Elizabeth Earl

DSC680-T301

Project 1 – Disney+ & Reviews

Business Problem

Reviews are a beneficial way for a company like Disney to try and predict how future planning will impact the company throughout. Analyzing review data will help Disney grow in different branches of their business such as theme parks, standalone, online sales, etc. Additionally, while analyzing the popularity of films throughout the years will help guide Disney on what brings in money in the box office and what fails to bring revenue or good reviews.

In recent years Disney began shifting to an online system in terms of merchandise sale as well as film releases this has made an impact in how people react to Disney. This project will aim to find whether it was a positive or negative impact. The aim of my project is to discover which branches' changes have made the most positive impact to Disney's company.

Background/History

After a yearlong of theme park and cinema closures due to the COVID-19 Disney suffered alongside other companies. Disney's revenue did in fact plumet as there was no guests entering the park or spending on merchandise. COVID-19 in a sense pushed companies like Disney into a corner where they had to make the choice to move most of their productions online.

Because Disney's theme parks require in person experiences the reviews guests leave will help Disney change (or stay the same in) their ways to adapt to the changing needs of guests. As for merchandise sales, because there are so many lingering restrictions of COVID-19 many physical stores are still suffering, to avoid this, online sales and how guests (who might not be local react) will play a major role in Disney's revenue. When it comes to films, because cinemas have been closed Disney as well will have to rely on previous film's success (based on reviews and monetary revenue) to determine what the public wants.

Data Explanation (Data Prep/Data Dictionary/etc)

Dataset -- Disney quarterly revenue:

- Revenue
- Cost of goods sold
- Gross Profit
- General and administrative expense
- Operating expense total

Dataset – Disney Parks reviews:

- Rating
- Review Text
- Branch
- Sentiment

Dataset – Disney+ content

- Date (Year & month)
- Count of movies

Methods

Using revenue data, which had to be corrected from an HTML chart to a csv file, I used a correlation matrix to determine what where the variables that were most impacting to Disney's revenue. Focusing then on my second dataset, involving park reviews, I ran a TextBlob to have sentiment values based on positive (indicated by value 1) and negative (indicated by value 0) rating.

The biggest dataset used was the Disney Parks review dataset, which I used a subset of 1000 random values. I ran three models on my data: Logistic Regression, multinomial Naive Bayes classifier and Decision Tree Classifier. Overall, my accuracies were not any better than when I tried smaller datasets (of the same dataset).

Analysis

With the initial revenue dataset, I wanted to focus more on the "Cost of goods sold" variable as I determined that would be the closest to included revenue based off merchandise. Unfortunately, this variable did not have much correlation with the revenue, so it became a dead end in my research. (see figure 1)

With sentiment analysis I sorted the ratings from positive and negative. The sorting resulted in Disneyland California parks receiving the most positive reviews, next was Hong Kong Disney and coming in last was Disneyland Paris. (see figure 2)

Most of the analysis was done through the three models mentioned previously. The accuracy of using each model were as follows: (see figures 3a-3c)

- Logistic Regression: 51% (figure 3a)
- Multinomial Naive Bayes classifier: 50% (figure 3b)
- Decision Tree Classifier: 46% (figure 3c)

For the Logistic Regression model, which had the best accuracy, I noticed a majority of the reviews being used where on the higher side of the scale. Logistic regression saw a majority of 3-5 (on a scale of 1-5) reviews versus my Decision tree Classifier which had scores from 2-5. To have more accurate results we would want to use a model that includes a larger variety of scores so for this reason I ignored my Multinomial Naive Bayes classifier (which only had ratings of 5 included).

The Disney+ dataset results were quite interesting but not surprising. “Disney+ was launched on November 12, 2019, in the United States, Canada, and the Netherlands, and expanded to Australia, New Zealand, and Puerto Rico a week later” ("Disney+ - Wikipedia", 2022). Because Disney+ was founded in November 2019 it is expected that an abundance of films would be pushed on the site; after 2019 Disney+ slowed down the release of content on the site. From 2019 to 2020 the film content lowered by more than double (630->230) while from 2020 to 2021 content went down less than half (230->131). (see figure 4)

Conclusion

Disney has a lot of branches where their revenue comes from, and I concluded that regardless of outside factors (such as COVID) Disney still thrives. On a small scale

I noticed reviews typically fall towards a higher rating which shows how much people are devoted to Disney. Disney does not have to rely on in-person and online sales, they can switch between one or the other and still bring in high ratings from guests. As seen in my Disney+ dataset analysis, Disney has come up with great tactics to intice people to spend their money. When Disney+ was released Disney essentially dumped content on the site to get people interested and give them a sense of FOMO where people are more likely to sign up for a streaming service, they feel is worth it.

