In [1]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [142…

```
#List of possible encoding to try

encodings = ['utf-8', 'latin1','ISO-8859-1','cp1252']
file_path = 'sms_spam.csv' #Change this to the path of your CSV file
```

In [143…

```
  df
#Attemp to read the csv file with different encoding

for encoding in encodings:
    try:
        df = pd.read_csv(file_path,encoding=encoding)
        print(f"File successfully read with encoding: {encoding}")
        break #stop the Loop if successful
    except UnicodeDecodeError:
        print(f"Failed to read with encoding: {encoding}")
        continue #Try the next encoding

# if the loop completes without success, df will not be defined
if 'df' in locals():
    print("CSV file has been successfully loaded.")
else:
    print("All encoding attemps failed. Unable to read the CSV file.")
```

```
File successfully read with encoding: utf-8
CSV file has been successfully loaded.
```

In [20]:

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

Out[20]:

	type	text
1895	ham	I dled 3d its very imp
4747	ham	Been up to ne thing interesting. Did you have ...
3002	ham	Got it. Seventeen pounds for seven hundred ml ...
2885	ham	Ill call u 2mrw at ninish, with my address tha...
3443	ham	IM REALY SOZ IMAT MY MUMS 2NITE WHAT ABOUT 2MORO

In [21]:

```
df.shape
```

Out[21]:

```
(5559, 2)
```

Data Cleaning

In [25]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5559 entries, 0 to 5558
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   type    5559 non-null   object
 1   text    5559 non-null   object
dtypes: object(2)
memory usage: 87.0+ KB
```

In [28]:

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

Out[28]:

	type	text
977	ham	u takin linear algebra today?
4843	ham	Wat time u finish ur lect today?
288	ham	And is there a way you can send shade's stuff ...
1385	ham	Never blame a day in ur life. Good days give u...
1024	spam	Got what it takes 2 take part in the WRC Rally...

In [32]:

```
df.rename(columns={'type':'target'},inplace = True)
df.sample(5)
```

Out[32]:

	target	text
2879	ham	I had been hoping i would not have to send you...
4467	ham	Anyway i'm going shopping on my own now. Cos m...
4778	spam	You are being ripped off! Get your mobile cont...
5510	ham	Good evening! How are you?
1074	ham	Nah im goin 2 the wrks with j wot bout u?

In [30]:

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

In [33]:

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

In [34]:

```
df.head()
```

Out[34]:

	target	text
0	0	Hope you are having a good week. Just checking in
1	0	K..give back my thanks.
2	0	Am also doing in cbe only. But have to pay.
3	1	complimentary 4 STAR Ibiza Holiday or £10,000 ...
4	1	okmail: Dear Dave this is your final notice to...

In [35]:

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

Out[35]:

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

In [36]:

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

Out[36]:

```
403
```

In [37]:

```
df = df.drop_duplicates(keep = 'first')
```

In [38]:

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

Out[38]:

```
0
```

In [39]:

```
df.shape
```

Out[39]:

```
(5156, 2)
```

In [40]:

```
df.head()
```

Out[40]:

	target	text
0	0	Hope you are having a good week. Just checking in
1	0	K..give back my thanks.
2	0	Am also doing in cbe only. But have to pay.
3	1	complimentary 4 STAR Ibiza Holiday or £10,000 ...
4	1	okmail: Dear Dave this is your final notice to...

In [41]:

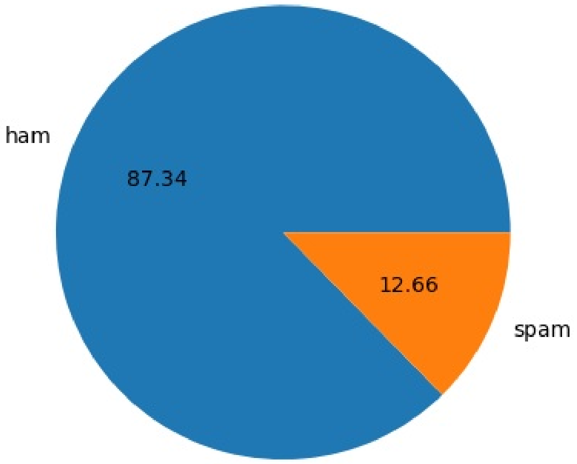
```
df['target'].value_counts()
```

Out[41]:

```
target
0    4503
1     653
Name: count, dtype: int64
```

In [42]:

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



In [43]:

```python
import nltk
```

In [46]:

```python
df['num_characters'] = df['text'].apply(len) #number of char
```

In [47]:

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

Out[47]:

	target	text	num_characters
0	0	Hope you are having a good week. Just checking in	49
1	0	K..give back my thanks.	23
2	0	Am also doing in cbe only. But have to pay.	43
3	1	complimentary 4 STAR Ibiza Holiday or £10,000 ...	149
4	1	okmail: Dear Dave this is your final notice to...	161

In [64]:

```python
df.tail()
```

Out[64]:

	target	text	num_characters
5554	0	You are a great role model. You are giving so ...	245
5555	0	Awesome, I remember the last time we got someb...	88
5556	1	If you don't, your prize will go to another cu...	145
5557	1	SMS. ac JSco: Energy is high, but u may not kn...	154
5558	0	Shall call now dear having food	31

In [71]:

```python
df.describe()
```

Out[71]:

	target	num_characters
count	5156.000000	5156.000000
mean	0.126649	78.658844
std	0.332611	57.615904
min	0.000000	2.000000
25%	0.000000	35.000000
50%	0.000000	60.000000
75%	0.000000	117.250000
max	1.000000	910.000000

In [72]:

```python
#targeting ham
df[df['target']==0][['num_characters']].describe()
```

Out[72]:

	num_characters
count	4503.000000
mean	70.104375
std	55.626601
min	2.000000
25%	33.000000
50%	52.000000
75%	90.000000
max	910.000000

In [73]:
```
#targeting ham
df[df['target']==1][['num_characters']].describe()
```

Out[73]:

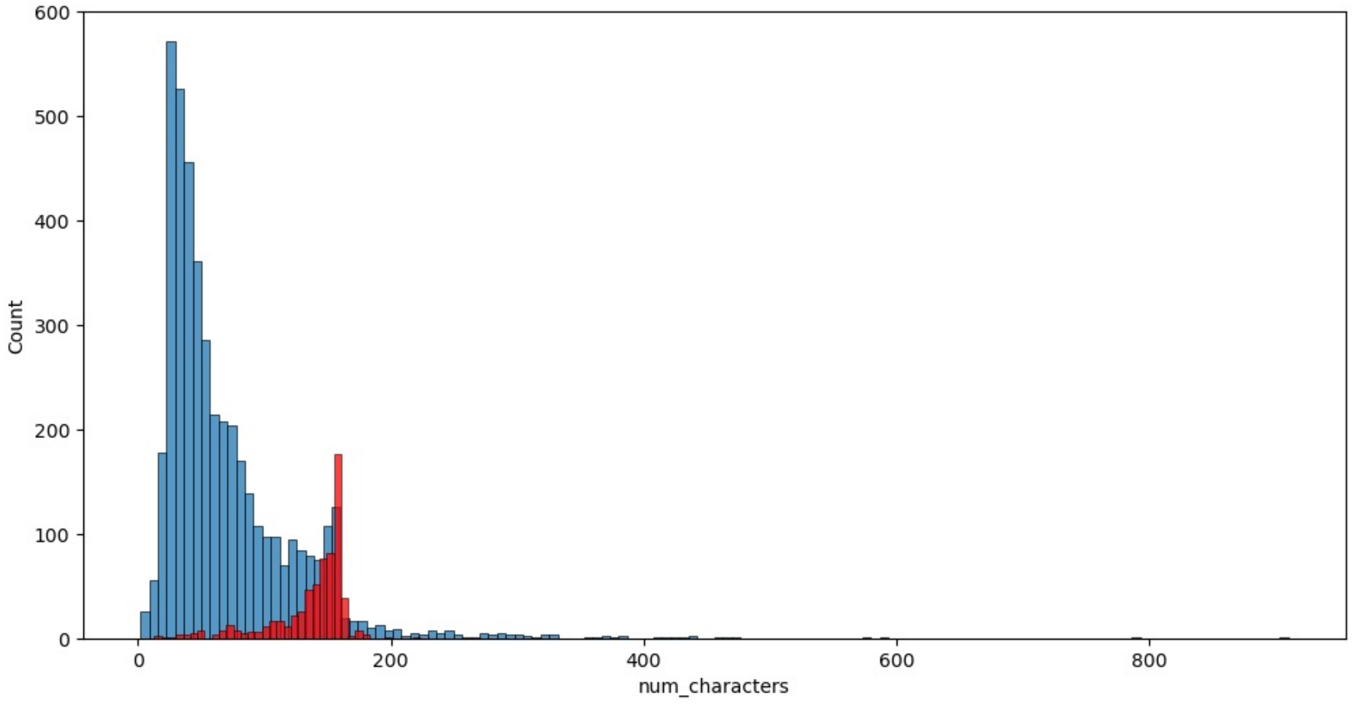
	num_characters
count	653.000000
mean	137.649311
std	29.825481
min	13.000000
25%	132.000000
50%	148.000000
75%	157.000000
max	223.000000

In [75]:
```
import seaborn as sns
```

In [76]:
```
plt.figure(figsize = (12,6))
sns.histplot(df[df['target'] == 0]['num_characters'])
sns.histplot(df[df['target'] == 1]['num_characters'],color = 'red')
```

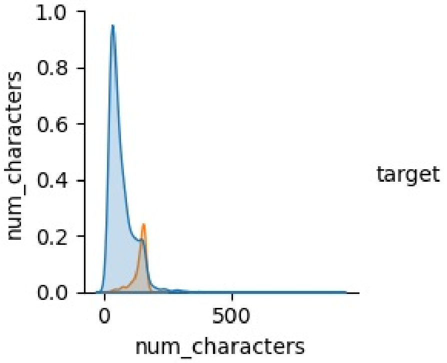
```
C:\Users\Sukesh\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is dep
recated and will be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Sukesh\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119: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

Out[76]:
```
<Axes: xlabel='num_characters', ylabel='Count'>
```



In [110…
```
# relation of columns
sns.pairplot(df,hue='target')
plt.show()
```

```
C:\Users\Sukesh\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is dep
recated 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):
```

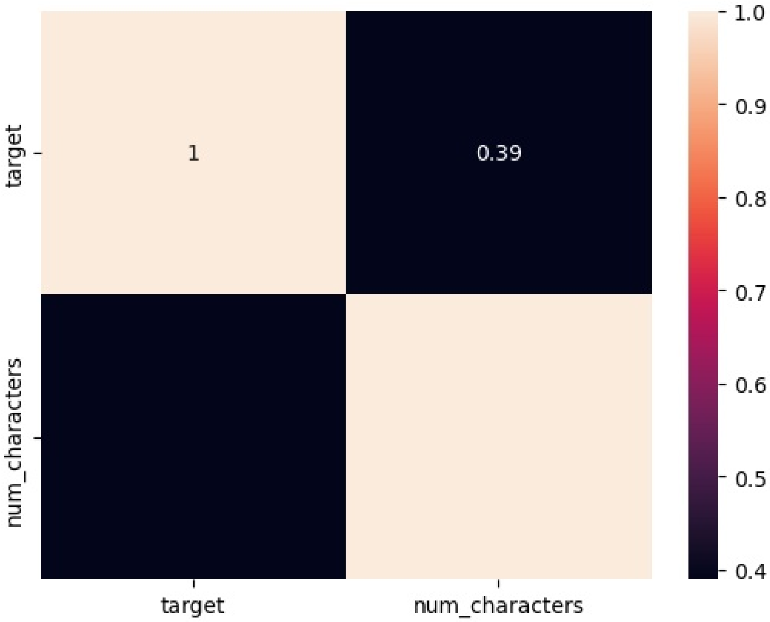


In [111…

```
sns.heatmap(df.drop('text',axis=1).corr(),annot=True)
```

Out[111…

```
<Axes: >
```



Data Preprocessing

In [100…

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

In [106…

```
import nltk
nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\Sukesh\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping corpora\stopwords.zip.
```

Out[106…

```
True
```

In [104…

```
def transform_text(text):
    text = text.lower()
    text = nltk.word_tokenize(text)
    # since text is converted into list so we will loop through it now onwards
    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)
```

In [107…

```
transform_text('Hi How Are You? @Nice that is Great &*. I loved your videos on ML, I would like coding with you
```

Out[107…

```
'hi nice great love video ml would like code sometim'
```

In [125…

```
transform_text("Just forced myself to eat a slice. I'm really not hungry tho. This sucks. Mark is getting worri
```

Out[125…

```
'forc eat slice realli hungri tho suck mark get worri know sick turn pizza lol'
```

In [80]:

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

Out[80]:

```
'love'
```

In [89]:

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

Out[89]:

```
'Sure thing big man. i have hockey elections at 6, shouldn€˜t go on longer than an hour though'
```

In [126…

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

In [127…

```
df.head()
```

Out[127…

	target	text	num_characters	transformed_text
0	0	Hope you are having a good week. Just checking in	49	hope good week check
1	0	K..give back my thanks.	23	k give back thank
2	0	Am also doing in cbe only. But have to pay.	43	also cbe pay
3	1	complimentary 4 STAR Ibiza Holiday or £10,000 ...	149	complimentari 4 star ibiza holiday cash need u...
4	1	okmail: Dear Dave this is your final notice to...	161	okmail dear dave final notic collect 4 tenerif...

Creating WordCloud of ham and spam

In [129…

```
# let's see what are the top 30 common words in spam
spam_corpus = []
for msg in df[df['target'] == 1]['transformed_text'].tolist():
    for word in msg.split():
        spam_corpus.append(word)
```

In [130…

```
len(spam_corpus)
```

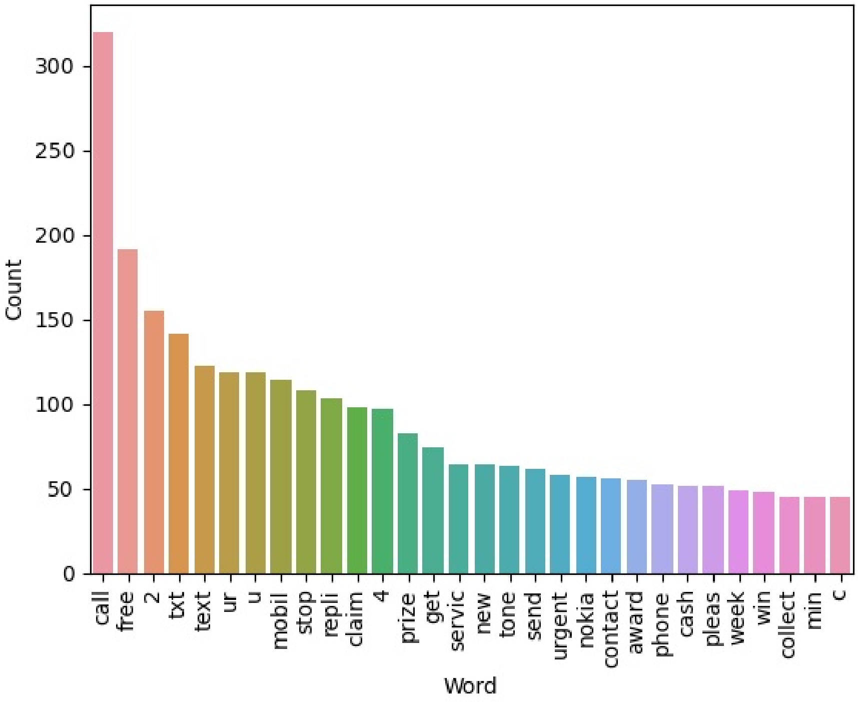
Out[130…

```
9978
```

In [131…

```
from collections import Counter
spam_counts = pd.DataFrame(Counter(spam_corpus).most_common(30), columns=['Word', 'Count'])

# Plot using seaborn barplot
sns.barplot(x='Word', y='Count', data=spam_counts)
plt.xticks(rotation='vertical')
plt.show()
```



In [132…

```
# let's see what are the top 30 common words in ham
ham_corpus = []
for msg in df[df['target'] == 0]['transformed_text'].tolist():
    for word in msg.split():
        ham_corpus.append(word)
```

In [133…

```
len(ham_corpus)
```

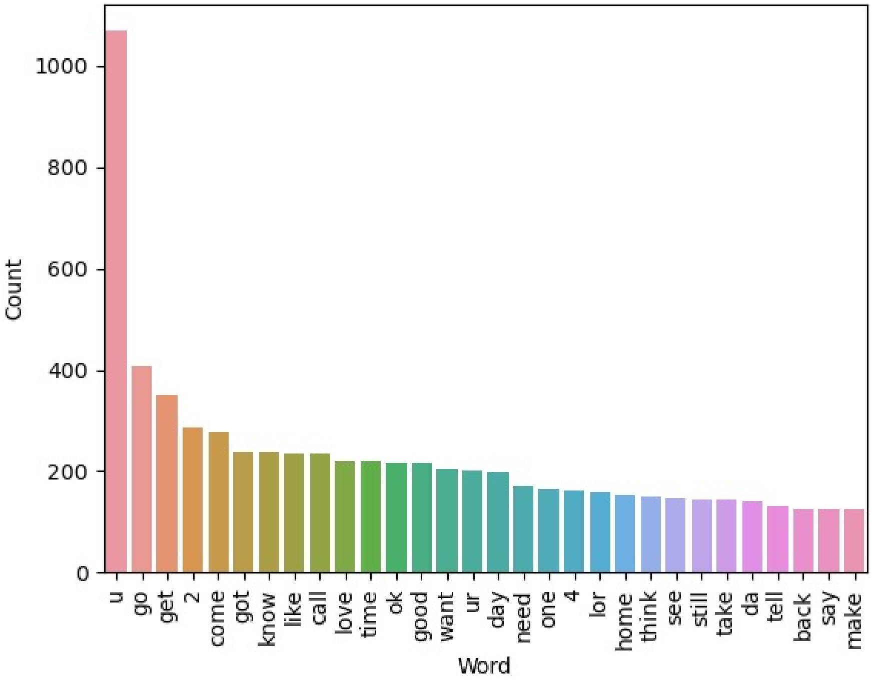
Out[133…

```
35091
```

In [134…

```
ham_counts = pd.DataFrame(Counter(ham_corpus).most_common(30), columns=['Word', 'Count'])

# Plot using seaborn barplot
sns.barplot(x='Word', y='Count', data=ham_counts)
plt.xticks(rotation='vertical')
plt.show()
```



Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js