Author : Harish Nakireddy

GRIP (Graduate Rotational Internship Program) - The Sparks Foundation

Exploratory Data Analysis - Retail

Task 3

In [2]:
```
#importing required libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

In [3]:
```
df= pd.read_csv("C:\\Users\\Asus\\Desktop\\SampleSuperstore.csv")
df.head()
```

Out[3]:

	Ship Mode	Segment	Country	City	State	Postal Code	Region	Category	Sub-Category	Sa
0	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Bookcases	261.9
1	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Chairs	731.9
2	Second Class	Corporate	United States	Los Angeles	California	90036	West	Office Supplies	Labels	14.6
3	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Furniture	Tables	957.5
4	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Office Supplies	Storage	22.3

In [4]:
```
df.shape
```

Out[4]: (9994, 13)

In [5]:
```
df.describe()
```

Out[5]:

	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 [6]:
```
df.isnull().sum()
```

Out[6]:
```
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 [7]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9994 entries, 0 to 9993
Data columns (total 13 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   Ship Mode     9994 non-null   object 
 1   Segment       9994 non-null   object 
 2   Country       9994 non-null   object 
 3   City          9994 non-null   object 
 4   State         9994 non-null   object 
 5   Postal Code   9994 non-null   int64  
 6   Region        9994 non-null   object 
 7   Category      9994 non-null   object 
 8   Sub-Category  9994 non-null   object 
 9   Sales         9994 non-null   float64
 10  Quantity      9994 non-null   int64  
 11  Discount      9994 non-null   float64
 12  Profit        9994 non-null   float64
dtypes: float64(3), int64(2), object(8)
memory usage: 1015.1+ KB
```

In [8]:
```
df.nunique()
```

Out[8]:
```
Ship Mode          4
Segment            3
Country            1
City             531
State             49
Postal Code      631
Region             4
Category           3
Sub-Category      17
Sales           5825
Quantity          14
Discount          12
Profit          7287
dtype: int64
```

Checking for duplicate data

In [9]:
```
df.duplicated().sum()
```

Out[9]: 17

In [10]:
```
#removing duplicate rows from the dataset

df.drop_duplicates(inplace = True)
df
```

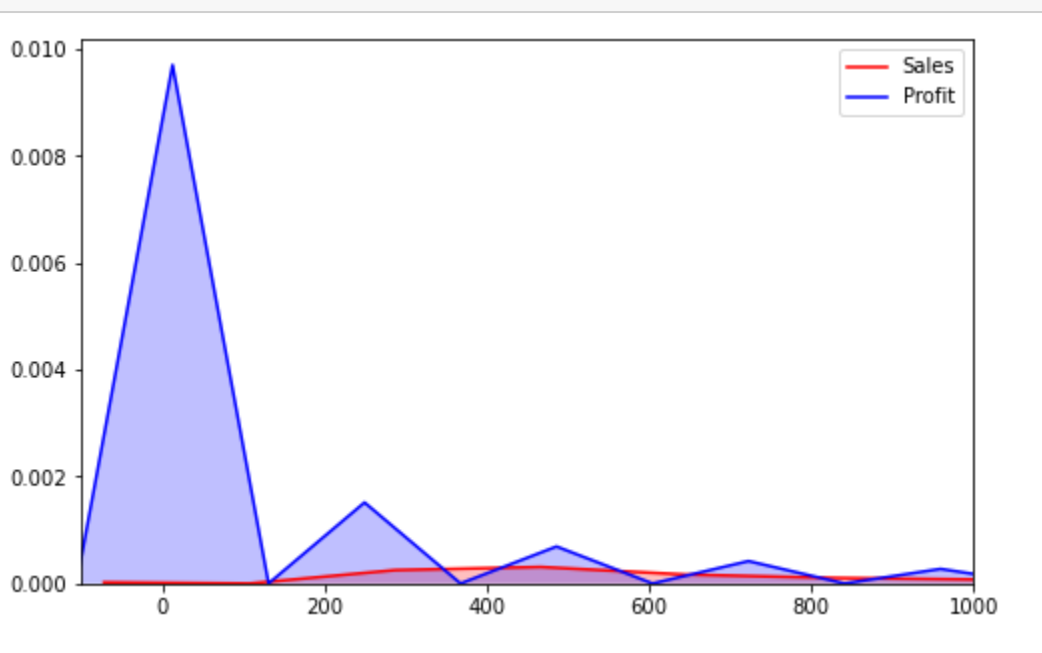
Out[10]:

	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
...	...	...	...	...	...	...	...	...	...
