Pontificia Universidad Católica de Chile
Facultad de Ciencias Económicas y Administrativas

Análisis de Big Data:
EAA361A

Laboratorio 5

Profesor: Cristián Vásquez Ayudante: Pablo González

Aclaración:

Hoy principalmente buscaremos replicar los ejercicios 1 y 3 del segundo laboratorio mediante el uso de R. Las librerías principales con las que se trabajará pertenecen al "tidyverse" y, dentro de estas, se recomienda familiarizarse especialmente con "dplyr" y su interacción con "ggplot".

Ejercicio 1: Caso Pokémon

Luego de una larga travesía, Ash Ketchum (**サトシ** para los amigos) vuelve a casa con su Pokédex completa. Ésta es luego entregada al profesor Oak para su posterior análisis, quien le pide a usted, que, utilizando sus bastos conocimientos de R, responda las siguientes preguntas:

1. Obtenga el número, nombre, nombre en japonés, clasificación y si es o no legendario, de aquellos Pokémon cuyo ataque triplique su defensa.
2. ¿Cuáles son los nombres (en ambos idiomas) de los 10 Pokémon con una mayor felicidad base? ¿Son estos legendarios? (Responda "Sí" o "No")
3. De aquellos Pokémon legendarios, ¿qué porcentaje pertenece a cada generación?
4. Genere una función en R que le permita generar gráficos circulares con ggplot.
5. Utilice la función creada para generar gráficos en relación con los ejercicios anteriores.

** Datos extraídos de www.kaggle.com/rounakbanik/pokemon/data con fecha 23/03/2020*

Ejercicio 2: Recomendación de películas

Utilizando la versión ligera de "MovieLens", una base de datos con información de calificaciones y etiquetados para aproximadamente 10.000 películas, responda las siguientes preguntas:

0. Genere una nueva tabla, separando el título y año de la película en distintas variables.
1. Utilizando una medida arbitraria $\mu_{Rating} \times Log(N° \, Ratings)$, muestre los títulos y puntajes obtenidos por película, ordenándolos descendentemente. Grafique el promedio anual de nuestra métrica para los últimos 20 años, considerando únicamente a aquellos registros con una medida mayor a 0.
2. Obtenga los nombres de las 10 películas con mayor cantidad de votaciones en la categoría comedia. Grafique.
3. Obtenga los títulos, calificación promedio y géneros de las 10 mejores películas de animación no infantil producidas dentro de los últimos 3 años. Utilice solo el rating. Posteriormente, grafique la calificación promedio de dichas películas.
4. Muestre el título, etiqueta más común y rating promedio de aquellas 10 películas con un etiquetado más consistente (con más repeticiones).
5. **Propuesto 1:** Obtenga las 50 películas con menor calificación promedio según los 3 usuario que han escrito más reseñas.
6. **Propuesto 2:** Utilizando la medida arbitraria descrita en el ejercicio 1 y, considerando las 100 mejores películas según ésta, obtenga las 5 etiquetas más repetidas en sus descripciones.

** Datos extraídos https://grouplens.org/datasets/movielens/ con fecha 28/03/2020*