```
import numpy as np
import pandas as pd
import nltk
from nltk.sentiment.vader import SentimentIntensityAnalyzer
import re
from textblob import TextBlob
from  wordcloud import WordCloud
import seaborn as sns
import matplotlib.pyplot as plt
import cufflinks as cf
%matplotlib inline
from plotly.offline import init_notebook_mode
init_notebook_mode(connected=True)
cf.go_offline();
import plotly.graph_objects as go
from plotly.subplots import make_subplots
import warnings
warnings.filterwarnings('ignore')
warnings.warn('this will show')
pd.set_option('display.max_columns',None)
```

```
df=pd.read_csv('amazon.csv')
df.head()
```

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

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

	reviewerName	overall	reviewText	reviewTime	day_diff	helpful_
2031	Hyoun Kim "Faluzure"	5	[[UPDATE - 6/19/2014]]So my lovely wife boug...	05-01-2013	702	19
3449	NLee the Engineer	5	I have tested dozens of SDHC and micro-SDHC ca...	26-09-2012	803	14
4212	SkincareCEO	1	NOTE: please read the last update (scroll to ...	08-05-2013	579	15
	Amazon		If your card gets hot			

```
def missinng_analysis(df):
  mi_columns=[col for col in df.columns if df[col].isnull().sum()==0]
  n_miss=df[mi_columns].isnull().sum().sort_values(ascending=True)
  ratio=(df[mi_columns].isnull().sum()/df.shape[0]*100).sort_values(ascending=True)
  missing_df=pd.concat([n_miss,np.round(ratio,2)],axis=1,keys=['missinfvalues','ratio'])
  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(62,'-'))
  print(df.dtypes)
  print(''.center(82,'-'))
  print(missinng_analysis(df))
  print('DUPLICATE VALUES'.center(83,'-'))
  print(df.duplicated().sum())
  print('QUARTILES'.center(82,'-'))
  print(df.quantile([0,0.05,0.50,0,.95,0.99,1]).T)


check_dataframe(df)
```

```
--------------------------------------SHAPE---------------------------------------
rows:()
columns:()
----------------------------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
----------------------------------------------------------------------------------
                      missinfvalues  ratio
overall                           0    0.0
reviewTime                        0    0.0
day_diff                          0    0.0
helpful_yes                       0    0.0
helpful_no                        0    0.0
total_vote                        0    0.0
score_pos_neg_diff                0    0.0
score_average_rating              0    0.0
wilson_lower_bound                0    0.0
----------------------------------DUPLICATE VALUES---------------------------------
0
------------------------------------QUARTILES-------------------------------------
                       0.00  0.05   0.50   0.00        0.95       0.99  \
overall                 1.0   2.0    5.0    1.0    5.000000    5.00000
day_diff                1.0  98.0  431.0    1.0  748.000000  943.00000
helpful_yes             0.0   0.0    0.0    0.0    1.000000    3.00000
helpful_no              0.0   0.0    0.0    0.0    0.000000    2.00000
total_vote              0.0   0.0    0.0    0.0    1.000000    4.00000
score_pos_neg_diff   -130.0   0.0    0.0 -130.0    1.000000    2.00000
```

```
score_average_rating    0.0   0.0    0.0    0.0    1.000000    1.00000
wilson_lower_bound      0.0   0.0    0.0    0.0    0.206549    0.34238

                           1.00
overall                5.000000
day_diff            1064.000000
helpful_yes         1952.000000
helpful_no           183.000000
total_vote          2020.000000
score_pos_neg_diff  1884.000000
score_average_rating   1.000000
wilson_lower_bound     0.957544
```

```
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)
```

	variable	classes
0	reviewText	4912

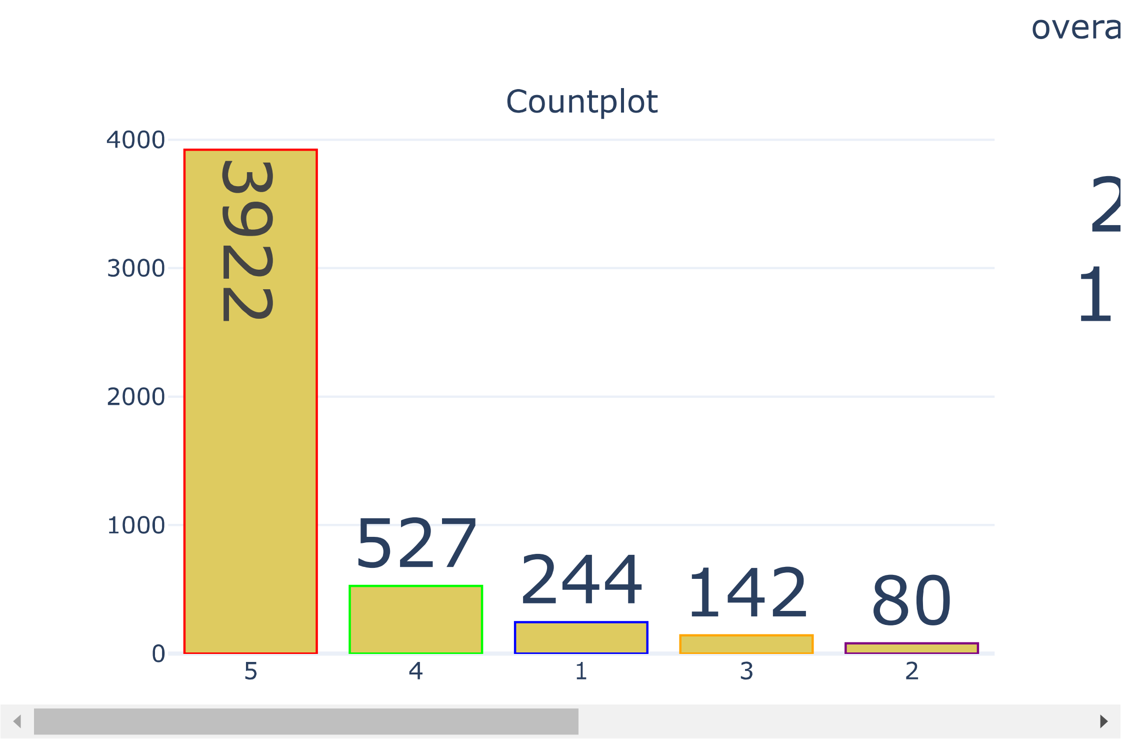
```
!pip install plotly
```

```
Requirement already satisfied: plotly in /usr/local/lib/python3.10/dist-packages (5.13.1)
Requirement already satisfied: tenacity>=6.2.0 in /usr/local/lib/python3.10/dist-packages (from plotly) (8.2.2)
```

```
4    wilson_lower_bound    40
```

```
constraints=['#FF0000', '#00FF00', '#0000FF', '#FFA500', '#800080']

import plotly.io as pio
from IPython.display import display, HTML
import plotly.offline as pyo




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

    fig.add_trace(
        go.Bar(
            y=df[column_name].value_counts().values.tolist(),
            x=[str(s) for s in df[column_name].value_counts().index],
            text=df[column_name].value_counts().values.tolist(),
            textfont=dict(size=34),
            name=column_name,
            textposition='auto',
            showlegend=False,
            marker=dict(color='#decb60', line=dict(color=constraints, 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=38),
            textposition='auto',
            showlegend=False,
            name=column_name,
            marker=dict(colors=['#decb60', 'lightgrey', 'darkblue', 'orange'])
        ),
        row=1, col=2
    )

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

    # pio.show(fig)
    html_file = f"{column_name}_plot.html"
    pyo.plot(fig, filename=html_file, auto_open=False)

    display(HTML(html_file))
```

```
categorical_variable_summary(df, 'overall')
```



```
df.reviewText.head()
```

```
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
```

```
review_example=df.reviewText[2031]
review_example
```

```
'[[ 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 w
ith Samsung products, the Galaxy Tab 4 has the ability to add a microSD c
ard 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 bes
t (performance-wise) are the as follows:SanDiskSanDisk UltraSanDisk Ultra
```

```
review_example=review_example.lower().split()
review_example
```

```
  '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'.
```

```
rt=lambda x: re.sub('[a-z-Z]',' ',str(x))
df['reviewText']=df['reviewText'].map(rt)
df['reviewText']=df['reviewText'].str.lower()
df.head()
```

	reviewerName	overall	reviewText	reviewTime	day_diff	helpful_yes
2031	Hyoun Kim "Faluzure"	5	[[update 6/19/2014]]s ...	05-01-2013	702	1952
3449	NLee the Engineer	5	i sdhc sdhc ...	26-09-2012	803	1428
4212	SkincareCEO	1	note: (...	08-05-2013	579	1568

```
pip install vaderSentiment
```

