Name: - Mansi Matele

Roll no: - ET2-28

PRN: - 202401070162

Dataset Link: https://www.kaggle.com/datasets/mohinurabdurahimova/maildataset

import pandas as pd

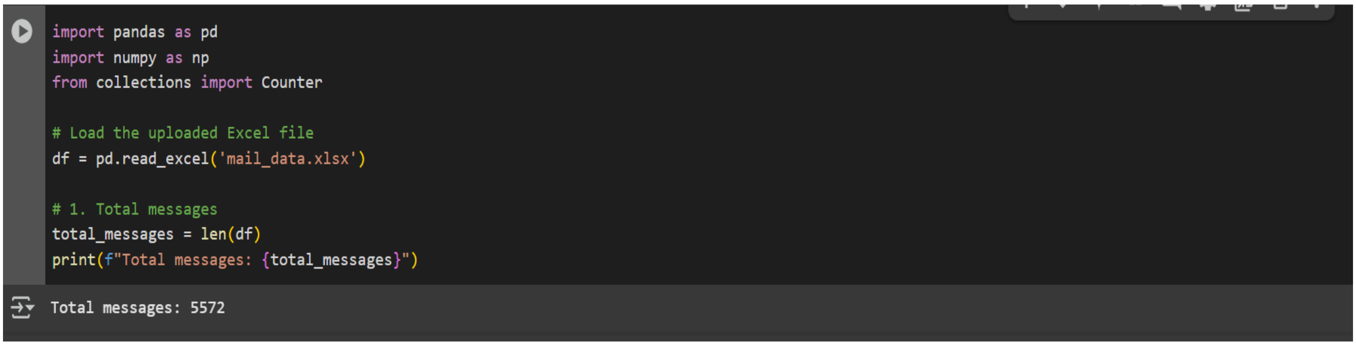
import numpy as np

from collections import Counter

# Load the uploaded Excel file

df = pd.read_excel('mail_data.xlsx')

1) Total messages



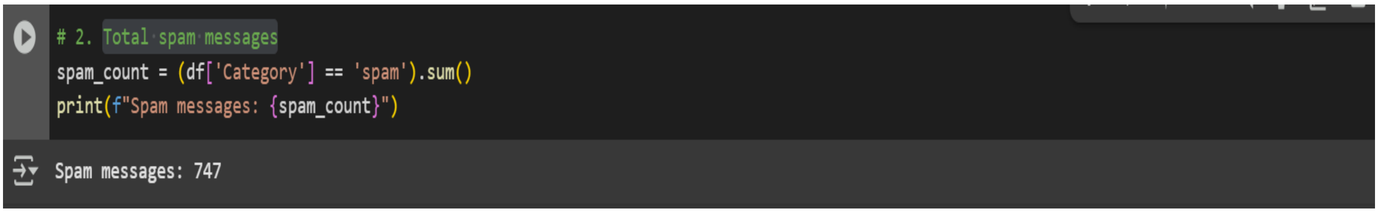
```
import pandas as pd
import numpy as np
from collections import Counter

# Load the uploaded Excel file
df = pd.read_excel('mail_data.xlsx')

# 1. Total messages
total_messages = len(df)
print(f"Total messages: {total_messages}")
```

```
Total messages: 5572
```

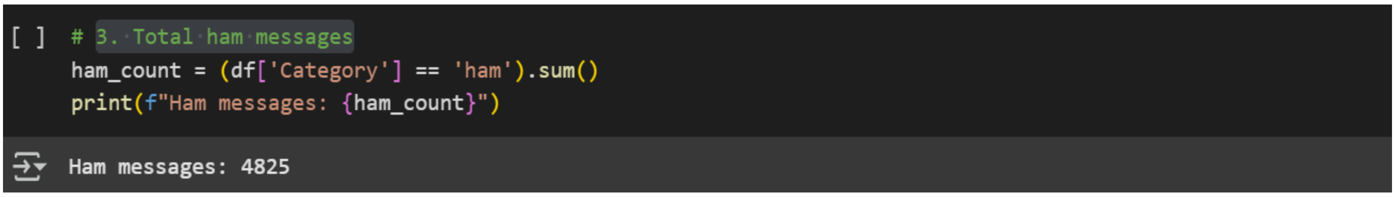
2) Total spam messages



```
# 2. Total spam messages
spam_count = (df['Category'] == 'spam').sum()
print(f"Spam messages: {spam_count}")
```

```
Spam messages: 747
```

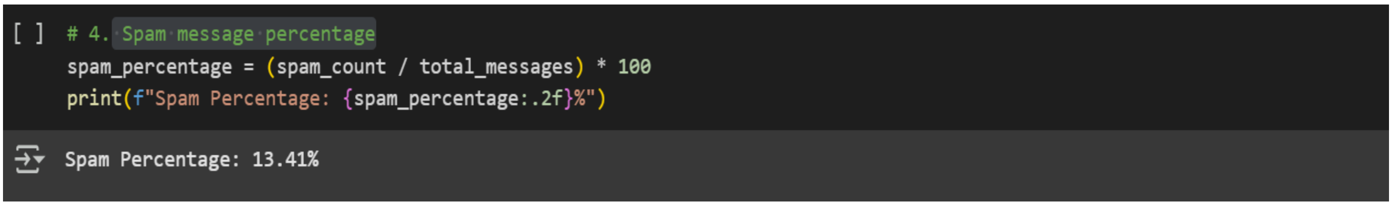
3) Total ham messages



```
# 3. Total ham messages
ham_count = (df['Category'] == 'ham').sum()
print(f"Ham messages: {ham_count}")
```

```
Ham messages: 4825
```

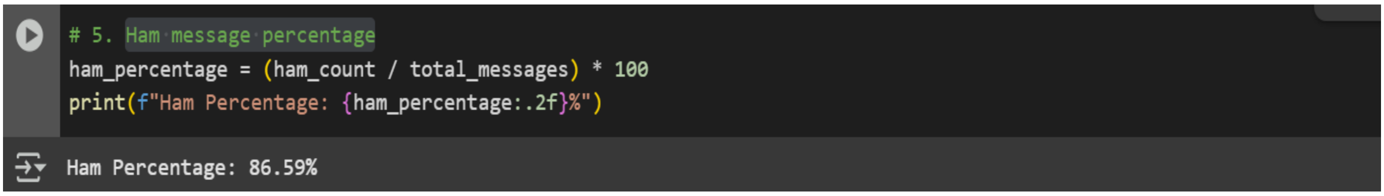
4) Spam message percentage



```
[ ] # 4. Spam message percentage
    spam_percentage = (spam_count / total_messages) * 100
    print(f"Spam Percentage: {spam_percentage:.2f}%")
```

```
Spam Percentage: 13.41%
```

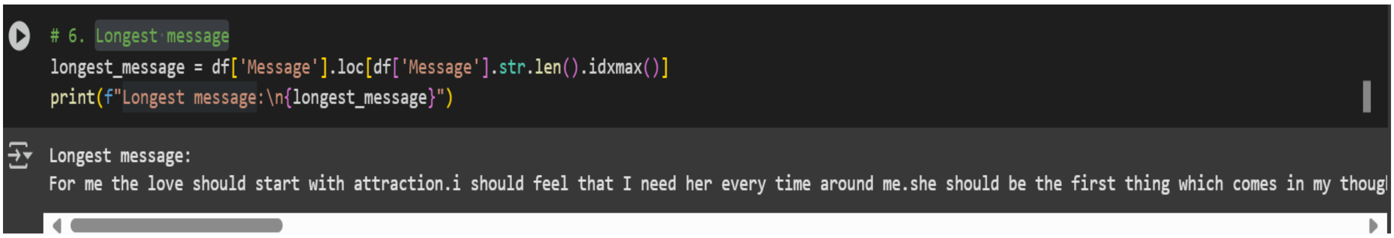
5) Ham message percentage



```
# 5. Ham message percentage
ham_percentage = (ham_count / total_messages) * 100
print(f"Ham Percentage: {ham_percentage:.2f}%")
```

```
Ham Percentage: 86.59%
```

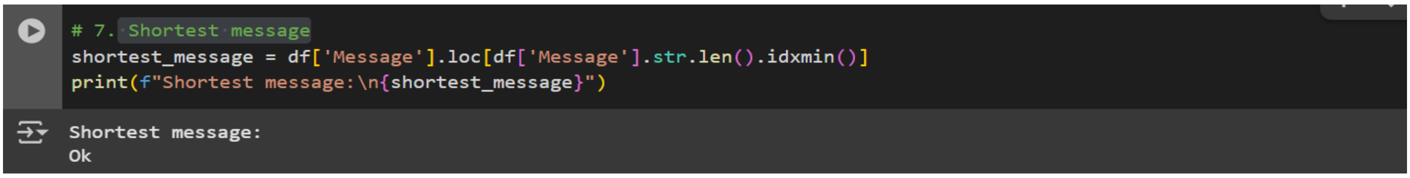
6) Longest message



```
# 6. Longest message
longest_message = df['Message'].loc[df['Message'].str.len().idxmax()]
print(f"Longest message:\n{longest_message}")
```

```
Longest message:
For me the love should start with attraction.i should feel that I need her every time around me.she should be the first thing which comes in my thoug
```

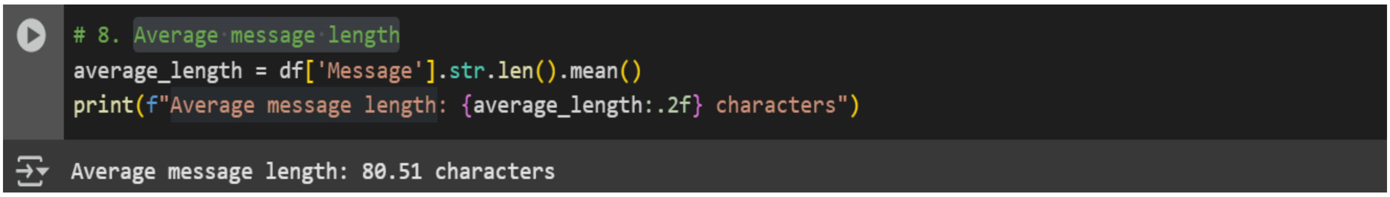
7) Shortest message



```
# 7. Shortest message
shortest_message = df['Message'].loc[df['Message'].str.len().idxmin()]
print(f"Shortest message:\n{shortest_message}")
```

```
Shortest message:
Ok
```

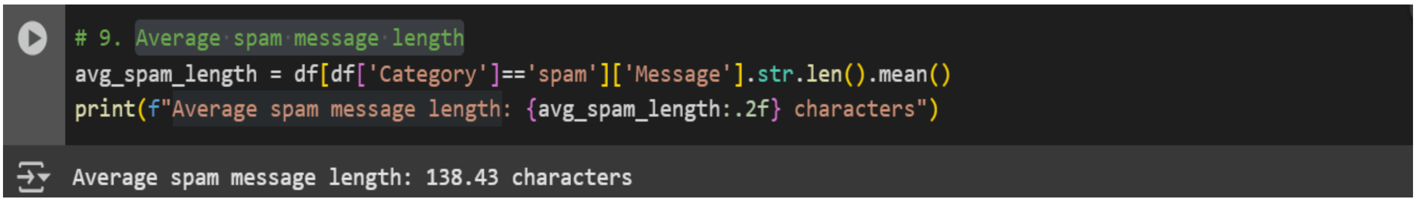
8) Average message length



```
# 8. Average message length
average_length = df['Message'].str.len().mean()
print(f"Average message length: {average_length:.2f} characters")
```

```
Average message length: 80.51 characters
```

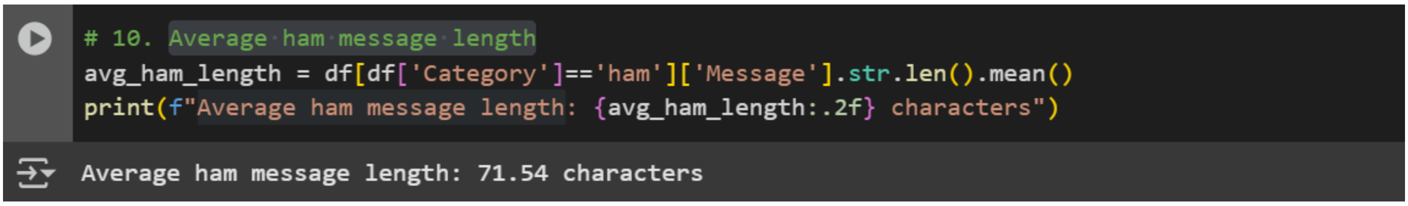
9) Average spam message length



```
# 9. Average spam message length
avg_spam_length = df[df['Category']=='spam']['Message'].str.len().mean()
print(f"Average spam message length: {avg_spam_length:.2f} characters")
```

```
Average spam message length: 138.43 characters
```

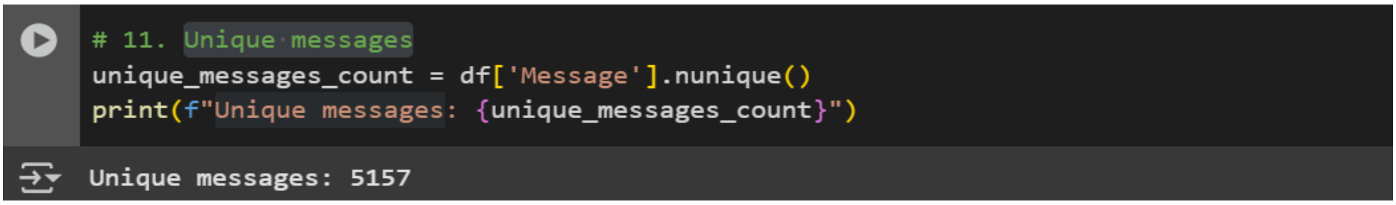
10) Average ham message length



```
# 10. Average ham message length
avg_ham_length = df[df['Category']=='ham']['Message'].str.len().mean()
print(f"Average ham message length: {avg_ham_length:.2f} characters")
```

```
Average ham message length: 71.54 characters
```

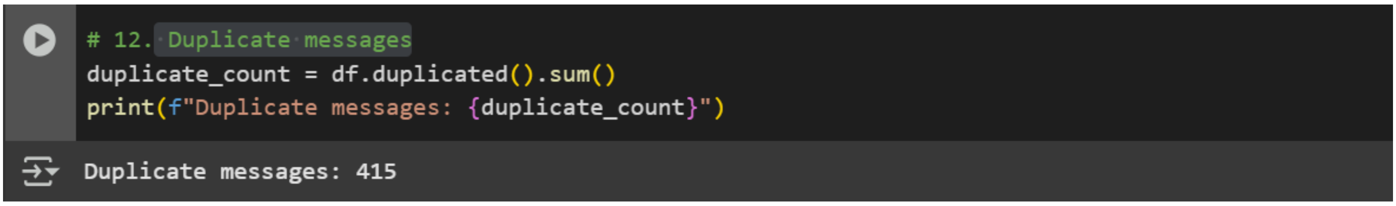
11) Unique messages



```
# 11. Unique messages
unique_messages_count = df['Message'].nunique()
print(f"Unique messages: {unique_messages_count}")
```

```
Unique messages: 5157
```

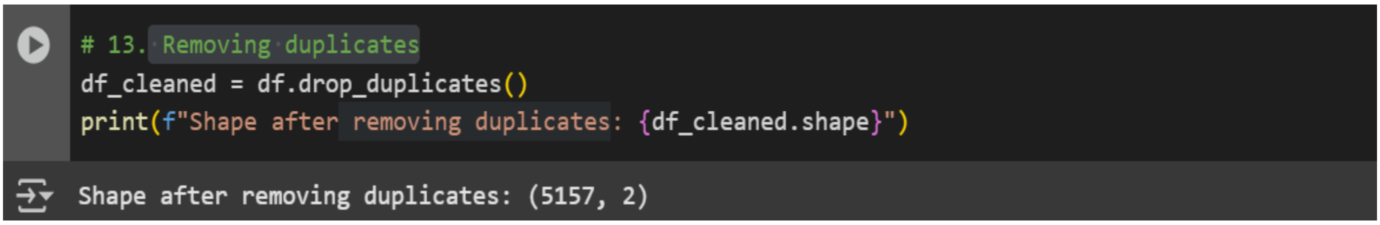
12) Duplicate messages



```
# 12. Duplicate messages
duplicate_count = df.duplicated().sum()
print(f"Duplicate messages: {duplicate_count}")
```

```
Duplicate messages: 415
```

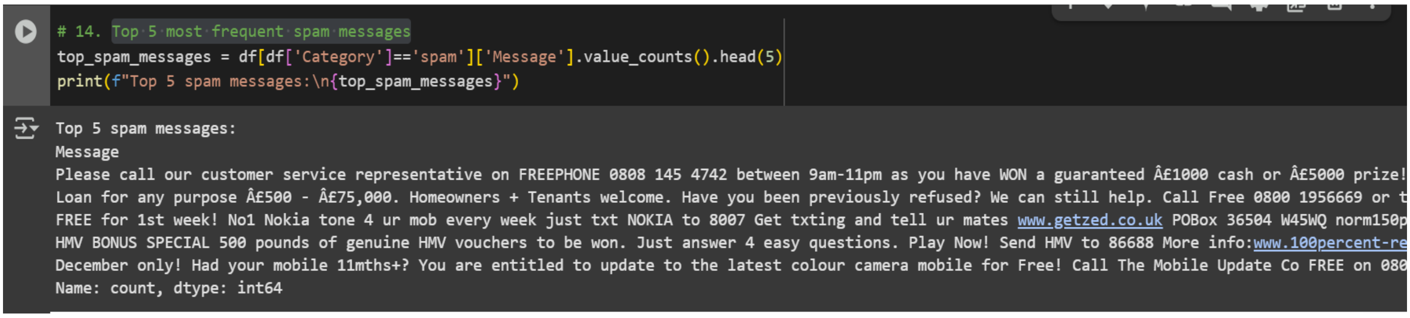
13) Removing duplicates



```
# 13. Removing duplicates
df_cleaned = df.drop_duplicates()
print(f"Shape after removing duplicates: {df_cleaned.shape}")
```

```
Shape after removing duplicates: (5157, 2)
```

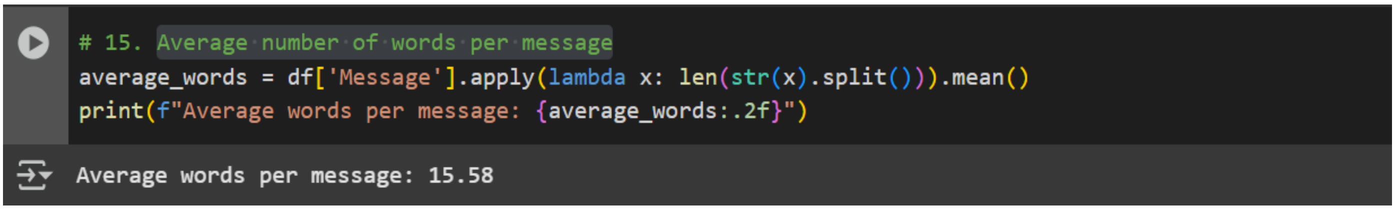
14) Top 5 most frequent spam messages



```
# 14. Top 5 most frequent spam messages
top_spam_messages = df[df['Category']=='spam']['Message'].value_counts().head(5)
print(f"Top 5 spam messages:\n{top_spam_messages}")
```

```
Top 5 spam messages:
Message
Please call our customer service representative on FREEPHONE 0808 145 4742 between 9am-11pm as you have WON a guaranteed Â£1000 cash or Â£5000 prize!
Loan for any purpose Â£500 - Â£75,000. Homeowners + Tenants welcome. Have you been previously refused? We can still help. Call Free 0800 1956669 or t
FREE for 1st week! No1 Nokia tone 4 ur mob every week just txt NOKIA to 8007 Get txting and tell ur mates www.getzed.co.uk POBox 36504 W45WQ norm150p
HMV BONUS SPECIAL 500 pounds of genuine HMV vouchers to be won. Just answer 4 easy questions. Play Now! Send HMV to 86688 More info:www.100percent-re
December only! Had your mobile 11mths+? You are entitled to update to the latest colour camera mobile for Free! Call The Mobile Update Co FREE on 080
Name: count, dtype: int64
```

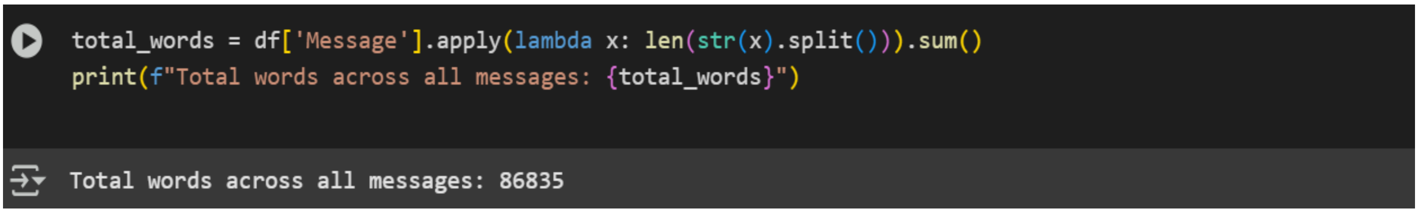
15) Average number of words per message



```
# 15. Average number of words per message
average_words = df['Message'].apply(lambda x: len(str(x).split())).mean()
print(f"Average words per message: {average_words:.2f}")
```

```
Average words per message: 15.58
```

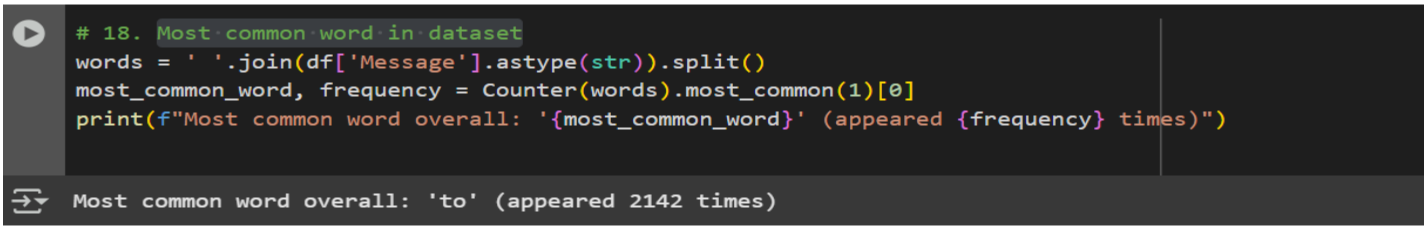
16) Total number of words across all messages



```
total_words = df['Message'].apply(lambda x: len(str(x).split())).sum()
print(f"Total words across all messages: {total_words}")
```

```
Total words across all messages: 86835
```

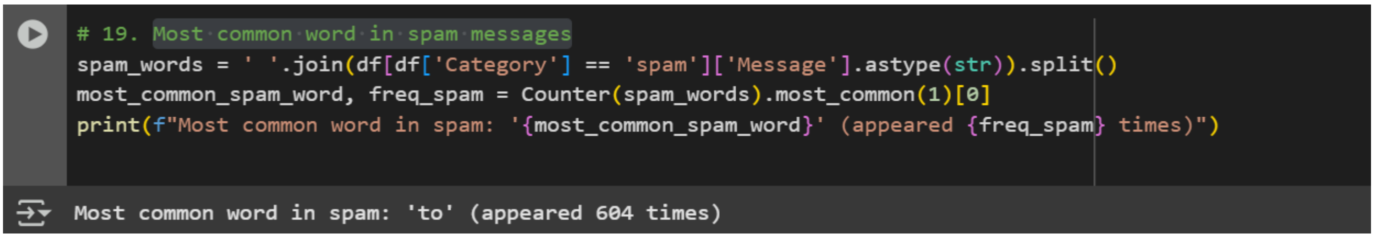
17) Most common word in dataset



```
# 18. Most common word in dataset
words = ' '.join(df['Message'].astype(str)).split()
most_common_word, frequency = Counter(words).most_common(1)[0]
print(f"Most common word overall: '{most_common_word}' (appeared {frequency} times)")
```

```
Most common word overall: 'to' (appeared 2142 times)
```

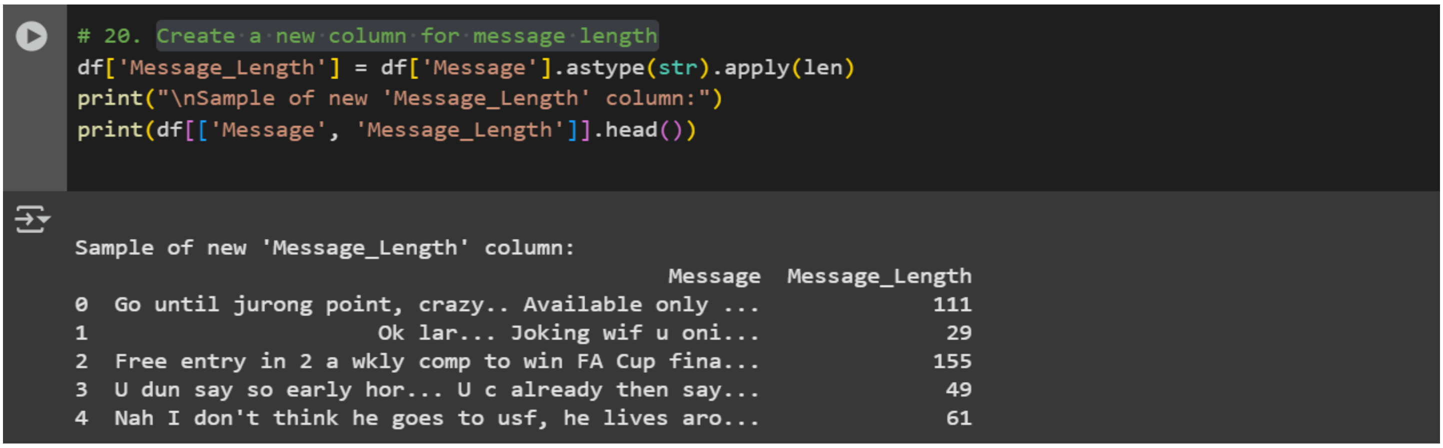
18) Most common word in spam messages



```
# 19. Most common word in spam messages
spam_words = ' '.join(df[df['Category'] == 'spam']['Message'].astype(str)).split()
most_common_spam_word, freq_spam = Counter(spam_words).most_common(1)[0]
print(f"Most common word in spam: '{most_common_spam_word}' (appeared {freq_spam} times)")
```

```
Most common word in spam: 'to' (appeared 604 times)
```

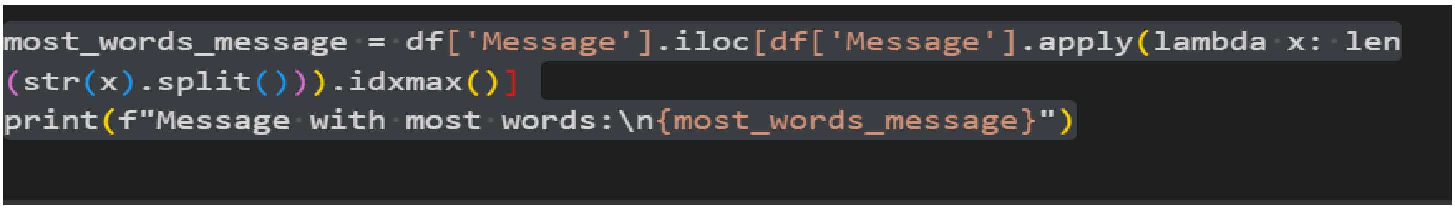
19) Create a new column for message length



```
# 20. Create a new column for message length
df['Message_Length'] = df['Message'].astype(str).apply(len)
print("\nSample of new 'Message_Length' column:")
print(df[['Message', 'Message_Length']].head())
```

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

20) Message with most words



```
most_words_message = df['Message'].iloc[df['Message'].apply(lambda x: len
(str(x).split())).idxmax()]
print(f"Message with most words:\n{most_words_message}")
```