9989	Second Class	Consumer	United States	Miami	Florida	33180	South	Furniture	Furnishings
9990	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Furniture	Furnishings
9991	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Technology	Phones
9992	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Office Supplies	Paper
9993	Second Class	Consumer	United States	Westminster	California	92683	West	Office Supplies	Appliances

9977 rows × 13 columns

Exploratory Data Analysis

In [11]:
```
import seaborn as sns
plt.figure(figsize=(8,5))
sns.kdeplot(df["Sales"], color='red', label='Sales', shade=True)
sns.kdeplot(df["Profit"], color='blue', label='Profit', shade=True)
plt.xlim([-100,1000])
plt.legend()
plt.show()
```



It tell us that profit is higher than sales.

In [12]:
```
#Pairplot based on segment column
sns.pairplot(df, hue="Segment")
```

Out[12]: <seaborn.axisgrid.PairGrid at 0x214b9f98df0>

In [13]:
```
cor = df.corr()
cor
```

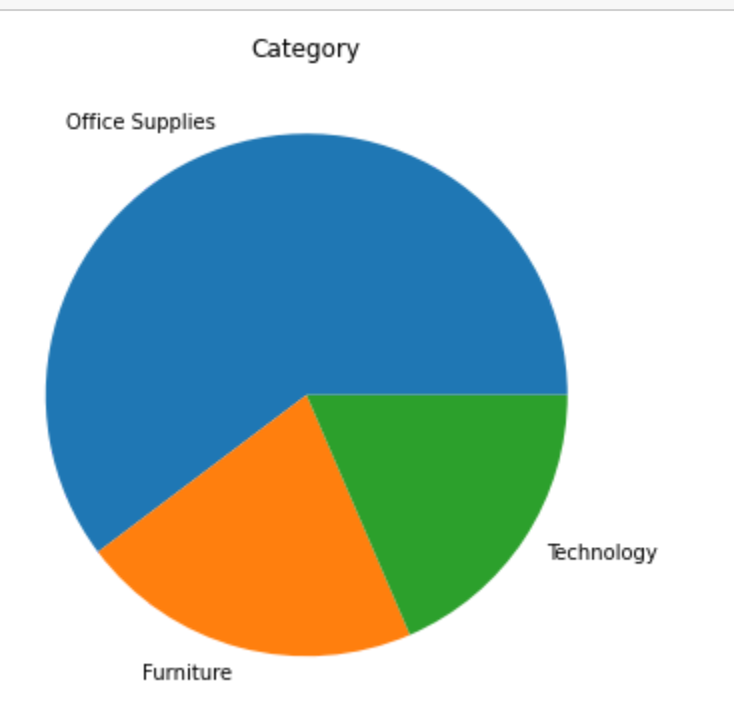
Out[13]:

	Postal Code	Sales	Quantity	Discount	Profit
Postal Code	1.000000	-0.023476	0.013110	0.059225	-0.029892
Sales	-0.023476	1.000000	0.200722	-0.028311	0.479067
Quantity	0.013110	0.200722	1.000000	0.008678	0.066211
Discount	0.059225	-0.028311	0.008678	1.000000	-0.219662
Profit	-0.029892	0.479067	0.066211	-0.219662	1.000000

From this you can say profit and sales has positive correlation with each other i.e; if the sales increase profit increases and vice versa.

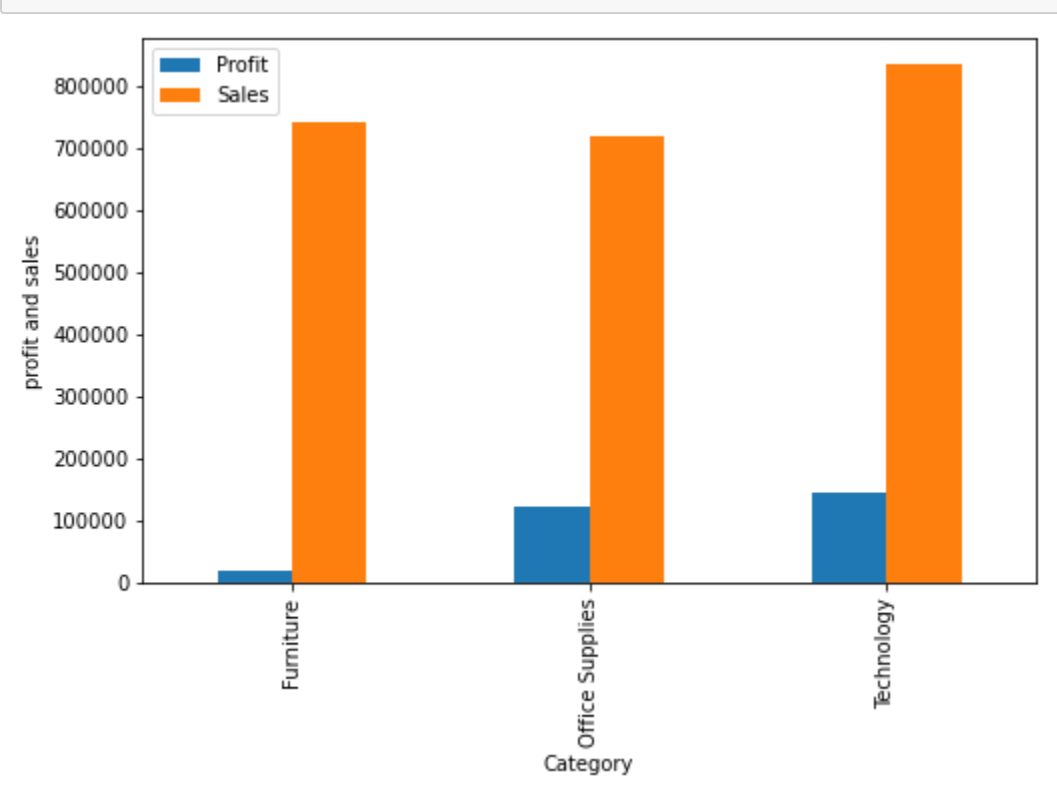
Also, the discount and profit are negatively correlated.

In [15]:
```
#pie chart for category column
plt.figure(figsize=(6,6))
plt.pie(df["Category"].value_counts(), labels = df["Category"].value_counts().index)
plt.title("Category")
plt.show()
```



Office supplies have highest demand.

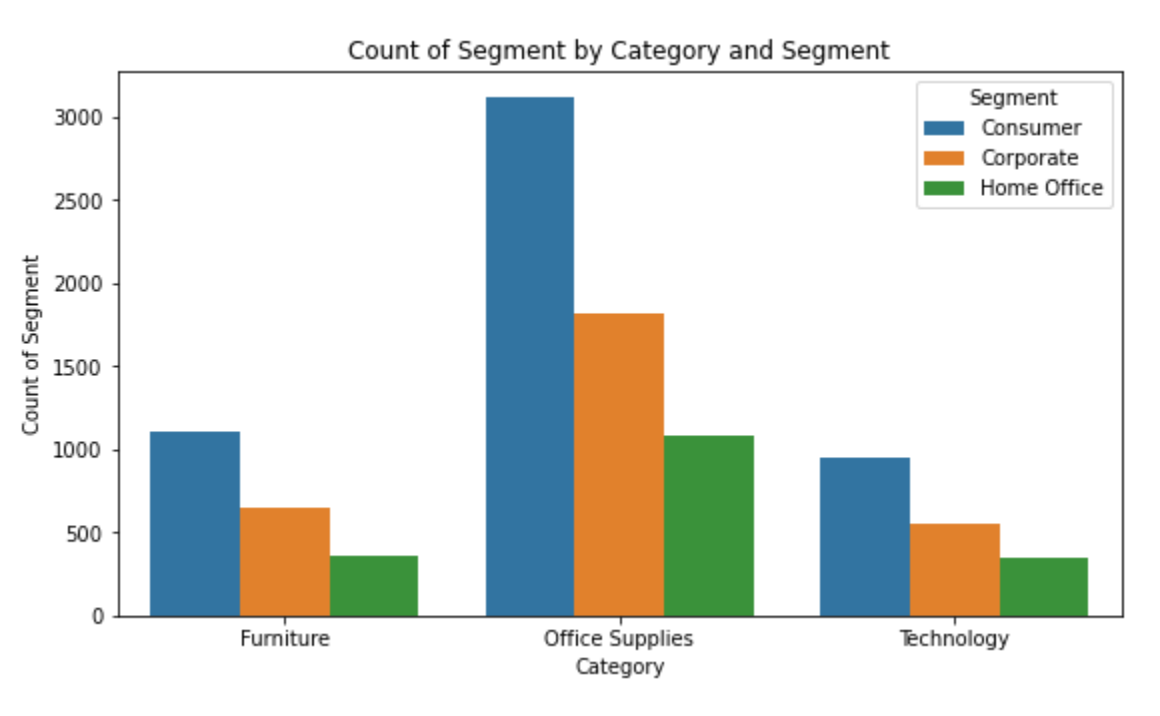
In [16]:
```
#Bar plot to check profit and sales in each category
df.groupby("Category")[["Profit","Sales"]].agg(sum).plot(kind = "bar",figsize=(8,5))
plt.ylabel("profit and sales")
plt.show()
```



