

Task-3: Exploratory Data Analysis-Retail

Author: Rudra Pratap Singh

DATA SCIENCE AND BUSINESS ANALYTICS INTERN @ THE SPARKS FOUNDATION

Importing Libraries

```
In [2]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Reading Data

```
In [4]: file=pd.read_csv('SampleSuperstore.csv')
```

Out[4]:	Ship Mode	Segment	Country	City	State	Postal Code	Region	Category	Sub-Category	Sales	Quantity	Discount	Profit
0	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Bookcases	261.9600	2	0.00	41.9136
1	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Chairs	731.9400	3	0.00	219.5820
2	Second Class	Corporate	United States	Los Angeles	California	90036	West	Office Supplies	Labels	14.6200	2	0.00	6.8714
3	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Furniture	Tables	957.5775	5	0.45	-383.0310
4	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Office Supplies	Storage	22.3680	2	0.20	2.5164
...
9989	Second Class	Consumer	United States	Miami	Florida	33180	South	Furniture	Furnishings	25.2480	3	0.20	4.1028
9990	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Furniture	Furnishings	91.9600	2	0.00	15.6332
9991	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Technology	Phones	258.5760	2	0.20	19.3932
9992	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Office Supplies	Paper	29.6000	4	0.00	13.3200
9993	Second Class	Consumer	United States	Westminster	California	92683	West	Office Supplies	Appliances	243.1600	2	0.00	72.9480

9994 rows x 13 columns

```
In [5]: file.tail(10)
```

Out[5]:	Ship Mode	Segment	Country	City	State	Postal Code	Region	Category	Sub-Category	Sales	Quantity	Discount	Profit
9984	Standard Class	Consumer	United States	Long Beach	New York	11561	East	Office Supplies	Labels	31.500	10	0.0	15.1200
9985	Standard Class	Consumer	United States	Long Beach	New York	11561	East	Office Supplies	Supplies	65.600	4	0.0	16.1240
9986	Standard Class	Consumer	United States	Los Angeles	California	90008	West	Technology	Accessories	36.240	1	0.0	15.2208
9987	Standard Class	Corporate	United States	Georgia	Georgia	30605	South	Technology	Accessories	79.990	1	0.0	28.7964
9988	Standard Class	Corporate	United States	Athens	Georgia	30605	South	Furniture	Phones	206.100	5	0.45	55.6470
9989	Second Class	Consumer	United States	Miami	Florida	33180	South	Furniture	Furnishings	25.248	3	0.2	4.1028
9990	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Furniture	Furnishings	91.960	2	0.0	15.6332
9991	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Technology	Phones	258.576	2	0.2	19.3932
9992	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Office Supplies	Paper	29.600	4	0.0	13.3200

9993	Second Class	Consumer	United States	Westminster	California	92683	West	Office Supplies	Appliances	243.160	2	0.0	72.9480
------	--------------	----------	---------------	-------------	------------	-------	------	-----------------	------------	---------	---	-----	---------

```
In [9]: file.info
```

Out[9]:	#	bound method DataFrame.info of	Ship Mode	Segment	Country	City	State	Postal Code	Region	Category	Sub-Category	Sales	Quantity	Discount	Profit
0	bound method DataFrame.info of	Second Class	Consumer	United States	Henderson	Kentucky	Kentucky	42420	South	Furniture	Bookcases	261.9600	2	0.00	41.9136
1	bound method DataFrame.info of	Second Class	Consumer	United States	Henderson	Kentucky	Kentucky	42420	South	Furniture	Chairs	731.9400	3	0.00	219.5820
2	bound method DataFrame.info of	Second Class	Corporate	United States	Los Angeles	California	California	90036	West	Office Supplies	Labels	14.6200	2	0.00	6.8714
3	bound method DataFrame.info of	Standard Class	Consumer	United States	Fort Lauderdale	Florida	Florida	33311	South	Furniture	Tables	957.5775	5	0.45	-383.0310
4	bound method DataFrame.info of	Standard Class	Consumer	United States	Fort Lauderdale	Florida	Florida	33311	South	Office Supplies	Storage	22.3680	2	0.20	2.5164
...
9989	bound method DataFrame.info of	Second Class	Consumer	United States	Miami	Florida	Florida	33180	South	Furniture	Furnishings	25.2480	3	0.20	4.1028
9990	bound method DataFrame.info of	Standard Class	Consumer	United States	Costa Mesa	California	California	92627	West	Furniture	Furnishings	91.9600	2	0.00	15.6332
9991	bound method DataFrame.info of	Standard Class	Consumer	United States	Costa Mesa	California	California	92627	West	Technology	Phones	258.5760	2	0.20	19.3932
9992	bound method DataFrame.info of	Standard Class	Consumer	United States	Costa Mesa	California	California	92627	West	Office Supplies	Paper	29.6000	4	0.00	13.3200
9993	bound method DataFrame.info of	Second Class	Consumer	United States	Westminster	California	California	92683	West	Office Supplies	Appliances	243.1600	2	0.00	72.9480
...
0	bound method DataFrame.info of	Postal Code	Region	Category	Sub-Category	Sales	Quantity	Discount	Profit
1	bound method DataFrame.info of	42420	South	Furniture	Bookcases	261.9600	2	0.00	41.9136
2	bound method DataFrame.info of	42420	South	Furniture	Chairs	731.9400	3	0.00	219.5820
3	bound method DataFrame.info of	90036	West	Office Supplies	Labels	14.6200	2	0.00	6.8714
4	bound method DataFrame.info of	33311	South	Furniture	Tables	957.5775	5	0.45	-383.0310
...
9989	bound method DataFrame.info of	33180	South	Furniture	Furnishings	25.2480	3	0.20	4.1028
9990	bound method DataFrame.info of	92627	West	Furniture	Furnishings	91.9600	2	0.00	15.6332
9991	bound method DataFrame.info of	92627	West	Technology	Phones	258.5760	2	0.20	19.3932
9992	bound method DataFrame.info of	92627	West	Office Supplies	Paper	29.6000	4	0.00	13.3200
9993	bound method DataFrame.info of	92683	West	Office Supplies	Appliances	243.1600	2	0.00	72.9480
...
9994	bound method DataFrame.info of	rows	x	13	columns	>

```
In [11]: file.shape
```

```
Out[11]: (9994, 13)
```

```
In [13]: file.describe()
```

Out[13]:	Postal Code	Sales	Quantity	Discount	Profit
count	9994.000000	9994.000000	9994.000000	9994.000000	9994.000000
mean	55190.379428	229.858001	3.789574	0.156203	28.656896
std	32063.693350	623.245101	2.225110	0.206452	234.260108
min	1040.000000	0.444000	1.000000	0.000000	-6599.978000
25%	23223.000000	17.280000	2.000000	0.000000	1.728750
50%	56430.500000	54.490000	3.000000	0.200000	8.666500
75%	90008.000000	209.940000	5.000000	0.200000	29.364000
max	99301.000000	22638.480000	14.000000	0.800000	8399.976000

```
In [16]: file.duplicated().sum()
```

```
Out[16]: 17
```

```
In [19]: file.nunique().sum()
```

```
Out[19]: 14381
```

```
In [21]: file.drop_duplicates()
```

Out[21]:	Ship Mode	Segment	Country	City	State	Postal Code	Region	Category	Sub-Category	Sales	Quantity	Discount	Profit
0	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Bookcases	261.9600	2	0.00	41.9136
1	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Chairs	731.9400	3	0.00	219.5820
2	Second Class	Corporate	United States	Los Angeles	California	90036	West	Office Supplies	Labels	14.6200	2	0.00	6.8714
3	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Furniture	Tables	957.5775	5	0.45	-383.0310
4	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Office Supplies	Storage	22.3680	2	0.20	2.5164
...
9989	Second Class	Consumer	United States	Miami	Florida	33180	South	Furniture	Furnishings	25.2480	3	0.20	4.1028
9990	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Furniture	Furnishings	91.9600	2	0.00	15.6332
9991	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Technology	Phones	258.5760	2	0.20	19.3932
9992	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Office Supplies	Paper	29.6000	4	0.00	13.3200
9993	Second Class	Consumer	United States	Westminster	California	92683	West	Office Supplies	Appliances	243.1600	2	0.00	72.9480

9973 rows x 13 columns

```
In [23]: useless_col=['Postal Code']
file=file.drop(columns=useless_col,axis=1)
```

```
In [25]: file.head(10)
```

Out[25]:	Ship Mode	Segment	Country	City	State	Region	Category	Sub-Category	Sales	Quantity	Discount	Profit
0	Second Class	Consumer	United States	Henderson	Kentucky	South	Furniture	Bookcases	261.9600	2	0.00	41.9136
1	Second Class	Consumer	United States	Henderson	Kentucky	South	Furniture	Chairs	731.9400	3	0.00	219.5820
2	Standard Class	Corporate	United States	Los Angeles	California	West	Office Supplies	Labels	14.6200	2	0.00	6.8714
3	Standard Class	Consumer	United States	Fort Lauderdale	Florida	South	Office Supplies	Storage	957.5775	5	0.45	-383.0310
4	Standard Class	Consumer	United States	Fort Lauderdale	Florida	South	Office Supplies	Storage	22.3680	2	0.20	2.5164
5	Standard Class	Consumer	United States	Los Angeles	California	West	Furnishings	Appliances	48.6600	7	0.00	14.1694
6	Standard Class	Consumer	United States	Los Angeles	California	West	Office Supplies	Art	7.2800	4	0.00	1.9656
7	Standard Class	Consumer	United States	Los Angeles	California	West	Technology	Phones	907.1520	6	0.20	90.7152
8	Standard Class	Consumer	United States	Los Angeles	California	West	Binders	Binders	18.5040	3	0.20	5.7825
9	Standard Class	Consumer	United States	Los Angeles	California	West	Office Supplies	Appliances	114.9000	5	0.00	34.4700

```
In [27]: file.nunique()
```

Out[27]:	Ship Mode	Segment	Country	City	State	Region	Category	Sub-Category	Sales	Quantity	Discount	Profit
0	Second Class	Consumer	United States	Henderson	Kentucky	South	Furniture	Bookcases	261.9600	2	0.00	41.9136
1	Second Class	Consumer	United States	Henderson	Kentucky	South	Furniture	Chairs	731.9400	3	0.00	219.5820
2	Standard Class	Corporate	United States	Los Angeles	California	West	Office Supplies	Labels	14.6200	2	0.00	6.8714
3	Standard Class	Consumer	United States	Fort Lauderdale	Florida	South	Office Supplies	Storage	957.5775	5	0.45	-383.0310
4	Standard Class	Consumer	United States	Fort Lauderdale	Florida	South	Office Supplies	Storage	22.3680	2	0.20	2.5164
5	Standard Class	Consumer	United States	Los Angeles	California	West	Furnishings	Appliances	48.6600	7	0.00	14.1694
6	Standard Class	Consumer	United States	Los Angeles	California	West	Office Supplies	Art	7.2800	4	0.00	1.9656
7	Standard Class	Consumer	United States	Los Angeles	California	West	Technology	Phones	907.1520	6	0.20	90.7152
8	Standard Class	Consumer	United States	Los Angeles	California	West	Binders	Binders	18.5040	3	0.20	5.7825
9	Standard Class	Consumer	United States	Los Angeles	California	West	Office Supplies	Appliances	114.9000	5	0.00	34.4700

```
In [29]: file.describe()
```

Out[29]:	Sales	Quantity	Discount	Profit
count	9994.000000	9994.000000	9994.000000	9994.000000
mean	229.858001	3.789574	0.156203	28.656896
std	623.245101	2.225110	0.206452	234.260108
min	0.444000	1.000000	0.000000	-6599.978000
25%	17.280000	2.000000	0.000000	1.728750
50%	54.490000	3.000000	0.200000	8.666500
75%	209.940000	5.000000	0.200000	29.364000
max	22638.480000	14.000000	0.800000	8399.976000

Visualization of Data

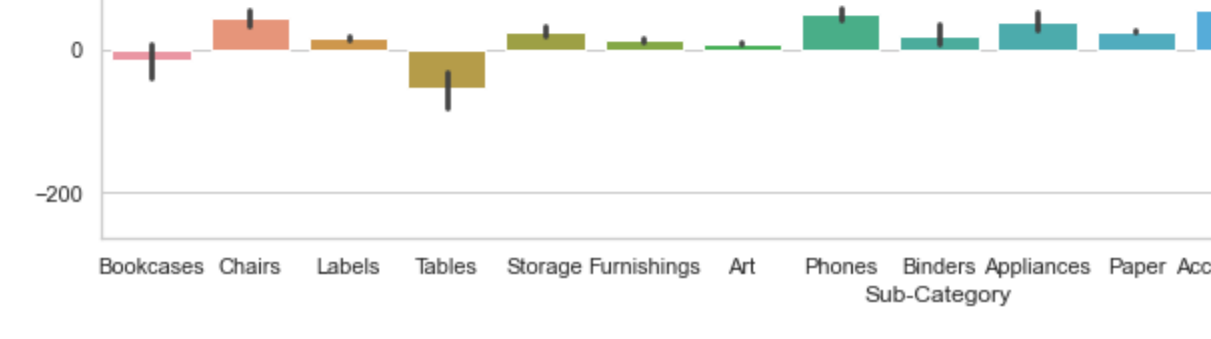
```
In [43]: plt.figure(figsize=(20,10))
plt.bar(file['Sub-Category'],file['Category'])
plt.xlabel('Sub-Cat',size=50,color='red')
plt.ylabel('Cat',size=30,color='red')
plt.title('Sub-Cat VS Cat',size=30,color='red')
plt.xticks(rotation=90)
plt.grid(color='green')
plt.show()
```



```
In [34]: file['Region'].value_counts()
```

Region	count
West	3203
East	2848
Central	2323
South	1628

```
In [42]: plt.figure(figsize=(10,5))
sns.countplot(x=file['Region'])
plt.title('Region',size=20)
plt.xticks(rotation=40)
plt.grid(color='red')
plt.show()
```



Analysing using barplot

```
In [45]: sns.set(style='whitegrid')
plt.figure(figsize=(15,10))
sns.barplot(x=file['Sub-Category'],y=file['Profit'])
plt.title('Consumption')
plt.show()
```