**Mayhem at DinoFunWorld - Visitor Movement Analysis**

Ian O’Keeffe

Department of Computer Science, University of Denver

COMP 4449 - Data Science Capstone

Professor Claudio Delrieux

August 14, 2023

**Introduction**

In 2013, Disney released the most expensive project they had ever built at Walt Disney World in Orlando, Florida. It wasn’t a new roller coaster, or a new area of a theme park. Disney spent $2 billion to build a system to track guest movement and activity, called “MyMagic+”. For context, the two most recently built US Disney parks at that time were Disney’s Animal Kingdom (1998) and Disney’s California Adventure (2001), costing the company $1 billion and $600 million, respectively.

Disney is the undisputed industry leader in theme parks: 7 of the top 10 attended theme parks in the world are Disney parks, and their total annual attendance across parks is 2.5 times higher than the #2 company, Merlin Entertainment. It’s fair to say they knew the importance of granular guest data. Although the MyMagic+ project also included upgrades to the Disney World Resort’s hotel, dining, and ticketing systems, the primary purpose was tracking guest movement by issuing each guest a “Magic Band,” an RFID enabled wristband that was constantly tracking where the guest was and what they were doing.

This project investigates a similar system at the fictional “DinoFunWorld” park. In this fictional scenario, an olympic gold medalist was doing a special event at the park, when his memorabilia was vandalized and his gold medal stolen. The primary purpose of this project is to analyze visitor movement behavior, to find groups of visitor types, discern trends across the weekend, and find anomalies in the data that may be related to the crime.

Having a good grasp on guest movement is a huge boon to a theme park. Which areas are frequented and which aren’t can help determine which rides need refreshing or demolishing, whether a blank expansion area would be a good place to build a refreshment stand, and general user flow.

**Data Exploration**

The data given included guest ID, a timestamp for when the event was tracked, an X coordinate, a Y coordinate, and the event type, either “movement” or “check-in,” separated into three separate CSVs for Friday, Saturday, and Sunday. The data was clean, there was one missing row on Sunday that was dropped. There were a total of 26,021,962 rows across the three days, making this a large dataset. The total number of unique ids across the weekend was 11,374. Sunday was the busiest day of the weekend, with a bit more than twice as many unique IDs as were present Friday. 7,024 visitors stayed for one day only, 2,537 spent two days at the park, and 1,813 spent all three days of the weekend at the park.

The vast majority of events were movement events. 328,838 events were checkins, and 25,693,123 were movement events, 78 times as many. Although the park map has 68 locations listed, only 42 of those were ever checked into. The same 42 locations were checked into each day, indicating no failure of any sensors. The movement events were more robust, with over 1300 unique pairs of X and Y coordinates listed.

Also provided was a typical park map, with different locations on the map listed as a location number, and a legend saying what the name of that numbered location was. For example, location 16 is the “Kiddie Ride” Stegocycles.

**Data Preprocessing**

The first step to processing the data into a usable form is to convert the X,Y coordinate pairs into a more useful form. A park map was provided giving us a list of attractions. The X and Y values span the range of 0-100. The park map PNG with an included list of locations was converted to a high contrast black and white image, and then the pytesseract package was used to extract the text. This provided a list of location names, with only minimal manual adjusting.

Overlaying the “check-in” events on the map, a new image was produced including the coordinates as well as an indicator on the map where that event happened. The list of location names with numbers and this new image were used to create dictionaries of location numbers keys pointing to coordinates, and location numbers pointing to location names. This was used to build a new dataframe for the locations, consisting of location number, coordinates of the location, name, location type, and whether the location was secondary or not.

The secondary field was an attempt to include the locations which did not have explicit check-in events. After a few analysis attempts, it became clear that this column would end up being useless, so it was dropped and only primary locations were used. These locations included all food and beverage locations, information stands, restrooms, shops, and one of the show buildings.

This location dataframe was then combined with the original to compress the IDs to snapshots of their daily behavior. Each ID / Day combination was given a count of check-in events at each location type, as well as the total time they spent in the park in seconds (computed by subtracting their first registered event from their last registered event), and the time of their Entry (first event, which in nearly all cases was a check in at a park entrance) or Exit (last event).

**Data Analysis**

Now that the data was wrangled into a more usable form, it was possible to start tracking trends. Since our data unfortunately didn’t include check-ins for food, beverages, or stores we’d be looking at the attractions instead, which are classified as being either thrill rides, kiddie rides, rides for everyone, or shows. In a real world setting, we could probably get supplemental sales data for restaurant and store sales.

Thrill rides were the most popular attraction type, despite the fact that there were only 9 thrill rides compared to 11 kiddie rides and 12 rides for everyone. 157,929 total guests checked in to a thrill ride throughout the weekend, 64,679 checked into rides for everyone, 42,527 for kiddie rides, and 24,804 for the shows. The Keimosaurus Big Spin thrill ride was the single most popular attraction, and all 9 of the thrill rides where the most popular 9 attractions. The shows were also very popular, despite low total attendance, each individual show had high check-in counts. These trends were the same across individual days as well, with not much variation in proportion of attractions being checked into, although the overall numbers did increase as the park attendance got higher each day.

The average visitor checked into 25.6 attractions each day. Considering there was a crime that took place this weekend, this would be a good place to begin investigation into who might be suspicious. Although I have no background in criminal justice or psychology, stealing an olympic gold medal doesn’t seem like an impulsive action, so tracking down outliers in the data could help pinpoint abnormal behavior and then possibly the culprit.

With such a large dataset, we could be picky about what was considered suspicious. The guest ids were filtered down to guests who were 4 standard deviations below the average number of check-ins to attractions, which gave us 35 ids. The culprit was probably there to steal a medal, not have a fun day at the park. Some people might not be fans of rides, or possibly season pass holders who had free admission and only stopped by for a little while, so this wasn’t enough.

For all guests, the average time in between check in events was of 9 minutes and 47 seconds. (Incidentally, this shows that DinoFunWorld has very high guest throughput for attractions. Most attractions are between 1.5 and 3 minutes long, and including walking time, guests spent almost no time waiting in line for their next ride.) We took guests who’s average time between checkins was two standard deviations higher than normal, and cross referenced that with the previous list. We came out with two Ids, which was much more manageable.

Finally, our last filtering operation was checking to see if any guests’ first registered activity was movement. In our previous analysis, there were some IDs for whom that was the case. Although there are several possible explanations for this happening, one is that the guest snuck into the park. Sure enough, one of those two ids had a movement event as their first event on Saturday.

Looking at guest ID 657863, there seems to be enough evidence to investigate them as the possible culprit. They spent 5 hours and 40 minutes in the park on Friday, during which they rode no rides, and saw no shows. There were only two singular movement events over the course of the entire day. Additionally, they had one singular event on Saturday, which was a movement event. Since the movement data is tracked by the DinoFunWorld phone app, turning the phone off, turning it on airplane mode, or uninstalling the app would all hide user movement from the system. I would recommend this guest be investigated by authorities.

Moving away from the investigation of the criminal activity, we also tried to categorize guest behavior. We categorized guest ride activity based on exclusion of ride types, as well as guests who only prefer a single ride type. The three largest categories here are guests who never rode a kiddie ride (2591), guests who never saw a show (2050), and guests who never rode rides for everyone (369). There were only minor numbers of guests who exclusively preferred one ride type, with 27 guests only riding thrill rides and 10 only riding rides for everyone. No guest only rode kiddie rides or saw shows.

Guests who don’t ride kiddie rides preferred thrill rides compared to other ride preference groups by a wide margin. They also spent more time in the park than other ride preference groups.

We also analyzed when guests entered the park. Most guests entered before noon, even on Friday (3,456 vs 101 afternoon), showing that very few people are entering the park after school or work. DinoFunWorld doesn’t offer a late entry ticket, unlike many theme parks, where entry after a certain time is discounted, so this might be something worth looking into to bump up those numbers.

The morning entry group spent much more time in the park than afternoon entries, which makes sense logistically. They also had a higher average ride count per hour.

Looking at multi-day vs single day guests, we have categories for Friday only, Saturday only, Sunday only, Friday and Saturday, Saturday and Sunday, and all three days. There was very little variation in visitor activity no matter which kind of guest they were. On a positive note, the largest single group was the three day guests. Not only that, they were in the park for an average of 10-12 hours, in line with the other guest types. This is a very good indicator. DinoFunWorld has a lot of very loyal fans who will go for three days and spend a lot of time there.

Surprisingly, not a single guest went to the park both Friday and Sunday, but not Saturday. Many theme parks are surrounded local tourist industries. Anaheim and Orlando both have thriving hotel, shopping, entertainment, and restaurant areas in the neighborhood of the theme parks. It might be worthwhile for DinoFunWorld to look into developing the surrounding area, as this is good for both the local economy and the park. Guests who stay at one of DinoFunWorld’s two lodging accommodations would be more likely to stay an additional night even if they only go to the park for two days if there are other local activities.

Ride popularity by time stayed relatively consistent from day to day. Thrill rides were the most popular rides across days and times. There tended to be a large spike for riding right as the park opened, and again in the mid-afternoon between 2 and 4pm. The shows had large spikes in attendance at 9:45 and 2 o'clock for Friday and Saturday, most likely indicating the show times. On Sunday, the show had a morning spike, but not one in the afternoon. Our data doesn’t indicate when the theft of the gold medal took place, but it’s possible it was stolen Saturday after the show, so the 2 o'clock olympic show was canceled on Sunday.

Every day showed a similar trend in total number of visitors in the park by time of day. There was a massive spike as the park opened, and then a relatively consistent downward trend in attendance as the day went on. This information is useful for staffing considerations, and also another indicator that there is a missed opportunity for driving more guests into the park in the afternoon.

Looking at the top ten attractions and check-in counts by days, there is again a lot of consistency. The one thing to note here is that the Sabretooth Theater, the 10th most checked into attraction, was only open on Sunday. This feels like a missed opportunity, as there is clearly high demand for the show. Opening the show on other days of the weeks as well could lead to increased park attendance. Whatever the show in this theater is isn’t related to the olympic gold medalist event either, as the supplemental data indicates his signings and meet and greets would be taking place in the Creighton Pavillion.

Several visualizations of visitor movement were created. The first, a simple scatterplot of X,Y coordinates, outlined where the paths were nicely, but didn’t offer much more. The Density heatmap showed two hotspots, one in the central west part of the park, where several paths converge, and another along the south part of the park cutting in to the middle. The bottom area is especially of note, as there is plenty of unused space there and no shopping or food locations. An area with that much foot traffic is a prime spot for sales opportunities.

Looking at movement frequency, there tends to be a lull each day between 3 and four o’clock before guests start moving again. There’s not a particularly high volume of show check ins at that time, and it’s not a meal time, so it might just be a time when people take a break from the heat.

We did a DBScan clustering analysis to see visitor movement activity clusters. There weren’t any surprises here, with strong clustering around thrill ride areas as well as the kiddie land where the kiddie rides were found.

**Model Building and Evaluation**

Without a clear output variable, it would seem the best way to classify our data would be to use clustering techniques. The data was fit and transformed with a PCA (Principal Component Analysis) model. This would reduce the number of dimensions while preserving as much of the variance as possible. An elbow plot was also created as a step in building out a K-nearest neighbors model, which indicated 3 as the optimal number of clusters. The K-nearest neighbors algorithm measures distance between points in the feature space, choose a number of neighbors (3 in our case), and classifies each point with a majority voting process according to what its K neighbors are classified as. The K-nearest neighbors clustering visualization unfortunately did not seem to give us clear cluster relationships for our dataset.

The second clustering model we built was agglomerative clustering. This is a hierarchical clustering algorithm used to group similar data points into clusters. The algorithm starts by considering each data point as an individual cluster and then merges clusters together based on their similarity until all data points belong to a single cluster. Once the agglomerative clustering model was built, we then used it to create a dendogram to separate.

Finally, we built out a DBScan model which we then overlaid onto the park map to find clusters. We visualized this with plotly express, and can see a few interesting patterns. For example, there are clusters in the Entry corridors between shopping locations, showing that guests don’t necessarily just stop at one store for souvenirs - they go to several.

**Notable Observations**

The park does not require a check-in event for guest exit of the park. This is a problem for a few reasons. First off, it means exit time for guests is fuzzy. We can only track the last move in or check in event of the guest, and these are often not at park exits. Additionally, there’s a possibility that some guests could be staying in the park after hours, which is a security risk.

The park has three entrances. The entry corridor is built out similarly to many other theme parks, where upon entering the park, guests move forward and see a variety of shops and restaurants. The entrances on the east and west side are not used as frequently as the entry corridor entrance, but they are still used. Funneling guests through the primary entrance for both entrances and exits would lead to more sales in both merchandise and food / beverage.

Thrill rides are by far the most popular kind of attraction at the park. Although this dataset doesn’t include any long term data, it is fair to say that the primary draw to the park is its selection of thrill rides. Comparatively, nearly half of all visitors skip kiddie rides entirely. In the future, if park attendance begins to wane, or if they would like attendance bumps, a large thrill ride seems the way to go.

**Recommendations**

Despite the health of the park, there are definitely areas that could be improved. Tracking check-ins at shops and restaurants seems like an easy implementation. Although there would be purchasing data available, that won’t paint the full picture. A guest could stop in a store, look around, and not purchase anything. This is an important number to have access to, as it is a major factor in gauging guest interest for the merchandise and food options.

Showing guests shops on both the way in and out of the park helps drive sales. If someone sees an item on their way into and considers purchasing it but doesn’t, we are missing a second chance opportunity to sell that item if they are leaving through the side exits. The entry path from the east leads directly into Kiddie Land, which is the both the least popular area of the park. According to park density, the entry corridor is one of the most frequented areas of the park though, so this may not be that pressing of an issue.

The overwhelming majority of parkgoers enter in the morning, right as the park opens. Over the course of the day, there are very few new entries from attendees. This is a real missed opportunity. A discounted afternoon ticket or something similar could drive up park attendance throughout the day.

The Sabretooth theater is the 10th most popular attraction, yet it is only open on Sundays. There is clearly guest demand for more frequent shows there.

There is a wonderful opportunity for building out new food or shopping options in the central-west and south-central parts of the park. They are the areas with the most guest foot traffic, and both are missing any sales options.

**Difficulties / Problems**

The Data set was very large. A few times, while wrangling and merging my dataframes, I had the joins entered incorrectly, which crashed my computer a few times until I was able to get the join corrected.

The park map and numbering system was poorly designed, leading to an unnecessary headache in connecting coordinates to locations. The location numbers aren’t laid out in any meaningful way on the map, meaning there’s no way to find location 42 without playing Where’s Waldo.

Since there were so many movement events, running processes on them was taking an enormous amount of time. I took random samples of visitor IDs to track rather than run analyses on the full set in order to expedite while building and coding the plots and clustering. Even after running these on the full data sets, it turns out the results were extremely similar, leading me to prefer the sampling of movement data by ID rather than the full dataset.

Some of the clustering techniques didn’t pay off with good clean clusters, like the Kmeans and Agglomerative clustering.

**Conclusion**

DinoFunWorld seems to be a park with healthy trends across visitor activity. There are many visitors coming for multiple days, and three day park visitors are the most frequent type of guest across this dataset. They spend a lot of time in the park and engage in as many attractions as other types of guests. This type of brand loyalty should be cultivated.

Despite the health of the park, there are definitely areas that could be improved. Tracking check-ins at shops and restaurants seems like an easy implementation. Although there would be purchasing data available, that won’t paint the full picture. A guest could stop in a store, look around, and not purchase anything. If that number seems to be growing over time, it would seem to be a good reason to overhaul menus or in store selections.

Although we can’t be 100% sure, guest #657863’s unusual pattern of activity should be noted and reported to the proper authorities for additional investigation.