# aneefdev

August 25, 2024

[3]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

[7]:
```
data=pd.read_csv("C:/Users/MANOJ/Downloads/spam_ham_dataset.csv")
```

[9]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5171 entries, 0 to 5170
Data columns (total 4 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   Unnamed: 0  5171 non-null   int64 
 1   label       5171 non-null   object
 2   text        5171 non-null   object
 3   label_num   5171 non-null   int64 
dtypes: int64(2), object(2)
memory usage: 161.7+ KB
```

[11]:
```
data.head()
```

[11]:

| | Unnamed: 0 | label | text | label_num |
|---|---|---|---|---|
| 0 | 605 | ham | Subject: enron methanol ; meter # : 988291\r\n… | 0 |
| 1 | 2349 | ham | Subject: hpl nom for january 9 , 2001\r\n( see… | 0 |
| 2 | 3624 | ham | Subject: neon retreat\r\nho ho ho , we ' re ar… | 0 |
| 3 | 4685 | spam | Subject: photoshop , windows , office . cheap … | 1 |
| 4 | 2030 | ham | Subject: re : indian springs\r\nthis deal is t… | 0 |

[13]:
```
data.tail()
```

[13]:
```
      Unnamed: 0 label                                               text  \
5166        1518   ham  Subject: put the 10 on the ft\r\nthe transport…
5167         404   ham  Subject: 3 / 4 / 2000 and following noms\r\nhp…
5168        2933   ham  Subject: calpine daily gas nomination\r\n>\r\n…
5169        1409   ham  Subject: industrial worksheets for august 2000…
5170        4807  spam  Subject: important online banking alert\r\ndea…

      label_num
5166          0
5167          0
5168          0
5169          0
5170          1
```

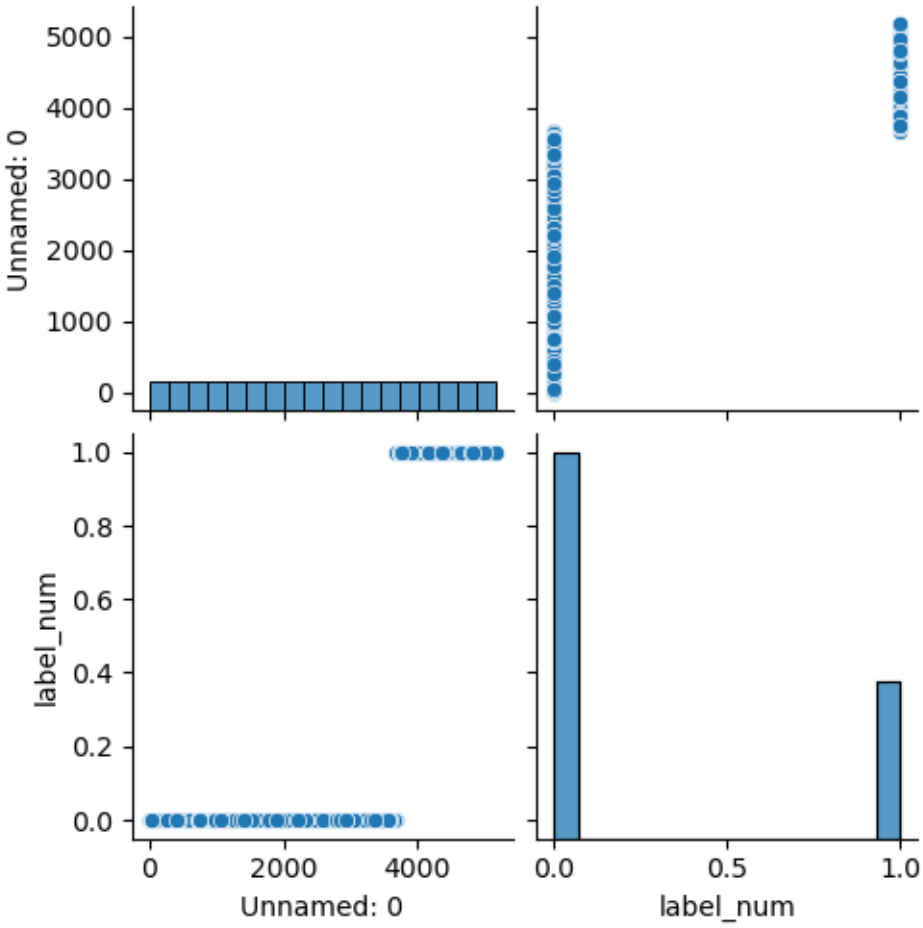
[15]:
```
data.describe()
```

[15]:
```
         Unnamed: 0    label_num
count   5171.000000  5171.000000
mean    2585.000000     0.289886
std     1492.883452     0.453753
min        0.000000     0.000000
25%     1292.500000     0.000000
50%     2585.000000     0.000000
75%     3877.500000     1.000000
max     5170.000000     1.000000
```

[17]:
```
data.isnull().sum()
```

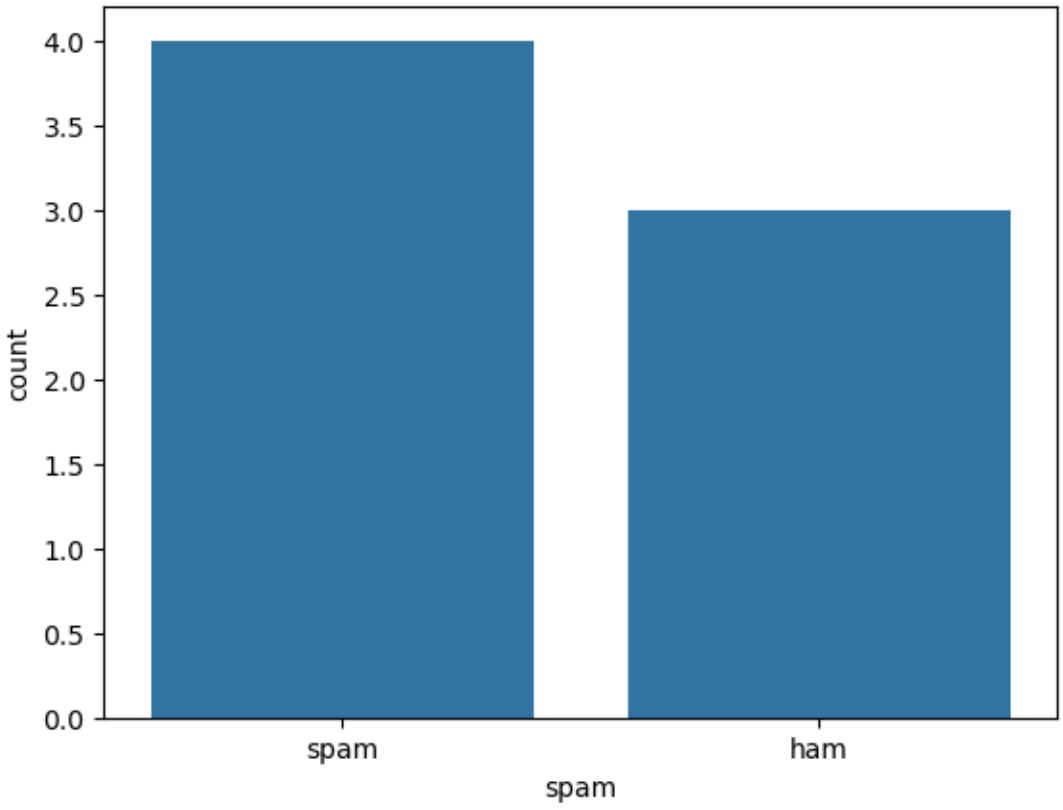
[17]:
```
Unnamed: 0    0
label         0
text          0
label_num     0
dtype: int64
```

[21]:
```
sns.pairplot(data)
plt.show()
```




[79]:
```
data = pd.DataFrame({'spam': ['spam', 'ham', 'spam', 'ham', 'spam', 'spam',␣
↪'ham']})
```

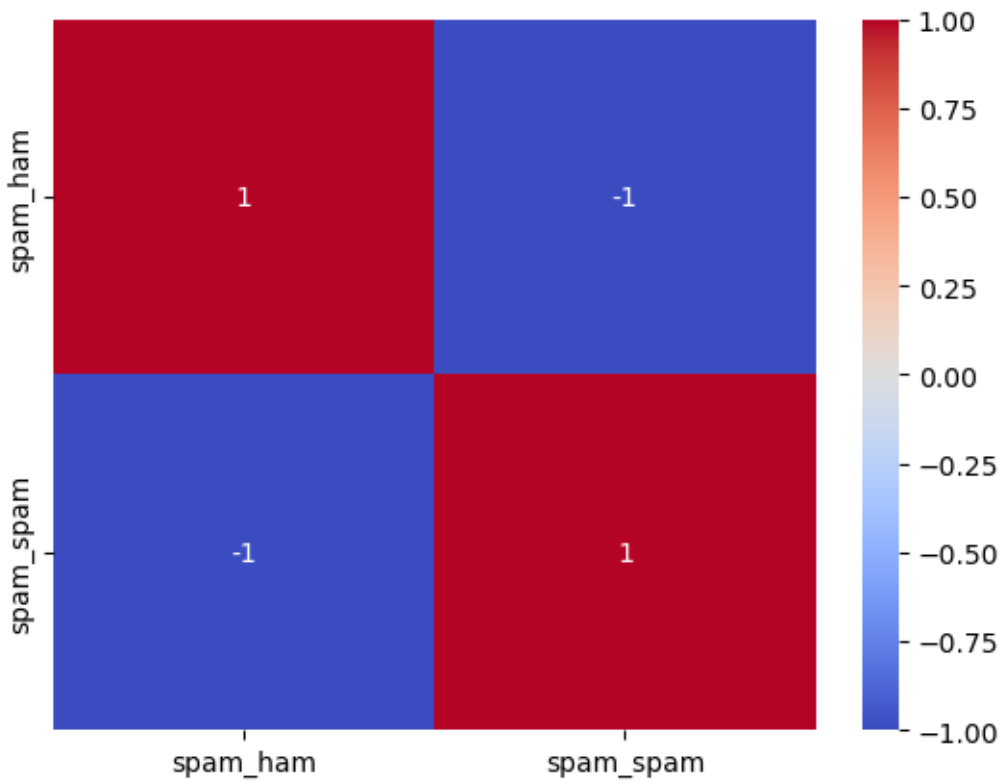
[81]:
```
sns.countplot(x='spam',data=data)
plt.show()
```




[91]:
```
data_encoded = pd.get_dummies(data)
```

[93]:
```
correlation_matrix = data_encoded.corr()
```

[95]:
```
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.show()
```


spam_ham
spam_spam
1
-1
-1
1
spam_ham
spam_spam
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

[ ]: