Disney_EDA_Report

December 29, 2021

[24]:
```
import numpy as np
import pandas as pd
import altair as alt
from pandas_profiling import ProfileReport
from sklearn.model_selection import train_test_split
import string
import nltk
from nltk.stem import PorterStemmer, WordNetLemmatizer
from nltk.corpus import wordnet
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from wordcloud import WordCloud, STOPWORDS
import tkinter
import matplotlib
import matplotlib.pyplot as plt
from PIL import Image
from textblob import TextBlob

alt.renderers.enable('mimetype')
alt.data_transformers.disable_max_rows()
matplotlib.use('TkAgg')
```

[2]:
```
disney_df = pd.read_csv("../data/raw/DisneylandReviews.csv")
disney_df
```

[2]:

	Review_ID	Rating	Year_Month	Reviewer_Location	\
0	670772142	4	2019-4	Australia	
1	670682799	4	2019-5	Philippines	
2	670623270	4	2019-4	United Arab Emirates	
3	670607911	4	2019-4	Australia	
4	670607296	4	2019-4	United Kingdom	
…	…	…	…	…	
42651	1765031	5	missing	United Kingdom	
42652	1659553	5	missing	Canada	
42653	1645894	5	missing	South Africa	
42654	1618637	4	missing	United States	
42655	1536786	4	missing	United Kingdom	

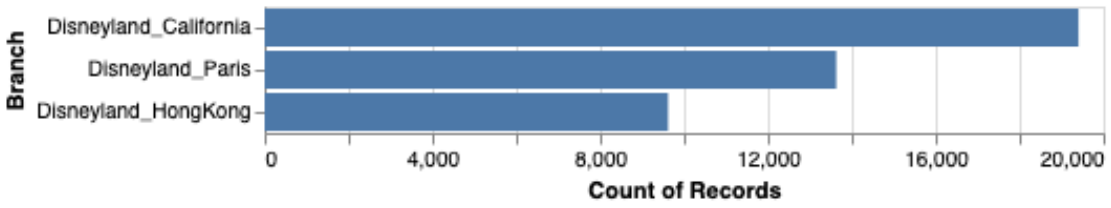
```
                                                Review_Text                 Branch
0       If you've ever been to Disneyland anywhere you…  Disneyland_HongKong
1       Its been a while since d last time we visit HK…  Disneyland_HongKong
2       Thanks God it wasn   t too hot or too humid wh…  Disneyland_HongKong
3       HK Disneyland is a great compact park. Unfortu…  Disneyland_HongKong
4       the location is not in the city, took around 1…  Disneyland_HongKong
…                                                     …                   …
42651   i went to disneyland paris in july 03 and thou…     Disneyland_Paris
42652   2 adults and 1 child of 11 visited Disneyland …     Disneyland_Paris
42653   My eleven year old daughter and myself went to…     Disneyland_Paris
42654   This hotel, part of the Disneyland Paris compl…     Disneyland_Paris
42655   I went to the Disneyparis resort, in 1996, wit…     Disneyland_Paris

[42656 rows x 6 columns]
```

[3]:
```
disney_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 42656 entries, 0 to 42655
Data columns (total 6 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Review_ID          42656 non-null  int64
 1   Rating             42656 non-null  int64
 2   Year_Month         42656 non-null  object
 3   Reviewer_Location  42656 non-null  object
 4   Review_Text        42656 non-null  object
 5   Branch             42656 non-null  object
dtypes: int64(2), object(4)
memory usage: 2.0+ MB
```

[4]:
```
alt.Chart(disney_df).mark_bar().encode(
    x='count()',
    y=alt.Y('Branch', sort='-x'))
```

[4]:



[5]:
```
disney_df['Review_Text']
```

[5]:
```
0        If you've ever been to Disneyland anywhere you…
1        Its been a while since d last time we visit HK…
2        Thanks God it wasn   t too hot or too humid wh…
3        HK Disneyland is a great compact park. Unfortu…
4        the location is not in the city, took around 1…
                              …
42651    i went to disneyland paris in july 03 and thou…
42652    2 adults and 1 child of 11 visited Disneyland …
42653    My eleven year old daughter and myself went to…
42654    This hotel, part of the Disneyland Paris compl…
42655    I went to the Disneyparis resort, in 1996, wit…
Name: Review_Text, Length: 42656, dtype: object
```

