```python
import pandas as pd
import seaborn as sns
```

```python
df=pd.read_csv("spam.csv")
```

```python
df
```

	Category	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 ü 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

5572 rows × 2 columns

```python
df.groupby('Category').describe()
```

			Message	
	count	unique	top	freq
Category				
ham	4825	4516	Sorry, I'll call later	30
spam	747	641	Please call our customer service representativ...	4

```python
def add(value):
    if value=="ham":
        return 0
    else:
        return 1
df['newtype']=df['Category'].apply(add)
```

```python
df    # we convert the category to 0, 1 bcz the model runs on numeriv value
```

	Category	Message	newtype
0	ham	Go until jurong point, crazy.. Available only ...	0
1	ham	Ok lar... Joking wif u oni...	0
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	1
3	ham	U dun say so early hor... U c already then say...	0
4	ham	Nah I don't think he goes to usf, he lives aro...	0
...	...	...	...
5567	spam	This is the 2nd time we have tried 2 contact u...	1
5568	ham	Will ü b going to esplanade fr home?	0
5569	ham	Pity, * was in mood for that. So...any other s...	0
5570	ham	The guy did some bitching but I acted like i'd...	0
5571	ham	Rofl. Its true to its name	0

5572 rows × 3 columns

```python
df.drop('Category',axis=1)
```

	Message	newtype
0	Go until jurong point, crazy.. Available only ...	0
1	Ok lar... Joking wif u oni...	0
2	Free entry in 2 a wkly comp to win FA Cup fina...	1
3	U dun say so early hor... U c already then say...	0
4	Nah I don't think he goes to usf, he lives aro...	0
...	...	...
5567	This is the 2nd time we have tried 2 contact u...	1
5568	Will ü b going to esplanade fr home?	0
5569	Pity, * was in mood for that. So...any other s...	0
5570	The guy did some bitching but I acted like i'd...	0
5571	Rofl. Its true to its name	0

5572 rows × 2 columns

```python
from sklearn.model_selection import train_test_split
```

```python
x_train,x_test,y_train,y_test=train_test_split(df.Message,df.newtype,test_size=0.2)
```

```python

```

```python

```

```python
from sklearn.pipeline import Pipeline
```

```python
from sklearn.naive_bayes import MultinomialNB
```

```python
from sklearn.feature_extraction.text import CountVectorizer
```

```python
ram=Pipeline([
    ("vector",CountVectorizer()),         # pipeline do two job first to convert the text to vector then pass to model
    ("class",MultinomialNB())
])
```

```python
ram.fit(x_train,y_train)
```

▸ **Pipeline**

▸ CountVectorizer

▸ MultinomialNB

```python
email=["100% discount","HI, My friend"]
```

```python
ram.predict(email)
```

array([1, 0], dtype=int64)

```python
#1 shows the email is spam and 0 means no spam
```

```python
ram.score(x_test,y_test)
```

0.9838565022421525

here the accuracy is high but when we run the code same time it will change because the train data will be random each time.

we can use the multinomialNB to find out which category the x_train data is like message, wine, laptop etc.

In [ ]: