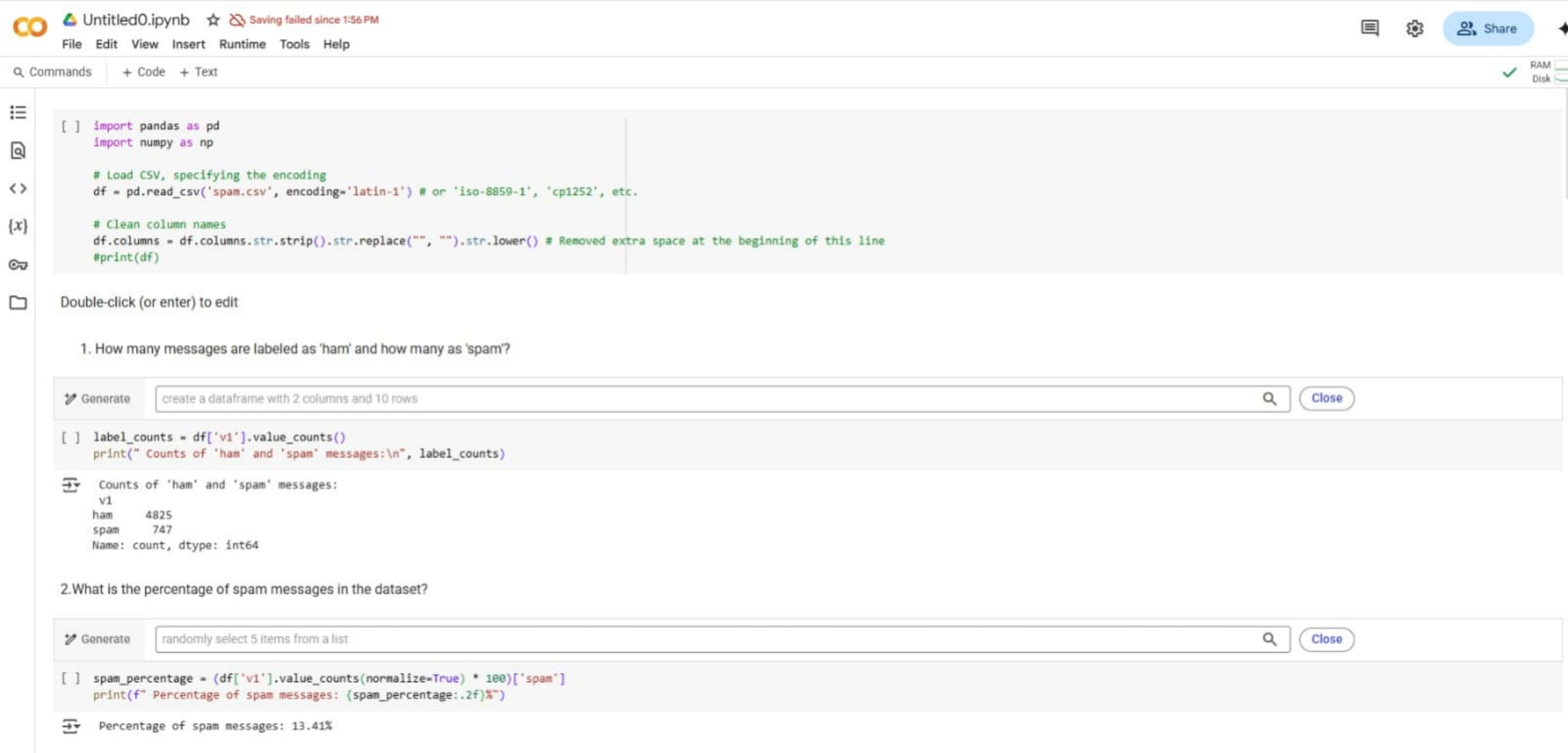


```
[ ] import pandas as pd
    import numpy as np

    # Load CSV, specifying the encoding
    df = pd.read_csv('spam.csv', encoding='latin-1') # or 'iso-8859-1', 'cp1252', etc.

    # Clean column names
    df.columns = df.columns.str.strip().str.replace("", "").str.lower() # Removed extra space at the beginning of this line
    #print(df)
```

