Sentiment Analysis on Amazon Product Reviews

Importing Libraries

In [1]:
```
import pandas as pd
from nltk.sentiment.vader import SentimentIntensityAnalyzer
import nltk
import re
from textblob import TextBlob
from wordcloud import WordCloud, STOPWORDS
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import cufflinks as cf
%matplotlib inline
from plotly.offline import init_notebook_mode, iplot
init_notebook_mode(connected = True)
cf.go_offline();
import plotly.graph_objs as go
from plotly.subplots import make_subplots

import warnings
warnings.filterwarnings("ignore")
warnings.warn("this will not show")

pd.set_option('display.max_columns', None)
```

Importing Dataset

In [2]:
```
df =pd.read_csv(r'D:\amazon.csv')
```

Data Preprocessing/Cleaning

In [3]:
```
df.head()
```

Out[3]:

	Unnamed: 0	reviewerName	overall	reviewText	reviewTime	day_diff	helpful_yes	helpful_no
0	0	NaN	4	No issues.	23-07-2014	138	0	0
1	1	0mie	5	Purchased this for my device, it worked as adv...	25-10-2013	409	0	0
2	2	1K3	4	it works as expected. I should have sprung for...	23-12-2012	715	0	0
3	3	1m2	5	This think has worked out great.Had a diff. br...	21-11-2013	382	0	0
4	4	2&1/2Men	5	Bought it with Retail Packaging, arrived legit...	13-07-2013	513	0	0

In [4]:
```
df.tail()
```

Out[4]:

	Unnamed: 0	reviewerName	overall	reviewText	reviewTime	day_diff	helpful_yes	helpful_n
4910	4910	ZM "J"	1	I bought this Sandisk 16GB Class 10 to use wit...	23-07-2013	503	0	
4911	4911	Zo	5	Used this for extending the capabilities of my...	22-08-2013	473	0	
4912	4912	Z S Liske	5	Great card that is very fast and reliable. It ...	31-03-2014	252	0	
4913	4913	Z Taylor	5	Good amount of space for the stuff I want to d...	16-09-2013	448	0	
4914	4914	Zza	5	I've heard bad things about this 64gb Micro SD...	01-02-2014	310	0	

In [5]:
```
list(df)
```

Out[5]:
```
['Unnamed: 0',
 'reviewerName',
 'overall',
 'reviewText',
 'reviewTime',
 'day_diff',
 'helpful_yes',
 'helpful_no',
 'total_vote',
 'score_pos_neg_diff',
 'score_average_rating',
 'wilson_lower_bound']
```

In [6]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4915 entries, 0 to 4914
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Unnamed: 0            4915 non-null   int64
 1   reviewerName          4914 non-null   object
 2   overall               4915 non-null   int64
 3   reviewText            4914 non-null   object
 4   reviewTime            4915 non-null   object
 5   day_diff              4915 non-null   int64
 6   helpful_yes           4915 non-null   int64
 7   helpful_no            4915 non-null   int64
 8   total_vote            4915 non-null   int64
 9   score_pos_neg_diff    4915 non-null   int64
 10  score_average_rating  4915 non-null   float64
 11  wilson_lower_bound    4915 non-null   float64
dtypes: float64(2), int64(7), object(3)
memory usage: 460.9+ KB
```

In [7]:
```
print(df.shape)
```

```
(4915, 12)
```

Sentiment Analysis

In [8]:
```
df = df.sort_values('wilson_lower_bound', ascending = False)
df.drop('Unnamed: 0', inplace = True, axis = 1)
df.head()
```

Out[8]:

	reviewerName	overall	reviewText	reviewTime	day_diff	helpful_yes	helpful_no	total_
2031	Hyoun Kim "Faluzure"	5	[[UPDATE - 6/19/2014]]So my lovely wife boug...	05-01-2013	702	1952	68	
3449	NLee the Engineer	5	I have tested dozens of SDHC and micro-SDHC ca...	26-09-2012	803	1428	77	
4212	SkincareCEO	1	NOTE: please read the last update (scroll to ...	08-05-2013	579	1568	126	
317	Amazon Customer "Kelly"	1	If your card gets hot enough to be painful, it...	09-02-2012	1033	422	73	
4672	Twister	5	Sandisk announcement of the first 128GB micro ...	03-07-2014	158	45	4	

In [9]:
```
def missing_values_analysis(df):
    na_columns_ = [col for col in df.columns if df[col].isnull().sum() > 0]
    n_miss = df[na_columns_].isnull().sum().sort_values(ascending=True)
```

```
    ratio_ = (df[na_columns_].isnull().sum() / df.shape[0] * 100).sort_values(asce
    missing_df = pd.concat([n_miss, np.round(ratio_, 2)], axis=1, keys=['Total Mis
    missing_df = pd.DataFrame(missing_df)
    return missing_df

def check_dataframe(df, head=5, tail = 5):

    print(" SHAPE ".center(82,'~'))
    print('Rows: {}'.format(df.shape[0]))
    print('Columns: {}'.format(df.shape[1]))
    print(" TYPES ".center(82,'~'))
    print(df.dtypes)
    print("".center(82,'~'))
    print(missing_values_analysis(df))
    print(' DUPLICATED VALUES '.center(83,'~'))
    print(df.duplicated().sum())
    print(" QUANTILES ".center(82,'~'))
    print(df.quantile([0, 0.05, 0.50, 0.95, 0.99, 1]).T)

check_dataframe(df)
```

```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ SHAPE ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Rows: 4915
Columns: 11
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ TYPES ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
reviewerName             object
overall                   int64
reviewText               object
reviewTime               object
day_diff                  int64
helpful_yes               int64
helpful_no                int64
total_vote                int64
score_pos_neg_diff        int64
score_average_rating    float64
wilson_lower_bound      float64
dtype: object
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
              Total Missing Values  Ratio
reviewerName                     1   0.02
reviewText                       1   0.02
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ DUPLICATED VALUES ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
~
0
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ QUANTILES ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
                        0.00  0.05   0.50        0.95       0.99         1.00
overall                  1.0   2.0    5.0    5.000000    5.00000     5.000000
day_diff                 1.0  98.0  431.0  748.000000  943.00000  1064.000000
helpful_yes              0.0   0.0    0.0    1.000000    3.00000  1952.000000
helpful_no               0.0   0.0    0.0    0.000000    2.00000   183.000000
total_vote               0.0   0.0    0.0    1.000000    4.00000  2020.000000
score_pos_neg_diff    -130.0   0.0    0.0    1.000000    2.00000  1884.000000
score_average_rating     0.0   0.0    0.0    1.000000    1.00000     1.000000
wilson_lower_bound       0.0   0.0    0.0    0.206549    0.34238     0.957544
```

In [10]:

```
def check_class(dataframe):
    nunique_df = pd.DataFrame({'Variable': dataframe.columns,
                               'Classes': [dataframe[i].nunique() \
                                           for i in dataframe.columns]})

    nunique_df = nunique_df.sort_values('Classes', ascending=False)
    nunique_df = nunique_df.reset_index(drop = True)
    return nunique_df
```

```
check_class(df)
```

Out[10]:

	Variable	Classes
0	reviewText	4912
1	reviewerName	4594
2	reviewTime	690
3	day_diff	690
4	wilson_lower_bound	40
5	score_average_rating	28
6	score_pos_neg_diff	27
7	total_vote	26
8	helpful_yes	23
9	helpful_no	17
10	overall	5

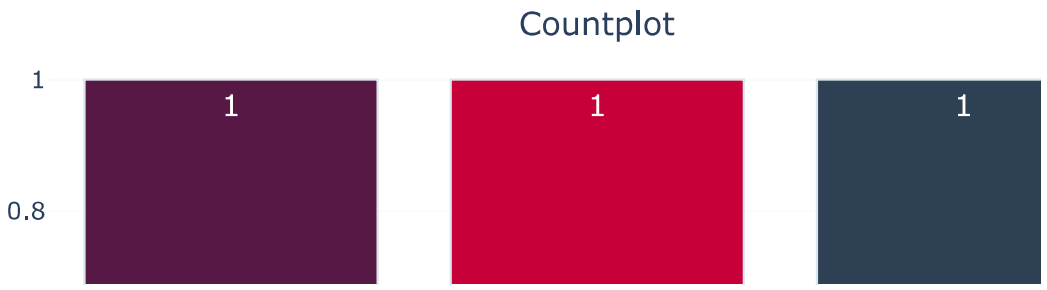
Data Visualization

In [11]:

```
# categorical variable analysis ---> overall

constraints = ['#581845','#C70039','#2E4053','#1ABC9C','#7F8C8D']

def categorical_variable_summary(df, column_name):
    fig = make_subplots(rows=1,cols=2,
                        subplot_titles=('Countplot','Percentages'),
                        specs=[[{"type": "xy"}, {'type':'domain'}]])

    fig.add_trace(go.Bar( y = df[column_name].value_counts().values.tolist(),
                          x = [str(i) for i in df[column_name].value_counts().inde
                          text = df[column_name].value_counts().values.tolist(),
                          textfont = dict(size=15),
                          name = column_name,
                          textposition = 'auto',
                          showlegend=False,
                          marker=dict(color = constraints,
                                      line=dict(color='#DBE6EC',
                                                width=1))),
                  row = 1, col = 1)

    fig.add_trace(go.Pie(labels= df[column_name].value_counts().keys(),
                         values= df[column_name].value_counts().values,
                         textfont = dict(size = 20),
                         textposition='auto',
                         showlegend = False,
                         name = column_name,
                         marker=dict(colors=constraints)),
                  row = 1, col = 2)

    fig.update_layout(title={'text': column_name,
                             'y':0.9,
                             'x':0.5,
                             'xanchor': 'center',
                             'yanchor': 'top'},
```

```
                      template='plotly_white')

    iplot(fig)
```

In [29]:
```
categorical_variable_summary(df,'overall')
```



Model Training, Evaluation and Prediction

In [13]:
```
# sample for cleaning
df.reviewText.head()
```

Out[13]:
```
2031    [[ UPDATE - 6/19/2014 ]]So my lovely wife boug...
3449    I have tested dozens of SDHC and micro-SDHC ca...
4212    NOTE:  please read the last update (scroll to ...
317     If your card gets hot enough to be painful, it...
4672    Sandisk announcement of the first 128GB micro ...
Name: reviewText, dtype: object
```

In [14]:
```
example_review = df.reviewText[2031]
example_review
```

Out[14]: '[[UPDATE - 6/19/2014]]So my lovely wife bought me a Samsung Galaxy Tab 4 for Father\'s Day and I\'ve been loving it ever since. Just as other with Samsung products, the Galaxy Tab 4 has the ability to add a microSD card to expand the memory on the device. Since it\'s been over a year, I decided to do some more research to see if SanDisk offered anything new. As of 6/19/2014, their product lineup for microSD cards from worst to best (performance-wise) are the as follows:SanDiskSanDisk UltraSanDisk Ultra PLUSSanDisk ExtremeSanDisk Extreme PLUSSanDisk Extreme PRONow, the difference between all of these cards are simply the speed in which you can read/write data to the card. Yes, the published rating of most all these cards (except the SanDisk regular) are Class 10/UHS-I but that\'s just a rating... Actual real world performance does get better with each model, but with faster cards come more expensive prices. Since Amazon doesn\'t carry the Ultra PLUS model of microSD card, I had to do direct comparisons between the SanDisk Ultra ($34.27), Extreme ($57.95), and Extreme PLUS ($67.95).As mentioned in my earlier review, I purchased the SanDisk Ultra for my Galaxy S4. My question was, did I want to pay over $20 more for a card that is faster than the one I already owned? Or I could pay almost double to get SanDisk\'s 2nd-most fastest microSD card.The Ultra works perfectly fine for my style of usage (storing/capturing pictures & HD video and movie playback) on my phone. So in the end, I ended up just buying another SanDisk Ultra 64GB card. I use my cell phone *more* than I do my tablet and if the card is good enough for my phone, it\'s good enough for my tablet. I don\'t own a 4K HD camera or anything like that, so I honestly didn\'t see a need to get one of the faster cards at this time.I am now a proud owner of 2 SanDisk Ultra cards and have absolutely 0 issues with it in my Samsung devices.[[ORIGINAL REVIEW - 5/1/2013]]I haven\'t had to buy a microSD card in a long time. The last time I bought one was for my cell phone over 2 years ago. But since my cellular contract was up, I knew I would have to get a newer card in addition to my new phone, the Samsung Galaxy S4. Reason for this is because I knew my small 16GB microSD card wasn\'t going to cut it.Doing research on the Galaxy S4, I wanted to get the best card possible that had decent capacity (32 GB or greater). This led me to find that the Galaxy S4 supports the microSDXC Class 10 UHS-I card, which is the fastest possible given that class. Searching for that specifically on Amazon gave me results of only 3 vendors (as of April) that makes these microSDXC Class 10 UHS-1 cards. They are Sandisk (the majority), Samsung and Lexar. Nobody else makes these that are sold on Amazon.Seeing how SanDisk is a pretty good name out of the 3 (I\'ve used them the most), I decided upon the SanDisk because Lexar was overpriced and the Samsung one was overpriced (as well as not eligible for Amazon Prime).But the scary thing is that when you filter by the SanDisk, you literally get DOZENS of options. All of them have different model numbers, different sizes, etc. Then there\'s that confusion of what\'s the difference between SDHC & SDXC?SDHC vs SDXC:SDHC stand for "Secure Digital High Capacity" and SDXC stands for "Secure Digital eXtended Capacity". Essentially these two cards are the same with the exception that SDHC only supports capcities up to 32GB and is formated with the FAT32 file system. The SDXC cards are formatted with the exFAT file system. If you use an SDXC card in a device, it must support that file system, otherwise it may not be recognizable and/or you have to reformat the card to FAT32.FAT32 vs exFAT:The differences between the two file systems means that FAT32 has a maximum file size of 4GB, limited by that file system. exFAT on the otherhand, supports file sizes up to 2TB (terabytes). The only thing you need to know here really is that it\'s possible your device doesn\'t support exFAT. If that\'s the case, just reformat it to FAT32. REMEMBER FORMATTING ERASES ALL DATA!To clarify the model numbers, I I hopped over to the SanDisk official webpage. What I found there is that they offer two "highspeed" options for SanDisk cards. These are SanDisk Extreme Pro and SanDisk Ultra. SanDisk Extreme Pro is a line that supports read speeds up to 95MB/sec, however they are SDHC only. To make things worse, they are currently only available in 16GB & 8GB capacities. Since one of my requirements was to have a lot of storage, I ruled these out.The remaining devices listed on Amazon\'s search were the SanDisk Ultra line. But here, confusion sets in because SanDisk separates these cards to two different devices. Cameras & mobile devices. Is there a real difference between the two or is this just a marketing stunt? Unfortunately I\'m not sure but I do know the price difference between the two range from a couple cents to a few dollars. Since I wasn\'t sure, I opted for the one specifically targeted for mobile devices (just in case there is some kind of compatibility issue). To find the exact model number, I would go to Sandisk\'s webpage (sandisk.com) and compare their existing product lineup. From there, you get exact mo

