Revisión 2021

Lorena Pérez 4926489-9

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("Datos/Prueba", "muestra.csv"))
```

No es reproducible, estás leyendo los datos con el directorio que funciona solo en tu compu. No estás usando correctamente la función here

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)
library(tidyverse)
library(xtable)
datos<-datos %>% mutate(Abandono=recode(Abandono,`0`="No",`1`="Sí"))
prop<-datos %>%group_by(Abandono)%>%summarise(n=n())%>%xtable()
prop
```

% latex table generated in R 4.0.2 by xtable 1.8-4 package % Fri Jun 11 01:15:05 2021

	Abandono	n
1	No	3762
2	Sí	261

Saqué el eval=FALSE para que salga la tabla (10 Puntos)

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

```
prop2<-datos %>%filter(nombre_departamento=="Montevideo")%>%group_by(Abandono)%>%summarise(proporcion=(
print(paste0("El porcentaje de abandono en Montevideo es ",round(prop2[2,2],2),"%"))
```

[1] "El porcentaje de abandono en Montevideo es 1.67%"

El denominador en el cálculo del porcentaje es incorrecto, pusiste el total y no las observaciones de Montevideo. (5 Puntos)

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)**

```
prop_depto<-datos %>%
  group_by(nombre_departamento,Abandono)%>%summarise(Conteo=n())%>%mutate(Proporción=round(Conteo/sum(C
prop_depto%>%ggplot(aes(x = Proporción,y=reorder(nombre_departamento, Proporción)))+geom_point()+theme(
```

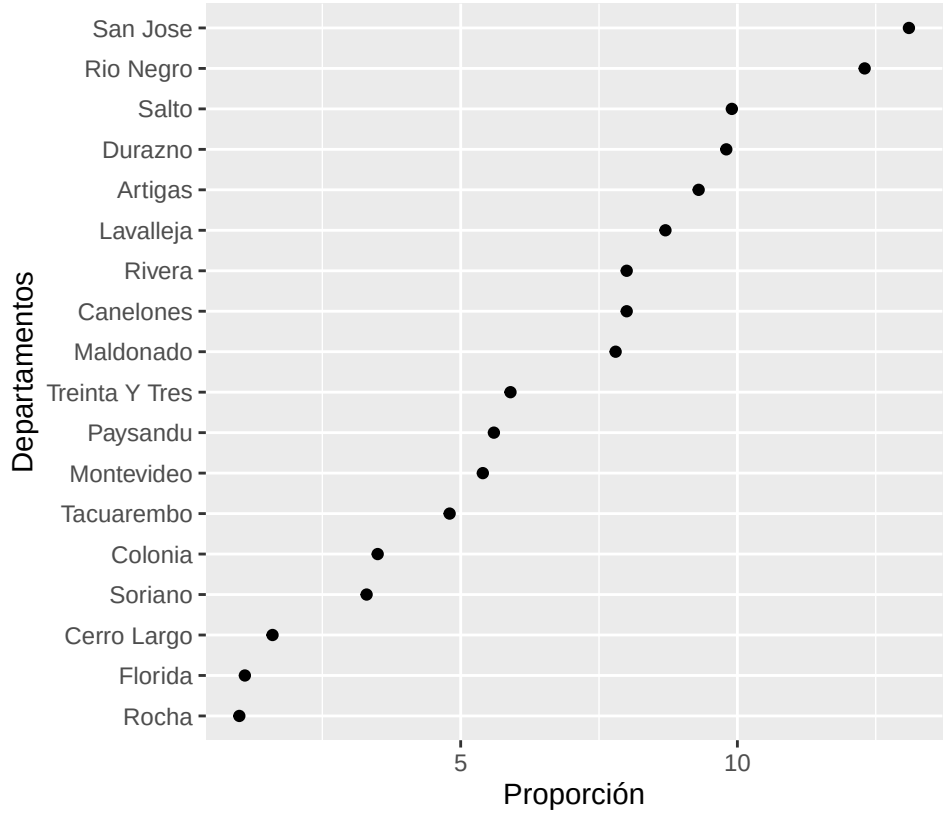

Figure 1: Porcentaje de abandono por departamento

Se observa que el porcentaje de abandono en Flores es nulo, mientras que Río Negro y San José son los únicos dos departamentos con un porcentaje de abandono mayor a 10.

El eje x es Porcentaje de abandono, no proporción (9 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)**

```
datos<-rename(datos,"Género"="Sexo")
prop_depto_sex<-datos %>%filter(Abandono=="Sí") %>%
  group_by(nombre_departamento,Género)%>%summarise(Conteo=n())%>%mutate(Proporción=round(Conteo/sum(Con

prop_depto_sex%>%ggplot(aes(x = Proporción,fill = Género,y=reorder(nombre_departamento, -Proporción)))+
```

A excepción de Canelones, San José, Soriano y Rocha, más del 50% de abandono se corresponde al género

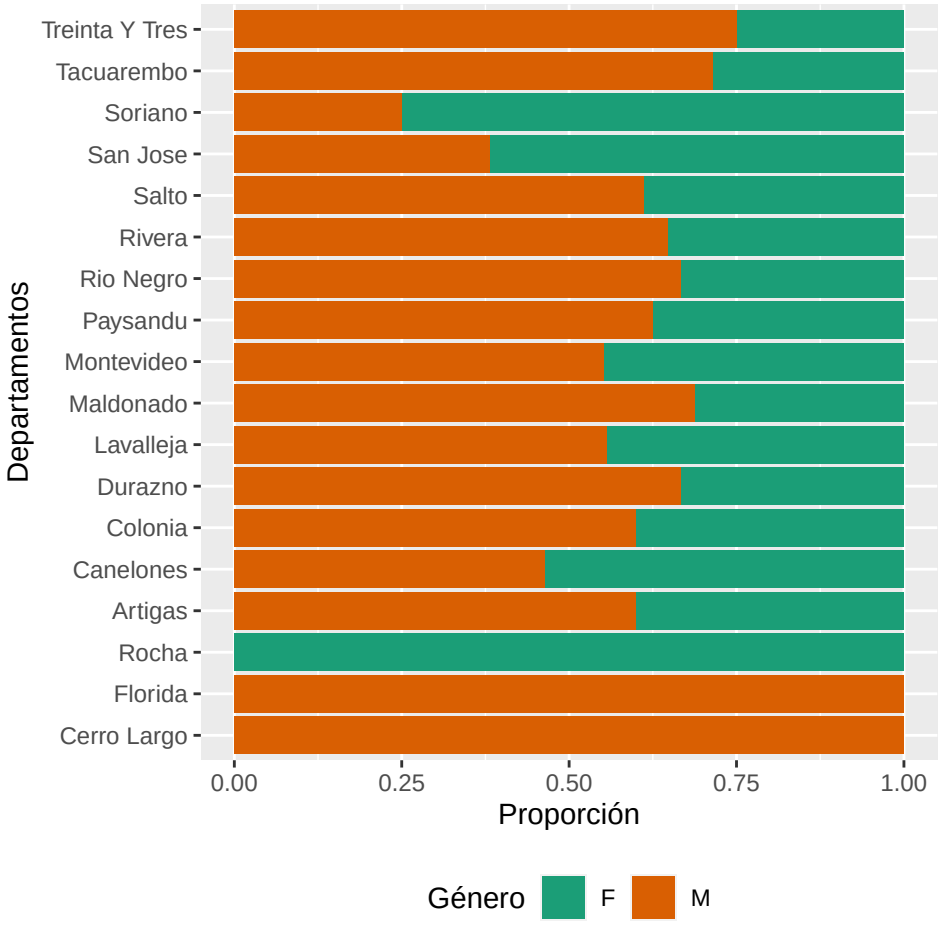

Figure 2: Proporción de abandono por departamento y por sexo

Masculino.

\textbf{\textcolor{violet}{Falta ordenar,incluir el nombre del gráfico: “Gráfico de barras apilaadas al 100\%...” Ver sol. (8 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)**

```
prop_clus<-datos %>%
  group_by(nombre_departamento,Abandono)%>%summarise(Conteo=n())%>%mutate(Proporción=round(Conteo/sum(C

datos%>%ggplot(aes(x = cl,y=reorder(nombre_departamento, Proporción)))+geom_point()+theme(aspect.ratio =
```

El eje x es Porcentaje de abandono,no proporción (3 Puntos)

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)**

```
datos<-datos %>% mutate(grupo_desc=ifelse(datos$grupo_desc%in% c("1ro. G. 1","1ro. G. 2","1ro. G. 3","1
head(datos$grupo_desc)
```

(5 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)**

```
datos %>% mutate(datos,año=separate())
```

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)**

```
datos$`Fecha nacimiento`<-as.Date(datos$`Fecha nacimiento`)
class(datos$`Fecha nacimiento`)
```

(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)**

```
datos$extra<-ifelse(datos$`Fecha nacimiento`>as.Date('2003-04-30'),1,0)
head(datos$extra,6)
```

Te quedó la condición alrevez (9 Puntos)

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)**

```
reducida<-datos%>%select(documento, Grado2013, Grado2014,Grado2015,
Grado2016 )
#reducida<-pivot_longer(reducida,names_to = "Grado",values_to = "Nivel")
```

Te faltó el argumento cols (3 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
```

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)** Porque al tener muchos niveles, la visualización se pierde un poco en el pie chart, en cambio usando un bar chart, aún teniendo muchos niveles, se podrá comparar visualmente uno con otro. Esto sucede incluso cuando los valores para cada nivel son parecidos similares, se pierde la visualización con el pie chart.

Incompleto, ver sol (3 Puntos)

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

Porque nos garantiza que las unidades en ambos ejes, x y y, sean de igual longitud. Permite una mejor visualización de los datos, dado que no "achata" o "estira" el panel del gráfico. **(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
```

Ver sol

Tu prueba no fue reproducible porque pusiste un directorio absoluto, tu repositorio estaba ordenado hasta la prueba, algunas partes de tu código se puede restructurar para que sea más sencilla su lectura (11 Puntos). TOTAL DE PUNTOS 76/130