Asignatura	Datos del Equipo		Fecha
Herramienta de Visualización	Integrantes:	Román Cedillo Angeles, Damian Menchaca Muñoz Armando Gómez Garrigós	23-05-2021

Maestría en Análisis y Visualización de Datos Masivos

Universidad Internacional de la Rioja

Trabajo: Visualizando nuestro primer *dataset*

Por medio del presente se pretende tener un primer acercamiento con las herramientas y el proceso de visualización, desde seleccionar un dataset, limpiar y analizar los datos, representarlos gráficamente y verificar que el mensaje que pretendo se entiende.

Curso: Herramientas de Visualización.

Presenta: Román Cedillo Angeles

Damian Menchaca Muñoz

Armando Gómez Garrigós

Profesor: Luis Fernando Franco Jimenez.

Asignatura	Datos del Equipo		Fecha
Herramienta de Visualización	Integrantes:	Román Cedillo Angeles, Damian Menchaca Muñoz Armando Gómez Garrigós	23-05-2021

Índice.

Índice.

Introducción.

Desarrollo.

- El dataset, limpieza y análisis de los datos.
- Las herramientas utilizadas.

Visualización.

Conclusiones.

Participación de los integrantes.

Fuentes de información.

Introducción.

En esta primera visualización se escogió un dataset que consta de programas de televisión y películas disponibles en Netflix a partir de 2019, el cual fue descargado de Kaggle en la url: https://www.kaggle.com/shivamb/netflix-shows.

El dataset en cuestión consta de 7,777 registros de programación agregada en Netflix y que está disponible a partir de 2019, los campos que contiene el dataset son (Kaggle,2021):

show_id: Identificación única para cada película, programa de televisión.
type: categoría para película o programa de televisión [movie | TV Show].
title: título de la película o programa de televisión.
director: Director de la película.
cast: Actores involucrados en la película o programa de TV.
country: País donde la película o programa de TV fue producida.
date_Added: Fecha en que se agregó en Netflix.
release_year: Año de lanzamiento de la película o programa de TV.
rating: Clasificación de la película o programa de TV.
duration: Duración total en episodios o número de minutos.
listed_in: define en qué clasificación se ha puesto la película o programa de TV, ej. TV Commedies, TV Drama, ids, etc.
description: Breve descripción de la película o capítulo de programa de TV.

Como herramientas para presentar gráficamente se eligen varios programas gratuitos para presentar estos datos.

Asignatura	Datos del Equipo	Fecha
Herramienta de Visualización	Integrantes: Román Cedillo Angeles, Damian Menchaca Muñoz Armando Gómez Garrigós	23-05-2021

Desarrollo.

El dataset, limpieza y análisis de los datos.

El dataset en cuestión contiene las recomendaciones descritas en el material de UNIR del tema 1, las cuales son: Contiene varias variables que pueden ser comparadas entre sí, Contiene valores de tiempo, contiene información de varios países (UNIR, 2021).

Por otra parte presenta los siguientes errores o falta de estándares:

- El campo director es nulo en la mayoría de los programas de TV, sin embargo no puede considerarse una regla ya que tiene excepciones.
- Los campos cast, country y listed_in son arreglos es decir se constituyen de uno o varios elementos.
- El campo date_added no contiene valores de fecha, es decir son fechas en formato de texto y no presenta estándares, las fechas están descritas en español, en ingles, algunas abreviadas, etc.

En esta primera actividad se convirtió el campo date_added en fecha para poder representar los registros a través del tiempo, al concluir esta actividad existen 10 registros con valor nulo para este campo, por lo que se tomó la decisión de agregarlos al 01 de Enero de su año de lanzamiento.

Las herramientas utilizadas.

En esta primera actividad se eligen varios programas para realizar la visualización todas gratuitas.
Para iniciar el proceso de limpieza de los datos, se utiliza únicamente Google Sheets, este software se utilizó también para hacer las gráficas.
Con el propósito de integrar la visualización se utilizó la herramienta de dibujos de Google.

Asignatura	Datos del Equipo		Fecha
Herramienta de Visualización	Integrantes:	Román Cedillo Angeles, Damian Menchaca Muñoz Armando Gómez Garrigós	23-05-2021

Visualización.

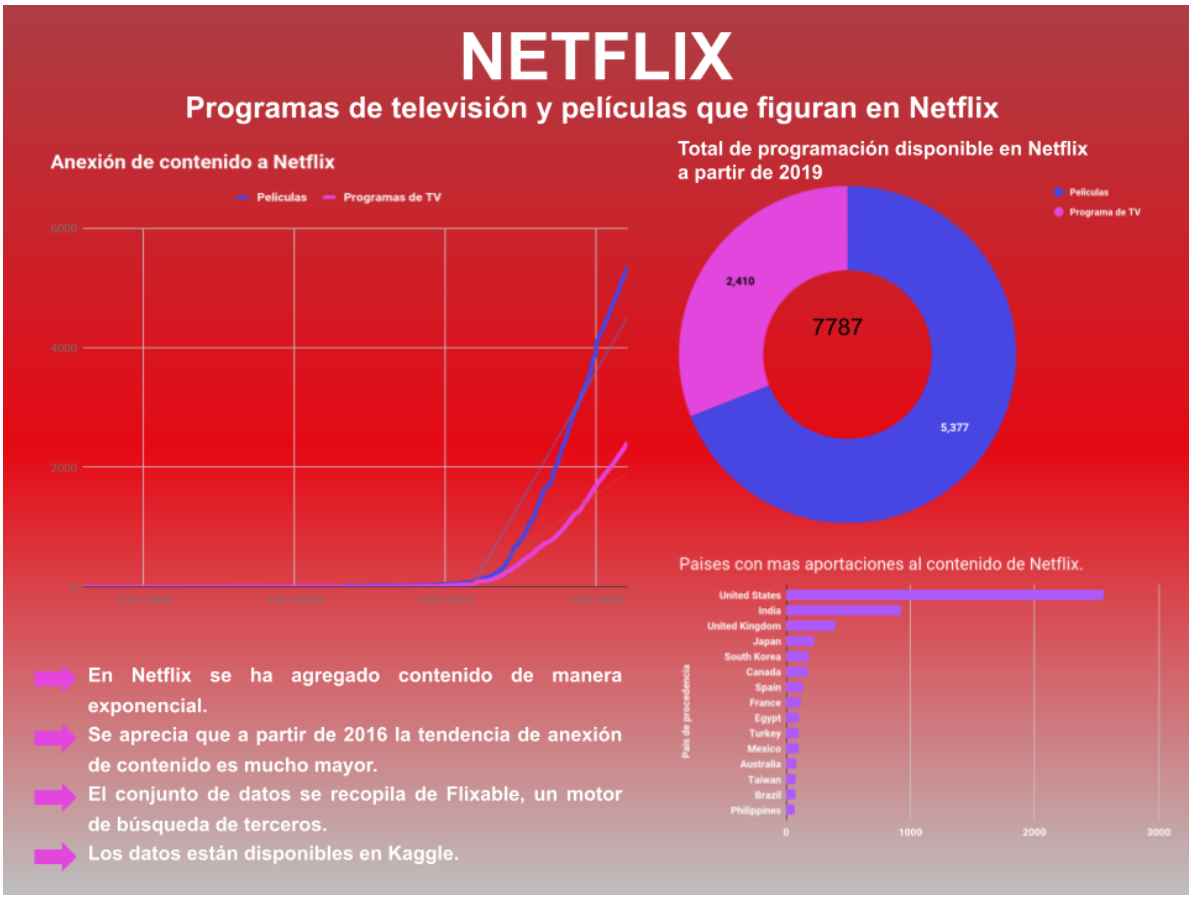

Figura 1. Visualización de la dataset de netflix. Kaggle 2019

Conclusiones.

Las fuentes de información en Big Data pueden ser de terceros y por ello no tienes control sobre estas, por lo que es de suma importancia el trabajo de limpieza de los datos para poder trabajar el análisis y visualización de los mismos.

Para tener un primer acercamiento con la visualización en la mayoría de los casos se utiliza software muy básico pero que nos puede dar líneas de acción para lograr la visualización que pretendemos.

Participación de los integrantes.

Román Cedillo Angeles: Realización de la investigación, realización del trabajo de visualización, Búsqueda del dataset de netflix en Kaggle, depuración de los datos, utilización de google spreadsheet.

Damian Menchaca Muñoz: Integración del equipo de trabajo, retroalimentación y mejora del documento, investigación y verificación de los datos de Kaggle.

Armando Gómez Garrigós: Revisión de documentación, proposición de ideas de mejora para posibles trabajos a futuro, búsqueda de diversos dataset para el entregable.

Fuentes de información.

Anonimo. (2020). TV Shows and Movies listed on Netflix. 27/05/2021, de Kaggle LLC Sitio web: https://www.kaggle.com/shivamb/netflix-show.

UNIR. (2021). Seleccionar un dataset. 27/05/2021, de Universidad Internacional de La Rioja Sitio web: https://micampus.unir.net/courses/19315/external_tools/104869.