```
del numbers and you can then search Amazon for these model numbers. That is how I
got mine (SDSDQUA-064G).As for speed tests, I haven\'t run any specific testing, b
ut copying 8 GB worth of data from my PC to the card literally took just a few min
utes.One last note is that Amazon attaches additional characters to the end (for e
xample SDSDQUA-064G-AFFP-A vs SDSDQUA-064G-U46A). The difference between the two i
s that the "AFFP-A" means "Amazon Frustration Free Packaging". Other than that, th
ese are exactly the same.  If you\'re wondering what I got (and want to use it in
your Galaxy S4), I got the SDSDQUA-064G-u46A and it works like charm.'
```

In [15]:
```
# we clean it from punctuation and numbers - using regex. {Regular expression}
example_review = re.sub("[^a-zA-Z]",' ',example_review)
example_review
```

Out[15]: ' UPDATE So my lovely wife bought me a Samsung Galaxy Tab for Fa ther s Day and I ve been loving it ever since Just as other with Samsung product s the Galaxy Tab has the ability to add a microSD card to expand the memory on the device Since it s been over a year I decided to do some more research to se e if SanDisk offered anything new As of their product lineup for micr oSD cards from worst to best performance wise are the as follows SanDiskSanDisk UltraSanDisk Ultra PLUSSanDisk ExtremeSanDisk Extreme PLUSSanDisk Extreme PRONow the difference between all of these cards are simply the speed in which you can re ad write data to the card Yes the published rating of most all these cards exc ept the SanDisk regular are Class UHS I but that s just a rating Actual rea l world performance does get better with each model but with faster cards come mo re expensive prices Since Amazon doesn t carry the Ultra PLUS model of microSD c ard I had to do direct comparisons between the SanDisk Ultra Extreme and Extreme PLUS As mentioned in my earlier review I purchased the SanDi sk Ultra for my Galaxy S My question was did I want to pay over more for a card that is faster than the one I already owned Or I could pay almost double to get SanDisk s nd most fastest microSD card The Ultra works perfectly fine for my style of usage storing capturing pictures HD video and movie playback on my ph one So in the end I ended up just buying another SanDisk Ultra GB card I us e my cell phone more than I do my tablet and if the card is good enough for my p hone it s good enough for my tablet I don t own a K HD camera or anything like that so I honestly didn t see a need to get one of the faster cards at this time I am now a proud owner of SanDisk Ultra cards and have absolutely issues with it in my Samsung devices ORIGINAL REVIEW I haven t had to buy a mi croSD card in a long time The last time I bought one was for my cell phone over years ago But since my cellular contract was up I knew I would have to get a new er card in addition to my new phone the Samsung Galaxy S Reason for this is bec ause I knew my small GB microSD card wasn t going to cut it Doing research on th e Galaxy S I wanted to get the best card possible that had decent capacity G B or greater This led me to find that the Galaxy S supports the microSDXC Class UHS I card which is the fastest possible given that class Searching for that spe cifically on Amazon gave me results of only vendors as of April that makes the se microSDXC Class UHS cards They are Sandisk the majority Samsung and Le xar Nobody else makes these that are sold on Amazon Seeing how SanDisk is a prett y good name out of the I ve used them the most I decided upon the SanDisk bec ause Lexar was overpriced and the Samsung one was overpriced as well as not eligi ble for Amazon Prime But the scary thing is that when you filter by the SanDisk you literally get DOZENS of options All of them have different model numbers dif ferent sizes etc Then there s that confusion of what s the difference between SD HC SDXC SDHC vs SDXC SDHC stand for Secure Digital High Capacity and SDXC stan ds for Secure Digital eXtended Capacity Essentially these two cards are the sam e with the exception that SDHC only supports capcities up to GB and is formated with the FAT file system The SDXC cards are formatted with the exFAT file syste m If you use an SDXC card in a device it must support that file system otherwis e it may not be recognizable and or you have to reformat the card to FAT FAT v s exFAT The differences between the two file systems means that FAT has a maximu m file size of GB limited by that file system exFAT on the otherhand supports file sizes up to TB terabytes The only thing you need to know here really is t hat it s possible your device doesn t support exFAT If that s the case just refo rmat it to FAT REMEMBER FORMATTING ERASES ALL DATA To clarify the model numbers I I hopped over to the SanDisk official webpage What I found there is that they o ffer two highspeed options for SanDisk cards These are SanDisk Extreme Pro and SanDisk Ultra SanDisk Extreme Pro is a line that supports read speeds up to MB sec however they are SDHC only To make things worse they are currently only ava ilable in GB GB capacities Since one of my requirements was to have a lot of storage I ruled these out The remaining devices listed on Amazon s search were th e SanDisk Ultra line But here confusion sets in because SanDisk separates these cards to two different devices Cameras mobile devices Is there a real differen ce between the two or is this just a marketing stunt Unfortunately I m not sure b ut I do know the price difference between the two range from a couple cents to a f ew dollars Since I wasn t sure I opted for the one specifically targeted for mob ile devices just in case there is some kind of compatibility issue To find the exact model number I would go to Sandisk s webpage sandisk com and compare thei r existing product lineup From there you get exact model numbers and you can the

```
n search Amazon for these model numbers  That is how I got mine  SDSDQUA    G  As
for speed tests  I haven t run any specific testing  but copying   GB worth of dat
a from my PC to the card literally took just a few minutes One last note is that A
mazon attaches additional characters to the end  for example SDSDQUA    G AFFP A v
s SDSDQUA    G U  A   The difference between the two is that the  AFFP A  means  A
mazon Frustration Free Packaging   Other than that  these are exactly the same   I
f you re wondering what I got  and want to use it in your Galaxy S    I got the SD
SDQUA    G u  A and it works like charm '
```

In [16]:
```
example_review = example_review.lower().split()
```

In [17]:
```
example_review
```

Out[17]:
```
['update',
 'so',
 'my',
 'lovely',
 'wife',
 'bought',
 'me',
 'a',
 'samsung',
 'galaxy',
 'tab',
 'for',
 'father',
 's',
 'day',
 'and',
 'i',
 've',
 'been',
 'loving',
 'it',
 'ever',
 'since',
 'just',
 'as',
 'other',
 'with',
 'samsung',
 'products',
 'the',
 'galaxy',
 'tab',
 'has',
 'the',
 'ability',
 'to',
 'add',
 'a',
 'microsd',
 'card',
 'to',
 'expand',
 'the',
 'memory',
 'on',
 'the',
 'device',
 'since',
 'it',
 's',
 'been',
 'over',
 'a',
 'year',
 'i',
 'decided',
 'to',
 'do',
 'some',
 'more',
 'research',
 'to',
 'see',
 'if',
```