It tell us that furniture is generating very less profit.

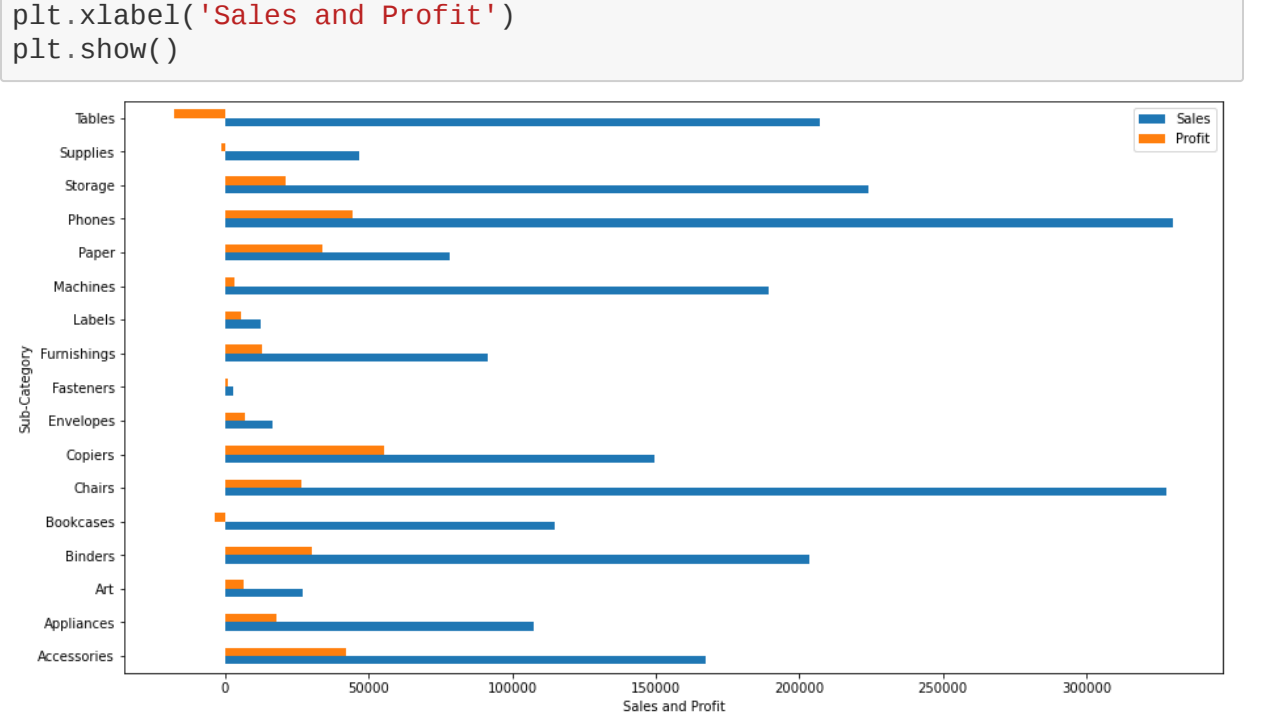
In [17]:
```
#Count plot to check purchases in each segement for each category
plt.figure(figsize=(9,5))
sns.countplot(x='Category',hue='Segment',data=df)
plt.ylabel('Count of Segment')
plt.title('Count of Segment by Category and Segment')
```

Out[17]: Text(0.5, 1.0, 'Count of Segment by Category and Segment')



The highest purchases are in consumer segment whereas home office segment has the least purchases in each category. We need to focus on the corporate and home office segment as well to increase our sales and profit.

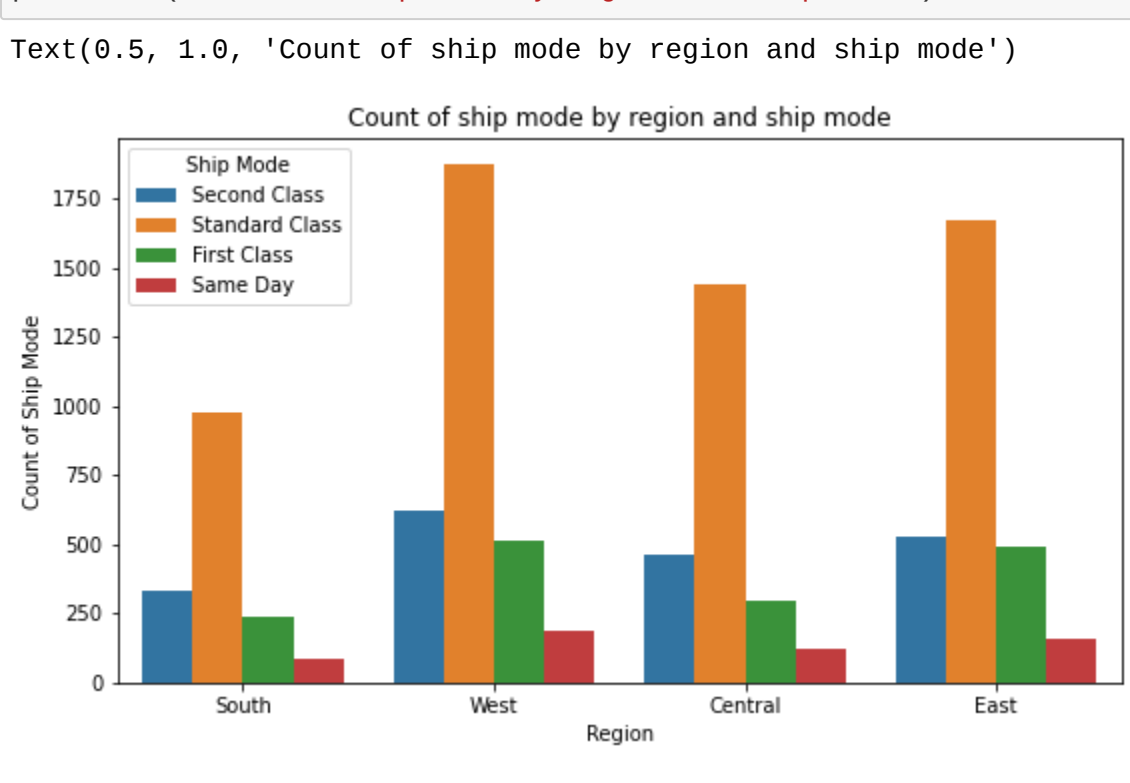
In [18]:
```
#Horizontal Bar plot for sub category
df.groupby('Sub-Category')[['Sales','Profit']].agg(sum).plot(kind='barh',figsize=(15, 8))
plt.ylabel('Sub-Category')
plt.xlabel('Sales and Profit')
plt.show()
```



We have 3 weak areas Tables, Supplies and Bookcases as we are facing loss on the sales of these items. On the sales of tables we are facing highest loss.

