Universidad Costa Rica
Facultad de Ciencias Económicas
Escuela de Estadística

XS-0130 Programación para Estadística II
Proyecto 2

"Análisis descriptivo de filmes mediante una aplicación de Shiny"

Informe

Profesor:
M.Sc. Carlos Francisco Solís Fonseca

Estudiantes:
Fabiola Fang Sánchez | C32851
Catalina Monge Rubí | B94988
Marisol Quesada Madrigal | C26111

II-2025

Explicación y justificación

Dentro de las diversas expresiones artísticas creadas por la humanidad con el paso del tiempo, el cine destaca entre ellas debido a que en cada filme debe darse un delicado balance entre múltiples elementos creativos como fotografía, sonido, escritura y, por supuesto, las artes dramáticas. Cada película representa no solo una obra de arte, sino también un producto comercial sujeto a dinámicas de mercado, en el cual resulta fundamental equilibrar la calidad artística con la viabilidad financiera.

La industria cinematográfica contemporánea, particularmente aquel sector ejecutivo encargado de producir cada largometraje, no solo está interesada en presentarle a la audiencia que visita una sala de cine una historia cautivadora y una puesta en escena deslumbrante, también existe el objetivo de generar riqueza a partir de la película realizada para cada una las partes involucradas. Algunos autores como De Vany & Walls (1999) señalan que el éxito de un producto cinematográfico no puede atribuirse a una serie de factores causales individuales, por el contrario, plantean que obtener ingresos que puedan ser catalogados como favorables tras la venta de entradas en los cines dependerá solamente de la relación impredecible entre la obra y su público. Esta complejidad ha motivado múltiples aproximaciones teóricas y metodológicas para identificar variables que permitan comprender el desempeño de una película.

Entre los predictores más recurrentes en la literatura se encuentran el presupuesto (Galvão & Henriques, 2018; García del Barrio & Zarco, 2016), el género cinematográfico (Karniouchina, 2011; Zhang et al., 2009), la recepción crítica y de audiencia (Hennig‐Thurau et al., 2007; Quader et al., 2017), y factores contextuales como el país de origen y el idioma (Zhang et al., 2009). Asimismo, métricas derivadas de la popularidad redes sociales y del llamado word of mouth han ganado relevancia como indicadores de resonancia cultural previa al estreno (Ting Liu et al., 2016; Ding et al., 2017).

Frente a esta diversidad de enfoques que suelen privilegiar modelos explicativos o predictivos, el presente estudio adopta una perspectiva descriptiva y exploratoria. Su objetivo no es establecer relaciones causales ni predecir resultados, sino caracterizar de manera sistemática un conjunto de variables clave en una muestra significativa de películas, con el fin de identificar patrones, distribuciones y tendencias que enriquezcan la comprensión del ecosistema cinematográfico actual.

Descripción, origen y análisis del conjunto de datos

Para ello, se ha construido una base de datos integrada a partir de fuentes públicas especializadas (TMDB y Kaggle), que permite observar de manera interactiva a través de una aplicación desarrollada en Shiny para Python cómo se relacionan variables como presupuesto, ingresos, género, idioma y valoración del público. Este abordaje busca ofrecer una mirada panorámica y accesible, que sirva como punto de partida para análisis posteriores y como herramienta de divulgación para cinéfilos, estudiantes y profesionales del sector.

Los datos utilizados en este estudio provienen de dos fuentes principales: el dataset Full TMDB Movies Dataset 2024 (1M Movies) (Asaniczka, 2025) y el dataset Top 10000 Popular Movies (Borikar, 2021), y aplicando una unión interna para conservar únicamente los registros presentes en ambos conjuntos. Posteriormente se seleccionó un subconjunto relevante de variables relacionadas con características técnicas, desempeño comercial y recepción crítica de las producciones:

- ***title***: Título original de la película
- ***vote_average***: Rating promedio
- ***vote_count***: Número de votos
- ***revenue***: Ingresos generados
- ***budget***: Presupuesto
- ***popularity***: Índice de popularidad
- ***runtime***: Duración en minutos
- ***genre***: Género cinematográfico
- ***language***: Idioma original
- ***country***: Países productores

Tras la unión de las dos bases de datos, se obtuvo un conjunto final de 9952 observaciones el cual se trabajará mediante el aplicativo Shiny (Posit PBC, 2024). Para conocer con mayor profundidad la base de datos resultante del proceso de unión de los dos insumos originales, se presenta una tabla con las medidas de posición principales para las variables cuantitativas:

Tabla 1. Análisis descriptivo de las variables numéricas

Variable	Conteo	Media	Desviación	Mínimo	25%	50%	75%	Máximo
vote_average	9952	6.42	1.13	0	5.9	6.51	7.12	10
vote_count	9952	1624.81	2951.10	0	117.0	576.50	1649.25	34495
revenue_mill*	9952	60.94	155.30	0**	0	2.46	51.04	2923.71
budget_mill*	5481	36.34	46.92	0**	7.0	20	45	600
popularity	9952	19.99	25.80	0.6	10.9	15.57	22.55	1175.27
runtime	9845	101.27	24.52	2	90	100	113	400

* ajuste de las variables originales revenue y budget a millones de dólares para facilitar su ajuste
** al las unidades estar en millones de dólares, este 0 no es un absoluto sino un valor significativamente menor al millón de dólares, esto ya que los resultados se encuentran redondeados con dos decimales.

En general, se observa que las puntuaciones promedio de los usuarios se concentran en valores medios, sin grandes extremos y con una varianza entre observaciones reducida; hecho

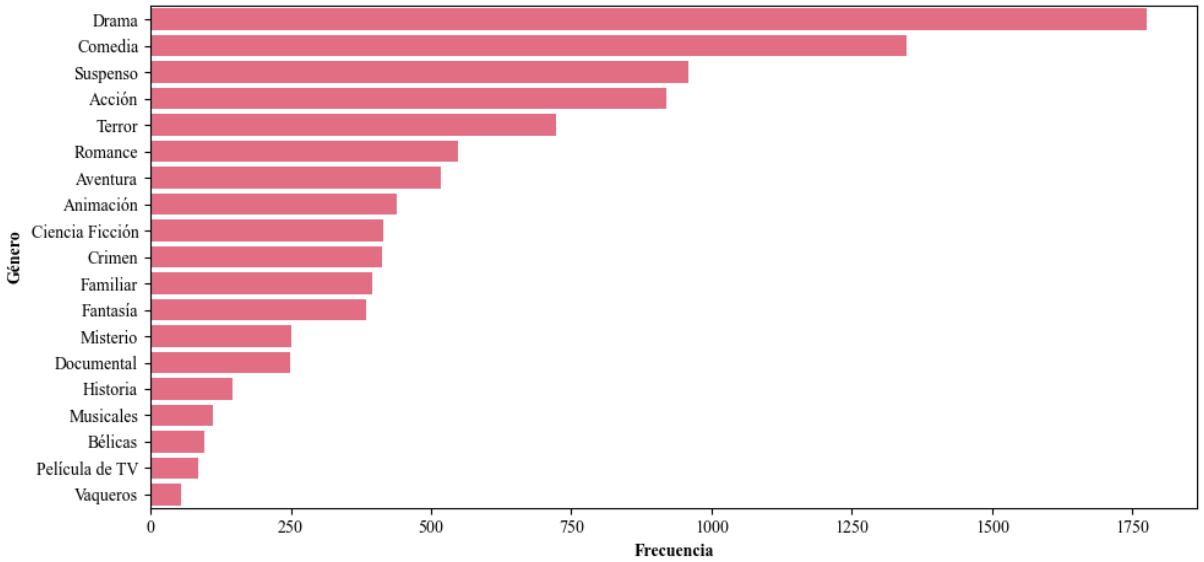
que puede responder a que es una escala previamente definida. Por su parte, la cantidad de votos presenta una alta variabilidad: algunas películas reciben muy pocos votos, mientras que otras alcanzan cifras bastante altas, lo que indica una distribución fuertemente asimétrica.

En cuanto al rendimiento económico, los ingresos en millones muestran una distribución muy dispersa: la mitad de las películas recauda menos de 2.46 millones, pero un grupo reducido alcanza valores altísimos en la taquilla, hecho que explica su desviación alta. Una tendencia similar se obtiene con el presupuesto ya que hay pequeñas pequeñísimas producciones cuyo presupuesto no alcanza el millón de dólares, mientras que se registran largometrajes con un presupuesto de 600 millones de dólares.

La popularidad de cada película también varía considerablemente, en una industria sumamente competitiva resulta razonable que haya películas que logren posicionarse como protagonistas para el público general, mientras que otros largometrajes apenas reciben atención por su audiencia de nicho. Finalmente, la duración se mantiene más estable con la mayoría de las películas situándose entre 90 y 113 minutos, lo cual coincide con la duración típica del cine comercial.

En cuanto a las variables categóricas, para aquellas películas con múltiples valores registrados para alguna de ellas, se implementó un ajuste de selección aleatoria para asignar una sola categoría por observación, con el fin de evitar conteos repetidos y facilitar la visualización del análisis comparativo. La técnica de selección por el azar podría ser refinada con mayor conocimiento sobre los filmes particulares o de la propia industria cinematográfica como tal; no obstante, ante la carencia de un criterio experto se consideró como una alternativa interesante.

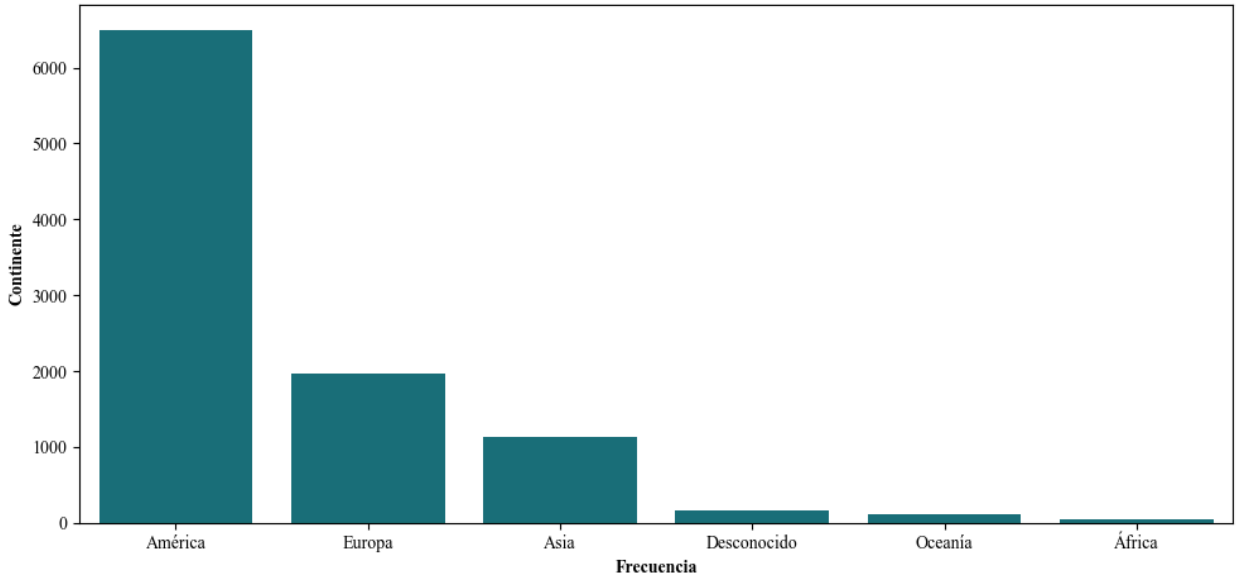
Gráfico 1. Frecuencia del género de las películas



El drama, suspenso y la comedia son los géneros cinematográficos más frecuentes, con dominio del género dramático probablemente por su versatilidad para abordar temáticas variadas para audiencias masivas, así como su predominancia en premiaciones de alto prestigio como los Premios de la Academia (OSCARS); por lo que los estudios se interesen en producir largometrajes con mayor posibilidad de ser reconocidos. En el extremo opuesto, los filmes del nicho de vaqueros, TV y bélicas son poco frecuentes en la muestra ya que abordan temáticas muy puntuales y su audiencia es limitada. En general, se observa una distribución asimétrica donde pocos géneros concentran la mayoría de las producciones.

Ahora bien, ¿de dónde provienen los filmes analizados? Para facilitar la representación gráfica descriptiva inicial, se clasificó cada país en su respectivo continente:

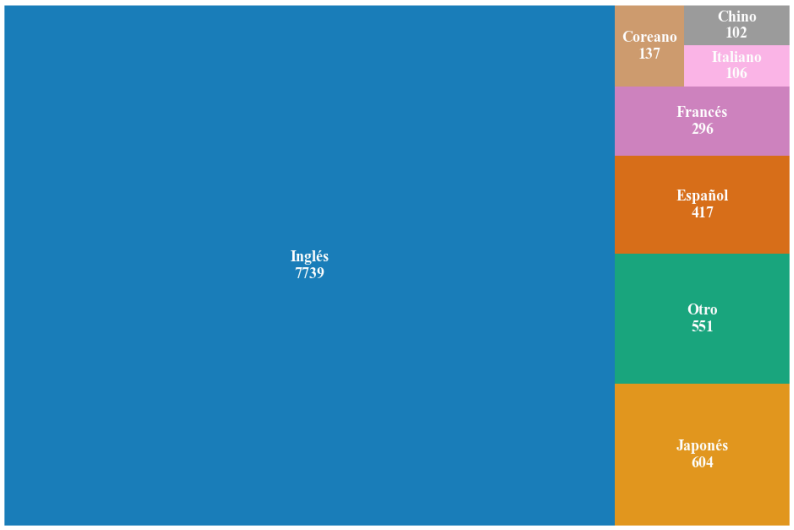
Gráfico 2. Frecuencia de películas por continente



A partir de este, es posible determinar que los largometrajes producidos en países del continente americano lideran significativamente la distribución, lo cuál podría implicar la prominencia y el dominio del cine estadounidense en la industria cinematográfica global, hecho representado en el imaginario mediante *Hollywood*. Europa, si bien se posiciona en segundo lugar, representa menos de la mitad del volumen de América, indicando la brecha pronunciada en el mercado. Esta diferencia se extiende hacia los otros tres continentes quienes logran posicionar una cantidad sumamente limitada de filmes en el mercado, e incluso son mayores las películas de origen “desconocido” que las historias contadas desde África u Oceanía.

Finalmente, también se analizó el comportamiento de las películas según su idioma original. Para ello, previo al abordaje visual se identificaron los 7 lenguajes más frecuentes y el resto de alternativas (50 idiomas adicionales) se agruparon en el nivel “otros”.

Gráfico 3. Frecuencia del idioma original de las películas



Este "treemap" concluye en la misma línea del gráfico 2, existe una concentración importante de la industria cinematográfica desde los países angloparlantes, o incluso, países con otros idiomas optan por realizar sus películas en inglés en aras de apelar a un público mayor. Seguidamente destacan las producciones en japonés, para las cuales se ha logrado configurar una audiencia estable a través de productos audiovisuales, el conjunto de idiomas agrupados y el español. Así, la gran mayoría del cine producido, consumido y, por ende, exitosamente económico son aquellos en inglés.

Objetivos

- Describir sistemáticamente las propiedades, distribuciones y relaciones básicas de las variables seleccionadas en el conjunto de datos cinematográficos integrado
- Desarrollar visualizaciones interactivas que faciliten la comprensión exploratoria de los datos

Preguntas de investigación

- ¿Cómo se distribuyen las variables numéricas clave (presupuesto, revenue, rating) en la muestra?
- ¿Qué géneros, idiomas y países predominan en el conjunto de datos?
- ¿Cómo se distribuye la producción de películas por países y cómo afecta esto a su éxito comercial?

Uso de Git

Este trabajo fue elaborado colaborativamente mediante el sistema Git, puntualmente a través de la plataforma Github en un repositorio público, en el cual la totalidad el equipo figura

como colaboradores, con un sistema de ramas, commits, pushes y pulls para actualizar de manera segura y transparente las versiones estables del proyecto. A partir del siguiente enlace es posible acceder a dicho repositorio, en aras de presentar evidencia en torno a las tareas realizadas: https://github.com/fabiolafang/Proyecto_Shiny

Descripción del grupo y roles

El grupo de trabajo está conformado por Fabiola Fang Sánchez, Catalina Monge Rubí y Marisol Quesada Madrigal, quienes son estudiantes del Bachillerato en Estadística de la Universidad de Costa Rica. Para la creación del producto programado final y el presente informe, no se optó por la asignación de roles y tareas específicas para alguna de las integrantes, sino que el trabajo fue resuelto de forma equitativa según las necesidades puntuales del proceso en un determinado momento.

Breve conclusión

La principal conclusión que deriva de este análisis descriptivo es el hecho de que existe un claro monopolio por parte de la industria cinematográfica estadounidense. No solo es el país del mundo donde, con amplia diferencia, se producen la gran mayoría de largometrajes, sino que también sostienen un amplio dominio sobre en qué idioma deben contarse las historias. Si bien el cine, como forma de arte, puede existir en cualquier rincón del planeta, como producto mercantil está claramente confinado a cierta localidad. De esta forma, *Hollywood*, sus ideas y propuestas, se convierten en la cotidianidad y la verdad absoluta, en claro detrimento de la diversidad de la experiencia humana.

En cuanto a la relación entre las variables exploradas, el scatterplot en el aplicativo shiny deja entrever una débil relación lineal positiva entre la ganancia de las películas y cuánto presupuesto estas poseen. No obstante, gracias al tamaño de los puntos que representan qué tan popular es el filme, es sencillo observar que una taquilla más abultada no necesariamente se traduce en un mayor impacto en el público general, puesto que se evidencian películas populares que finalmente no tuvieron una recaudación exorbitante.

El mapa elaborado con el conteo de las películas por país que las produjo reafirma el dominio estadounidense. Incluso, fue necesario truncar la escala del conteo de las películas con la finalidad de representar una mayor variación entre los países del mundo: al establecer un máximo como 100, se pudo observar más claramente la producción de otras naciones frente al abrumante control de Estados Unidos, quienes se adjudican más de 3800 largometrajes. Asimismo, se evidenció como en el continente africano se obtienen los mayores vacíos.

Sobre los histogramas de las variables numéricas en torno al género de los filmes, fue posible analizar que todos los géneros siguen un promedio más o menos normal en cuanto a su valoración, con leves excepciones más marcadas para las de tipo bélicas y las animaciones. Asimismo, se demuestra que la mayoría de las películas recaudan entre 100 y 400 millones,

siendo aquellas de tipo ciencia ficción, fantasía y animación donde se pueden dar valores más altos; esto debido a tener un público más numeroso. Finalmente, el presupuesto de la mayoría de las películas se agrupa entre 0 y 100 millones de dólares, con picos más elevados en casos muy puntuales que no pueden tomarse como referencia válida para su género.

Referencias bibliográficas

Asaniczka. (2025). TMDB Movies Dataset 2024 (1M Movies) (2025) [Data set]. Kaggle. https://www.kaggle.com/datasets/asaniczka/tmdb-movies-dataset-2023-930k-movies

Borikar, Omkar. (2021). Top 10000 Popular Movies [Data set]. Kaggle. https://www.kaggle.com/datasets/omkarborikar/top-10000-popular-movies

De Vany, Arthur., & Walls, W. David. (1999). Uncertainty in the Movie Industry: Does Star Power Reducethe Terror of the Box Office?Journal of Cultural Economics,23(4), 285‐318.https://doi.org/10.1023/a:1007608125988

Ding, Chao., Cheng, Hsing Kenneth., Duan, Yang., & Jin, Yong. (2017). The power of the «like» button: Theimpact of social media on box office.Decision Support Systems,94, 77‐84.https://doi.org/10.1016/j.dss.2016.11.002

Galvão, Marta., & Henriques, Roberto. (2018). Forecasting Movie Box Office Profitability.Journal ofInformation Systems Engineering & Management,3(3).https://doi.org/10.20897/jisem/2658

García del Barrio, Pedro., & Zarco, Hugo. (2016). Do movie contents influence box‐office revenues?AppliedEconomics,49(17), 1679‐1688.https://doi.org/10.1080/00036846.2016.1223828

Hennig‐Thurau, Thorsten., Houston, Mark B., & Walsh, Gianfranco. (2007). Determinants of motion picturebox office and profitability: an interrelationship approach.Review of Managerial Science,1(1), 65‐92.https://doi.org/10.1007/s11846‐007‐0003‐9

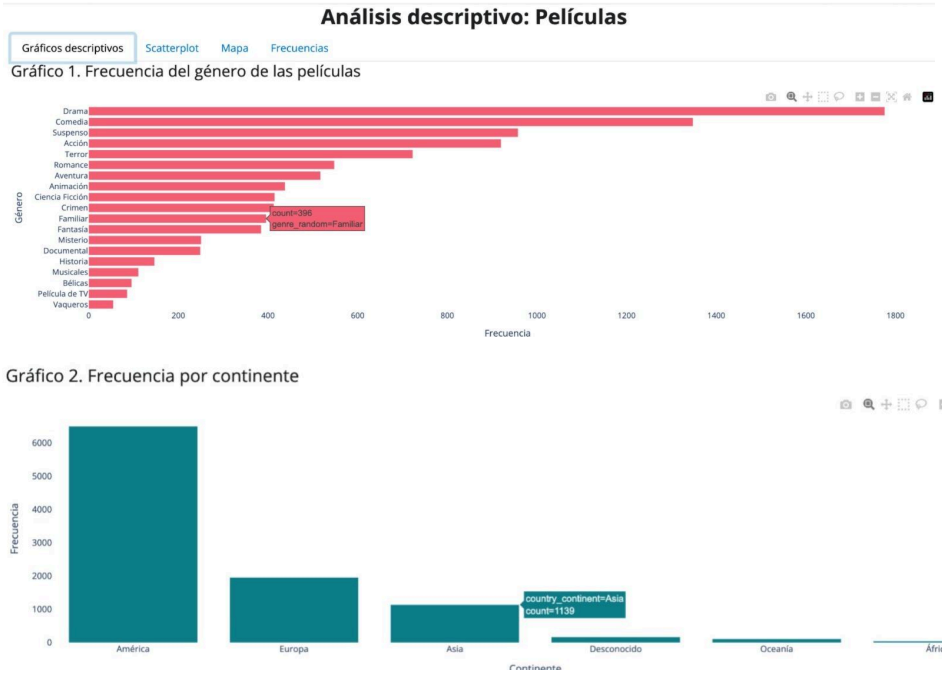
Karniouchina, Ekaterina V. (2011). Impact of star and movie buzz on motion picture distribution and boxoffice revenue.International Journal of Research in Marketing,28(1), 62‐74.https://doi.org/10.1016/j.ijresmar.2010.10.001

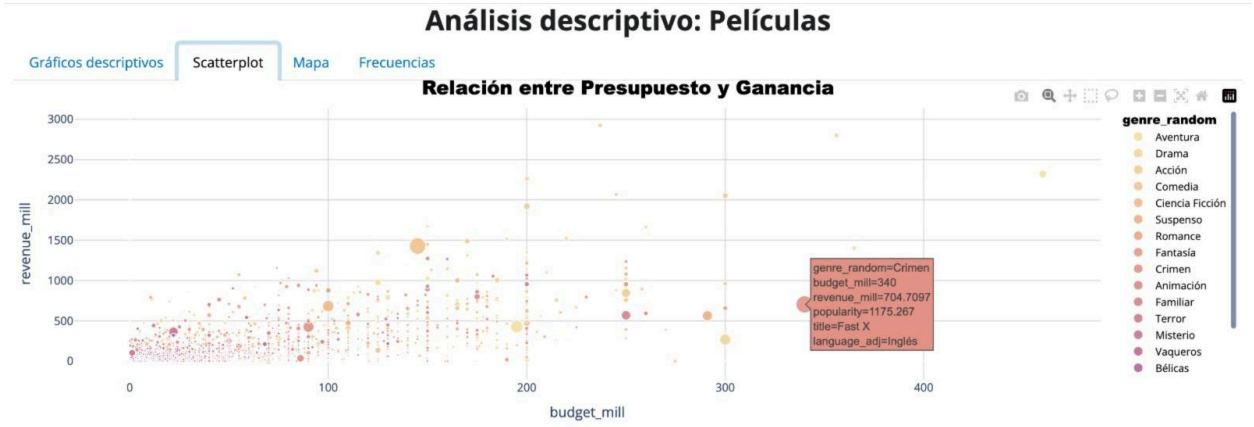
Liu, Ting., Ding, Xiao., Chen, Yiheng., Chen, Haochen., & Guo, Maosheng. (2016). Predicting movie Box‐office revenues by exploiting large‐scale social media content.Multimedia Tools and Applications,75(3), 1509‐1528.https://doi.org/10.1007/s11042‐014‐2270‐1

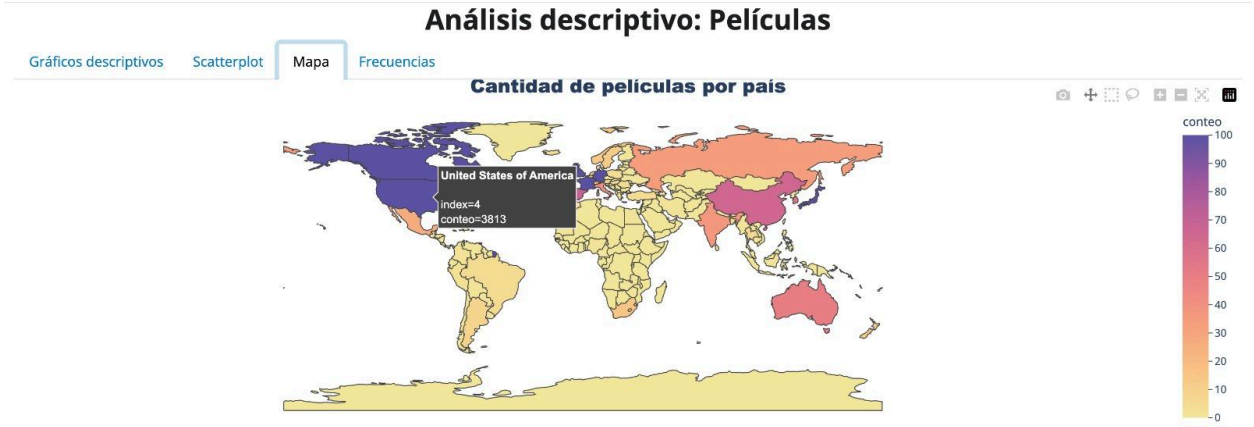
Zhang, Li., Luo, Jianhua., & Yang, Suying. (2009). Forecasting box office revenue of movies with BP neuralnetwork.Expert Systems with Applications,36(3), 6580‐6587.https://doi.org/10.1016/j.eswa.2008.07.064

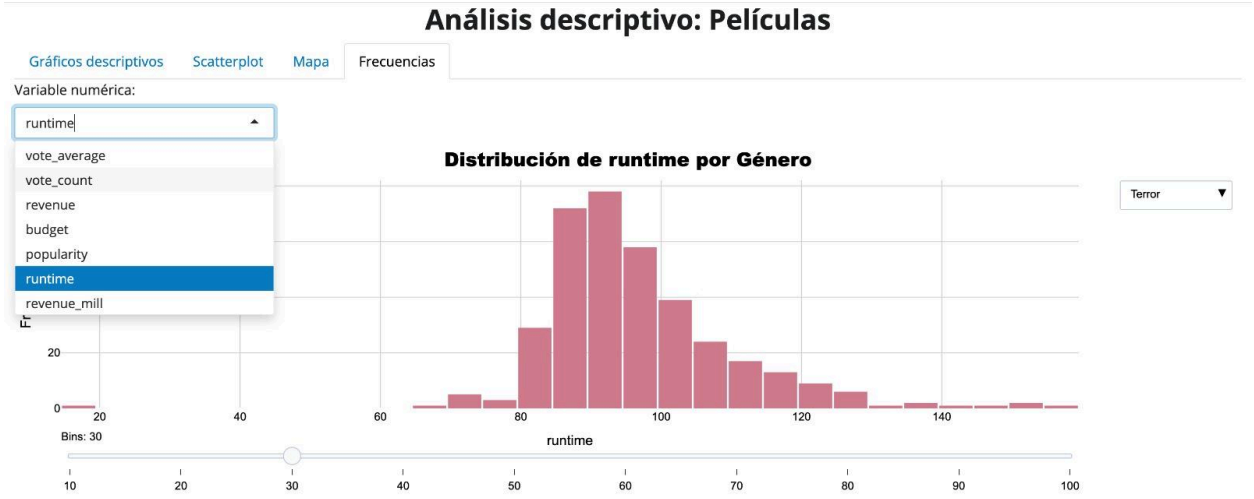
Anexos

Anexo 1. Capturas de pantalla del shiny en funcionamiento






Análisis descriptivo: Películas
Gráficos descriptivos
Scatterplot
Mapa
Frecuencias
Cantidad de películas por país
United States of America
index=4
conteo=3813
conteo
100
90
80
70
60
50
40
30
20
10
0


Análisis descriptivo: Películas
Gráficos descriptivos
Scatterplot
Mapa
Frecuencias
Variable numérica:
runtime
vote_average
vote_count
revenue
budget
popularity
runtime
revenue_mill
Distribución de runtime por Género
Terror
Fr
20
0
20
40
60
80
100
120
140
Bins: 30
runtime
10
20
30
40
50
60
70
80
90
100