```
 'sandisk',
 'offered',
 'anything',
 'new',
 'as',
 'of',
 'their',
 'product',
 'lineup',
 'for',
 'microsd',
 'cards',
 'from',
 'worst',
 'to',
 'best',
 'performance',
 'wise',
 'are',
 'the',
 'as',
 'follows',
 'sandisksandisk',
 'ultrasandisk',
 'ultra',
 'plussandisk',
 'extremesandisk',
 'extreme',
 'plussandisk',
 'extreme',
 'pronow',
 'the',
 'difference',
 'between',
 'all',
 'of',
 'these',
 'cards',
 'are',
 'simply',
 'the',
 'speed',
 'in',
 'which',
 'you',
 'can',
 'read',
 'write',
 'data',
 'to',
 'the',
 'card',
 'yes',
 'the',
 'published',
 'rating',
 'of',
 'most',
 'all',
 'these',
 'cards',
 'except',
 'the',
 'sandisk',
```

```
 'regular',
 'are',
 'class',
 'uhs',
 'i',
 'but',
 'that',
 's',
 'just',
 'a',
 'rating',
 'actual',
 'real',
 'world',
 'performance',
 'does',
 'get',
 'better',
 'with',
 'each',
 'model',
 'but',
 'with',
 'faster',
 'cards',
 'come',
 'more',
 'expensive',
 'prices',
 'since',
 'amazon',
 'doesn',
 't',
 'carry',
 'the',
 'ultra',
 'plus',
 'model',
 'of',
 'microsd',
 'card',
 'i',
 'had',
 'to',
 'do',
 'direct',
 'comparisons',
 'between',
 'the',
 'sandisk',
 'ultra',
 'extreme',
 'and',
 'extreme',
 'plus',
 'as',
 'mentioned',
 'in',
 'my',
 'earlier',
 'review',
 'i',
 'purchased',
 'the',
```

```
'sandisk',
'ultra',
'for',
'my',
'galaxy',
's',
'my',
'question',
'was',
'did',
'i',
'want',
'to',
'pay',
'over',
'more',
'for',
'a',
'card',
'that',
'is',
'faster',
'than',
'the',
'one',
'i',
'already',
'owned',
'or',
'i',
'could',
'pay',
'almost',
'double',
'to',
'get',
'sandisk',
's',
'nd',
'most',
'fastest',
'microsd',
'card',
'the',
'ultra',
'works',
'perfectly',
'fine',
'for',
'my',
'style',
'of',
'usage',
'storing',
'capturing',
'pictures',
'hd',
'video',
'and',
'movie',
'playback',
'on',
'my',
'phone',
```

```
'so',
'in',
'the',
'end',
'i',
'ended',
'up',
'just',
'buying',
'another',
'sandisk',
'ultra',
'gb',
'card',
'i',
'use',
'my',
'cell',
'phone',
'more',
'than',
'i',
'do',
'my',
'tablet',
'and',
'if',
'the',
'card',
'is',
'good',
'enough',
'for',
'my',
'phone',
'it',
's',
'good',
'enough',
'for',
'my',
'tablet',
'i',
'don',
't',
'own',
'a',
'k',
'hd',
'camera',
'or',
'anything',
'like',
'that',
'so',
'i',
'honestly',
'didn',
't',
'see',
'a',
'need',
'to',
'get',
```

```
 'one',
 'of',
 'the',
 'faster',
 'cards',
 'at',
 'this',
 'time',
 'i',
 'am',
 'now',
 'a',
 'proud',
 'owner',
 'of',
 'sandisk',
 'ultra',
 'cards',
 'and',
 'have',
 'absolutely',
 'issues',
 'with',
 'it',
 'in',
 'my',
 'samsung',
 'devices',
 'original',
 'review',
 'i',
 'haven',
 't',
 'had',
 'to',
 'buy',
 'a',
 'microsd',
 'card',
 'in',
 'a',
 'long',
 'time',
 'the',
 'last',
 'time',
 'i',
 'bought',
 'one',
 'was',
 'for',
 'my',
 'cell',
 'phone',
 'over',
 'years',
 'ago',
 'but',
 'since',
 'my',
 'cellular',
 'contract',
 'was',
 'up',
```

```
 'i',
 'knew',
 'i',
 'would',
 'have',
 'to',
 'get',
 'a',
 'newer',
 'card',
 'in',
 'addition',
 'to',
 'my',
 'new',
 'phone',
 'the',
 'samsung',
 'galaxy',
 's',
 'reason',
 'for',
 'this',
 'is',
 'because',
 'i',
 'knew',
 'my',
 'small',
 'gb',
 'microsd',
 'card',
 'wasn',
 't',
 'going',
 'to',
 'cut',
 'it',
 'doing',
 'research',
 'on',
 'the',
 'galaxy',
 's',
 'i',
 'wanted',
 'to',
 'get',
 'the',
 'best',
 'card',
 'possible',
 'that',
 'had',
 'decent',
 'capacity',
 'gb',
 'or',
 'greater',
 'this',
 'led',
 'me',
 'to',
 'find',
```

```
 'that',
 'the',
 'galaxy',
 's',
 'supports',
 'the',
 'microsdxc',
 'class',
 'uhs',
 'i',
 'card',
 'which',
 'is',
 'the',
 'fastest',
 'possible',
 'given',
 'that',
 'class',
 'searching',
 'for',
 'that',
 'specifically',
 'on',
 'amazon',
 'gave',
 'me',
 'results',
 'of',
 'only',
 'vendors',
 'as',
 'of',
 'april',
 'that',
 'makes',
 'these',
 'microsdxc',
 'class',
 'uhs',
 'cards',
 'they',
 'are',
 'sandisk',
 'the',
 'majority',
 'samsung',
 'and',
 'lexar',
 'nobody',
 'else',
 'makes',
 'these',
 'that',
 'are',
 'sold',
 'on',
 'amazon',
 'seeing',
 'how',
 'sandisk',
 'is',
 'a',
 'pretty',
```

```
 'good',
 'name',
 'out',
 'of',
 'the',
 'i',
 've',
 'used',
 'them',
 'the',
 'most',
 'i',
 'decided',
 'upon',
 'the',
 'sandisk',
 'because',
 'lexar',
 'was',
 'overpriced',
 'and',
 'the',
 'samsung',
 'one',
 'was',
 'overpriced',
 'as',
 'well',
 'as',
 'not',
 'eligible',
 'for',
 'amazon',
 'prime',
 'but',
 'the',
 'scary',
 'thing',
 'is',
 'that',
 'when',
 'you',
 'filter',
 'by',
 'the',
 'sandisk',
 'you',
 'literally',
 'get',
 'dozens',
 'of',
 'options',
 'all',
 'of',
 'them',
 'have',
 'different',
 'model',
 'numbers',
 'different',
 'sizes',
 'etc',
 'then',
 'there',
```

