British Airline Sentiment Analysis By Matthew Kusto

This is a project that was done with BA. This is part one of two in the job simulation. This first part utilizes web scrapping (BeautifulSoup), pandas, matplotlib, seaborn, nltk (Natural Language ToolKit), and textblob. This project was meant to scrape the BA website for reviews, import the reviews into a .csv file and filter the words (and in certain reviews, numbers), then graph the hot words as well as the polarity and subjectivity.

Imports

```
from bs4 import BeautifulSoup
import pandas as pd
import numpy as np
import nltk
import matplotlib.pyplot as plt
import seaborn as sns
import requests

base_url = "https://www.airlinequality.com/airline-reviews/british-
airways"
pages = 10
page_size = 100

reviews = []

# for i in range(1, pages + 1):
for i in range(1, pages + 1):

    print(f"Scraping page {i}")

    # Create URL to collect links from paginated data
    url = f"{base_url}/page/{i}/?sortby=post_date
%3ADesc&pagesize={page_size}"

    # Collect HTML data from this page
    response = requests.get(url)

    # Parse content
    content = response.content
    parsed_content = BeautifulSoup(content, 'html.parser')
    for para in parsed_content.find_all("div", {"class":
"text_content"}):
        reviews.append(para.get_text())

    print(f"   ---> {len(reviews)} total reviews")
```

```
Scraping page 1
   ---> 100 total reviews
Scraping page 2
```

```
   ---> 200 total reviews
Scraping page 3
   ---> 300 total reviews
Scraping page 4
   ---> 400 total reviews
Scraping page 5
   ---> 500 total reviews
Scraping page 6
   ---> 600 total reviews
Scraping page 7
   ---> 700 total reviews
Scraping page 8
   ---> 800 total reviews
Scraping page 9
   ---> 900 total reviews
Scraping page 10
   ---> 1000 total reviews
```

```
df = pd.DataFrame()
df["reviews"] = reviews
df.head()
```

```
                                                reviews
0  ✅ Trip Verified |  My daughter and I were deni...
1  ✅ Trip Verified | Despite boarding being the u...
2  Not Verified |  Flight cancelled, no crew! 9th...
3  Not Verified |  The worst service ever, my bag...
4  ✅ Trip Verified |  4/4 flights we booked this ...
```

```
df.to_csv("d:BA_reviews.csv")
```

Analysing Data

```
df = pd.read_csv("D:data/BA_reviews.csv") # Using thumbstick to
df
```

```
                                                reviews
0      The worst service ever, my baggage did not a...
1       4/4 flights we booked this holiday were del...
2       British Airways has a total lack of respect...
3      London Heathrow to Keflavik, Iceland in Busi...
4      Mumbai to London Heathrow in Business Class ...
..                                                   ...
995     Amsterdam to London arrived 33 minutes late...
996     Buenos Aires to London. We flew overnight F...
997     Business Class space is ridiculously narrow...
998     Aberdeen to Heathrow to connect to a flight...
999    I would not recommend this airline. I travel...
```

```
[1000 rows x 1 columns]
```

word count in cell

```
df['word_count'] = df['reviews'].apply(lambda x: len(x.split()))
df
```

```
                                                reviews  word_count
0      The worst service ever, my baggage did not a...          30
1       4/4 flights we booked this holiday were del...          28
2       British Airways has a total lack of respect...         414
3      London Heathrow to Keflavik, Iceland in Busi...         102
4      Mumbai to London Heathrow in Business Class ...         171
..                                                   ...         ...
995     Amsterdam to London arrived 33 minutes late...          76
996     Buenos Aires to London. We flew overnight F...          86
997     Business Class space is ridiculously narrow...          34
998     Aberdeen to Heathrow to connect to a flight...         111
999    I would not recommend this airline. I travel...         267

[1000 rows x 2 columns]
```

