Theory Activity No. 1

Name: Vaibhavi M. Aochar

Division: CS2

Roll no.: 35

PRN: 202401040342

Dataset: SMS Spam Collection

1. Find the total number of messages in the dataset.

```
import pandas as pd
import numpy as np

# Load dataset (adjust path as needed)
df = pd.read_csv('spam.csv', encoding='latin-1')
df = df[['v1', 'v2']]  # only necessary columns
df.columns = ['label', 'message']

print("the total no of messages are",len(df))
```

```
the total no of messages are 5572
```

2. Find the number of spam messages and ham (non-spam) messages separately.

```
[12] import pandas as pd
import numpy as np

# Load dataset (adjust path as needed)
df = pd.read_csv('spam.csv', encoding='latin-1')
df = df[['v1', 'v2']]  # only necessary columns
df.columns = ['label', 'message']

# 2. Number of spam and ham messages
print("spam count:",(df['label'] == 'spam').sum())
print("ham count:",(df['label'] == 'ham').sum())
```

```
spam count: 747
ham count: 4825
```

3. Calculate the percentage of spam messages.

```
[13] import pandas as pd
     import numpy as np

     # Load dataset (adjust path as needed)
     df = pd.read_csv('spam.csv', encoding='latin-1')
     df = df[['v1', 'v2']]  # only necessary columns
     df.columns = ['label', 'message']

     print("Percentage of spam messages:",(spam_count / total_messages) * 100)
```

```
Percentage of spam messages: 13.406317300789663
```

4. Find the length (character count) of each message and add it as a new column.

```
import pandas as pd
import numpy as np

# Load dataset (adjust path as needed)
df = pd.read_csv('spam.csv', encoding='latin-1')
df = df[['v1', 'v2']]  # only necessary columns
df.columns = ['label', 'message']

print("character count of the file is:", df['message'].apply(len))
```

```
character count of the file is: 0       111
1        29
2       155
3        49
4        61
       ...
5567    161
5568     37
5569     57
5570    125
5571     26
Name: message, Length: 5572, dtype: int64
```

5 Find the average message length for spam and ham messages separately.

```
[20] import pandas as pd
     import numpy as np

     # Load dataset (adjust path as needed)
     df = pd.read_csv('spam.csv', encoding='latin-1')
     df = df[['v1', 'v2']]  # only necessary columns
     df.columns = ['label', 'message']
     df['message_length'] = df['message'].apply(len)

     avg_spam_length = df[df['label'] == 'spam']['message_length'].mean()
     avg_ham_length = df[df['label'] == 'ham']['message_length'].mean()

     print("Average message lenght for spam is:",avg_spam_length)
     print("Average message lenght for ham is:",avg_ham_length )
```

```
Average message lenght for spam is: 138.8661311914324
Average message lenght for ham is: 71.02362694300518
```

6 Find the longest message and its label (spam/ham).

```
[25] import pandas as pd
     import numpy as np

     # Load dataset (adjust path as needed)
     df = pd.read_csv('spam.csv', encoding='latin-1')
     df = df[['v1', 'v2']]  # only necessary columns
     df.columns = ['label', 'message']

     df['message_length'] = df['message'].apply(len)
     print(df.loc[df['message_length'].idxmax()])
```

```
label                                                       ham
message           For me the love should start with attraction.i...
message_length                                              910
Name: 1084, dtype: object
```

7 Find the shortest message and its label.

```
[24] import pandas as pd
     import numpy as np

     # Load dataset (adjust path as needed)
     df = pd.read_csv('spam.csv', encoding='latin-1')
     df = df[['v1', 'v2']]  # only necessary columns
     df.columns = ['label', 'message']

     df['message_length'] = df['message'].apply(len)
     print(df.loc[df['message_length'].idxmin()])
```

```
label             ham
message            Ok
message_length      2
Name: 1924, dtype: object
```

8 Find how many messages contain the word "free".

```
import pandas as pd
import numpy as np

# Load dataset (adjust path as needed)
df = pd.read_csv('spam.csv', encoding='latin-1')
df = df[['v1', 'v2']]  # only necessary columns
df.columns = ['label', 'message']

print("No. of messages containing word'Free':",df['message'].str.contains('free', case=False).sum())
```

```
No. of messages containing word'Free': 265
```

9 Find how many spam messages contain the word "win".

```
[29] import pandas as pd
     import numpy as np

     # Load dataset (adjust path as needed)
     df = pd.read_csv('spam.csv', encoding='latin-1')
     df = df[['v1', 'v2']]  # only necessary columns
     df.columns = ['label', 'message']

     print("spam messages containing the word 'win':", df[(df['label'] == 'spam') & (df['message'].str.contains('win', case=False))].shape[0])
```

```
spam messages containing the word 'win': 100
```

10 Create a new column that shows whether a message contains a number (0/1).

```
[31] import pandas as pd
     import numpy as np

     # Load dataset (adjust path as needed)
     df = pd.read_csv('spam.csv', encoding='latin-1')
     df = df[['v1', 'v2']]  # only necessary columns
     df.columns = ['label', 'message']

     print(df['message'].str.contains(r'\d').astype(int))
```

```
0       0
1       0
2       1
3       0
4       0
       ..
5567    1
5568    0
5569    0
5570    0
5571    0
Name: message, Length: 5572, dtype: int64
```

11 Find the average length of messages that contain numbers.

```
[34] import pandas as pd
     import numpy as np

     # Load dataset (adjust path as needed)
     df = pd.read_csv('spam.csv', encoding='latin-1')
     df = df[['v1', 'v2']]  # only necessary columns
     df.columns = ['label', 'message']
     df['message_length'] = df['message'].apply(len)
     df['contains_number'] = df['message'].str.contains(r'\d').astype(int)
     avg_length_with_numbers = df[df['contains_number'] == 1]['message_length'].mean()


     print("Finding average length of message that contains the number:",avg_length_with_numbers)
```

```
Finding average length of message that contains the number: 119.48901098901099
```

12 Find the most common word across all messages.

```
[35] import pandas as pd
     import numpy as np

     # Load dataset (adjust path as needed)
     df = pd.read_csv('spam.csv', encoding='latin-1')
     df = df[['v1', 'v2']]  # only necessary columns
     df.columns = ['label', 'message']

     from collections import Counter
     all_words = ' '.join(df['message']).lower().split()
     most_common_word = Counter(all_words).most_common(1)

     print("Most common word across all messages:",most_common_word)
```

```
Most common word across all messages: [('to', 2226)]
```

13 Find the most common word specifically in spam messages.

```
[36] import pandas as pd
     import numpy as np

     # Load dataset (adjust path as needed)
     df = pd.read_csv('spam.csv', encoding='latin-1')
     df = df[['v1', 'v2']]  # only necessary columns
     df.columns = ['label', 'message']

     spam_words = ' '.join(df[df['label'] == 'spam']['message']).lower().split()
     most_common_spam_word = Counter(spam_words).most_common(1)

     print("most common word specifically in spam messages:",most_common_spam_word)
```

```
most common word specifically in spam messages: [('to', 682)]
```

14 Find the number of messages that are entirely in lowercase.

```
[37] import pandas as pd
     import numpy as np

     # Load dataset (adjust path as needed)
     df = pd.read_csv('spam.csv', encoding='latin-1')
     df = df[['v1', 'v2']]  # only necessary columns
     df.columns = ['label', 'message']

     print(" number of messages that are entirely in lowercase.:",df['message'].apply(lambda x: x.islower()).sum())
```

```
number of messages that are entirely in lowercase.: 82
```

15 Find the total number of unique words across all messages.

```
[38] import pandas as pd
     import numpy as np

     # Load dataset (adjust path as needed)
     df = pd.read_csv('spam.csv', encoding='latin-1')
     df = df[['v1', 'v2']]  # only necessary columns
     df.columns = ['label', 'message']

     unique_words = set(all_words)
     total_unique_words = len(unique_words)

     print(" total number of unique words across all messages:",total_unique_words)
```

```
total number of unique words across all messages: 13496
```

16 Find how many messages have more than 100 characters.

```
[39] import pandas as pd
     import numpy as np

     # Load dataset (adjust path as needed)
     df = pd.read_csv('spam.csv', encoding='latin-1')
     df = df[['v1', 'v2']]  # only necessary columns
     df.columns = ['label', 'message']

     df['message_length'] = df['message'].apply(len)

     print("messages having more than 100 characters:",(df['message_length'] > 100).sum())
```

```
messages having more than 100 characters: 1744
```

17 Calculate the correlation between message length and whether the message is spam (1) or ham (0).

```
[41] import pandas as pd
     import numpy as np

     # Load dataset (adjust path as needed)
     df = pd.read_csv('spam.csv', encoding='latin-1')
     df = df[['v1', 'v2']]  # only necessary columns
     df.columns = ['label', 'message']
     df['message_length'] = df['message'].apply(len)
     df['spam_label'] = np.where(df['label'] == 'spam', 1, 0)
     correlation = df['message_length'].corr(df['spam_label'])

     print("correlation between message length and whether the message is spam (1) or ham (0):",correlation)
```

```
correlation between message length and whether the message is spam (1) or ham (0): 0.3872852892684781
```

18 Find the average number of words per message.

```
[42] import pandas as pd
     import numpy as np

     # Load dataset (adjust path as needed)
     df = pd.read_csv('spam.csv', encoding='latin-1')
     df = df[['v1', 'v2']]  # only necessary columns
     df.columns = ['label', 'message']

     df['word_count'] = df['message'].apply(lambda x: len(x.split()))
     avg_words_per_message = df['word_count'].mean()

     print("average number of words per message:",avg_words_per_message)
```

```
average number of words per message: 15.494436468054559
```

19 Find the top 5 most frequent words in ham messages.

```
[44] import pandas as pd
     import numpy as np

     # Load dataset (adjust path as needed)
     df = pd.read_csv('spam.csv', encoding='latin-1')
     df = df[['v1', 'v2']]  # only necessary columns
     df.columns = ['label', 'message']

     ham_words = ' '.join(df[df['label'] == 'ham']['message']).lower().split()
     top_5_ham_words = Counter(ham_words).most_common(5)

     print(top_5_ham_words)
```

```
[('i', 2172), ('you', 1665), ('to', 1544), ('the', 1113), ('a', 1046)]
```

20 Create a NumPy array of message lengths and find the 25th, 50th, and 75th percentiles.

```
import pandas as pd
import numpy as np

# Load dataset (adjust path as needed)
df = pd.read_csv('spam.csv', encoding='latin-1')
df = df[['v1', 'v2']]  # only necessary columns
df.columns = ['label', 'message']
df['message_length'] = df['message'].apply(len)
length_array = df['message_length'].values
percentiles = np.percentile(length_array, [25, 50, 75])

print(percentiles)
```

```
[ 36.  61. 121.]
```