```
 's',
 'that',
 'confusion',
 'of',
 'what',
 's',
 'the',
 'difference',
 'between',
 'sdhc',
 'sdxc',
 'sdhc',
 'vs',
 'sdxc',
 'sdhc',
 'stand',
 'for',
 'secure',
 'digital',
 'high',
 'capacity',
 'and',
 'sdxc',
 'stands',
 'for',
 'secure',
 'digital',
 'extended',
 'capacity',
 'essentially',
 'these',
 'two',
 'cards',
 'are',
 'the',
 'same',
 'with',
 'the',
 'exception',
 'that',
 'sdhc',
 'only',
 'supports',
 'capcities',
 'up',
 'to',
 'gb',
 'and',
 'is',
 'formated',
 'with',
 'the',
 'fat',
 'file',
 'system',
 'the',
 'sdxc',
 'cards',
 'are',
 'formatted',
 'with',
 'the',
 'exfat',
 'file',
```

```
'system',
'if',
'you',
'use',
'an',
'sdxc',
'card',
'in',
'a',
'device',
'it',
'must',
'support',
'that',
'file',
'system',
'otherwise',
'it',
'may',
'not',
'be',
'recognizable',
'and',
'or',
'you',
'have',
'to',
'reformat',
'the',
'card',
'to',
'fat',
'fat',
'vs',
'exfat',
'the',
'differences',
'between',
'the',
'two',
'file',
'systems',
'means',
'that',
'fat',
'has',
'a',
'maximum',
'file',
'size',
'of',
'gb',
'limited',
'by',
'that',
'file',
'system',
'exfat',
'on',
'the',
'otherhand',
'supports',
'file',
'sizes',
```

```
 'up',
 'to',
 'tb',
 'terabytes',
 'the',
 'only',
 'thing',
 'you',
 'need',
 'to',
 'know',
 'here',
 'really',
 'is',
 'that',
 'it',
 's',
 'possible',
 'your',
 'device',
 'doesn',
 't',
 'support',
 'exfat',
 'if',
 'that',
 's',
 'the',
 'case',
 'just',
 'reformat',
 'it',
 'to',
 'fat',
 'remember',
 'formatting',
 'erases',
 'all',
 'data',
 'to',
 'clarify',
 'the',
 'model',
 'numbers',
 'i',
 'i',
 'hopped',
 'over',
 'to',
 'the',
 'sandisk',
 'official',
 'webpage',
 'what',
 'i',
 'found',
 'there',
 'is',
 'that',
 'they',
 'offer',
 'two',
 'highspeed',
 'options',
```

```
 'for',
 'sandisk',
 'cards',
 'these',
 'are',
 'sandisk',
 'extreme',
 'pro',
 'and',
 'sandisk',
 'ultra',
 'sandisk',
 'extreme',
 'pro',
 'is',
 'a',
 'line',
 'that',
 'supports',
 'read',
 'speeds',
 'up',
 'to',
 'mb',
 'sec',
 'however',
 'they',
 'are',
 'sdhc',
 'only',
 'to',
 'make',
 'things',
 'worse',
 'they',
 'are',
 'currently',
 'only',
 'available',
 'in',
 'gb',
 'gb',
 'capacities',
 'since',
 'one',
 'of',
 'my',
 'requirements',
 'was',
 'to',
 'have',
 'a',
 'lot',
 'of',
 'storage',
 'i',
 'ruled',
 'these',
 'out',
 'the',
 'remaining',
 'devices',
 'listed',
 'on',
```

```
 'amazon',
 's',
 'search',
 'were',
 'the',
 'sandisk',
 'ultra',
 'line',
 'but',
 'here',
 'confusion',
 'sets',
 'in',
 'because',
 'sandisk',
 'separates',
 'these',
 'cards',
 'to',
 'two',
 'different',
 'devices',
 'cameras',
 'mobile',
 'devices',
 'is',
 'there',
 'a',
 'real',
 'difference',
 'between',
 'the',
 'two',
 'or',
 'is',
 'this',
 'just',
 'a',
 'marketing',
 'stunt',
 'unfortunately',
 'i',
 'm',
 'not',
 'sure',
 'but',
 'i',
 'do',
 'know',
 'the',
 'price',
 'difference',
 'between',
 'the',
 'two',
 'range',
 'from',
 'a',
 'couple',
 'cents',
 'to',
 'a',
 'few',
 'dollars',
```

```
 'since',
 'i',
 'wasn',
 't',
 'sure',
 'i',
 'opted',
 'for',
 'the',
 'one',
 'specifically',
 'targeted',
 'for',
 'mobile',
 'devices',
 'just',
 'in',
 'case',
 'there',
 'is',
 'some',
 'kind',
 'of',
 'compatibility',
 'issue',
 'to',
 'find',
 'the',
 'exact',
 'model',
 'number',
 'i',
 'would',
 'go',
 'to',
 'sandisk',
 's',
 'webpage',
 'sandisk',
 'com',
 'and',
 'compare',
 'their',
 'existing',
 'product',
 'lineup',
 'from',
 'there',
 'you',
 'get',
 'exact',
 'model',
 'numbers',
 'and',
 'you',
 'can',
 'then',
 'search',
 'amazon',
 'for',
 'these',
 'model',
 'numbers',
 'that',
```

```
'is',
'how',
'i',
'got',
'mine',
'sdsdqua',
'g',
'as',
'for',
'speed',
'tests',
'i',
'haven',
't',
'run',
'any',
'specific',
'testing',
'but',
'copying',
'gb',
'worth',
'of',
'data',
'from',
'my',
'pc',
'to',
'the',
'card',
'literally',
'took',
'just',
'a',
'few',
'minutes',
'one',
'last',
'note',
'is',
...]
```

In [18]:
```
rt = lambda x: re.sub("[^a-zA-Z]",' ',str(x))
df["reviewText"] = df["reviewText"].map(rt)
df["reviewText"] = df["reviewText"].str.lower()
df.head(10)
```

Out[18]:

	reviewerName	overall	reviewText	reviewTime	day_diff	helpful_yes	helpful_no	total_
2031	Hyoun Kim "Faluzure"	5	update so my lovely wife boug...	05-01-2013	702	1952	68	2
3449	NLee the Engineer	5	i have tested dozens of sdhc and micro sdhc ca...	26-09-2012	803	1428	77	1
4212	SkincareCEO	1	note please read the last update scroll to ...	08-05-2013	579	1568	126	1
317	Amazon Customer "Kelly"	1	if your card gets hot enough to be painful it...	09-02-2012	1033	422	73	
4672	Twister	5	sandisk announcement of the first gb micro ...	03-07-2014	158	45	4	
1835	goconfigure	5	bought from bestbuy online the day it was anno...	28-02-2014	283	60	8	
3981	R. Sutton, Jr. "RWSynergy"	5	the last few days i have been diligently shopp...	22-10-2012	777	112	27	
3807	R. Heisler	3	i bought this card to replace a lost gig in...	27-02-2013	649	22	3	
4306	Stellar Eller	5	while i got this card as a deal of the day o...	06-09-2012	823	51	14	
4596	Tom Henriksen "Doggy Diner"	1	hi i ordered two card and they arrived the nex...	22-09-2012	807	82	27	

Sentiment Analysis using TextBlob

In [19]:
```
'''
# Sentiment analysis
# TextBlob Exit will return polarity and subjectivity.
# Polarity indicates your mood, that is, whether it is positive.
# It returns a value between 0 and 1. The closer to 1 the more positive, the closer
'''

df[['polarity', 'subjectivity']] = df['reviewText'].apply(lambda Text: pd.Series(T

for index, row in df['reviewText'].iteritems():
```

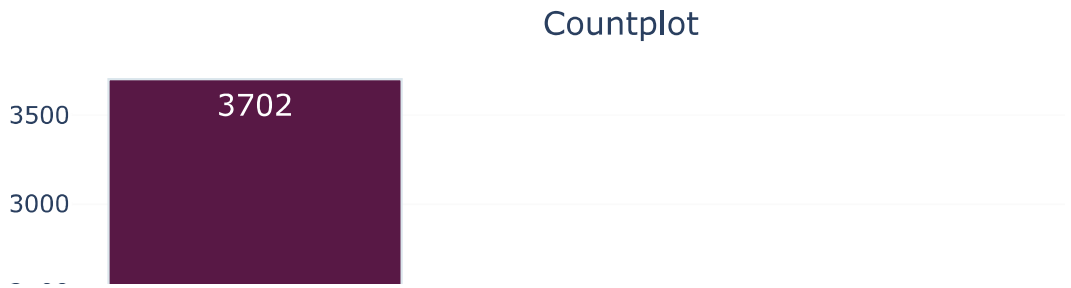
```
    score = SentimentIntensityAnalyzer().polarity_scores(row)

    neg = score['neg']
    neu = score['neu']
    pos = score['pos']
    if neg > pos:
        df.loc[index, 'sentiment'] = "Negative"
    elif pos > neg:
        df.loc[index, 'sentiment'] = "Positive"
    else:
        df.loc[index, 'sentiment'] = "Neutral"
```

In [20]:
```
# 20 Identifying the interpretation, now we can include the positive, negative and

df[df["sentiment"] == "Positive"].sort_values("wilson_lower_bound", ascending=Fals
```

Out[20]:

	reviewerName	overall	reviewText	reviewTime	day_diff	helpful_yes	helpful_no	total_
2031	Hyoun Kim "Faluzure"	5	update so my lovely wife boug...	05-01-2013	702	1952	68	2
3449	NLee the Engineer	5	i have tested dozens of sdhc and micro sdhc ca...	26-09-2012	803	1428	77	1
4212	SkincareCEO	1	note please read the last update scroll to ...	08-05-2013	579	1568	126	1
317	Amazon Customer "Kelly"	1	if your card gets hot enough to be painful it...	09-02-2012	1033	422	73	
4672	Twister	5	sandisk announcement of the first gb micro ...	03-07-2014	158	45	4	

In [21]:
```
# Let's see if we have an unbalanced data problem
categorical_variable_summary(df,'sentiment')
```



In [22]:
```
# Let's see if there is an imbalance in the scoring?
df.groupby(["sentiment"])[['overall']].mean()
```

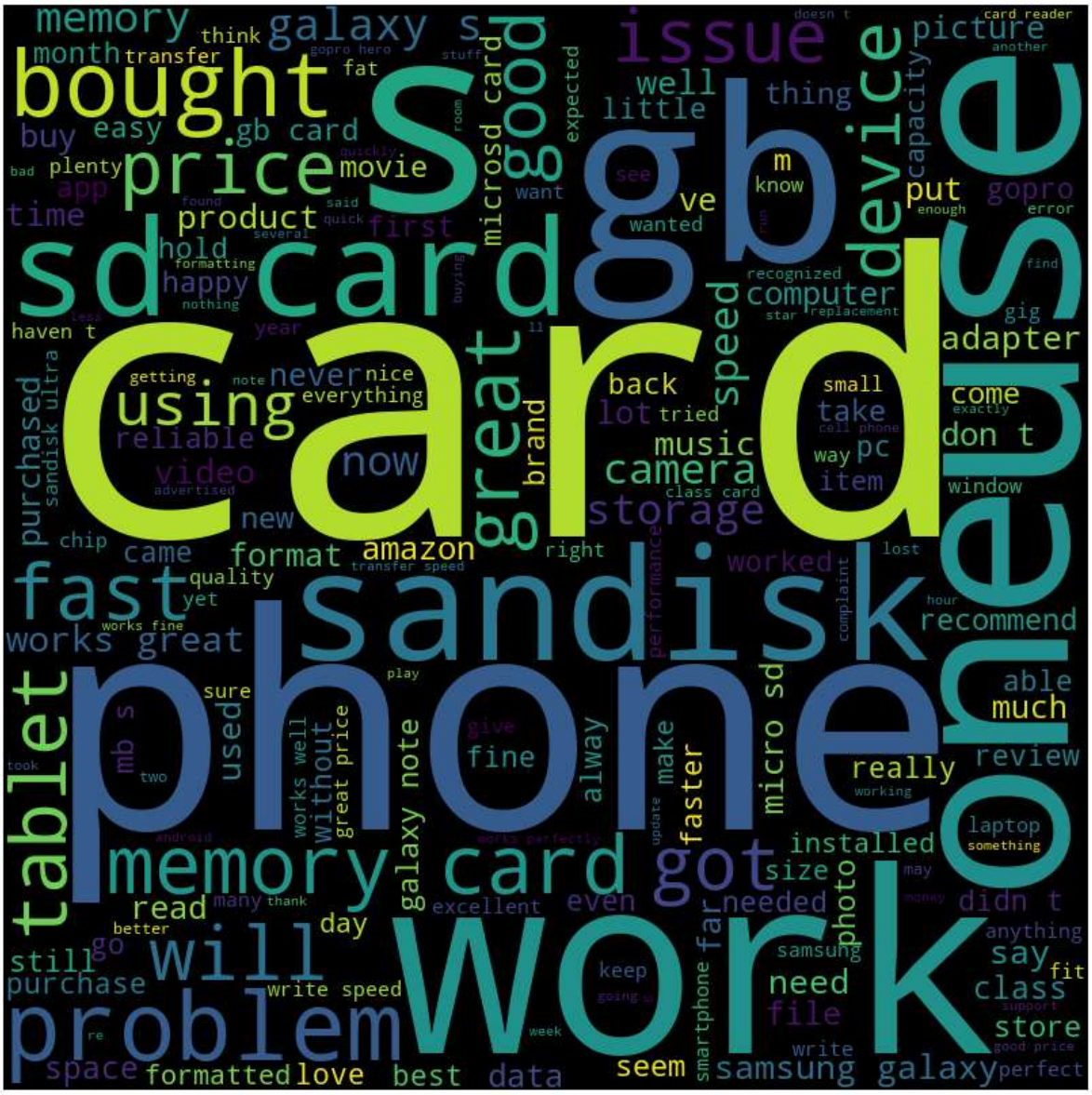
Out[22]:

sentiment	overall
Negative	3.968085
Neutral	4.688645
Positive	4.737439

Genearting Word Cloud

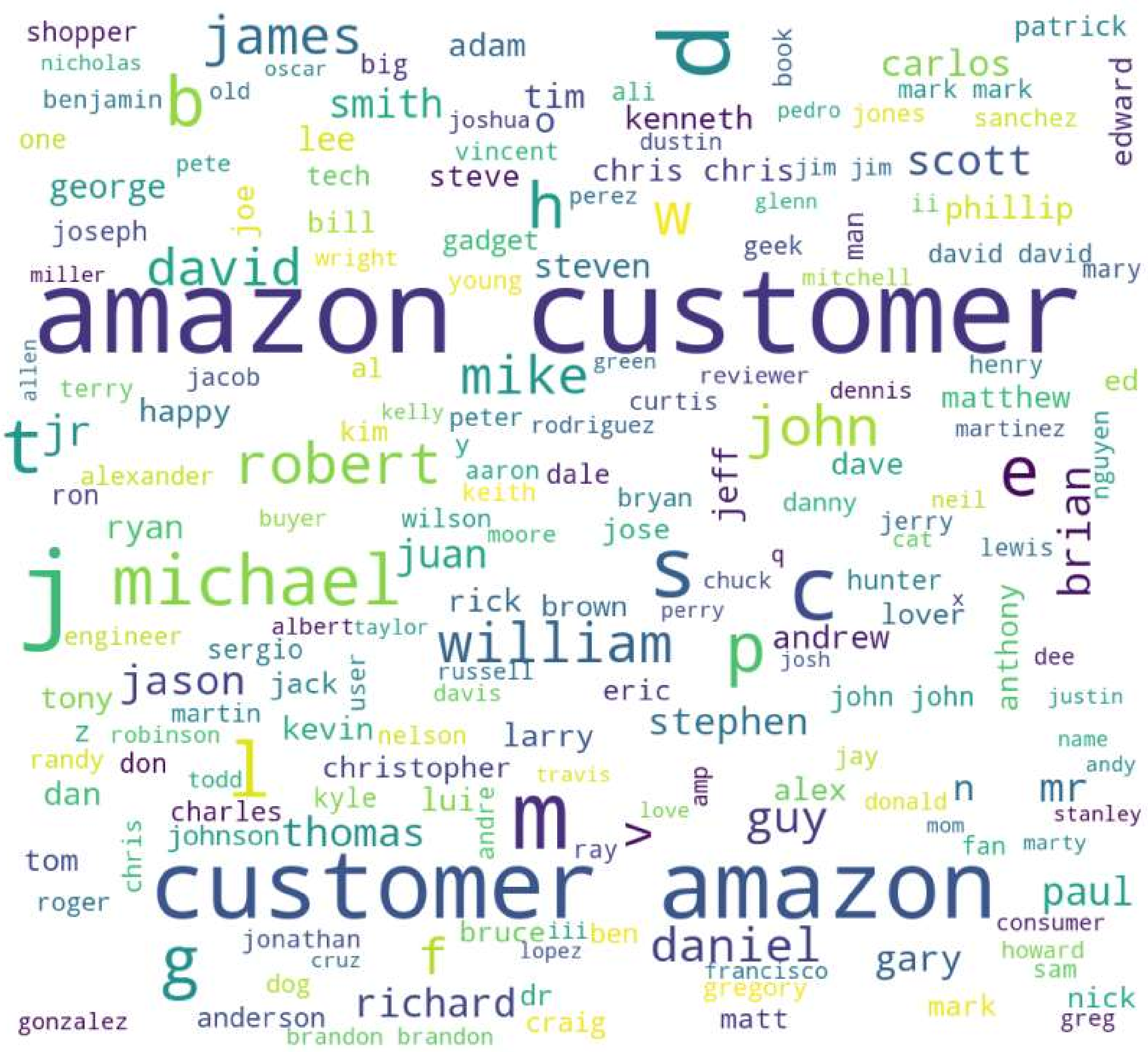
In [23]:
```
comment_words = ''
stopwords = set(STOPWORDS)

# iterate through the csv file
for val in df.reviewText:

    # typecaste each val to string
    val = str(val)

    # split the value
    tokens = val.split()

    # Converts each token into lowercase
```

```
    for i in range(len(tokens)):
        tokens[i] = tokens[i].lower()

    comment_words += " ".join(tokens)+" "

wordcloud = WordCloud(width = 800, height = 800,
                background_color ='black',
                stopwords = stopwords,
                min_font_size = 10).generate(comment_words)

# plot the WordCloud image
plt.figure(figsize = (8, 8), facecolor = None)
plt.imshow(wordcloud)
plt.axis("off")
plt.tight_layout(pad = 0)

plt.show()
```



In [24]:
```
comment_words = ''
stopwords = set(STOPWORDS)

# iterate through the csv file
for val in df.reviewerName:

    # typecaste each val to string
    val = str(val)
```

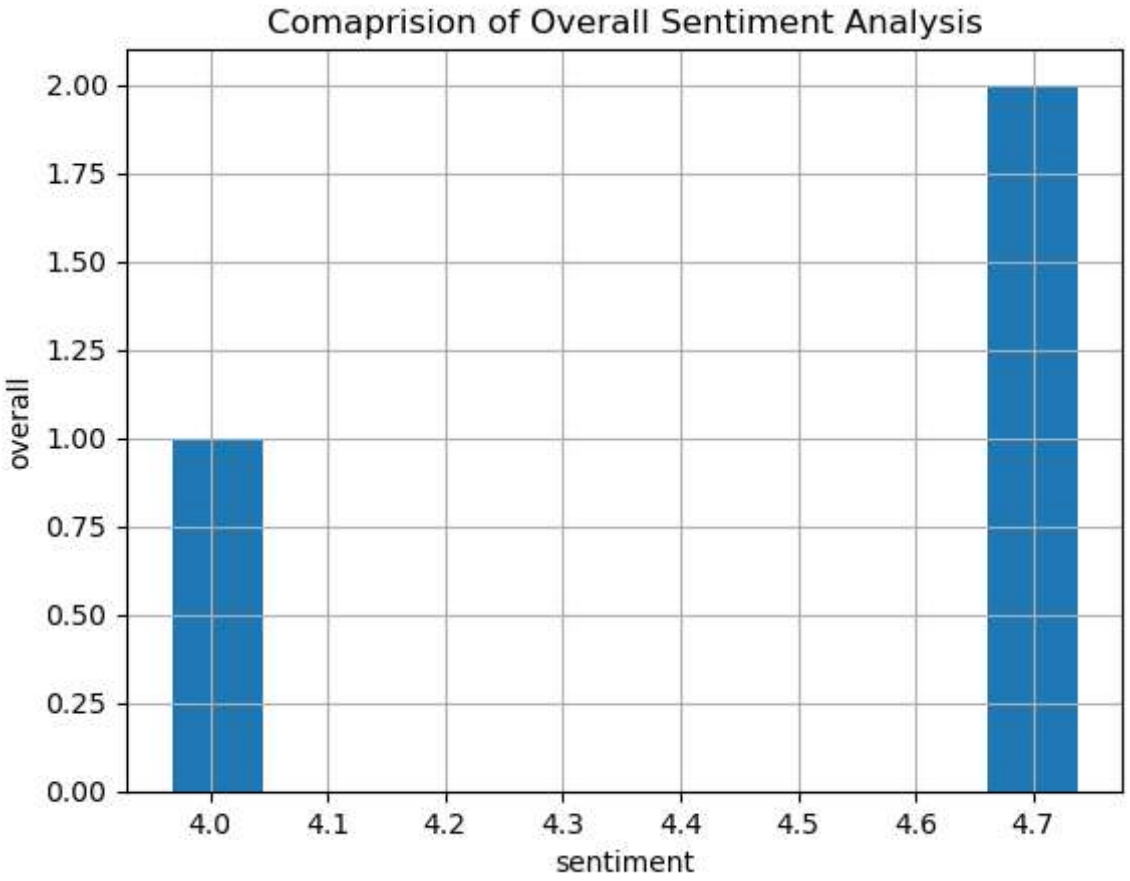
```
    # split the value
    tokens = val.split()

    # Converts each token into lowercase
    for i in range(len(tokens)):
        tokens[i] = tokens[i].lower()

    comment_words += " ".join(tokens)+" "

wordcloud = WordCloud(width = 1000, height = 900,
                background_color ='white',
                stopwords = stopwords,
                min_font_size = 12).generate(comment_words)

# plot the WordCloud image
plt.figure(figsize = (8, 8), facecolor = None)
plt.imshow(wordcloud)
plt.axis("off")
plt.tight_layout(pad = 0)

plt.show()
```



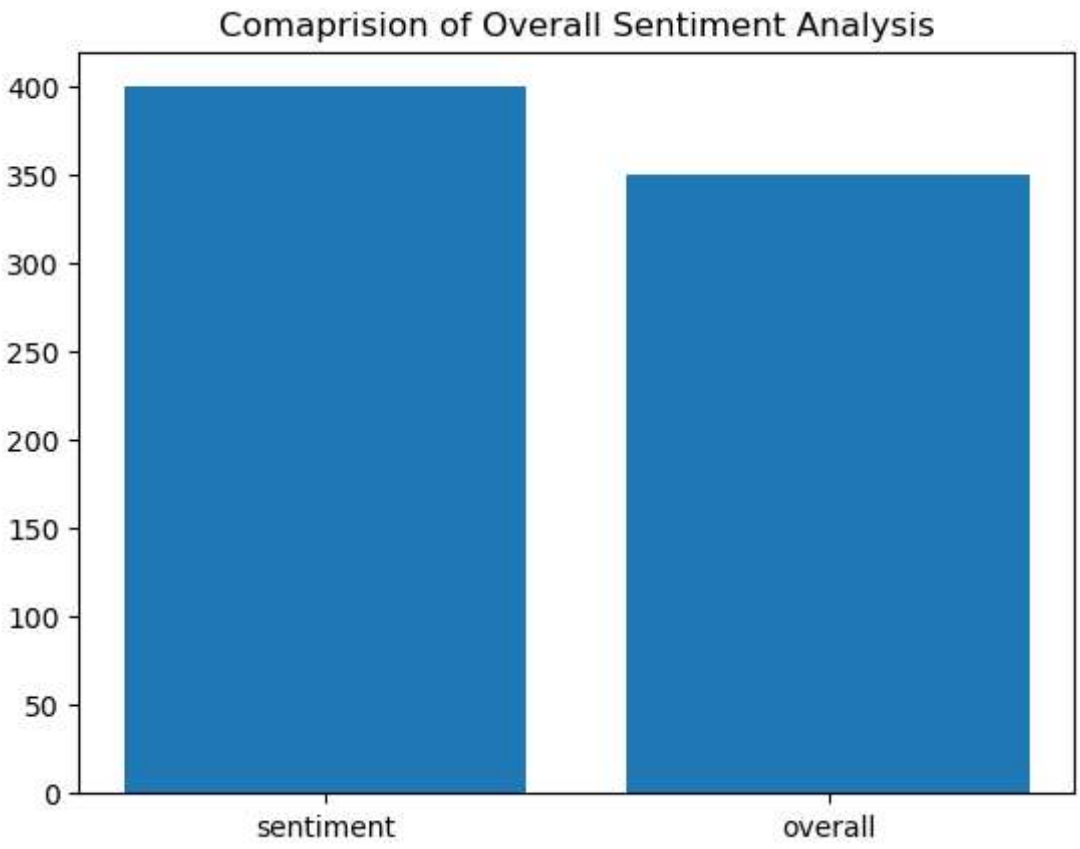
Generating Histograms

In [25]:
```
# create 2D array of table given above
data = [['Negative', 3.968085],
        ['Neutral', 4.688645],
```

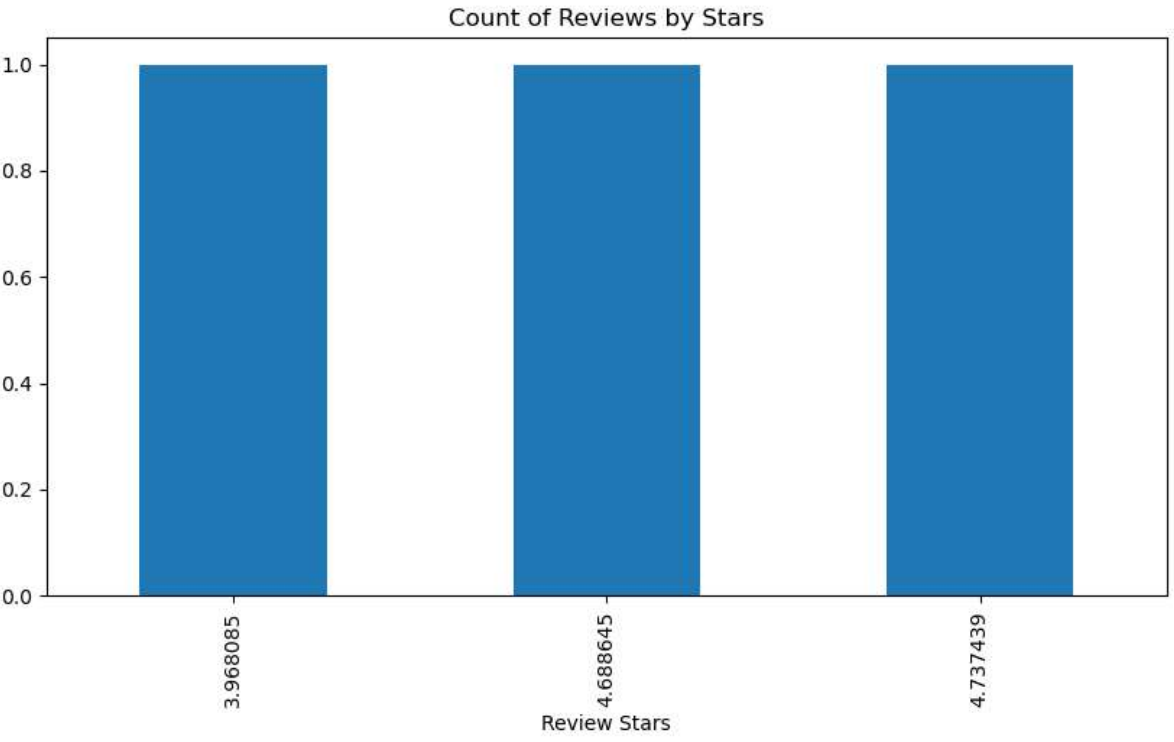
```
        ['Positive', 4.737439]]

# dataframe created with
# the above data array
df = pd.DataFrame(data, columns = ['sentiment', 'overall'] )

# create histogram for numeric data
df.hist()

plt.xlabel("sentiment")
plt.ylabel("overall")
plt.title("Comaprision of Overall Sentiment Analysis")

# show plot
plt.show()
```



In [26]:
```
x = ["sentiment", "overall"]
y = [400, 350]
plt.title("Comaprision of Overall Sentiment Analysis")

plt.bar(x, y)
plt.show()
```



In [27]:
```
#create bar plot from data
ax = df['overall'].value_counts().sort_index() \
    .plot(kind='bar',
          title='Count of Reviews by Stars',
          figsize=(10, 5))

ax.set_xlabel('Review Stars')
plt.show()
```



In [ ]: