Tarea N°4

Nicolás Ratto

05-11-2021

```
library(dplyr)
library(ggplot2)
```

Instrucciones generales

La tarea debe ser escrita en Rmarkdown (`.rmd`) y debe ser entregada con su respectiva salida en `.html`.

Los documentos deben ser comprimidos en un archivo .zip y enviados a correo de profesor y ayudante. El nombre del archivo **debe** ser el primer apellido y nombre de el o la estudiante.

Entrega a más tardar el **martes 16 de noviembre a las 23:59 hrs.**

Toda pregunta es bien recibida vía correo electrónico. Mandar ejemplos reproducibles (a lo menos el código utilizado y una descripción/foto del error).

Para hacer todos los gráficos que se solicitan utilice el paquete *ggplot2*.

I. Visualización de datos

Escoga un microdato de su interés que trate de temas relacionados a las ciencias sociales. Idealmente el microdato debe ser provenir de una encuesta, pero también se aceptan microdatos producidos desde el análisis de registros administrativos, análisis de prensa, entre otros.

Acá dejamos un listado de posibles datos a utilizar:

- Latinobarómetro
- CASEN
- Encuesta Laboral ENCLA
- Encuesta Nacional de Empleo
- Encuesta Nacional de Discapacidad
- Estudio Longitudinal Social de Chile
- Datos del Observatorio de Huelgas Laborales
- Datos del Observatorio de Conflictos sociales
- Encuesta CEP
- Encuesta Mundial de Valores
- Institutional Characteristics of Trade Unions, Wage Setting, State Intervention and Social Pacts, 1960-2018 ICTWSS

Si tiene dudas sobre un microdato a utilizar, no dude en consultar a profesor y/o ayudante.

No utilizar los datos de gapminder o de paquetes similares. Tampoco los datos del Banco Mundial vistos en clases

Pregunta 1

Plantee formalmente un objetivo de investigación descriptivo al que pueda dar respuesta desde el conjunto de datos que seleccionó. Los restantes análisis que se le solicitan deben ir en la línea de responder a este objetivo.

Intente articular el objetivo en una introducción más amplia que presente el documento y el conjunto de datos que utilizará.

Pregunta 2 - 4

Haga los siguientes gráficos univariados: de barras, de cajas y un histograma (cada uno con variables diferentes). Describa la distribución de las variables a partir de los gráficos. Si gusta puede apoyarse de estadísticos como media, mediana, etc.

Pregunta 5

Haga un histograma o un gráfico de cajas bivariado. Interprete.

Pregunta 6

Haga un diagrama de dispersión con dos variables de interés. Describa. Si es pertinente, reporte e interprete brevemente el coeficiente de correlación.

Pregunta 7

Haga un gráfico de alta calidad y que llame la atención de una audiencia. La idea es que el gráfico resuma sus principales hallazgos descriptivos y que tenga varias capas, como etiquetas de variables, título, leyendas, etiquetas de datos, colores apropiados, paneles, entre otros. El gráfico debe ser interpretado. Para dar con el código adecuado a la hora de configurar el tema, puede apoyarse en el paquete `ggThemeAssist`.

II. Combinar data frames y luego visualizar (opcional)

El presente módulo es opcional, la idea es que le permita subir hasta 1,5 puntos en la nota de la prueba 2.

Descarga las bases de datos de países de América Latina y de Afganistán vistas en clases (CHL, ARG, BOL, MEX, HTI, AFG): https://www.dropbox.com/scl/fi/kzujrqzmn9qfoh5kbz9ed/paises_banco_mundial.xlsx?dl=0&rlkey=vvpcu0ic88p9n0w4xjb8iu18v.

Estos datos fueron descargados de la página de datos del Banco Mundial y ya fueron ordenados. **Tu tarea es combinarlos entre sí** y crear una data frame unificada que se llame `data`.

```
library(readxl)
arg <- read_excel("data/datos_bm/paises_banco_mundial.xlsx",sheet = 1)
bol <- read_excel("data/datos_bm/paises_banco_mundial.xlsx",sheet = 2)
chl <- read_excel("data/datos_bm/paises_banco_mundial.xlsx",sheet = 3)
hti <- read_excel("data/datos_bm/paises_banco_mundial.xlsx",sheet = 4)
mex <- read_excel("data/datos_bm/paises_banco_mundial.xlsx",sheet = 5)
afg <- read_excel("data/datos_bm/paises_banco_mundial.xlsx",sheet = 6)
data <- rbind(arg,chl,bol,hti,mex,afg)
dim(data)
```

```
## [1] 366   5
```

```
table(data$country)
```

```
##
## Afghanistan   Argentina     Bolivia       Chile       Haiti      Mexico
##          61          61          61          61          61          61
```

1. Pegar variable continente

Luego carga el paquete de gapminder para obtener los datos de todos los países. Agrega la variable `continent` de `gapminder` al conjunto de datos que combinaste. La idea es que la data siga teniendo solo información de los seis países originales.

```
library(gapminder)
continente <- gapminder %>%
  filter(year==1952) %>%
  select(country,continent)
table(continente$continent)
```

```
##
##   Africa Americas     Asia   Europe  Oceania
##       52       25       33       30        2
```

```
data <- merge(data,continente,by=c("country"),all.x = TRUE)
head(data)
```

```
##       country year     pop lifeExp gdpPercap continent
## 1 Afghanistan 1960 8996967  32.446  59.77323      Asia
## 2 Afghanistan 1961 9169406  32.962  59.86090      Asia
## 3 Afghanistan 1962 9351442  33.471  58.45801      Asia
## 4 Afghanistan 1963 9543200  33.971  78.70643      Asia
## 5 Afghanistan 1964 9744772  34.463  82.09531      Asia
## 6 Afghanistan 1965 9956318  34.948 101.10833      Asia
```

```
tail(data)
```

```
##     country year       pop lifeExp gdpPercap continent
```

```
## 361  Mexico 2015 121858251  74.904  9616.646  Americas
## 362  Mexico 2016 123333379  74.917  8744.516  Americas
## 363  Mexico 2017 124777326  74.947  9287.850  Americas
## 364  Mexico 2018 126190782  74.992  9686.514  Americas
## 365  Mexico 2019 127575529  75.054  9946.034  Americas
## 366  Mexico 2020 128932753      NA  8346.702  Americas
```

2. Crea un gráfico por países

Con el conjunto de seis países de `data`, que ahora tiene la variable continente, crea un gráfico de líneas que presente la evolución en el tiempo de la variable esperanza de vida al nacer (*lifeExp*), desde 1960 a 2020. Las líneas del país deben estar en un panel distinto al de las líneas de los países de América Latina (*facet_wrap*).

Antes de intentarlo transforma a numérica la variable año, de esta forma:

```
data$year <- as.numeric(data$year)
```

```
data %>%
  ggplot(aes(x=year,y=lifeExp,color=country))+
  geom_line() +
  facet_wrap(~continent)
```