This idea that Disney creates services to entice a fear of missing out on guests thus bringing in revenue can be seen throughout the parks as well. Disney created new ticketing schemes to get people to feel they need to purchase. Initially the Magic Key Program Disney introduced did not go down well but, this negative response to Magic Key pushed Disney into creating incentives to join with Magic Key exclusive merchandise.

Assumptions

Before running any models on the review's dataset, I expected my Decision Tree Classifier to have the best accuracy but, it was the opposite. It is assumed that no matter what Disney creates, or gets rid of, people will always have a deep love for the company. Disney pushes content and merchandise when ratings are low and this way, they are able to keep a high average of guests happy throughout the year.

Limitations

Because Disney is a big corporation, a lot of the data available regarding revenue and sales is limited. There is a big focus on data privacy on their end so when

looking for initial datasets it was quite difficult. The dataset used for quarterly revenue for example was created (by me) using revenue data listed in chart form rather than the formatting required for this project.

Another limitation was regarding the parks being reviewed as there are more than just three (California, Paris, and Honk Kong) Disney parks. My assumption for the cause of this limitation is that perhaps the other parks are not as visited yearly versus the included parks (yet this would not apply to Disney World Florida).

Something that played a major role in not getting accuracies with good numbers is the size of the dataset used. I realized this limitation as the file was unable to be uploaded to Github for it was too big, this goes to show how the quantity of values used in the models could drastically affect the quality.

Challenges

After getting past the limited data issue, I was stumped with the issue of abundance of rating data. I had to cut my dataset down to preserve memory on my device; I ran a small-scale version of my dataset which only included 1000 values. I ran into a challenge with my third dataset as I was not able to find a dataset that I wanted to work with that was similar to the Disney Park review data. The dataset I used simply included content data on what was on Disney's streaming platform Disney+. Although this was limited data provided this did give me an insight on how the times impacted how much content was pushed onto Disney+.

Future Uses/Additional Applications

This idea that reviews are most of the time positive is helpful for Disney as they can determine when a guests happiness peaks. Being such a big company, Disney can use this information to predict when they need to push out new products to keep consumers happy. If they use these trends to track consumer's happiness, they can continue to maintain happiness and positive revenue throughout the year and essentially never worry about global emergencies to negatively affect their company.

Answering some of the questions I proposed in the last milestone opened the door to what could be done with the information I received. I noticed that when I forced my models to use negative review values the accuracy was less, I was getting 30% accuracy versus 50%. This drastic difference in accuracy shows how the negative reviews are not as common as the positive and therefore do not impact Disney as much. Because negative reviews are limited Disney can use this information to assume they can simply continue to push programs they already have in place.

Implementation Plan

Q&A:

1. What other models would be helpful in analyzing reviews?
    a. Because of my limited data and low accuracies, I determined, if I had a more capable machine, a regression tree could be used to get better accuracy. This assumption is made as the decision tree I used was included more review values versus all other models. Having a variety of review values allows for a better scale of opinions to be considered.
2. Can negative reviews be forced into the models?
    a. To do this I sorted the sentiments from lowest rating to highest; the limitation of this is that only the lowest reviews are used rather than a random number from reviews under a 3 rating. But by getting the lowest rating values I concluded I would get a very extreme accuracy.
3. What is the accuracy of negative reviews versus the positive?
    a. The accuracies when running on a subset focused on negative reviews yielded the following accuracies with the previously used models:
        i. Logistic Regression: 36% (figure 5a)
        ii. Multinomial Naive Bayes classifier: 30% (figure 5b)
        iii. Decision Tree Classifier: 32% (figure 5c)
4. What about tv shows released by Disney+?
    a. TV shows were not surprisingly less likely to be pushed onto Disney+, I assumed this was due in part to the fact Disney had already produced more films throughout the years versus TV shows. Although this was the

case I did notice a trend of TV shows being constistently being pushed out unlike films. Disney+ is focusing on producing TV Shows that produce weekly episodes so it made sense that this was more of a constant decrease versus the drastic drop of films dropped in Disney+.

5. What trends are noticeable on Disney+ with Tv shows?
    a. As mentioned in the previous question TV shows proved to be less likely to be put onto Disney+ but their production/addition was slightly diminishing in content through the years. (See figure 6a & 6b)
6. Going off the previous question, how do TV show trends compare to film trends?
    a. Versus films we see that TV shows also were decreasing in content on Disney+. Although this is the case TV shows did not drop in major numbers like films did per year. I assumed that TV shows were more likely to be pushed onto Disney+ as they help maintain a constant audience as they push episodes weekly versus a film that is a simple one day watch.
7. Data shows Disney+ content in initial countries what about the others?
    a. Unfortunately there was not much I could do with this column of data as many of the same countries were written in various ways. A simple example is USA being written as US, United States, U. States. There was already too much data to work with using this dataset so finding out the different ways would be too tedious and provide no real information as the country was vague on whether it was the content's origin or destination.
8. Do guests consider outside factors such as COVID when basing their reviews of Disney?

    a. I can conclude that this dataset was prior to COVID as there was a total of 0 reviews that contained any trace of “covid”. I figured maybe other key words like ‘sick’ or ‘ill’ would contribute to a negative review of feeling sick but this is not a direct correlation to COVID. From sick/ill reviews I found that 41% of people complained against Disney dude to an outside factor of feeling sick.

9. Are reviews solely on the park experience or do they include hotel/streaming services opinions?
    a. 11% of the total reviews were either on Disney’s hotels or online services. This percentage of reviews being so low did not prove to have any important impact on the given data.
10. Can revenue on Disney’s streaming services be compared against Disney Park revenue?
    a. Because there is such limited data on how Disney gains revenue or how their revenue is doing I was unable to find datasets containing the revenue that Disney+ solely bring to Disney as a company.

