Program:

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from wordcloud import WordCloud

import nltk

from nltk.corpus import stopwords

from collections import Counter

import string

# Download stopwords if not already downloaded

import nltk

nltk.download('stopwords')

# Load SMS Spam Collection dataset directly from UCI repository

url = "https://raw.githubusercontent.com/justmarkham/pycon-2016-tutorial/master/data/sms.tsv"

df = pd.read\_csv(url, sep='\t', header=None, names=['label', 'text'])

print("First 5 rows:")

print(df.head())

print("\nData Info:")

print(df.info())

print("\nLabel counts:")

print(df['label'].value\_counts())

# Add word count column

df['word\_count'] = df['text'].apply(lambda x: len(str(x).split()))

# Plot 1: Spam vs Ham counts

plt.figure(figsize=(6,4))

sns.countplot(data=df, x='label', palette=['green', 'red'])

plt.title("Count of Spam vs Ham Messages")

plt.show()

# Plot 2: Word count distribution by label

plt.figure(figsize=(8,5))

sns.histplot(data=df, x='word\_count', hue='label', bins=50, kde=True, palette=['green', 'red'])

plt.title("Word Count Distribution by Message Type")

plt.show()

# Preprocess text function

stop\_words = set(stopwords.words('english'))

def clean\_text(text):

text = text.lower()

text = ''.join([c for c in text if c not in string.punctuation])

tokens = text.split()

tokens = [word for word in tokens if word not in stop\_words]

return tokens

# Separate spam and ham words

spam\_words = df[df['label']=='spam']['text'].apply(clean\_text).sum()

ham\_words = df[df['label']=='ham']['text'].apply(clean\_text).sum()

# Most common words

spam\_common = Counter(spam\_words).most\_common(15)

ham\_common = Counter(ham\_words).most\_common(15)

# Plot 3: Bar charts for most common words

fig, axes = plt.subplots(1, 2, figsize=(15,6))

spam\_words\_, spam\_counts = zip(\*spam\_common)

axes[0].bar(spam\_words\_, spam\_counts, color='red')

axes[0].set\_title('Top 15 Words in Spam Messages')

axes[0].tick\_params(axis='x', rotation=45)

ham\_words\_, ham\_counts = zip(\*ham\_common)

axes[1].bar(ham\_words\_, ham\_counts, color='green')

axes[1].set\_title('Top 15 Words in Ham Messages')

axes[1].tick\_params(axis='x', rotation=45)

plt.tight\_layout()

plt.show()

# Plot 4 & 5: Word Clouds for spam and ham

spam\_text = ' '.join(spam\_words)

ham\_text = ' '.join(ham\_words)

plt.figure(figsize=(12,6))

plt.subplot(1,2,1)

wc\_spam = WordCloud(background\_color='white', max\_words=100, colormap='Reds').generate(spam\_text)

plt.imshow(wc\_spam, interpolation='bilinear')

plt.axis('off')

plt.title('Word Cloud for Spam Messages')

plt.subplot(1,2,2)

wc\_ham = WordCloud(background\_color='white', max\_words=100, colormap='Greens').generate(ham\_text)

plt.imshow(wc\_ham, interpolation='bilinear')

plt.axis('off')

plt.title('Word Cloud for Ham Messages')

plt.show()

Output:







