Disneyland & Universal Studios Review Analysis

Émilie Brazeau, Nicholas Gin, Gordon Tang

01.

Planning our Data Mart

Disneyland Reviews

Reviews and Ratings of 3 Disneyland branches - California, Hong Kong and Paris

Last Updated: 2 years ago (Version 1)

About this Dataset

The dataset includes 42,000 reviews of 3 Disneyland branches - Paris, California and Hong Kong, posted by visitors on Trip Advisor.

Column Description:

1. Review_ID: unique id given to each review
2. Rating: ranging from 1 (unsatisfied) to 5 (satisfied)
3. Year_Month: when the reviewer visited the theme park
4. Reviewer_Location: country of origin of visitor
5. Review_Text: comments made by visitor
6. Disneyland_Branch: location of Disneyland Park

Review_ID
Rating
Year_Month
Reviewer_Location
Review_Text
Branch

Review_ID	Rating	Year_Month	Reviewer_...	Review_Text	Branch
670772142	4	2019-4	Australia	If you've ever been to Disneyland anywhere you'll find Disneyland Hong Kong very similar in the layo...	Disneyland_HongKong

Why this dataset?

1

Find correlations between ratings and branch.

“Which Disneyland branch has the best reviews?”

2

Find correlations between ratings and other factors.

“Is there a time of year when people leave the most positive reviews?”

3

Find trends in how reviews are written.

“Which words are most commonly found in reviews left by visitors?”

4

End goal:

Machine learning model that can predict a review’s rating if it is given a review text.

Enriching the dataset

Reviews of Universal Studios

Reviews and Ratings of 3 Universal Studios branches

About Dataset

Context

Universal Studio gets a vast amount of reviews from visitors. To go through all the reviews can be a tedious job. We have to categorize reviews expressed. This can be utilized for the reviews management system. We determining overall reviews based on individual comments. So that company can get a complete idea of reviews provided by visitors and can take care of those particular fields. This makes more loyal visitors to the company, increase business, fame, brand value, and also profit.

Content

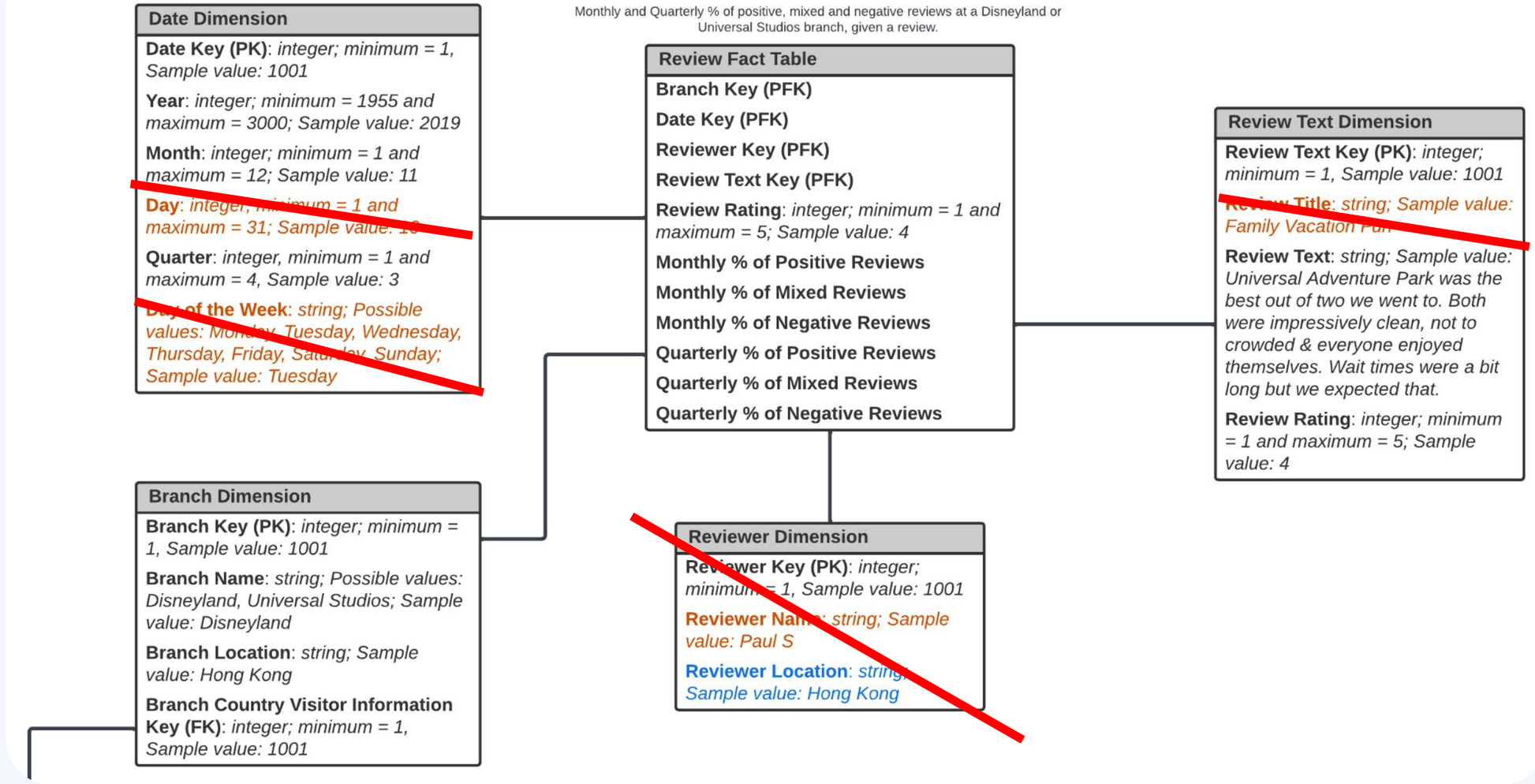
The dataset includes 50,000++ reviews of 3 Universal Studios branches (Florida, Singapore, Japan), posted by visitors on the Trip Advisor website.

Description of columns:

1. reviewer - account name of the reviewer
2. rating - rating from the reviewer, from 1 (unsatisfied) to 5 (satisfied)
3. written_date - date of the review
4. title - the title of the review
5. review_text - review made by the visitor
6. branch - location of Universal Studios

reviewer
rating
written_date
title
review_text
branch

reviewer	rating	written_date	title	review_text	branch
Kelly B	2.0	May 30, 2021	Universal is a complete Disaster - stick with Disney!	We went to Universal over Memorial Day weekend and it was a total train wreck. We waited to get in t...	Universal Studios Florida



Our Initial Dimensional Model

Challenge:

Scale up the datasets?

Our Initial Dimensional Model

Monthly and Quarterly % of positive, mixed and negative reviews at a Disneyland or Universal Studios branch, given a review.

Review Fact Table
- **Branch Key (PFK)**
- **Date Key (PFK)**
- **Reviewer Key (PFK)**
- **Review Text Key (PFK)**
- **Review Rating**: *integer; minimum = 1 and maximum = 5; Sample value: 4*
- **Monthly % of Positive Reviews**
- **Monthly % of Mixed Reviews**
- **Monthly % of Negative Reviews**
- **Quarterly % of Positive Reviews**
- **Quarterly % of Mixed Reviews**
- **Quarterly % of Negative Reviews**

Date Dimension
- **Date Key (PK)**: *integer; minimum = 1, Sample value: 1001*
- **Year**: *integer; minimum = 1955 and maximum = 3000; Sample value: 2019*
- **Month**: *integer; minimum = 1 and maximum = 12; Sample value: 11*
- **Day**: *integer, minimum = 1 and maximum = 31; Sample value: 10*
- **Quarter**: *integer, minimum = 1 and maximum = 4, Sample value: 3*
- **Day of the Week**: *string; Possible values: Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday; Sample value: Tuesday*

Review Text Dimension
- **Review Text Key (PK)**: *integer; minimum = 1, Sample value: 1001*
- **Review Title**: *string; Sample value: Family Vacation Fun*
- **Review Text**: *string; Sample value: Universal Adventure Park was the best out of two we went to. Both were impressively clean, not to crowded & everyone enjoyed themselves. Wait times were a bit long but we expected that.*
- **Review Rating**: *integer; minimum = 1 and maximum = 5; Sample value: 4*

Branch Dimension
- **Branch Key (PK)**: *integer; minimum = 1, Sample value: 1001*
- **Branch Name**: *string; Possible values: Disneyland, Universal Studios; Sample value: Disneyland*
- **Branch Location**: *string; Sample value: Hong Kong*
- **Branch Country Visitor Information Key (FK)**: *integer; minimum = 1, Sample value: 1001*

Reviewer Dimension
- **Reviewer Key (PK)**: *integer; minimum = 1, Sample value: 1001*
- **Reviewer Name**: *string; Sample value: Paul S*
- **Reviewer Location**: *string; Sample value: Hong Kong*

Branch Country Visitor Information Outrigger Dimension
- **Branch Country Visitor Information Key (PK)**: *integer; minimum = 1, Sample value: 1001*
- **Number of Arrivals (yearly?)**: *integer; minimum = 1, Sample value: 21821*
- **Income Group**: *string; Sample value: High income*

Challenge:

Scale up the datasets?

Our Final Dimensional Model

Grain: Monthly rating scaling at Disneyland or Universal Studios for a given branch.

Review Fact Table

Date Key (PFK)

Branch Key (PFK)

Review Text Key (PFK)

Monthly % of Positive Reviews: *double; minimum = 0.0 and maximum = 100.0; Sample value = 8.8*

Monthly % of Mixed Reviews: *double; minimum = 0.0 and maximum = 100.0; Sample value = 8.8*

Monthly % of Negative Reviews: *double; minimum = 0.0 and maximum = 100.0; Sample value = 8.8*

Date Dimension

Date Key (PK): *integer; minimum = 1, Sample value: 1001*

Month: *integer; minimum = 1 and maximum = 12; Sample value: 11*

Quarter: *integer, minimum = 1 and maximum = 4, Sample value: 3*

Year: *integer; minimum = 2002; Sample value: 2019*

Review Text Dimension

Review Text Key (PK): *integer; minimum = 1, Sample value: 1001*

Review Text: *string; Sample value: Universal Adventure Park was the best out of two we went to. Both were impressively clean, not to crowded & everyone enjoyed themselves. Wait times were a bit long but we expected that.*

Rating: *integer; minimum = 1 and maximum = 5; Sample value: 4*

Branch Attendance Outrigger Dimension

Branch Attendance Key (PK): *integer; minimum = 1, Sample value: 1001*

Attendance 2006 (millions): *integer; minimum = 1, Sample value: 21821*

Attendance 2007 (millions): *integer; minimum = 1, Sample value: 21821*

Attendance 2008 (millions): *integer; minimum = 1, Sample value: 21821*

Attendance 2009 (millions): *integer; minimum = 1, Sample value: 21821*

Attendance 2010 (millions): *integer; minimum = 1, Sample value: 21821*

Attendance 2011 (millions): *integer; minimum = 1, Sample value: 21821*

Attendance 2012 (millions): *integer; minimum = 1, Sample value: 21821*

Attendance 2013 (millions): *integer; minimum = 1, Sample value: 21821*

Attendance 2014 (millions): *integer; minimum = 1, Sample value: 21821*

Attendance 2015 (millions): *integer; minimum = 1, Sample value: 21821*

Attendance 2016 (millions): *integer; minimum = 1, Sample value: 21821*

Attendance 2017 (millions): *integer; minimum = 1, Sample value: 21821*

Attendance 2018 (millions): *integer; minimum = 1, Sample value: 21821*

Attendance 2019 (millions): *integer; minimum = 1, Sample value: 21821*

Attendance 2020 (millions): *integer; minimum = 1, Sample value: 21821*

Attendance 2021 (millions): *integer; minimum = 1, Sample value: 21821*

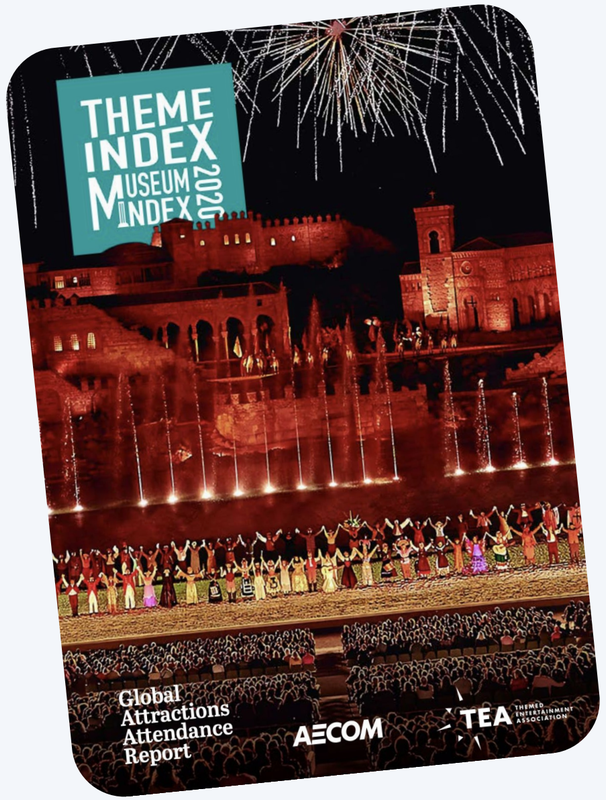
Branch Dimension

Branch Key (PK): *integer; minimum = 1, Sample value: 1001*

Branch Name: *string; Possible values: Disneyland California, Disneyland Hong Kong, Disneyland Paris, Universal Studios Florida, Universal Studios Japan, Universal Studios Singapore; Sample value: Disneyland Paris*

Branch Attendance Key (FK): *integer; minimum = 1, Sample value: 1001*

Where are we getting this data?

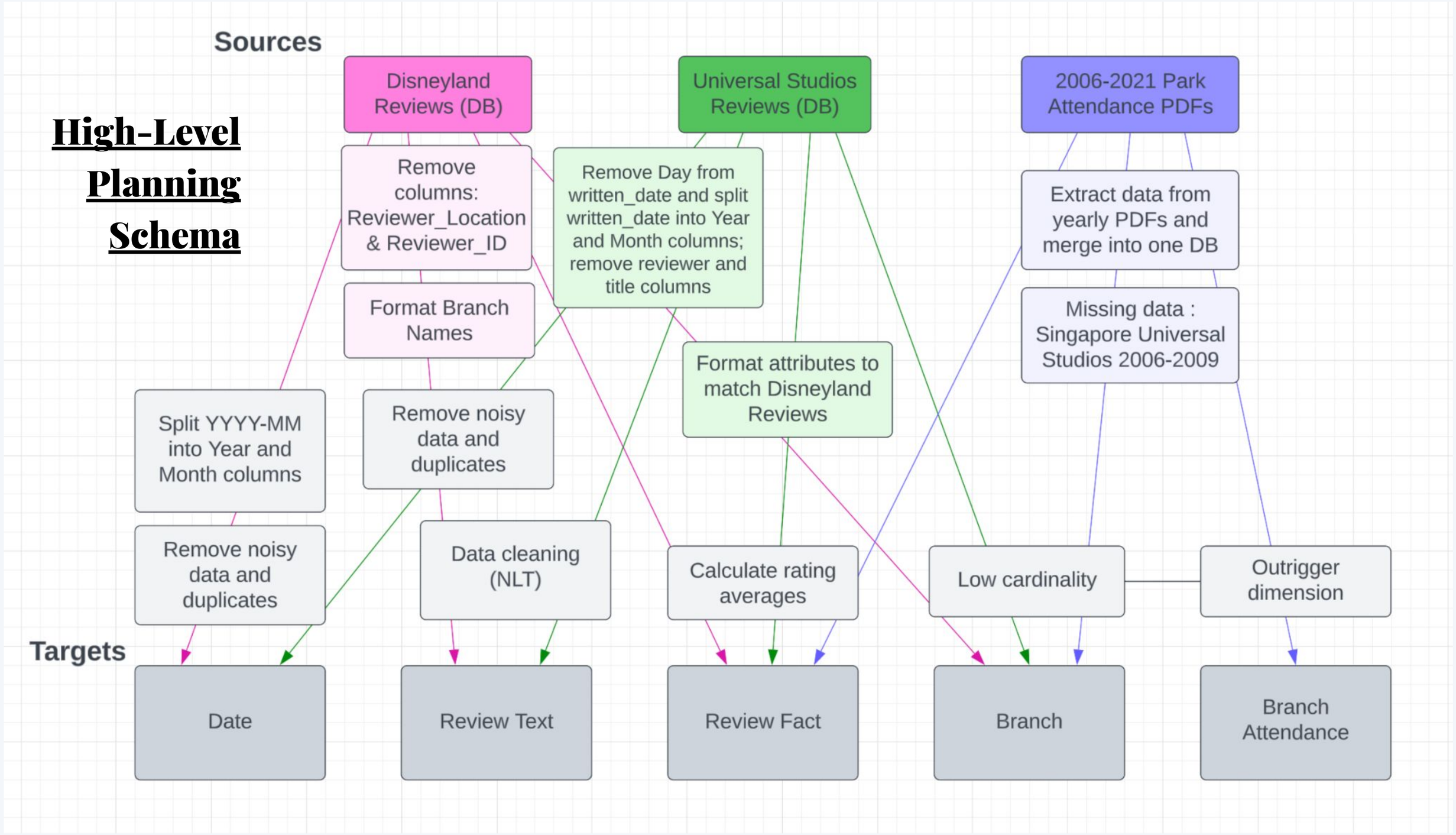


Park Attendance Dataset

- Manually compiled a Park Attendance Dataset ourselves using data from published Global Attraction Attendance Reports from the years 2006-2021.

