THE SPARKS FOUNDATION

DATA SCIENCE AND BUSINESS ANALYTICS

AUTHOR: MEGHA

TASK 3:Exploratory Data Analysis - Retail

Objective:To find out the weak areas where you can work to make more profit and to derive all the business problems by exploring the data

In [1]:

```
import pandas as pd
import seaborn as sns
import numpy as np
import matplotlib.pyplot as plt
```

In [6]:

```
Dataset=pd.read_csv("SampleSuperstore.csv")
```

In [7]:

```
Dataset
```

Out[7]:

	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
...	...	...	...	...	...	...	...	...	...	...	...	...	...

	Ship Mode	Segment	Country	City	State	Postal Code	Region	Category	Sub-Category	Sales	Quantity	Discount	Profit
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

Regarding the attributes included in our data sets are:

- Ship mode: It is a method of shipping physical good and products to the customers
- Segment: It specifies the customer segment or marketing groups according to their interests and demographics
- Country: All the deliveries are made in the USA
- City: Tells us that in what cities deliveries are been made.
- State: Tells us that in what states the deliveries are made.
- Postal Code: denotes the postal destination address sorted by the regions in USA.
- Region: The geographic in the USA continent
- Category: It divides the whole delivered products into three parts namely- Office Supplies, Technology and Furniture.
- Sub-Category: Subdivision with respect to Categories.
- Sales: are the monetary value of transactions between the Superstore and its customers of physical goods in American Dollars (USD$)measurements
- Quantity: number of product in each transaction
- Profit: It's the profit made on each sale

In [8]:
```
Dataset.drop_duplicates(inplace=True)
Dataset.head()
```

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	261.9600	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

In [9]:

```
Dataset_new=Dataset.drop(columns="Postal Code",axis=1)
Dataset_new
```

Out[9]:

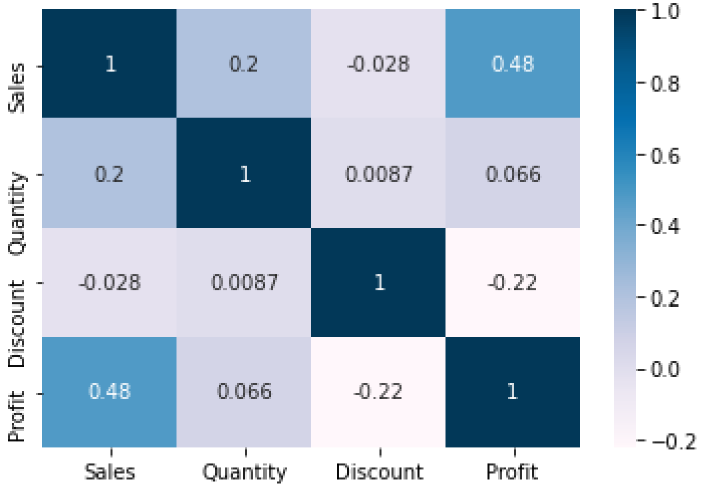
	Ship Mode	Segment	Country	City	State	Region	Category	Sub-Category	Sales	Quantity	Discount	Profit
0	Second Class	Consumer	United States	Henderson	Kentucky	South	Furniture	Bookcases	261.9600	2	0.00	41.9136
1	Second Class	Consumer	United States	Henderson	Kentucky	South	Furniture	Chairs	731.9400	3	0.00	219.5820
2	Second 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	Furniture	Tables	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
...	...	...	...	...	...	...	...	...	...	...	...	...
9989	Second Class	Consumer	United States	Miami	Florida	South	Furniture	Furnishings	25.2480	3	0.20	4.1028
9990	Standard Class	Consumer	United States	Costa Mesa	California	West	Furniture	Furnishings	91.9600	2	0.00	15.6332

	Ship Mode	Segment	Country	City	State	Region	Category	Sub-Category	Sales	Quantity	Discount	Profit
9991	Standard Class	Consumer	United States	Costa Mesa	California	West	Technology	Phones	258.5760	2	0.20	19.3932
9992	Standard Class	Consumer	United States	Costa Mesa	California	West	Office Supplies	Paper	29.6000	4	0.00	13.3200
9993	Second Class	Consumer	United States	Westminster	California	West	Office Supplies	Appliances	243.1600	2	0.00	72.9480

9977 rows × 12 columns

In [10]:

```
sns.heatmap(Dataset_new.corr(), cmap = 'PuBu', annot = True)
plt.show()
```



Sales and Profit have a moderate positive linear correlation. When Sales increase by USD$1, Profit will increase by USD$0.48.This shows that the management have a good ability to manage costs and get a good profit on each item.

Discount and Profit have a weak negative linear correlation. When Discount increase by 1%, Profit will decrease by USD$0.22.

Quantity and Profit have little to no linear correlation. This shows us that the Superstore may sell a considerable number of products at low profit margins. This could be due to the fact the the company uses the promotional strategies such as providing higher discounts

on larger quantities of products in a single transaction, and Buy One Get One Free promotions. As a result, average profit is lower. Another likely explanation is that Profit might increase as Quantity increases, but not on a linear scale.

Discount and Sales have little no linear correlation.This shows that even discounts do not attract the consumers.This could be due to the fact that most of the consumers might think that higher discount are provided because of the bad quality of the products and many people tend to choose qulatity over anything.

Discount and Quantity have little to no linear correlation. This is indicative that implementing discounts as Superstore's promotional strategy is not attractive or motivating. As discussed in the previous point, this may result in bad consumer psychology. Superstore is advised to consider other promotional strategies

In [11]:

```
Dataset_new.describe()
```

Out[11]:

	Sales	Quantity	Discount	Profit
count	9977.000000	9977.000000	9977.000000	9977.00000
mean	230.148902	3.790719	0.156278	28.69013
std	623.721409	2.226657	0.206455	234.45784
min	0.444000	1.000000	0.000000	-6599.97800
25%	17.300000	2.000000	0.000000	1.72620
50%	54.816000	3.000000	0.200000	8.67100
75%	209.970000	5.000000	0.200000	29.37200
max	22638.480000	14.000000	0.800000	8399.97600

In [12]:

```
round(sum(Dataset_new['Sales']),2)
```

Out[12]:

```
2296195.59
```

In [13]:

```
sum(Dataset_new['Quantity'])
```

Out[13]:

```
37820
```

In [14]:

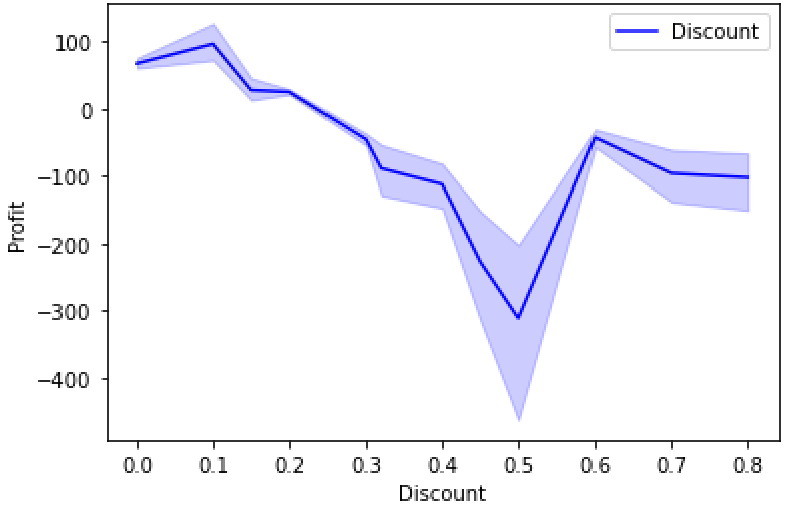
```
sum(Dataset_new['Profit'])
```

```
Out[14]: 286241.4226000013
```

- From the above data description of data it can be noted that a total of 2296195 sale has been made where the minimum sale was of 0.444 and the maximum sale was of 22638.48 and an average of 230.14 was made on each item.
- The total of 37,820 products were sold.The quantity varied from 1 to 14.
- The total profit of 286241.42 was made by the Superstore.On an average a profit of 28.69 was made on each item.However it can range from a loss of 6599.97 to a profit of 8399.97. As can be seen from the correlation table that profit and discount have a negative correlation and hence losses can be due to higher discounts.

```
In [15]: sns.lineplot('Discount','Profit', data=Dataset_new , color='b',label='Discount')
         plt.legend()
         plt.show()
```

```
C:\Users\ASUS\anaconda3\anaconda\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following varia
bles as keyword args: x, y. From version 0.12, the only valid positional argument will be `data`, and passing other a
rguments without an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```



- The following line chart as well as the above correlation table tells us that Profit and Discount have a negative relationship. It means that if one increases the other decreases.So in order to make more profit the company can lower down the discounts.

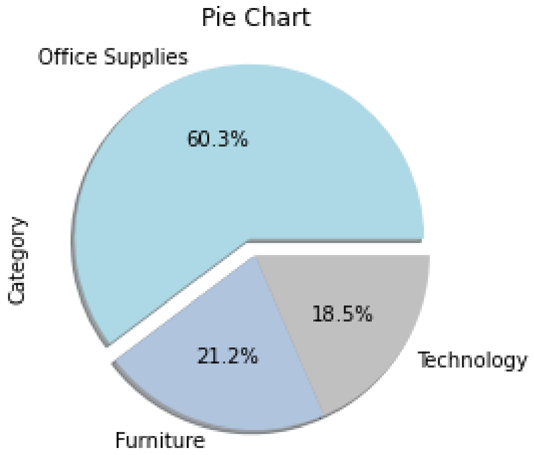
- **Losses are likely to occur for higher discount levels above 20%. In other words, between 0% and 20% of discounts, a profit can be made. As discussed before, higher discounts also encouraged bad consumer psychology, instead of targeting to implement discounts as a promotional strategy to successfully increase sales. A considerable number of customers may develop the thinking that Superstore's products are defective or low quality. In a long run, this can negatively affect Superstore's brand image. All in all, Superstore is advised to consider other promotional strategies.**

- **The worst losses occured when discount approximates 50%. Such discounts might be because of festivals, end-of-season sales, and in order to clean the store for new upcoming items.**

Super store earns the most profit when the discounts are less than 10%.

In [18]:

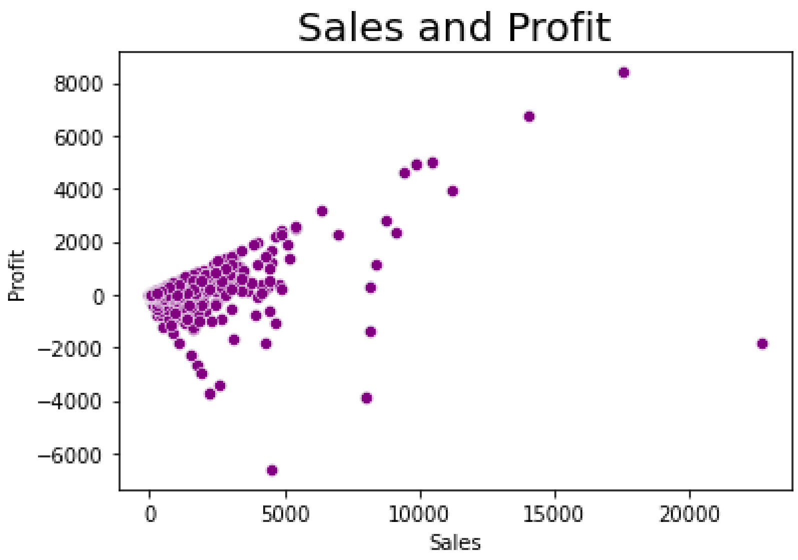
```
my_colors = ['lightblue','lightsteelblue','silver']
my_explode = (0.1, 0, 0)
Dataset_new['Category'].value_counts().plot.pie(autopct="%1.1f%%",shadow=True,colors=my_colors,explode=my_explode)
plt.title('Pie Chart')
plt.show()
```



- **The above pie-chat shows that 60.3% of the total sale was of the office supplies**
- **21.2% was furniture**
- **18.5% was Technology**

In [19]:

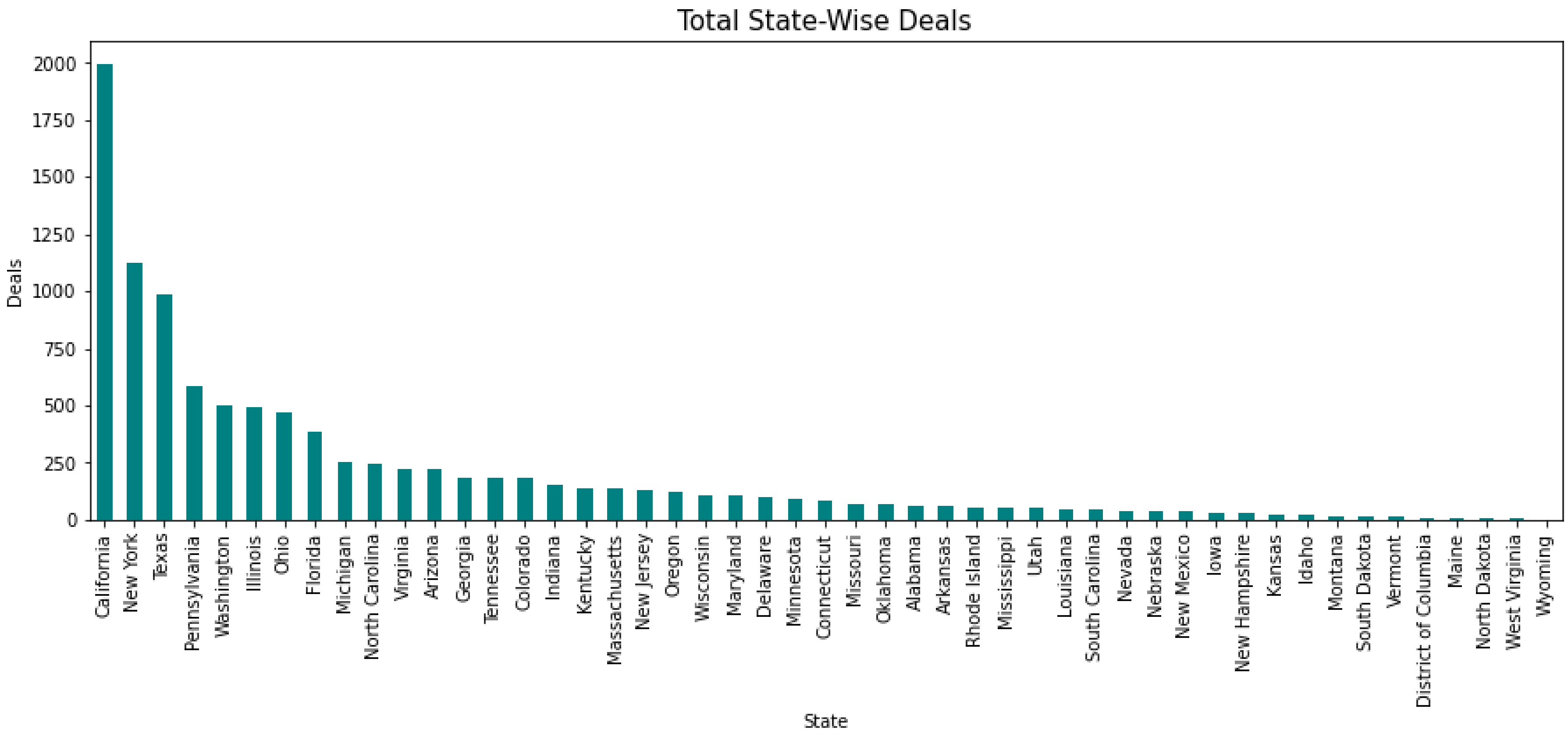
```
sns.scatterplot(x=Dataset_new["Sales"] ,y= Dataset_new["Profit"], color = 'Purple')
plt.title('Sales and Profit', fontsize = 20)
plt.show()
```



From the above scatter plot it can be interpreted that most of the sales were under 5000. It can also be noted that most of the losses were under 2.5k a possible reason for this can be higher discounts which results in price reduction to an extent that it causes losses.

In [20]:

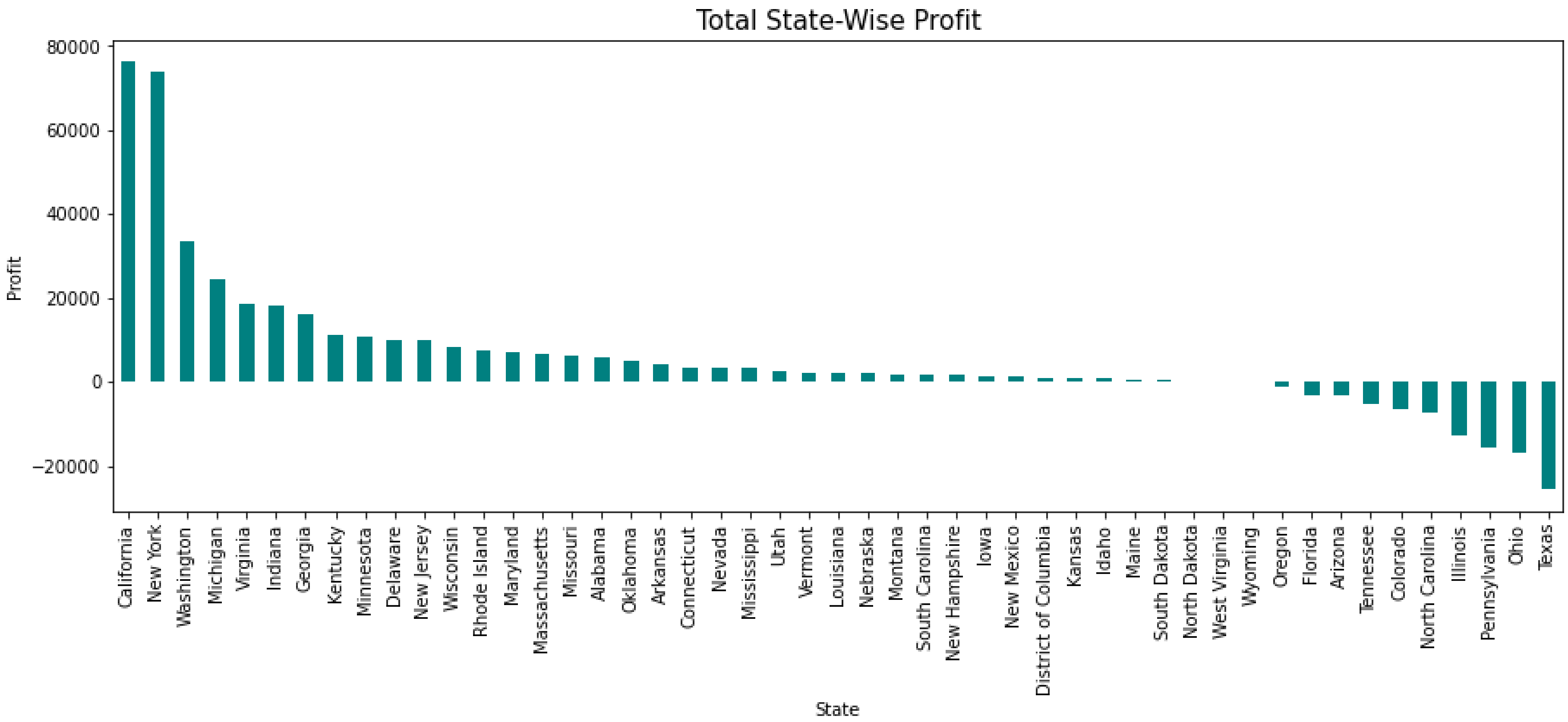
```
states = Dataset_new.groupby('State')
states['Quantity'].count().sort_values(ascending = False).plot.bar(figsize = (15, 5), color = 'Teal')
plt.ylabel('Deals')
plt.title('Total State-Wise Deals', fontsize = 15)
plt.show()
```



The states like that of California,New York and Texas have the highest number of dealings whereas the states Wyoming, West Virginia, North Dakota, Maine, District of Columbia have very less dealings. So some promotional strategies must be looked upon in order to increase the number of dealings here.This can help the company to increase it's sales and hence the profit.

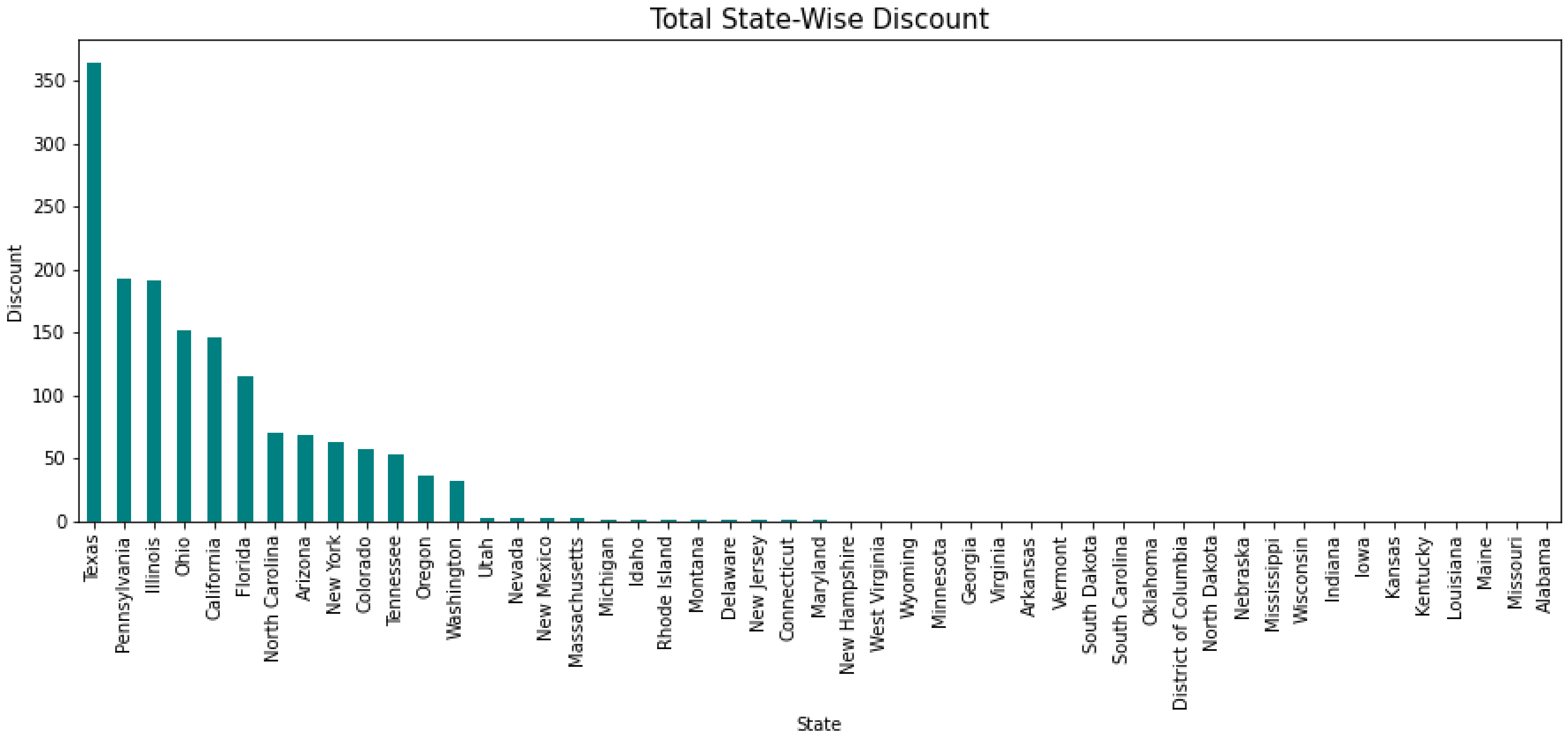
In [21]:

```
states = Dataset_new.groupby('State')
states['Profit'].sum().sort_values(ascending = False).plot.bar(figsize = (15, 5), color = 'Teal')
plt.ylabel('Profit')
plt.title('Total State-Wise Profit', fontsize = 15)
plt.show()
```



Superstore have earned a good amount of profit in the state of California. New York has the second highest profit earned and Washington is third. Most of the losses that can be seen are in the states of Texas, Ohio, Pennsylvania. This could be becaues of the higher discounts provided at these states.

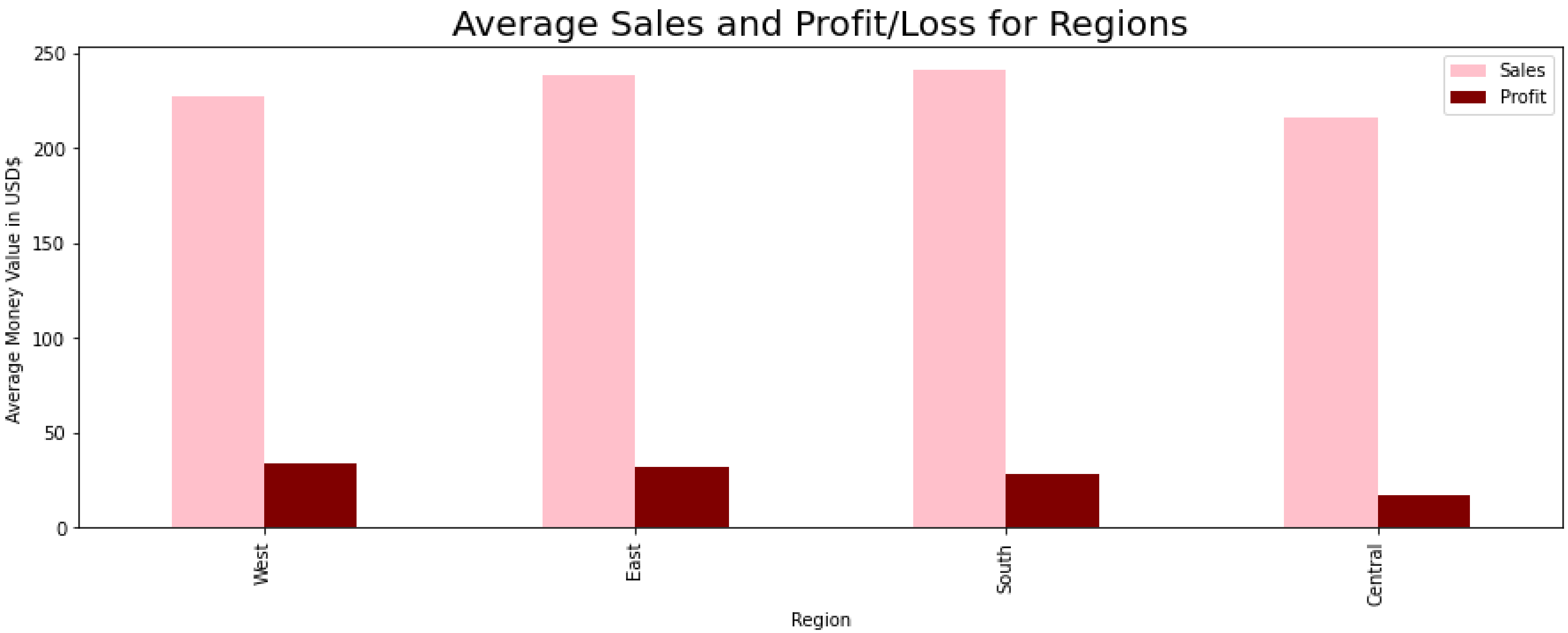
In [22]:
```
states = Dataset_new.groupby('State')
states['Discount'].sum().sort_values(ascending = False).plot.bar(figsize = (15, 5), color = 'Teal')
plt.ylabel('Discount')
plt.title('Total State-Wise Discount', fontsize = 15)
plt.show()
```



As can be seen from the two graphs that the states which were given more discounts suffered from higher losses but these discounts even managed to earn higher deals as well.This could be considered to be a good promotional strategy for the store in a long runas they can earn a good number of customers.

In [23]:

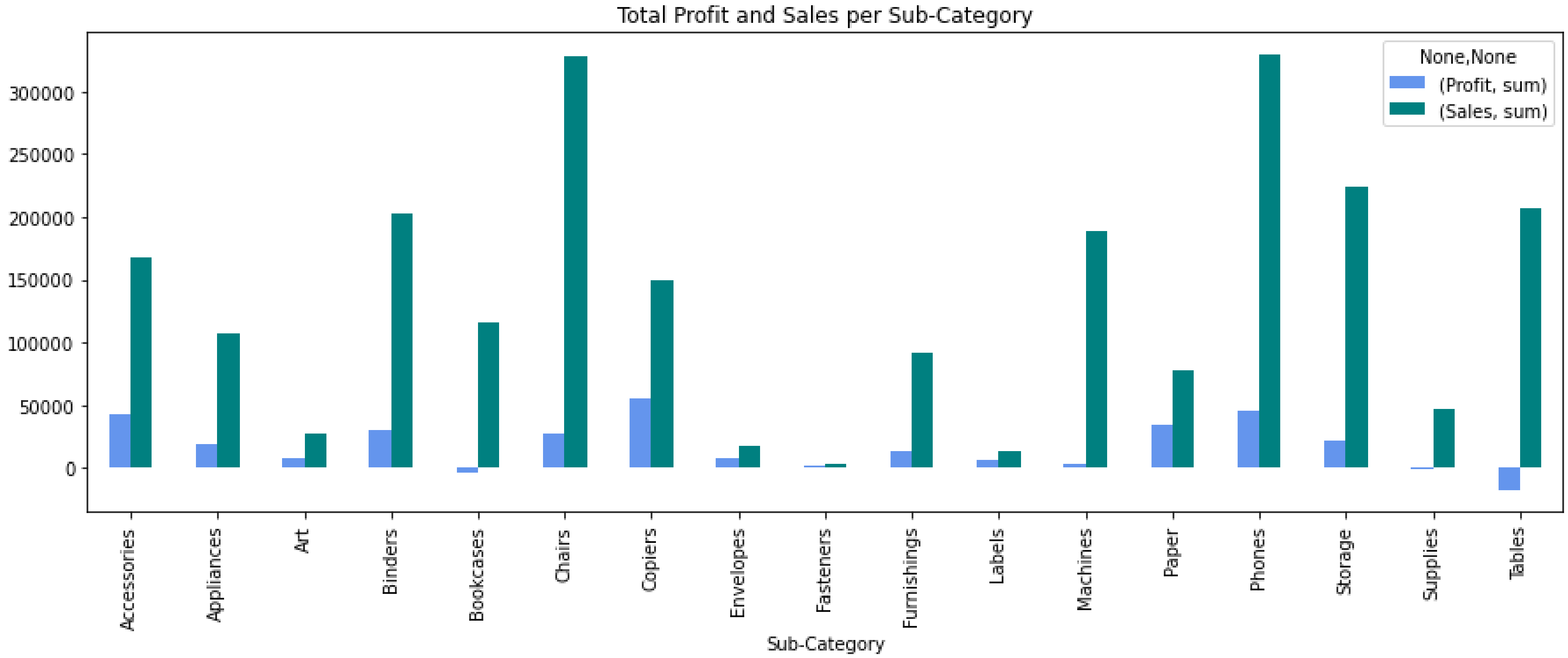
```
colors = ['Pink', 'Maroon']
region = Dataset_new.groupby(['Region'])[['Sales', 'Discount', 'Profit']].mean()
region.sort_values('Profit', ascending = False)[['Sales', 'Profit']].plot(kind = 'bar',figsize = (15, 5),color = colo
plt.ylabel('Average Money Value in USD$')
plt.xlabel('Region')
plt.title('Average Sales and Profit/Loss for Regions', fontsize = 20)
plt.show()
```



As could be clearly seen that region does not have much influence on sales and profit.

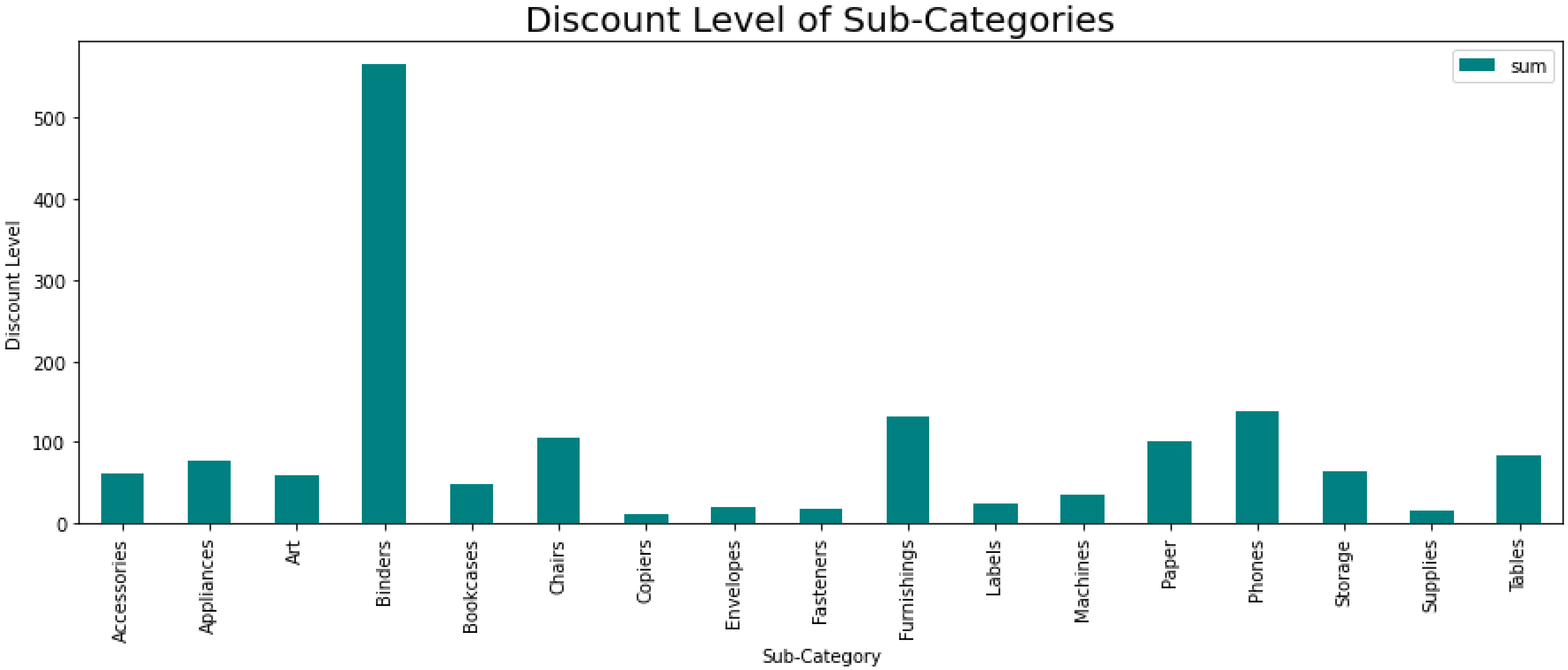
In [24]:
```
colors = ['Cornflowerblue', 'Teal']
Dataset_new.groupby('Sub-Category')['Profit','Sales'].agg(['sum']).plot.bar(figsize=(15,5),color=colors)
plt.title('Total Profit and Sales per Sub-Category')
plt.show();
```

```
<ipython-input-24-86a0f001d1fe>:2: FutureWarning: Indexing with multiple keys (implicitly converted to a tuple of key
s) will be deprecated, use a list instead.
  Dataset_new.groupby('Sub-Category')['Profit','Sales'].agg(['sum']).plot.bar(figsize=(15,5),color=colors)
```



In [25]:

```
Dataset_new.groupby('Sub-Category')['Discount'].agg(['sum']).plot.bar(figsize=(15,5),color='Teal')
plt.ylabel('Discount Level')
plt.xlabel('Sub-Category')
plt.title("Discount Level of Sub-Categories", fontsize = 20)
plt.show()
```

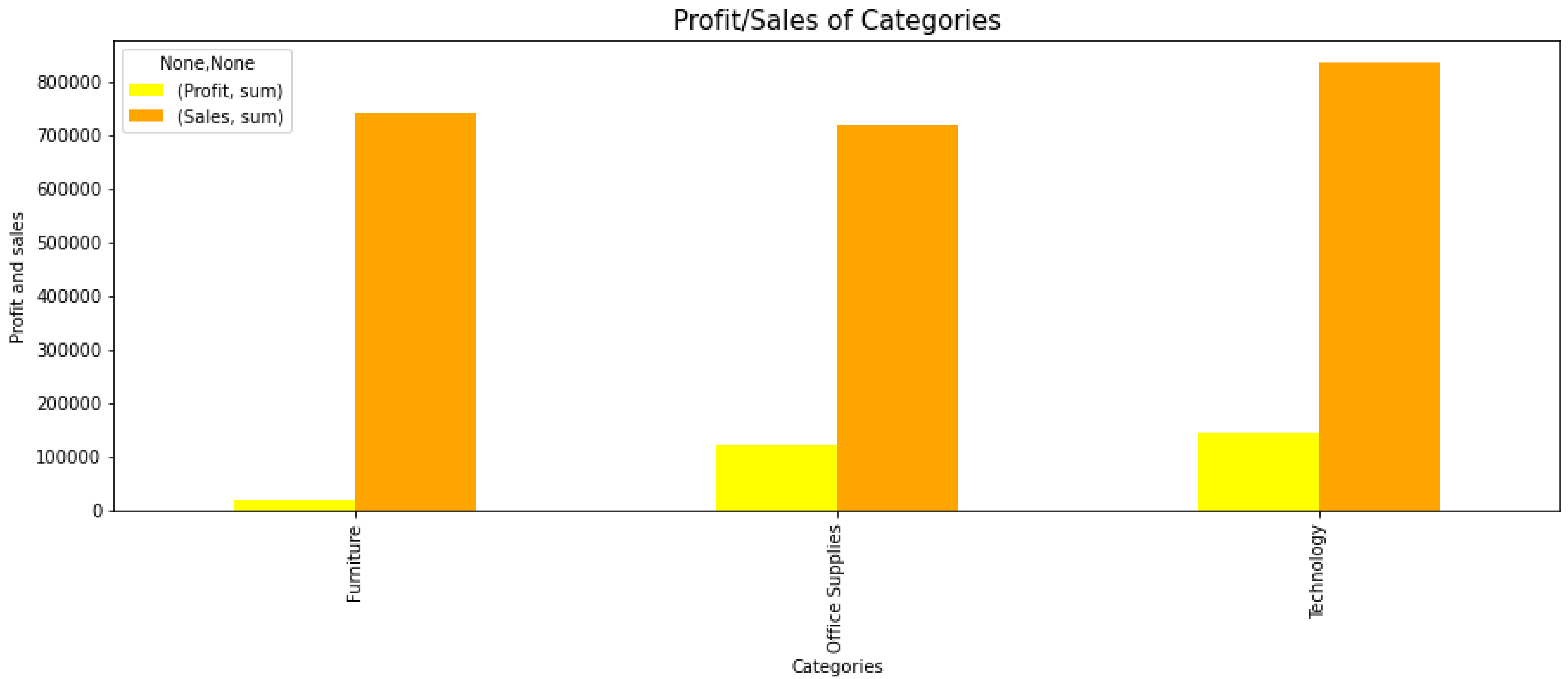


- **From the following graphs it can be seen that although the sales of Bookcases and Tables were decent but the resulted in losses. This could be because of the discounts provided on these items.**
- **The highest discounts were provided on the Binders followed by Furnishings and phones and this could be the reason that despite the higher sales of Furnishings and phones the profits weren't that high.**

In [26]:

```
Dataset_new.groupby("Category")['Profit','Sales'].agg(['sum']).plot.bar(figsize=(15,5),color=('Yellow','Orange'))
plt.ylabel('Profit and sales')
plt.xlabel('Categories')
plt.title("Profit/Sales of Categories", fontsize = 15)
plt.show()
```

```
<ipython-input-26-1b3a84bb0871>:1: FutureWarning: Indexing with multiple keys (implicitly converted to a tuple of key
s) will be deprecated, use a list instead.
  Dataset_new.groupby("Category")['Profit','Sales'].agg(['sum']).plot.bar(figsize=(15,5),color=('Yellow','Orange'))
```

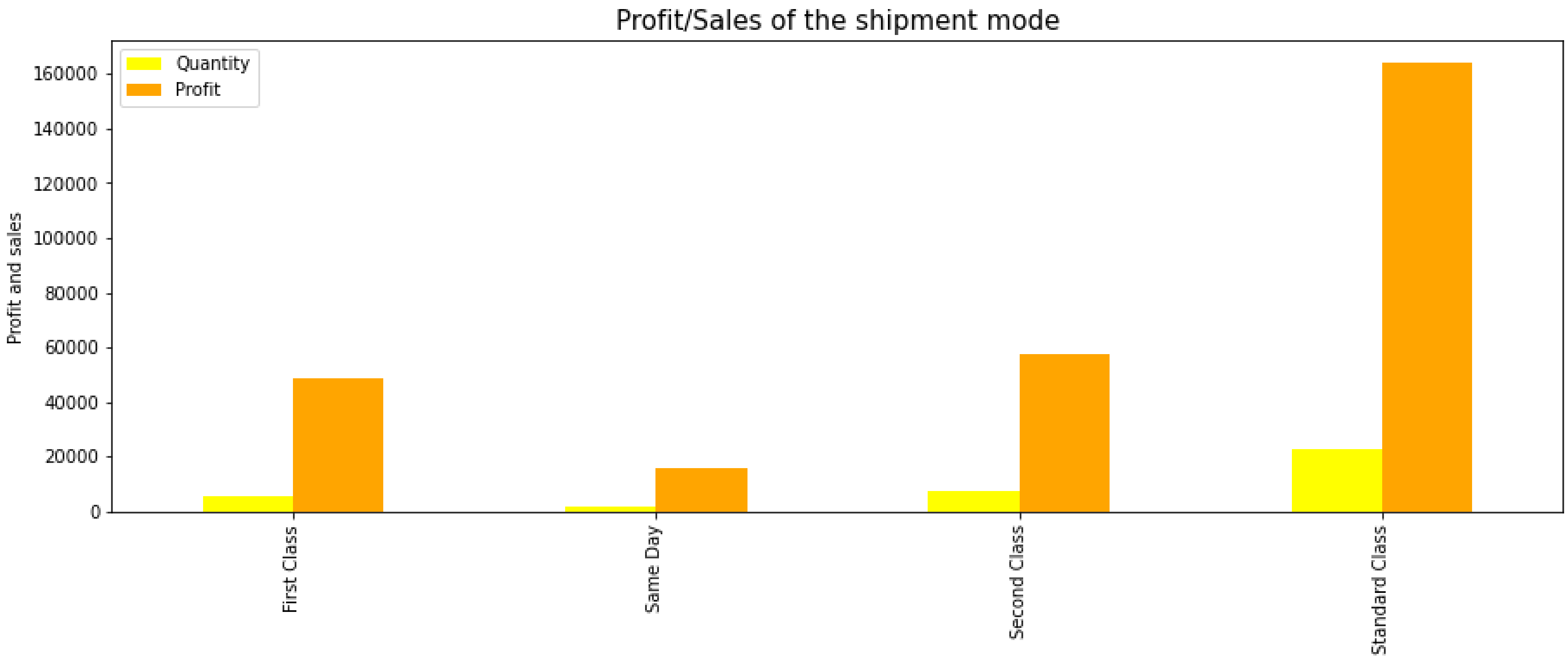


As it can be seen that all of the three categories have almost equal sales but the technology department managed to earn the maximum profit followes by office supplies. The least profit was earned by the Furniture category this could be because of the higher discounts and maybe due to more transportation cost.

In [27]:

```
Dataset_new.groupby('Ship Mode')["Quantity","Profit"].sum().plot.bar(figsize=(15,5),color=('Yellow','Orange'))
plt.ylabel('Profit and sales')
plt.xlabel('Shipment Mode')
plt.title("Profit/Sales of the shipment mode", fontsize = 15)
plt.show()
```

```
<ipython-input-27-0af2d9ec2df1>:1: FutureWarning: Indexing with multiple keys (implicitly converted to a tuple of key
s) will be deprecated, use a list instead.
  Dataset_new.groupby('Ship Mode')["Quantity","Profit"].sum().plot.bar(figsize=(15,5),color=('Yellow','Orange'))
```



CONCLUSION

The Superstore cannot stop from giving out discounts on their products, especially during festivals, end-of-season sales, and clearance sales that are necessary to make space in their warehouses for newer and more current stocks. Instead, It is recommended to reduce overall discount levels to minimise losses, and even focus on implementing some new promotional strategies to ensure increasing maximum profits and consistent future gains in long term customers.