MIT | Academy of Engineering

(An Autonomous Institute Affiliated to Savitribai Phule Pune University)

EDS Theory

Activity I

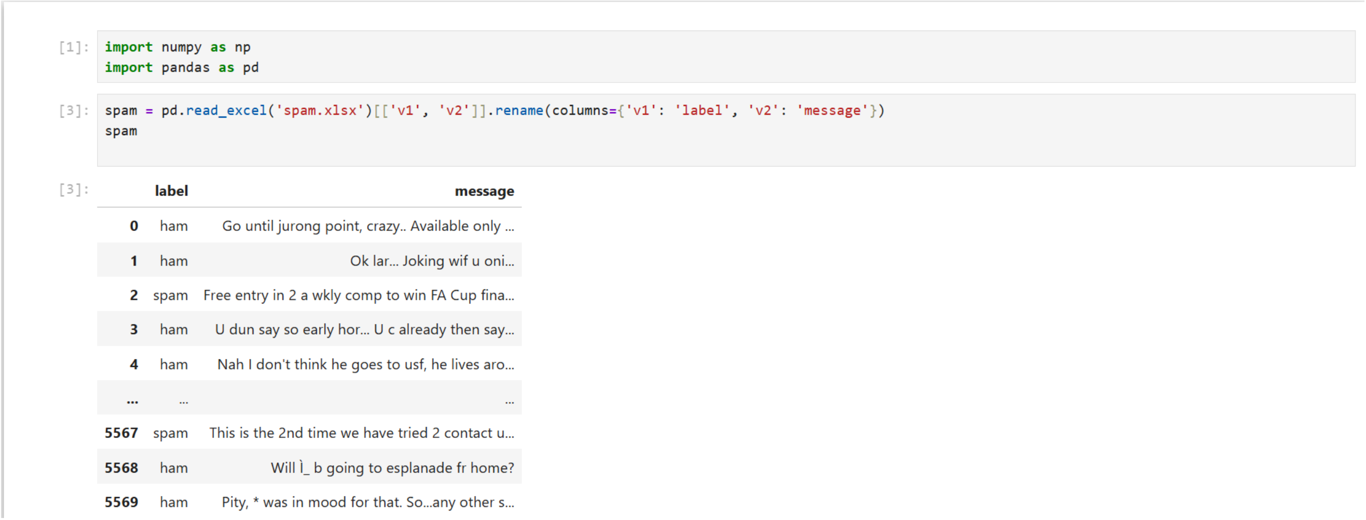
Name: Avishkar B. Bhapkar

DIV: CS6

Batch: C64

Roll no.: 78

PRN: 202401080056



[1]:

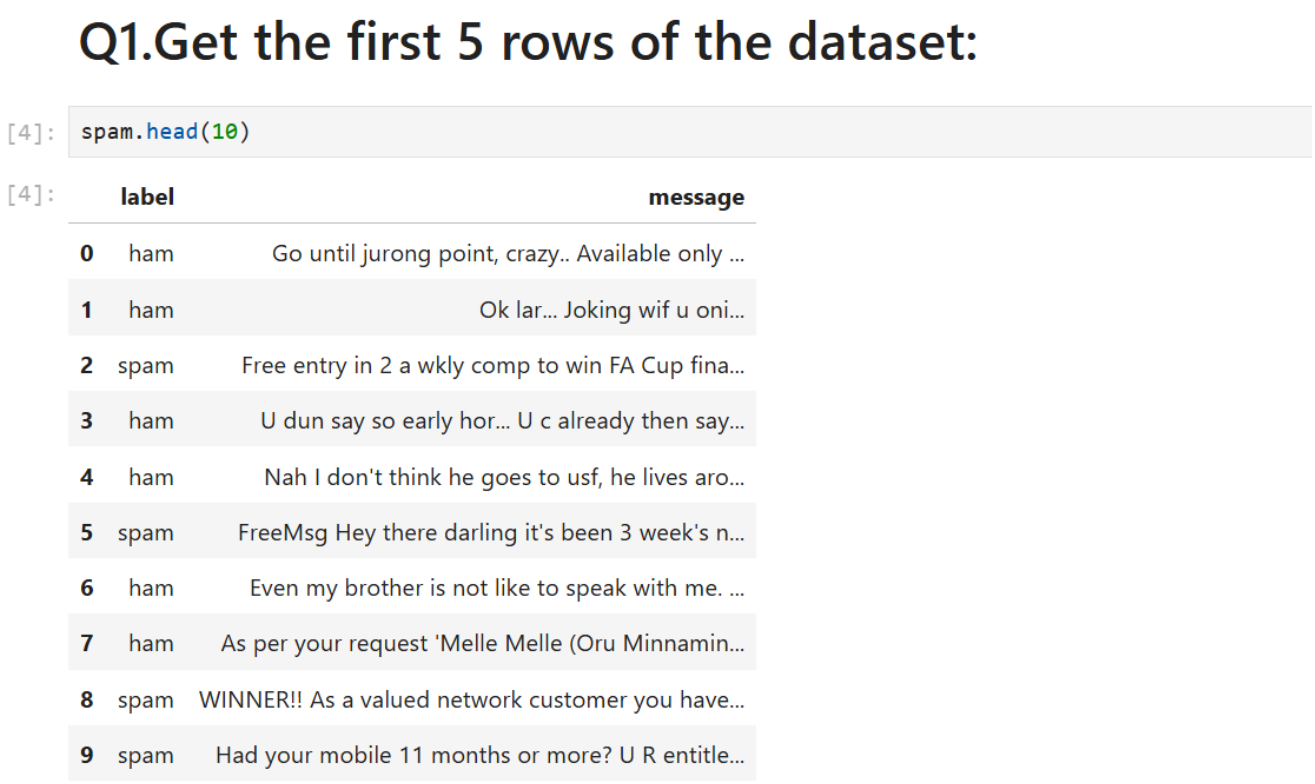
```
import numpy as np
import pandas as pd
```

[3]:

```
spam = pd.read_excel('spam.xlsx')[['v1', 'v2']].rename(columns={'v1': 'label', 'v2': 'message'})
spam
```

[3]:

	label	message
0	ham	Go until jurong point, crazy.. Available only ...
1	ham	Ok lar... Joking wif u oni...
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to usf, he lives aro...
...	...	...
5567	spam	This is the 2nd time we have tried 2 contact u...
5568	ham	Will Ì_ b going to esplanade fr home?
5569	ham	Pity, * was in mood for that. So...any other s...



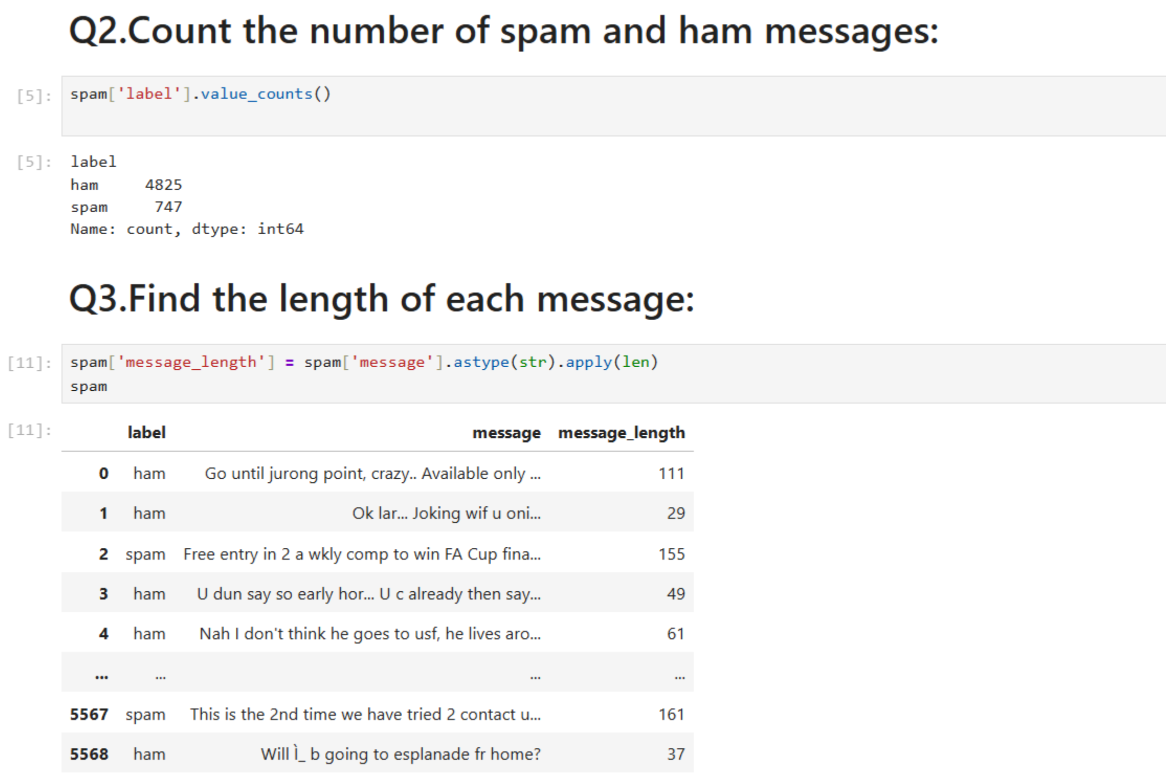
Q1.Get the first 5 rows of the dataset:

[4]:

```
spam.head(10)
```

[4]:

	label	message
0	ham	Go until jurong point, crazy.. Available only ...
1	ham	Ok lar... Joking wif u oni...
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to usf, he lives aro...
5	spam	FreeMsg Hey there darling it's been 3 week's n...
6	ham	Even my brother is not like to speak with me. ...
7	ham	As per your request 'Melle Melle (Oru Minnamin...
8	spam	WINNER!! As a valued network customer you have...
9	spam	Had your mobile 11 months or more? U R entitle...



Q2.Count the number of spam and ham messages:

```
[5]: spam['label'].value_counts()
```

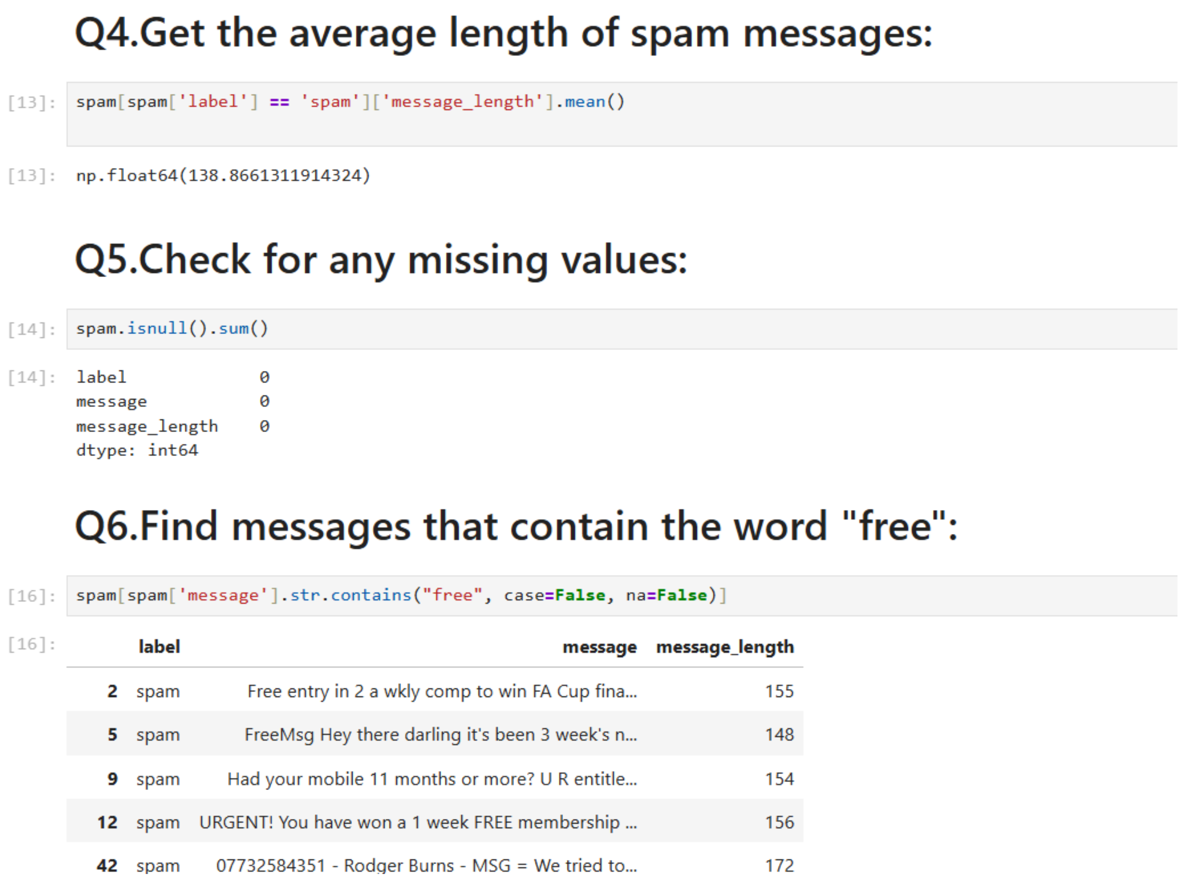
```
[5]: label
     ham     4825
     spam     747
     Name: count, dtype: int64
```

Q3.Find the length of each message:

```
[11]: spam['message_length'] = spam['message'].astype(str).apply(len)
      spam
```

[11]:

	label	message	message_length
0	ham	Go until jurong point, crazy.. Available only ...	111
1	ham	Ok lar... Joking wif u oni...	29
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	155
3	ham	U dun say so early hor... U c already then say...	49
4	ham	Nah I don't think he goes to usf, he lives aro...	61
...	...	...	...
5567	spam	This is the 2nd time we have tried 2 contact u...	161
5568	ham	Will Ì_ b going to esplanade fr home?	37



Q4.Get the average length of spam messages:

```
[13]: spam[spam['label'] == 'spam']['message_length'].mean()
```

```
[13]: np.float64(138.8661311914324)
```

Q5.Check for any missing values:

```
[14]: spam.isnull().sum()
```

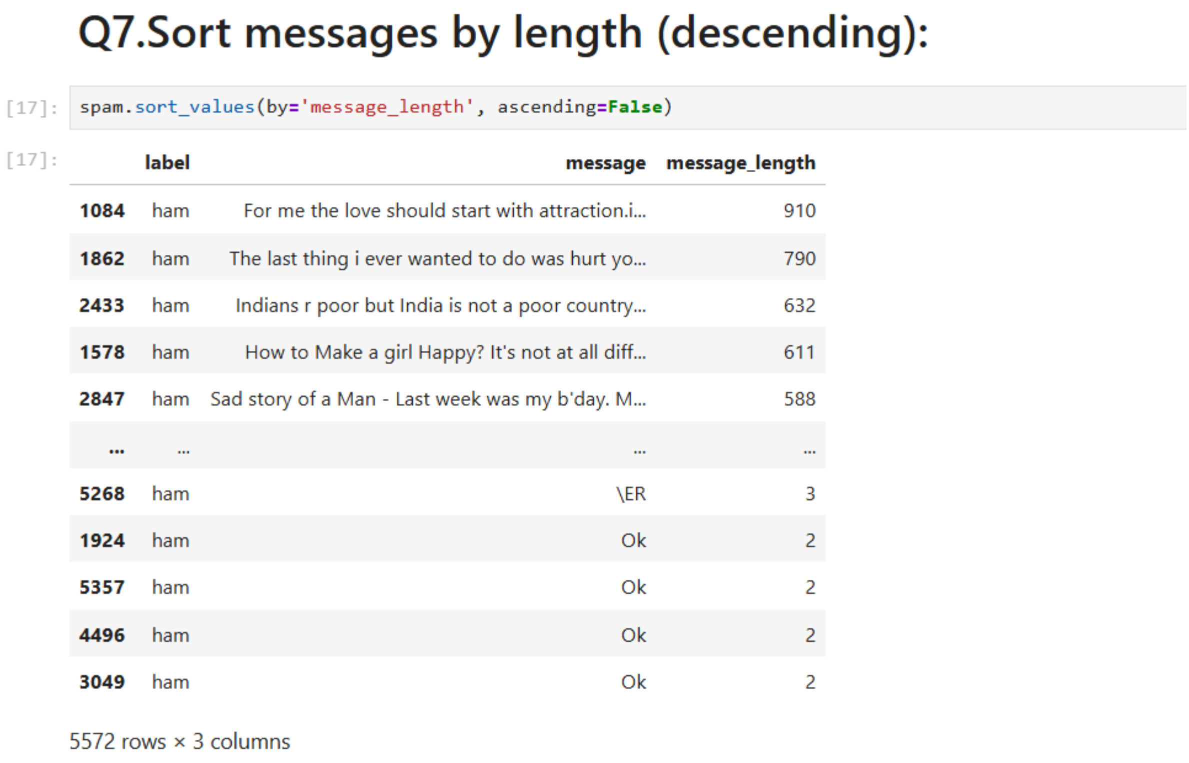
```
[14]: label             0
      message           0
      message_length    0
      dtype: int64
```

Q6.Find messages that contain the word "free":

```
[16]: spam[spam['message'].str.contains("free", case=False, na=False)]
```

[16]:

	label	message	message_length
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	155
5	spam	FreeMsg Hey there darling it's been 3 week's n...	148
9	spam	Had your mobile 11 months or more? U R entitle...	154
12	spam	URGENT! You have won a 1 week FREE membership ...	156
42	spam	07732584351 - Rodger Burns - MSG = We tried to...	172



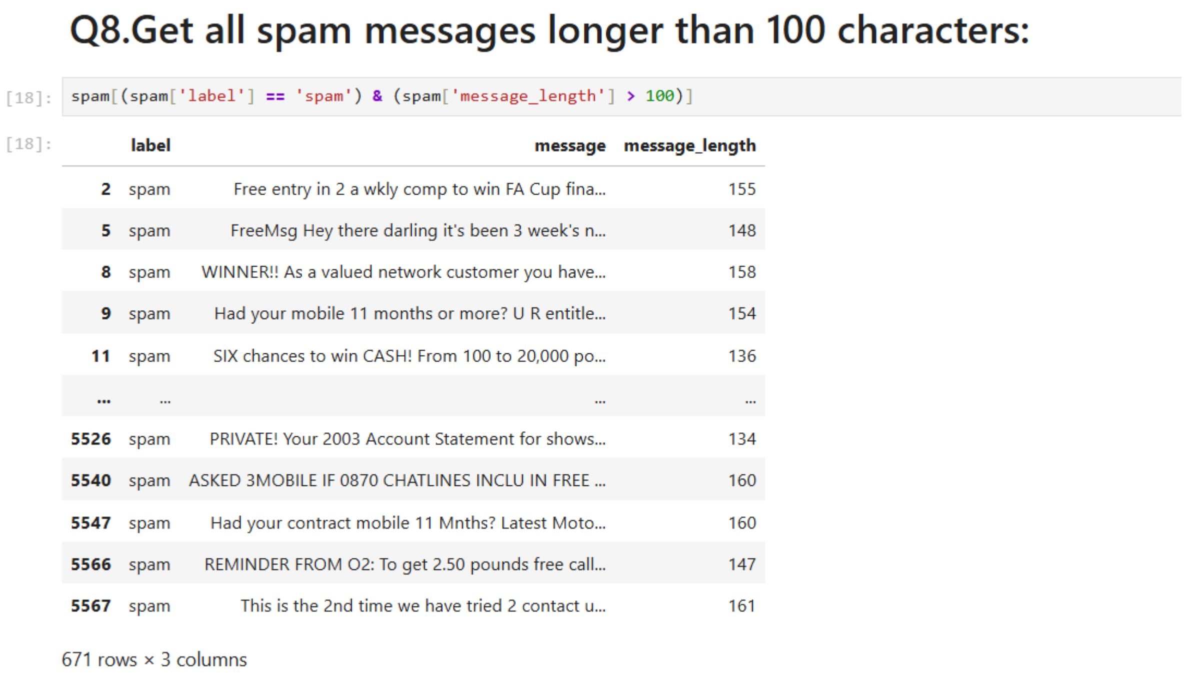
Q7.Sort messages by length (descending):

```
[17]: spam.sort_values(by='message_length', ascending=False)
```

[17]:

	label	message	message_length
1084	ham	For me the love should start with attraction.i...	910
1862	ham	The last thing i ever wanted to do was hurt yo...	790
2433	ham	Indians r poor but India is not a poor country...	632
1578	ham	How to Make a girl Happy? It's not at all diff...	611
2847	ham	Sad story of a Man - Last week was my b'day. M...	588
...	...	...	...
5268	ham	\ER	3
1924	ham	Ok	2
5357	ham	Ok	2
4496	ham	Ok	2
3049	ham	Ok	2

5572 rows × 3 columns



Q8.Get all spam messages longer than 100 characters:

```
[18]: spam[(spam['label'] == 'spam') & (spam['message_length'] > 100)]
```

[18]:

	label	message	message_length
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	155
5	spam	FreeMsg Hey there darling it's been 3 week's n...	148
8	spam	WINNER!! As a valued network customer you have...	158
9	spam	Had your mobile 11 months or more? U R entitle...	154
11	spam	SIX chances to win CASH! From 100 to 20,000 po...	136
...	...	...	...
5526	spam	PRIVATE! Your 2003 Account Statement for shows...	134
5540	spam	ASKED 3MOBILE IF 0870 CHATLINES INCLU IN FREE ...	160
5547	spam	Had your contract mobile 11 Mnths? Latest Moto...	160
5566	spam	REMINDER FROM O2: To get 2.50 pounds free call...	147
5567	spam	This is the 2nd time we have tried 2 contact u...	161

671 rows × 3 columns

Q9.Add a column with the number of words in each message:

[22]:

```
spam['word_count'] = spam['message'].astype(str).apply(lambda x: len(x.split()))
spam
```

[22]:

	label	message	message_length	word_count
0	ham	Go until jurong point, crazy.. Available only ...	111	20
1	ham	Ok lar... Joking wif u oni...	29	6
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	155	28
3	ham	U dun say so early hor... U c already then say...	49	11
4	ham	Nah I don't think he goes to usf, he lives aro...	61	13
...	...	...	...	...
5567	spam	This is the 2nd time we have tried 2 contact u...	161	30
5568	ham	Will Ì_ b going to esplanade fr home?	37	8
5569	ham	Pity, * was in mood for that. So...any other s...	57	10
5570	ham	The guy did some bitching but I acted like i'd...	125	26
5571	ham	Rofl. Its true to its name	26	6

5572 rows × 4 columns

Q10.Group by label and calculate average word count:

[23]:

```
spam.groupby('label')['word_count'].mean()
```

[23]:

```
label
ham      14.200207
spam     23.851406
Name: word_count, dtype: float64
```

Q11.Convert message lengths to a NumPy array:

[29]:

```
lengths = spam['message'].astype(str).apply(len).to_numpy()
```

Q12.Find mean message length:

[34]:

```
print('mean message length is',np.mean(lengths))
```

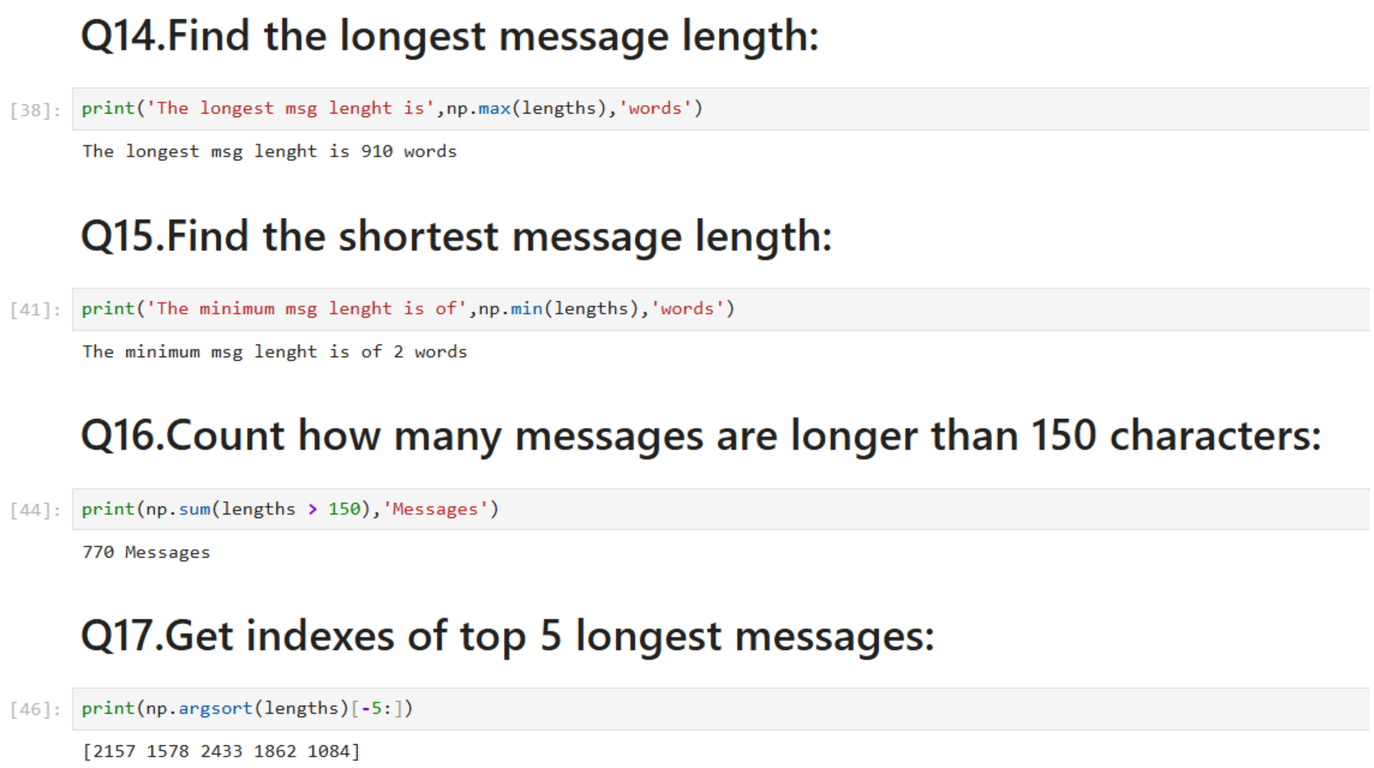
```
mean message length is 80.12096195262025
```

Q13.Find standard deviation of message lengths:

[36]:

```
print('standard deviation of msg lenght is',np.std(lengths))
```

```
standard deviation of msg lenght is 59.688258208117176
```



Q14.Find the longest message length:

```
[38]: print('The longest msg lenght is',np.max(lengths),'words')
```

```
The longest msg lenght is 910 words
```

Q15.Find the shortest message length:

```
[41]: print('The minimum msg lenght is of',np.min(lengths),'words')
```

```
The minimum msg lenght is of 2 words
```

Q16.Count how many messages are longer than 150 characters:

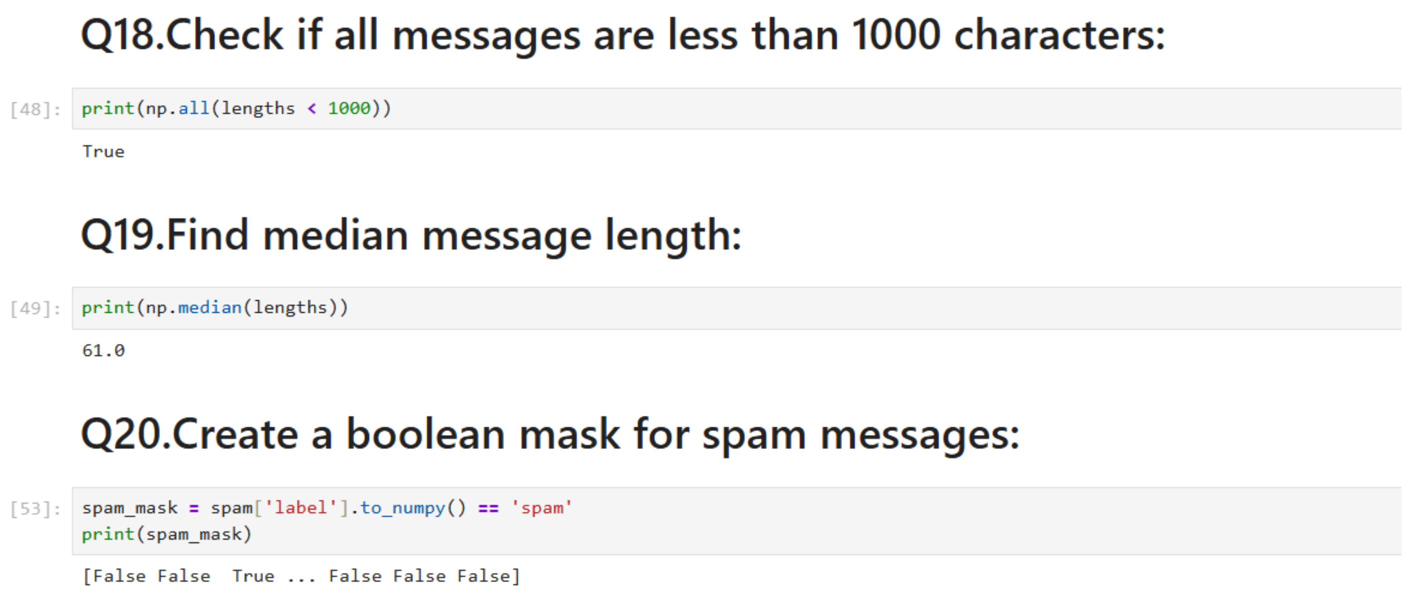
```
[44]: print(np.sum(lengths > 150),'Messages')
```

```
770 Messages
```

Q17.Get indexes of top 5 longest messages:

```
[46]: print(np.argsort(lengths)[-5:])
```

```
[2157 1578 2433 1862 1084]
```



Q18.Check if all messages are less than 1000 characters:

```
[48]: print(np.all(lengths < 1000))
```

```
True
```

Q19.Find median message length:

```
[49]: print(np.median(lengths))
```

```
61.0
```

Q20.Create a boolean mask for spam messages:

```
[53]: spam_mask = spam['label'].to_numpy() == 'spam'
      print(spam_mask)
```

```
[False False  True ... False False False]
```