Branch	2006 Attendance	2007 Attendance	2008 Attendance	2009 Attendance	2010 Attendance	2011 Attendance	2012 Attendance	2013 Attendance	2014 Attendance	2015 Attendance	2016 Attendance	2017 Attendance	2018 Attendance	2019 Attendance	2020 Attendance	2021 Attendance (millions)
Disneyland Califo	14.73	14.87	14.72	15.9	15.98	16.14	15.96	16.2	16.77	18.28	17.94	18.3	18.66	18.66	3.67	8.5
Disneyland Paris	10.6	12	12.688	12.74	10.5	10.99	11.2	10.43	9.94	10.36	8.4	9.66	9.843	9.745	2.62	3.5
Disneyland Hong	5.2	4.15	4.5	4.6	5.2	5.9	6.7	7.4	7.6	6.8	6.1	6.2	6.7	5.695	1.7	2.6
Universal Studios	8.5	8.713	8.3	8	8.16	8.5	9.7	10.1	11.8	13.9	14.5	14.935	14.3	14.5	4.901	5.5
Universal Studios	6	6.2	6.231	5.53	5.925	6.044	6.195	7.062	8.263	9.585	9.998	10.198	10.708	10.922	4.096	8.987
Universal Studios Singapore					2	3.411	3.48	3.65	3.84	4.2	4.1	4.22	4.4	4.5	1.098	1.2

High-Level Planning Schema



02.

Extract, Transform & Load

Extracting our Datasets

- Loaded datasets (as csvs) into dataframes

```
[ ] # Loading the data
    url='https://drive.google.com/file/d/1atEjlCz6cNs4_bKLd2Mwe-W-MIKwz3AU/view?usp=sharing'
    url='https://drive.google.com/uc?id=' + url.split('/')[-2]
    df_dis = pd.read_csv(url, encoding='cp1252')

    df_dis.head(9)
```

	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
5	670591897	3	2019-4	Singapore	Have been to Disney World, Disneyland Anaheim ...	Disneyland_HongKong
6	670585330	5	2019-4	India	Great place! Your day will go by and you won't...	Disneyland_HongKong
7	670574142	3	2019-3	Malaysia	Think of it as an intro to Disney magic for th...	Disneyland_HongKong
8	670571027	2	2019-4	Australia	Feel so let down with this place,the Disneylan...	Disneyland_HongKong

```
[ ] # Loading the data
    url='https://drive.google.com/file/d/1Avu5OYvmN8kWFzICK3_KE13Ob1V0uPwP/view?usp=sharing'
    url='https://drive.google.com/uc?id=' + url.split('/')[-2]
    df_uni = pd.read_csv(url, encoding='UTF-8')

    df_uni
```

	reviewer	rating	written_date	title	review_text	branch
0	Kelly B	2.0	May 30, 2021	Universal is a complete Disaster - stick with ...	We went to Universal over Memorial Day weekend...	Universal Studios Florida
1	Jon	1.0	May 30, 2021	Food is hard to get.	The food service is horrible. I'm not reviewin...	Universal Studios Florida
2	Nerdy P	2.0	May 30, 2021	Disappointed	I booked this vacation mainly to ride Hagrid m...	Universal Studios Florida
3	ran101278	4.0	May 29, 2021	My opinion	When a person tries the test seat for the ride...	Universal Studios Florida
4	tammies20132015	5.0	May 28, 2021	The Bourne Stuntacular...MUST SEE	Ok, I can't stress enough to anyone and everyo...	Universal Studios Florida
...	...	...	...	...	...	...
50899	vinz20	4.0	March 29, 2010	I'll Be Back Only If ...	This is my first visit to a Universal Studio t...	Universal Studios Singapore
50900	betty l	4.0	March 29, 2010	Universal Studios Singapore Experience	We finally visited Singapore's very first them...	Universal Studios Singapore
50901	spoonos65	4.0	March 28, 2010	Impressive but not quite finished!	We visited during the first week of its 'soft ...	Universal Studios Singapore
50902	HeatSeekerWrexham_UK	4.0	March 22, 2010	Small but beautifully marked	We visited on the 3rd day of the 'soft' openin...	Universal Studios Singapore
50903	sc_myinitial	5.0	February 24, 2010	Excellent Sneak Preview	My group managed to get the tickets for the 16...	Universal Studios Singapore

50904 rows × 6 columns

Transforming our Datasets

On review texts:

- **Strip whitespaces**
- **Decapitalize all review texts**
- **Remove an unicode characters**
- **Remove stop words**
- **Lemmatize the words**
  (we originally stemmed the words, but changed to lemmatization - more on this later)

	Review_ID	Rating	Year_Month	Reviewer_Location	Review_Text	Branch
0	670772142	4	2019-4	Australia	ever disneyland anywhere find disneyland hong ...	Disneyland Hong Kong
1	670682799	4	2019-5	Philippines	since last time visit hk disneyland yet time s...	Disneyland Hong Kong
2	670623270	4	2019-4	United Arab Emirates	thanks god hot humid visiting park otherwise w...	Disneyland Hong Kong
3	670607911	4	2019-4	Australia	hk disneyland great compact park unfortunately...	Disneyland Hong Kong
4	670607296	4	2019-4	United Kingdom	location city took around 1 hour kowlon kid li...	Disneyland Hong Kong
5	670591897	3	2019-4	Singapore	disney world disneyland anaheim tokyo disneyla...	Disneyland Hong Kong
6	670585330	5	2019-4	India	great place day go even know obviously went da...	Disneyland Hong Kong
7	670574142	3	2019-3	Malaysia	think intro disney magic little one almost att...	Disneyland Hong Kong
8	670571027	2	2019-4	Australia	feel let place disneyland train fantastic get ...	Disneyland Hong Kong
9	670570869	5	2019-3	India	go talking disneyland whatever say le disneyla...	Disneyland Hong Kong
10	670443403	5	2019-4	United States	disneyland never cease amaze disneyland florid...	Disneyland Hong Kong
11	670435886	5	2019-4	Canada	spent day grown kid admit great time seems kid...	Disneyland Hong Kong
12	670376905	4	2019-4	Australia	spend two day second day went early went strai...	Disneyland Hong Kong
13	670324965	5	2019-4	Philippines	indeed happiest place earth family really fun ...	Disneyland Hong Kong
14	670274554	5	2018-9	Australia	place huge definately need one day 3 child age...	Disneyland Hong Kong
15	670205135	3	2019-1	United Kingdom	brought ticket left got 2 day le price 1 visit...	Disneyland Hong Kong
16	670199487	4	2019-4	Myanmar (Burma)	huge enough visit one day 2 day pas scene amaz...	Disneyland Hong Kong
17	670129921	3	2019-4	United Kingdom	around 60 per person want eat drink point cost...	Disneyland Hong Kong
18	670099231	4	2019-4	Australia	disneyland need reviewing place speaks however...	Disneyland Hong Kong
19	670033848	5	2018-11	Hong Kong	nothing say except become child step inside di...	Disneyland Hong Kong

	reviewer	Rating	written_date	title	Review_Text	Branch
0	Kelly B	2	May 30, 2021	Universal is a complete Disaster - stick with ...	went universal memorial day weekend total trai...	Universal Studios Florida
1	Jon	1	May 30, 2021	Food is hard to get.	food service horrible reviewing food wait time...	Universal Studios Florida
2	Nerdy P	2	May 30, 2021	Disappointed	booked vacation mainly ride hagrid motorcycle ...	Universal Studios Florida
3	ran101278	4	May 29, 2021	My opinion	person try test seat ride get green light go l...	Universal Studios Florida
4	tammies20132015	5	May 28, 2021	The Bourne Stuntacular...MUST SEE	ok stress enough anyone everyone go universal ...	Universal Studios Florida
5	John	1	May 28, 2021	This is not a vacation	worst experience ever ride outdated whole plac...	Universal Studios Florida
6	annapN7702ZW	2	May 27, 2021	Expected More	expected alot waiting around lack staffing pri...	Universal Studios Florida
7	Deb P	2	May 27, 2021	Disapointing.....	4th trip daughter universal unfortunately disa...	Universal Studios Florida
8	Chuck N	1	May 27, 2021	Greed makes for a terrible guest experience	universal one thing disney everything disney w...	Universal Studios Florida
9	Jen	4	May 26, 2021	Good first time visit with kids	spent 6 night site sapphire fall family 6 2 ad...	Universal Studios Florida
10	Paul S	1	May 26, 2021	Same old Orlando experience.	literally standing line hagrid thing scheduled...	Universal Studios Florida
11	Mandee L	5	May 25, 2021	Family Vacation Fun	universal adventure park best two went impress...	Universal Studios Florida
12	Kate Z	2	May 25, 2021	Crowded, unhelpful staff, and difficult to get...	disney tell much returned one day visit univer...	Universal Studios Florida
13	Kimberly T	1	May 24, 2021	Parking and Guest Services TERRIBLE!!!!!!!!!	went city walk due quadriplegic son nothing ch...	Universal Studios Florida
14	Nancy C	5	May 24, 2021	Excellent Vacation	age ride restaurant lot photo opportunity fant...	Universal Studios Florida
15	SunilnRiss	5	May 22, 2021	Stay at one of 3 resorts that offer EXPRESS PA...	park always awesome vet pretty crowded harry p...	Universal Studios Florida
16	Natasha L	5	May 19, 2021	Despite the Circumstances	humid windy wear mask universal still disappoi...	Universal Studios Florida
17	khali m	1	May 19, 2021	AWFUL	rude worker yelled ton people multiple ride br...	Universal Studios Florida
18	Linkstrips	4	May 19, 2021	Over priced	went park wednesday got early early pas get ah...	Universal Studios Florida
19	Aleisha523	1	May 19, 2021	Don't buy a package with one of their partner ...	first visit universal disney passholders year ...	Universal Studios Florida

Transforming our Datasets

- Drop any unique columns in either dataset
- Split any date-related columns in “year” and “month” columns (and remove null values)
- Create a “quarter” based on the “month” column
- Remove duplicated rows or rows with reviews that were duplicated
- Generate our facts

	Review_Text	Rating	Month	Quarter	Year	Branch	Monthly % of Positive Reviews	Monthly % of Mixed Reviews	Monthly % of Negative Reviews
0	ever disneyland anywhere find disneyland hong ...	4	4	2	2019	Disneyland Hong Kong	78.688525	18.032787	3.278689
1	thanks god hot humid visiting park otherwise w...	4	4	2	2019	Disneyland Hong Kong	78.688525	18.032787	3.278689
2	hk disneyland great compact park unfortunately...	4	4	2	2019	Disneyland Hong Kong	78.688525	18.032787	3.278689
3	location city took around 1 hour kowlon kid li...	4	4	2	2019	Disneyland Hong Kong	78.688525	18.032787	3.278689
4	disney world disneyland anaheim tokyo disneyla...	3	4	2	2019	Disneyland Hong Kong	78.688525	18.032787	3.278689

	Review_Text	Rating	Month	Quarter	Year	Branch	Monthly % of Positive Reviews	Monthly % of Mixed Reviews	Monthly % of Negative Reviews
0	went universal memorial day weekend total trai...	2	5	2	2021	Universal Studios Florida	36.734694	8.163265	55.102041
1	food service horrible reviewing food wait time...	1	5	2	2021	Universal Studios Florida	36.734694	8.163265	55.102041
2	booked vacation mainly ride hagrid motorcycle ...	2	5	2	2021	Universal Studios Florida	36.734694	8.163265	55.102041
3	person try test seat ride get green light go l...	4	5	2	2021	Universal Studios Florida	36.734694	8.163265	55.102041
4	ok stress enough anyone everyone go universal ...	5	5	2	2021	Universal Studios Florida	36.734694	8.163265	55.102041

Transforming our Datasets

- **Dataset Integration (merged Disneyland Reviews dataset with Universal Studios Reviews dataset)**

	Review_Text	Rating	Month	Quarter	Year	Branch	Monthly % of Positive Reviews	Monthly % of Mixed Reviews	Monthly % of Negative Reviews
0	ever disneyland anywhere find disneyland hong ...	4	4	2	2019	Disneyland Hong Kong	78.688525	18.032787	3.278689
1	thanks god hot humid visiting park otherwise w...	4	4	2	2019	Disneyland Hong Kong	78.688525	18.032787	3.278689
2	hk disneyland great compact park unfortunately...	4	4	2	2019	Disneyland Hong Kong	78.688525	18.032787	3.278689
3	location city took around 1 hour kowlon kid li...	4	4	2	2019	Disneyland Hong Kong	78.688525	18.032787	3.278689
4	disney world disneyland anaheim tokyo disneyla...	3	4	2	2019	Disneyland Hong Kong	78.688525	18.032787	3.278689
...	...	...	...	...	...	...	...	...	...
90875	first visit universal studio theme park went p...	4	3	1	2010	Universal Studios Singapore	100.000000	0.000000	0.000000
90876	finally visited singapore first theme park uni...	4	3	1	2010	Universal Studios Singapore	100.000000	0.000000	0.000000
90877	visited first week soft opening unfortunately ...	4	3	1	2010	Universal Studios Singapore	100.000000	0.000000	0.000000
90878	visited 3rd day soft opening ticket sale limit...	4	3	1	2010	Universal Studios Singapore	100.000000	0.000000	0.000000
90879	group managed get ticket 16 february 2010 snea...	5	2	1	2010	Universal Studios Singapore	100.000000	0.000000	0.000000

90880 rows × 9 columns

Extract & Integrate Park Attendance Dataset

```
[ ] # Loading the data
    # (We manually compiled this data ourselves with the goal of using the data
    # for data for this project, so it is imported clean).
    url='https://drive.google.com/file/d/1m3_DJbGPaYLTtgtfVB0k9Mz5GY23toAH/view?usp=sharing'
    url='https://drive.google.com/uc?id=' + url.split('/')[-2]
    df_att = pd.read_csv(url, encoding='UTF-8')

    df_att
```

	Branch	2006 Attendance (millions)	2007 Attendance (millions)	2008 Attendance (millions)	2009 Attendance (millions)	2010 Attendance (millions)	2011 Attendance (millions)	2012 Attendance (millions)	2013 Attendance (millions)	2014 Attendance (millions)	2015 Attendance (millions)	2016 Attendance (millions)	2017 Attendance (millions)	2018 Attendance (millions)	2019 Attendance (millions)	2020 Attendance (millions)	2021 Attendance (millions)
0	Disneyland California	14.73	14.870	14.720	15.90	15.980	16.140	15.960	16.200	16.770	18.280	17.940	18.300	18.660	18.660	3.670	8.500
1	Disneyland Paris	10.60	12.000	12.688	12.74	10.500	10.990	11.200	10.430	9.940	10.360	8.400	9.660	9.843	9.745	2.620	3.500
2	Disneyland Hong Kong	5.20	4.150	4.500	4.60	5.200	5.900	6.700	7.400	7.600	6.800	6.100	6.200	6.700	5.695	1.700	2.600
3	Universal Studios Japan	8.50	8.713	8.300	8.00	8.160	8.500	9.700	10.100	11.800	13.900	14.500	14.935	14.300	14.500	4.901	5.500
4	Universal Studios Florida	6.00	6.200	6.231	5.53	5.925	6.044	6.195	7.062	8.263	9.585	9.998	10.198	10.708	10.922	4.096	8.987
5	Universal Studios Singapore	NaN	NaN	NaN	NaN	2.000	3.411	3.480	3.650	3.840	4.200	4.100	4.220	4.400	4.500	1.098	1.200

Load

	Review_Text	Rating	Month	Quarter	Year	Branch	Monthly % of Positive Reviews	Monthly % of Mixed Reviews	Monthly % of Negative Reviews	2006 Attendance (millions)	...	2012 Attendance (millions)	2013 Attendance (millions)	2014 Attendance (millions)	2015 Attendance (millions)	2016 Attendance (millions)	2017 Attendance (millions)	2018 Attendance (millions)	2019 Attendance (millions)	2020 Attendance (millions)	2021 Attendance (millions)
0	ever disneyland anywhere find disneyland hong ...	4	4	2	2019	Disneyland Hong Kong	78.688525	18.032787	3.278689	5.2	...	6.70	7.40	7.60	6.8	6.1	6.20	6.7	5.695	1.700	2.6
1	thanks god hot humid visiting park otherwise w...	4	4	2	2019	Disneyland Hong Kong	78.688525	18.032787	3.278689	5.2	...	6.70	7.40	7.60	6.8	6.1	6.20	6.7	5.695	1.700	2.6
2	hk disneyland great compact park unfortunately...	4	4	2	2019	Disneyland Hong Kong	78.688525	18.032787	3.278689	5.2	...	6.70	7.40	7.60	6.8	6.1	6.20	6.7	5.695	1.700	2.6
3	location city took around 1 hour kowlon kid li...	4	4	2	2019	Disneyland Hong Kong	78.688525	18.032787	3.278689	5.2	...	6.70	7.40	7.60	6.8	6.1	6.20	6.7	5.695	1.700	2.6
4	disney world disneyland anaheim tokyo disneyla...	3	4	2	2019	Disneyland Hong Kong	78.688525	18.032787	3.278689	5.2	...	6.70	7.40	7.60	6.8	6.1	6.20	6.7	5.695	1.700	2.6
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
90875	first visit universal studio theme park went p...	4	3	1	2010	Universal Studios Singapore	100.000000	0.000000	0.000000	NaN	...	3.48	3.65	3.84	4.2	4.1	4.22	4.4	4.500	1.098	1.2
90876	finally visited singapore first theme park uni...	4	3	1	2010	Universal Studios Singapore	100.000000	0.000000	0.000000	NaN	...	3.48	3.65	3.84	4.2	4.1	4.22	4.4	4.500	1.098	1.2
90877	visited first week soft opening unfortunately ...	4	3	1	2010	Universal Studios Singapore	100.000000	0.000000	0.000000	NaN	...	3.48	3.65	3.84	4.2	4.1	4.22	4.4	4.500	1.098	1.2
90878	visited 3rd day soft opening ticket sale limit...	4	3	1	2010	Universal Studios Singapore	100.000000	0.000000	0.000000	NaN	...	3.48	3.65	3.84	4.2	4.1	4.22	4.4	4.500	1.098	1.2
90879	group managed get ticket 16 february 2010 snea...	5	2	1	2010	Universal Studios Singapore	100.000000	0.000000	0.000000	NaN	...	3.48	3.65	3.84	4.2	4.1	4.22	4.4	4.500	1.098	1.2

90880 rows × 25 columns

Load using PostgreSQL

Load

Branch Dimension

branch_key [PK] integer	branch_name character varying (255)	branch_attendance_key integer
1	Disneyland California	4
2	Disneyland Hong Kong	6
3	Disneyland Paris	2
4	Universal Studios Florida	5
5	Universal Studios Japan	3
6	Universal Studios Singapore	1

Date Dimension

date_key [PK] integer	month integer	quarter integer	year integer
1	4	2	2012
2	7	3	2009
3	5	2	2005
4	5	2	2006
5	3	1	2011
6	1	1	2016
7	4	2	2021
8	7	3	2004
9	9	3	2004

Branch Outrigger Dimension

branch_attendance_key [PK] integer	attendance_millions_2006 numeric	attendance_millions_2007 numeric	attendance_millions_2008 numeric	attendance_millions_2009 numeric	attendance_millions_2010 numeric	attendance_millions_2011 numeric	attendance_millions_2012 numeric	attendance_millions_2013 numeric
1	[null]	[null]	[null]	[null]	2.0	3.411	3.48	3.65
2	10.6	12.0	12.688	12.74	10.5	10.99	11.2	10.43
3	8.5	8.713	8.3	8.0	8.16	8.5	9.7	10.1
4	14.73	14.87	14.72	15.9	15.98	16.14	15.96	16.2
5	6.0	6.2	6.231	5.53	5.925	6.044	6.195	7.062
6	5.2	4.15	4.5	4.6	5.2	5.9	6.7	7.4

Load

Review Text Dimension

review_text_key [PK] integer	review_text text
1	daughter spent sunny hot day disneyland hong kong overall good california version certainly fun day think older kid teenager find park little tame timid folk enjoy park liked fact separate line english
2	location infront mtr station time required cover place 1 full dayits possible cover entire place event show one day still better go early cover almost everything really fabulous place age people went 1
3	spent day disneyland great time family friendly good age big park good variety ride
4	recent tour hong kong limited time choose different tourist attraction finally zeroed disney land bought ticket senior citizen apprehension whether would find enjoyable experience turned quite amaz
5	hong kong young kid place go fun obviously piece america asia well organised plenty keep kid interested really need two day experience whole place make sure catch paint night parade firework
6	like disneyland must nut let kid adult plenty option age group keep engage throughout day previously visited anaheim disneyland california small compared one hong kong still plenty option loved sp

Fact Table

surrogate_key integer	date_key [PK] integer	branch_key [PK] integer	review_text_key [PK] integer	monthly_percent_positive_reviews numeric	monthly_percent_mixed_reviews numeric	monthly_percent_negative_reviews numeric
10307	1	1	7581	94.85714285714286	2.857142857142857	2.2857142857142856
82461	1	1	7762	94.85714285714286	2.857142857142857	2.2857142857142856
83439	1	1	7908	94.85714285714286	2.857142857142857	2.2857142857142856
47327	1	1	7914	94.85714285714286	2.857142857142857	2.2857142857142856
50142	1	1	8139	94.85714285714286	2.857142857142857	2.2857142857142856
19068	1	1	8475	94.85714285714286	2.857142857142857	2.2857142857142856
2903	1	1	9227	94.85714285714286	2.857142857142857	2.2857142857142856
2981	1	1	9341	94.85714285714286	2.857142857142857	2.2857142857142856
17830	1	1	9396	94.85714285714286	2.857142857142857	2.2857142857142856
47250	1	1	9502	94.85714285714286	2.857142857142857	2.2857142857142856
40457	1	1	9552	94.85714285714286	2.857142857142857	2.2857142857142856

03.

