

Revisión 2023

Explicativo sobre la prueba:

Esta prueba es de carácter *INDIVIDUAL*, cada estudiante debe completarla y no interactuar con el resto de los estudiantes, si tienen dudas nos pueden preguntar a Mauro o a mi durante la prueba. Pueden consultar durante la prueba el libro del curso (<https://r4ds.had.co.nz/>), las slides, actividades y deberes del curso pero ningún otro material adicional. Sin embargo *NO ESTA PERMITIDO* utilizar otro material que el sugerido ni las respuestas, comentarios o sugerencias de otras personas que no sean los docentes del curso, cualquier apartamiento de esto invalidará la prueba.

Por favor completá tu nombre y CI en el YAML del archivo donde dice author: "NOMBRE Y CI:". Los archivos y la información necesaria para desarrollar la prueba se encuentran en Eva en la pestaña Revisión_2023. La revisión debe quedar en tu repositorio del curso GitHub en una carpeta que se llame Revisión_2023 con el resto de las actividades y tareas del curso. Parte de los puntos de la prueba consisten en que tu revisión sea reproducible y tu repositorio de GitHub esté bien organizado como se indica en el Ejercicio 1.

La prueba dura 2 horas con posibilidad de extensión de media hora adicional.

Hay puntos parciales por lo que si algunas de las respuestas de código te queda incompleta agregá en el entorno de código `eval=FALSE` para que no se evalúe el mismo y evitar posibles errores.

EJERCICIO 1 (GitHub y entrega de la Revisión)

(10 puntos)

Esta pregunta es sobre el uso de GitHub y es la forma que van a entregar la prueba. Recordar que para que tengas la última versión de tu repositorio debes hacer `pull` a tu repositorio para no generar inconsistencias y antes de terminar subir tus cambios con `commit` y `push`.

1.a) (1 Punto) En tu repositorio del curso creá una carpeta que se llame `Revision_2023` (sin tilde).

1.b) (1 Punto) Asegurate que soy colaboradora del mismo y si no lo soy mandame la invitación, mi usuario es natydasilva.

1.c) (1 Punto) Subí el archivo `.Rmd` de esta revisión y los necesarios para reproducir el documento. Actualizá el repositorio regularmente durante prueba para que todo ande bien, asegurate al final de la prueba que el archivo `Rmd` compila.

1.d) (3 Puntos) Asegurate que tu prueba sea reproducible cuando clone tu repositorio. Para ello deberás subir los archivos necesarios para la reproducibilidad (los datos van a ser necesarios). Poné los datos en la carpeta Datos de tu repositorio.

1.e) (2 Puntos) A parte de subir el archivo al repositorio debés mandarme por correo (natalia.dasilva@fcea.edu.uy) el `.Rmd` y el `.pdf` de tu versión final. Escribime en el asunto del correo `Rev23_STAT_NT` y en el cuerpo tu usuario de GitHub para que sea más sencillo encontrar tu repositorio con la prueba.

1.f) (2 Puntos) Presentá tu código de forma ordenada así como las respuestas a cada pregunta.

Ejercicio 2 (R Base)

(35 Puntos)

2.a) (8 Puntos) En cada uno de los siguientes casos tenés que arreglar algún problema común cuando subdividimos `data.frame`

```
mtcars[mtcars$cyl == 8, ]  
mtcars[-1:5, ] # ayuda el - tiene que aparecer en tu propuesta de arreglo  
mtcars[mtcars$cyl <= 6]  
mtcars[mtcars$cyl == 4 | 2, ] # el 4 y el 2 tienen que aparecer en tu propuesta de arreglo
```

2.b) (5 Puntos) ¿Porqué `mtcars[1:30]` da un error? y ¿Cómo esto es distinto de `mtcars[1:30,]`?

2.c) (10 Puntos) Usando un for loop recodificá los valores del siguiente vector `c(2, 8, 5, 7, 2, 3, 1, NA)` al que deberás llamar `v_orig` como 1 es 'Azul', 2 es 'Verde', 3 es 'Amarillo', 4 es 'Rojo', 5 es 'Violeta', NA es NA y el resto como 'Otro'. El vector resultante deberá ser `c("Verde", "Otro", "Violeta", "Otro", "Verde", "Amarillo", "Azul", NA)` al que debes guardar en el objeto `v_recod`

Ahora incorporá lo anterior en un loop que sirva para cualquier vector numérico y si el vector no es numérico que frene e indique un error con sentido.

2.d) (7 Puntos) Implementá una función que extraiga la diagonal principal de una matriz. Debes comprobar que se comporta como la función `diag(x)`, donde `x` es una matriz.

2.e) (5 Puntos) ¿Qué propiedad importante de `&&` hace funcionar a `f1`?

```
f1 <- function(x) {  
  !is.null(x) && length(x) == 1 && x > 0  
}
```

¿Qué es diferente con este código? Porqué este comportamiento no es deseable aquí?

```
f1 <- function(x) {  
  !is.null(x) & length(x) == 1 & x > 0  
}
```

Ejercicio 3 (tidyverse)

(35 Puntos)

En este ejercicio usaremos los datos del **Estudio Longitudinal de Bienestar en el Uruguay** (IECON) utilizados en la Actividad 12 de clase. Estos datos consisten en un relevamiento longitudinal representativo de los niños que concurren al sistema de educación primaria pública.

Se recoge información referente a múltiples dimensiones del bienestar de los niños en la muestra así como también del resto de los integrantes de sus hogares, entre los que se destacan:

- situación nutricional.
- habilidades cognitivas y no cognitivas.
- imaginación; razonamiento y sentimientos.
- actividades de ocio e interacción social.
- logros educativos.
- ingresos.
- calidad de la vivienda y bienestar subjetivo, entre otros.

La información es relevada en Olas, la primera fue en 2004 a 3000 niños del primer grado de escuela. En las distintas olas se los sigue en el tiempo a los mismos niños (por eso es longitudinal).

- De **personas**, con información referente al niño y personas del hogar donde reside.

- De **hogares**, con información más enfocada al niño, sus capacidades de relacionamiento, emociones, amigos, etc.

Vamos a usar datos de la tercer ola (2012). La base de personas contiene información sobre las personas del hogar del niño que salió en la muestra. Contiene la variable `nform` (número de formulario) que permite identificar al **hogar** (es una *key*). Para cada hogar existe la variable `nper` (número de persona) que le asigna un número a cada **persona del hogar**.

La base de hogares contiene información de los hogares.

Tiene la variable `nform` (número de formulario) que permite identificar al hogar y la variable `nper` (número de persona encuestada) que permite identificar a la persona del hogar que fue encuestada.

Todas las respuestas a este problema deberán ser respondidas con funciones de algún paquete de tidyverse

Todas las transformaciones y recodificaciones de variables que se hicieron en la Actividad 12 están en el conjunto de datos que vamos a trabajar para personas.

Trabajaremos solamente con las siguientes variables y sus nombres:

- Número de formulario: `nform`
- Número de persona: `nper`
- Sexo: `sexo`
- Situación conyugal: `sit.conyugal`
- Nivel educativo que cursa o cursó: `nivel.educ`
- Relación de parentesco con el jefe/a: `parent.jefe`
- Edad: `edad`
- Sexo: `sexo`
- Ocupación principal: `dedicacion.p`
- Tipo de ocupación principal: `tipo.p`
- Tareas de su ocupación principal: `tareas.p`
- Jubilaciones pensiones: `jub.pen`
- Transferencia de otros hogares: `transf.pais`
- Transferencias externas: `transf.ext`
- Monto asignación familiar: `afam`
- Beneficio en especie: `benef.esp`

3.a) (1 Punto) Usando la función `read_csv` del paquete `readr` cargá la base de datos `persona_recode_ed.csv` que se encuentra disponible en el EVA y a estos datos nombrálos `personas`. Usá la función `read_csv()` y `here()` para asegurar la reproducibilidad.

3.b) (5 Puntos) La variable `sexo` tiene tres valores, recodificala usando la función `case_when` para que el 1 sea Hombre el 2 sea Mujer y 9 sea NS/NC (no sabe). Guardá los nuevos datos en `personas_reco`. La nueva variable nombrala `sexo_reco`

3.c) (7 Puntos) Usando funciones de `dplyr` respondé ¿Cuál es la proporción de personas según `sexo_reco`? Reportá una tabla (con `xtable`), La tabla debe contener tres columnas (Sexo, Conteo y Proporción). Guardá el objeto generado con nombre `tabla`. Recordá poner en el chunk de código `results='asis'` y adentro del entorno `options(xtable.comment = FALSE)` para que salga la tabla en el pdf.

3.d) (5 Puntos) Nos gustaría tener la variable **departamento** que indica el departamento al cuál pertenece el hogar de la muestra así como la variable `a4` que indica propiedad de la vivienda y `a4a` que indica el monto del alquiler.

Tenemos esta información en la **base de Hogares** y una variable de macheo que es `nform`, la cuál indica el **número de formulario** y vale lo mismo para todos los miembros del hogar.

Nos interesa mantener todas las observaciones de la base de personas, y “pegarle” el departamento, `a4` y `a4a`.

La base de hogares es `hogares_reco.csv` con alguna función de `dplyr` relacioná estas dos `data.frames` y al nuevo objeto nombralo `presonas_hogar`

3.e) (7 Puntos) Recodificá la variable `a4` (propiedad de la vivienda) de `personas_hogar` usando `case_match` guardá los datos en el objeto `personas_recode` como sigue

- 1 `prop_vivter_pagando`
- 2 `prop_vivter`
- 3 `prop_viv_pagando`
- 4 `prop_viv`
- 5 `alquila`
- 6 `ocupa_dep`
- 7 `ocupa_gr`
- 8 `ocupa_noper`
- 9 `pension`
- 10 `otro`
- `ns/nc` “NA”

Usando `ggplot2` hacé un gráfico de barras con la distribución de dicha variable indicando la más frecuente y en caso que las etiquetas no se vean bien proponé alguna forma de mejorarlo visualmente.

3.f) (10 Puntos) Mediante un gráfico describí como se distribuye la variable `a4a` por departamento en distintos paneles (4 columnas y dos filas), notá que deberá convertirla a numérica y no considerar los NA.

1. Cambiá el nombre de los ejes apropiadamente.
2. Cambiar la paleta de colores a una paleta apropiada según el tipo de variable, usando el paquete `paletteer` y la función `scale_color_paletteer_d`
3. Cambiá el nombre de la leyenda de color a “Depto” usando la función de tipo `scale_aes_type` apropiada.
4. Cambiá la posición de la leyenda en la parte inferior del gráfico usando `theme`, el título del eje x e y de color `darkgray` y tamaño 12, el título y texto de la leyenda también de tamaño 12 y con color `darkgray`.
5. Incluí un caption a tu figura y una descripción de lo que ves en ella, algo interesante que surja del gráfico.

Ejercicio 4 (Popurrí)

(20 Puntos)

4.a) (5 Puntos) Indique cuál de las siguientes opciones NO corresponde a partes de un gráfico estadístico.

Seleccione una:

- a) Datos, mapeo estético y sistema de coordenadas
- b) Datos, mapeo estético y escalas
- c) Datos, mapeo estético y modelo estadístico
- d) Datos, mapeo estético y temas

4.b) (5 Puntos) Indica cuál de las opciones describe mejor lo que hace la función

```
func1 <-function(a){  
  for(i in 1:a){  
    b <- i^2  
  }  
  print(b)}  
}
```

- a. Calcula un número al cuadrado
- b. Calcula una secuencia de números al cuadrado

- c. Calcula el producto de a por 2
- d. Calcula una secuencia de números al cuadrado para los valores positivos hasta a

4.c) (10 Puntos) Hacé una función llamada `simulaplot` que tenga como argumentos `n` tamaño de muestra, `sigma` el desvío y `forma`. Esta función simula datos para distintos tamaños de muestra con distinta media y desvío de una distribución normal usando `rnorm` y simula la misma cantidad de observaciones n de una distribución con `rgamma` cambiando el parámetro de forma. Como resultado de esta función se hace un diagrama de dispersión (usando `ggplot2`) con x los datos simulados normales y y los datos gamma. A su vez si el argumento es `sigma` es negativo la función no debe ser evaluada y debe imprimir el mensaje “sigma debe ser positivo”.