Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación

IIC2115 – Programación como Herramienta para la Ingeniería (II/2022)

Ejercicio Capítulo 2a

Aspectos generales

- **Objetivos:** Aplicar los contenidos de análisis exploratorio de datos para completar y expandir una base de datos incompleta y responder consultas sobre la misma.
- **Lugar de entrega:** lunes 11 de abril a las 22:00 hrs. en repositorio privado.
- **Formato de entrega:** archivo Python Notebook (**C2a.ipynb**) con el avance logrado durante la sesión. El archivo debe estar ubicado en la carpeta **C2a**. Utilice múltiples celdas de texto y código para facilitar el trabajo del cuerpo docente.

Introducción

Con el fin de ejercitar los contenidos de análisis exploratorio de datos en Python, en este ejercicio deberá realizar los pasos básicos para expandir y completar una base de datos con información faltante. El cómo hacerlo en cada caso será una decisión de uds., que deberá ser tomada y **JUSTIFICADA** en base a las características de los datos analizados. Además de esto, una vez teniendo la base de datos completa, deberá contestar una serie de consultas con respecto a los datos, que requerirán el uso de técnicas de agregación, agrupación y visualización.

Descripción del problema

Netflix es un servico de *streaming* muy conocido con vasto catálogo de películas, series. La empresa ha hecho un concurso de "Director/a por 1 día!!" y usted ha ganado el ticket dorado! Como persona cinéfila, se ha propuesto crear una película de taquilla y para eso, decide analizar las tendencias en el mundo del *streaming*.

Ha descubierto que han subido a la plataforma de datasets públicos Kaggle con el catálogo de Netflix hasta el 2021. Su objetivo es analizarlo para ayudarle con su nueva película.

La base de datos

La base da datos se encuentra disponible en el sitio del curso, en el archivo `netflix.csv`. Esta contiene información del catálogo de series y películas de Netflix por medio de las siguientes columnas:

show_id	Identificador único del programa
type	Indica si es una película o serie
title	Título del programa
director	Nombre de quién(es) dirigieron la producción
country	País de producción
date_added	Año en que se añade al catálogo de Netflix
release_year	Año de estreno
rating	Clasificación por edades
duration	Duración en minutos
listed_in	Lista de géneros/clasificaciones a las que pertenece la película

Misiones

1. **Carga y exploración**: cargue el archivo con los datos y describa su contenido, indicando qué columnas tienen información incompleta. Debe eliminar la presencia de elementos duplicados. Finalmente, visualice las variables, con el fin de evaluar la existencia de *outliers*.

2. **Imputación y eliminación**: para cada una de las columnas con elementos faltantes, impute los valores en base a algún criterio basado en los datos. Además de esto, analice la posible eliminación de filas y columnas completas, en base a los valores faltantes y la relación entre las columnas.

3. **Expandir el dataset**: considerando los datos que contiene la columna `listed_in`, genere una nueva tabla que para cada `show_id` tenga columnas binarias indicando si el programa pertenece a una de las siguientes categorías:

    - `Action & Adventure`
    - `Comedy`

- `Drama`
- `Horror`
- `Mystery`
- `Sci-Fi & Fantasy`
- `Thrillers`
- `Independent`
- `Children & Family`

Se recomienda usar el comando `str.contains()`. Hay que tener cuidado con categorías que sean sinónimos o la palabra se escriba de formas diferentes. En el archivo `categories.csv` puede ver una lista de las 42 categorías que existen. Finalmente, debe unir esta tabla con la original.

4. **Consultas**: conteste cada una de las siguientes consultas, justificando los análisis y supuestos realizados:

   a) ¿Cuál es el género más popular en los últimos 5 años? ¿En qué década hay más películas de ese género? ¿Y qué país ha producido más películas de este tipo?

   b) En promedio, ¿cuánto demora en llegar un programa al catálogo de Netflix considerando desde la creación del catálogo de *streaming* (2008)?

   c) Analice la tendencia en el tiempo del género de `Horror` en comparación con la clasificación por edades.