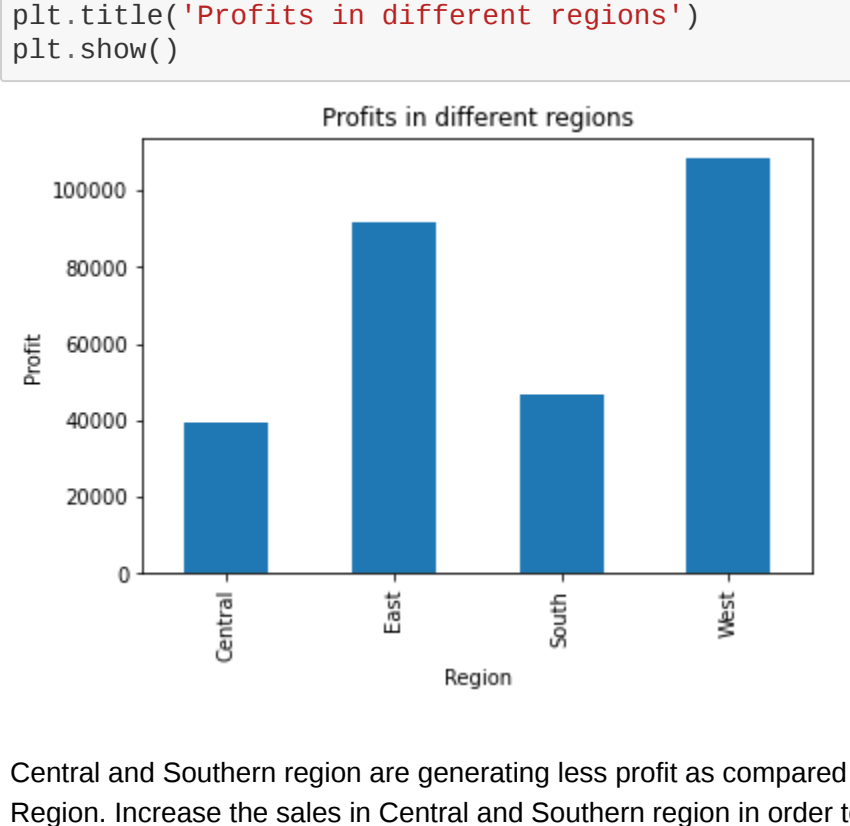
In [19]:
```
#Count plot for region and ship mode
plt.figure(figsize=(9,5))
sns.countplot(x='Region',hue='Ship Mode',data=df)
plt.ylabel('Count of Ship Mode')
plt.title('Count of ship mode by region and ship mode')
```

Out[19]: Text(0.5, 1.0, 'Count of ship mode by region and ship mode')



Least sales are in Southern region, we can attract this region by offering more discount. Standard mode of shipment is opted by most of the customers in each region.

