

Lp1Da4

December 10, 2021

```
[0]: import numpy as np
import pandas as pd
%matplotlib inline

from sklearn.preprocessing import LabelEncoder

import matplotlib.pyplot as plt
import seaborn as sns

data=pd.read_csv('store.csv')
```

```
[0]: data.head() #####info about data
```

```
[0]:
```

	Duration	Start date	End date	Start station number \
0	1012	2010-09-20 11:27:04	2010-09-20 11:43:56	31208
1	61	2010-09-20 11:41:22	2010-09-20 11:42:23	31209
2	2690	2010-09-20 12:05:37	2010-09-20 12:50:27	31600
3	1406	2010-09-20 12:06:05	2010-09-20 12:29:32	31600
4	1413	2010-09-20 12:10:43	2010-09-20 12:34:17	31100

	Start station	End station number \
0	M St & New Jersey Ave SE	31108
1	1st & N St SE	31209
2	5th & K St NW	31100
3	5th & K St NW	31602
4	19th St & Pennsylvania Ave NW	31201

	End station	Bike number	Member type
0	4th & M St SW	W00742	Member
1	1st & N St SE	W00032	Member
2	19th St & Pennsylvania Ave NW	W00993	Member
3	Park Rd & Holmead Pl NW	W00344	Member
4	15th & P St NW	W00883	Member

```
[0]: data.info() ##### info about data
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 115597 entries, 0 to 115596
```

```
Data columns (total 9 columns):
Duration                115597 non-null int64
Start date              115597 non-null object
End date                115597 non-null object
Start station number    115597 non-null int64
Start station           115597 non-null object
End station number      115597 non-null int64
End station             115597 non-null object
Bike number             115597 non-null object
Member type             115597 non-null object
dtypes: int64(3), object(6)
memory usage: 7.9+ MB
```

```
[0]: data.dtypes      ##### info about types
```

```
[0]: Duration                int64
Start date                  object
End date                    object
Start station number        int64
Start station               object
End station number          int64
End station                 object
Bike number                 object
Member type                 object
dtype: object
```

```
[0]: data['Duration'].describe()  ##### statistics
```

```
[0]: count      115597.000000
mean          1254.649956
std           2914.317998
min            60.000000
25%           403.000000
50%           665.000000
75%          1120.000000
max           85644.000000
Name: Duration, dtype: float64
```

```
[0]: data=data.drop('Start date',axis=1)
data=data.drop('End date',axis=1)
data=data.drop('Start station',axis=1)
data=data.drop('End station',axis=1)
```

```
[0]: data.head() ##### label encoder
le = LabelEncoder()
le.fit(data['Member type'])
data['Member type'] = le.transform(data['Member type'])
```

```
[0]: le = LabelEncoder()
le.fit(data['Bike number'])
data['Bike number'] = le.transform(data['Bike number'])
```

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

```
[0]:
```

	Duration	Start station number	End station number	Bike number	\
0	1012	31208	31108	614	
1	61	31209	31209	41	
2	2690	31600	31100	836	
3	1406	31600	31602	282	
4	1413	31100	31201	734	

```

Member type
0          1
1          1
2          1
3          1
4          1
```

```
[0]: data.shape      ##### data.size
```

```
[0]: (115597, 5)
```

```
[0]: train=np.array(data.iloc[0:85000])    ### spitting into training and tetsign
test=np.array(data.iloc[85000:,])
```

```
[0]: train.shape,test.shape      ##### train and test
```

```
[0]: ((85000, 5), (30597, 5))
```

```
[0]: from sklearn.naive_bayes import GaussianNB    ##### guassinan
model=GaussianNB()
```

```
[0]: model.fit(train[:,0:4],train[:,4])
predicted=model.predict(test[:,0:4])
```

```
[0]: predicted.shape
```

```
[0]: (30597,)
```

```
[0]: predicted
```

```
[0]: array([1, 1, 1, ..., 1, 1, 0])
```

```
[0]: count=0      ### accuracy
for l in range(30597):
    if(predicted[l]==test[l,4]):
```

```
count=count+1
```

```
[0]: count
```

```
[0]: 28086
```

```
[0]: print(count/30597)
```

```
0.9179331306990881
```