Revisión 2021

NOMBRE Y CI: Martin Olivera 44845488-3

4/6/2021

Explicativo sobre la prueba

Por favor completá tu nombre y CI en el YAML del archivo donde dice `author: "NOMBRE Y CI: "`. El examen es individual y cualquier apartamiento de esto invalidará la prueba. Puede consultar el libro del curso durante la revisión http://r4ds.had.co.nz así como el libro de ggplot2 pero no consultar otras fuentes de información.

Los archivos y la información necesaria para desarrollar la prueba se encuentran en Eva en la pestaña Prueba.

La revisión debe quedar en tu repositorio PRIVADO de GitHub en una carpeta que se llame Prueba con el resto de las actividades y tareas del curso. Parte de los puntos de la prueba consisten en que la misma sea reproducible y tu repositorio de GitHub esté bien organizado.
Además una vez finalizada la prueba debes mandarme el archivo pdf y Rmd a natalia@iesta.edu.uy y por favor recordame tu usuario de GitHub para que sea más sencillo encontrar tu repositorio, asegurate que haya aceptado la invitación a tu repositorio y de no ser así enviame nuevamente la invitación a natydasilva.

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`.

La Revisión vale 130 puntos donde 15 de los puntos son de reproducibilidad de la misma, organización del repositorio en GitHub, órden y organización en el código y respuestas.

Ejercicio 1 (90 puntos)

Explicativo sobre los datos

Los datos que vamos a utilizar en este ejercicio son una muestra de datos a nivel nacional sobre abandono escolar en los años 2016 que ya utilizamos en la Tarea 2.

En el Cuadro 1 se presentan las variables en el conjunto de datos **muestra.csv**.

Table 1: Variables en **muestra.csv**

Variable	Descripción
documento	Cédula de Identidad del alumno
nro_doc_centro_educ	Liceo que concurre el alumno en 2016
nombre_departamento	Nombre del Departamento del centro educativo
grupo_desc	Grupo del alumno en 2016
coberturaT	Cobertura en el primer semestre de 2016
Centro_Grupo	Liceo y grupo del alumno en 2016
cl	Cluster - contexto sociocultural del liceo en 1016
Grado_2016_UE	Grado del alumno en el 2016 según UE
Grado2013	Grado del alumno en 2013 según CRM
Grado2014	Grado del alumno en 2014 según CRM
Grado2015	Grado del alumno en 2015 según CRM
Grado 2016	Grado del alumno en 2016 según CRM
Sexo	Sexo del alumno
Fecha.nacimiento	Fecha de nacimiento del alumno
Grupo_UE_2017	Grupo del alumno en 2017
inasistencias	cantidad de inasistencias en el primer semestre de 2016
asistencias	cantidad de asistencias en el primer semestre de 2016

1. Dentro de tu proyecto de RStudio creá un subdirectorio llamado Datos y copiá el archivo muestra.csv. Lee los datos usando alguna función de la librería `readr` y `here`. **(5 puntos)**

```
library(readr)
library(here)
datos <- read_csv(here("Prueba/Datos/muestra.csv"))
```

El archivo no compila, porque no pusiste la ruta donde están los datos, probablemente en tu computadora funcionaba porque tenías los datos en otro lado. Se pedía usar alguna función de readr y here que no lo hiciste (1 Puntos)

2. Utilizando funciones de `dplyr` transformá la variable Abandono para que sea un factor con dos niveles donde el 0 se recodifique a No y el 1 a Si. Mostrame el resultado resumido en una tabla con la cantidad de observaciones para cada categoría usando `xtable`, recordá incluir en el chunk `results='asis'`. **(10 puntos)**

```
library(dplyr)
datos_recod <- datos %>%
mutate(Abandono = recode(Abandono, `0` = "No",
`1` = "Si"))
datos$Abandono <- as.factor(datos$Abandono)

library(xtable)
# options(xtable.comment = FALSE)
# tabla <- datos_recod %>%
# group_by(Abandono) %>%
# mutate(prop = n/sum(n, na.rm = TRUE))

# tabla %>% xtable() %>% print(include.rownames = FALSE)
```

Luego de agrupar deberias haber usado un summarize(Cantidad = n()) para tener el conteo total de observaciones. Ver sol. Comenté parte del código y lo evalué porque sino no funciona el resto (5 Puntos)}

3. Usando funciones de `dplyr` respondé ¿Cuál es el porcentaje de abandono en Montevideo? **(10 puntos)**

```
datos_recod %>%
group_by(Abandono) %>%
filter(nombre_departamento=="Montevideo") %>%
summarise(n = n()) %>%
mutate(prop = n/sum(n, na.rm = TRUE))
```

```
## # A tibble: 2 x 3
##   Abandono     n   prop
##   <chr>    <int>  <dbl>
## 1 No        1181 0.946
## 2 Si          67 0.0537
```

(10 Puntos)

Para el departamento, se observa un porcentaje de abandono de estudios del 5.37 %

4. Reproducí el siguiente gráfico y en vez de "Gráfico a replicar" (`caption`) debes agregar un título que describa la figura y algún comentario interesante de lo que observás en la misma.**(10 puntos)**

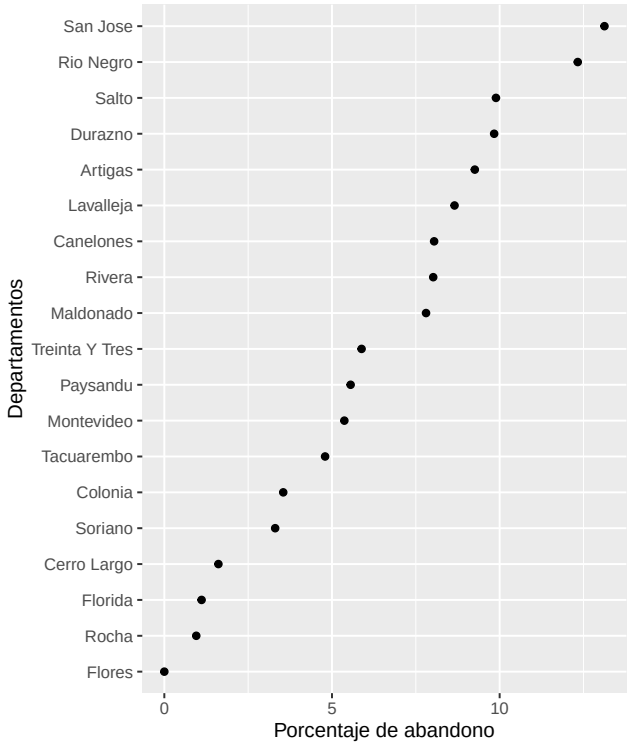

Figure 1: Gráfico a replicar

```
library(ggplot2)
datos %>% filter(Abandono==1) %>% group_by(nombre_departamento) %>% summarise(n = n()) %>% mutate(prop =
  labs(x = "Porcentaje de Abandonos", y = "Departamentos")
```

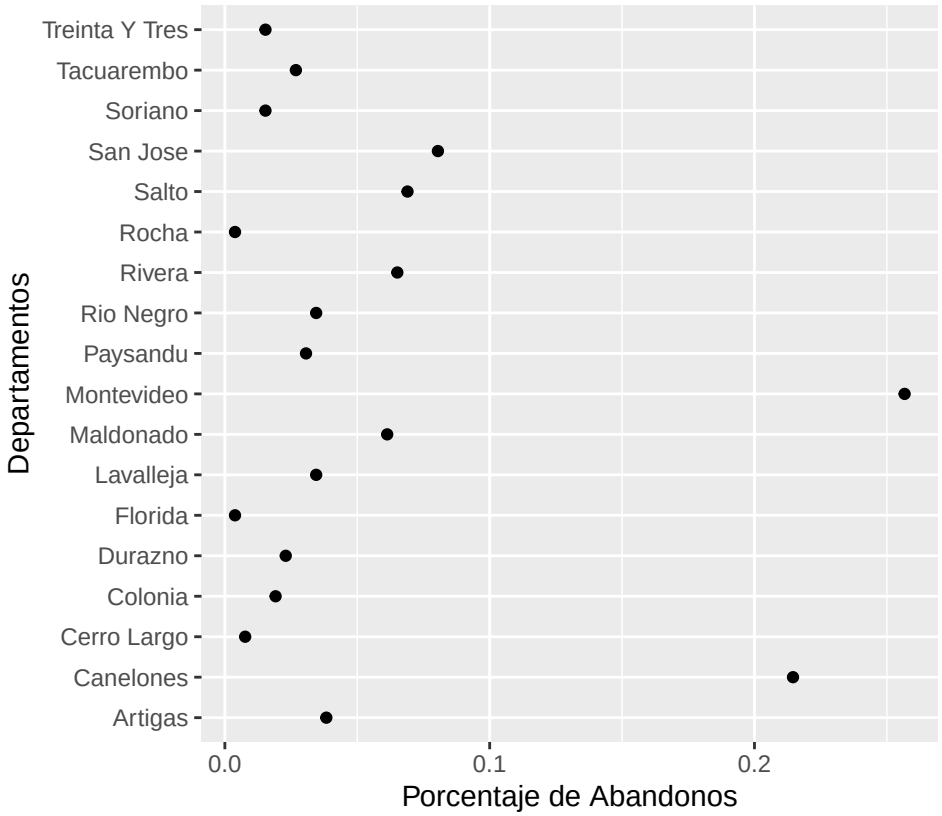

Figure 2: Figura 1: Porcentaje de abandono de estudiantes por departamento.

Se observa que Montevideo y Artigas son los departamentos con mayor tasa de abandono. Luego, el resto de los departamentos, son similares en cuanto a dicha tasa con niveles que en ningún caso superan el 10%. Rocha, Florida y Cerro Largo, son los que presentan niveles de abandono inferiores.

No calculaste el porcentaje de abandono en cada departamento, te faltó ordenar y en el eje x no es un porcentaje en tu caso (6 Puntos)

5. Reproducí el siguiente gráfico realizado solo con los estudiantes que abandonaron y en vez de "Gráfico a replicar" (`caption`) debes agregar un título que describa la figura y algún comentario interesante de lo que observás en la misma. La paleta usada es Dark2. **(10 puntos)**

```
library(ggplot2)
datos %>% filter(Abandono==1) %>% group_by(nombre_departamento, Sexo) %>% summarise(n = n()) %>% mutate
  labs(x = "Porcentaje de Abandonos", y = "Departamentos")
```

\textbf{\textcolor{violet}{Nuevamente el problema son los datos que usá para graficar, no necesitá resumir los datos aquí con los datos filtrando por abandono funcionaba. Faltaba ordenar, te quedaron los nombres de los ejes cambiados, si usás fig.cap no necesitas numerar las figuras, lo hace directamente. Te falta poner el nombre del gráfico, es de barras apiladas al 100%... Ver sol. (4 Puntos)}}

6. Reproducí el siguiente gráfico y en vez de "Gráfico a replicar" (`caption`) debes agregar un título que describa la figura y algún comentario interesante de lo que observás en la misma. La paleta usada es Dark2.**(15 puntos)**