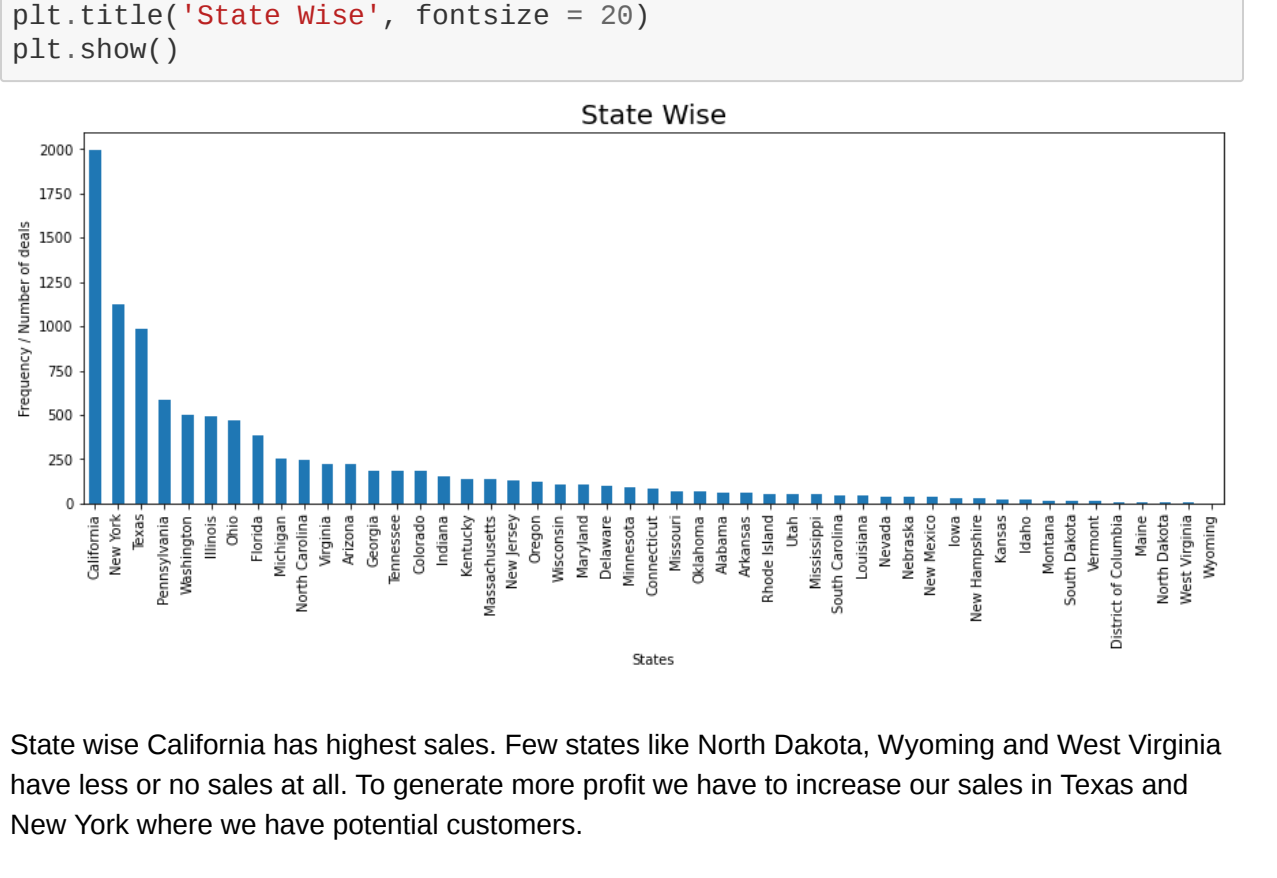
In [20]:
```
#Bar plot to check which region has least profit
df.groupby('Region')['Profit'].agg(sum).plot(kind='bar')
plt.ylabel('Profit')
plt.title('Profits in different regions')
plt.show()
```



Central and Southern region are generating less profit as compared to the Eastern and Western Region. Increase the sales in Central and Southern region in order to generate more profit in these region.

In [21]:
```
#Bar plot for state
df["State"].value_counts().plot(kind='bar',figsize=(15,5))
plt.ylabel('Frequency / Number of deals')
plt.xlabel('States')

plt.title('State Wise', fontsize = 20)
plt.show()
```



State wise California has highest sales. Few states like North Dakota, Wyoming and West Virginia have less or no sales at all. To generate more profit we have to increase our sales in Texas and New York where we have potential customers.

Conclusion

1.We should limit the sales of furniture and increase that of office supplies and technology as furtinute is generating very less profit as compared to its sales. 2.We have highest purchases in consumer segment but we should also concentrate on corporate and home office segments to increase our sales and profit. 3.In the sub-categories we are facing huge loss on the sale of tables so its sale should be minimized. 4.The sales and profit in Southern and Central region are less so we should give more incentives like discount in these states in order to increase sales, hence profit will increase. 5.Few states like North Dakota, Wyoming and West Virginia have less or no sales at all so we also need to focus on these state. 6.After the highest sales in California, we have high value of potenial customers in Texas and New York so we should concentrate more on these states to generate more profit.