char count in cell

```
df['char_count'] = df['reviews'].apply(lambda x: len(x))
df
```

```
                                                reviews  word_count
char_count
0      The worst service ever, my baggage did not a...          30
160
1       4/4 flights we booked this holiday were del...          28
162
2       British Airways has a total lack of respect...         414
2300
3      London Heathrow to Keflavik, Iceland in Busi...         102
594
4      Mumbai to London Heathrow in Business Class ...         171
1003
..                                                   ...         ...
...
995     Amsterdam to London arrived 33 minutes late...          76
405
996     Buenos Aires to London. We flew overnight F...          86
503
997     Business Class space is ridiculously narrow...          34
194
998     Aberdeen to Heathrow to connect to a flight...         111
551
```

```
999    I would not recommend this airline. I travel...          267
1406

[1000 rows x 3 columns]
```

average count in cell

```
# Created a function
def average_words(x):
    words = x.split() # number of words in cell
    return sum(len(word) for word in words) / len(words)

df['avg_word_len'] = df['reviews'].apply(lambda x: average_words(x))
df
```

```
                                                reviews  word_count
0      The worst service ever, my baggage did not a...          30  \
1       4/4 flights we booked this holiday were del...          28
2       British Airways has a total lack of respect...         414
3      London Heathrow to Keflavik, Iceland in Busi...         102
4      Mumbai to London Heathrow in Business Class ...         171
..                                                   ...         ...
995     Amsterdam to London arrived 33 minutes late...          76
996     Buenos Aires to London. We flew overnight F...          86
997     Business Class space is ridiculously narrow...          34
998     Aberdeen to Heathrow to connect to a flight...         111
999    I would not recommend this airline. I travel...         267

     char_count  avg_word_len
0           160      4.266667
1           162      4.714286
2          2300      4.550725
3           594      4.813725
4          1003      4.859649
..          ...           ...
995         405      4.276316
996         503      4.825581
997         194      4.647059
998         551      3.945946
999        1406      4.262172

[1000 rows x 4 columns]
```

stop words

```
from nltk.corpus import stopwords

stop_words = stopwords.words('english')
```

```
df['stopword_count'] = df['reviews'].apply(lambda x: len([word for
word in x.split() if word.lower() in stop_words]))
df
```

```
                                                reviews  word_count
0      The worst service ever, my baggage did not a...          30  \
1       4/4 flights we booked this holiday were del...          28
2       British Airways has a total lack of respect...         414
3      London Heathrow to Keflavik, Iceland in Busi...         102
4      Mumbai to London Heathrow in Business Class ...         171
..                                                   ...         ...
995     Amsterdam to London arrived 33 minutes late...          76
996     Buenos Aires to London. We flew overnight F...          86
997     Business Class space is ridiculously narrow...          34
998     Aberdeen to Heathrow to connect to a flight...         111
999    I would not recommend this airline. I travel...         267

     char_count  avg_word_len  stopword_count
0           160      4.266667              15
1           162      4.714286              10
2          2300      4.550725             191
3           594      4.813725              44
4          1003      4.859649              69
..          ...           ...             ...
995         405      4.276316              34
996         503      4.825581              39
997         194      4.647059              16
998         551      3.945946              64
999        1406      4.262172             145

[1000 rows x 5 columns]
```

```
df['stopword_rate'] = df['stopword_count'] / df['word_count']
df
```

```
                                                reviews  word_count
0      The worst service ever, my baggage did not a...          30  \
1       4/4 flights we booked this holiday were del...          28
2       British Airways has a total lack of respect...         414
3      London Heathrow to Keflavik, Iceland in Busi...         102
4      Mumbai to London Heathrow in Business Class ...         171
..                                                   ...         ...
995     Amsterdam to London arrived 33 minutes late...          76
996     Buenos Aires to London. We flew overnight F...          86
997     Business Class space is ridiculously narrow...          34
998     Aberdeen to Heathrow to connect to a flight...         111
999    I would not recommend this airline. I travel...         267

     char_count  avg_word_len  stopword_count  stopword_rate
0           160      4.266667              15       0.500000
```

```
1           162      4.714286                10       0.357143
2          2300      4.550725               191       0.461353
3           594      4.813725                44       0.431373
4          1003      4.859649                69       0.403509
..          ...           ...               ...            ...
995         405      4.276316                34       0.447368
996         503      4.825581                39       0.453488
997         194      4.647059                16       0.470588
998         551      3.945946                64       0.576577
999        1406      4.262172               145       0.543071

[1000 rows x 6 columns]
```

```
df.sort_values(by='stopword_rate')
```

```
                                                  reviews  word_count
182     Good lounge at Cape Town. On time departure...          23  \
266     Full afternoon flight. Ready to fly on time...          39
35      Check-in Desk rude and dismissive. Flight l...          35
942     Gatwick to Fort Lauderdale. Charging to cho...          34
284    Routine typical BA domestic shuttle flight. ...          42
..                                                 ...         ...
743    No problems at the airport in Vienna, it was...          88
431     After 1h queuing at the check-in desk, I am...          27
882     San Francisco to London. After paying £4000...          42
500    I entered the plane with a bunch of white fl...          63
590     Gatwick to Cancun. Flight was late. My food...          43

     char_count  avg_word_len  stopword_count  stopword_rate
182         154      5.608696               4       0.173913
266         241      5.128205               7       0.179487
35          245      5.942857               8       0.228571
942         207      5.029412               8       0.235294
284         250      4.928571              10       0.238095
..          ...           ...             ...            ...
743         472      4.352273              52       0.590909
431         133      3.851852              16       0.592593
882         227      4.333333              25       0.595238
500         330      4.222222              38       0.603175
590         216      3.976744              27       0.627907

[1000 rows x 6 columns]
```

```
df.describe()
```

```
        word_count   char_count  avg_word_len  stopword_count
stopword_rate
count  1000.000000  1000.000000   1000.000000     1000.000000
1000.000000
mean    144.739000   797.573000      4.561031       66.866000
```

```
0.449125
std     102.650251   554.291174      0.334865       50.743975
0.062763
min      23.000000   132.000000      3.677083        4.000000
0.173913
25%      74.000000   416.750000      4.339289       32.000000
0.415873
50%     117.000000   649.500000      4.526825       53.000000
0.457729
75%     183.250000   996.250000      4.736573       86.000000
0.490171
max     654.000000  3466.000000      6.457143      347.000000
0.627907
```

Data Cleaning

lower casing every word

```
df['reviews']

0         The worst service ever, my baggage did not a...
1          4/4 flights we booked this holiday were del...
2          British Airways has a total lack of respect...
3         London Heathrow to Keflavik, Iceland in Busi...
4         Mumbai to London Heathrow in Business Class ...
                              ...
995        Amsterdam to London arrived 33 minutes late...
996        Buenos Aires to London. We flew overnight F...
997        Business Class space is ridiculously narrow...
998        Aberdeen to Heathrow to connect to a flight...
999       I would not recommend this airline. I travel...
Name: reviews, Length: 1000, dtype: object

df['lowercase'] = df['reviews'].apply(lambda x: " ".join(word.lower()
for word in x.split()))
```

removing punctuation

(can combine lowercase and the removal of puncuation... only separate to compare)

```
df['punc_removed'] = df['lowercase'].str.replace(r'[^\w\s]', '',
regex=True)

from nltk.corpus import stopwords

stop_words = stopwords.words('english')

df['stopwords']= df['punc_removed'].apply(lambda x: " ".join(word for
word in x.split() if word not in stop_words))
```

(The bottom cell shows the "hot words" that were used in the reviews)

(this is just a small side note)

```
someStopWords = pd.Series("
".join(df['stopwords']).split()).value_counts()
someStopWords
```

```
flight       1842
ba           1080
service       710
london        584
seat          501
             ...
purport         1
topclass        1
reluctant       1
ruination       1
misplace        1
Name: count, Length: 7848, dtype: int64
```

```
otherStopWords = pd.Series("
".join(df['stopwords']).split()).value_counts()[:-30:-1]
otherStopWords
```

```
misplace                 1
ruination                1
reluctant                1
topclass                 1
purport                  1
fivehour                 1
reconcile                1
cancellingrearranging    1
oxymoron                 1
supposedly               1
midday                   1
sprung                   1
miracle                  1
530am                    1
aug                      1
incurs                   1
crises                   1
savvy                    1
shutting                 1
incapable                1
polenta                  1
canapé                   1
manger                   1
whichever                1
student                  1
brim                     1
```

```
sketch                    1
suffice                   1
lhrmle                    1
Name: count, dtype: int64
```

Cleaning Data

```
df['cleaned_review'] = df['stopwords'].apply(lambda x: " ".join(word
for word in x.split() if word not in otherStopWords))

df['cleaned_review_count'] = df['cleaned_review'].apply(lambda x:
len(x.split()))
df['clean_rate'] = df['cleaned_review_count'] / df['word_count']
```

Lemmatization

```
from textblob import Word

df['lemmatized'] = df['cleaned_review'].apply(lambda x: "
".join(Word(word).lemmatize() for word in x.split()))
```

Sentiment Analysis

```
from textblob import TextBlob
```

(on average)

```
df['polarity']= df['lemmatized'].apply(lambda
x:TextBlob(x).sentiment[0]) #(polarity, subjectivity)

df['subjectivity']= df['lemmatized'].apply(lambda
x:TextBlob(x).sentiment[1]) #(polarity, subjectivity)

# df.drop(['lowercase','punc_removed', 'stopwords',
'cleaned_review','lemmatized'], axis=1, inplace=True)

# df.sort_values(by='polarity')
df.describe()
```

```
        word_count   char_count  avg_word_len  stopword_count
stopword_rate
count  1000.000000  1000.000000   1000.000000     1000.000000
1000.000000  \
mean    144.739000   797.573000      4.561031       66.866000
0.449125
std     102.650251   554.291174      0.334865       50.743975
0.062763
min      23.000000   132.000000      3.677083        4.000000
```

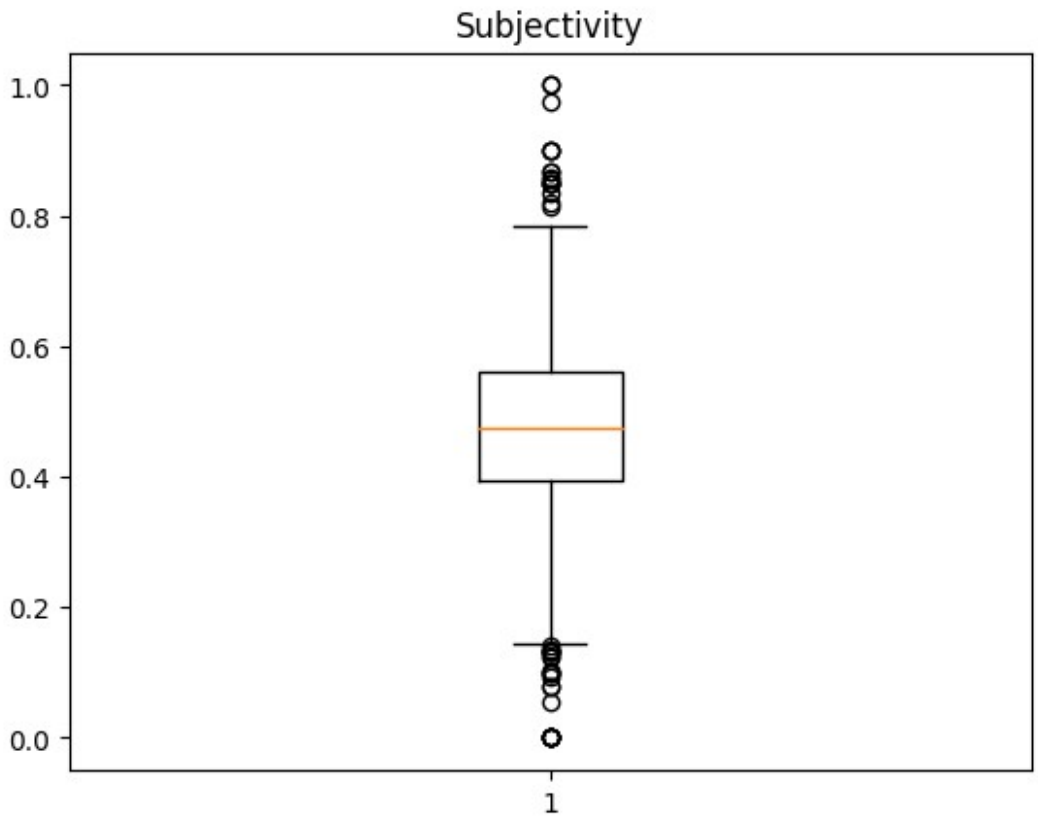
```
0.173913
25%      74.000000   416.750000      4.339289       32.000000
0.415873
50%     117.000000   649.500000      4.526825       53.000000
0.457729
75%     183.250000   996.250000      4.736573       86.000000
0.490171
max     654.000000  3466.000000      6.457143      347.000000
0.627907
```

```
       cleaned_review_count   clean_rate     polarity  subjectivity
count           1000.000000  1000.000000  1000.000000   1000.000000
mean              76.004000     0.538734     0.074324      0.477406
std               51.557329     0.062580     0.218752      0.141455
min               10.000000     0.348837    -1.000000      0.000000
25%               40.000000     0.496168    -0.048661      0.394575
50%               63.000000     0.530843     0.066181      0.475000
75%               95.250000     0.571429     0.211915      0.561594
max              322.000000     0.826087     0.800000      1.000000
```

Graphs

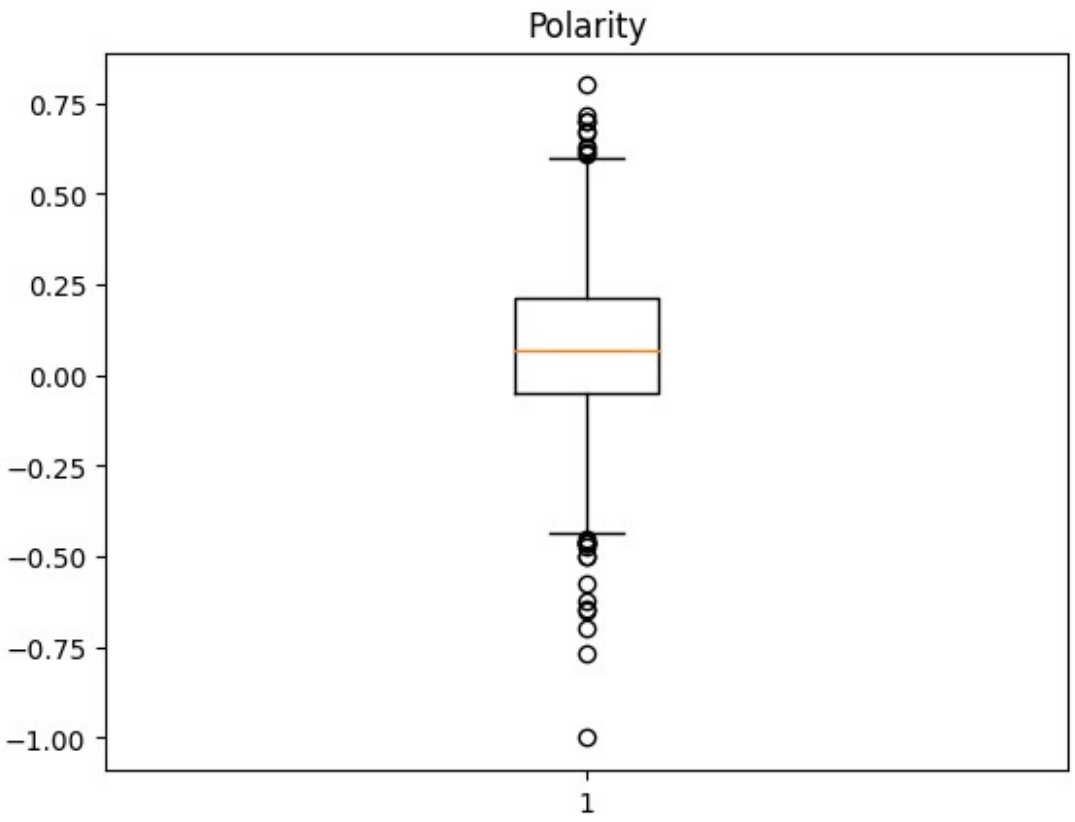
```
plt.boxplot(df['subjectivity'])
plt.title('Subjectivity')
```

```
Text(0.5, 1.0, 'Subjectivity')
```



