# EDS Theory Activity No. 1
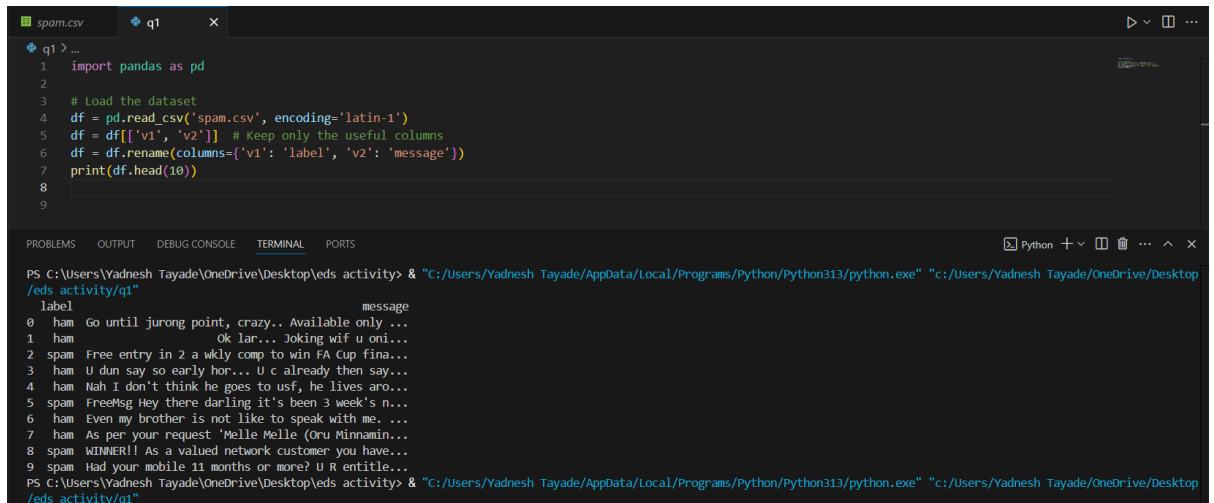
Name: Yadnesh Nilkanth Tayade

Roll no: CS7 – 78

PRN: 202401110057

# 1. View first 10 rows

```python
import pandas as pd

# Load the dataset
df = pd.read_csv('spam.csv', encoding='latin-1')
df = df[['v1', 'v2']]  # Keep only the useful columns
df = df.rename(columns={'v1': 'label', 'v2': 'message'})
print(df.head(10))
```

```
PS C:\Users\Yadnesh Tayade\OneDrive\Desktop\eds activity> & "C:/Users/Yadnesh Tayade/AppData/Local/Programs/Python/Python313/python.exe" "c:/Users/Yadnesh Tayade/OneDrive/Desktop
/eds activity/q1"
  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...
PS C:\Users\Yadnesh Tayade\OneDrive\Desktop\eds activity> & "C:/Users/Yadnesh Tayade/AppData/Local/Programs/Python/Python313/python.exe" "c:/Users/Yadnesh Tayade/OneDrive/Desktop
/eds activity/q1"
```

# 2. How many rows and columns are there?

```python
# Load the dataset
df = pd.read_csv('spam.csv', encoding='latin-1')
df = df[['v1', 'v2']]  # Keep only the useful columns
df = df.rename(columns={'v1': 'label', 'v2': 'message'})
print(df.shape)
```

```
PS C:\Users\Yadnesh Tayade\OneDrive\Desktop\eds activity> & "C:/Users/Yadnesh Tayade/AppData/Local/Programs/Python/Python313/python.exe" "c:/Users/Yadnesh Tayade/OneDrive/De
/eds activity/q2"
(5572, 2)
```

# 3. Check column names

```python
import pandas as pd

# Load the dataset
df = pd.read_csv('spam.csv', encoding='latin-1')
df = df[['v1', 'v2']]  # Keep only the useful columns
df = df.rename(columns={'v1': 'label', 'v2': 'message'})
print(df.columns)
```

```
PS C:\Users\Yadnesh Tayade\OneDrive\Desktop\eds activity> & "C:/Users/Yadnesh Tayade/AppData/Local/Programs/Python/Python313/python.exe" "c:/Users/Yadnesh Tayade/OneDrive/Deskt
/eds activity/q3"
Index(['label', 'message'], dtype='object')
```

# 4. Check data types

```
q4 > ...
1    import pandas as pd
2
3    df = pd.read_csv('spam.csv', encoding='latin-1')
4    df = df[['v1', 'v2']]  # Keep only the useful columns
5    df = df.rename(columns={'v1': 'label', 'v2': 'message'})
6    print(df.dtypes)
7
8
```

```
PROBLEMS    OUTPUT    DEBUG CONSOLE    TERMINAL    PORTS

PS C:\Users\Yadnesh Tayade\OneDrive\Desktop\eds activity> & "C:/Users/Yadnesh Tayade/AppData/Local/Programs/Python/Python313/python.exe" "c:/Users/Yadnesh Tayade/OneDrive/Desktop
/eds activity/q4"
label      object
message    object
dtype: object
```

# 5. Check if there are any missing values

```
q5 > ...
1    import pandas as pd
2
3    # Load the dataset
4    df = pd.read_csv('spam.csv', encoding='latin-1')
5    df = df[['v1', 'v2']]  # Keep only the useful columns
6    df = df.rename(columns={'v1': 'label', 'v2': 'message'})
7    print(df.isnull().sum())
8
9
```

```
PROBLEMS    OUTPUT    DEBUG CONSOLE    TERMINAL    PORTS

PS C:\Users\Yadnesh Tayade\OneDrive\Desktop\eds activity> & "C:/Users/Yadnesh Tayade/AppData/Local/Programs/Python/Python313/python.exe" "c:/Users/Yadnesh Tayade/OneDrive/Desktop
/eds activity/q5"
label      0
message    0
dtype: int64
```

# 6. Count how many spam and ham messages

```
q6 > ...
1    import pandas as pd
2
3    # Load the dataset
4    df = pd.read_csv('spam.csv', encoding='latin-1')
5    df = df[['v1', 'v2']]  # Keep only the useful columns
6    df = df.rename(columns={'v1': 'label', 'v2': 'message'})
7    print(df['label'].value_counts())
8
9
```

```
PROBLEMS    OUTPUT    DEBUG CONSOLE    TERMINAL    PORTS

PS C:\Users\Yadnesh Tayade\OneDrive\Desktop\eds activity> & "C:/Users/Yadnesh Tayade/AppData/Local/Programs/Python/Python313/python.exe" "c:/Users/Yadnesh Tayade/OneDrive/Desktop
/eds activity/q6"
label
ham     4825
spam     747
Name: count, dtype: int64
```

# 7. Display all spam messages

```
q7 > ...
1    import pandas as pd
2
3    # Load the dataset
4    df = pd.read_csv('spam.csv', encoding='latin-1')
5    df = df[['v1', 'v2']]  # Keep only the useful columns
6    df = df.rename(columns={'v1': 'label', 'v2': 'message'})
7    print(df[df['label'] == 'spam'])
8
9
```

```
PROBLEMS    OUTPUT    DEBUG CONSOLE    TERMINAL    PORTS

PS C:\Users\Yadnesh Tayade\OneDrive\Desktop\eds activity> & "C:/Users/Yadnesh Tayade/AppData/Local/Programs/Python/Python313/python.exe" "c:/Users/Yadnesh Tayade/OneDrive/Desktop
/eds activity/q7"
      label                                          message
2      spam  Free entry in 2 a wkly comp to win FA Cup fina...
5      spam  FreeMsg Hey there darling it's been 3 week's n...
8      spam  WINNER!! As a valued network customer you have...
9      spam  Had your mobile 11 months or more? U R entitle...
11     spam  SIX chances to win CASH! From 100 to 20,000 po...
...     ...                                              ...
5537   spam  Want explicit SEX in 30 secs? Ring 02073162414...
5540   spam  ASKED 3MOBILE IF 0870 CHATLINES INCLU IN FREE ...
5547   spam  Had your contract mobile 11 Mnths? Latest Moto...
5566   spam  REMINDER FROM O2: To get 2.50 pounds free call...
5567   spam  This is the 2nd time we have tried 2 contact u...

[747 rows x 2 columns]
```

# 8. Display all ham (non-spam) messages

```
q8 > ...
1    import pandas as pd
2
3    # Load the dataset
4    df = pd.read_csv('spam.csv', encoding='latin-1')
5    df = df[['v1', 'v2']]  # Keep only the useful columns
6    df = df.rename(columns={'v1': 'label', 'v2': 'message'})
7    print(df[df['label'] == 'ham'])
8
9
```

```
PROBLEMS    OUTPUT    DEBUG CONSOLE    TERMINAL    PORTS

PS C:\Users\Yadnesh Tayade\OneDrive\Desktop\eds activity> & "C:/Users/Yadnesh Tayade/AppData/Local/Programs/Python/Python313/python.exe" "c:/Users/Yadnesh Tayade/OneDrive/Desktop
/eds activity/q8"
      label                                          message
0       ham  Go until jurong point, crazy.. Available only ...
1       ham                      Ok lar... Joking wif u oni...
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...
6       ham  Even my brother is not like to speak with me. ...
...     ...                                              ...
5565    ham                                      Huh y lei...
5568    ham              Will Ì_ b going to esplanade fr home?
5569    ham  Pity, * was in mood for that. So...any other s...
5570    ham  The guy did some bitching but I acted like i'd...
5571    ham                         Rofl. Its true to its name

[4825 rows x 2 columns]
```

# 9. Sort messages alphabetically

```python
import pandas as pd

# Load the dataset
df = pd.read_csv('spam.csv', encoding='latin-1')
df = df[['v1', 'v2']]  # Keep only the useful columns
df = df.rename(columns={'v1': 'label', 'v2': 'message'})
sorted_df = df.sort_values(by='message')
print(sorted_df.head())
```

```
PS C:\Users\Yadnesh Tayade\OneDrive\Desktop\eds activity> & "C:/Users/Yadnesh Tayade/AppData/Local/Programs/Python/Python313/python.exe" "c:/Users/Yadnesh Tayade/OneDrive/Desktop
/eds activity/q8"
      label                                          message
4024  ham              &lt;#&gt;  in mca. But not conform.
3014  ham    &lt;#&gt;  mins but i had to stop somewhere f...
4264  ham    &lt;DECIMAL&gt; m but its not a common car he...
2765  ham            and  picking them up from various points
3671  ham    came to look at the flat, seems ok, in his 50...
```

# 10. Find the longest message

```python
import pandas as pd

# Load the dataset
df = pd.read_csv('spam.csv', encoding='latin-1')
df = df[['v1', 'v2']]  # Keep only the useful columns
df = df.rename(columns={'v1': 'label', 'v2': 'message'})
df['length'] = df['message'].apply(len)
print(df[df['length'] == df['length'].max()])
```

```
PS C:\Users\Yadnesh Tayade\OneDrive\Desktop\eds activity> & "C:/Users/Yadnesh Tayade/AppData/Local/Programs/Python/Python313/python.exe" "c:/Users/Yadnesh Tayade/OneDrive/Desktop
/eds activity/q10"
      label                                    message  length
1084  ham  For me the love should start with attraction.i...     910
```

# 11. Add a column with word count in each message

```python
import pandas as pd

# Load the dataset
df = pd.read_csv('spam.csv', encoding='latin-1')
df = df[['v1', 'v2']]  # Keep only the useful columns
df = df.rename(columns={'v1': 'label', 'v2': 'message'})
df['word_count'] = df['message'].apply(lambda x: len(x.split()))
print(df[['message', 'word_count']])
```

```
PS C:\Users\Yadnesh Tayade\OneDrive\Desktop\eds activity> & "C:/Users/Yadnesh Tayade/AppData/Local/Programs/Python/Python313/python.exe" "c:/Users/Yadnesh Tayade/OneDrive/Desktop
/eds activity/q11"
                                          message  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
...                                           ...         ...
5567  This is the 2nd time we have tried 2 contact u...          30
5568            Will Ì_ b going to esplanade fr home?           8
5569  Pity, * was in mood for that. So...any other s...          10
5570  The guy did some bitching but I acted like i'd...          26
5571                     Rofl. Its true to its name           6

[5572 rows x 2 columns]
```

## 12. Find length of each message

```
q12 > ...
 1   import pandas as pd
 2
 3   # Load dataset
 4   df = pd.read_csv('spam.csv', encoding='latin-1')
 5   df = df[['v1', 'v2']]  # Keeping useful columns only
 6   df = df.rename(columns={'v1': 'label', 'v2': 'message'})
 7   lengths = df['message'].apply(len)
 8   print(lengths.head())
 9
10
```

```
PROBLEMS   OUTPUT   DEBUG CONSOLE   TERMINAL   PORTS

PS C:\Users\Yadnesh Tayade\OneDrive\Desktop\eds activity> & "C:/Users/Yadnesh Tayade/AppData/Local/Programs/Python/Python313/python.exe" "c:/Users/Yadnesh Tayade/OneDrive/Desktop
/eds activity/q12"
0    111
1     29
2    155
3     49
4     61
Name: message, dtype: int64
```

## 13. Convert message lengths into a NumPy array

```
q13 > ...
 1   import pandas as pd
 2   import numpy as np   # <-- You forgot this!
 3
 4   # Load dataset
 5   df = pd.read_csv('spam.csv', encoding='latin-1')
 6   df = df[['v1', 'v2']]  # Keeping useful columns only
 7   df = df.rename(columns={'v1': 'label', 'v2': 'message'})
 8
 9   # Create message lengths array
10   lengths = df['message'].apply(len)
11   lengths_array = np.array(lengths)
12
13   # Print result
14   print(lengths_array)
15
```

```
PROBLEMS   OUTPUT   DEBUG CONSOLE   TERMINAL   PORTS

PS C:\Users\Yadnesh Tayade\OneDrive\Desktop\eds activity> & "C:/Users/Yadnesh Tayade/AppData/Local/Programs/Python/Python313/python.exe" "c:/Users/Yadnesh Tayade/OneDrive/Desktop
/eds activity/q13"
[111  29 155 ...  57 125  26]
PS C:\Users\Yadnesh Tayade\OneDrive\Desktop\eds activity>
```

## 14. Find the average (mean) message length

```
 1   import pandas as pd
 2   import numpy as np   # <-- You forgot this!
 3
 4   # Load dataset
 5   df = pd.read_csv('spam.csv', encoding='latin-1')
 6   df = df[['v1', 'v2']]  # Keeping useful columns only
 7   df = df.rename(columns={'v1': 'label', 'v2': 'message'})
 8
 9   # Create message lengths array
10   lengths = df['message'].apply(len)
11   lengths_array = np.array(lengths)
12
13   average_length = np.mean(lengths_array)
14   print(average_length)
15
```

```
PROBLEMS   OUTPUT   DEBUG CONSOLE   TERMINAL   PORTS

PS C:\Users\Yadnesh Tayade\OneDrive\Desktop\eds activity> & "C:/Users/Yadnesh Tayade/AppData/Local/Programs/Python/Python313/python.exe" "c:/Users/Yadnesh Tayade/OneDrive/Desktop
/eds activity/q13"
80.11880832735105
PS C:\Users\Yadnesh Tayade\OneDrive\Desktop\eds activity>
```

## 15. Find the shortest message

```
1   import pandas as pd
2   import numpy as np    # <-- You forgot this!
3
4   # Load dataset
5   df = pd.read_csv('spam.csv', encoding='latin-1')
6   df = df[['v1', 'v2']]  # Keeping useful columns only
7   df = df.rename(columns={'v1': 'label', 'v2': 'message'})
8
9   # Create message lengths array
10  lengths = df['message'].apply(len)
11  lengths_array = np.array(lengths)
12  shortest = np.min(lengths_array)
13  print(shortest)
14
```

```
PROBLEMS   OUTPUT   DEBUG CONSOLE   TERMINAL   PORTS
PS C:\Users\Yadnesh Tayade\OneDrive\Desktop\eds activity> & "C:/Users/Yadnesh Tayade/AppData/Local/Programs/Python/Python313/python.exe" "c:/Users/Yadnesh Tayade/OneDrive/Desktop
/eds activity/q13"
2
```

## 16. Find the longest message

```
1   import pandas as pd
2   import numpy as np    # <-- You forgot this!
3
4   # Load dataset
5   df = pd.read_csv('spam.csv', encoding='latin-1')
6   df = df[['v1', 'v2']]  # Keeping useful columns only
7   df = df.rename(columns={'v1': 'label', 'v2': 'message'})
8
9   # Create message lengths array
10  lengths = df['message'].apply(len)
11  lengths_array = np.array(lengths)
12  longest = np.max(lengths_array)
13  print(longest)
14
```

```
PROBLEMS   OUTPUT   DEBUG CONSOLE   TERMINAL   PORTS
PS C:\Users\Yadnesh Tayade\OneDrive\Desktop\eds activity> & "C:/Users/Yadnesh Tayade/AppData/Local/Programs/Python/Python313/python.exe" "c:/Users/Yadnesh Tayade/OneDrive/Desktop
/eds activity/q13"
910
```

## 17. Standard deviation of message lengths

```
1   import pandas as pd
2   import numpy as np    # <-- You forgot this!
3
4   # Load dataset
5   df = pd.read_csv('spam.csv', encoding='latin-1')
6   df = df[['v1', 'v2']]  # Keeping useful columns only
7   df = df.rename(columns={'v1': 'label', 'v2': 'message'})
8
9   # Create message lengths array
10  lengths = df['message'].apply(len)
11  lengths_array = np.array(lengths)
12  std_dev = np.std(lengths_array)
13  print(std_dev)
14
```

```
PROBLEMS   OUTPUT   DEBUG CONSOLE   TERMINAL   PORTS
PS C:\Users\Yadnesh Tayade\OneDrive\Desktop\eds activity> & "C:/Users/Yadnesh Tayade/AppData/Local/Programs/Python/Python313/python.exe" "c:/Users/Yadnesh Tayade/OneDrive/Desktop
/eds activity/q13"
59.6854842151988
```

## 18. Find how many short messages (< 20

characters)

```
1   import pandas as pd
2   import numpy as np   # <-- You forgot this!
3
4   # Load dataset
5   df = pd.read_csv('spam.csv', encoding='latin-1')
6   df = df[['v1', 'v2']]  # Keeping useful columns only
7   df = df.rename(columns={'v1': 'label', 'v2': 'message'})
8
9   # Create message lengths array
10  lengths = df['message'].apply(len)
11  lengths_array = np.array(lengths)
12
13  short_messages_count = np.sum(lengths_array < 20)
14  print(short_messages_count)
15
16
```

```
PROBLEMS   OUTPUT   DEBUG CONSOLE   TERMINAL   PORTS                                                    >_ Python + ∨ □ �识 ... ∧ ✕
PS C:\Users\Yadnesh Tayade\OneDrive\Desktop\eds activity> & "C:/Users/Yadnesh Tayade/AppData/Local/Programs/Python/Python313/python.exe" "c:/Users/Yadnesh Tayade/OneDrive/Desktop
/eds activity/q14"
156
PS C:\Users\Yadnesh Tayade\OneDrive\Desktop\eds activity> & "C:/Users/Yadnesh Tayade/AppData/Local/Programs/Python/Python313/python.exe" "c:/Users/Yadnesh Tayade/OneDrive/Desktop
/eds activity/q14"
156
PS C:\Users\Yadnesh Tayade\OneDrive\Desktop\eds activity>
```

## 19. Randomly select 5 message lengths

```
1   import pandas as pd
2   import numpy as np   # <-- You forgot this!
3
4   # Load dataset
5   df = pd.read_csv('spam.csv', encoding='latin-1')
6   df = df[['v1', 'v2']]  # Keeping useful columns only
7   df = df.rename(columns={'v1': 'label', 'v2': 'message'})
8
9   # Create message lengths array
10  lengths = df['message'].apply(len)
11  lengths_array = np.array(lengths)
12
13  random_lengths = np.random.choice(lengths_array, 5)
14  print(random_lengths)
15
16
```

```
PROBLEMS   OUTPUT   DEBUG CONSOLE   TERMINAL   PORTS                                                    >_ Python + ∨ □ ⩍ ... ∧ ✕
PS C:\Users\Yadnesh Tayade\OneDrive\Desktop\eds activity> & "C:/Users/Yadnesh Tayade/AppData/Local/Programs/Python/Python313/python.exe" "c:/Users/Yadnesh Tayade/OneDrive/Desktop
/eds activity/q14"
[  7 143 100  37  54]
PS C:\Users\Yadnesh Tayade\OneDrive\Desktop\eds activity>
```

## 20. Grain: Count Total Spam and Ham Messages

```
1   import pandas as pd
2   import numpy as np   # <-- You forgot this!
3
4   # Load dataset
5   df = pd.read_csv('spam.csv', encoding='latin-1')
6   df = df[['v1', 'v2']]  # Keeping useful columns only
7   df = df.rename(columns={'v1': 'label', 'v2': 'message'})
8
9   # Create message lengths array
10  lengths = df['message'].apply(len)
11  lengths_array = np.array(lengths)
12
13  counts = df['label'].value_counts()
14  print(counts)
15
```

```
nesh Tayade\OneDrive\Desktop\eds activity\spam.csv
PROBLEMS   OUTPUT   DEBUG CONSOLE   TERMINAL   PORTS                                                    >_ Python + ∨ □ ⩍ ... ∧ ✕
PS C:\Users\Yadnesh Tayade\OneDrive\Desktop\eds activity> & "C:/Users/Yadnesh Tayade/AppData/Local/Programs/Python/Python313/python.exe" "c:/Users/Yadnesh Tayade/OneDrive/Desktop
/eds activity/q14"
label
ham     4825
spam     747
Name: count, dtype: int64
```