0.0.1 Preprocess corpus

[6]:
```
corpus=''
for i in disney_df['Review_Text']:
    corpus+=str(i)
```

[7]:
```
def preprocess(corpus):
    corpus = corpus.lower()
    corpus = corpus.replace('hk', 'hongkong')
    corpus = corpus.replace('hong kong', 'hongkong')
    corpus = corpus.replace("n't", " not")
    corpus = ''.join([i for i in corpus if not i.isdigit()])
    new_corpus =''
    for i in corpus:
        if i not in string.punctuation:
            new_corpus+=i
    return new_corpus
```

[8]:
```
def lemmatize_stem(corpus):
    lemmatizer = WordNetLemmatizer()
    stemmer = PorterStemmer()
    new_corpus = lemmatizer.lemmatize(corpus)
    new_corpus = stemmer.stem(new_corpus)
    return new_corpus
```

[9]:
```
print(stopwords.words("english"))
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're",
"you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he',
'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's",
'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what',
'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those', 'am', 'is',
'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having',
'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or',
'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about',
```

```
'against', 'between', 'into', 'through', 'during', 'before', 'after', 'above',
'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under',
'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why',
'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some',
'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very',
's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now',
'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'couldn',
"couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn',
"hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't",
'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn',
"shouldn't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn',
"wouldn't"]
```

[10]:
```
stopwords_700 = pd.read_csv('stopwords.txt', header= None, delimiter = "\t")
stopwords_700 = set(stopwords_700[0])
```

[11]:
```
def remove_stopwords(corpus):
    words = nltk.word_tokenize(corpus)
    stop_words = set(stopwords.words("english"))
    stop_words_customize = {'disney', 'disneyland', 'land', 'park', 'parks',␣
 ↪'world', 'disneyworld', 'disney world'}
    stop_words_all = stop_words.union(stopwords_700, stop_words_customize)
    new_corpus = [word for word in words if not word in stop_words_all]
    return new_corpus
```

[12]:
```
new_corpus = preprocess(corpus)
new_corpus = lemmatize_stem(new_corpus)
new_corpus = remove_stopwords(new_corpus)
```

[13]:
```
#lemmatizer = WordNetLemmatizer()
#lemmatizer.lemmatize(a)
```

[14]:
```
#stemmer = PorterStemmer()
#stemmer.stem(a)
```

[15]:
```
from collections import Counter
new_disney_df = pd.DataFrame()
new_disney_df['word'] = list(Counter(new_corpus).keys())
new_disney_df['count']= list(Counter(new_corpus).values())
new_disney_df.sort_values(by=['count'], ascending=False).head(20)
```

[15]:

	word	count
10	rides	33898
24	time	28441
15	day	27421
420	ride	17498
38	great	15659
125	food	13985

```
161    kids  13824
83    place  13108
55     like  11868
201     see  11566
25    visit  11257
181  people  10878
214    went  10792
945    long  10086
184    fast  10001
409    pass   9690
402    wait   9565
365    back   9432
272     fun   9265
462   first   8777
```

0.0.2 WordCloud

[16]:
```
micky_mask = np.array(Image.open('mickey.png'))
```

[17]:
```
text = ''
for i in new_corpus:
    text+=' '
    text+=i
```

[18]:
```
wordcloud = WordCloud(stopwords = STOPWORDS,
                      mask = micky_mask,
                      random_state = 123,
                      collocations=False,
                      max_words = 100).generate(text)
```

[19]:
```
wordcloud.to_file('wordcloud.png')
```

[19]:
```
<wordcloud.wordcloud.WordCloud at 0x1595120d0>
```

0.0.3 sentiment polarity

[20]:
```
def clean(Review):
    Review = Review.str.lower()
    Review = Review.replace('hk', 'hongkong', regex = True)
    Review = Review.replace('hong kong', 'hongkong', regex = True)
    Review = Review.replace("n't", " not", regex = True)
    return Review
```

[21]:
```
disney_df['Review_Text'] = clean(disney_df['Review_Text'])
```

[25]:
```
disney_df['polarity'] = disney_df['Review_Text'].map(lambda text:␣
↪TextBlob(text).sentiment.polarity)
```

[26]:
```
disney_df
```

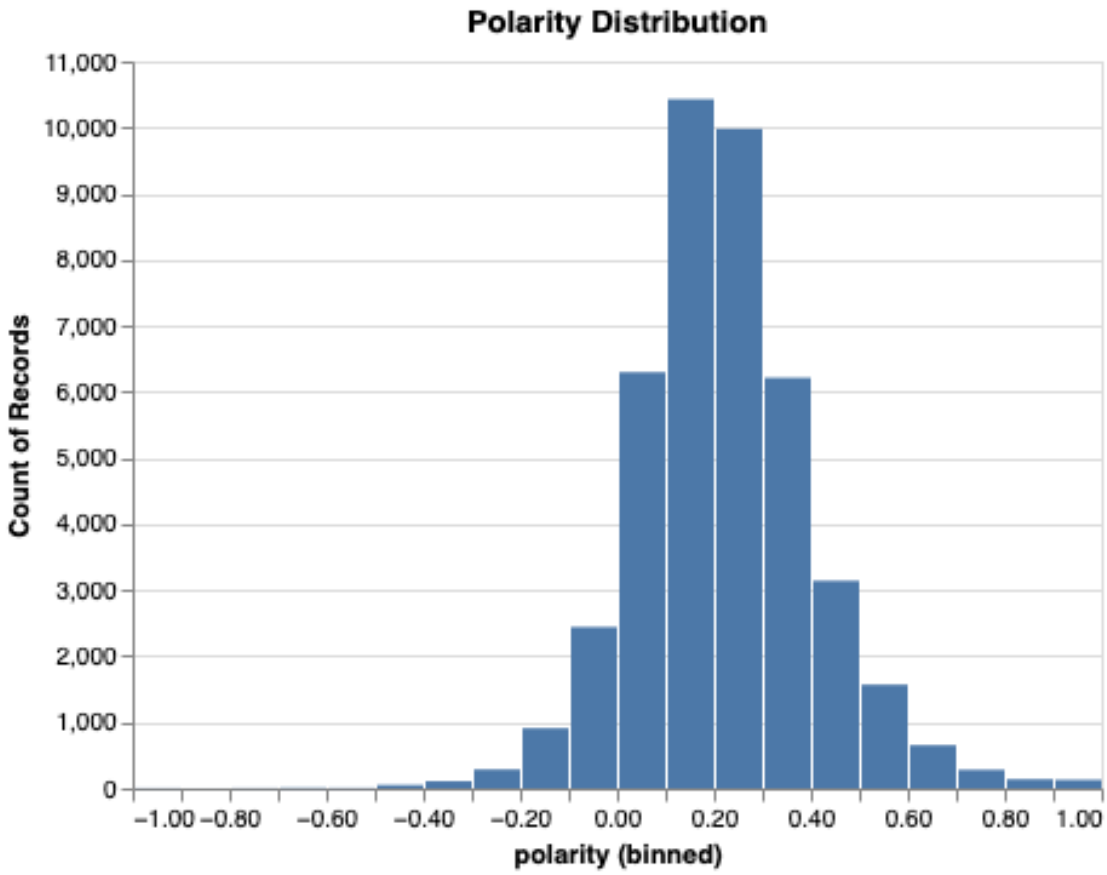
[26]:

	Review_ID	Rating	Year_Month	Reviewer_Location	Review_Text	Branch	polarity
0	670772142	4	2019-4	Australia	if you've ever been to disneyland anywhere you…	Disneyland_HongKong	0.243981
1	670682799	4	2019-5	Philippines	its been a while since d last time we visit ho…	Disneyland_HongKong	0.236131
2	670623270	4	2019-4	United Arab Emirates	thanks god it wasn t too hot or too humid wh…	Disneyland_HongKong	0.160498
3	670607911	4	2019-4	Australia	hongkong disneyland is a great compact park. u…	Disneyland_HongKong	0.189286
4	670607296	4	2019-4	United Kingdom	the location is not in the city, took around 1…	Disneyland_HongKong	0.266667
…	…	…	…	…	…	…	…
42651	1765031	5	missing	United Kingdom	i went to disneyland paris in july 03 and thou…	Disneyland_Paris	0.238636
42652	1659553	5	missing	Canada	2 adults and 1 child of 11 visited disneyland …	Disneyland_Paris	0.179250
42653	1645894	5	missing	South Africa	my eleven year old daughter and myself went to…	Disneyland_Paris	0.153205
42654	1618637	4	missing	United States	this hotel, part of the disneyland paris compl…	Disneyland_Paris	0.265327
42655	1536786	4	missing	United Kingdom	i went to the disneyparis resort, in 1996, wit…	Disneyland_Paris	0.254881

[42656 rows x 7 columns]

[28]:
```
(alt.Chart(disney_df, title ="Polarity Distribution").mark_bar().encode(
    alt.X('polarity', bin=alt.Bin(maxbins=30)),
    y='count()'))
```

[28]:



[39]:
```
print('3 random reviews with the positive sentiment polarity: \n')
pos_review = disney_df.loc[disney_df.polarity > 0, ['Review_Text']].sample(3).
 ↪values
for review in pos_review:
    print(review)
```

```
3 random reviews with the positive sentiment polarity:

['went with my hubby and two children age 12 and 14. we have been to disneyland
orlando and this park is very similar, just not as big, that is not a bad thing.
we had a great time and went on many rides, some more often then others. i
recommend that you pack some snacks to take with you because it is pricey to eat
at the restaurants and cafes in disneyland. and do stay for the fireworks final
show it is better than in orlando.']
["ive been to disneyland paris with my parents and brother 3 4 times and we have
loved every single visit. theres something for everyone and, although the ques
may be long during the school holidays, there is so much to look at the time
goes by very quickly. both parks are really clean and the staff really make your
holiday. the only downside would be that some nationalities tend to push in more
than other. unlike us brits who will just complain about ques others will simply
```

try to slyly push past you. however this is not the fault of disneyland itself. there are plenty of places to eat from lower budget 'fast food' places to more expensive sit down buffetts and waiter service food venues. the food can be quite expensive (but hey you are in disneyland after all) there is a picnic area out side the park on the way from the carpark however if your in a group or travelling as a family it can be quite a walk for little ones, and personally i would save time and effort and but food. the rides are great, food is great, staff are friendly, the parades are definatley not to miss. plently of photo oportunitiesoverall as four adults my brother and myself being 13 and 10 went we first went and 18 and 15 when we last went we have so much fun!disneyland paris is certainly not just a kids holiday!!"]
["even though we knew it would be crowded, the lines are still a surprise. 45 60 minutes wait for a photo with a character or a ride sre the standards. many rides are dated, really have not moved on since last century early last century. giiven the long wait for photos with characters, i'm not sure why they ca not scatter more of the same arounds. do kids really care thst minnie was back over there? and why do some princesses need a day off? they're pretend!all in all an average experience, though the kids did have a good time. guess that sums it up."]

[40]:
```
print('3 random reviews with the negative sentiment polarity: \n')
neg_review = disney_df.loc[disney_df.polarity < 0, ['Review_Text']].sample(3).
 ↪values
for review in neg_review:
    print(review)
```

3 random reviews with the negative sentiment polarity:

['pathetically small and uninspiring. not worth the effort to visit, unless you are tired of shopping in hongkong.']
['there were no fireworks because the castle was undergoing transformation : (since it was already a small version, the transformation of some areas made it much smaller : (']
["it's expensive, it's crowded, it's everything everyone says in these reviews, but hey, it's disneyland!"]

[ ]: