Revisión 2024

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.hadley.nz), las presentaciones de clase, tus apuntes, 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 otros 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_2024. La revisión debe quedar en tu repositorio del curso GitHub en una carpeta que se llame Revision_2024 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 debajo del 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`.

1.1) (1 Punto) En tu repositorio del curso creá una carpeta que se llame Revision_2024 (sin tilde).

1.2) (1 Punto) Asegurate que soy colaboradora del mismo y si no lo soy mandame la invitación, mi usuario es natydasilva.

1.3) (1 Punto) Subí el archivo .qmd 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 qmd compila adecuadamente, es decir, se genera el archivo revision.pdf adecuadamente.

1.4) (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.

1.5) (2 Puntos) A parte de subir el archivo al repositorio debés mandarme por correo (natalia.dasilva@fcea.edu.uy) el archivo revision.qmd y el revision.pdf de tu versión final. Escribime en el asunto del correo Rev24_STAT_NT y en el cuerpo tu usuario de GitHub para que sea más sencillo encontrar tu repositorio con la prueba.

1.6) (2 Puntos) Presentá tu código de forma ordenada así como las respuestas a cada pregunta.

Ejercicio 2 (R Base)

(25 Puntos)

Este ejercicio debe ser resuelto con R base en caso contrario se especificará el paquete o función específica a utilizar.

2.1 (3 Puntos) Generá un `data.frame` con tres columnas nombradas como `var1`, `var2`, `var3`, y 12 filas siendo `var1` un vector numérico con posibles valores 2, 3 y 4, con más de una observación para cada posible valor, `var2` un factor con tres niveles, `alto`, `medio` y `bajo` con más de una observación para cada posible valor y `var3` un vector lógico Guardá el `data.frame` en un objeto llamado `datos`

2.2 (2 Puntos) Usando R base, seleccioná del conjunto datos la tercera columna en base a su nombre (`var3`)

2.3 (5 Puntos) Usando R base, seleccioná las filas de `datos` donde `var1` es 3, `var2` es medio y `var3` es FALSE.

2.4 (3 Puntos) Usando R base, seleccioná las filas donde `var2` toma los valores medio y alto

2.5 (2 Puntos) Hacer una tabla que cuente la cantidad de observaciones para cada combinación de `var2` y `var3`

2.6 (10 Puntos) Usando un for loop generá una nueva variable en el objeto `datos` llamada `var4` que valga 1 si (`var2` es alto y `var3` es TRUE) o (`var2` es medio y `var3` es TRUE) y vale 0 en otro caso

Ejercicio 3 (tidyverse)

(35 Puntos)

Los datos que vamos a utilizar son del catálogo de datos abertos (https://catalogodatos.gub.uy). Especificamente utilizaremos los datos de e las Encuestas EUTIC (Encuestas de uso de TIC) que realiza el Instituto Nacional de Estadística en conjunto con la división Agenda Digital y Observatorio de la Sociedad de la Información de AGESIC. Por cada año de las encuestas realizadas se publica el recurso con los datos, además se publican por cada año el diccionario de variables y el formulario de la encuesta. En la carpeta Revisión_2024 encontrarás dos archivos `eutic.xlsx` que contiene los datos de la encuesta para el 2019 y el archivo `diccionario.xls` que contiene los metadatos de dicha encuesta (descripción de variables, codificación, etc).

3.1 (2 Puntos)

Leer los datos `eutic.xlsx` usando el paquete `readxl` y `here` para asegurar la reproducibilidad en caso que lo necesites, guardá los datos en un objeto llamado `tic`.

3.2 (2 Puntos)

Vamos a trabajar con un subconjunto de los datos. Generá un nuevo objeto llamado `tic_red` que contenga las variables desde `C7` hasta `C9` y `C14_1` hasta `C14_18`(en el archivo `diccionario.xls` podés encontrar el nombre de todas las variables junto con su descripción). El conjunto de datos reducidos (`tic_red`) debe tener 2341 filas y 21 columnas.

