Miembro: Luis Castro Badilla

1. Contexto. Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información.

   Personalmente soy un gran aficionado al cine, todas las semanas visito el cine y miro un par de películas por televisión o servicios de streaming.

   El sitio IMDB.com es la referencia mas importante en el internet para el tema cinematográfico. Me intereso obtener un dataset de las películas más taquilleras es Estados Unidos de América. Los valores de recaudación son muy interesantes y me interesa automatizar esta extracción para más adelante hacer un análisis e incluso tratar de predecir taquillas.

2. Definir un título para el dataset. Elegir un título que sea descriptivo.

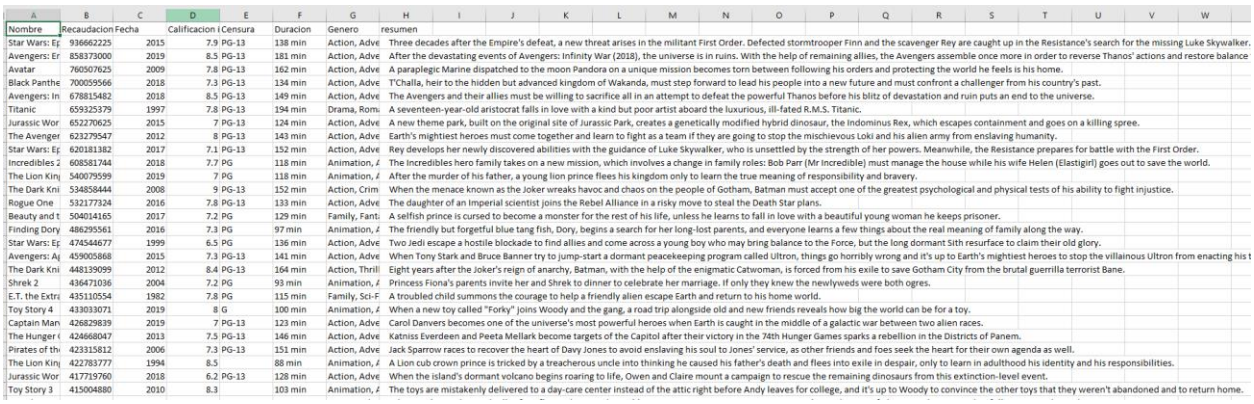
   Top 50 películas más taquilleras en USA según IMDB

3. Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido).

   El dataset consta de 50 filas que equivalen a las 50 películas que más dinero han recaudado en Estados Unidos de América, cuenta con 8 columnas (Nombre, Recaudación, Fecha, Calificación imdb, Censura, Duración, Genero, resumen) todas estas características describen a las películas.

4. Representación gráfica. Presentar una imagen o esquema que identifique el dataset visualmente

Nombre	Recaudacion	Fecha	Calificacion imdb	Censura	Duracion	Genero	resumen
pelicula 1	entero representa el total de dinero en dolares US	año	float del 0 al 10	rating segun la MPAA: PG, R, etc	duracion en tiempo de la pelicula	genero segun IMDB	resumen de la pelicula segun IMDB
...							



