```python
import numpy as np
import pandas as pd

# df = pd.read_csv('spam.csv')
df=pd.read_csv("C:\\Users\\hp\\Downloads\spam.csv",encoding="latin1")
df
```

```
        v1                                                 v2 Unnamed:
2  \
0      ham  Go until jurong point, crazy.. Available only ...
NaN
1      ham                      Ok lar... Joking wif u oni...
NaN
2     spam  Free entry in 2 a wkly comp to win FA Cup fina...
NaN
3      ham  U dun say so early hor... U c already then say...
NaN
4      ham  Nah I don't think he goes to usf, he lives aro...
NaN
...    ...                                                ...           .
..
5567  spam  This is the 2nd time we have tried 2 contact u...
NaN
5568   ham              Will Ì_ b going to esplanade fr home?
NaN
5569   ham  Pity, * was in mood for that. So...any other s...
NaN
5570   ham  The guy did some bitching but I acted like i'd...
NaN
5571   ham                         Rofl. Its true to its name
NaN

     Unnamed: 3 Unnamed: 4
0           NaN        NaN
1           NaN        NaN
2           NaN        NaN
3           NaN        NaN
4           NaN        NaN
...         ...        ...
5567        NaN        NaN
5568        NaN        NaN
5569        NaN        NaN
5570        NaN        NaN
5571        NaN        NaN

[5572 rows x 5 columns]
```

```python
df.sample(5)
```

```
        v1                                               v2 Unnamed:
2  \
138   spam  You'll not rcv any more msgs from the chat svc...
NaN
2508   ham                                              Ok...
NaN
3446   ham  Sitting ard nothing to do lor. U leh busy w work?
NaN
1214   ham  I'll text now! All creepy like so he won't thi...
NaN
997    ham  Not a lot has happened here. Feels very quiet....
NaN

     Unnamed: 3 Unnamed: 4
138         NaN        NaN
2508        NaN        NaN
3446        NaN        NaN
1214        NaN        NaN
997         NaN        NaN
```

```
df.shape
```

```
(5572, 5)
```

```
# 1. Data cleaning
# 2. EDA
# 3. Text Preprocessing
# 4. Model building
```

1. Data Cleaning

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5572 entries, 0 to 5571
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   v1          5572 non-null   object
 1   v2          5572 non-null   object
 2   Unnamed: 2  50 non-null     object
 3   Unnamed: 3  12 non-null     object
 4   Unnamed: 4  6 non-null      object
dtypes: object(5)
memory usage: 217.8+ KB
```

```
# drop last 3 cols
df.drop(columns=['Unnamed: 2','Unnamed: 3','Unnamed: 4'],inplace=True)
```

```
df.sample(5)
```

```
       v1                                                 v2
172   ham                    What time you coming down later?
5279  ham  Helloooo... Wake up..! \Sweet\" \"morning\" \"...
4267  ham                     Hey so whats the plan this sat?
2215  ham       Prabha..i'm soryda..realy..frm heart i'm sory
1669  ham  Very hurting n meaningful lines ever: \I compr...
```

```
# renaming the cols
df.rename(columns={'v1':'target','v2':'text'},inplace=True)
df.sample(5)
```

```
     target                                               text
1893    ham                        Good Morning plz call me sir
2039    ham  Dont pack what you can buy at any store.like c...
3882    ham  Gumby's has a special where a  &lt;#&gt; \ che...
92      ham  Smile in Pleasure Smile in Pain Smile when tro...
4679    ham  That's cool he'll be here all night, lemme kno...
```

```
from sklearn.preprocessing import LabelEncoder
encoder = LabelEncoder()

df['target'] = encoder.fit_transform(df['target'])

df.head()
```

```
   target                                               text
0       0  Go until jurong point, crazy.. Available only ...
1       0                      Ok lar... Joking wif u oni...
2       1  Free entry in 2 a wkly comp to win FA Cup fina...
3       0  U dun say so early hor... U c already then say...
4       0  Nah I don't think he goes to usf, he lives aro...
```

```
# missing values
df.isnull().sum()
```

```
target    0
text      0
dtype: int64
```

```
# check for duplicate values
df.duplicated().sum()
```

```
403
```

```
# remove duplicates
df = df.drop_duplicates(keep='first')

df.duplicated().sum()
```

```
0
```

```
df.shape
```

```
(5169, 2)
```

2.EDA

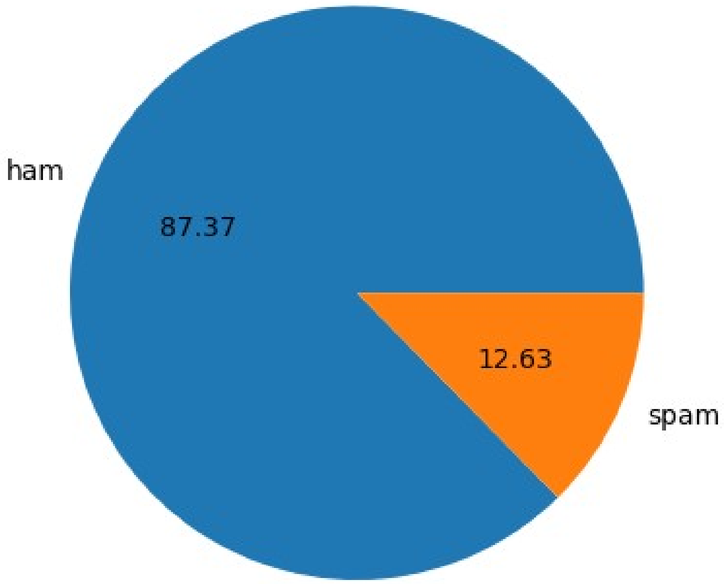
```
df.head()
```

```
   target                                               text
0       0  Go until jurong point, crazy.. Available only ...
1       0                      Ok lar... Joking wif u oni...
2       1  Free entry in 2 a wkly comp to win FA Cup fina...
3       0  U dun say so early hor... U c already then say...
4       0  Nah I don't think he goes to usf, he lives aro...
```

```
df['target'].value_counts()
```

```
target
0    4516
1     653
Name: count, dtype: int64
```

```
import matplotlib.pyplot as plt
plt.pie(df['target'].value_counts(),
labels=['ham','spam'],autopct="%0.2f")
plt.show()
```



```
# Data is imbalanced
```

```python
import nltk
```

```python
!pip install nltk
```

```python
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\hp\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

```
True
```

```python
df['num_characters'] = df['text'].apply(len)
```

```
C:\Users\hp\AppData\Local\Temp\ipykernel_26344\253964734.py:1:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation:
https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#
returning-a-view-versus-a-copy
  df['num_characters'] = df['text'].apply(len)
```

```python
df.head()
```

```
   target                                               text
num_characters
0       0  Go until jurong point, crazy.. Available only ...
111
1       0                      Ok lar... Joking wif u oni...
29
2       1  Free entry in 2 a wkly comp to win FA Cup fina...
155
3       0  U dun say so early hor... U c already then say...
49
4       0  Nah I don't think he goes to usf, he lives aro...
61
```

```python
# num of words
df['num_words'] = df['text'].apply(lambda
x:len(nltk.word_tokenize(x)))
```

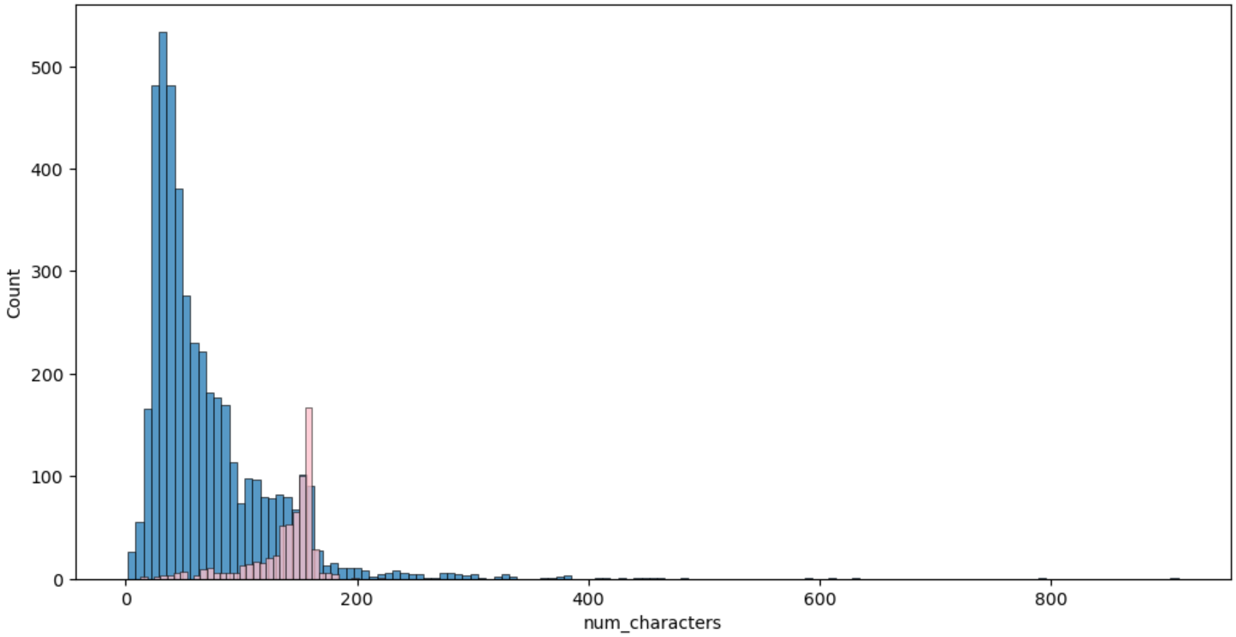
```
C:\Users\hp\AppData\Local\Temp\ipykernel_26344\192676766.py:2:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation:
https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#
returning-a-view-versus-a-copy
```

```
  df['num_words'] = df['text'].apply(lambda
x:len(nltk.word_tokenize(x)))

df.head()

   target                                               text
num_characters  \
0       0  Go until jurong point, crazy.. Available only ...
111
1       0                      Ok lar... Joking wif u oni...
29
2       1  Free entry in 2 a wkly comp to win FA Cup fina...
155
3       0  U dun say so early hor... U c already then say...
49
4       0  Nah I don't think he goes to usf, he lives aro...
61

   num_words
0         24
1          8
2         37
3         13
4         15

# num of sentences
df['num_sentences'] = df['text'].apply(lambda
x:len(nltk.sent_tokenize(x)))

C:\Users\hp\AppData\Local\Temp\ipykernel_26344\3097215481.py:2:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation:
https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#
returning-a-view-versus-a-copy
  df['num_sentences'] = df['text'].apply(lambda
x:len(nltk.sent_tokenize(x)))

df.head()

   target                                               text
num_characters  \
0       0  Go until jurong point, crazy.. Available only ...
111
1       0                      Ok lar... Joking wif u oni...
29
2       1  Free entry in 2 a wkly comp to win FA Cup fina...
155
3       0  U dun say so early hor... U c already then say...
```

```
49
4        0  Nah I don't think he goes to usf, he lives aro...
61

   num_words  num_sentences
0         24              2
1          8              2
2         37              2
3         13              1
4         15              1

df[['num_characters','num_words','num_sentences']].describe()

       num_characters    num_words  num_sentences
count     5169.000000  5169.000000    5169.000000
mean        78.977945    18.455794       1.965564
std         58.236293    13.324758       1.448541
min          2.000000     1.000000       1.000000
25%         36.000000     9.000000       1.000000
50%         60.000000    15.000000       1.000000
75%        117.000000    26.000000       2.000000
max        910.000000   220.000000      38.000000

# ham
df[df['target'] == 0]
[['num_characters','num_words','num_sentences']].describe()

       num_characters    num_words  num_sentences
count     4516.000000  4516.000000    4516.000000
mean        70.459256    17.123782       1.820195
std         56.358207    13.493970       1.383657
min          2.000000     1.000000       1.000000
25%         34.000000     8.000000       1.000000
50%         52.000000    13.000000       1.000000
75%         90.000000    22.000000       2.000000
max        910.000000   220.000000      38.000000

#spam
df[df['target'] == 1]
[['num_characters','num_words','num_sentences']].describe()

       num_characters   num_words  num_sentences
count      653.000000  653.000000     653.000000
mean       137.891271   27.667688       2.970904
std         30.137753    7.008418       1.488425
min         13.000000    2.000000       1.000000
25%        132.000000   25.000000       2.000000
50%        149.000000   29.000000       3.000000
75%        157.000000   32.000000       4.000000
max        224.000000   46.000000       9.000000
```

```
import seaborn as sns

plt.figure(figsize=(12,6))
sns.histplot(df[df['target'] == 0]['num_characters'])
sns.histplot(df[df['target'] == 1]['num_characters'],color='pink')
```

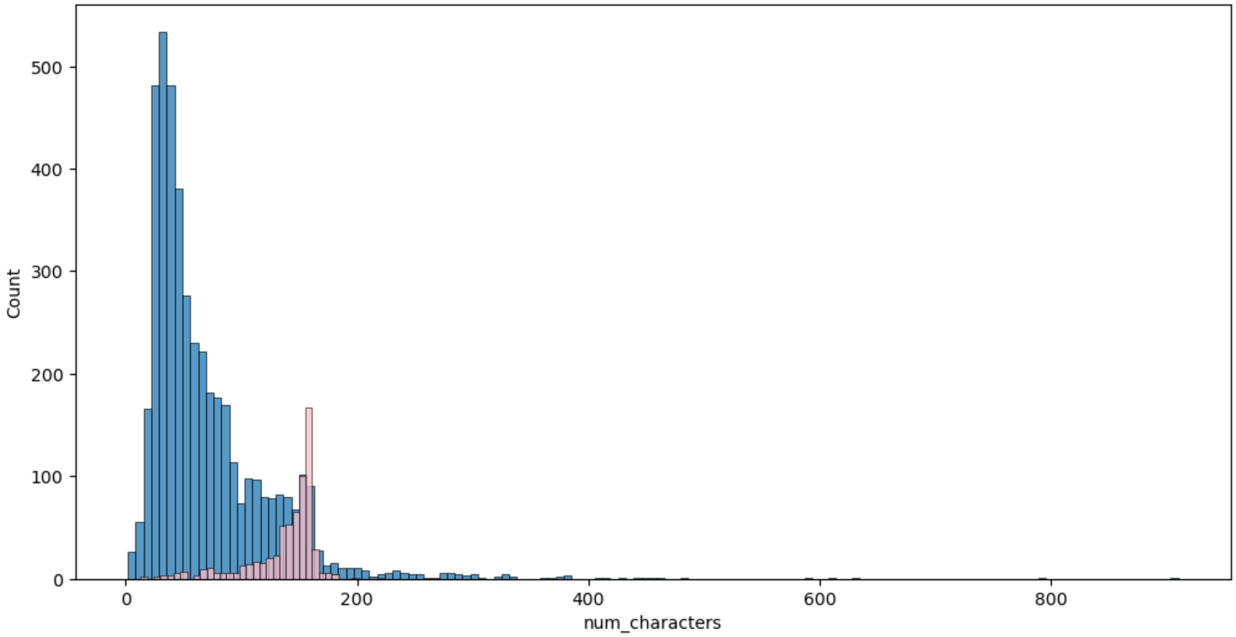
```
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed
in a future version. Convert inf values to NaN before operating
instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed
in a future version. Convert inf values to NaN before operating
instead.
  with pd.option_context('mode.use_inf_as_na', True):

<Axes: xlabel='num_characters', ylabel='Count'>
```



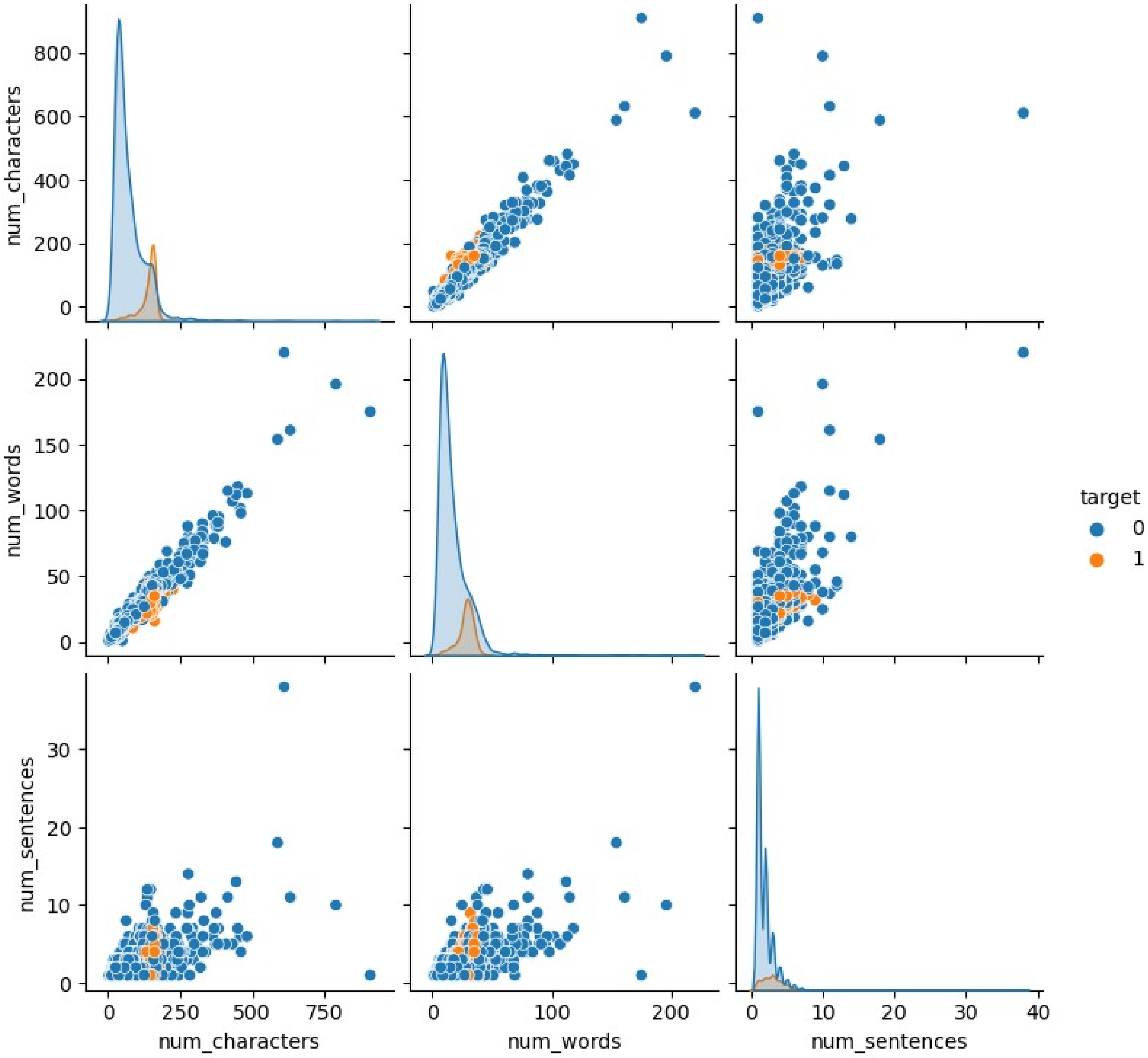
```
plt.figure(figsize=(12,6))
sns.histplot(df[df['target'] == 0]['num_characters'])
sns.histplot(df[df['target'] == 1]['num_characters'],color='pink')
```

```
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed
in a future version. Convert inf values to NaN before operating
instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed
```

```
in a future version. Convert inf values to NaN before operating
instead.
  with pd.option_context('mode.use_inf_as_na', True):

<Axes: xlabel='num_characters', ylabel='Count'>
```



```
sns.pairplot(df,hue='target')

C:\ProgramData\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed
in a future version. Convert inf values to NaN before operating
instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed
in a future version. Convert inf values to NaN before operating
instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed
in a future version. Convert inf values to NaN before operating
instead.
  with pd.option_context('mode.use_inf_as_na', True):

<seaborn.axisgrid.PairGrid at 0x21461793b50>
```



3. Data Preprocessing

- Lower case
- Tokenization
- Removing special characters
- Removing stop words and punctuation
- Stemming

```
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
import string
import nltk
nltk.download('stopwords')
ps = PorterStemmer()
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\hp\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

```
def transform_text(text):
    text = text.lower()
    text = nltk.word_tokenize(text)

    y = []
    for i in text:
        if i.isalnum():
            y.append(i)

    text = y[:]
    y.clear()

    for i in text:
        if i not in stopwords.words('english') and i not in
string.punctuation:
            y.append(i)

    text = y[:]
    y.clear()

    for i in text:
        y.append(ps.stem(i))


    return " ".join(y)
```

```
transform_text("I'm gonna be home soon and i don't want to talk about
this stuff anymore tonight, k? I've cried enough today.")
```

```
'gon na home soon want talk stuff anymor tonight k cri enough today'
```

```
df['text'][10]
```

```
"I'm gonna be home soon and i don't want to talk about this stuff
anymore tonight, k? I've cried enough today."
```

```
from nltk.stem.porter import PorterStemmer
ps = PorterStemmer()
ps.stem('loving')
```

```
'love'
```

```
df['transformed_text'] = df['text'].apply(transform_text)
```

```
C:\Users\hp\AppData\Local\Temp\ipykernel_26344\283536690.py:1:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
```

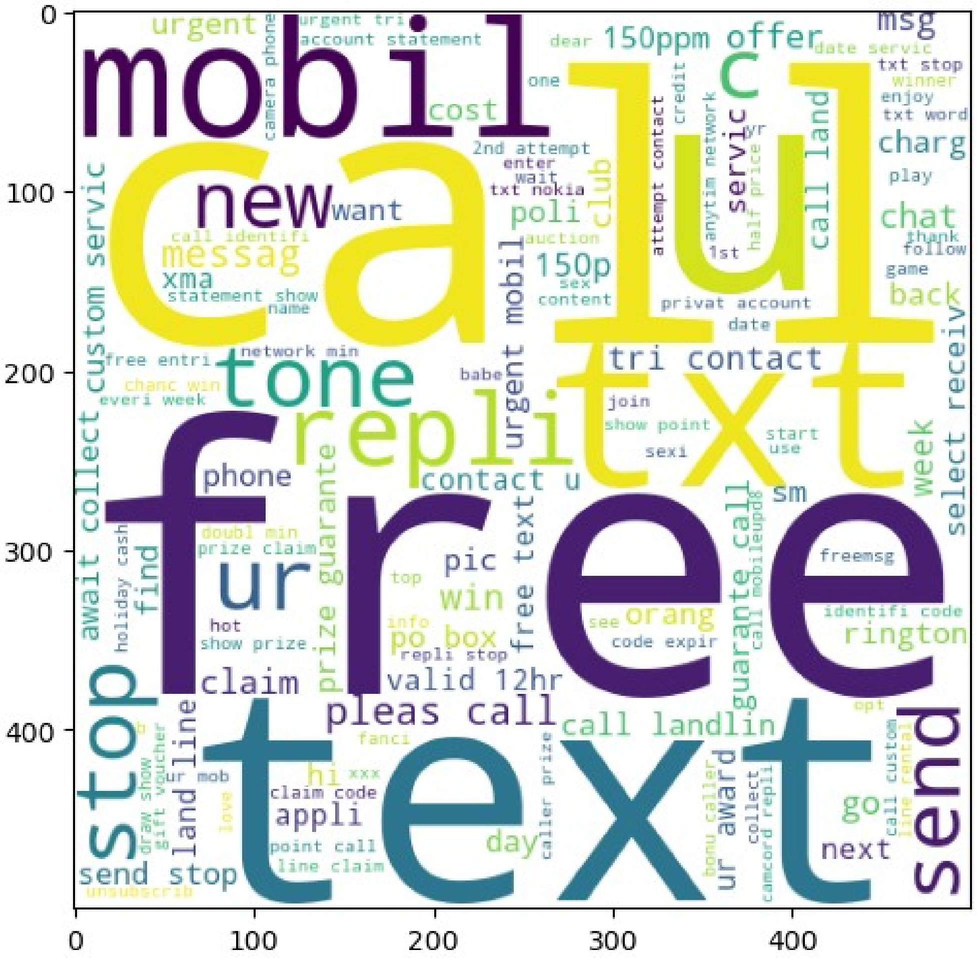
```
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation:
https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#
returning-a-view-versus-a-copy
  df['transformed_text'] = df['text'].apply(transform_text)
```

```python
df.head()
```

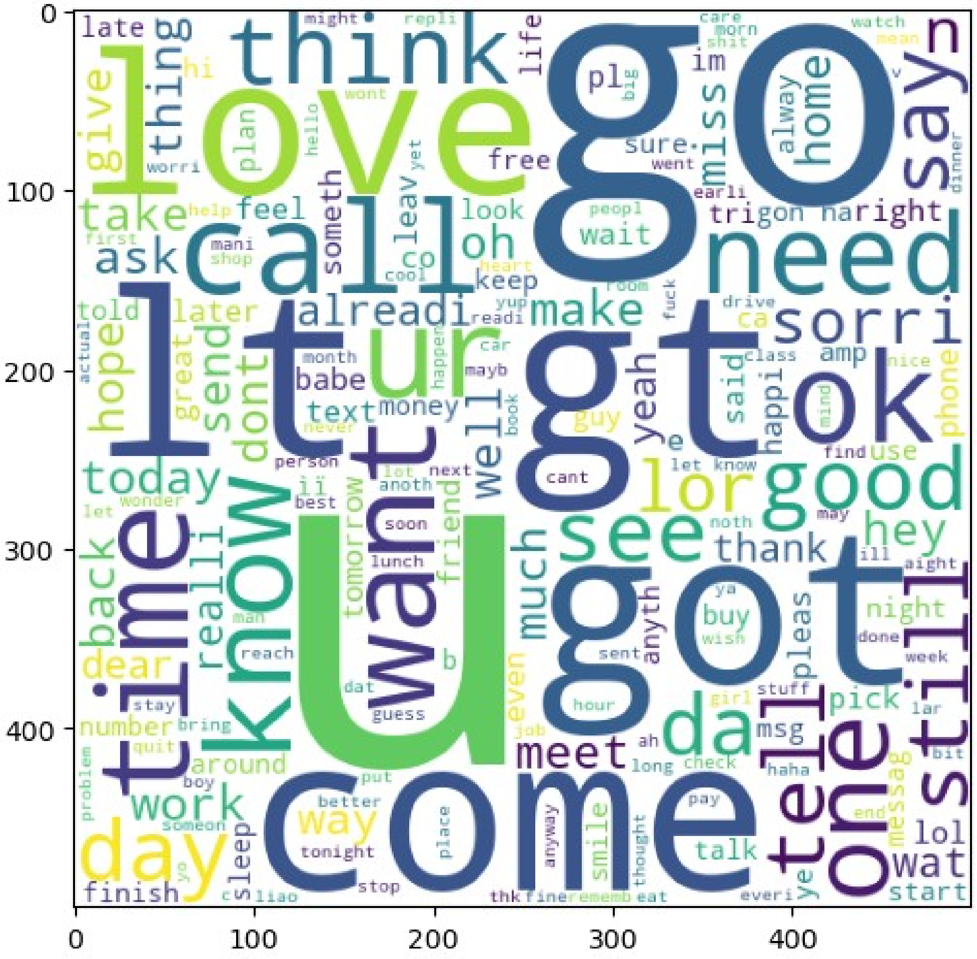
```
   target                                               text
num_characters  \
0       0  Go until jurong point, crazy.. Available only ...
111
1       0                      Ok lar... Joking wif u oni...
29
2       1  Free entry in 2 a wkly comp to win FA Cup fina...
155
3       0  U dun say so early hor... U c already then say...
49
4       0  Nah I don't think he goes to usf, he lives aro...
61

   num_words  num_sentences
transformed_text
0         24              2  go jurong point crazi avail bugi n great
world...
1          8              2                              ok lar joke
wif u oni
2         37              2  free entri 2 wkli comp win fa cup final
tkt 21...
3         13              1                u dun say earli hor u c
alreadi say
4         15              1               nah think goe usf live
around though
```

```python
from wordcloud import WordCloud
wc =
WordCloud(width=500,height=500,min_font_size=10,background_color='whit
e')

spam_wc = wc.generate(df[df['target'] == 1]
['transformed_text'].str.cat(sep=" "))

plt.figure(figsize=(15,6))
plt.imshow(spam_wc)
```

```
<matplotlib.image.AxesImage at 0x2146a216690>
```



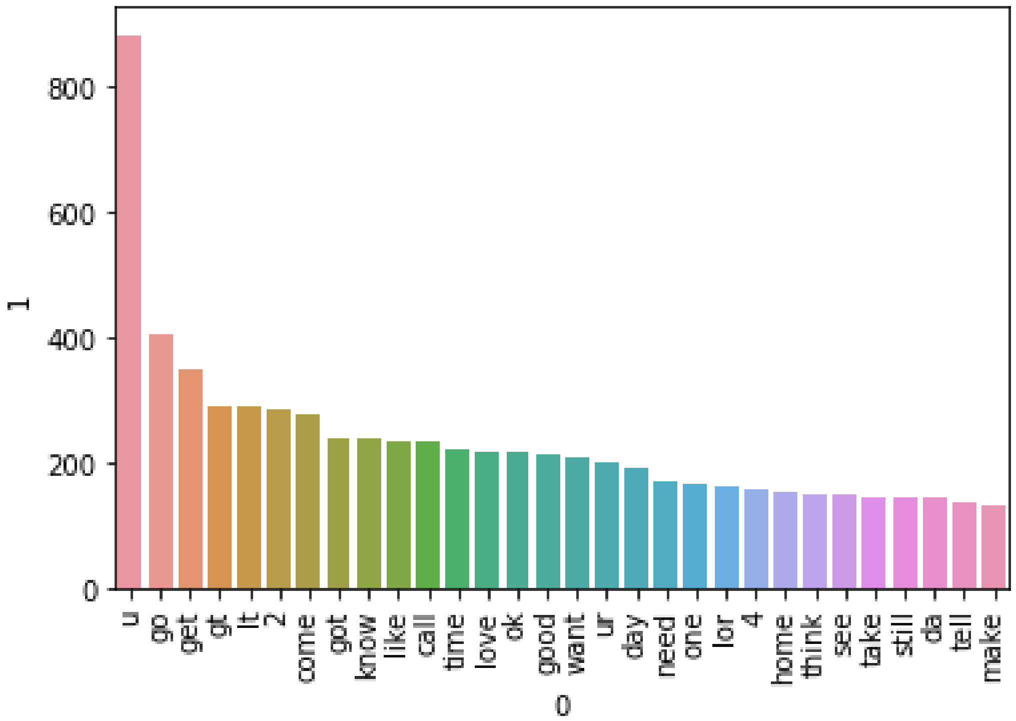
```
ham_wc = wc.generate(df[df['target'] == 0]
['transformed_text'].str.cat(sep=" "))

plt.figure(figsize=(15,6))
plt.imshow(ham_wc)

<matplotlib.image.AxesImage at 0x2146a15e290>
```



```
df.head()

   target                                               text  
num_characters  \
0       0  Go until jurong point, crazy.. Available only ...    
111   
1       0                      Ok lar... Joking wif u oni...     
29   
2       1  Free entry in 2 a wkly comp to win FA Cup fina...    
155   
3       0  U dun say so early hor... U c already then say...    
49   
4       0  Nah I don't think he goes to usf, he lives aro...    
61   

   num_words  num_sentences  
transformed_text  
0         24              2  go jurong point crazi avail bugi n great 
world...  
1          8              2                                ok lar joke
```

```
wif u oni
2         37              2  free entri 2 wkli comp win fa cup final
tkt 21...
3         13              1                u dun say earli hor u c
alreadi say
4         15              1               nah think goe usf live
around though

spam_corpus = []
for msg in df[df['target'] == 1]['transformed_text'].tolist():
    for word in msg.split():
        spam_corpus.append(word)


len(spam_corpus)

9939

from collections import Counter
sns.barplot(pd.DataFrame(Counter(spam_corpus).most_common(30))
[0],pd.DataFrame(Counter(spam_corpus).most_common(30))[1])
plt.xticks(rotation='vertical')
plt.show()

----------------------------------------------------------------------
-----
TypeError                                 Traceback (most recent call
last)
Cell In[62], line 2
      1 from collections import Counter
----> 2 sns.barplot(pd.DataFrame(Counter(spam_corpus).most_common(30))
[0],pd.DataFrame(Counter(spam_corpus).most_common(30))[1])
      3 plt.xticks(rotation='vertical')
      4 plt.show()

TypeError: barplot() takes from 0 to 1 positional arguments but 2 were
given

ham_corpus = []
for msg in df[df['target'] == 0]['transformed_text'].tolist():
    for word in msg.split():
        ham_corpus.append(word)

len(ham_corpus)

35303

from collections import Counter
sns.barplot(pd.DataFrame(Counter(ham_corpus).most_common(30))
[0],pd.DataFrame(Counter(ham_corpus).most_common(30))[1])
```

```
plt.xticks(rotation='vertical')
plt.show()

C:\Users\91842\anaconda3\lib\site-packages\seaborn\_decorators.py:36:
FutureWarning: Pass the following variables as keyword args: x, y.
From version 0.12, the only valid positional argument will be `data`,
and passing other arguments without an explicit keyword will result in
an error or misinterpretation.
  warnings.warn(
```



```
# Text Vectorization
# using Bag of Words
df.head()

   target                                               text
num_characters  \
0       0  Go until jurong point, crazy.. Available only ...
111
1       0                      Ok lar... Joking wif u oni...
29
2       1  Free entry in 2 a wkly comp to win FA Cup fina...
155
3       0  U dun say so early hor... U c already then say...
49
4       0  Nah I don't think he goes to usf, he lives aro...
61
```

```
   num_words  num_sentences
transformed_text
0         24              2  go jurong point crazi avail bugi n great
world...
1          8              2                              ok lar joke
wif u oni
2         37              2  free entri 2 wkli comp win fa cup final
tkt 21...
3         13              1                u dun say earli hor u c
alreadi say
4         15              1               nah think goe usf live
around though
```

4. Model Building

```
from sklearn.feature_extraction.text import
CountVectorizer,TfidfVectorizer
cv = CountVectorizer()
tfidf = TfidfVectorizer(max_features=3000)

X = tfidf.fit_transform(df['transformed_text']).toarray()

#from sklearn.preprocessing import MinMaxScaler
#scaler = MinMaxScaler()
#X = scaler.fit_transform(X)

# appending the num_character col to X
#X = np.hstack((X,df['num_characters'].values.reshape(-1,1)))

X.shape

(5169, 3000)

y = df['target'].values

from sklearn.model_selection import train_test_split

X_train,X_test,y_train,y_test =
train_test_split(X,y,test_size=0.2,random_state=2)

from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import
accuracy_score,confusion_matrix,precision_score

gnb = GaussianNB()

gnb.fit(X_train,y_train)
y_pred1 = gnb.predict(X_test)
print(accuracy_score(y_test,y_pred1))
print(confusion_matrix(y_test,y_pred1))
print(precision_score(y_test,y_pred1))
```

```
0.8694390715667312
[[788 108]
 [ 27 111]]
0.5068493150684932
```