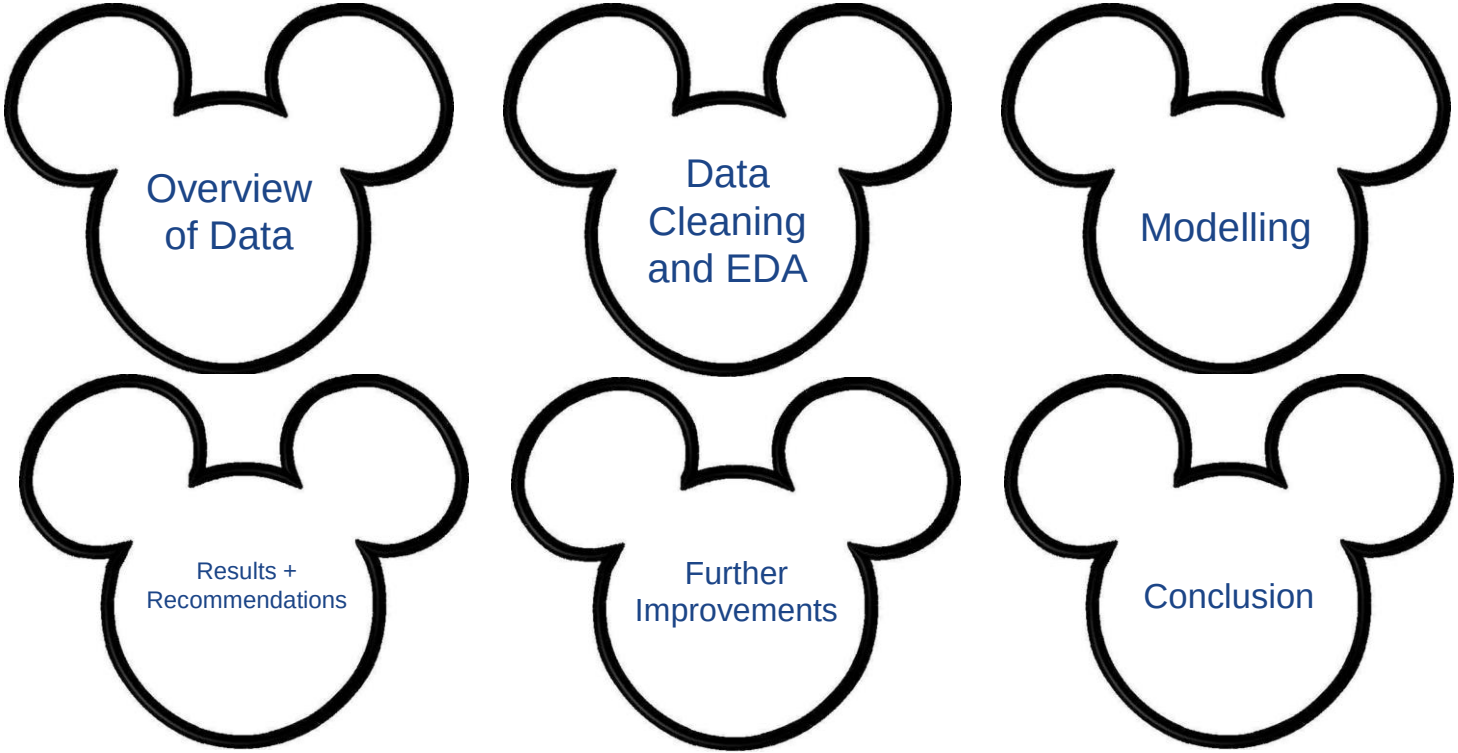
DSO 560 Final Project

Hanniyah Bilwani, Rebecca Bland, Kofi Buahin, Jena Lim, Suraj Swarup

Agenda


Overview of Data
Data Cleaning and EDA
Modelling
Results + Recommendations
Further Improvements
Conclusion

Overview of Data

This dataset includes reviews on three Disneyland branches (California, Paris, and Hong Kong).

These reviews are posted by visitors on Trip Advisor and show the Rating, Reviews, the Disney Branch being reviewed, the Month and Year of the Review, and the Reviewer Location.

We downloaded our dataset of 42,000 total reviews from Kaggle.

	Review_ID	Rating	Year_Month	Reviewer_Location	Review_Text	Branch
0	670772142	4	2019-4	Australia	If you've ever been to Disneyland anywhere you...	Disneyland_HongKong
1	670682799	4	2019-5	Philippines	Its been a while since d last time we visit HK...	Disneyland_HongKong
2	670623270	4	2019-4	United Arab Emirates	Thanks God it wasn t too hot or too humid wh...	Disneyland_HongKong
3	670607911	4	2019-4	Australia	HK Disneyland is a great compact park. Unfortu...	Disneyland_HongKong
4	670607296	4	2019-4	United Kingdom	the location is not in the city, took around 1...	Disneyland_HongKong

Business Objective

Our objective is to derive high level meaning from the thousands of Disney Parks reviews and develop business recommendations that can prevent the churn of customers that had a negative experience. Investing in these recommendations will save Disney Parks millions of dollars over the next decade by preventing a higher global churn rate and bringing all customers back to the Happiest Place On Earth and increasing global attendance.

Model Text Preprocessing

1. **General Cleaning**

   Removed punctuations, converted texts to lowercase, and replaced common named entities (such as urls, currency, etc.)

2. **Removing Stopwords**

   Elected to use the gensim library to remove stop words in comparison to nltk because it doesn't require you to tokenize the document first. Gensim includes 337 words in their stopword collection.

3. **Lemmatization**

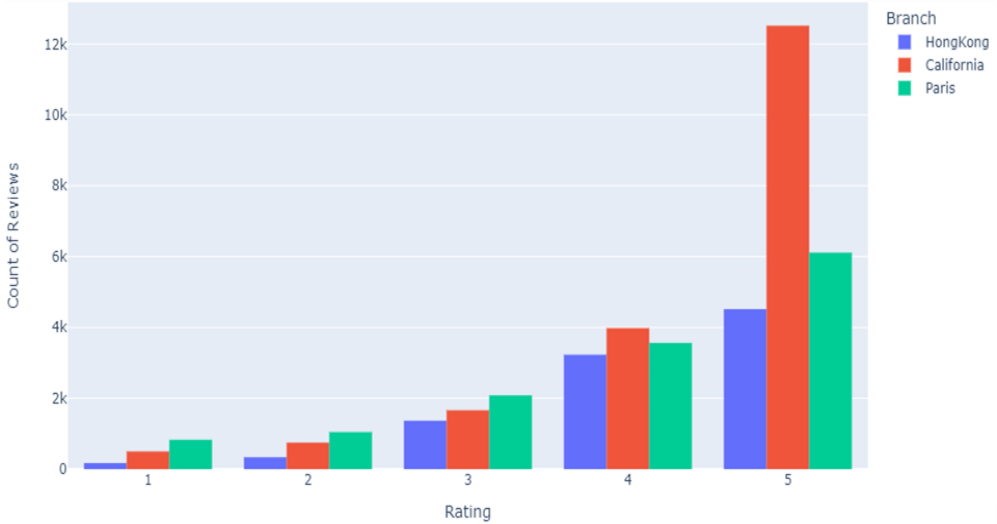
   To reduce words down to their inflection and help standardize our review text, we chose to do lemmatization. We prefer this over stemming as it takes into account the usage of the word and provides a more interpretable base word.

4. **Regex groupings**

   Did some regex groupings to help make ride names into one token, standardize different types of visitors into one group (e.g. families, kids, etc.) and standardized different names for Disney/Disneyland/Disneyworld and the cities it is located in.

EDA

Histogram of Count of Reviews by Rating and Branch



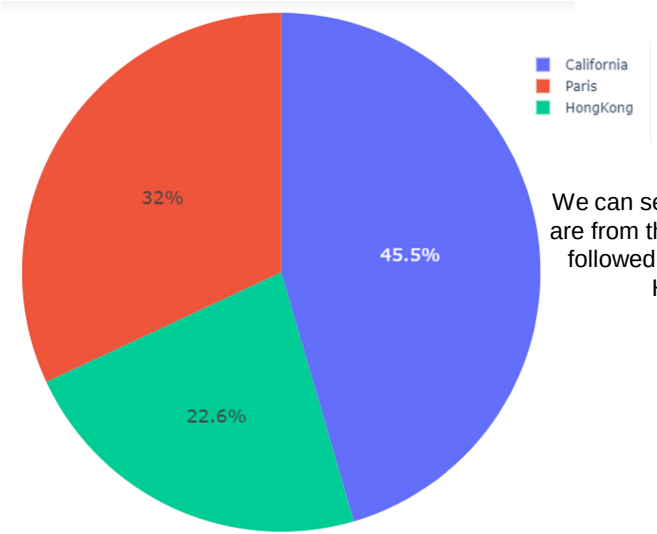
For all branches we can see that most reviews are 5 stars, and for each branch this data is left skewed. California has more reviews compared to other branches, but they also have far more 5 star reviews as well whereas the Paris branch has the most amount of 1,2,and 3 star reviews compared to the other branches.

Histogram of Average of Rating by Branch for Each Year

We can see that for each of the branches and across the past 10 years the average ratings by year have stayed relatively constant, with very slight fluctuations.

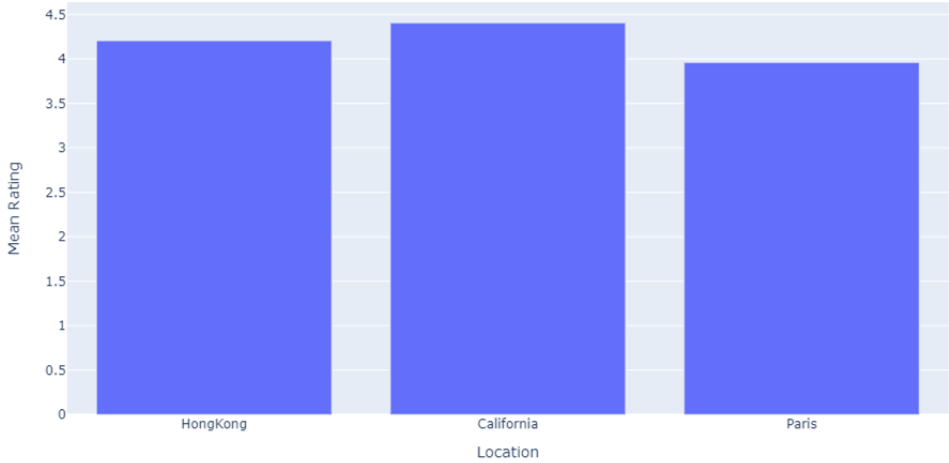
EDA (By Branch)

Pie Chart Showing Review Count by Branch



We can see that most reviews are from the California branch, followed by Paris, and then Hong Kong

Bar Chart Showing Mean Rating by Branch



We can see that the California Branch has the highest average rating, followed by Hong Kong, and Paris has the lowest average rating at 3.96

Sentiment Analysis

- Created Sentiment Analysis models to predict if a given review from TripAdvisor's reviews on the three Disney branches is positive or negative.
  - Using pre-trained BERT model
  - Using Logistic Regression

- Use case: can use this model to scrape data from social media sites regarding Disneyland and can help to find the sentiment analysis (if the review is positive or negative) when there is no quantifiable rating metric available

