Intern: Kinza Shabbir

GRIP Task 3: SampleSuperstore Exploratory Data Analysis

Perform ‘Exploratory Data Analysis’ on dataset ‘SampleSuperstore’

As a business manager, try to find out the weak areas where you can work to make more profit.

What all business problems you can derive by exploring the data?

In [6]:

```
#Importing libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
import warnings
warnings.filterwarnings('ignore')
```

In [7]:

```
# Reading data file
store=pd.read_csv("SampleSuperstore.csv")
store.head(5)
```

Out[7]:

	Ship Mode	Segment	Country	City	State	Postal Code	Region	Category	Sub-Category	
0	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Bookcases	2
1	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Chairs	7
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	9
4	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Office Supplies	Storage	

Dataset Structure

In [8]:

```
#Number of Records and columns
store.shape
```

Out[8]:

```
(9994, 13)
```

In [9]:

```
#Number of Values across three categories
store["Category"].value_counts()
```

Out[9]:

```
Office Supplies    6026
Furniture          2121
Technology         1847
Name: Category, dtype: int64
```

In [10]:

```
#Checking the number of sub categories of products
store["Sub-Category"].value_counts()
```

Out[10]:

```
Binders        1523
Paper          1370
Furnishings     957
Phones          889
Storage         846
Art             796
Accessories     775
Chairs          617
Appliances      466
Labels          364
Tables          319
Envelopes       254
Bookcases       228
Fasteners       217
Supplies        190
Machines        115
Copiers          68
Name: Sub-Category, dtype: int64
```

In [11]:

```
#checking total number of sub category
store["Category"].unique()
```

Out[11]:

```
array(['Furniture', 'Office Supplies', 'Technology'], dtype=object)
```

In [12]:

```
#checking total number of sub category
store["Sub-Category"].unique()
```

Out[12]:

```
array(['Bookcases', 'Chairs', 'Labels', 'Tables', 'Storage',
       'Furnishings', 'Art', 'Phones', 'Binders', 'Appliances', 'Paper',
       'Accessories', 'Envelopes', 'Fasteners', 'Supplies', 'Machines',
       'Copiers'], dtype=object)
```

In [13]:

```
# names of columns in the dataset
store.columns
```

Out[13]:

```
Index(['Ship Mode', 'Segment', 'Country', 'City', 'State', 'Postal Code',
       'Region', 'Category', 'Sub-Category', 'Sales', 'Quantity', 'Discoun
t',
       'Profit'],
      dtype='object')
```

In [14]:

```
#checking the columns datatypes
store.dtypes
```

Out[14]:

```
Ship Mode        object
Segment          object
Country          object
City             object
State            object
Postal Code       int64
Region           object
Category         object
Sub-Category     object
Sales           float64
Quantity          int64
Discount        float64
Profit          float64
dtype: object
```

Missing values

In [15]:

```
# Checking missing values for all columns
store.isnull().sum()
```

Out[15]:

```
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
```

Removing unnecessary columns

In [16]:

```
#country value counts
store["Country"].value_counts()
```

Out[16]:

```
United States    9994
Name: Country, dtype: int64
```

The dataset contains only US country. Removing country column

In [17]:

```
store_new=store.drop("Country",axis=1)
store_new.head(5)
```

Out[17]:

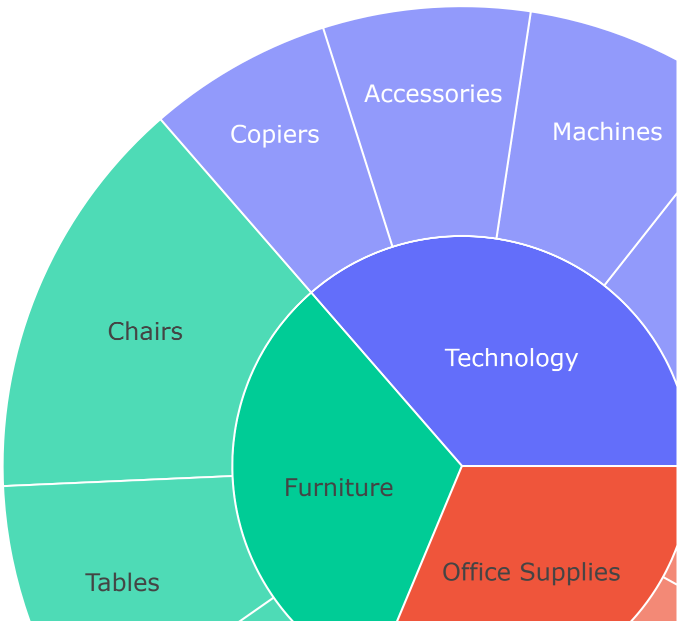
	Ship Mode	Segment	City	State	Postal Code	Region	Category	Sub-Category	Sales
0	Second Class	Consumer	Henderson	Kentucky	42420	South	Furniture	Bookcases	261.9600
1	Second Class	Consumer	Henderson	Kentucky	42420	South	Furniture	Chairs	731.9400
2	Second Class	Corporate	Los Angeles	California	90036	West	Office Supplies	Labels	14.6200
3	Standard Class	Consumer	Fort Lauderdale	Florida	33311	South	Furniture	Tables	957.5775
4	Standard Class	Consumer	Fort Lauderdale	Florida	33311	South	Office Supplies	Storage	22.3680

