

```
In [2]: greetings ="Assalam-o-Alaikum!"
print(greetings)
```

Assalam-o-Alaikum!

```
In [3]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

import Dataset

```
In [4]: df = pd.read_csv("SampleSuperstore.csv")
df
```

```
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 × 13 columns

What is the total profit generated for the "Office Supplies" category?

```
In [5]: office_supplies_df = df[df["Category"] == "Office Supplies"]
total_profit_supplies = office_supplies_df["Profit"].sum().round(2)
print("Total Profit generated for Office Supplies = $" + str(total_profit_supplies))
```

Total Profit generated for Office Supplies = \$122490.8

Which sub-category has the highest average sales? What is the average sales amount?

```
In [6]: highest_avg_sale = df.groupby(["Sub-Category"])["Sales"].agg("mean").to_frame().reset_index().sort_values("Sales")
highest_avg_sale
```

Out[6]:

	Sub-Category	Sales
6	Copiers	2198.941618
11	Machines	1645.553313
16	Tables	648.794771
5	Chairs	532.332420
4	Bookcases	503.859633
13	Phones	371.211534
14	Storage	264.590553
15	Supplies	245.650200
1	Appliances	230.755710
0	Accessories	215.974604
3	Binders	133.560560
9	Furnishings	95.825668
7	Envelopes	64.867724
12	Paper	57.284092
10	Labels	34.303055
2	Art	34.068834
8	Fasteners	13.936774

In [7]:

```
high_sub_category_avg_sales = highest_avg_sale.iloc[0]
print("Sub-Category" , "'" + high_sub_category_avg_sales["Sub-Category"] + "'" , "Has Highest Average Sale = $"
Sub-Category 'Copiers' Has Highest Average Sale = $2198.94
```

In [8]:

```
df.head(1)
```

Out[8]:

	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	261.96	2	0.0	41.9136

3. In which region does the "Technology" category have the highest sales?

In [9]:

```
category_technology = df[df["Category"] == "Technology"]
reigon_sales = category_technology.groupby(["Region"])[["Sales"]].agg("sum").astype("int").to_frame().reset_index
reigon_sales
```

Out[9]:

	Region	Sales
1	East	264973
3	West	251991
0	Central	170416
2	South	148771

In [10]:

```
highest_sale = reigon_sales.iloc[0]
print("Region" , "'" + highest_sale["Region"] + "'" , "has High Sale by Technology category = $" + str(highest_sale["Sales"])
Region 'East' has High Sale by Technology category = $264973
```

4. What is the total quantity sold for the "Furniture" category in the "West" region?

In [11]:

```
category_region = df[(df["Category"] == "Furniture") & (df["Region"] == "West")]
category_region
```

Out[11]:

	Ship Mode	Segment	Country	City	State	Postal Code	Region	Category	Sub-Category	Sales	Quantity	Discount	Profit
5	Standard Class	Consumer	United States	Los Angeles	California	90032	West	Furniture	Furnishings	48.860	7	0.00	14.1694
10	Standard Class	Consumer	United States	Los Angeles	California	90032	West	Furniture	Tables	1706.184	9	0.20	85.3092
24	Standard Class	Consumer	United States	Orem	Utah	84057	West	Furniture	Tables	1044.630	3	0.00	240.2649
65	Standard Class	Consumer	United States	Los Angeles	California	90004	West	Furniture	Furnishings	79.760	4	0.00	22.3328
104	Standard Class	Consumer	United States	Aurora	Colorado	80013	West	Furniture	Furnishings	102.360	3	0.20	-3.8385
...
9928	Same Day	Consumer	United States	San Francisco	California	94122	West	Furniture	Furnishings	24.270	3	0.00	8.7372
9931	Standard Class	Consumer	United States	San Bernardino	California	92404	West	Furniture	Bookcases	683.332	4	0.15	-40.1960
9937	Second Class	Corporate	United States	Los Angeles	California	90049	West	Furniture	Tables	71.088	2	0.20	-1.7772
9955	Second Class	Corporate	United States	Los Angeles	California	90045	West	Furniture	Tables	273.568	2	0.20	10.2588
9990	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Furniture	Furnishings	91.960	2	0.00	15.6332

707 rows × 13 columns

In [12]:

```
Total_Quantity_Sold =category_region["Quantity"].sum()
print("Total Quantity Sold for Furniture In West = " + str(int(Total_Quantity_Sold)))

Total Quantity Sold for Furniture In West = 2696
```

5. How many unique cities are there in the dataset?

In [13]:

```
unique_cities =df["City"].nunique()
print("Total No.Of Unique Cities in the Dataset = " + str(int(unique_cities)))

Total No.Of Unique Cities in the Dataset = 531
```

6. Calculate the average discount for the "Consumer" segment.

In [14]:

```
consumer_segment = df[df["Segment"] == "Consumer"]
consumer_segment
```

Out[14]:

	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
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
5	Standard Class	Consumer	United States	Los Angeles	California	90032	West	Furniture	Furnishings	48.8600	7	0.00	14.1694
...
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

5191 rows × 13 columns

In [15]:

```
avg_discount_consumer =consumer_segment["Discount"].mean().round(2)
print("Average Discount For Consumer segment = " + str(float(avg_discount_consumer)))
```

Average Discount For Consumer segment = 0.16

7. Which state has the highest total sales? What is the total sales amount?

```
In [16]: state_highest_sale = df.groupby("State")["Sales"].agg("sum").to_frame().reset_index().sort_values("Sales", ascen
state_highest_sale
```

Out[16]:

	State	Sales
3	California	457687.63
30	New York	310876.27
41	Texas	170188.05
45	Washington	138641.27
36	Pennsylvania	116511.91
8	Florida	89473.71
11	Illinois	80166.10
33	Ohio	78258.14
20	Michigan	76269.61
44	Virginia	70636.72
31	North Carolina	55603.16
12	Indiana	53555.36
9	Georgia	49095.84
15	Kentucky	36591.75
28	New Jersey	35764.31
1	Arizona	35282.00
47	Wisconsin	32114.61
4	Colorado	32108.12
40	Tennessee	30661.87
21	Minnesota	29863.15
19	Massachusetts	28634.43
6	Delaware	27451.07
18	Maryland	23705.52
37	Rhode Island	22627.96
23	Missouri	22205.15
34	Oklahoma	19683.39
0	Alabama	19510.64
35	Oregon	17431.15
26	Nevada	16729.10
5	Connecticut	13384.36
2	Arkansas	11678.13
42	Utah	11220.06
22	Mississippi	10771.34
16	Louisiana	9217.03
43	Vermont	8929.37
38	South Carolina	8481.71
25	Nebraska	7464.93
27	New Hampshire	7292.52
24	Montana	5589.35
29	New Mexico	4783.52
13	Iowa	4579.76
10	Idaho	4382.49
14	Kansas	2914.31
7	District of Columbia	2865.02
48	Wyoming	1603.14
39	South Dakota	1315.56
17	Maine	1270.53
46	West Virginia	1209.82
32	North Dakota	919.91

In [17]:

```
total_highest_sale_of_state = state_highest_sale.iloc[0]
print("State", "" + total_highest_sale_of_state["State"]+ "", "has highest sale = $" + str(total_highest_sale
State 'California' has highest sale = $457687.63
```

Q Identify the top 5 cities with the highest sales

9. Identify the top 3 cities with the highest sales.

```
In [18]: Highest_sales = df.groupby("City")["Sales"].agg(sum).to_frame().astype(int).reset_index().sort_values("Sales",
Highest_sales
```

```
Out[18]:
```

	City	Sales
329	New York City	256368
266	Los Angeles	175851
452	Seattle	119540
438	San Francisco	112669
374	Philadelphia	109077

10. What is the correlation coefficient between the "Quantity" and "Profit" columns?

```
In [19]: correlation = df[["Quantity", "Profit"]].corr()
correlation
```

```
Out[19]:
```

	Quantity	Profit
Quantity	1.000000	0.066253
Profit	0.066253	1.000000

11. How many transactions had a negative profit?

```
In [27]: negative_profit = len(df[df["Profit"] < 0])
print("Total transactions had a negative profit =", str(negative_profit))
```

Total transactions had a negative profit = 1871

12. What is the average profit for the "Technology" category in the "South" region?

```
In [31]: df.head(2)
```

```
Out[31]:
```

	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	261.96	2	0.0	41.9136
1	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Chairs	731.94	3	0.0	219.5820

```
In [43]: average_profit = df[(df["Category"] == "Technology") & (df["Region"] == "South")]
profit = average_profit["Profit"].agg("mean").round(2)
print("Total Average Profit for the 'Technology' in 'South' = $" + str(profit))
```

Total Average Profit for the 'Technology' in 'South' = \$68.23

13. Identify the top 3 sub-categories with the highest quantity sold in the "Standard Class" shipping mode.

```
In [56]: top_categories = df[(df["Ship Mode"] == "Standard Class")]
top_categories.head(1)
```

```
Out[56]:
```

	Ship Mode	Segment	Country	City	State	Postal Code	Region	Category	Sub-Category	Sales	Quantity	Discount	Profit
3	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Furniture	Tables	957.5775	5	0.45	-383.031

```
In [55]: quantity_category = top_categories.groupby("Sub-Category")["Quantity"].agg("sum").to_frame().reset_index().sort
quantity_category
```

```
Out[55]:
```

	Sub-Category	Quantity
3	Binders	3777
12	Paper	2984
9	Furnishings	2186

14. What is the average discount for transactions with a profit less than 0?

```
In [76]: average_discount =(df[df["Profit"] < 0])
Aveg_Discount = average_discount["Discount"].mean().round(2)
print("Total Average Discount for Transactions with profit less then 0 = $" + str(Aveg_Discount))
```

Total Average Discount for Transactions with profit less then 0 = \$0.48

15. Create a bar chart showing the sales amount for each category.

```
In [84]: plt.figure(figsize =(18, 7))
graph =sns.barplot(x ="Category", y ="Sales", data =df, palette ="muted", errorbar = None)
for p in graph.patches:
    graph.annotate('{:.0f}'.format(p.get_height()),
                    (p.get_x()+0.41, p.get_height()),
                    ha='center', va='bottom',color= 'black', size = 15)
plt.title("The Sales Amount For Each Category", weight ="bold", size =15)
plt.xticks(size =12)
plt.yticks(size =12)
plt.xlabel("Category", weight ="bold", size =15)
plt.ylabel("Sales", weight ="bold", size =15)
plt.show()
```

