Data Analytics Bootcamp by EDEX.org

Chicago: Movies in the Park

Project 3

Renton, William

Rodriguez, Fatima

Schutz, Samantha

White, Henry

DATA-PT-EAST-APRIL-041524-MTTH

Instructor: Booth, Alexander

August 20th , 2024

Table of contents

Introduction

In an era where data drives decision-making and enhances our understanding of urban life, the ability to analyze and visualize information has become crucial. This project focuses on a fascinating dataset known as "Movies in the Park," which captures the essence of community engagement in the city of Chicago through outdoor film screenings. Our objective was to leverage this dataset to uncover insights and present them in a meaningful way.

To achieve this, we took a comprehensive approach. The initial phase involved selecting the dataset and preparing it for detailed analysis. This required meticulous cleaning to ensure the data's accuracy and consistency. Following this, we transformed the dataset into a structured SQL database, which facilitated efficient querying and analysis.

Our exploration included formulating and addressing specific research questions that aimed to reveal patterns and trends within the data. To make our findings accessible and engaging, we developed an interactive web application using Flask. This application features dynamic visualizations that allow users to explore the dataset in a user-friendly manner.

Data Set

The data set was retrieved on August 1st, 2024, from Kaggle.com, the name was: Chicago Parks: Movies in the Park 2014-2019. The dataset includes a list of all "Movies in the Parks" events. Each year, the Parks District does a one-time upload of the year's movie screenings across the city. Since the file is a one-time upload, it does not include information about cancellations or updates. Rather a list of the originally-planned schedule. All movie screenings begin at dusk. Estimate 8:30 from June-July 15, 8:15 from July 15 - August 15 and 8:00 after August 15. This data base contains 1,406 rows and the columns are listed as follows:

1. Title
2. park
3. date
4. phone
5. rating
6. cc
7. location.latitude
8. location.longitude
9. location.human_address2
10. datayear
11. day
12. park_address
13. Address
14. city
15. state
16. zip
17. community
18. geocode_address
19. lat
20. long
21. osm_type
22. display_name
23. class
24. type

Data Cleaning/Data base creation

After selecting the data set, we proceed to download the csv file and use jupyter notebooks to perform the data cleaning. The first step was to import the csv file and explore the shape of the data looking for null values and select and evaluate the columns that we will keep for this analysis.

It was recognized that in order to create the choropleth layer for the map, the community names in our database would need to match exactly with the community names of the geoJSON used to draw the neighborhood boundaries. An initial comparison of the values of the community column in the data frame with a list extracted from the geoJSON showed that only about a third of the community names matched. Several issues were identified: some community names in the data frame had leading or trailing space characters, some community names included multiple comma-separated names, and some referenced names that were not part of the 77 official community names (such as police precincts). This was addressed by using the .strip() method to remove white space and manual inspection and reassignment of the comma-separated names based on a check of which neighborhood the latitude and longitude correlated to on google maps. Rows that still did not match after this process were dropped, along with rows containing nulls. 1090 rows remained following this cleanup. Unnecessary columns were also dropped from the dataset, and an SQLite database was created.

Color design considerations

For our website, we utilized the Lux template, renowned for its neutral background and structured layout that effectively showcases our visualizations. Our data visualizations consist of a bar chart, a sunburst diagram, and a choropleth map. To ensure clarity and visual appeal, we selected a color gradient ranging from teal to green for these elements, complementing the overarching theme of our presentation. In crafting our presentation materials, we chose to highlight the city of Chicago through a distinct use of color, specifically red, which is one of the city's most characteristic hues.

Website architecture

Dashboard

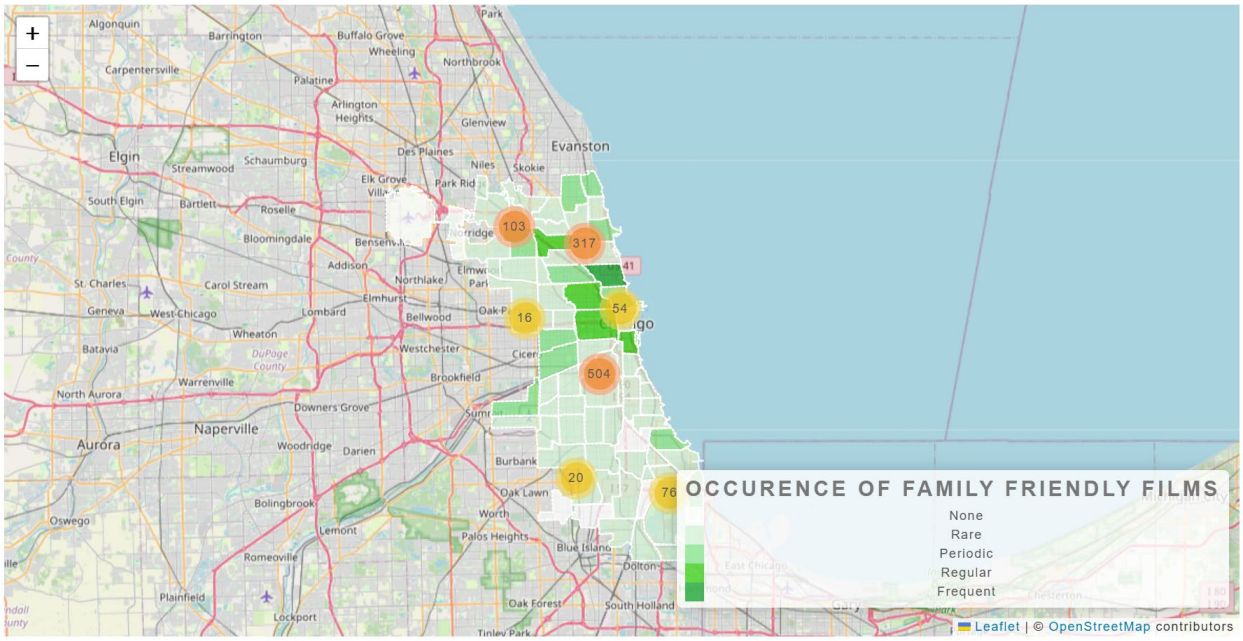
The dashboard calls a query using the flask API that returns films filtered by month of their showing. This data is used to construct both a bar chart and a sunburst chart. Given the sheer volume of different movie titles, two considerations improved the usability of this page. First, movies with only one showing were filtered out of the bar chart display, to prevent overcrowding. Second, plotly was used to allow popups with metadata of the movie title to appear when the user interacted with the chart.

Earlier iterations of the page allowed for filtering based on year, but in an effort to link the two data visualizations to a single query, the parameters were changed to allow for filtering based on month.

Map

Three files were used to create the map page: an html file to create the structure, a css file to provide the styling of the map size, position, and legend, and a javascript file to dictate the interactions with the map feature. The map data is queried from SQLlite through the flask app, and returns all columns based on the year filters. For the marker layer, each row of the returned data correlates to a map marker, with a popup that shows the metadata for the title, park, date, rating, and closed captioning of the row. The leaflet marker cluster feature made this data more user friendly.

For the choropleth layer, the geoJSON was used to create the polygons for the layer, each identified by the community name. When the map is created, an empty dictionary is initialized to store the counts of PG and G films showing by community based on the bounds of the query. For each film in the returned data from the flask API, if the film ratings is G or PG, the total count of the associated community in the dictionary is increased by one. Once all the data is looped through, the community counts are used in conjunction with a color function to determine the shading of each layer. The shading is relative to the max value in the dictionary.



Additional Pages

Three additional pages were created for the web app: the home page (landing page), the about us page, and the works cited page. Each is a static page only requiring an html file to function. The photo of the landing page was created in Canva. The about us page pulls images from each group members' public LinkedIn profile and links to the same. Links are included on the works cited page.

DASHBOARD MAP ABOUT US SOURCES

CHICAGO: MOVIES IN THE PARK

Movies in the Park offer a wonderful opportunity for families to enjoy something different, have fun, explore new experiences, and socialize with their community and friends. We believe this initiative fosters family togetherness and strengthens the community

DASHBOARD MAP ABOUT US SOURCES

ABOUT US

We are a dedicated team of four passionate individuals who have embarked on an exciting journey to bring the story telling of outdoor movie screenings in Chicago's parks. Our project began with a comprehensive dataset of movies shown in city parks from 2014 to 2019. With this valuable information in hand, we set out to transform it into something truly special.

Thank you for visiting our website. We hope you find our visualizations both helpful and inspiring as you plan your next movie night in the park!

Warm regards,


FATIMA RODRIGUEZ


HENRY WHITE


SAMANTHA SCHUTZ


WILLIAM GRAY RENTON

DASHBOARD MAP ABOUT US SOURCES

SOURCES

Levy, Jonathan. Chicago Parks: Movies in the Park 2014-2019. Kaggle, 2021, www.kaggle.com/datasets/abrambeyer/chicago-parks-movies-in-the-park-20142019.
Plotly JavaScript Open Source Graphing Library. plotly.com/javascript/. Accessed Aug. 2024.
Bootswatch. Free themes for Bootstrap. Accessed Aug. 2024.
Title page background: openair-cinema-vector-illustration-people-watching-movie-in-night-vector-id1202503070 (612×408) (istockphoto.com)
Images: https://backiee.com/static/wallpapers/1920x1080/206919.jpg
Images 2: question_mark_PNG128.png (976×580) (pngimg.com)
Images 3: blank-clapper-board-png.png (1920×1920) (vecteezy.com)
Images 4: film-reel-png-roddes-movie-and-music-reviews-5672.png (5672×1080) (pluspng.com)

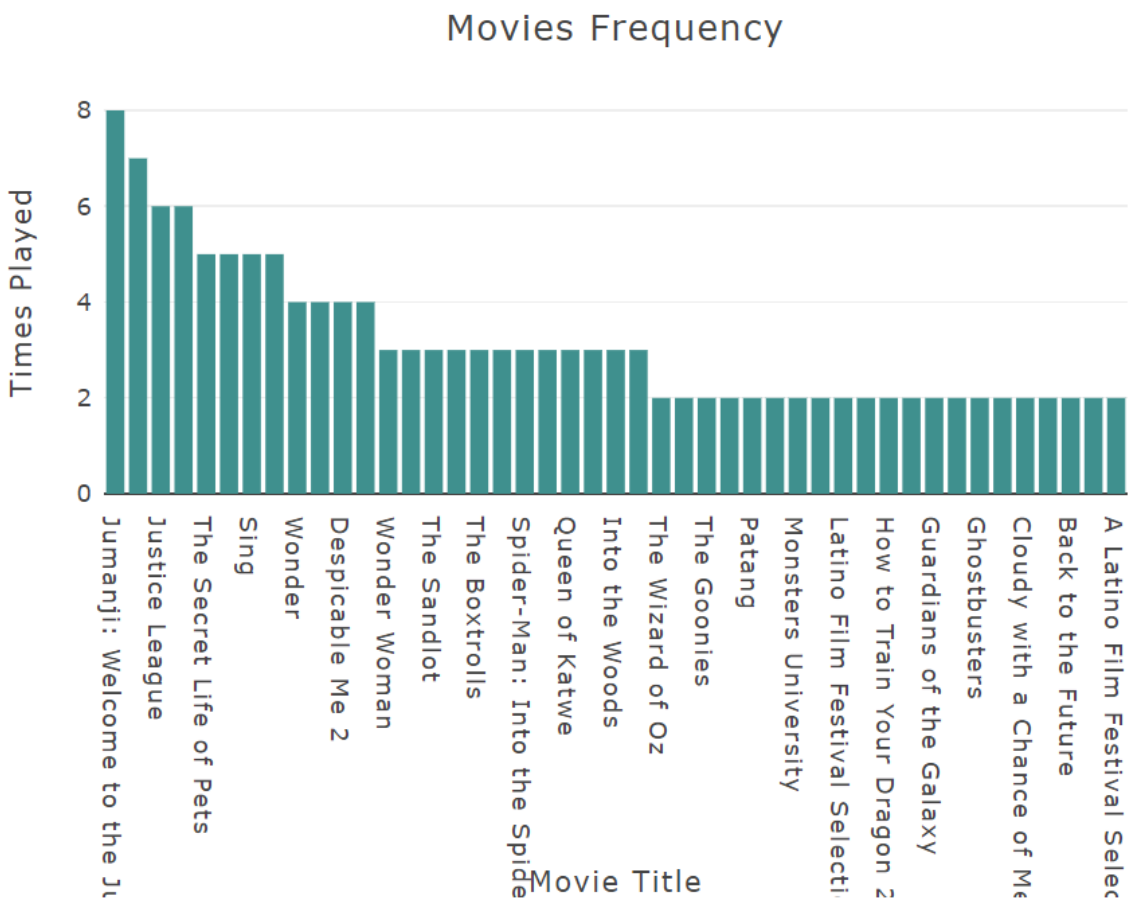
Dashboard design concepts (How does the dashboard answer our research questions)

After reviewing our data, we came up with three major questions:

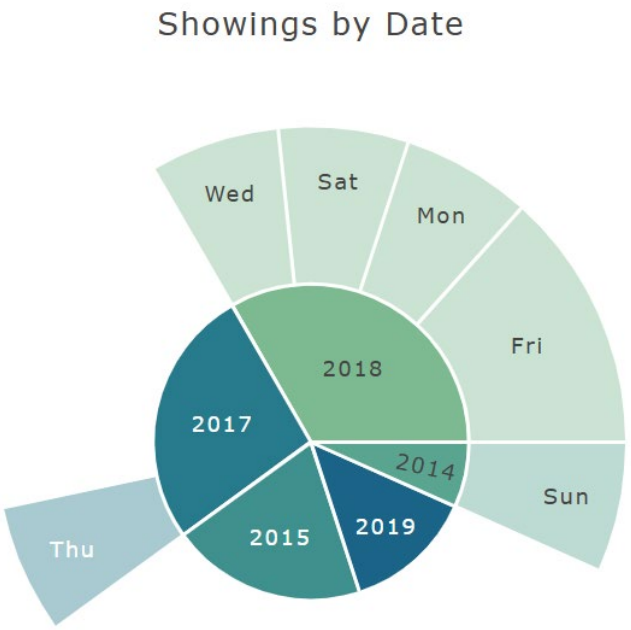
1. What are the frequencies of the most frequently shown movies from 2014 to 2019? How do these frequencies change by month?
2. What are the best days of the week to catch a movie throughout the year?
3. Which Park has shown the most movies? Are there locations that show more movies than others, and are they situated in specific communities? Which communities are the best for family-friendly films?

The first consideration was to create a filter that would allow users to select the month of the year they wish to explore. Once a month is selected, this filter will query the database and retrieve the relevant data for both the bar chart and the sunburst diagram.

The bar chart will display the frequency, or number of times, a particular film was shown during the chosen month. On the X-axis, users can locate the movie titles, and by hovering over each bar, they will be able to see the exact number of showings as well as the movie name. The user can use this filter and chart to answer not only the most popular movie, but which month had which movie shown most often.

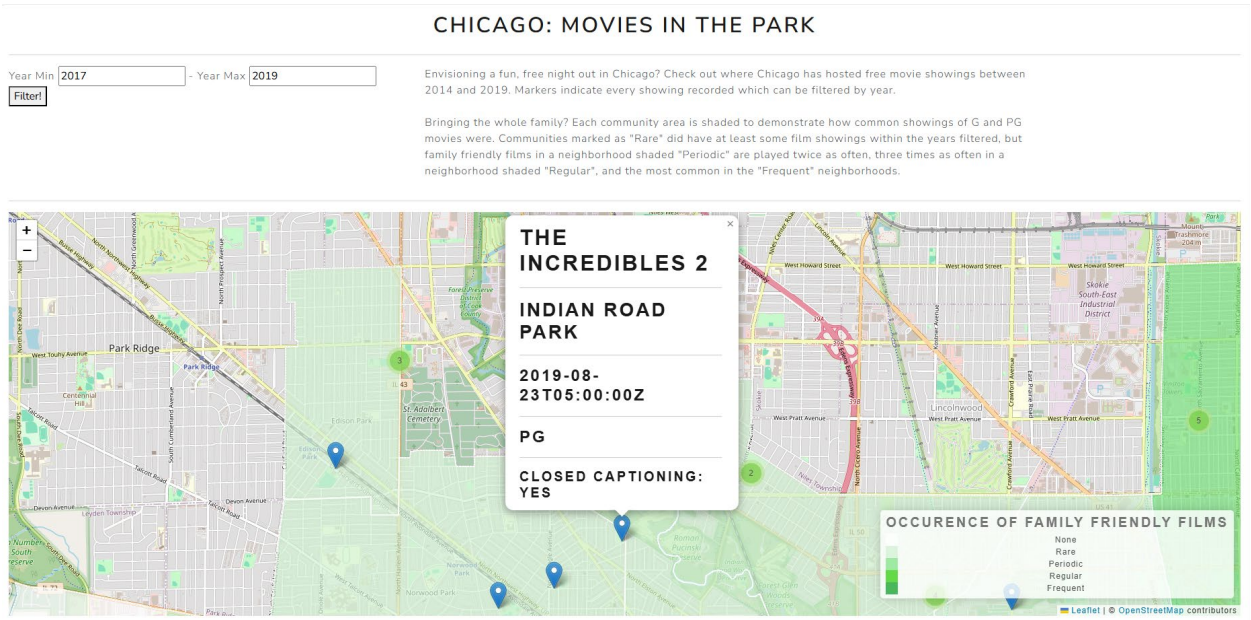


Once users select a month, the sunburst diagram will adjust to display the showings for each day of the week within that month. Users can further refine their exploration by selecting a specific year from the range of 2014 to 2019 (note that data for 2016 is missing from the dataset). By hovering over a particular day of the week within the sunburst diagram, users will be able to view the number of showings for that day, month, and year. This functionality directly addresses our second question by providing detailed insights into the distribution of movie showings across different days of the week.



For our third and final question, we chose to use a choropleth map to visually represent the frequency of family-friendly films shown in different communities. The map employs a color gradient to highlight communities with the highest frequency of such films. Users will first encounter a filter that allows them to select a minimum and maximum year for exploration.

The map's boundaries delineate each community, and as users zoom in, markers will appear on the map. These markers provide detailed information about the films shown, including the movie title, the name of the park where the film was displayed, the date and time of the screening, the movie's rating, and whether closed captions were available. This approach enables users to gain comprehensive insights into which communities are most active in offering family-friendly films and helps address our third question effectively.



Bias and Limitations

- Null Values: A high frequency of null values can skew analysis results, reduce the quality of insights, and complicate decision-making processes.
- Years Provided 2014-2019, Missing 2016 (No Dates in Data), Data is available for the years 2014 through 2019 but is missing for 2016, it creates a gap in the dataset. This missing year can affect time-series analysis, trend evaluations, and comparisons over time.
- Unrelated Parks Grouped Together: When unrelated parks are grouped together in data or visualizations, it can lead to confusion and misinterpretation. Parks might be grouped based on naming conventions, geographical proximity, or data entry errors, but this aggregation can mask meaningful differences and trends. For instance, Lincoln Park in Chicago should not be conflated with similarly named parks in other cities if they serve different communities or have different characteristics. Proper categorization and differentiation are necessary to ensure accurate analysis and reporting.

- Movies That Did Not Have a Title: Movies without titles in a dataset represent a significant issue for data integrity and usability. Titles are essential for identifying and referencing films, and their absence can hinder the ability to perform analyses, generate reports, or make comparisons. This issue might arise from incomplete data entry, data extraction problems, or errors in the database.

Conclusions

Friday: Most Frequent Day of the Week, Fridays are often the most popular day for events, gatherings, and leisure activities for several reasons. The end of the workweek typically signifies a transition into relaxation and personal time

The Lego Movie: Most Frequently Shown, "The Lego Movie," released in 2014, has seen a significant amount of airtime and showings due to its broad appeal and popularity. The film's combination of humor, creativity, and its appeal to both children and adults has made it a staple in family entertainment. Its frequent showings could be due to its success in theaters, its positive critical reception, or its appeal as a go-to family film for repeated viewings.

Lincoln Park: Most Common Park, Chicago, Lincoln Park is a large, well-known public park offering various recreational activities, cultural institutions, and green spaces.

2014 Had the Most Showings in the Year, If 2014 had the most showings of a particular film, event, or program, it suggests that year was notable for its high frequency or popularity. This could be attributed to a number of factors, including the film's success, special anniversaries, or a particularly strong interest in that year's events. It might also indicate that the film or event had a significant cultural impact or was part of a broader trend that made 2014 a standout year.

Family-Friendly Movies with a Few Films Rated for Adults Only: Family-friendly movies are designed to appeal to a broad audience, including children and adults, often featuring content that is appropriate for all ages. These films typically focus on positive messages, humor, and adventure suitable for younger viewers.

Future work

- Convert Months to Actual String Versions: To enhance the readability of our visualizations, we will convert numerical representations of months into their corresponding string names (e.g., January, February, March) on the images. This conversion will make the data more accessible and intuitive for users, allowing them to quickly interpret the time periods represented in the charts and diagrams.
- Compare Pre- and Post-COVID Showings: We will analyze and compare movie showings before and after the onset of the COVID-19 pandemic to identify any significant changes in trends. This comparison will involve examining data from the years prior to the pandemic (e.g., 2014-2019) against data from the pandemic period (e.g., 2020 onward). Key aspects of this analysis will include shifts in the frequency of movie showings, changes in popular movie genres, and variations in the number of screenings held in different parks or communities.
- Link IMDB to Movie Release Timeliness: To assess how long it takes for movies to be shown in parks after their release, we will integrate IMDB links into our data visualizations. These links will direct users to IMDB pages where they can find information about the release dates of movies. By comparing these release dates with the dates when the movies were shown in parks, users will be able to gauge the average time lag between a movie's theatrical release and its appearance in local screenings.

Works Cited

1. Levy, J. (2021). *Chicago parks: Movies in the park 2014-2019*. Kaggle. Retrieved from https://www.kaggle.com/datasets/abrambeyer/chicago-parks-movies-in-the-park-20142019
2. Plotly. (n.d.). *Plotly JavaScript open source graphing library*. Retrieved August 2024, from https://plotly.com/javascript/
3. Bootswatch. (n.d.). *Free themes for Bootstrap*. Retrieved August 2024, from https://bootswatch.com/
4. iStock. (n.d.). *Openair cinema vector illustration people watching movie in night* [Image]. Retrieved from https://www.istockphoto.com/photo/openair-cinema-vector-illustration-people-watching-movie-in-night-gm1202503070-171413821
5. Backiee. (n.d.). *1920x1080 wallpaper*. Retrieved from https://backiee.com/static/wallpapers/1920x1080/206919.jpg
6. PNGimg. (n.d.). *Question mark PNG*. Retrieved from https://pngimg.com/image/128
7. Vecteezy. (n.d.). *Blank clapper board PNG*. Retrieved from https://www.vecteezy.com/png/1920x1920/blank-clapper-board-png
8. Pluspng. (n.d.). *Film reel PNG*. Retrieved from https://www.pluspng.com/png/film-reel-png-roddes-movie-and-music-reviews-5672.png