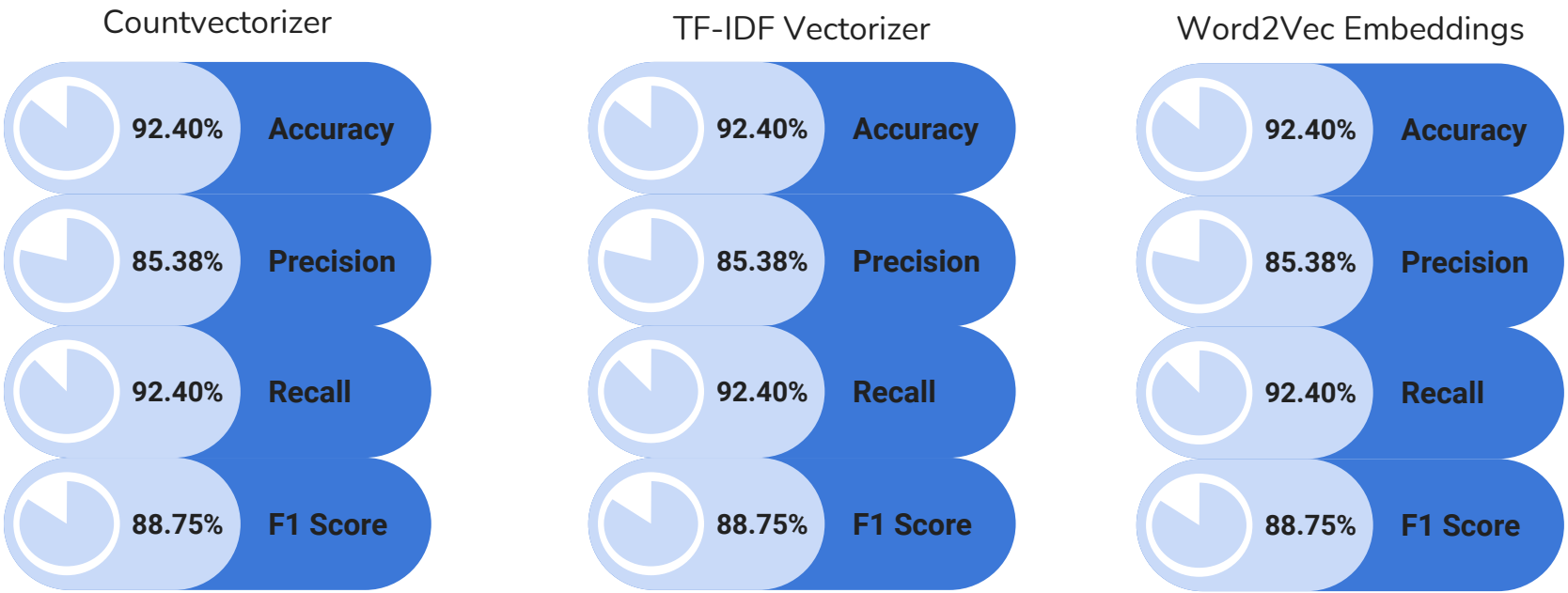
Sentiment Analysis BERT Model

- Used huggingface's pre-trained model and tokenizer: sentiment-analysis
- The results were skewed a bit towards precision, so used the f1 score (**83.26%**) to gauge our model's performance
- Result: Did better than the baseline f1 score of 50%
- Limitations: High number of false negatives and accuracy is less than baseline standard of around 80%

Metric	Result
Accuracy	73.66%
Precision	98.47%
Recall	72.13%
F1 Score	83.26%

Confusion Matrix	Actual Positives	Actual Negatives
Predicted Positives	3277	51
Predicted Negatives	1266	406

Sentiment Analysis with Logistic Regression



Limitations: due to only being able to run with a small dataset of heavily imbalanced data, we got the same scores for each vectorization technique; with more time, we will investigate further to fix this issue.

Topic Modeling

Based on our sentiment analysis results, we broke up our dataset into positive and negative reviews based on our three locations (California, Hong Kong, and Paris) and did topic modeling for each segment

Did a non-negative matrix factorization technique and opted to use TF-IDF for our vectorization because it gives us more information about the importance of a word in comparison to just count frequency

Ran topic modeling to see the highest topics as well as the top documents for each topic – tried different ngrams and components and ultimately decided on 3 topics with bigrams

Results and Recommendations - Paris

Results

Recommendations

Not meeting expectations in comparison to Disney Orlando and other US locations: *"We definitely won't be going back, and I can't recommend Disneyland Paris to anyone" , "Its a shame this park carry the name Disney... They build the place years back and they are now just collecting cash without any effort, it does not compare in anything with the American parks which are truly amazing"*

Disney Paris needs to upgrade their overall operations to provide a park experience that matches the American parks. This would include better manicured decorations/gardens, more characters taking photos with kids, staff monitoring people that are cutting in line, better trained staff, and overall more attention to detail to create a fun and magical environment

Closed Rides: *"A number of the attraction s being closed" , "There were also plenty of rides (like Autopia and Dumbo) that were closed for renovation" , "A number of the attraction s being closed and queues being very large*

Disney Paris needs to upgrade their communication with customers by letting them know beforehand of any ride closures, and also invest heavily in quality control methods to keep rides running as much as possible with minimal down time.

Disneyland Paris Recommendations ROI

Model Assumption	Value
% of total reviews about "crowded venue"	0.00674%
Average Annual Attendance Disneyland Paris (2009-2021)	9,181,538 (Refer to Appendix 1)
1 Day GA Ticket Price	$105.68
1 Day Child Ticket Price	$97.14
Average Ticket Price	$101.41
Low End Churn Rate	10%
High End Churn Rate	40%

$3,770,851.78 — ***$5,656,277.68***

Range of potentially recaptured annual revenue from Paris Recommendations

9,181,538 *Average visitors Annually* x 0.00674% *Customers lost to dissatisfaction over crowdedness* = 61,974 *Total re-capturable market of visitors*

Higher End of Recaptured Revenue = 61,974 *Total re-capturable market of visitors* x (0.9) *Return rate post churn (1 - High end churn rate)* = 55,776 *Actual Predicted Recaptured customers* x $101.41 *Average 1 Day Ticket Price* = **$5,656,277.68** *Max potentially recaptured revenue*

Lower End of Recaptured Revenue = 61,974 *Total re-capturable market of visitors* x (0.6) *Return rate post churn (1 - High end churn rate)* = 37,184 *Actual Predicted Recaptured customers* x $101.41 *Average 1 Day Ticket Price* = **$3,770,851.78** *Minimum potentially recaptured revenue*

Results and Recommendations - California

Results

Poor Disability Accommodations: "Disney is not disabled friendly any longer" "No disability parking available"

Rides break: "120 minute wait for Pirates of the Carribean! Splash Mountain, Bobsleds and Thunder Mountain were all down", "...even when Indiana Jones broke down (which I think is a feature of the ride at this point)."

Bad customer service: "Never have I seen such utter disregard for the handicapped, or such poor customer service", "My daughters disabilities are obvious really obvious unfortunately and this woman just was cold"

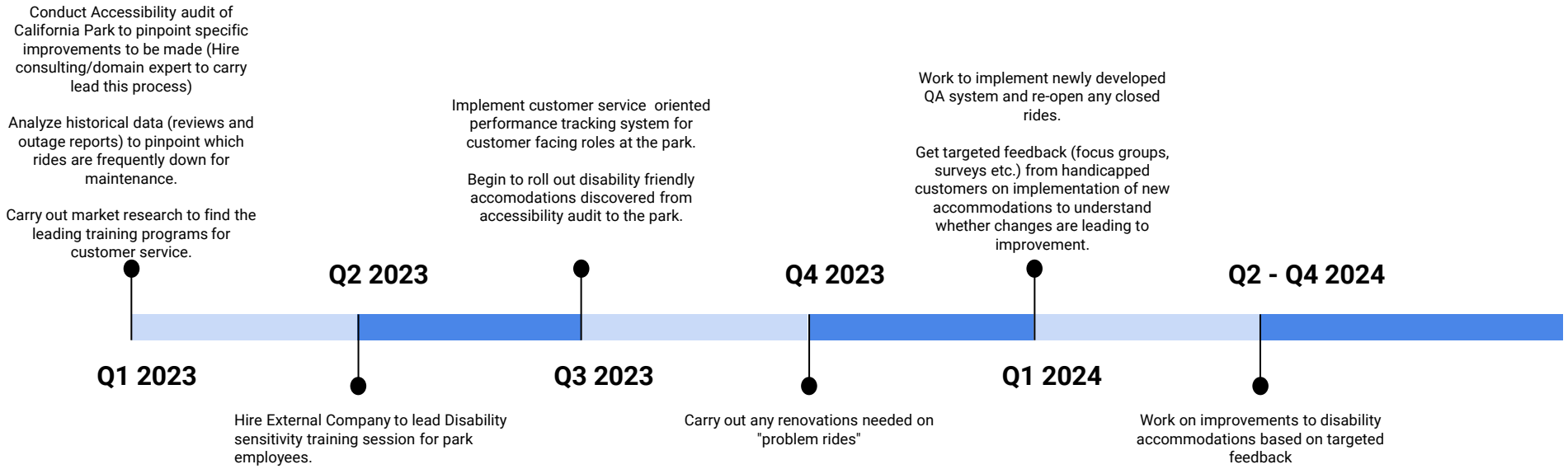
Recommendations

To address the poor disability accommodations, Disneyland California should work to come up with better accommodations and features/amenities to ensure their park offers benefits to handicapped customers.

To address rides breaking, Disneyland California should implement more frequent quality control checks to ensure the rides have as little downtime as possible.

Disneyland California must make sure to have more staff meetings and more trainings to keep up the high level of customer service, as poor customer service results in poor customer retention.

Disneyland California Implementation Roadmap



Disneyland California Recommendations ROI

Model Assumption	Value
% of total California reviews about bad "customer service" and "disability" accommodations	1.64%
Average Annual Attendance Disneyland California (2009-2021)	15,464,615 (Refer to Appendix 1)
1 Day GA Ticket Price	$138
1 Day Child Ticket Price	$130
Average Ticket Price	$134
Low End Churn Rate	10%
High End Churn Rate	40%

$20,438,538.06 — ***$30,657,807.09***

Range of potentially recaptured annual revenue from California Recommendations

15,464,615 *Average visitors Annually* x 1.64% *Customers lost to dissatisfaction over crowdedness* = 254,211 *Total re-capturable market of visitors*

Lower End of Recaptured Revenue = 254,211 *Total re-capturable market of visitors* x (0.9) *Return rate post churn (1 - High end churn rate)* = 228,790 *Actual Predicted Recaptured customers* x $134 *Average 1 Day Ticket Price* = **$30,657,807.09** *Max potentially recaptured revenue*

Higher End of Recaptured Revenue = 254,211 *Total re-capturable market of visitors* x (0.6) *Return rate post churn (1 - High end churn rate)* = 152,526 *Actual Predicted Recaptured customers* x $134 *Average 1 Day Ticket Price* = **$20,438,538.06** *Minimum potentially recaptured revenue*

Results and Recommendations - Hong Kong

Results

Recommendations

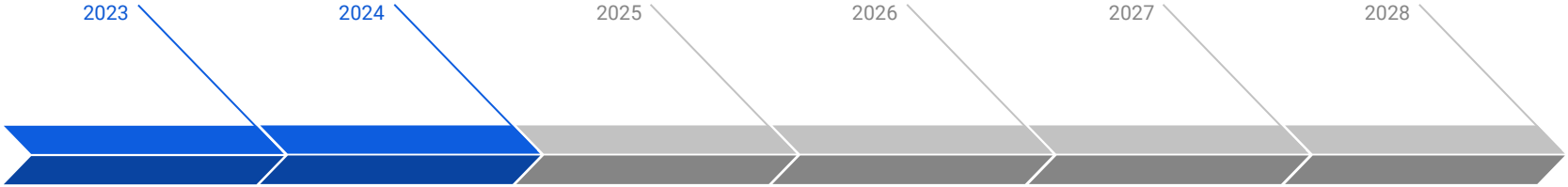
Small Theme Park: *"The park is so small, with so few people and performances that it is almost depressing" , "there is basically no reason for you to waste a day here.Disneyland Hong Kong is small... so small that on weekdays, if you arrived when the park opens, you can basically finish most of the best ride before noon"*

Disney Hong Kong has two options to address this issue. They could either work on an expansion plan to make their park bigger, or they could work to add additional amenities or smaller forms of entertainment to maximize the amount of entertainment they can offer given their limited space

Product offering limited/niche to younger demographic: *"Small kids only, teens might get bored" , "this park is definitely targeted at younger families" , "I would think the HK one is much smaller and for children only"*

Disney Hong Kong can address the fact that a lot of people think it is too targeted at younger kids by replacing certain rides/entertainment with those that are geared more towards teenagers and older kids

Disneyland Hong Kong Implementation Roadmap



Pre-planning

Market Study and Recommendations for level of expansion for theme park.

Concept Development to determine how the new section of the theme park fits into the already existing infrastructure of disney hong kong and how it can capture more demographics.

Master Planning to figure out the long term roadmap of what needs to be conducted to bring the project to fruition.

Early Stage Execution

Business Plan and Financing to solidify how the project is going to be funded

Design & Development of actual layout of park expansion

Construction Preparation and Documents (Securing permits, Government approvals etc.)

Construction Begins

Construction begins on actual expansion plans

Construction Concludes

Construction continues and is completed.

Testing and Training on the new expanded facilities starts

Post Construction

Soft Opening to get feedback from stakeholders and users.

Any updates made to facilities post soft launch.

Grand Opening

Hong Kong Disneyland Recommendations ROI

Model Assumption	Value
% of total reviews about "crowded venue"	7.98%
Average Annual Attendance Disneyland Hong Kong (2009-2020)	5,875,000 (Refer to Appendix 1)
1 Day GA Ticket Price	$82.18
1 Day Child Ticket Price	$61.09
Average Ticket Price	$71.64
Low End Churn Rate	10%
High End Churn Rate	40%

$20,151,985.33 — **$30,227,977.99**

Range of potentially recaptured revenue from Hong Kong Recommendations

5,875,000 *Average visitors Annually* x 7.98% *Customers lost to dissatisfaction over crowdedness* = 468,858 *Total re-capturable market of visitors*

Higher End of Recaptured Revenue

= 468,858 *Total re-capturable market of visitors* x (0.9) *Return rate post churn (1 - High end churn rate)* = 421,972 *Actual Predicted Recaptured customers* x $71.98 *Average 1 Day Ticket Price* = **$30,227,977.99** *Max potentially recaptured revenue*

Lower End of Recaptured Revenue

= 468,858 *Total re-capturable market of visitors* x (0.6) *Return rate post churn (1 - High end churn rate)* = 281,315 *Actual Predicted Recaptured customers* x $71.98 *Average 1 Day Ticket Price* = **$20,151,985.33** *Minimum potentially recaptured revenue*

Further Improvements

01	***Improve complexity of sentiment analysis model***	We used a pre-trained model from huggingface that provided baseline results, but can fine-tune a sentiment analysis model with different hyperparameters to improve our results
02	***Use more data***	Due to memory limitations, we only used a portion of the dataset to create our sentiment analysis model. Using the full 42,000 reviews can improve our model's accuracies by providing the model more data to learn from.While this data did not run on our computers, we believe with more RAM, it can be scalable to the full dataset.
03	***Balanced data***	Our data is highly skewed towards positive reviews and the majority of them are from California. Having a more balanced dataset in terms of positive/negative reviews and location can improve our topic modeling by giving us more insights into customers.

Conclusion

Overall, by analyzing the Disney Parks reviews with the sentiment analysis BERT model, logistic regression, and topic modeling, we were able to deliver 7 main recommendations, 2-3 per Disney Branch, that can make a huge impact in the future of Disney Parks. From impactful renovations, to meeting customer needs, Disney can invest in our recommendations to keep the customers coming back and ***save anywhere between $44,361,375.17 - $66,542,062.75 in annual revenue*** across its 3 parks in California, Paris and Hong Kong.

Sources

- https://www.disneytouristblog.com/new-castle-frozen-land-marvel-land-hong-kong-disneyland/
- https://www.statista.com/statistics/236154/attendance-at-the-disneyland-theme-park-california/
- https://www.statista.com/statistics/236162/attendance-at-the-paris-disneyland-park-theme-park/
- https://www.hongkongdisneyland.com/book/general-tickets/1day-tickets-o
- https://www.wdwinfo.com/disneyland/tickets.htm

Appendix 1: Disneyland Annual attendance by park from 2009 - 2021

	Annual Attendance		
Year	**Hong Kong**	**Paris**	**California**
2009	4,600,000	12,740,000	15,900,000
2010	5,200,000	10,500,000	15,980,000
2011	5,900,000	10,990,000	16,140,000
2012	6,700,000	11,200,000	15,960,000
2013	7,400,000	10,430,000	16,200,000
2014	7,500,000	9,940,000	16,770,000
2015	6,800,000	9,790,000	18,280,000
2016	6,100,000	8,400,000	17,940,000
2017	6,200,000	9,660,000	18,300,000
2018	6,700,000	9,840,000	18,660,000
2019	5,700,000	9,750,000	18,670,000
2020	1,700,000	2,620,000	3,670,000
2021	N/A	3,500,000	8,570,000

Source: Statista