FACULTAD DE CIENCIAS ECONÓMICAS Y DE ADMINISTRACIÓN

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

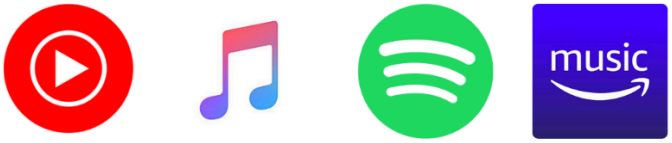
Informe de avance | Proyecto 'Estadística Descriptiva 2024'

Juan Carella, Claudio Ríos, Mariana Rodríguez, Faustino Valiente, Brandon Vidal.

Facultad de Ciencias Económicas y de Administración

Responsable del curso: Ignacio Alvarez-Castro

ignacio.alvarez@fcea.edu.uy



Entrega Informe Avance

Informe de avance | Trabajo grupal

En este informe mostraremos, los nombres de cada participante, el tema elegido, cierta información del 'dataset' seleccionado, la pregunta principal, entre otros.

Integrantes del grupo número 3.

Nombre	CI	email
Juan Carella	5.503.311-9	juanignaciocarella@gmail.com
Claudio Ríos	5.606.239-7	claudioriosstabile@gmail.com
Mariana Rodríguez	5.540.434-4	marodper05@gmail.com
Faustino Valiente	5.568.670-4	faustino.valiente130206@gmail.com
Brandon Vidal	5.465.968-7	brandonvidal677@gmail.com

Tema elegido - Preguntas orientadoras.

Nuestro trabajo se enfocará en el análisis de las canciones más escuchadas en el mundo entero en el año 2023, contamos con información acerca de sus características musicales, que explicaremos tanto en el video como en la parte escrita.

¿Qué analizamos? Trabajaremos comparando las canciones entre sí para, basándonos en sus características, descubrir **¿qué transforma a una canción en popular?, ¿con qué características cuenta una canción altamente reproducida?**

Durante este informe trabajaremos con la variable 'streams' (escuchas, reproducciones). Veremos un histograma que nos muestra la cantidad de canciones (eje Y) que tienen cierta cantidad de reproducciones (eje X).

Dataset.

Encontramos nuestro dataset en la siguiente página:

- Most Streamed Spotify Songs 2023

Este dataset contiene información sobre canciones populares y sus características musicales. A continuación mencionamos algunas de ellas:

- “track_name”: Nombre de la canción (cerca de 1000 canciones)
- “artist(s)_name”: Nombre del o de los artistas.
- “realeased_year”: Año de lanzamiento.
- “realeased_month”: Mes de lanzamiento.
- “streams”: Número total de reproducciones.
- “bpm”: Beats por minuto, tempo.
- “key”: Tono/Escala de la canción (nomenclatura anglosajona)
- “mode”: Modo, si el tono es mayor o menor.
- “danceability_%”: Porcentaje que mide la capacidad de una canción para ser bailada.
- “valence_%”: Medida de positividad o negatividad que se transmite a través de una pista musical. Mientras más alto, mayor es la positividad de la pista.
- “energy_%”: Esta variable, mide la energía de la canción, a mayor porcentaje, más intensa es la canción.
- “instrumentalness_%”: Porcentaje de instrumentalidad, a menor porcentaje, mayor cantidad de voces contiene la canción.
- “liveness_%”: Esta variable mide la detección de público en una canción. Si, por ejemplo, estamos hablando de una canción en vivo, con público presente, tendríamos un porcentaje alto de “liveness”.

Este dataset cuenta con 952 filas y 13 columnas.

Histograma variable ‘streams’

Captura de pantalla código R.

Tuvimos que eliminar una fila, puesto que tenía un error y generaba fallas en las gráficas.

```r
#Cargamos el dataset
spotifyr<- read.csv('C:/Users/Usuario/Documents/trabajoED24/spotify-2023.csv')

str(data.spotify)
head(data.spotify)

#Eliminamos fila con error
data.spotify<-spotifyr[-575,]
View(data.spotify)

#Cargamos librerias
library(ggplot2)
library(tidyverse)

```

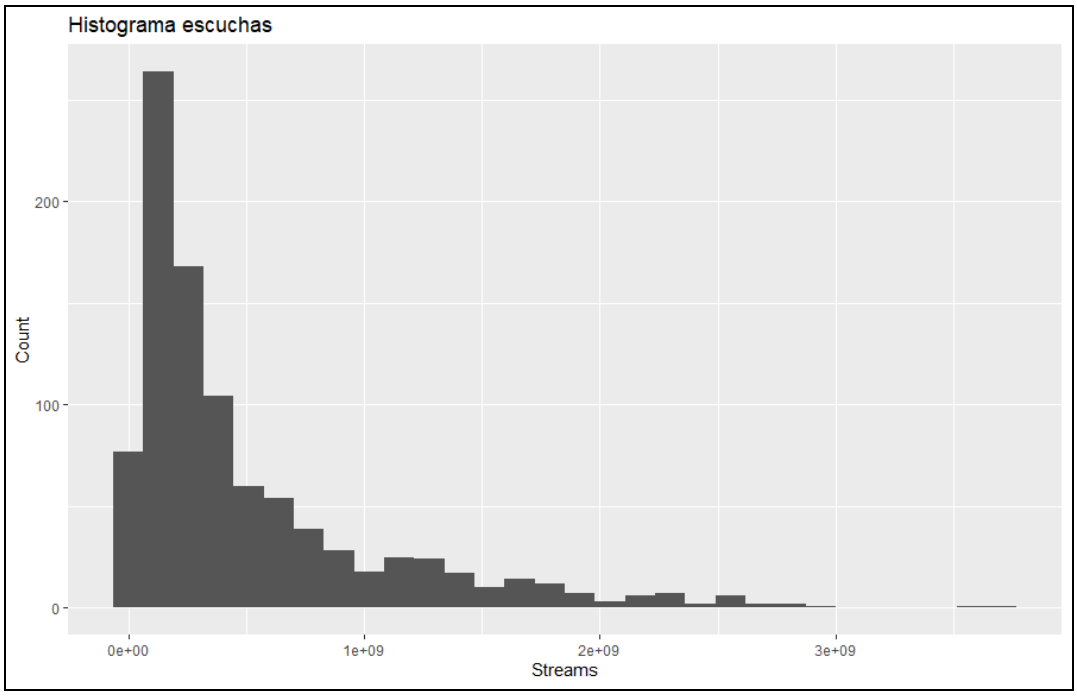
Tuvimos que convertir la variable ‘streams’ a numérica. Código para generar histograma:

```r
#Variable 'streams'
ggplot(data.spotify)+
  geom_histogram(aes(x=streams), bins = 30)+
  labs(title = 'Histograma escuchas', x='Streams', y='Count')

data.spotify$streams<-as.numeric(data.spotify$streams)

```

Histograma:



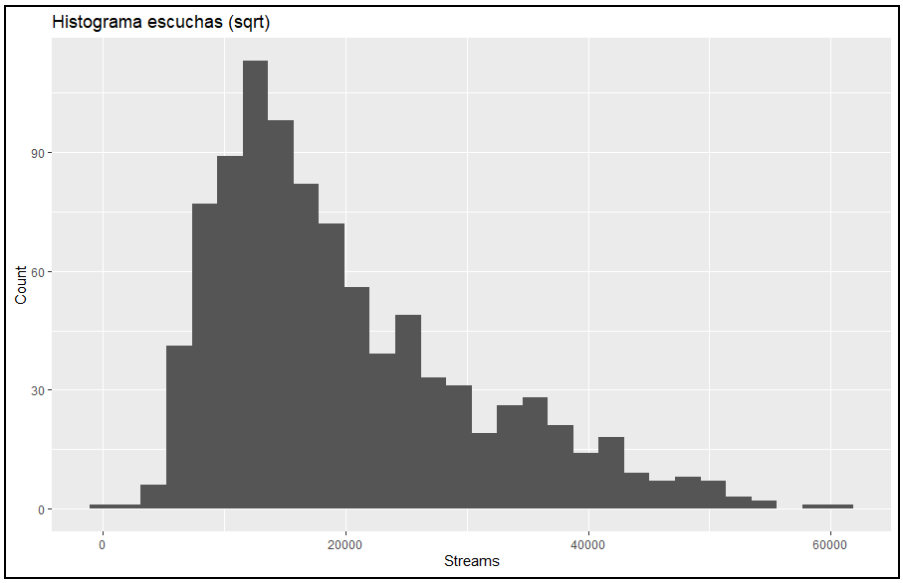
Entrega Informe Avance

Configuración del histograma.

Debido a comentarios del profesor, transformamos la variable para simplificar su descripción estadística. El cambio fue el siguiente:

```
26 ggplot(data.spotify)+
27   geom_histogram(aes(x= sqrt(streams)), bins = 30)+
28   labs(title = 'Histograma escuchas (sqrt)', x='Streams', y='Count')
29
```

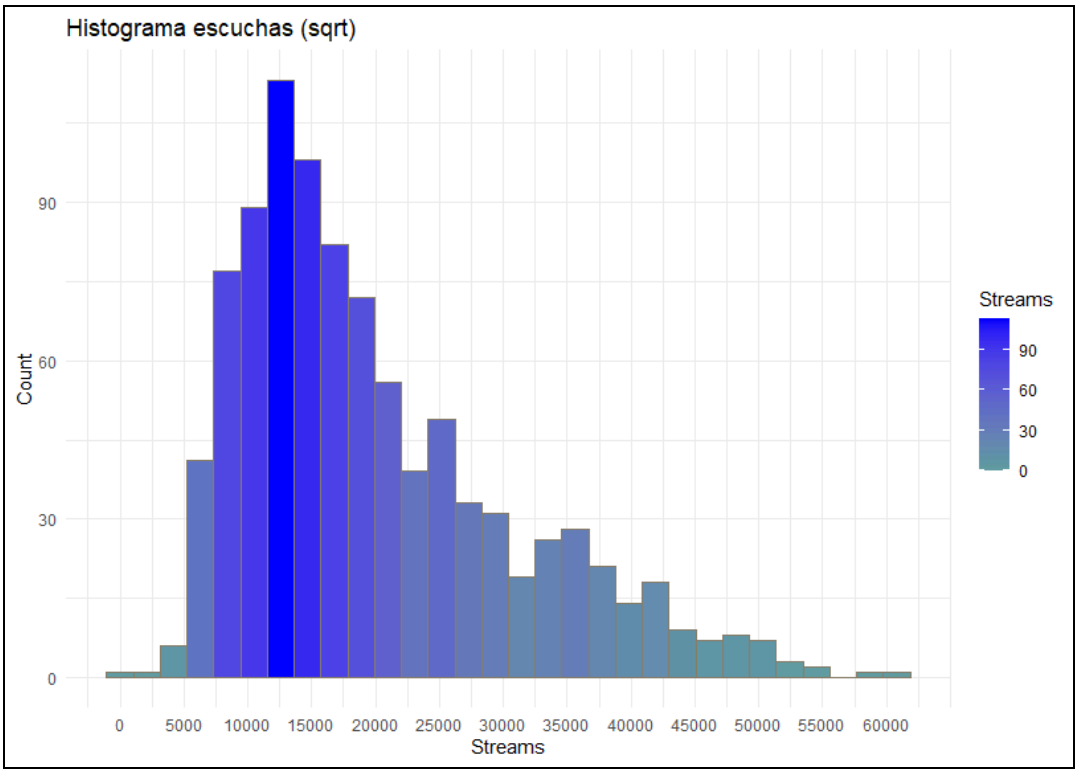
Intentamos cambiar la variable por '*log10*', pero preferimos utilizar la raíz cuadrada de cada elemento.



Agregamos algunos elementos estéticos y utilizamos la función "*scale_x_continous*" con el fin de aportar más información en el eje x.

```
35 ggplot(data.spotify)+
36   geom_histogram(aes(x= sqrt(streams), fill=..count..), bins = 30 , color='bisque4')+
37   scale_fill_gradient('Streams', low = 'cadetblue', high = 'blue')+
38   labs(title = 'Histograma escuchas (sqrt)', x='Streams', y='Count')+
39   scale_x_continuous(breaks = seq(0, max(sqrt(data.spotify$streams)), by = 5000)) +
40   theme_minimal()
```

El **resultado** es el siguiente:



¿Qué vemos en este histograma? Podemos sacar diferentes conclusiones viendo la gráfica, observamos una mayor concentración de escuchas entre 10.000 y 20.000 generando así un histograma unimodal asimétrico. También, utilizando las funciones de R (usando nuevamente la raíz cuadrada de cada valor) podemos calcular ciertas medidas de posición:

```
> mean(sqrt(data.spotify$streams))
[1] 19936.87
> median(sqrt(data.spotify$streams))
[1] 17044.97
> sd(sqrt(data.spotify$streams))
[1] 10806.54
> var(sqrt(data.spotify$streams))
[1] 116781274
> max(sqrt(data.spotify$streams))
[1] 60859.63
> min(sqrt(data.spotify$streams))
[1] 52.55473
```

Siguiendo el orden detallado en la imagen comentamos, la media del dataset es 19.936, la mediana es 17.044, la desviación estándar es 10.806, la varianza es 116.781.274, el valor máximo es 60.859 y por último, el valor mínimo es 52.

Discusión y conclusiones

En este breve informe comentamos los primeros pasos del proyecto final de curso, vimos ya un pequeño análisis de la variable ‘streams’. La transformación de esta variable utilizando la raíz cuadrada de cada valor nos permitió una mejor descripción estadística. La media, mediana, desviación estándar y varianza, nos proporcionaron una comprensión cuantitativa de la gráfica de la variable a analizar.

En las siguientes entregas, podremos responder a las preguntas planteadas en el inicio de este informe. Además de profundizar en análisis más complejos.

Concluimos que con esto establecemos una base para iniciar con las siguientes etapas del estudio.

Referencias - Bibliografía.

Irizarry, Rafael. Introducción a la ciencia de datos (Análisis de datos y algoritmos de predicción con R). Actualización 2023-02-09

Irizarry, Rafael. Ggplot2— Cheatsheet

Introducción a la ciencia de datos

CheatSheetA

CheatSheetB

Álvarez, Ignacio. (2024). Apuntes y diapositivas de clase “Estadística Descriptiva”.

EVA-FCEA

Paterno, Gustavo. (2018). Guía de Bolso: ggplot2 | Gráficos elegantes no R.

Guia de Bolso: ggplot2 | Gráficos elegantes no R