

```

import numpy as np
import pandas as pd
from datetime import datetime

df = pd.read_csv('Orders.csv')

In [57]:
df
Out[57]:
   OrderID  InvoiceDate  CustomerID  OrderVolume  ShipMode  Segment  Category  PaymentTerm  Discount  ExistingPurchaseOrder  FirstCustomerOrder  CashDate
0  O-100  2011-01-03  1024  276.1  Same Day  Corporate  Office Supplies  30  5%  Yes  Yes  2011-02-18
1  O-101  2011-01-03  1024  35.88  Same Day  Corporate  Office Supplies  30  5%  Yes  No  2011-02-18
2  O-102  2011-01-06  1006  66.12  Second Class  Consumer  Office Supplies  30  5%  No  Yes  2011-02-06
3  O-104  2011-01-06  1023  408.3  Standard Class  Consumer  Office Supplies  45  3%  No  Yes  2011-02-17
4  O-105  2011-01-06  1009  314.22  Standard Class  Consumer  Technology  45  3%  Yes  Yes  2011-02-10
...
37637 O-48951 2014-09-11 1029 50.09 Standard Class Corporate Office Supplies 14 3% No No 2014-12-31
37638 O-48953 2014-09-11 1010 21.12 Same Day Consumer Office Supplies 14 5% Yes No 2014-12-31
37639 O-48960 2014-11-30 1031 33.57 First Class Consumer Office Supplies 14 5% Yes No 2014-12-16
37640 O-48972 2014-11-30 1031 20.72 Second Class Consumer Office Supplies 30 3% Yes No 2014-12-15
37641 O-48977 2014-11-30 1000 1.2 Standard Class Consumer Office Supplies 14 0% Yes No 2014-12-30
37642 rows x 12 columns

In [57]:
df['value'] = np.arange(len(df))
df
Out[57]:
   OrderID  InvoiceDate  CustomerID  OrderVolume  ShipMode  Segment  Category  PaymentTerm  Discount  ExistingPurchaseOrder  FirstCustomerOrder  CashDate  value
0  O-100  2011-01-03  1024  276.1  Same Day  Corporate  Office Supplies  30  5%  Yes  Yes  2011-02-18  0
1  O-101  2011-01-03  1024  35.88  Same Day  Corporate  Office Supplies  30  5%  Yes  No  2011-02-18  1
2  O-102  2011-01-06  1006  66.12  Second Class  Consumer  Office Supplies  30  5%  No  Yes  2011-02-06  2
3  O-104  2011-01-06  1023  408.3  Standard Class  Consumer  Office Supplies  45  3%  No  Yes  2011-02-17  3
4  O-105  2011-01-06  1009  314.22  Standard Class  Consumer  Technology  45  3%  Yes  Yes  2011-02-10  4
...
37637 O-48951 2014-09-11 1029 50.09 Standard Class Corporate Office Supplies 14 3% No No 2014-12-31 37637
37638 O-48953 2014-09-11 1010 21.12 Same Day Consumer Office Supplies 14 5% Yes No 2014-12-31 37638
37639 O-48960 2014-11-30 1031 33.57 First Class Consumer Office Supplies 14 5% Yes No 2014-12-16 37639
37640 O-48972 2014-11-30 1031 20.72 Second Class Consumer Office Supplies 30 3% Yes No 2014-12-15 37640
37641 O-48977 2014-11-30 1000 1.2 Standard Class Consumer Office Supplies 14 0% Yes No 2014-12-30 37641
37642 rows x 13 columns

In [58]:
df.shape
Out[58]:
(37642, 13)

In [59]:
df.isnull().sum()
Out[59]:
OrderID      0
InvoiceDate   0
CustomerID    0
OrderVolume   0
ShipMode      0
Segment       0
Category      0
PaymentTerm   0
Discount      0
ExistingPurchaseOrder  0
FirstCustomerOrder  0
CashDate      0
dtype: int64

In [60]:
df['OrderVolume'].value_counts()
Out[60]:
15.06    47
25.92     39
32.0       2
19.44     29
15.55     29
212.65     1
862.46     1
90.99     1
1007.74     1
421.19     1
Name: OrderVolume, Length: 16952, dtype: int64

In [61]:
#extracting column from dataframe
df_test = df[['InvoiceDate', 'CashDate', 'value']]
df_test
Out[61]:
   InvoiceDate  CashDate  value
0  2011-01-03  2011-02-18    0
1  2011-01-03  2011-02-18    1
2  2011-01-06  2011-02-06    2
3  2011-01-06  2011-02-17    3
4  2011-01-06  2011-02-10    4
...
37637 2014-09-11 2014-12-31 37637
37638 2014-09-11 2014-12-31 37638
37639 2014-11-30 2014-12-16 37639
37640 2014-11-30 2014-12-15 37640
37641 2014-11-30 2014-12-30 37641
37642 rows x 3 columns

In [62]:
#converting column to numeric format
df['OrderVolume'] = pd.to_numeric(df['OrderVolume'], errors='coerce')
df['OrderVolume'].dtypes
Out[62]:
dtype('float64')

In [63]:
fill = df['OrderVolume'].mean()
df['OrderVolume'] = df['OrderVolume'].fillna(df['OrderVolume'].mean())

In [64]:
#filling null values with mean
df['OrderVolume'] = df['OrderVolume'].fillna(df['OrderVolume'].mean())

In [65]:
df['OrderVolume']
Out[65]:
0    276.10
1     35.88
2     66.12
3    408.30
4    314.22
...
37637    50.09
37638    21.12
37639    33.57
37640    20.72
37641     1.2
Name: OrderVolume, Length: 37642, dtype: float64

In [66]:
df_test.dtypes
Out[66]:
InvoiceDate    object
CashDate       object
dtype: object

In [67]:
df.loc[df['Discount'] == '5%','Discount'] = 5
df.loc[df['Discount'] == '3%','Discount'] = 3
df.loc[df['Discount'] == '0%','Discount'] = 0

In [68]:
df['Discount'].value_counts()
Out[68]:
5    13748
3     9440
0      4138
Name: Discount, dtype: int64

In [69]:
mode_dis = df['Discount'].mode()

In [70]:
mode_dis
Out[70]:
0
dtype: object

In [71]:
mode_dis = 5

In [72]:
df['Discount'] = df['Discount'].fillna(mode_dis)

In [73]:
df = df.drop(['OrderID'],axis=1)

In [74]:
df
Out[74]:
   InvoiceDate  CustomerID  OrderVolume  ShipMode  Segment  Category  PaymentTerm  Discount  ExistingPurchaseOrder  FirstCustomerOrder  CashDate  value
0  2011-01-03  1024  276.1  Same Day  Corporate  Office Supplies  30  5%  Yes  Yes  2011-02-18  0
1  2011-01-03  1024  35.88  Same Day  Corporate  Office Supplies  30  5%  Yes  No  2011-02-18  1
2  2011-01-06  1006  66.12  Second Class  Consumer  Office Supplies  30  5%  No  Yes  2011-02-06  2
3  2011-01-06  1023  408.30  Standard Class  Consumer  Office Supplies  45  3%  No  Yes  2011-02-17  3
4  2011-01-06  1009  314.22  Standard Class  Consumer  Technology  45  3%  Yes  Yes  2011-02-10  4
...
37637 2014-09-11 1029 50.09 Standard Class Corporate Office Supplies 14 3% No No 2014-12-31 37637
37638 2014-09-11 1010 21.12 Same Day Consumer Office Supplies 14 5% Yes No 2014-12-31 37638
37639 2014-11-30 1031 33.57 First Class Consumer Office Supplies 14 5% Yes No 2014-12-16 37639
37640 2014-11-30 1031 20.72 Second Class Consumer Office Supplies 30 3% Yes No 2014-12-15 37640
37641 2014-11-30 1000 1.2 Standard Class Consumer Office Supplies 14 0% Yes No 2014-12-30 37641
37642 rows x 12 columns

In [75]:
#transforming dataframe to numeric
df['FirstCustomerOrder'] = df['FirstCustomerOrder'].replace(['No','Yes'],[0,1])
df['ExistingPurchaseOrder'] = df['ExistingPurchaseOrder'].replace(['No','Yes'],[0,1])

In [76]:
from sklearn.preprocessing import LabelEncoder

In [77]:
#encoding few columns for future needs
label = LabelEncoder()
df['ShipMode'] = label.fit_transform(df['ShipMode'])
df['Segment'] = label.fit_transform(df['Segment'])
df['Category'] = label.fit_transform(df['Category'])

In [78]:
df
Out[78]:
   InvoiceDate  CustomerID  OrderVolume  ShipMode  Segment  Category  PaymentTerm  Discount  ExistingPurchaseOrder  FirstCustomerOrder  CashDate  value
0  2011-01-03  1024  276.1  1  1  1  30  5%  Yes  Yes  2011-02-18  0
1  2011-01-03  1024  35.88  1  1  1  30  5%  Yes  No  2011-02-18  1
2  2011-01-06  1006  66.12  2  0  1  30  5%  No  Yes  2011-02-06  2
3  2011-01-06  1023  408.30  3  0  1  30  5%  No  Yes  2011-02-17  3
4  2011-01-06  1009  314.22  3  0  2  45  3%  Yes  Yes  2011-02-10  4
...
37637 2014-09-11 1029 50.09 3  1  1  14  3%  No  No  2014-12-31 37637
37638 2014-09-11 1010 21.12 1  0  1  14  5%  Yes  No  2014-12-31 37638
37639 2014-11-30 1031 33.57 0  0  1  14  5%  Yes  No  2014-12-16 37639
37640 2014-11-30 1031 20.72 2  0  1  30  3%  Yes  No  2014-12-15 37640
37641 2014-11-30 1000 1.20 3  0  1  14  0%  Yes  No
```