Exploratory analysis

Categories & Sub-Categories

In [18]:

```
# to isplay the categories and sub-categories
import plotly.express as px
fig = px.sunburst(store,path=['Category','Sub-Category'],
                 values='Sales',color='Category',
                 hover_data =['Sales','Profit'])
fig.update_layout(height=600,title_text='Product Categories & Sub-Categories'
fig.show()
```



Product Categories & Sub-Categories

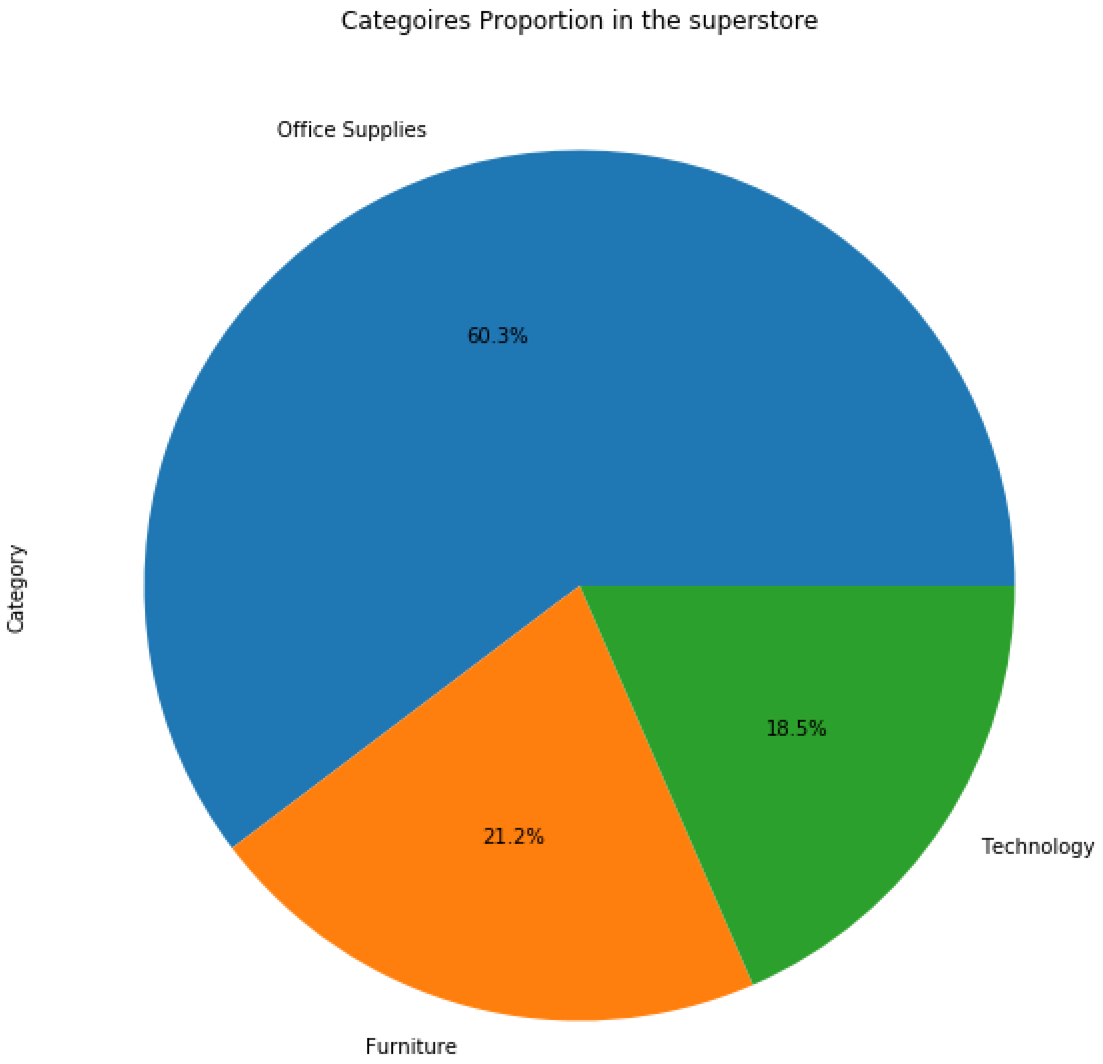
The above pie chart displays types of sub-categories for Categories.

1. **Furniture** includes **four** sub-categories which are bookcases, Chairs, Tableas, Furnishings.
2. **Office Supplies** contains Labels, Storage, Art, Binders, Appliances, Paper, Envelops, Fasteners, Supplies
3. **Technology** includes Phones, Accessories, Machines and Copiers.

Categories distribution

In [19]:

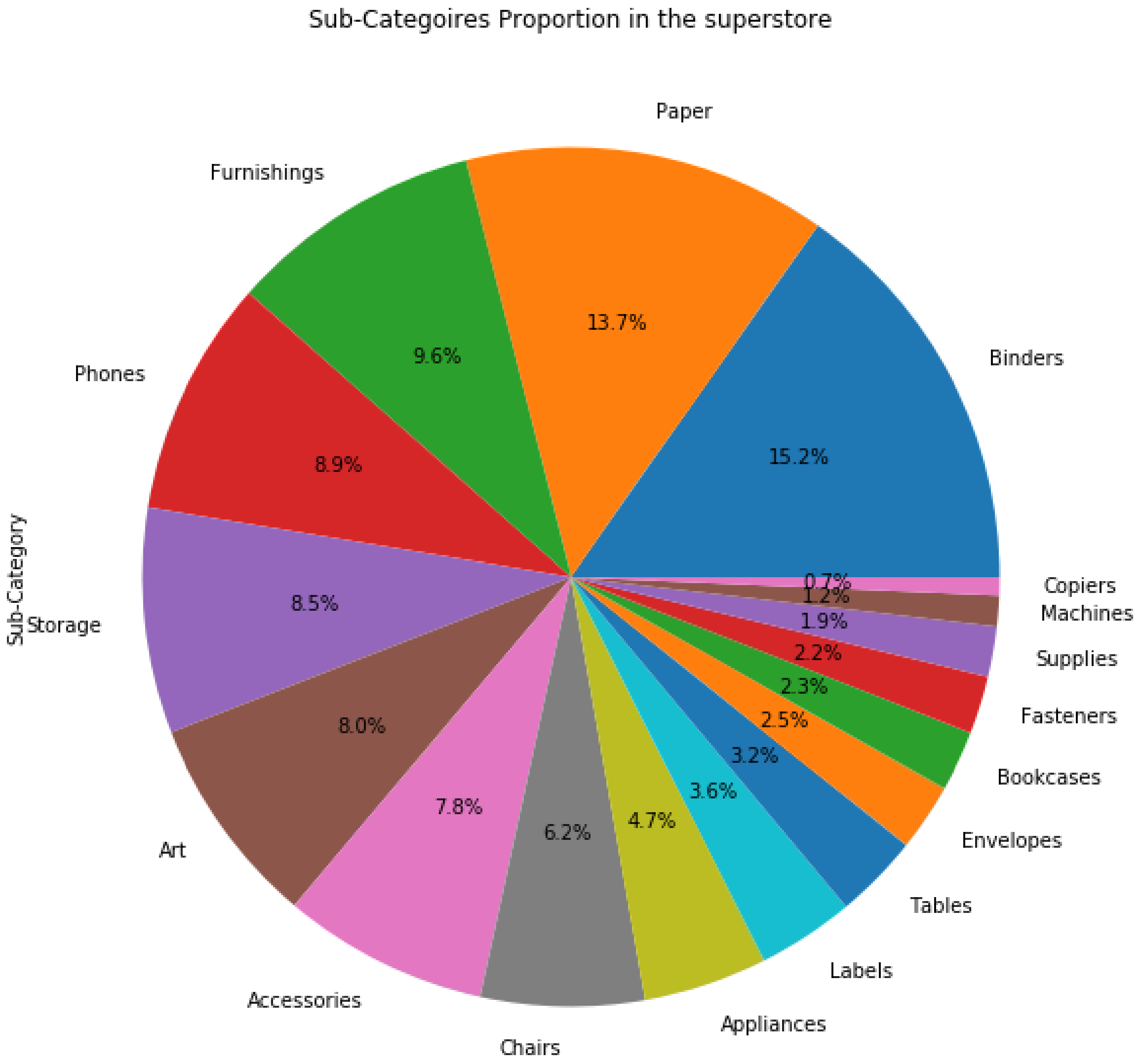
```
# pie chart to see the proportion of sub-categories
plt.figure(figsize=(12,10))
store_new['Category'].value_counts().plot.pie(autopct="%1.1f%%")
plt.title('Categoires Proportion in the superstore')
plt.show()
```



Above Pie chart shows that **Office supplies** makes the **60.3%** of total products which indicates there is chance that Office supplies benefitting superstore more than other categories.

Sub-Categories distribution

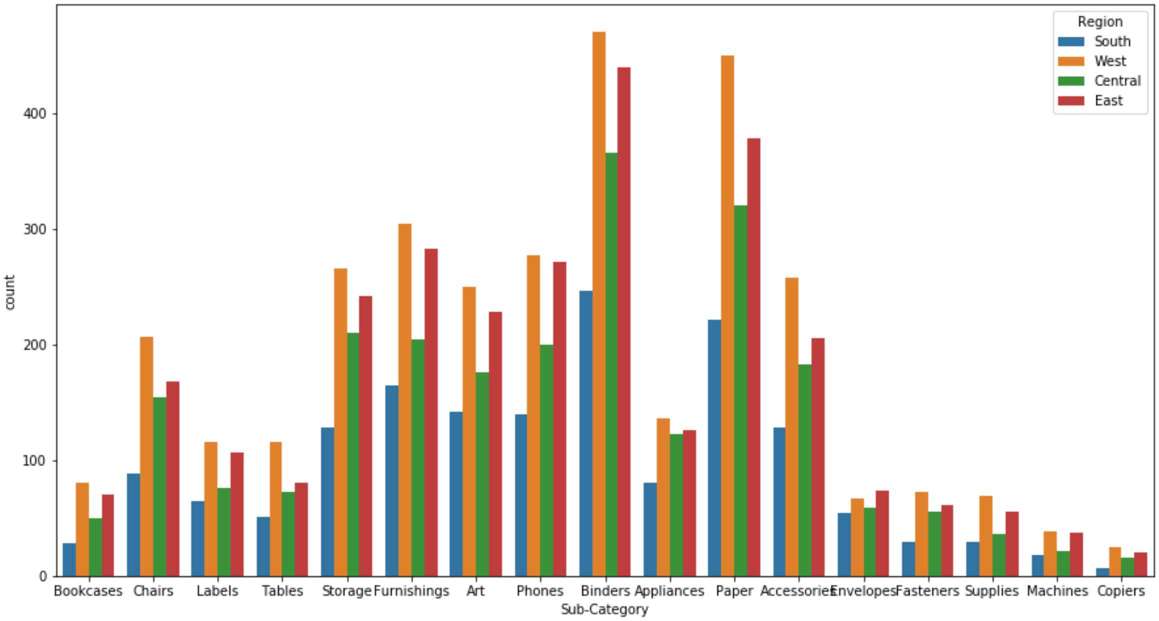
```
In [20]:  # pie chart to see the proportion of sub-categories
          plt.figure(figsize=(12,10))
          store_new['Sub-Category'].value_counts().plot.pie(autopct="%1.1f%%")
          plt.title('Sub-Categoires Proportion in the superstore')
          plt.show()
```



Sub-category distribution across regions of US

```
In [21]: #Count of sub-category region wise
         plt.figure(figsize=(15,8))
         sns.countplot(x="Sub-Category",hue="Region",data=store_new)
         plt.show

Out[21]: <function matplotlib.pyplot.show(*args, **kw)>
```



The above bar plot for sub-categories count for all the regions reveals:

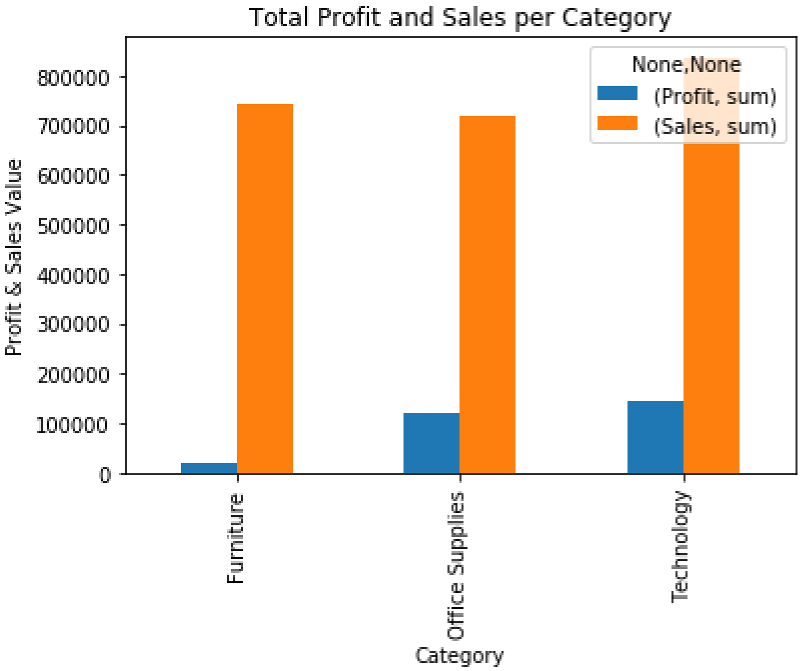
1. **West** has highest stock for all the sub-categories having highest sales and Profit.
2. However,**South** has lowest demand for the sub-categories.

Sales & Profit comparison

1. Category

In [26]:

```
store_new.groupby('Category')['Profit','Sales'].agg(['sum']).plot.bar()
plt.title('Total Profit and Sales per Category')
plt.xlabel("Category")
plt.ylabel("Profit & Sales Value")
plt.show()
```



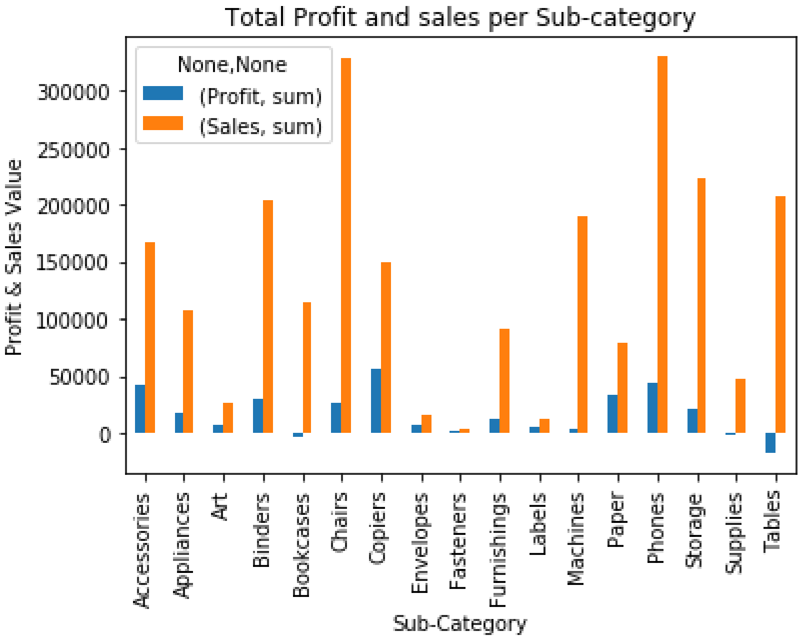
The above bar plot shows that

1. **Technology** is the most profitable category
2. **Office supplies** total sales are less than **furniture** but it makes more profit.
3. **Furnitue** is the category which is making least profit.

2. Sub-Categories

In [27]:

```
store_new.groupby("Sub-Category")["Profit","Sales"].agg(["sum"]).plot.bar()
plt.title("Total Profit and sales per Sub-category")
plt.xlabel("Sub-Category")
plt.ylabel("Profit & Sales Value")
plt.show()
```



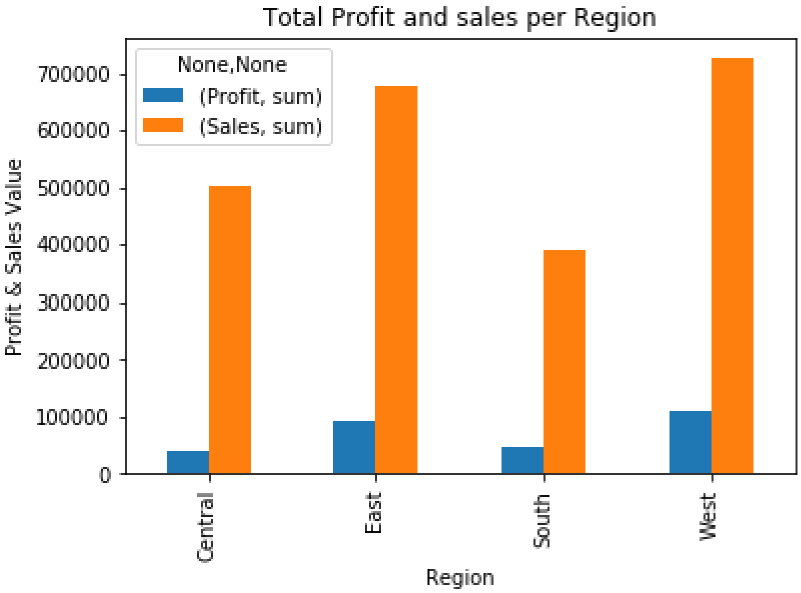
Interesting findings for sub-categories from above plot:

1. Despite the highest sales for **Chairs** and **Phones**, the most profit is made through **Copiers**
2. Overall,**Bookcases** and **Tables** sub-category are giving loss to superstore.

3. Region

In [28]:

```
store_new.groupby("Region")["Profit","Sales"].agg(["sum"]).plot.bar()
plt.title("Total Profit and sales per Region")
plt.xlabel("Region")
plt.ylabel("Profit & Sales Value")
plt.show()
```



The bar plot reveal:

1. The store in the **west** of US is getting more benefit. Moreover, South region makes more profit beside its least sales.

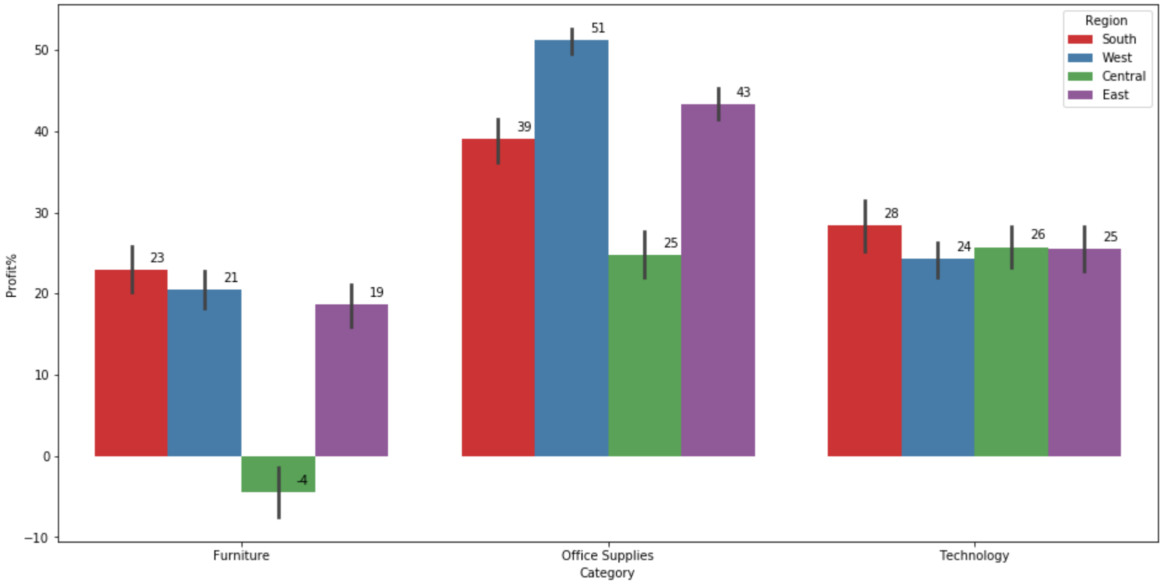
Profit Percentage/Loss in different regions of US

In [29]:

```
# Finding the cost of each product and Calculating Profit percentage for sub-
store['Cost'] = store['Sales'] - store['Profit']
store['Profit%'] = store['Profit']/store['Cost']*100
```