Double-click (or enter) to edit

1. How many messages are labeled as 'ham' and how many as 'spam'?

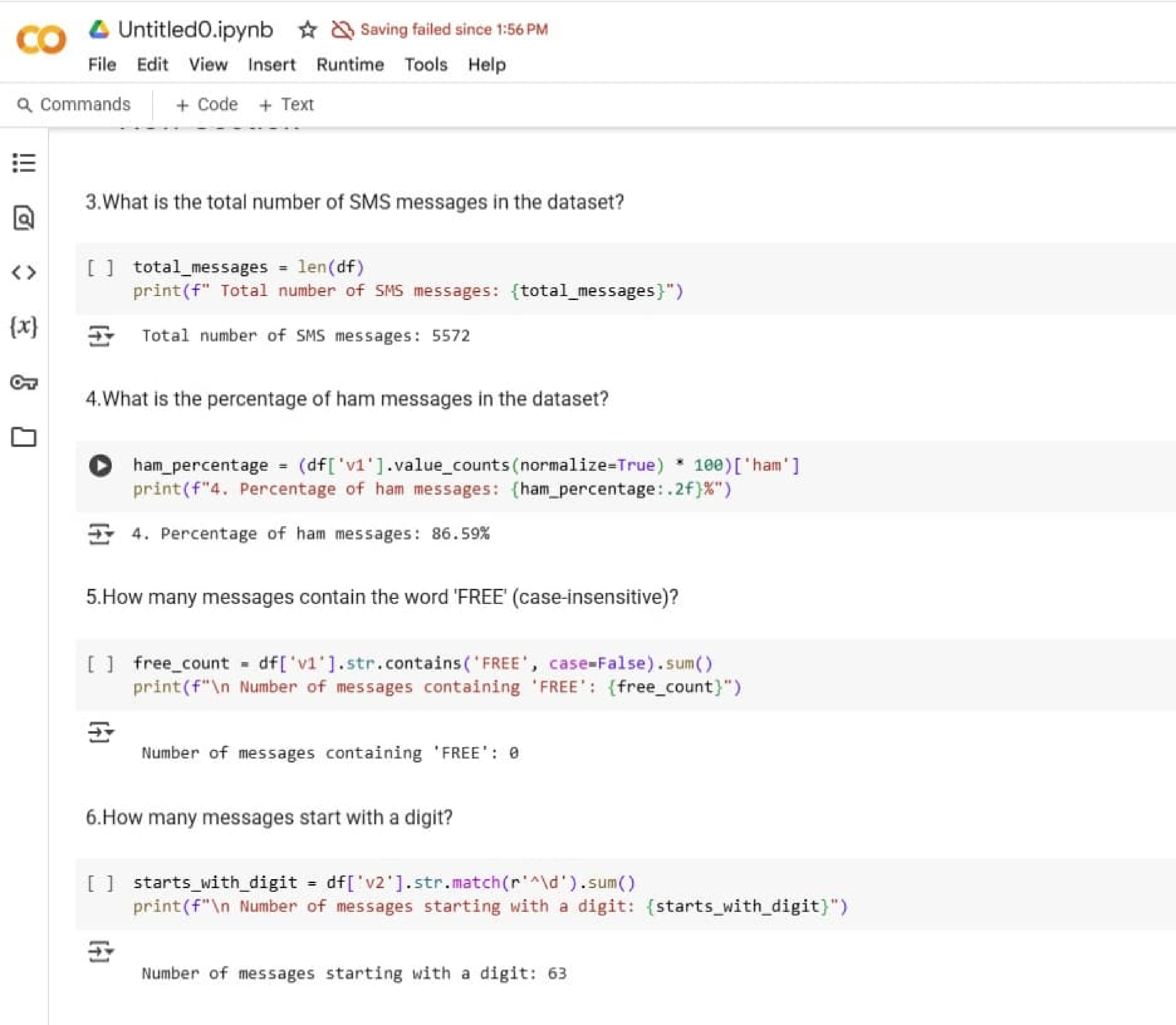
```
[ ] label_counts = df['v1'].value_counts()
    print(" Counts of 'ham' and 'spam' messages:\n", label_counts)
```

```
 Counts of 'ham' and 'spam' messages:
 v1
ham     4825
spam     747
Name: count, dtype: int64
```

2.What is the percentage of spam messages in the dataset?

```
[ ] spam_percentage = (df['v1'].value_counts(normalize=True) * 100)['spam']
    print(f" Percentage of spam messages: {spam_percentage:.2f}%")
```

```
 Percentage of spam messages: 13.41%
```



3.What is the total number of SMS messages in the dataset?

```
[ ] total_messages = len(df)
    print(f" Total number of SMS messages: {total_messages}")
```

```
 Total number of SMS messages: 5572
```

4.What is the percentage of ham messages in the dataset?

```
ham_percentage = (df['v1'].value_counts(normalize=True) * 100)['ham']
print(f"4. Percentage of ham messages: {ham_percentage:.2f}%")
```

```
4. Percentage of ham messages: 86.59%
```

5.How many messages contain the word 'FREE' (case-insensitive)?

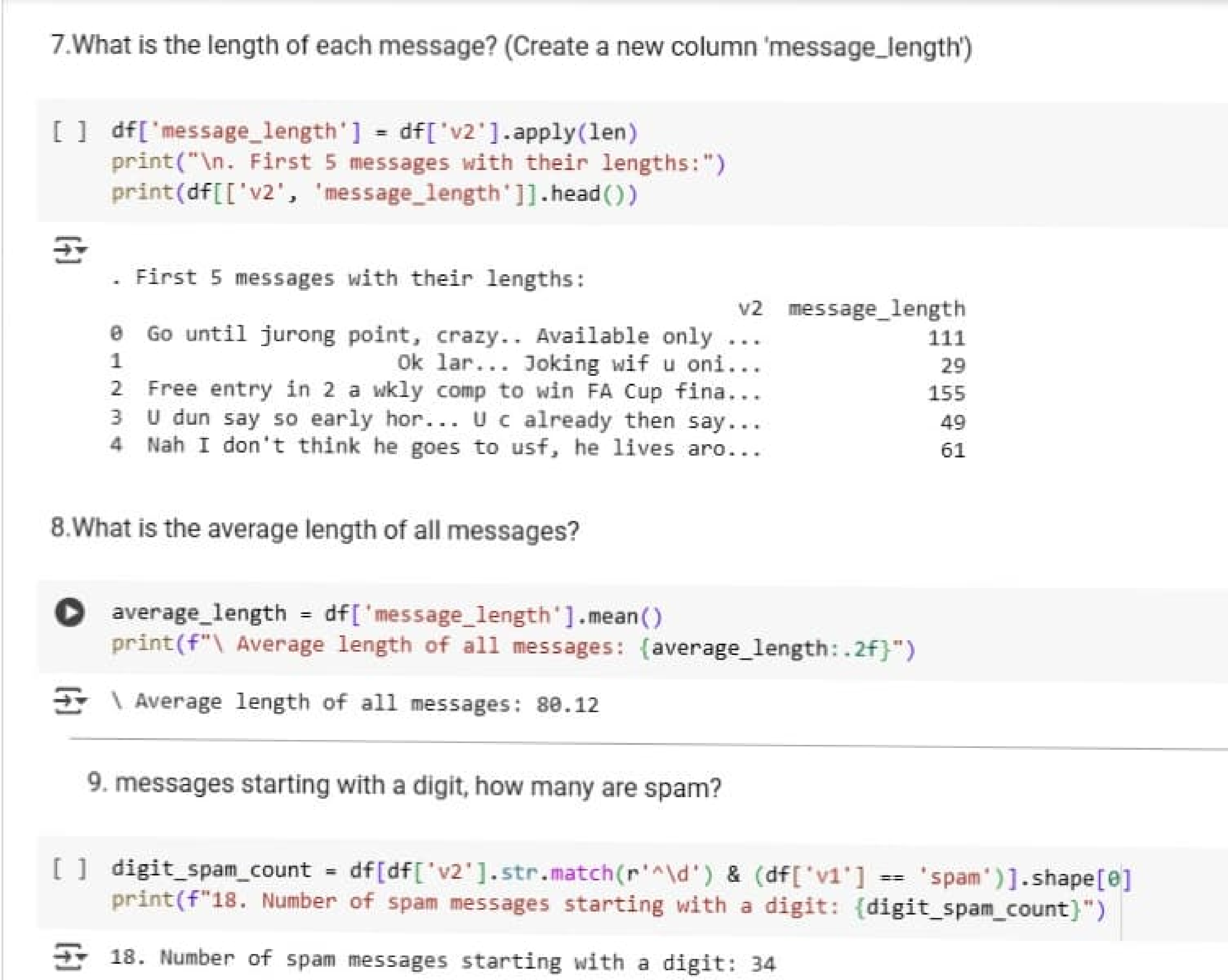
```
[ ] free_count = df['v1'].str.contains('FREE', case=False).sum()
    print(f"\n Number of messages containing 'FREE': {free_count}")
```

```

 Number of messages containing 'FREE': 0
```

6.How many messages start with a digit?

```
[ ] starts_with_digit = df['v2'].str.match(r'^\d').sum()
    print(f"\n Number of messages starting with a digit: {starts_with_digit}")
```

```

 Number of messages starting with a digit: 63
```



7.What is the length of each message? (Create a new column 'message_length')

```
df['message_length'] = df['v2'].apply(len)
print("\n. First 5 messages with their lengths:")
print(df[['v2', 'message_length']].head())
```

```
. First 5 messages with their lengths:
                                                  v2  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
```

8.What is the average length of all messages?

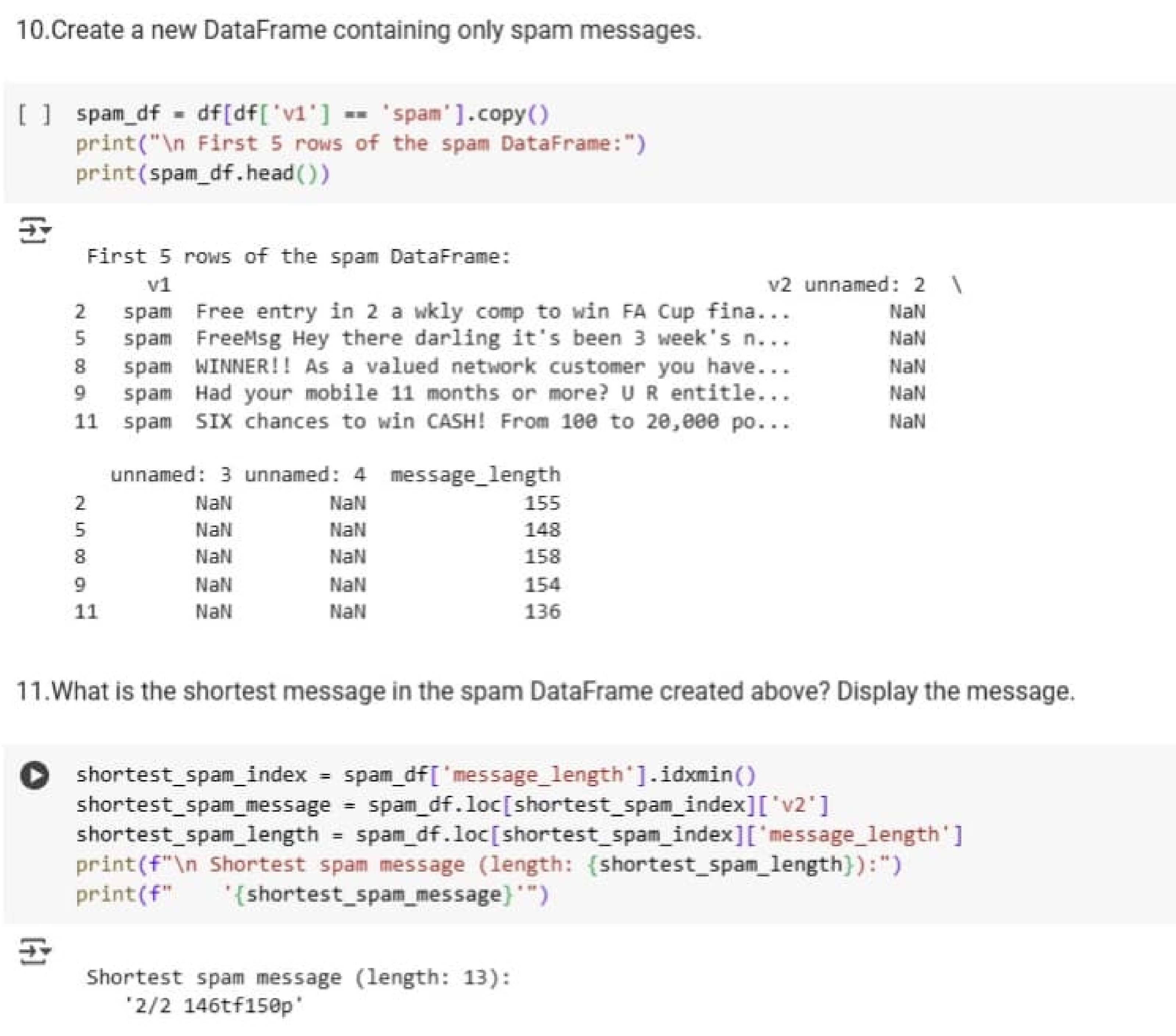
```
average_length = df['message_length'].mean()
print(f"\ Average length of all messages: {average_length:.2f}")
```

```
\ Average length of all messages: 80.12
```

9. messages starting with a digit, how many are spam?

```
digit_spam_count = df[df['v2'].str.match(r'^\d') & (df['v1'] == 'spam')].shape[0]
print(f"18. Number of spam messages starting with a digit: {digit_spam_count}")
```

```
18. Number of spam messages starting with a digit: 34
```



10.Create a new DataFrame containing only spam messages.

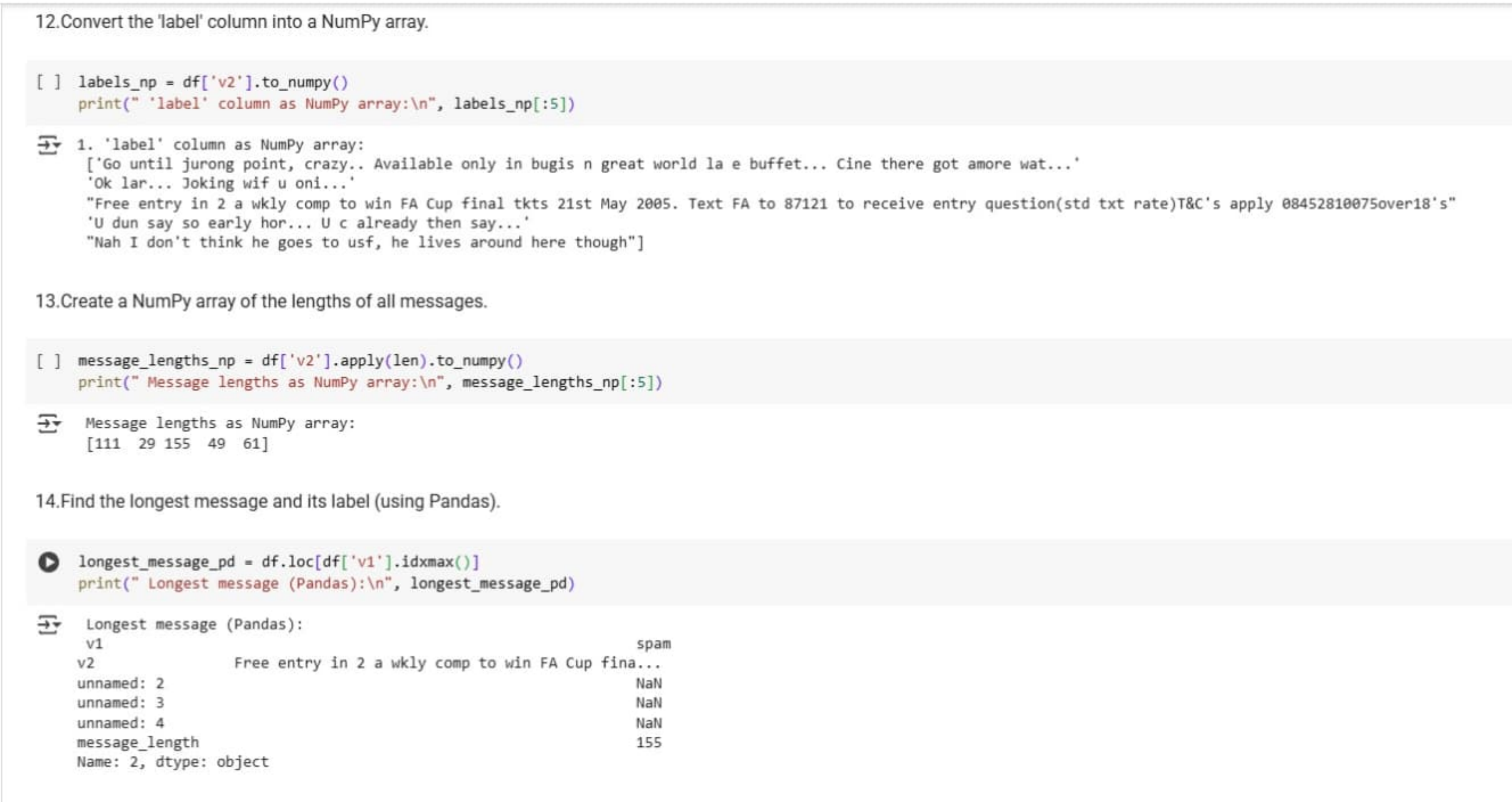
```
spam_df = df[df['v1'] == 'spam'].copy()
print("\n First 5 rows of the spam DataFrame:")
print(spam_df.head())
```

```
 First 5 rows of the spam DataFrame:
      v1                                                 v2 unnamed: 2  \
2   spam  Free entry in 2 a wkly comp to win FA Cup fina...        NaN   
5   spam  FreeMsg Hey there darling it's been 3 week's n...        NaN   
8   spam  WINNER!! As a valued network customer you have...        NaN   
9   spam  Had your mobile 11 months or more? U R entitle...        NaN   
11  spam  SIX chances to win CASH! From 100 to 20,000 po...        NaN   

   unnamed: 3 unnamed: 4  message_length  
2         NaN        NaN             155  
5         NaN        NaN             148  
8         NaN        NaN             158  
9         NaN        NaN             154  
11        NaN        NaN             136  
```

11.What is the shortest message in the spam DataFrame created above? Display the message.

```
shortest_spam_index = spam_df['message_length'].idxmin()
shortest_spam_message = spam_df.loc[shortest_spam_index]['v2']
shortest_spam_length = spam_df.loc[shortest_spam_index]['message_length']
print(f"\n Shortest spam message (length: {shortest_spam_length}):")
print(f"    '{shortest_spam_message}'")
```

```
 Shortest spam message (length: 13):
    '2/2 146tf150p'
```



12.Convert the 'label' column into a NumPy array.

```
[ ] labels_np = df['v2'].to_numpy()
    print(" 'label' column as NumPy array:\n", labels_np[:5])
```

```
1. 'label' column as NumPy array:
 ['Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...'
 'Ok lar... Joking wif u oni...'
 "Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's"
 'U dun say so early hor... U c already then say...'
 "Nah I don't think he goes to usf, he lives around here though"]
```

13.Create a NumPy array of the lengths of all messages.

