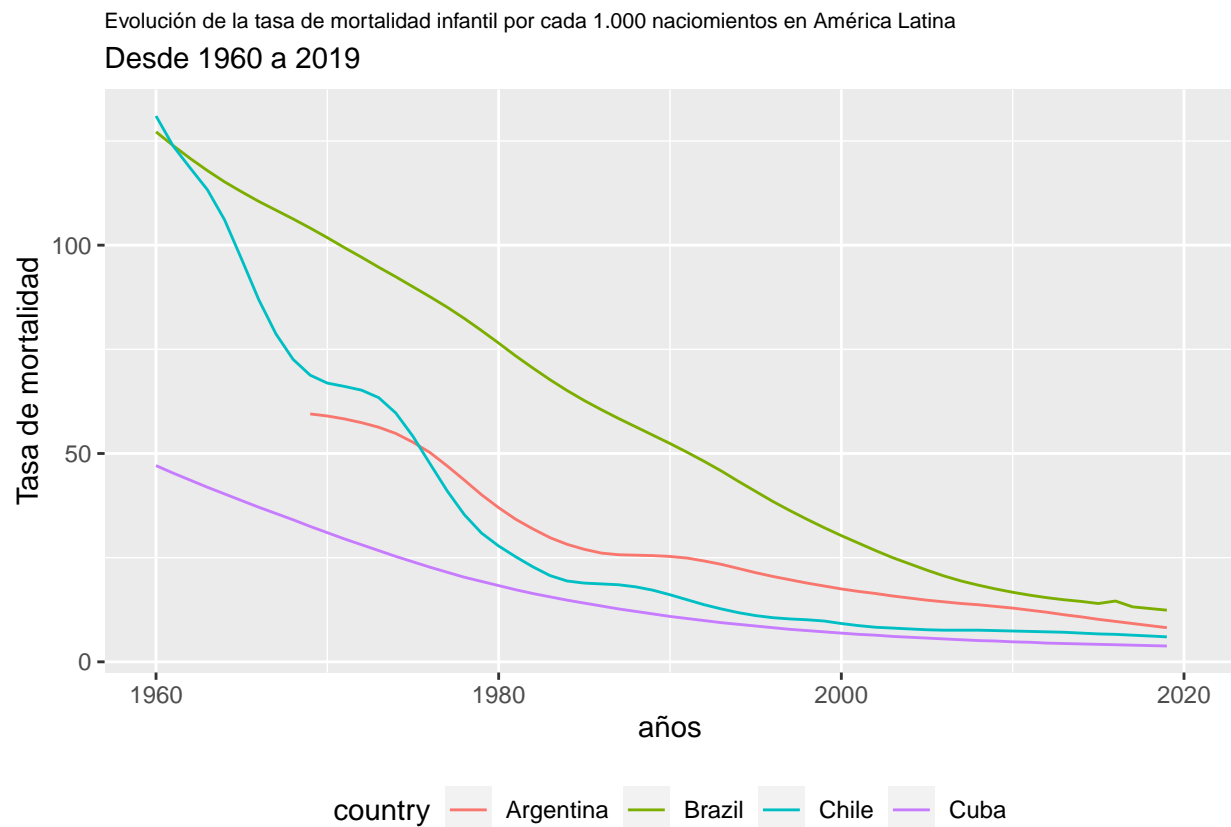
```
## Solo algunos
latam2_mortalidad %>%
  filter(country %in% c("Cuba","Chile","Brazil","Argentina")) %>%
  ggplot(aes(x=anio,y=value,color=country)) +
  geom_line() +
  labs(x="años",y="Tasa de mortalidad",
```

```

title = "Evolución de la tasa de mortalidad infantil por cada 1.000 nacimientos e
subtitle = "Desde 1960 a 2019" +
theme(plot.title = element_text(size=8),
      legend.position = "bottom")

```

```
## Warning: Removed 13 row(s) containing missing values (geom_path).
```



Quinto ejercicio

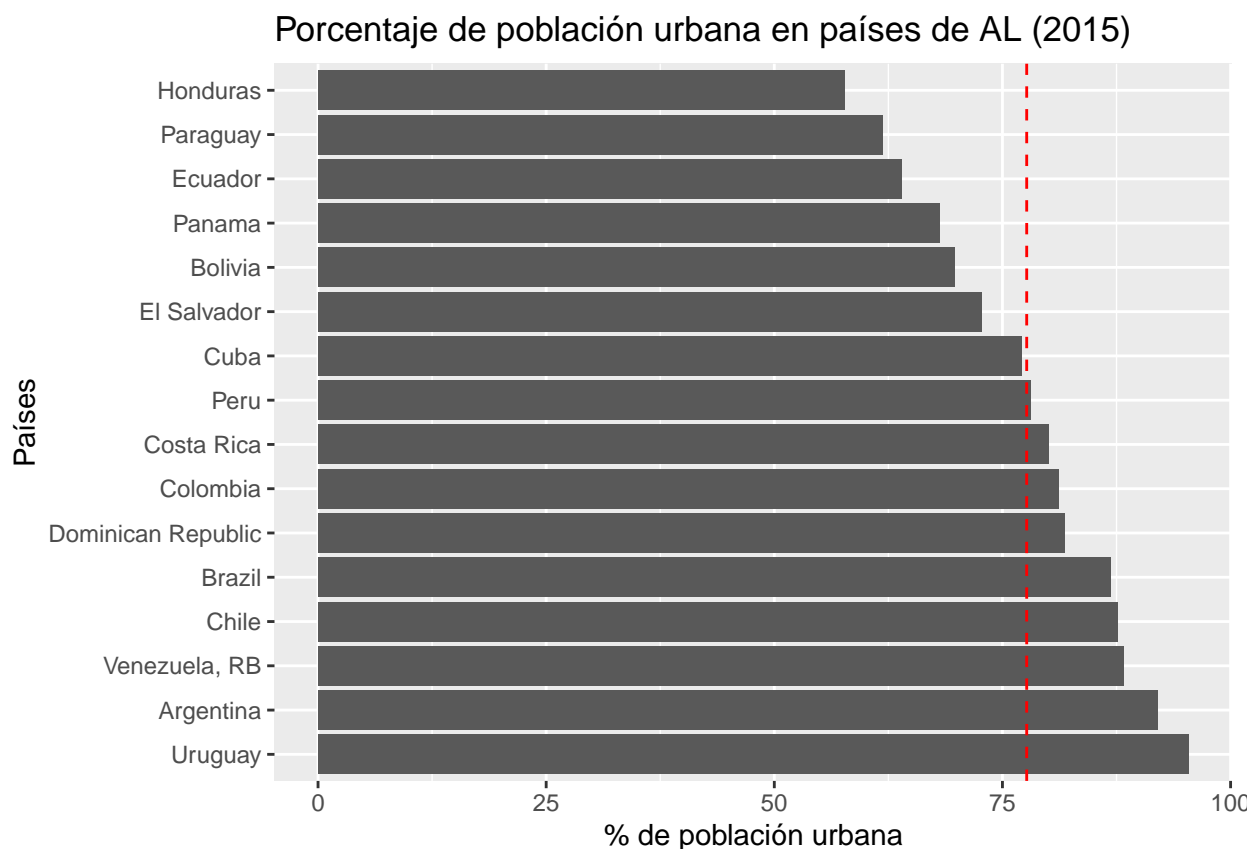
Filtre la base de datos de todos los países por otra variable y escoga un año, el que guste.

Haga un gráfico de barras, donde cada barra corresponda a un país. Si gusta, al final del código puede agregar el comando + `coord_flip()`, para invertir el gráfico y sea más clara su lectura.

```

latam2$anio<-as.numeric(latam2$anio)
latam2 %>% filter(anio==2019 & serie=="Urban population (% of total population)") %>%
  ggplot(aes(x = fct_reorder(country, desc(value)), y=value)) + geom_bar(stat = "identity") +
  geom_hline(aes(yintercept = mean(value)), linetype="dashed", color="red" ) +
  labs(x="Países", y="% de población urbana", title = "Porcentaje de población urbana en

```



Sexto ejercicio

Replique el ejercicio anterior, pero en vez de seleccionar un año, seleccione cuatro años (por ejemplo: 1960,1980,2000 y 2019). Haga un gráfico con 4 paneles, donde cada uno corresponda a un año.

Para mejorar la visualización puede seleccionar algunos países.

```
latam2 %>% filter(anio %in% c(1960,1980,2000,2019) &
  serie=="Urban population (% of total population)" &
  country %in% c("Cuba","Chile","Brazil","Argentina")) %>%
  ggplot(aes(x = fct_reorder(country, desc(value)),y=value)) + geom_bar(stat = "identity")
labs(x="Países",y="% de población urbana", title = "Porcentaje de población urbana en")
```


Porcentaje de población urbana en países de AL (2015)

