GRIP : THE SPARKS FOUNDATION

DATA SCIENCE AND BUSINESS ANALYTICS INTERN.

AUTHOR : SWETA PATEL

TASK 3 : Exploratory Data Analysis - Retail.

Aim : to focus on a business manager who will try to find out weak areas where he can work to make more profit.

In [7]:
```
import pandas as pd
import numpy as np
import seaborn as sns
from matplotlib import pyplot as plt
import sklearn.datasets as datasets
import warnings
warnings.filterwarnings('ignore')
```

Importing Data

In [14]:
```
pwd
```

Out[14]:
```
'/Users/apple'
```

In [16]:
```
df = pd.read_csv(r"/Users/apple/Desktop/task3/SampleSuperstore.csv")
df.head()
```

Out[16]:

	Ship Mode	Segment	Country	City	State	Postal Code	Region	Category	Sub-Category
0	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Bookcases
1	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Chairs
2	Second Class	Corporate	United States	Los Angeles	California	90036	West	Office Supplies	Labels
3	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Furniture	Tables
4	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Office Supplies	Storage

Exploring Data

In [17]:
```
df.shape
```

Out[17]: (9994, 13)

In [19]:
```
df.info
```

Out[19]:
```
<bound method DataFrame.info of            Ship Mode    Segment        Coun
try             City       State  \
0       Second Class   Consumer  United States        Henderson    Kentucky
1       Second Class   Consumer  United States        Henderson    Kentucky
2       Second Class  Corporate  United States      Los Angeles  California
3     Standard Class   Consumer  United States  Fort Lauderdale     Florida
4     Standard Class   Consumer  United States  Fort Lauderdale     Florida
...              ...        ...            ...              ...         ...
9989    Second Class   Consumer  United States            Miami     Florida
9990  Standard Class   Consumer  United States       Costa Mesa  California
9991  Standard Class   Consumer  United States       Costa Mesa  California
9992  Standard Class   Consumer  United States       Costa Mesa  California
9993    Second Class   Consumer  United States      Westminster  California

      Postal Code Region         Category Sub-Category     Sales  Quantity
\
0           42420  South        Furniture    Bookcases  261.9600         2
1           42420  South        Furniture       Chairs  731.9400         3
2           90036   West  Office Supplies       Labels   14.6200         2
3           33311  South        Furniture       Tables  957.5775         5
4           33311  South  Office Supplies      Storage   22.3680         2
...           ...    ...              ...          ...       ...       ...
9989        33180  South        Furniture  Furnishings   25.2480         3
9990        92627   West        Furniture  Furnishings   91.9600         2
9991        92627   West       Technology       Phones  258.5760         2
9992        92627   West  Office Supplies        Paper   29.6000         4
9993        92683   West  Office Supplies   Appliances  243.1600         2

      Discount    Profit
0         0.00   41.9136
1         0.00  219.5820
2         0.00    6.8714
3         0.45 -383.0310
4         0.20    2.5164
...        ...       ...
9989      0.20    4.1028
9990      0.00   15.6332
9991      0.20   19.3932
9992      0.00   13.3200
9993      0.00   72.9480

[9994 rows x 13 columns]>
```

Checking for Null Values

In [20]:
```
df.isnull().sum()
```

Out[20]:
```
Ship Mode       0
Segment         0
Country         0
City            0
State           0
Postal Code     0
Region          0
Category        0
Sub-Category    0
Sales           0
Quantity        0
Discount        0
Profit          0
dtype: int64
```

In [21]:
```
df.shape
```

Out[21]:
```
(9994, 13)
```

Duplicate Values

In [22]:
```
sum(df.duplicated(subset = None, keep = 'first'))
```

Out[22]:
```
17
```

In [23]:
```
df.drop_duplicates()
```

Out[23]:

	Ship Mode	Segment	Country	City	State	Postal Code	Region	Category	C
0	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Bo
1	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	
2	Second Class	Corporate	United States	Los Angeles	California	90036	West	Office Supplies	
3	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Furniture	
4	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Office Supplies	
...	...	...	...	...	...	...	...	...	
9989	Second Class	Consumer	United States	Miami	Florida	33180	South	Furniture	Fur
9990	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Furniture	Fur
9991	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Technology	
9992	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Office Supplies	
9993	Second Class	Consumer	United States	Westminster	California	92683	West	Office Supplies	Ap

9977 rows × 13 columns

In [24]:

```
df.corr()
```

Out[24]:

	Postal Code	Sales	Quantity	Discount	Profit
Postal Code	1.000000	-0.023854	0.012761	0.058443	-0.029961
Sales	-0.023854	1.000000	0.200795	-0.028190	0.479064
Quantity	0.012761	0.200795	1.000000	0.008623	0.066253
Discount	0.058443	-0.028190	0.008623	1.000000	-0.219487
Profit	-0.029961	0.479064	0.066253	-0.219487	1.000000

In [25]:

```
df.cov()
```

Out[25]:

	Postal Code	Sales	Quantity	Discount	Profit
Postal Code	1.028080e+09	-476682.766590	910.415885	386.870404	-225045.849445
Sales	-4.766828e+05	388434.455308	278.459923	-3.627228	69944.096586
Quantity	9.104159e+02	278.459923	4.951113	0.003961	34.534769
Discount	3.868704e+02	-3.627228	0.003961	0.042622	-10.615173
Profit	-2.250458e+05	69944.096586	34.534769	-10.615173	54877.798055

In [26]:

```
df.head()
```

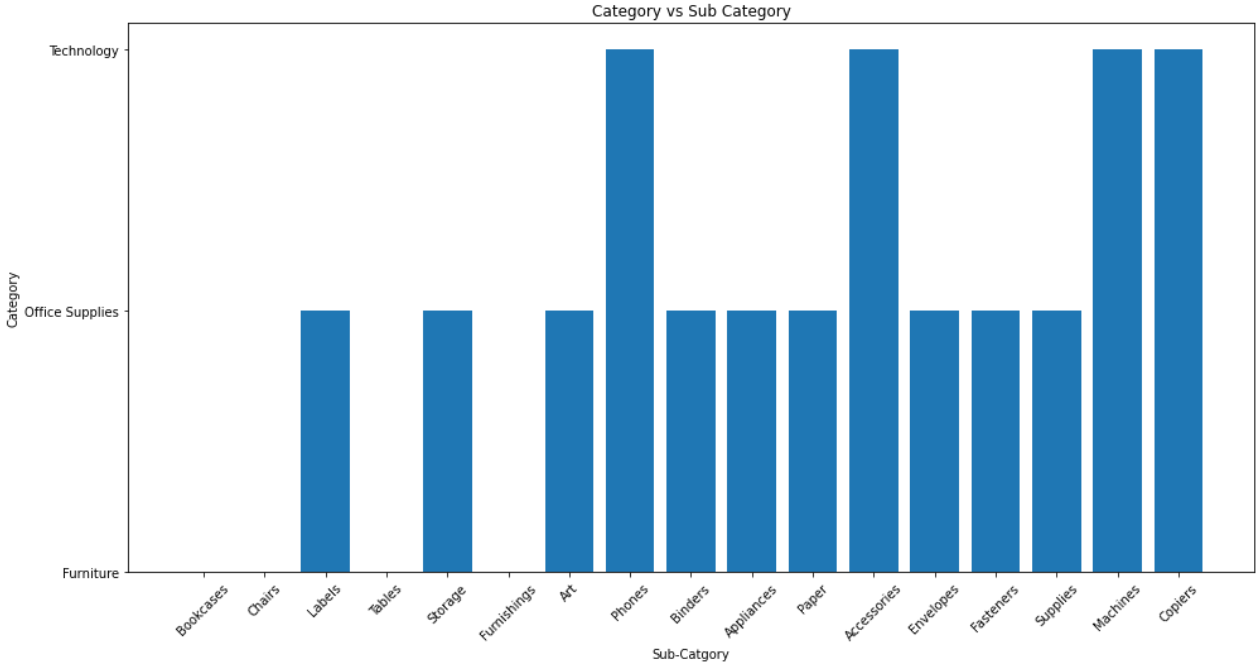
Out[26]:

	Ship Mode	Segment	Country	City	State	Postal Code	Region	Category	Sub-Category
0	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Bookcases
1	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Chairs
2	Second Class	Corporate	United States	Los Angeles	California	90036	West	Office Supplies	Labels
3	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Furniture	Tables
4	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Office Supplies	Storage

Data Visualisation

In [30]:

```
plt.figure(figsize=(16,8))
plt.bar('Sub-Category','Category', data=df)
plt.title('Category vs Sub Category')
plt.xlabel('Sub-Catgory')
plt.ylabel('Category')
plt.xticks(rotation=45)
plt.show()
```

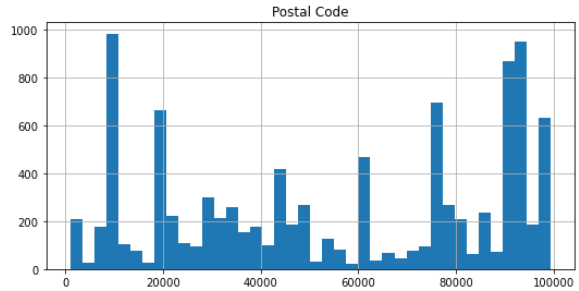


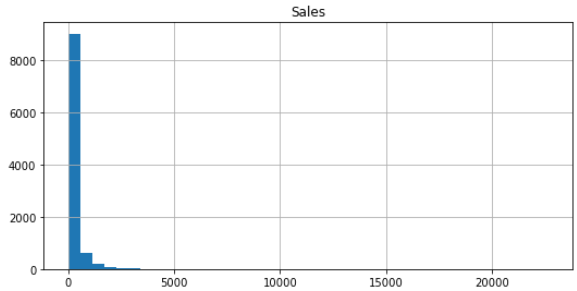
In [32]:
```
df.corr()
```

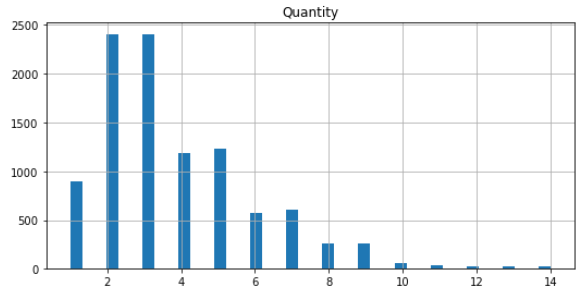
Out[32]:

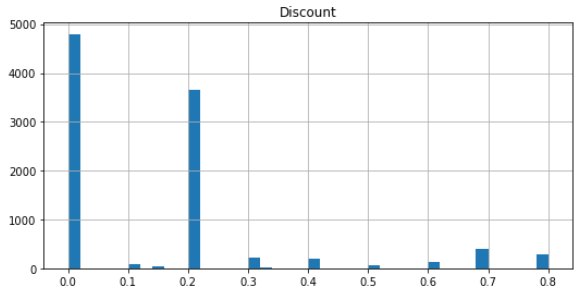
	Postal Code	Sales	Quantity	Discount	Profit
Postal Code	1.000000	-0.023854	0.012761	0.058443	-0.029961
Sales	-0.023854	1.000000	0.200795	-0.028190	0.479064
Quantity	0.012761	0.200795	1.000000	0.008623	0.066253
Discount	0.058443	-0.028190	0.008623	1.000000	-0.219487
Profit	-0.029961	0.479064	0.066253	-0.219487	1.000000

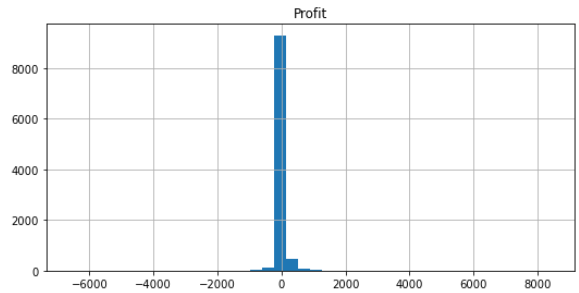
In [33]:
```
df.hist(bins=40,figsize=(20,15))
plt.show();
```









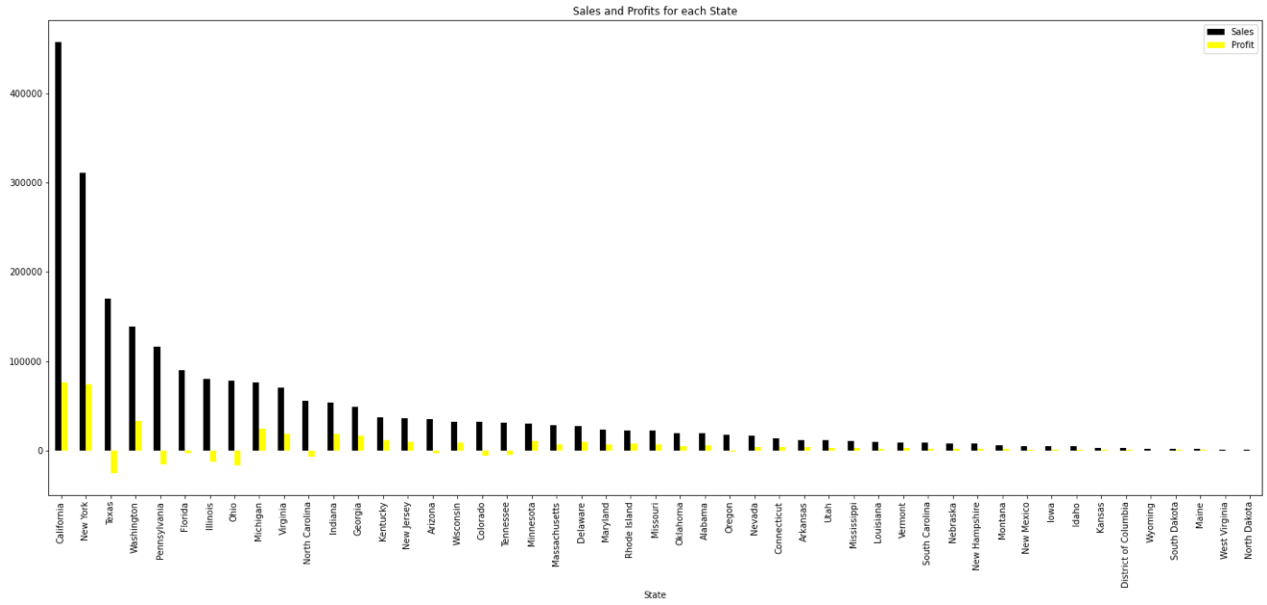


Repeated States

In [36]:
```
df['State'].value_counts()
```

Out[36]:
```
California                2001
New York                  1128
Texas                      985
Pennsylvania               587
Washington                 506
Illinois                   492
Ohio                       469
Florida                    383
Michigan                   255
North Carolina             249
Arizona                    224
Virginia                   224
Georgia                    184
Tennessee                  183
Colorado                   182
Indiana                    149
Kentucky                   139
Massachusetts              135
New Jersey                 130
Oregon                     124
Wisconsin                  110
Maryland                   105
Delaware                    96
Minnesota                   89
Connecticut                 82
Missouri                    66
Oklahoma                    66
Alabama                     61
Arkansas                    60
Rhode Island                56
Utah                        53
Mississippi                 53
South Carolina              42
Louisiana                   42
Nevada                      39
Nebraska                    38
New Mexico                  37
Iowa                        30
New Hampshire               27
Kansas                      24
Idaho                       21
Montana                     15
South Dakota                12
Vermont                     11
District of Columbia        10
Maine                        8
North Dakota                 7
West Virginia                4
Wyoming                      1
Name: State, dtype: int64
```

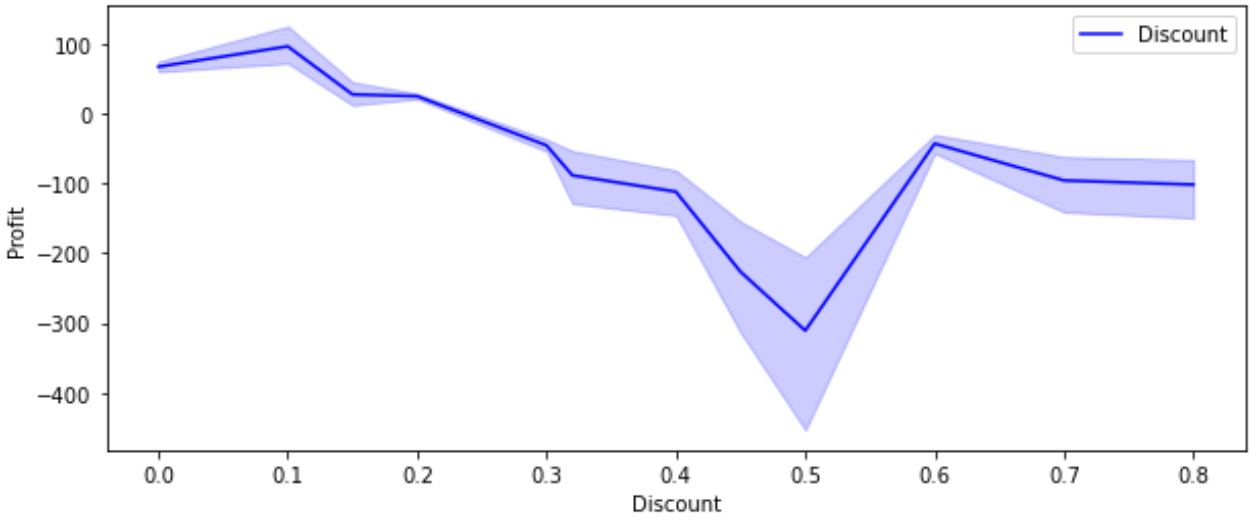
In [41]:
```
plt.rcParams["figure.figsize"] = [25,10]
df.groupby("State")[["Sales", "Profit"]].sum().sort_values(by = "Sales", as
plt.title("Sales and Profits for each State")
plt.show()
```



In [45]:
```
df['Discount'][df['Profit'] < 0].sort_values(ascending = True)
```

Out[45]:
```
9581    0.10
4645    0.10
6439    0.10
5827    0.10
5079    0.15
       ...
7786    0.80
2527    0.80
7781    0.80
8874    0.80
5097    0.80
Name: Discount, Length: 1871, dtype: float64
```

In [52]:
```
plt.figure(figsize=(10,4))
sns.lineplot('Discount','Profit', data=df , color='b',label='Discount')
plt.legend()
plt.show()
```



In [53]:

```
ps = df.groupby('Sub-Category')[['Sales','Profit']].sum().sort_values(by='S
ps[:].plot.bar(color=['red','lightblue'],figsize=(15,8))
plt.title('Profit/loss & Sales across states')
plt.xlabel('Sub-Category')
plt.ylabel('Profit/loss & Sales')
plt.show()
```