Control 2 - Computación Estadística con R

Joshua Kunst

22 de julio de 2019

## Instrucciones

* Esto se puede resolver con todo el material visto en clases. Pero no es restrictivo en el sentido de que puedes usar otros comandos que no hemos visto.
* Se puede realizar esta evaluación en grupo de hasta tres (3) personas.
* Usted(es) debe(n) crear un script en donde se especifique los pasos necesarios (código) y comentarios (#). En el caso ideal el script debiese correr sin problemas al ejecutar los comandos en orden.
* La fecha límite para enviar la tarea es el jueves 31 de Julio de 2019 hasta las 23:59 hrs. El método de envío es via correo electrónico a [jbkunst@gmail.com](mailto:jbkunst@gmail.com) con el asunto “Control 2 - Computación Estadística con R” (copiar y pegar por favor).
* Referencias que complementan las presentaciones y scripts vistos en clases:
  + Tutorial de Riva Quiroga <https://github.com/rivaquiroga/RLadies-Santiago/blob/master/primeros_pasos_en_R.md>
  + Traducción parcial al español de “R for data science” <https://es.r4ds.hadley.nz/>

## Los datos

Usando la función read\_csv lea un archivo desde la URL<https://raw.githubusercontent.com/jbkunst/puc-introduccion-a-R/master/data/clientes.csv>, y cree con ello una tabla llamadaclientes`.

hint:

library(tidyverse)  
  
clientes <- read\_csv("https://raw.githubusercontent.com/jbkunst/puc-introduccion-a-R/master/data/clientes.csv")

Estos datos corresponden a ventas en cierto banco. Los datos de los clientes corresponden a la información al momento de la entrega del producto financiero (digamos crédito de consumo), y la variable malo\_en\_los\_12m fue un registro a posterior indicado si el cliente en el producto vendido alcanzó una mora mayor a 90, es decir: malo :(

## Explorando:

* ¿Cuántas filas y cuantas columnas contienes los datos?
* Indique como puede conocer los tipos de cada columna

## Caracterizando a los clientes

* Obtenga la distribución de mujeres y hombres en las ventas, es decir cuantos créditos de consumo fueron vendidos a mujeres y cuantos a hombres.
* ¿Cual fue la proporción (número entre 0 y 1, porcentaje) de clientes que en los 12 meses posteriores a la venta que fueron *malos*?
* Genere un histograma de la variable ingreso y comente
* Estudie si existe diferencia del ingreso según género y comente.

## Agregando más datos

Nos dicen que la edad puede ser relevante para conocer mejor a nuestros clientes. La información está en otra tabla la cual se puede leer de la siguiente manera:

clientes\_edad <- read\_csv("https://raw.githubusercontent.com/jbkunst/puc-introduccion-a-R/master/data/clientes-edad.csv")

clientes\_edad

## # A tibble: 39,514 x 2  
## rut\_cliente edad  
## <dbl> <dbl>  
## 1 7701000 32  
## 2 6089500 22  
## 3 21271500 23  
## 4 25676500 42  
## 5 16584500 19  
## 6 11618500 28  
## 7 6669000 56  
## 8 22654000 19  
## 9 4834500 32  
## 10 8229000 46  
## # ... with 39,504 more rows

* El siguiente código obtiene la distribución de las ventas segun ciertos cortes de edad:

clientes\_edad <- clientes\_edad %>%   
 # CREAR (o modificar) variable  
 mutate(  
 edad\_categorizada = cut(edad, breaks = c(0, 40, 50, 60, 100))  
 )  
  
clientes\_edad

## # A tibble: 39,514 x 3  
## rut\_cliente edad edad\_categorizada  
## <dbl> <dbl> <fct>   
## 1 7701000 32 (0,40]   
## 2 6089500 22 (0,40]   
## 3 21271500 23 (0,40]   
## 4 25676500 42 (40,50]   
## 5 16584500 19 (0,40]   
## 6 11618500 28 (0,40]   
## 7 6669000 56 (50,60]   
## 8 22654000 19 (0,40]   
## 9 4834500 32 (0,40]   
## 10 8229000 46 (40,50]   
## # ... with 39,504 more rows

clientes\_edad %>%   
 count(edad\_categorizada)

## # A tibble: 3 x 2  
## edad\_categorizada n  
## <fct> <int>  
## 1 (0,40] 30152  
## 2 (40,50] 6636  
## 3 (50,60] 2726

¿Qué puede comentar/criticar respecto a los grupos generados?

* Ahora, nos mencionan que se requiere observar las edades usando una categorización de 20 en 20. Use el código anterior para obtener lo requerido.

## Juntando los datos

* Consolide la información en una sola tabla.
* Con los grupos de la pregunta del punto anterior calcule el ingreso promedio para cada grupo. Comente si hace sentido o no el resultado obtenido.