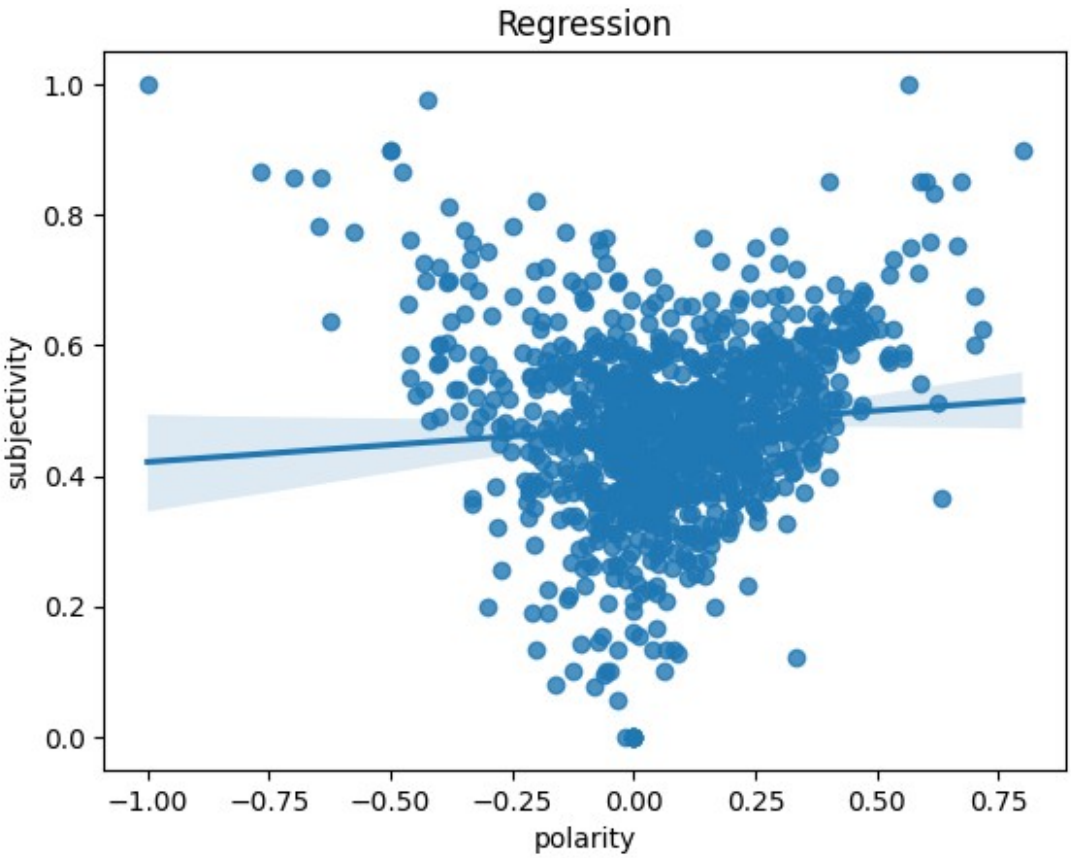

```
plt.boxplot(df['polarity'])
plt.title('Polarity')
```

```
Text(0.5, 1.0, 'Polarity')
```




```
plt.title("Regression")
sns.regplot(x=df['polarity'], y=df['subjectivity'])
```

```
<Axes: title={'center': 'Regression'}, xlabel='polarity',
ylabel='subjectivity'>
```


Regression
subjectivity
polarity
1.0
0.8
0.6
0.4
0.2
0.0
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75