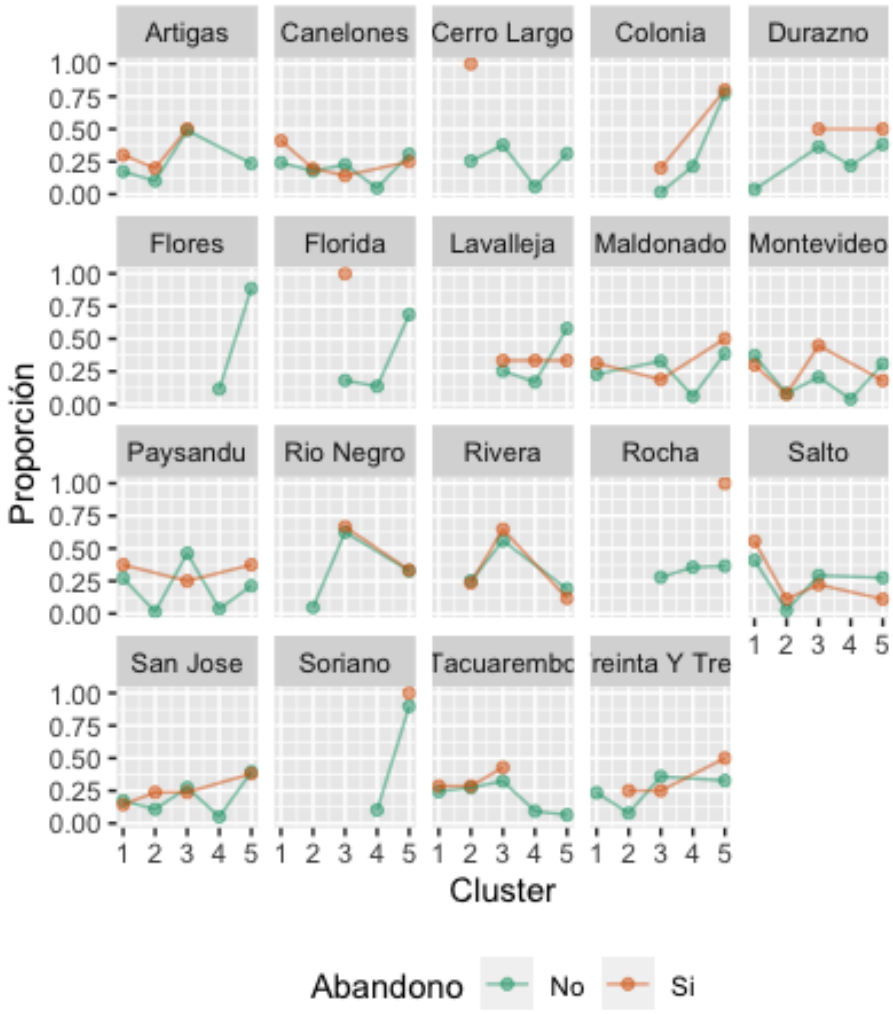

Figure 3: Gráfico a replicar

Ver sol

7. Recodificá la variable `grupo_desc` que tiene 17 niveles para que de 1ro.G.1 a 1ro.G5 sea A de 1ro.G.6 a 1ro G.11 sea B y los restantes C. Mostrá el resultado seleccionando la variable recodificada y las primeras 6 filas. **(5 puntos)**

```
aux <- c( "1ro.G.1" = "A"="1ro.G.1","1ro.G.2" = "A","1ro.G.3" = "A","1ro.G.4" = "A","1ro.G.5" = "A","1r
)

recode(datos$grupo_desc, !!!aux)
```

La recodificacion que armaste no funciona, ver sol. (2 Puntos)

8. Separá la variable Fecha.nacimiento en tres nuevas variables año, mes y dia, para ello usá la función `separate` de forma que sean numéricas. Mostrá el resultado seleccionando las variables documento, año, dia y mes con alguna función de `dplyr` y las primeras 6 filas.**(5 puntos)**

Ver sol.

9. Convertí la variable Fecha.nacimiento como objeto de tipo Date usando `as.Date` de R base y comprobá que la nueva variable Fecha.nacimiento es del tipo correcto. **(5 puntos)**

```
as.Date(datos$Fecha.nacimiento)
typeof(datos$Fecha.nacimiento)
```

Te quedó el nombre de la variable sin el punto por eso no corre (5 Puntos)

10. Usando la variable Fecha.nacimiento transformada, se considera que el alumno tiene extra-edad leve cuando nace antes del 30 de abril de 2003. Es decir, tiene un a~no m'as de la edad normativa para dicha generaci'on. En base a esta definición creá una nueva variable (nombrala extra) que valga 1 si el alumno tiene extra edad leve y 0 si no la tiene. Muestra solo el resultado de las primeras 6 filas. Pista para que la condición tome en cuenta el formato fecha podrías usar `as.Date('2003-04-30')`. **(10 puntos)**

Ver sol.

11. Trabajá con un subconjunto de datos que tenga documento, Grado2013, Grado2014,Grado2015, Grado2016 y llamale reducida. Con los datos reducidos restructuralos para que queden de la siguiente forma usando alguna de las funciones del paquete `tidyr` que vimos en la última clase.**(5 puntos)**

```
A tibble: 16,092 x 3
   documento  Grado   Nivel
       <int>  <chr>   <chr>
 1  52401872 Grado2013 4º
 2  52401872 Grado2014 5º
 3  52401872 Grado2015 6º
 4  52401872 Grado2016 1
 5  54975382 Grado2013 5º
 6  54975382 Grado2014 6º
 7  54975382 Grado2015 1u
 8  54975382 Grado2016 1
 9  54944549 Grado2013 4º
10  54944549 Grado2014 5º
```

Ver sol.

Ejercicio 2 (25 puntos)

1. En clase vimos distintas visualizaciones para variables categóricas y mencionamos como posibles el gráfico de barras y el gráficos de torta.

¿Cuál es el argumento teórico para decir que es siempre preferible un gráfico de barras a uno de tortas para ver la distribución de una variable categórica? **(5 puntos)**

Utilizar un gráfico de barras (en cualquiera de sus formas) permite observar más claramente la distribución de las categorías de cierta variable, pudiendo fácilmente ordenarlas y así, interpretarlas y que al fín, sea útil la visualización. El gráfico de torta, tiene un gran problema de visualización, cuando se trata con gran cantidad de categorías, donde no se logra dicernir las proporciones de cada una claramente y por lo tanto es menos intuitivo el órden de las categorías en términos de sus proporciones. Adicionalmente, surgen problemas de visualización con los gráficos de torta a la hora de contrastarlos con otros, por ejemplo, si quiero observar distribución por departamentos de cierta variable. Comparativamente con el gráfico de barras, el gráfico de torta sufre de gran ilegibilidad.

Ver sol. (3 Puntos)

2. ¿Porqué es necesario utilizar `aspect.ratio = 1` en un diagrama de dispersión? **(5 puntos)**

Utilizar la opción aspect.ratio=1 es útil en los diagramas de dispersión para evitar potenciales problemas con las escalas de los ejes y que el resultado del gráfico sea interpretable en términos de las dos variables incluidas. Problemas con las escalas en los ejes, no corregidas podría llevar a que la visualización arroge posibles conclusiones inadecuadas.

(5 Puntos)

3. Generá una función `compra` que tenga como argumentos un vector numérico `cprod` cantidad de productos a comprar de cada tipo y un vector numérico `cdisp` con la cantidad disponible de dichos productos (ambos vectores del mismo largo) que devuelva 1 si se pude hacer la compra y 0 en caso contrario. La compra se puede realizar siempre que haya stock suficiente para cada producto, es decir que la cantidad disponible sea igual o mayor a la cantidad comprada. A su vez si alguno de los argumentos no es un vector numérico la función no debe ser evaluada y debe imprimir el mensaje "Argumento no numérico". **(15 puntos)**

Comprobá que el resultado de la función sea

`compra(c(1,4,2), 1:3) = 0`

`compra(c("A","B"), 1:3)= Argumento no numérico`

```
compra=1
 i=1
 compra <- function(u,v) {

     if(typeof(u) == "character" | typeof(v) == "character") { compra = "Argumento no numérico"} else {
         while (compra==1 & i>=1) {
         if (u[i] <= v[i]) { compra=1} else
         {compra=0}
         i=i+1
         }
     }
     print(compra)
 }
compra(c(1,4,2), 1:3)
compra(c("A","B"), 1:3)
```

No está funcionando bien la parte del while, fijate en la sol que con ifelse sale más sencillo (10 Puntos)

A tu repositorio le falta un poco de orden, poner las actividades en carpetas así como las tareas. El documento no compilaba por varios motivos entre ellos la lectura de los datos y algún chunk que no evaluaste pero quedaban colgados objetos importantes para otras partes. Sea bueno que estructures un poco el código para que sea más entendible y sencilla la lectura. (10 Puntos). TOTAL DE PUNTOS 61/130