1. Category

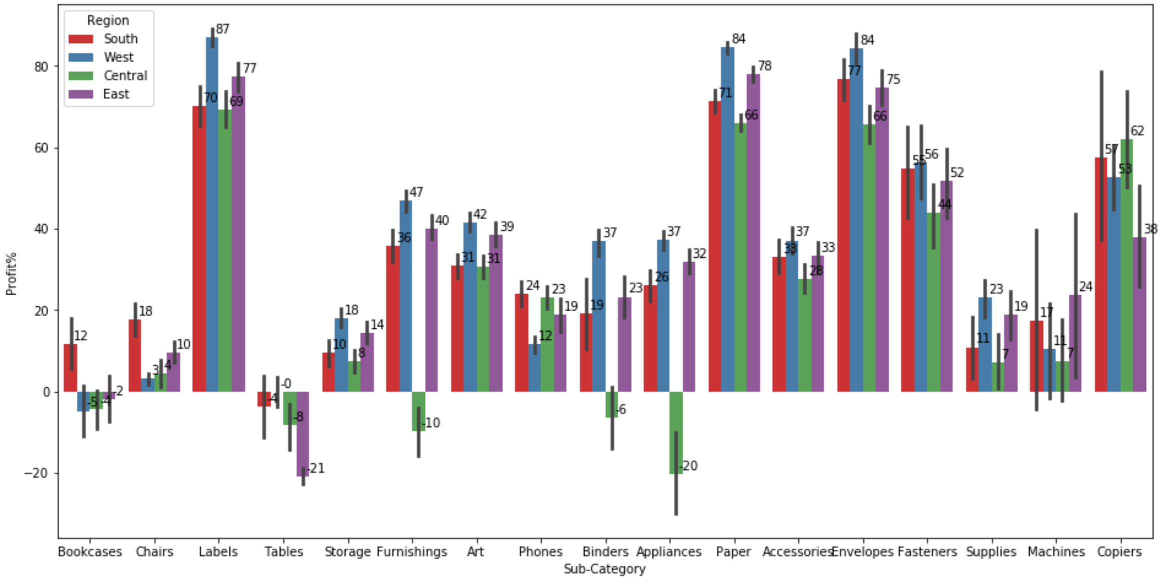
```
In [30]:
#displaying Profit percentage for each categories in four regions of US
fig=plt.figure(figsize=(16,8))
ax = fig.add_subplot(111)
sns.barplot('Category','Profit%',hue='Region',palette='Set1',data=store)
for o in ax.patches:
    ax.annotate('{:.0f}'.format(o.get_height()), (o.get_x()+0.15, o.get_heigh
plt.show()
```



The above plot shows that all the major categories furniture, office supplies and technology have benefitted the store in all regions, except the **furniture** category gave negative profit in **central** region of US

2. Sub-categories

```
In [31]:  #displaying Profit percentage for each sub-categories in four regions of US
          fig=plt.figure(figsize=(16,8))
          ax = fig.add_subplot(111)
          sns.barplot('Sub-Category','Profit%',hue='Region',palette='Set1',data=store)
          for o in ax.patches:
              ax.annotate('{:.0f}'.format(o.get_height()), (o.get_x()+0.15, o.get_heigh
          plt.show()
```



The above plot gives details about profit/loss for all sub-categories in four regions of US.

1. **Tables** sub-category from Furniture has performed worst on all regions giving most loss to the store. However, other sub-categories bookcases, tables, furnishings, Binders and appliances have been in loss on some locations.

Top 10 least profit to superstore

In [40]:

```
#Least profitable Sub-Category
store[['Region','Category','Sub-Category','Profit','Profit%']].sort_values('P
    .background_gradient(cmap='Greens',subset=['Profit'])\
    .background_gradient(cmap='RdPu',subset=['Profit%'])
```

Out[40]:

	Region	Category	Sub-Category	Profit	Profit%
7772	East	Technology	Machines	-6599.978000	-59.459459
683	South	Technology	Machines	-3839.990400	-32.432432
9774	Central	Office Supplies	Binders	-3701.892800	-62.962963
3011	West	Technology	Machines	-3399.980000	-57.142857
4991	Central	Office Supplies	Binders	-2929.484500	-60.784314
3151	East	Technology	Machines	-2639.991200	-59.459459
5310	Central	Office Supplies	Binders	-2287.782000	-60.000000
9639	South	Furniture	Tables	-1862.312400	-30.232558
1199	Central	Office Supplies	Binders	-1850.946400	-62.962963
2697	South	Technology	Machines	-1811.078400	-7.407407

Findings

As a business manager, try to find out the weak areas where you can work to make more profit

1. The least profitable sub-category is **Machines** in Technology in the South region of US.
2. Second sub-category is **Tables** giving loss to superstore in the South region of US.
3. The superstore can work on strategies to improve selling of Tables, Machines, Binders, Bookcases and Appliances which can increase sales and bring profit.