Nombre	Recaudacion	Fecha	Calificacion i	Censura	Duracion	Genero	resumen
Star Wars: Ep	936662225	2015	7.9	PG-13	138 min	Action, Adve	Three decades after the Empire's defeat, a new threat arises in the militant First Order. Defected stormtrooper Finn and the scavenger Rey are caught up in the Resistance's search for the missing Luke Skywalker.
Avengers: Er	858373000	2019	8.5	PG-13	181 min	Action, Adve	After the devastating events of Avengers: Infinity War (2018), the universe is in ruins. With the help of remaining allies, the Avengers assemble once more in order to reverse Thanos' actions and restore balance
Avatar	760507625	2009	7.8	PG-13	162 min	Action, Adve	A paraplegic Marine dispatched to the moon Pandora on a unique mission becomes torn between following his orders and protecting the world he feels is his home.
Black Panthe	700059566	2018	7.3	PG-13	134 min	Action, Adve	T'Challa, heir to the hidden but advanced kingdom of Wakanda, must step forward to lead his people into a new future and must confront a challenger from his country's past.
Avengers: In	678815482	2018	8.5	PG-13	149 min	Action, Adve	The Avengers and their allies must be willing to sacrifice all in an attempt to defeat the powerful Thanos before his blitz of devastation and ruin puts an end to the universe.
Titanic	659325379	1997	7.8	PG-13	194 min	Drama, Roma	A seventeen-year-old aristocrat falls in love with a kind but poor artist aboard the luxurious, ill-fated R.M.S. Titanic.
Jurassic Wor	652270625	2015	7	PG-13	124 min	Action, Adve	A new theme park, built on the original site of Jurassic Park, creates a genetically modified hybrid dinosaur, the Indominus Rex, which escapes containment and goes on a killing spree.
The Avenger	623279547	2012	8	PG-13	143 min	Action, Adve	Earth's mightiest heroes must come together and learn to fight as a team if they are going to stop the mischievous Loki and his alien army from enslaving humanity.
Star Wars: Ep	620181382	2017	7.1	PG-13	152 min	Action, Adve	Rey develops her newly discovered abilities with the guidance of Luke Skywalker, who is unsettled by the strength of her powers. Meanwhile, the Resistance prepares for battle with the First Order.
Incredibles 2	608581744	2018	7.7	PG	118 min	Animation, A	The Incredibles hero family takes on a new mission, which involves a change in family roles: Bob Parr (Mr Incredible) must manage the house while his wife Helen (Elastigirl) goes out to save the world.
The Lion Kin	540079599	2019	7	PG	118 min	Animation, A	After the murder of his father, a young lion prince flees his kingdom only to learn the true meaning of responsibility and bravery.
The Dark Kni	534858444	2008	9	PG-13	152 min	Action, Crim	When the menace known as the Joker wreaks havoc and chaos on the people of Gotham, Batman must accept one of the greatest psychological and physical tests of his ability to fight injustice.
Rogue One	532177324	2016	7.8	PG-13	133 min	Action, Adve	The daughter of an Imperial scientist joins the Rebel Alliance in a risky move to steal the Death Star plans.
Beauty and t	504014165	2017	7.2	PG	129 min	Family, Fanta	A selfish prince is cursed to become a monster for the rest of his life, unless he learns to fall in love with a beautiful young woman he keeps prisoner.
Finding Dory	486295561	2016	7.3	PG	97 min	Animation, A	The friendly but forgetful blue tang fish, Dory, begins a search for her long-lost parents, and everyone learns a few things about the real meaning of family along the way.
Star Wars: Ep	474544677	1999	6.5	PG	136 min	Action, Adve	Two Jedi escape a hostile blockade to find allies and come across a young boy who may bring balance to the Force, but the long dormant Sith resurface to claim their old glory.
Avengers: Ag	459005868	2015	7.3	PG-13	141 min	Action, Adve	When Tony Stark and Bruce Banner try to jump-start a dormant peacekeeping program called Ultron, things go horribly wrong and it's up to Earth's mightiest heroes to stop the villainous Ultron from enacting his t
The Dark Kni	448139099	2012	8.4	PG-13	164 min	Action, Thrill	Eight years after the Joker's reign of anarchy, Batman, with the help of the enigmatic Catwoman, is forced from his exile to save Gotham City from the brutal guerrilla terrorist Bane.
Shrek 2	436471036	2004	7.2	PG	93 min	Animation, A	Princess Fiona's parents invite her and Shrek to dinner to celebrate her marriage. If only they knew the newlyweds were both ogres.
E.T. the Extra	435110554	1982	7.8	PG	115 min	Family, Sci-F	A troubled child summons the courage to help a friendly alien escape Earth and return to his home world.
Toy Story 4	433033071	2019	8	G	100 min	Animation, A	When a new toy called "Forky" joins Woody and the gang, a road trip alongside old and new friends reveals how big the world can be for a toy.
Captain Mar	426829839	2019	7	PG-13	123 min	Action, Adve	Carol Danvers becomes one of the universe's most powerful heroes when Earth is caught in the middle of a galactic war between two alien races.
The Hunger (	424668047	2013	7.5	PG-13	146 min	Action, Adve	Katniss Everdeen and Peeta Mellark become targets of the Capitol after their victory in the 74th Hunger Games sparks a rebellion in the Districts of Panem.
Pirates of the	423315812	2006	7.3	PG-13	151 min	Action, Adve	Jack Sparrow races to recover the heart of Davy Jones to avoid enslaving his soul to Jones' service, as other friends and foes seek the heart for their own agenda as well.
The Lion Kin	422783777	1994	8.5		88 min	Animation, A	A Lion cub crown prince is tricked by a treacherous uncle into thinking he caused his father's death and flees into exile in despair, only to learn in adulthood his identity and his responsibilities.
Jurassic Wor	417719760	2018	6.2	PG-13	128 min	Action, Adve	When the island's dormant volcano begins roaring to life, Owen and Claire mount a campaign to rescue the remaining dinosaurs from this extinction-level event.
Toy Story 3	415004880	2010	8.3		103 min	Animation, A	The toys are mistakenly delivered to a day-care center instead of the attic right before Andy leaves for college, and it's up to Woody to convince the other toys that they weren't abandoned and to return home.

5. Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.

   Campos:

Campo	Descripción
Nombre	El título de la película
Recaudación	El total de USD recaudado por la película en el mercado USA
Fecha	El año de estreno de la película
Calificación imdb	Calificación del 1 al 10 brindada por la pagina
Censura	Rating según la MPAA que indica quienes pueden ver la película
Duración	Cuanto tiempo dura la película
Genero	Genero en el cual IMDB clasifico la película (puede pertenecer a más de un género)
Resumen	Un corto resumen de la película

   Periodo de Tiempo:

   El código extrae la información disponible, traerá los datos hasta la fecha en que se corrió.

   Método de recolección:

   El dataset es recolectado mediante la técnica de webscraping utilizando código en Python. El código lee de la página web utilizando la librería beautiful soup y luego guardados en formato csv.

6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay).

   Agradecer a la pagina IMDB y su comunidad por la información recopilada.

   IMDB.com

"IMDb es la fuente de contenido de películas, televisión y celebridades más popular y autorizada del mundo, diseñada para ayudar a los fanáticos a explorar el mundo de las películas y los programas y decidir qué ver.

Nuestra base de datos de búsqueda incluye millones de películas, programas de televisión y entretenimiento y miembros del elenco y el equipo. Te ayudamos a refrescar tu memoria sobre una película, un programa o una persona en la punta de tu lengua, a encontrar la mejor película o programa para ver a continuación, y te permitimos compartir tus conocimientos y opiniones de entretenimiento con la comunidad de fanáticos más grande del mundo."

IMDB inicio en 1990 y desde 1998 es una subsidiaria de Amazon, personalmente la uso para ayudarme a decidir qué película ver y también para investigar sobre actores o directores.

7. Inspiración. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder.

   El dataset está inspirado en mi gusto por el cine y mi interés personal de entender que hace a una película exitosa. El dataset te ayuda a responder que películas fueron las mas taquilleras en USA y que características tenían. Te permite buscar patrones en la lista: ¿cuáles géneros son más taquilleros? ¿Cuál es la duración promedio de una película taquillera? ¿Influye la censura en la recaudación (población más reducida)?

   Mas adelante me gustaría hacer más grande este dataset, más años, y aplicar machine learning para tratar de estimar la recaudación de una película aun no estrenada.

8. Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección: ○ Released Under CC0: Public Domain License ○ Released Under CC BY-NC-SA 4.0 License ○ Released Under CC BY-SA 4.0 License ○ Database released under Open Database License, individual contents under Database Contents License ○ Other (specified above) ○ Unknown License

   Unknown License: no me he es posible utilizar una de las licencias disponibles el dataset es para uso privado. No cuento con los derechos de algunos elementos del dataset como por ejemplo el resumen de la película por lo tanto no puedo tomar atribución sobre estos.

9. Código. Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.

   https://github.com/lcastroba/webScrapingIMDB/blob/master/imdb.py

10. Dataset. Presentar el dataset en formato CSV

https://github.com/lcastroba/webScrapingIMDB/blob/master/imdb.csv