De ahora en más trabajaremos con el conjunto de datos reducidos `tic_red`

3.3 (5 puntos)

Usando funciones del paquete `ggplot2`, hacé un gráfico de barras que contenga la cantidad de personas que utilizó alguna vez internet (`C9`). Cambiá los ticks del eje `x` donde dice 1 debe decir `Si` y donde dice `2` debe decir `No` y en el resto de los ticks nada. Para esta parte debes notar que la variable `C9` es numérica. Cambiá la etiqueta del eje `y` a Cantidad de personasy `la del ejexa`Usa internet‘.

3.4 (7 Puntos) Esta pregunta tiene tres partes:

1. Renombrá la variable `C9` como `usoint` y `C7` como `sexo`
2. Transformá la variable `usoint` y `sexo` en una variable de tipo factor

3. Recodificá la variable `usoint` para que `1` sea `Si` y `2` sea `No` y recodificá `sexo` para que `1` sea `Hombre` y `2` sea `Mujer`
4. Guardá los datos modificados en `tic_reco`
5. Luego de todos los pasos anteriores repetí el gráfico de la Pregunta 3.3 con los datos `tic_reco` notando el paso que ya no es necesario luego de las tranformaciones realizadas.

3.5 (4 puntos) Modifique el gráfico anterior para obtener las barras en términos de **proporción** en vez de cantidad modificando las leyendas de los ejes de forma apropiada.

3.6 (4 Puntos) Utilizando funciones de `dplyr` respondé ¿Qué proporción de personas que utilizan internet son Mujeres ?

3.7 (6 Punto)

Usando ggplot2 realiza una visualización apropiada a tu elección para ver la distribución de las edades de las personas que nunca usaron internet según sexo. Debes incluir etiquetas apropiadas para los ejes y título. Describe brevemente qué ves en el gráfico y comentá si hay o no diferencias en la edad entre hombres y mujeres.

A su vez usando funciones de `dplyr` respondé ¿Cuál es el promedio de edad de las personas que nunca usaron internet para cada sexo?

3.8 (5 Puntos)

De forma similar al punto 3.4 re-codificá las variables `C14_1` hasta `C14_18`. Son 18 columnas, utiliza alguna función de `dplyr` que opere sobre múltiples columnas al mismo tiempo. Al revisar los meta datos este grupo de variables tiene la misma codificación, es decir que `1` sea `Si`, `2` sea `No` y `3` sea `No contesta`. Guardá el nuevo conjunto de datos en un objeto `final`.

Nota: Con solo re-codificar la variable es suficiente, no es necesario transformarlo en factor aunque si querés podés hacerlo.

Ejercicio 4 (Varios)

(30 Puntos)

4.1 (15 Puntos)

Generá una función `pedidos` 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 el pedido y 0 en caso contrario. El pedido se puede realizar siempre que haya stock suficiente para cada producto, es decir que la cantidad disponible sea igual o mayor a la cantidad pedida. 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”.

Comprobá que el resultado de la función sea

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

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

4.2 (5 Puntos)

Generá un vector numérico de 24 valores simulados de una normal con media 20 y desvío 2 usando `rnorm` y nombralo `randnorm` y otro vector de la misma longitud generados con `rbinom` con probabilidad 0.5 y número de eventos posibles 10 nombrá al vector como `randbin`.

Utilizando un `while` hacé la suma de los dos vectores (`randnorm` y `randbin`)

4.3 (2.5 Puntos) 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?

4.4 (2.5 Puntos)

¿Porqué es necesario utilizar `aspect.ratio = 1` en un diagrama de dispersión?

4.5 (2.5 Puntos)

¿Con tus palabras definí qué es la investigación reproducible y mencioná alguna herramienta que lo facilita y porqué?

4.6 (2.5 Puntos)

¿Con tus palabras definí qué es la ciencia de datos y qué rol tiene la estadística en ella?