Visualization

Slicers

Branch

All

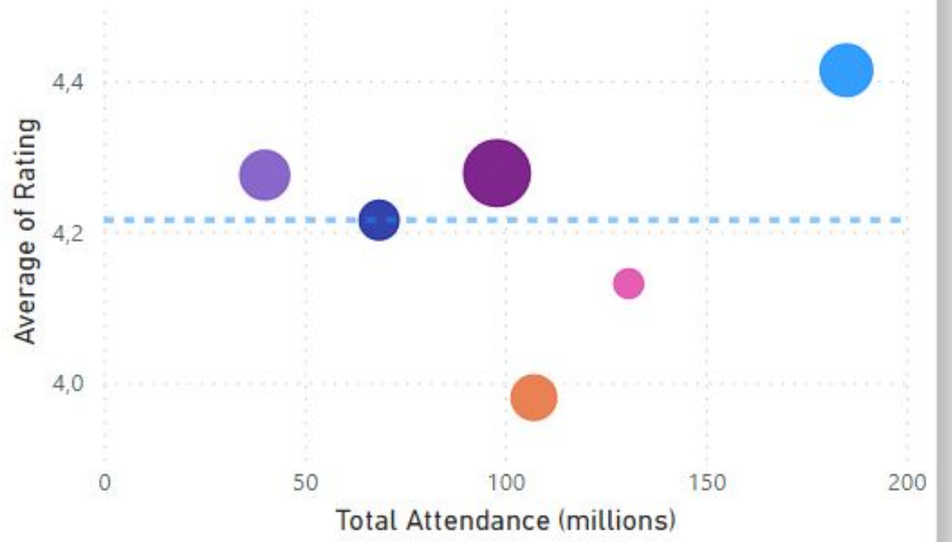
Rating

1 5

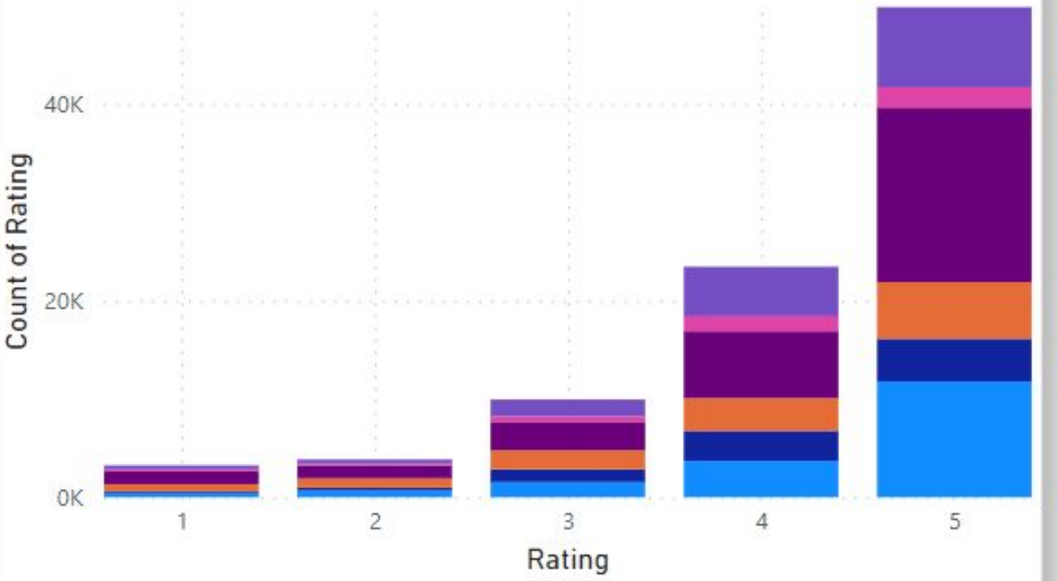
Year, Quarter, Month

All

Correlation between Count of Rating (size of circle) and Attendance from 2010 to 2020



Count of Rating by X-star Rating

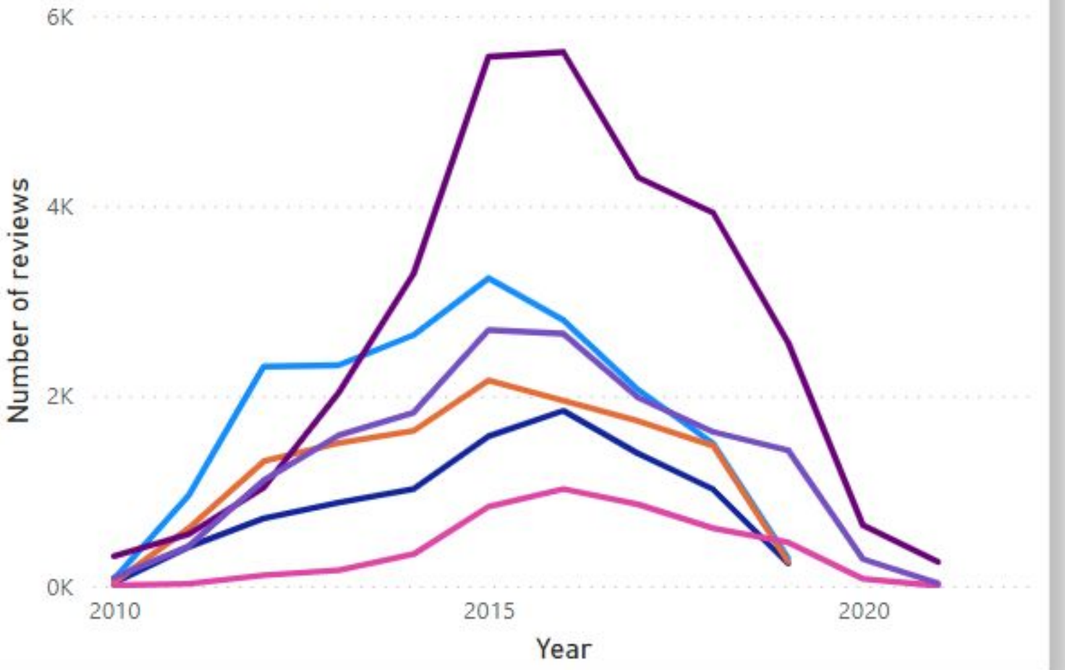


Colour Legend

Branch

- Disneyland California
- Disneyland Hong Kong
- Disneyland Paris
- Universal Studios Florida
- Universal Studios Japan
- Universal Studios Singapore

Count of Rating by Year and Branch



Word cloud

Branch

All

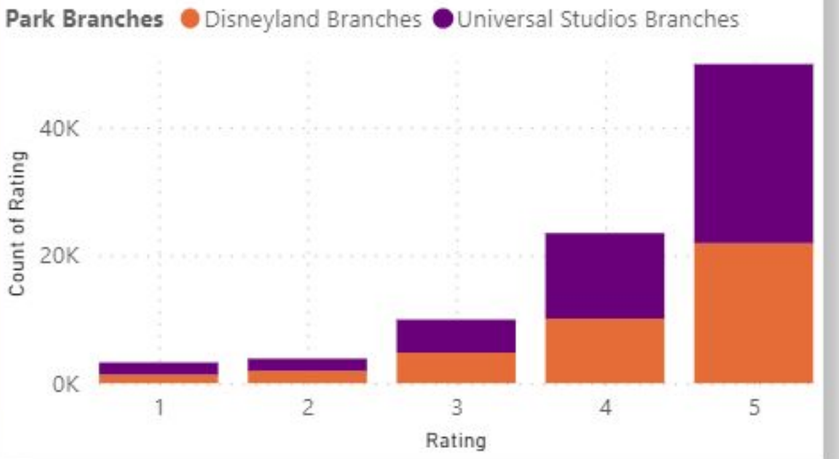
Rating

1 5

Year, Quarter, ...