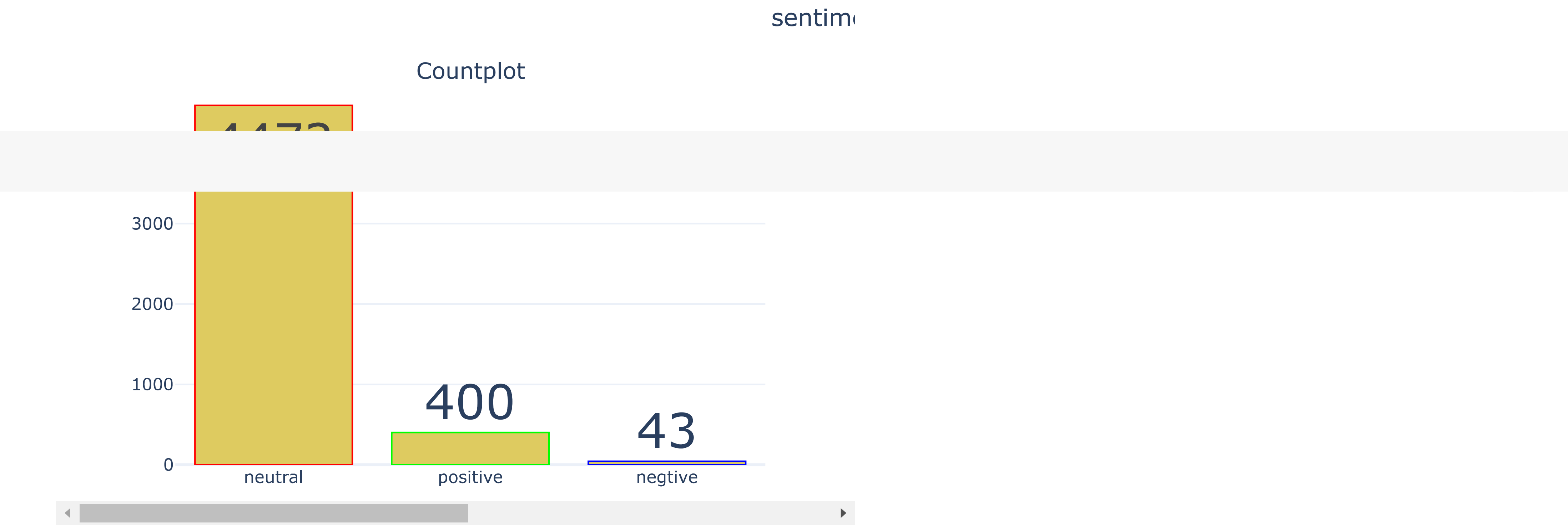
```
Collecting vaderSentiment
  Downloading vaderSentiment-3.3.2-py2.py3-none-any.whl (125 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 126.0/126.0 kB 5.2 MB/s eta 0:00:00
Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-packages (from vaderSentiment) (2.27.1)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in /usr/local/lib/python3.10/dist-packages (from requests->vaderSentiment) (1.26.16)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-packages (from requests->vaderSentiment) (2023.5.7)
Requirement already satisfied: charset-normalizer~=2.0.0 in /usr/local/lib/python3.10/dist-packages (from requests->vaderSentiment) (2.0.12)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests->vaderSentiment) (3.4)
Installing collected packages: vaderSentiment
Successfully installed vaderSentiment-3.3.2
```

```


from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
df[['polarity','subjectivity']]=df['reviewText'].apply(lambda Text:pd.Series(TextBlob(Text).sentiment))

for index,row in df['reviewText'].iteritems():
  score=SentimentIntensityAnalyzer().polarity_scores(row)
  neg=score['neg']
  pos=score['pos']
  neu=score['neu']
  if neg>pos:
    df.loc[index,'sentiment']='negtive'
  elif pos >neg:
    df.loc[index,'sentiment']='positive'
  else:
    df.loc[index,'sentiment']='neutral'
```

```
df[df['sentiment']=='positive'].sort_values('wilson_lower_bound',ascending=False).head(3)
```

	reviewerName	overall	reviewText	reviewTime	day_diff	helpful_yes
2031	Hyoun Kim "Faluzure"	5	[[update 6/19/2014]]s ...	05-01-2013	702	1952
3449	NLee the Engineer	5	i sdhc sdhc ...	26-09-2012	803	1428
4212	SkincareCEO	1	note: (...	08-05-2013	579	1568

```
categorical_variable_summary(df,'sentiment')
```



Colab paid products - Cancel contracts here

✓ 2s completed at 8:46 AM

Could not connect to the reCAPTCHA service. Please check your internet connection and reload to get a reCAPTCHA challenge.