Supporting illustrations

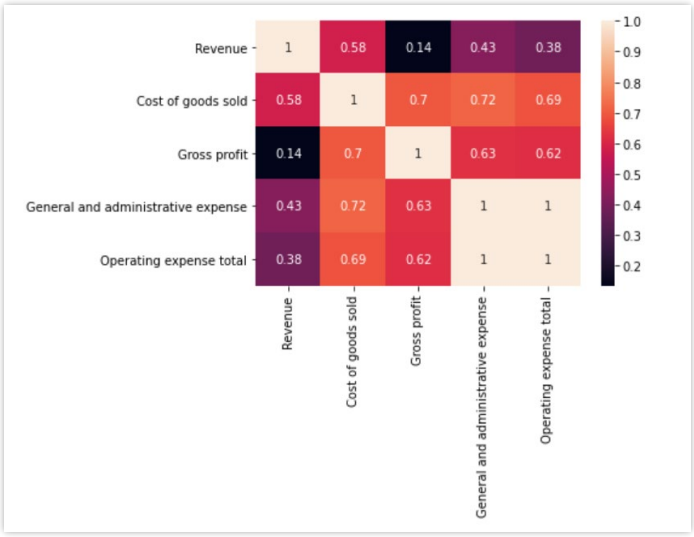

Figure 1

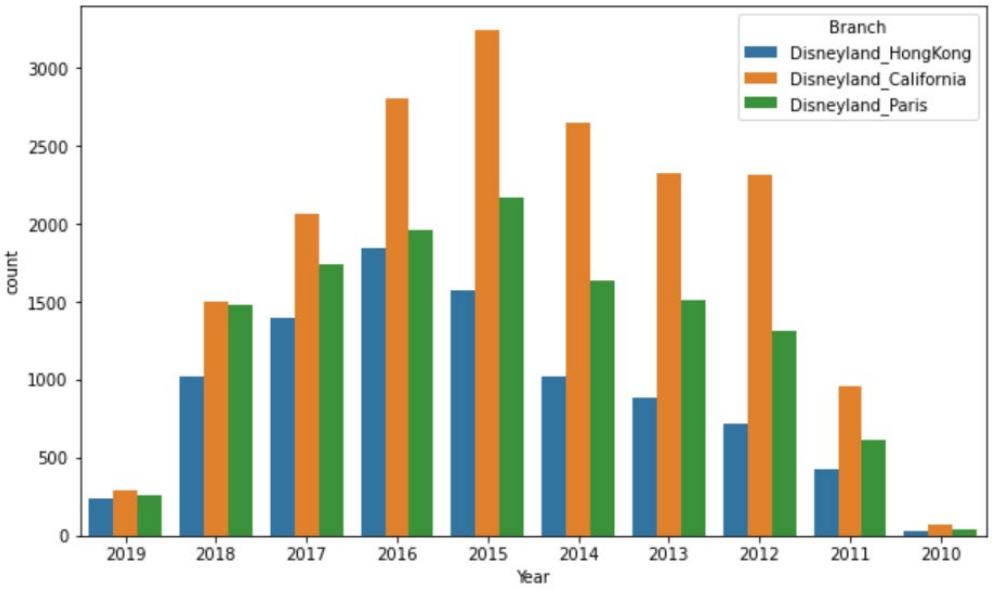

Figure 2

	precision	recall	f1-score	support
1	0.00	0.00	0.00	6
2	0.00	0.00	0.00	11
3	1.00	0.03	0.05	37
4	0.55	0.17	0.25	103
5	0.51	0.95	0.66	143
accuracy			0.51	300
macro avg	0.41	0.23	0.19	300
weighted avg	0.55	0.51	0.41	300

Figure 3a

	precision	recall	f1-score	support
1	0.00	0.00	0.00	4
2	0.00	0.00	0.00	19
3	0.00	0.00	0.00	68
4	0.00	0.00	0.00	158
5	0.50	1.00	0.67	251
accuracy			0.50	500
macro avg	0.10	0.20	0.13	500
weighted avg	0.25	0.50	0.34	500

Figure 3b

	precision	recall	f1-score	support
1	0.00	0.00	0.00	4
2	0.07	0.05	0.06	19
3	0.27	0.26	0.27	68
4	0.34	0.32	0.33	158
5	0.61	0.64	0.62	251
accuracy			0.46	500
macro avg	0.26	0.26	0.26	500
weighted avg	0.45	0.46	0.46	500

Figure 3c

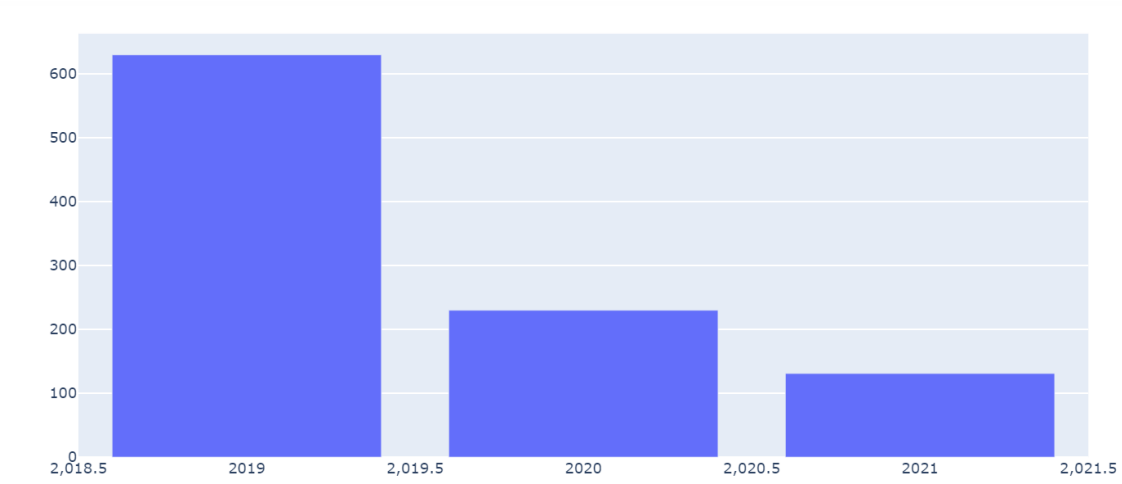

Figure 4

	precision	recall	f1-score	support
1	0.29	0.42	0.34	60
2	0.31	0.27	0.29	63
3	0.38	0.48	0.43	92
4	0.38	0.07	0.12	41
5	0.49	0.41	0.44	44
accuracy			0.36	300
macro avg	0.37	0.33	0.32	300
weighted avg	0.36	0.36	0.34	300

Figure 5a

	precision	recall	f1-score	support
1	0.46	0.17	0.25	111
2	0.23	0.21	0.22	105
3	0.29	0.76	0.42	133
4	0.00	0.00	0.00	69
5	0.83	0.12	0.21	82
accuracy			0.30	500
macro avg	0.36	0.25	0.22	500
weighted avg	0.36	0.30	0.25	500

Figure 5b

	precision	recall	f1-score	support
1	0.34	0.29	0.31	111
2	0.28	0.31	0.30	105
3	0.31	0.28	0.29	133
4	0.26	0.26	0.26	69
5	0.39	0.48	0.43	82
accuracy			0.32	500
macro avg	0.32	0.32	0.32	500
weighted avg	0.32	0.32	0.32	500

Figure 5c

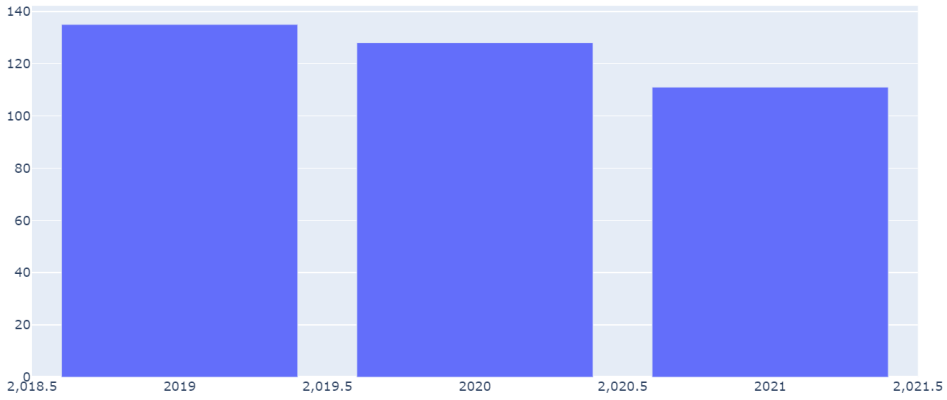

Figure 6a

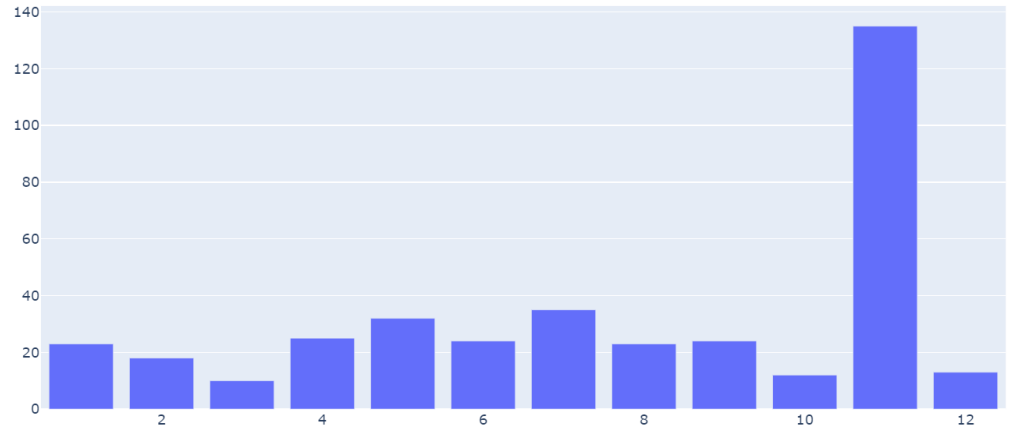

Figure 6b

Appendix

(2022). Retrieved 29 March 2022, from https://craft.co/the-walt-disney/metrics

The site provided chart formatting information on Disney revenue as a company. From charts/tables I was able to create a csv file containing quarterly revenue.

Disney+ - Wikipedia. (2022). Retrieved 30 March 2022, from https://en.wikipedia.org/wiki/Disney%2B?msclkid=767a5ecfb27911ec9498fe2d4b31277c

Website was used to find basic information on when and where Disney+ was rolled out.

Disney+ Shows and Movies - Exploratory Analysis. (2022). Retrieved 2 April 2022, from https://www.kaggle.com/code/shivamb/disney-shows-and-movies-exploratory-analysis/data

Disneyland Reviews. (2022). Retrieved 2 April 2022, from https://www.kaggle.com/datasets/arushchillar/disneyland-reviews?msclkid=1c5b9354a7c711ec968f3c333c2e4ac5

The previous two refences were used for datasets included throughout my research