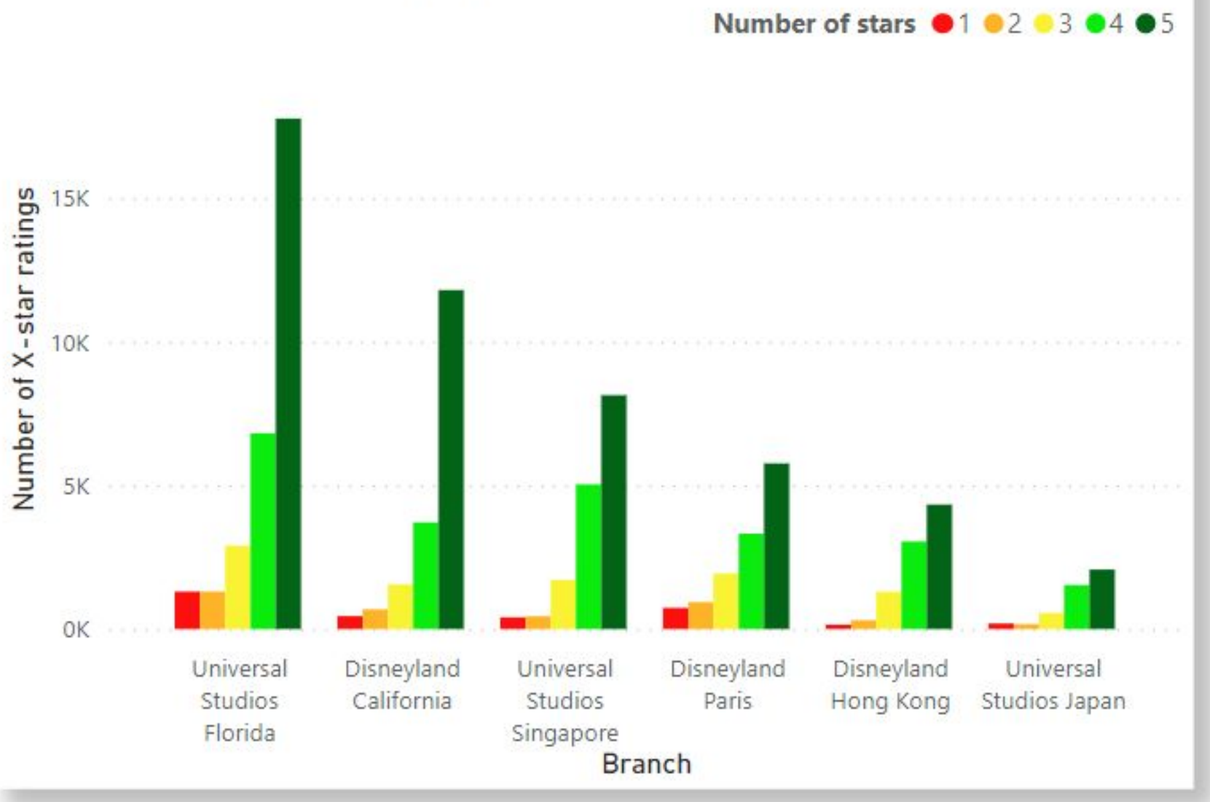
All

Count of Rating by Rating and Park Branches

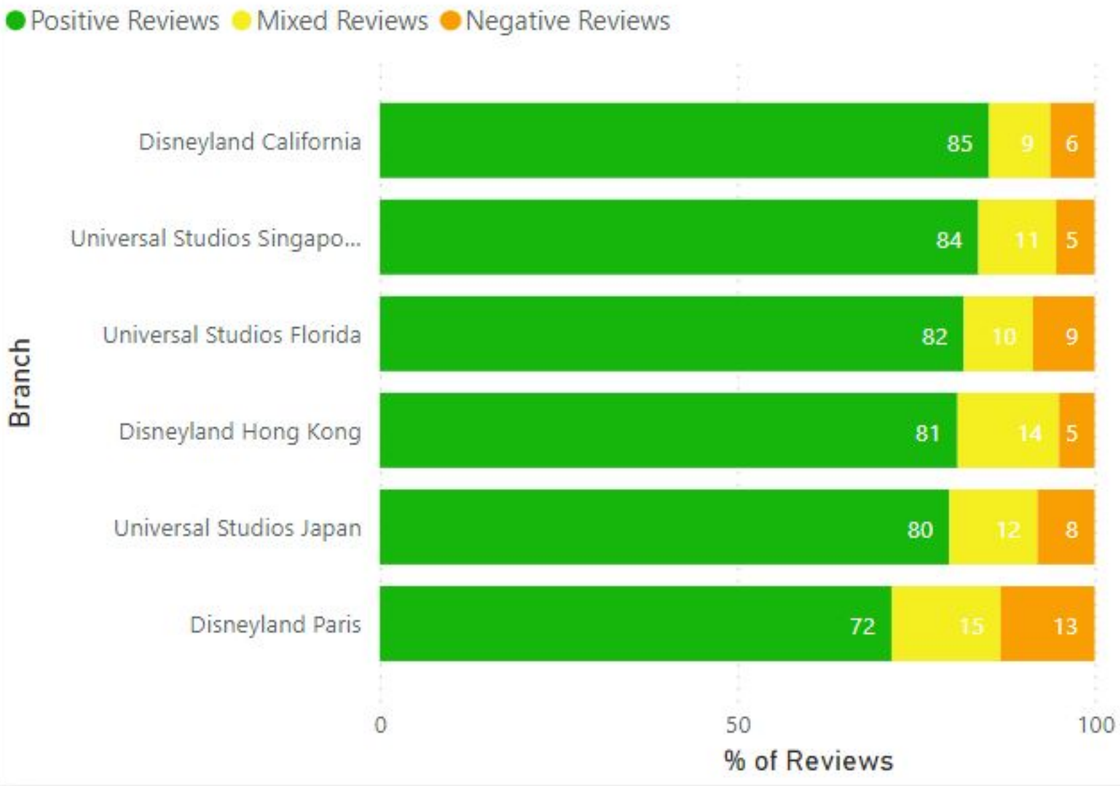


Average of Rating by Branch

Distribution of X-star Ratings by Branch



Positive Reviews, Mixed Reviews and Negative Reviews by Branch



04.

Modelling & Data Mining

Our Models

(The models selected – none of which were NLP models – were requested of us)

- Decision Tree
- Random Forest
- Gradient Boosting

Considerations When Pre-processing Review Texts for the Models

Embedding Words to Vectors

- Models do not accept strings
  (we can one-hot encode categorical data)
- Potential Methods: TF-IDF, CountVectorizer, word2vec/FastText

Typos or Rare Words

- These tokens will not count-vectorize correctly
- Will have lots of null vectors in word2vec

Stemming or Lemmatization?

- We initially stemmed the review texts when staging the data, but this may lead to issues with the word embeddings
- More on this later

Training Size

- Lemmatizing words could minimize the dataset (30 GB after running TF-IDF)
- We can sample and batch to train on a smaller dataset

Data is Skewed?

- From our visualizations, it appears our data is skewed
- It is not good practice to train models on imbalanced data!

Stemming vs. Lemmatization

- Stemming chops off prefix and suffix
  - Fast but can produce words that are not in the English dictionary

--Word--	--Stem--
Python	python
programmers	programm
often	often
tend	tend
like	like
programming	program
in	in
python	python
because	becaus
its	it
like	like
english	english
We	we
call	call
people	peopl
who	who
program	program
in	in
python	python
pythonistas	pythonista

- Lemmatization returns the base root of the word (lemma)
- Words are meaningful in our problem, so we chose to re-stage our data by using lemmatization instead of stemming before training the models on them.

--Word--	--Lemma--
Python	Python
programmers	programmers
often	often
tend	tend
like	like
programming	program
in	in
python	python
because	because
its	its
like	like
english	english
We	We
call	call
people	people
who	who
program	program
in	in
python	python
pythonistas	pythonistas

https://www.datacamp.com/tutorial/stemming-lemmatization-python

Results

Pre-processing steps for these models (None of the models are NLP models)

- Removed all reviews from before 2010 (these models do not accept null data – there is no attendance data for Universal Studios Singapore from before 2010, unlike other branches).
- Standardized all attendance data.
- One-hot encoded all branch names (since they are categorical and these models do not accept strings).
- Convert all review texts to tf-idf vectors (since these models do not accept strings).

Results When Predicting Ratings Given a Review Text (WITHOUT SAMPLING)

Decision Tree

```
Accuracy: 0.4844
Precision: 0.4785
Recall: 0.4844

Time taken to construct
model: 4.3144 mins
```

Random Forest

```
Accuracy: 0.5664
Precision: 0.4803
Recall: 0.5664

Time taken to construct
model: 5.9500 mins
```

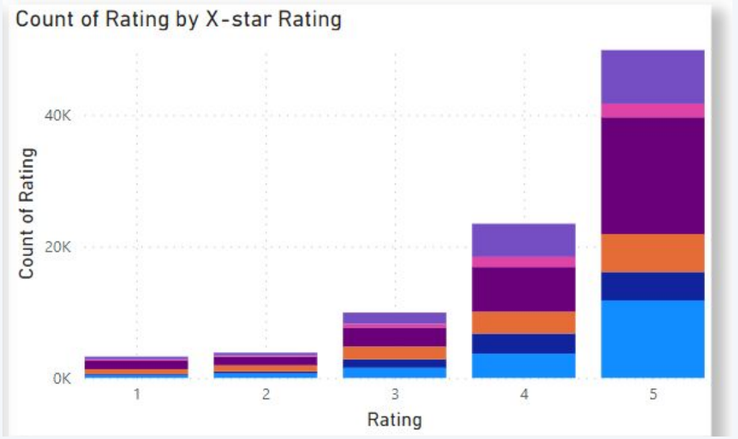
Gradient Boosting

```
Accuracy: 0.5986
Precision: 0.5410
Recall: 0.5986

Time taken to construct
model: 18.7993 hours
```

Results

- From our visualizations, we noticed that the reviews were heavily skewed toward being rated as positive (mainly 5 stars).
- We used **over and undersampling** to help combat this imbalance in our data.
- We achieved worse results, but they are more indicative of the models' actual capabilities.



Results When Predicting Ratings Given a Review Text (WITH OVER & UNDER SAMPLING)

Decision Tree

Accuracy: 0.2732
Precision: 0.4947
Recall: 0.2732

Time taken to construct model: 4.3507 mins

Random Forest

Accuracy: 0.2744
Precision: 0.7309
Recall: 0.2744

Time taken to construct model: 4.3279 mins

Gradient Boosting

Accuracy: 0.2875
Precision: 0.6381
Recall: 0.2875

Time taken to construct model: 17.7269 hours

05.

Next Steps

(Beyond the Scope of this Project)

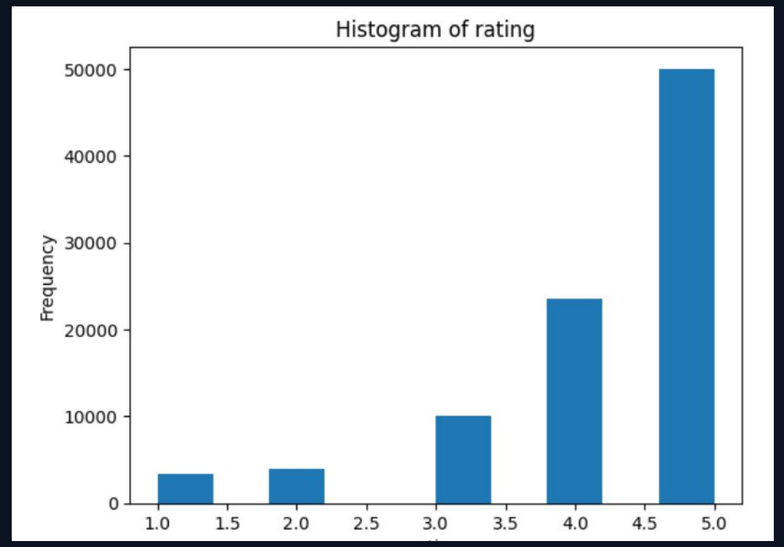
Try Different Sampling Techniques

Downsampling

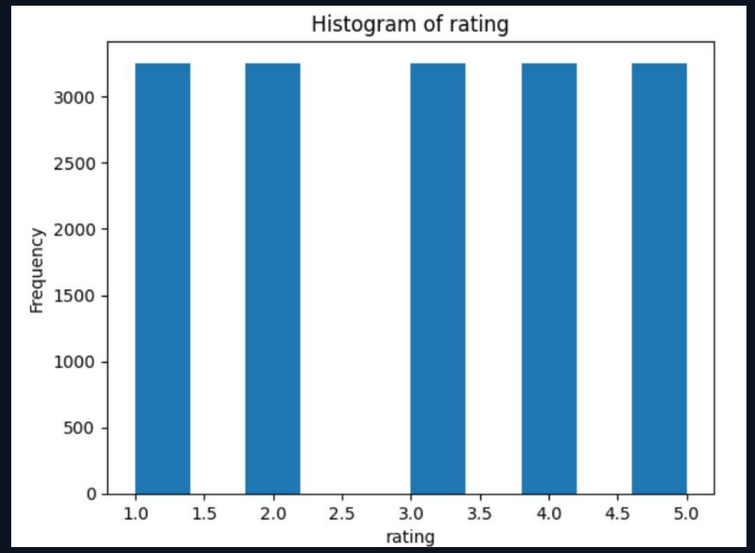
(randomly removing some of the samples from the majority class to create a more balanced dataset)

```
from sklearn.utils import resample
num_to_sample = 3000
df5 = resample(df[df['rating']==5],
               replace=True,
               n_samples=num_to_sample,
               random_state=42)
df4 = resample(df[df['rating']==4],
               replace=True,
               n_samples=num_to_sample,
               random_state=42)
df3 = resample(df[df['rating']==3],
               replace=True,
               n_samples=num_to_sample,
               random_state=42)
df2 = resample(df[df['rating']==2],
               replace=True,
               n_samples=num_to_sample,
               random_state=42)
df1 = resample(df[df['rating']==1],
               replace=True,
               n_samples=num_to_sample,
               random_state=42)
```

Before



After



downsampled 2.ipynb - Colaboratory (google.com)

Try Different Models

A work-in-progress CNN

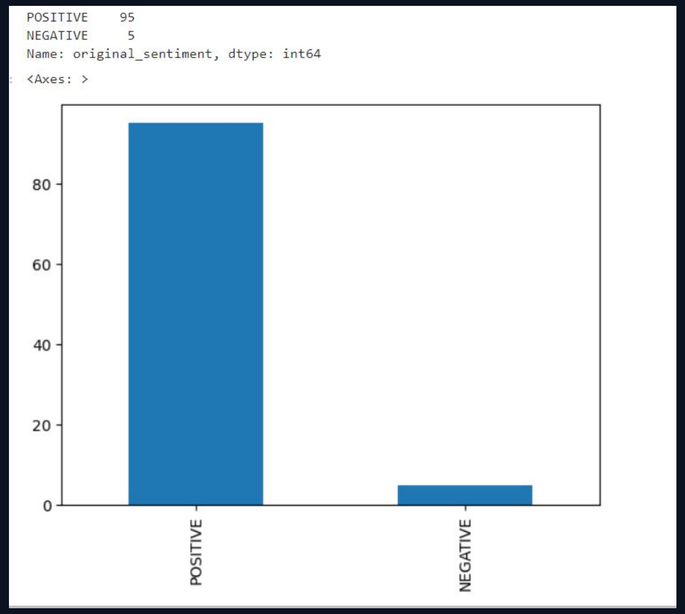
- Attempted a deep neural network approach to perform sentiment analysis
- Chose Convolutional Neural Network as it was the simplest type
- Used CountVectorizer to embed words into vectors
- Multi-class prediction problem
- Difficulty with training time: 4 hours+

```
# Define hyperparameters
vocab_size = 10000
embedding_dim = 100
num_filters = 100
filter_sizes = [3, 4, 5]
hidden_dim = 256
output_dim = 5
dropout = 0.5
num_epochs =5
```

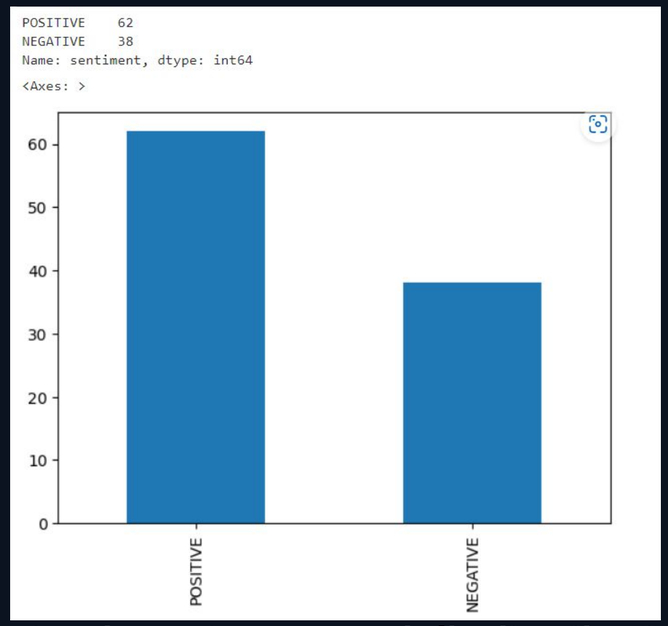
	precision	recall	f1-score	support
0	0.59	0.24	0.34	1388
1	0.01	0.71	0.02	7
2	0.03	0.14	0.06	143
3	0.00	0.00	0.00	2
4	0.64	0.27	0.38	1460
accuracy			0.25	3000
macro avg	0.25	0.27	[illegible]	3000
weighted avg	0.58	0.25	0.35	3000

neuralnetwork.ipynb - Colaboratory (google.com)

Sentiment Analysis



Reviewer's distribution



Pretrained sentiment model's distribution

Plot of sentiment to rating

	precision	recall	f1-score	support
NEGATIVE	0.13	1.00	0.23	5
POSITIVE	1.00	0.65	0.79	95
accuracy			0.67	100
macro avg	0.57	0.83	0.51	100
weighted avg	0.96	0.67	0.76	100

This means that the people rate higher than what they say with their words suggest.

(Performed on 100 data points due to time constraint running the Hugging Face model on 15k rows)

06.

Recommendation on How to Improve the Review System

Instead of assigning one 5 star rating, assign ratings to different categories!

Food

Rides

Fireworks

Service

Cleanliness

Allows for more nuanced, fine-grained analysis such as topical analysis.
All of the reviews should no longer be skewed to just being “5 stars.”

Thank you!