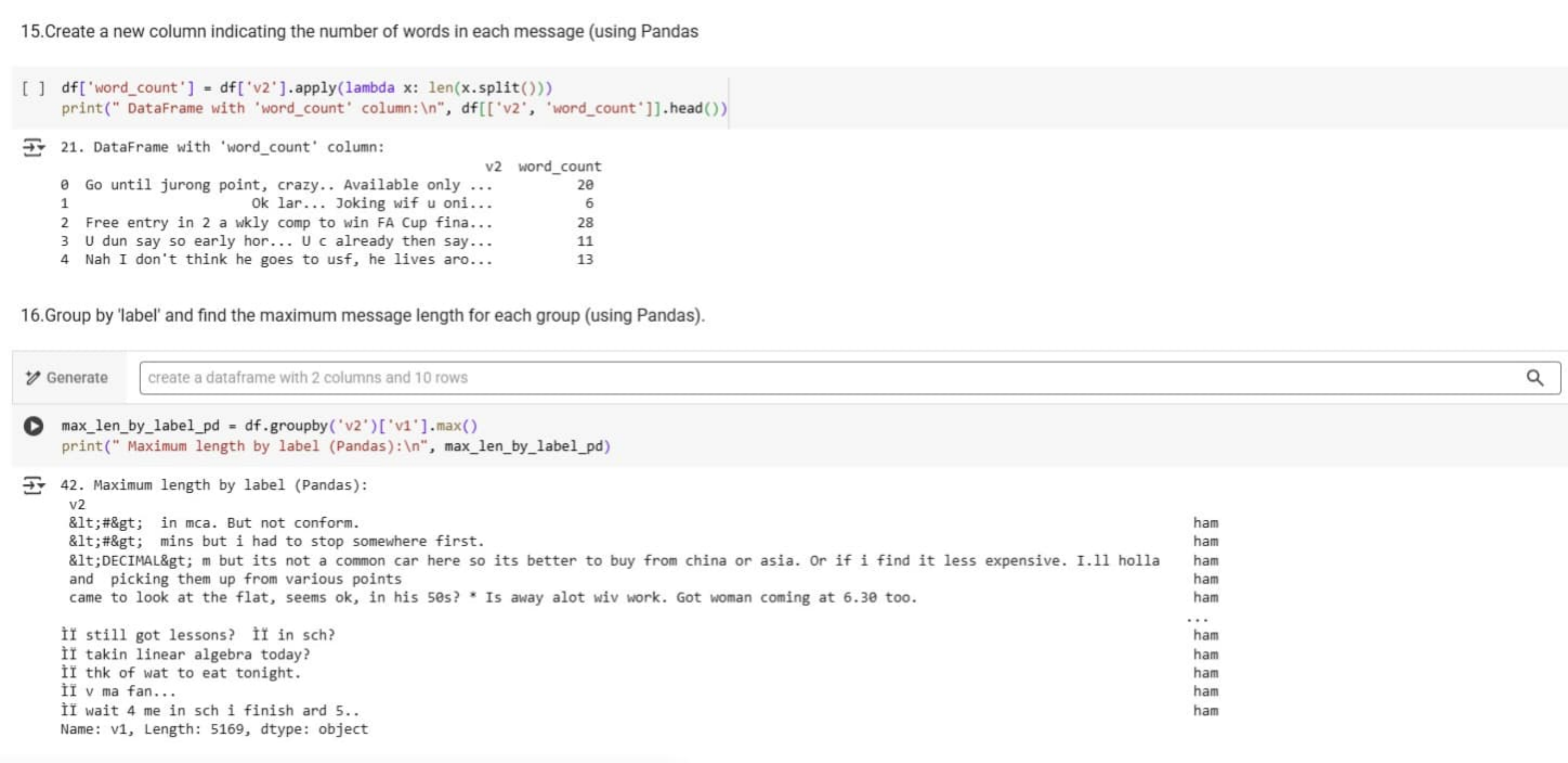
```
[ ] message_lengths_np = df['v2'].apply(len).to_numpy()
    print(" Message lengths as NumPy array:\n", message_lengths_np[:5])
```

```
 Message lengths as NumPy array:
 [111  29 155  49  61]
```

14.Find the longest message and its label (using Pandas).

```
longest_message_pd = df.loc[df['v1'].idxmax()]
print(" Longest message (Pandas):\n", longest_message_pd)
```

```
 Longest message (Pandas):
 v1                                                          spam
v2                Free entry in 2 a wkly comp to win FA Cup fina...
unnamed: 2                                                     NaN
unnamed: 3                                                     NaN
unnamed: 4                                                     NaN
message_length                                                 155
Name: 2, dtype: object
```



15.Create a new column indicating the number of words in each message (using Pandas

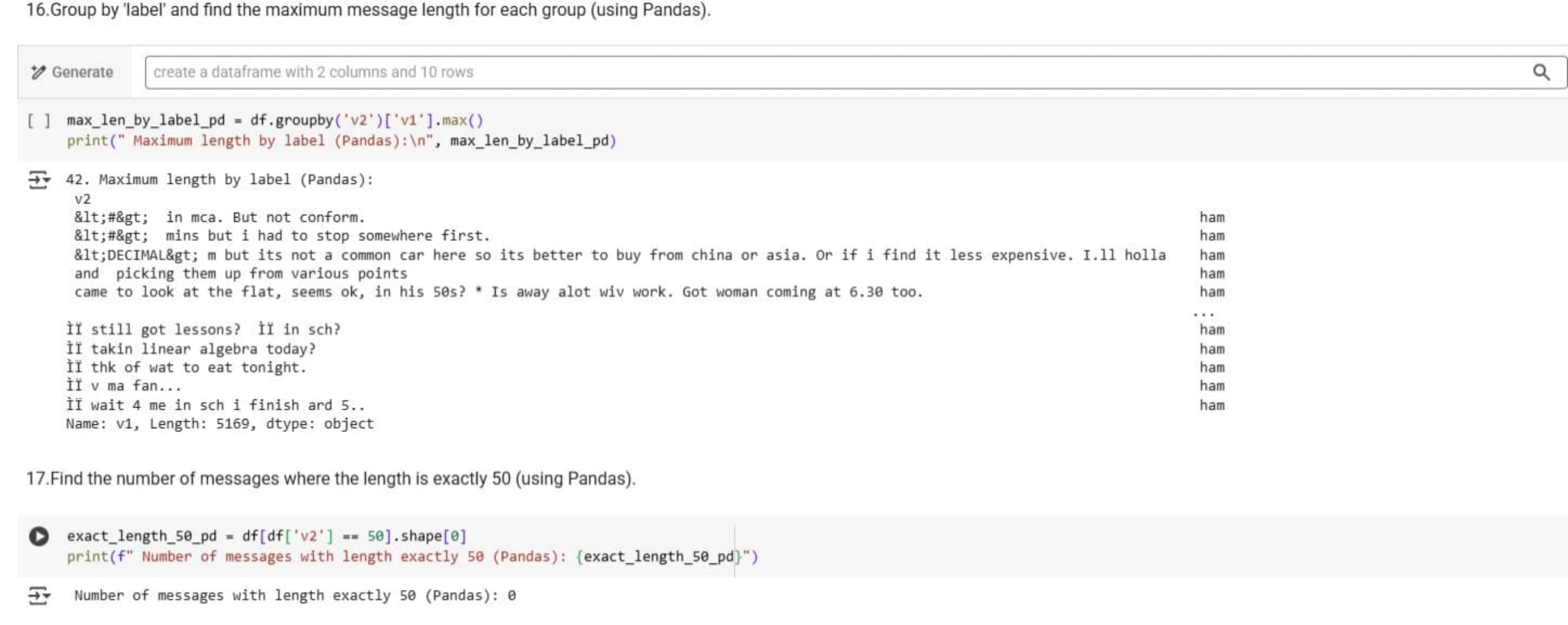
```
[ ] df['word_count'] = df['v2'].apply(lambda x: len(x.split()))
    print(" DataFrame with 'word_count' column:\n", df[['v2', 'word_count']].head())
```

```
21. DataFrame with 'word_count' column:
                                                   v2  word_count
0  Go until jurong point, crazy.. Available only ...          20
1                      Ok lar... Joking wif u oni...           6
2  Free entry in 2 a wkly comp to win FA Cup fina...          28
3  U dun say so early hor... U c already then say...          11
4  Nah I don't think he goes to usf, he lives aro...          13
```

16.Group by 'label' and find the maximum message length for each group (using Pandas).

```
max_len_by_label_pd = df.groupby('v2')['v1'].max()
print(" Maximum length by label (Pandas):\n", max_len_by_label_pd)
```

```
42. Maximum length by label (Pandas):
 v2
 &lt;#&gt;  in mca. But not conform.                                                                                  ham
 &lt;#&gt;  mins but i had to stop somewhere first.                                                                    ham
 &lt;DECIMAL&gt; m but its not a common car here so its better to buy from china or asia. Or if i find it less expensive. I.ll holla   ham
 and  picking them up from various points                                                                            ham
 came to look at the flat, seems ok, in his 50s? * Is away alot wiv work. Got woman coming at 6.30 too.              ham
                                                                                                                    ...
ÌÏ still got lessons?  ÌÏ in sch?                                                                                    ham
ÌÏ takin linear algebra today?                                                                                       ham
ÌÏ thk of wat to eat tonight.                                                                                        ham
ÌÏ v ma fan...                                                                                                       ham
ÌÏ wait 4 me in sch i finish ard 5..                                                                                 ham
Name: v1, Length: 5169, dtype: object
```



16.Group by 'label' and find the maximum message length for each group (using Pandas).

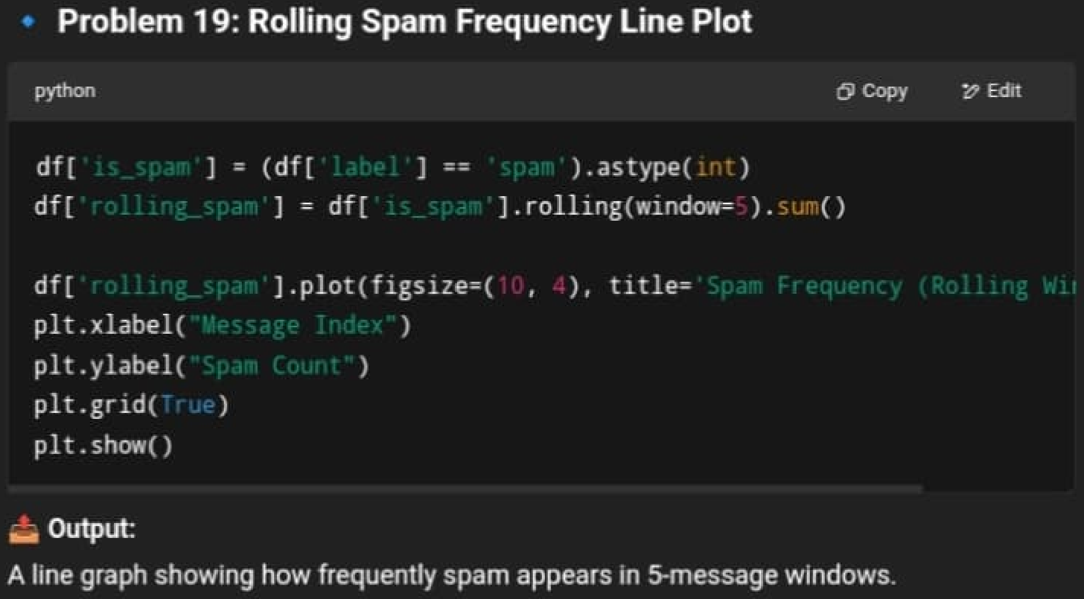
```
[ ] max_len_by_label_pd = df.groupby('v2')['v1'].max()
    print(" Maximum length by label (Pandas):\n", max_len_by_label_pd)
```

```
42. Maximum length by label (Pandas):
 v2
 &lt;#&gt;  in mca. But not conform.                                                                                                    ham
 &lt;#&gt;  mins but i had to stop somewhere first.                                                                                     ham
 &lt;DECIMAL&gt; m but its not a common car here so its better to buy from china or asia. Or if i find it less expensive. I.ll holla    ham
 and  picking them up from various points                                                                                               ham
 came to look at the flat, seems ok, in his 50s? * Is away alot wiv work. Got woman coming at 6.30 too.                                 ham
                                                                                                                                       ...
ÌÏ still got lessons?  ÌÏ in sch?                                                                                                       ham
ÌÏ takin linear algebra today?                                                                                                          ham
ÌÏ thk of wat to eat tonight.                                                                                                           ham
ÌÏ v ma fan...                                                                                                                          ham
ÌÏ wait 4 me in sch i finish ard 5..                                                                                                    ham
Name: v1, Length: 5169, dtype: object
```

17.Find the number of messages where the length is exactly 50 (using Pandas).

```
exact_length_50_pd = df[df['v2'] == 50].shape[0]
print(f" Number of messages with length exactly 50 (Pandas): {exact_length_50_pd}")
```

```
 Number of messages with length exactly 50 (Pandas): 0
```



🔹 Problem 19: Rolling Spam Frequency Line Plot

```python
df['is_spam'] = (df['label'] == 'spam').astype(int)
df['rolling_spam'] = df['is_spam'].rolling(window=5).sum()

df['rolling_spam'].plot(figsize=(10, 4), title='Spam Frequency (Rolling Wi
plt.xlabel("Message Index")
plt.ylabel("Spam Count")
plt.grid(True)
plt.show()
```

📤 **Output:**

A line graph showing how frequently spam appears in 5-message windows.

Problem 20: Unique Words in Spam vs Ham

```python
spam_words = set(' '.join(df[df['label']=='spam']['message']).lower().spli
ham_words = set(' '.join(df[df['label']=='ham']['message']).lower().split(

print("Unique spam words:", len(spam_words))
print("Unique ham words:", len(ham_words))
```

📤 **Output:**

```sql
Unique spam words: 312
Unique ham words: 421
```