Designing a Visualization for Your Manager

Project Description

The Sales - Superstore dataset contains detailed information about your company's sales. Your manager, Sylvia, has made a decision to cut the three worst performing sub-categories in their region in terms of Sales. To do this, she has asked you to create one data visualization that will identify which three sub-categories are the worst performers by region, and show how much worse they perform than other sub-categories. Sylvia will use this visualization to inform which product categories to cut, and in which regions.

Import Libraries

In [1]:

```
import numpy as np
from numpy import count_nonzero
from numpy import median
from numpy import mean
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
import random

import statsmodels.api as sm
import statsmodels.formula.api as smf
from statsmodels.formula.api import ols

import datetime
from datetime import datetime, timedelta

import scipy.stats

%matplotlib inline
#sets the default autosave frequency in seconds
%autosave 60
sns.set_style('dark')
sns.set(font_scale=1.2)

plt.rc('axes', titlesize=9)
plt.rc('axes', labelsize=14)
plt.rc('xtick', labelsize=12)
plt.rc('ytick', labelsize=12)

import warnings
warnings.filterwarnings('ignore')


pd.set_option('display.max_columns',None)
#pd.set_option('display.max_rows',None)
pd.set_option('display.width', 1000)
pd.set_option('display.float_format','{:.2f}'.format)

random.seed(0)
np.random.seed(0)
np.set_printoptions(suppress=True)
```

```
Autosaving every 60 seconds
```

Exploratory Data Analysis

In [2]:

```
df = pd.read_csv("Superstore.csv",parse_dates=['Order Date','Ship Date'])
```

In [3]:

```
df
```

Out[3]:

	Category	City	Country	Customer Name	Manufacturer	Order Date	Order ID	Postal Code	Product Name	Region	Segn
0	Office Supplies	Houston	United States	Darren Powers	Message Book	2011-04-01	CA-2011-103800	77095	Message Book, Wirebound, Four 5 1/2" X 4" Form...	Central	Consu
1	Office Supplies	Naperville	United States	Phillina Ober	GBC	2011-05-01	CA-2011-112326	60540	GBC Standard Plastic Binding Systems Combs	Central	H C
2	Office Supplies	Naperville	United States	Phillina Ober	Avery	2011-05-01	CA-2011-112326	60540	Avery 508	Central	H C
3	Office Supplies	Naperville	United States	Phillina Ober	SAFCO	2011-05-01	CA-2011-112326	60540	SAFCO Boltless Steel Shelving	Central	H C
4	Office Supplies	Philadelphia	United States	Mick Brown	Avery	2011-06-01	CA-2011-141817	19143	Avery Hi-Liter EverBold Pen Style Fluorescent ...	East	Consu
...	...	...	...	...	...	...	...	...	...	...	
9989	Office Supplies	Loveland	United States	Jill Matthias	Other	2014-12-31	CA-2014-156720	80538	Bagged Rubber Bands	West	Consu
9990	Office Supplies	Fairfield	United States	Erica Bern	Cardinal	2014-12-31	CA-2014-115427	94533	Cardinal Slant-D Ring Binder, Heavy Gauge Vinyl	West	Corpc
9991	Office Supplies	Fairfield	United States	Erica Bern	GBC	2014-12-31	CA-2014-115427	94533	GBC Binding covers	West	Corpc
9992	Technology	New York City	United States	Patrick O'Donnell	Other	2014-12-31	CA-2014-143259	10009	Gear Head AU3700S Headset	East	Consu

	Category	City	Country	Customer Name	Manufacturer	Order Date	Order ID	Postal Code	Product Name	Region	Segn
9993	Office Supplies	Columbus	United States	Chuck Clark	Eureka	2014-12-31	CA-2014-126221	47201	Eureka The Boss Plus 12-Amp Hard Box Upright V...	Central	H C

9994 rows × 21 columns

In [4]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9994 entries, 0 to 9993
Data columns (total 21 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Category           9994 non-null   object
 1   City               9994 non-null   object
 2   Country            9994 non-null   object
 3   Customer Name      9994 non-null   object
 4   Manufacturer       9994 non-null   object
 5   Order Date         9994 non-null   datetime64[ns]
 6   Order ID           9994 non-null   object
 7   Postal Code        9994 non-null   int64
 8   Product Name       9994 non-null   object
 9   Region             9994 non-null   object
 10  Segment            9994 non-null   object
 11  Ship Date          9994 non-null   datetime64[ns]
 12  Ship Mode          9994 non-null   object
 13  State              9994 non-null   object
 14  Sub-Category       9994 non-null   object
 15  Discount           9994 non-null   float64
 16  Number of Records  9994 non-null   int64
 17  Profit             9994 non-null   int64
 18  Profit Ratio       9994 non-null   float64
 19  Quantity           9994 non-null   int64
 20  Sales              9994 non-null   int64
dtypes: datetime64[ns](2), float64(2), int64(5), object(12)
memory usage: 1.6+ MB
```

In [5]:

```
df.describe(include='all')
```

Out[5]:

	Category	City	Country	Customer Name	Manufacturer	Order Date	Order ID	Postal Code	Product Name	Region	Segment
count	9994	9994	9994	9994	9994	9994	9994	9994.00	9994	9994	9994
unique	3	531	1	793	174	1238	5009	NaN	1841	4	3
top	Office Supplies	New York City	United States	William Brown	Other	2013-06-09 00:00:00	CA-2014-100111	NaN	Staples	West	Consumer
freq	6026	915	9994	37	2074	38	14	NaN	227	3203	5191
first	NaN	NaN	NaN	NaN	NaN	2011-01-02 00:00:00	NaN	NaN	NaN	NaN	NaN

	Category	City	Country	Customer Name	Manufacturer	Order Date	Order ID	Postal Code	Product Name	Region	Segment
last	NaN	NaN	NaN	NaN	NaN	2014-12-31 00:00:00	NaN	NaN	NaN	NaN	NaN
mean	NaN	NaN	NaN	NaN	NaN	NaN	NaN	55190.38	NaN	NaN	NaN
std	NaN	NaN	NaN	NaN	NaN	NaN	NaN	32063.69	NaN	NaN	NaN
min	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1040.00	NaN	NaN	NaN
25%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	23223.00	NaN	NaN	NaN
50%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	56430.50	NaN	NaN	NaN
75%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	90008.00	NaN	NaN	NaN
max	NaN	NaN	NaN	NaN	NaN	NaN	NaN	99301.00	NaN	NaN	NaN

In [6]:

```
df.columns
```

Out[6]:

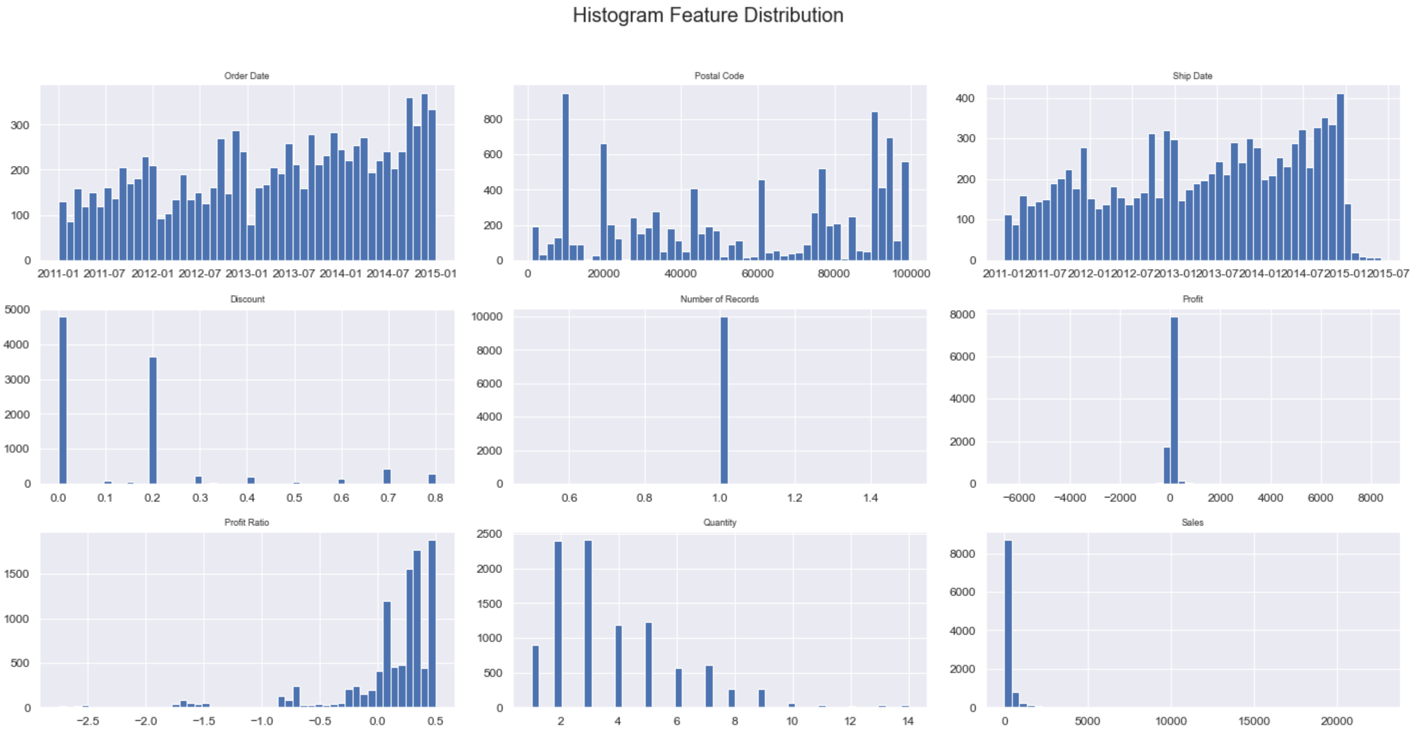
```
Index(['Category', 'City', 'Country', 'Customer Name', 'Manufacturer', 'Order Date', 'Orde
r ID', 'Postal Code', 'Product Name', 'Region', 'Segment', 'Ship Date', 'Ship Mode', 'Stat
e', 'Sub-Category', 'Discount', 'Number of Records', 'Profit', 'Profit Ratio', 'Quantity',
'Sales'], dtype='object')
```

Data Visualization

Univariate Data Exploration

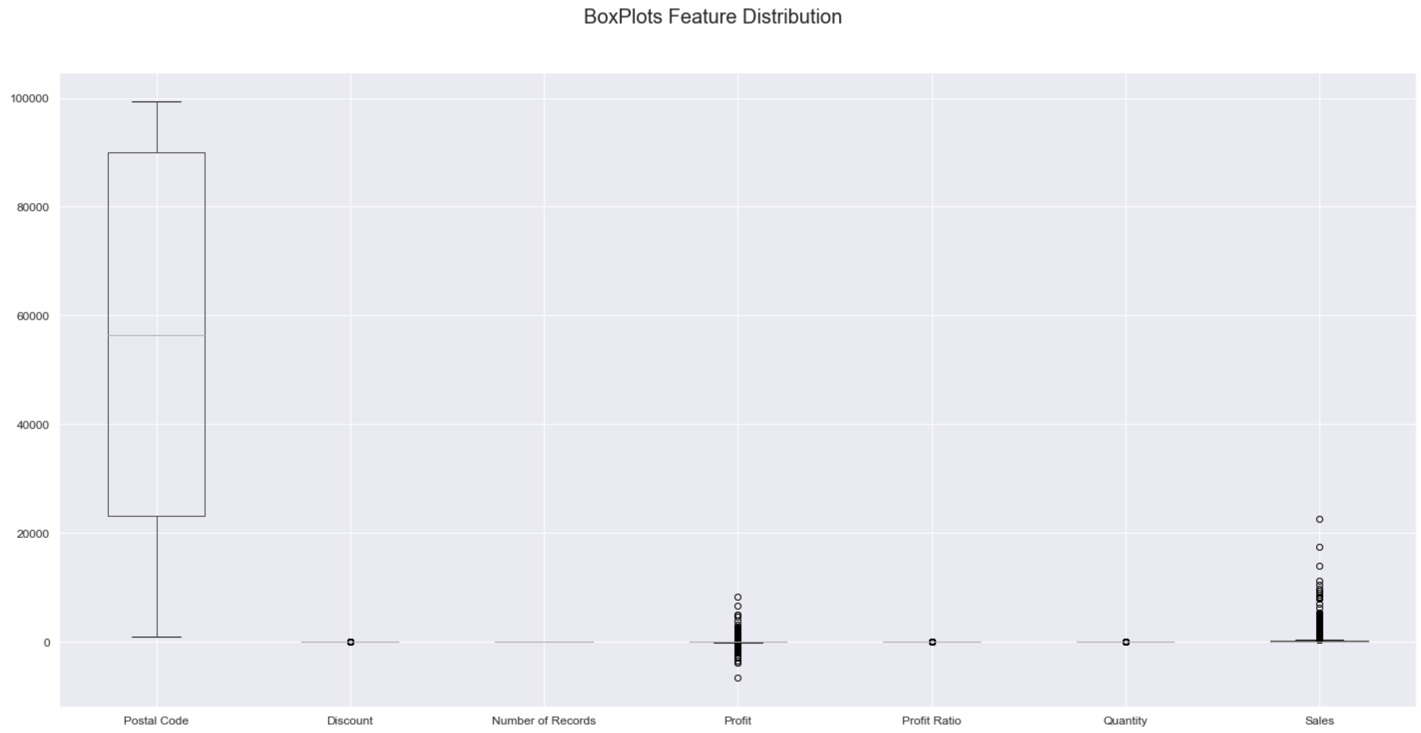
In [7]:

```
df.hist(bins=50, figsize=(20,10))
plt.suptitle('Histogram Feature Distribution', x=0.5, y=1.02, ha='center', fontsize=20)
plt.tight_layout()
plt.show()
```



In [8]:

```
df.boxplot(figsize=(20,10))
plt.suptitle('BoxPlots Feature Distribution', x=0.5, y=1.02, ha='center', fontsize=20)
plt.tight_layout()
plt.show()
```



In [9]:

```
df.columns
```

Out[9]:

```
Index(['Category', 'City', 'Country', 'Customer Name', 'Manufacturer', 'Order Date', 'Orde
r ID', 'Postal Code', 'Product Name', 'Region', 'Segment', 'Ship Date', 'Ship Mode', 'Stat
e', 'Sub-Category', 'Discount', 'Number of Records', 'Profit', 'Profit Ratio', 'Quantity',
'Sales'], dtype='object')
```

In [10]:

```
df2 = df[['Category','Region', 'Segment', 'State', 'Sub-Category', 'Profit', 'Quantity', '
df2
```

Out[10]:

	Category	Region	Segment	State	Sub-Category	Profit	Quantity	Sales
0	Office Supplies	Central	Consumer	Texas	Paper	6	2	16
1	Office Supplies	Central	Home Office	Illinois	Binders	-5	2	4
2	Office Supplies	Central	Home Office	Illinois	Labels	4	3	12
3	Office Supplies	Central	Home Office	Illinois	Storage	-65	3	273
4	Office Supplies	East	Consumer	Pennsylvania	Art	5	3	20
...	...	...	...	...	...	...	...	...
9989	Office Supplies	West	Consumer	Colorado	Fasteners	-1	3	3
9990	Office Supplies	West	Corporate	California	Binders	5	2	14
9991	Office Supplies	West	Corporate	California	Binders	6	2	21
9992	Technology	East	Consumer	New York	Phones	3	7	91
9993	Office Supplies	Central	Home Office	Indiana	Appliances	57	2	209

9994 rows × 8 columns

Groupby Function

```
In [11]:
d1 = df2.groupby(["Category"]).sum()
d1
```

```
Out[11]:
                 Profit  Quantity   Sales
        Category
       Furniture  18444      8028  742006
 Office Supplies 122474     22906  719127
      Technology 145429      6939  836221
```

```
In [12]:
d2 = df2.groupby(["Region"]).sum()
d2
```

```
Out[12]:
          Profit  Quantity   Sales
  Region
 Central   39719      8780  501256
    East   91521     10618  678834
   South   46721      6209  391750
    West  108386     12266  725514
```

```
In [13]:
d3 = df2.groupby(["Segment"]).sum()
d3
```

```
Out[13]:
              Profit  Quantity    Sales
     Segment
    Consumer  134113     19521  1161497
   Corporate   91965     11608   706183
 Home Office   60269      6744   429674
```

```
In [14]:
d4 = df2.groupby(["State"]).sum().nsmallest(10, columns=["Profit"])
d4
```

```
Out[14]:
                 Profit  Quantity   Sales
           State
           Texas -25714      3724  170187
            Ohio -16962      1759   78253
    Pennsylvania -15550      2153  116522
        Illinois -12607      1845   80162
  North Carolina  -7495       983   55604
        Colorado  -6527       693   32110
```

	Profit	Quantity	Sales
State			
Tennessee	-5347	681	30662
Arizona	-3432	862	35283
Florida	-3412	1379	89479
Oregon	-1187	499	17431

In [15]:

```
d5 = df2.groupby(["Sub-Category"]).sum().nsmallest(10, columns=["Profit"])
d5
```

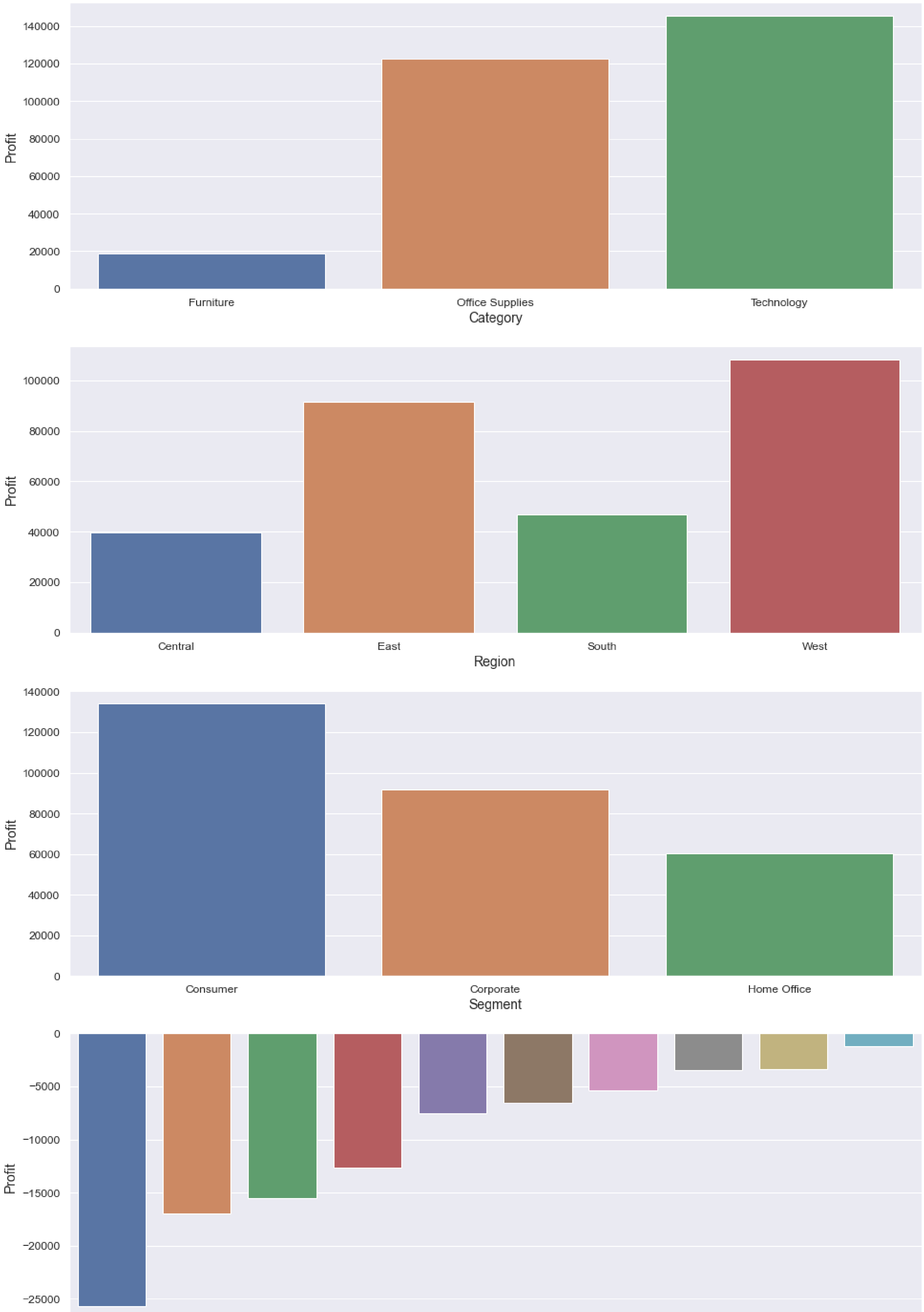
Out[15]:

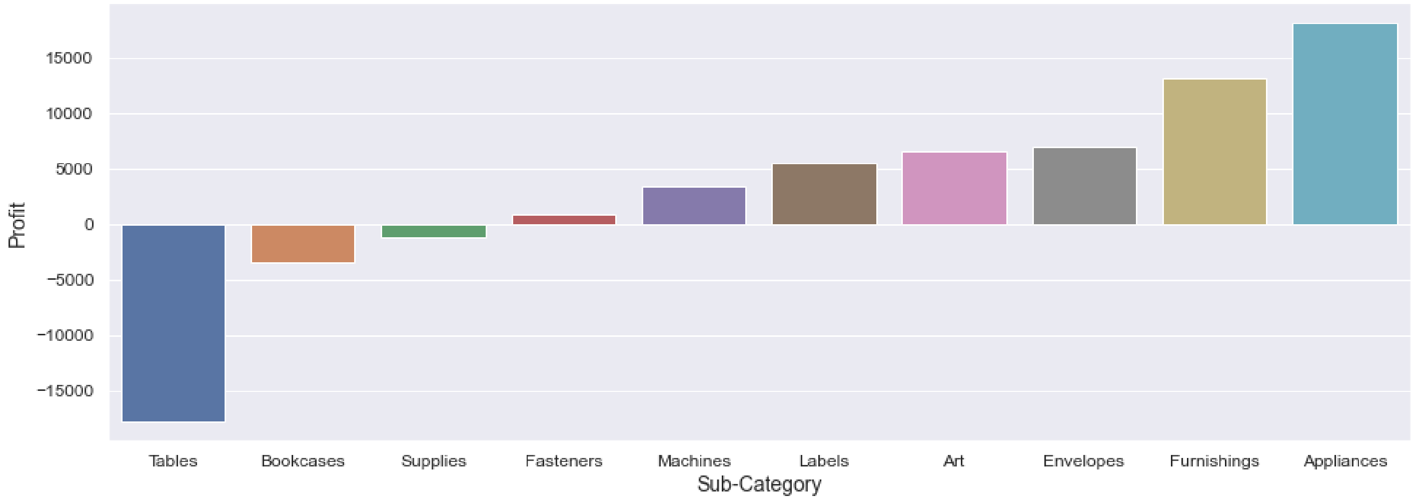
	Profit	Quantity	Sales
Sub-Category			
Tables	-17733	1241	206968
Bookcases	-3479	868	114879
Supplies	-1187	647	46679
Fasteners	952	914	3024
Machines	3387	440	189243
Labels	5558	1400	12507
Art	6530	3000	27137
Envelopes	6956	906	16477
Furnishings	13070	3563	91705
Appliances	18132	1729	107538

In [16]:

```
# Plot 4 rows and 1 column (can be expanded)

fig, ax = plt.subplots(5,1, sharex=False, figsize=(16,32))
#fig.suptitle('Bar Plots')


sns.barplot(x=d1.index, y="Profit", data=d1, ax=ax[0])
#ax[0].set_title('Title of the first chart')
#ax[0].tick_params('x', labelrotation=45)

sns.barplot(x=d2.index, y="Profit", data=d2, ax=ax[1])
#ax[1].set_title('Title of the second chart')
#ax[1].tick_params('x', labelrotation=45)

sns.barplot(x=d3.index, y="Profit", data=d3, ax=ax[2])
#ax[2].set_title('Title of the third chart')
#ax[2].tick_params('x', labelrotation=45)

sns.barplot(x=d4.index, y="Profit", data=d4, ax=ax[3])
#ax[3].set_title('Title of the fourth chart')
#ax[3].tick_params('x', labelrotation=45)

sns.barplot(x=d5.index, y="Profit", data=d5, ax=ax[4])
#ax[3].set_title('Title of the fourth chart')
#ax[3].tick_params('x', labelrotation=45)
```

```
plt.show()
```


140000
120000
100000
80000
60000
40000
20000
0
Profit
Furniture
Office Supplies
Technology
Category
100000
80000
60000
40000
20000
0
Profit
Central
East
South
West
Region
140000
120000
100000
80000
60000
40000
20000
0
Profit
Consumer
Corporate
Home Office
Segment
0
−5000
−10000
−15000
−20000
−25000
Profit



In [17]:

```
df3 = df2[(df2["Sub-Category"] == "Tables") | (df2["Sub-Category"] == "Bookcases") | (df2
df3
```

Out[17]:

	Category	Region	Segment	State	Sub-Category	Profit	Quantity	Sales
27	Furniture	West	Consumer	California	Bookcases	4	3	334
32	Furniture	East	Corporate	Pennsylvania	Bookcases	-53	4	62
39	Furniture	West	Consumer	Arizona	Bookcases	-321	5	181
63	Furniture	Central	Corporate	South Dakota	Bookcases	40	2	142
72	Furniture	West	Consumer	California	Tables	-17	3	333
...	...	...	...	...	...	...	...	...
9931	Furniture	East	Consumer	New York	Bookcases	5	2	192
9940	Furniture	East	Consumer	Ohio	Tables	-105	2	273
9949	Office Supplies	Central	Consumer	Texas	Supplies	5	7	45
9958	Furniture	Central	Consumer	Texas	Bookcases	-12	2	79
9987	Furniture	East	Consumer	New York	Bookcases	12	4	323

737 rows × 8 columns

In [18]:

```
df3.reset_index(inplace=True, drop=True)
```

In [19]:

```
df3
```

Out[19]:

	Category	Region	Segment	State	Sub-Category	Profit	Quantity	Sales
0	Furniture	West	Consumer	California	Bookcases	4	3	334
1	Furniture	East	Corporate	Pennsylvania	Bookcases	-53	4	62
2	Furniture	West	Consumer	Arizona	Bookcases	-321	5	181
3	Furniture	Central	Corporate	South Dakota	Bookcases	40	2	142
4	Furniture	West	Consumer	California	Tables	-17	3	333

	Category	Region	Segment	State	Sub-Category	Profit	Quantity	Sales
...	...	...	...	...	...	...	...	...
732	Furniture	East	Consumer	New York	Bookcases	5	2	192
733	Furniture	East	Consumer	Ohio	Tables	-105	2	273
734	Office Supplies	Central	Consumer	Texas	Supplies	5	7	45
735	Furniture	Central	Consumer	Texas	Bookcases	-12	2	79
736	Furniture	East	Consumer	New York	Bookcases	12	4	323

737 rows × 8 columns

Create Pivot Tables

In [20]:

```
table1 = pd.pivot_table(data=df3, values="Profit", index=["Sub-Category","Segment"], aggfu
                        columns=["State"])
```

In [21]:

```
table1
```

Out[21]:

	State	Alabama	Arizona	Arkansas	California	Colorado	Connecticut	Delaware	Florida	Georgia
Sub-Category	Segment									
Bookcases	Consumer	NaN	-513.00	172.00	454.00	-1856.00	NaN	139.00	-141.00	755.00
	Corporate	NaN	-357.00	NaN	690.00	-555.00	456.00	338.00	17.00	76.00
	Home Office	NaN	NaN	NaN	272.00	NaN	NaN	NaN	8.00	53.00
Supplies	Consumer	2.00	4.00	NaN	501.00	-292.00	22.00	NaN	-3.00	0.00
	Corporate	10.00	1.00	NaN	339.00	NaN	8.00	NaN	-194.00	10.00
	Home Office	NaN	-35.00	NaN	22.00	1.00	NaN	NaN	-1.00	NaN
Tables	Consumer	200.00	-992.00	45.00	-125.00	-509.00	-16.00	-49.00	-1300.00	NaN
	Corporate	355.00	-289.00	NaN	-69.00	-466.00	-4.00	-37.00	-843.00	NaN
	Home Office	NaN	-1000.00	NaN	-115.00	NaN	NaN	NaN	-350.00	138.00

In [22]:

```
table1.describe()
```

Out[22]:

State	Alabama	Arizona	Arkansas	California	Colorado	Connecticut	Delaware	Florida	Georgia	Idaho	Illino
count	4.00	8.00	2.00	9.00	6.00	5.00	4.00	9.00	6.00	1.00	8.
mean	141.75	-397.62	108.50	218.78	-612.83	93.20	97.75	-311.89	172.00	420.00	-606.
std	169.07	413.11	89.80	301.13	642.01	203.30	181.76	461.73	289.91	NaN	1001.
min	2.00	-1000.00	45.00	-125.00	-1856.00	-16.00	-49.00	-1300.00	0.00	420.00	-2970.

State	Alabama	Arizona	Arkansas	California	Colorado	Connecticut	Delaware	Florida	Georgia	Idaho	Illino
25%	8.00	-632.75	76.75	-69.00	-543.50	-4.00	-40.00	-350.00	20.75	420.00	-572.2
50%	105.00	-323.00	108.50	272.00	-487.50	8.00	51.00	-141.00	64.50	420.00	-259.5
75%	238.75	-26.00	140.25	454.00	-335.50	22.00	188.75	-1.00	122.50	420.00	-26.2
max	355.00	4.00	172.00	690.00	1.00	456.00	338.00	17.00	755.00	420.00	12.0

In [23]:

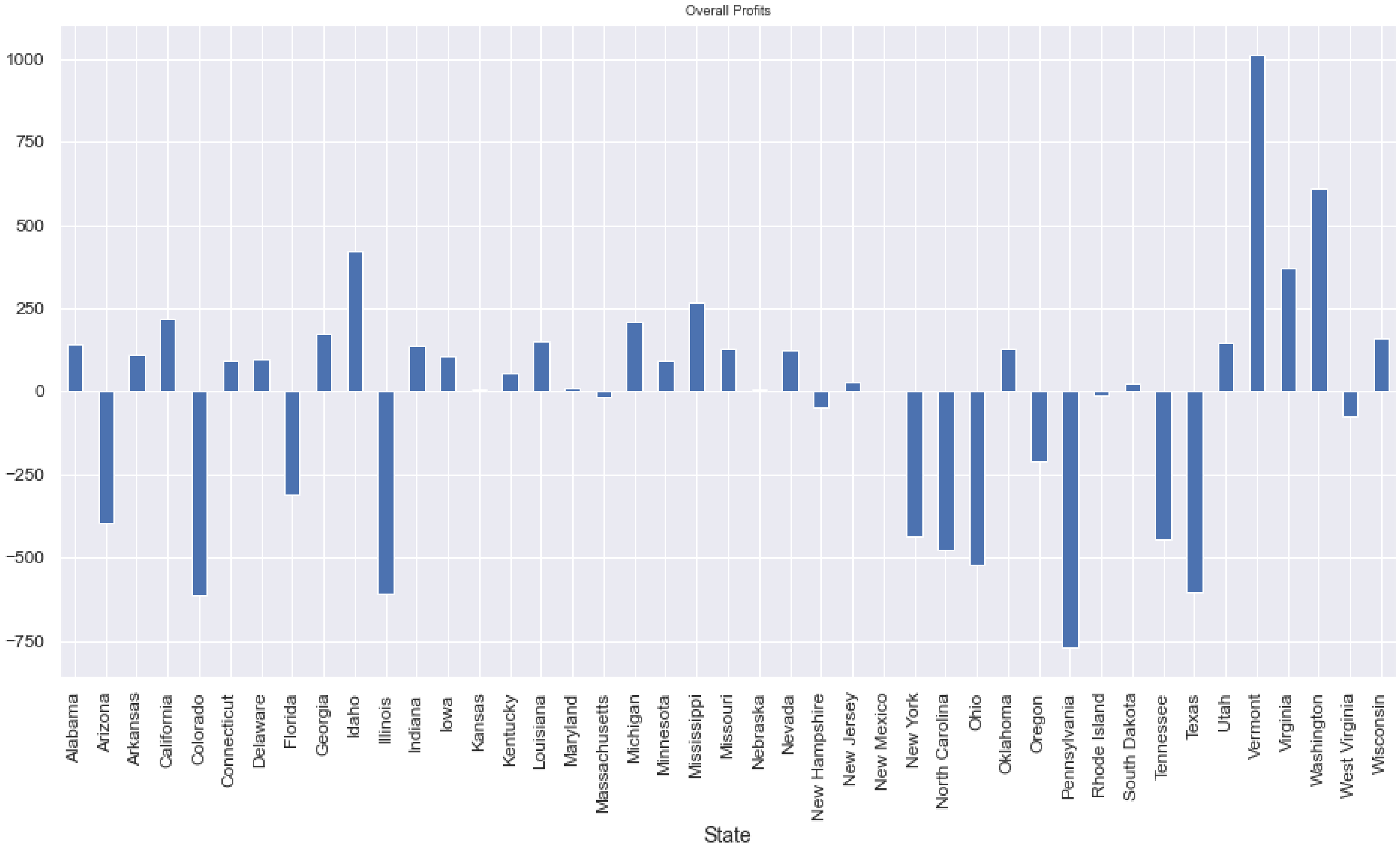
```
table1.mean()
```

Out[23]:

```
State
Alabama            141.75
Arizona           -397.62
Arkansas           108.50
California         218.78
Colorado          -612.83
Connecticut         93.20
Delaware            97.75
Florida           -311.89
Georgia            172.00
Idaho              420.00
Illinois          -606.38
Indiana            137.67
Iowa               107.00
Kansas               7.00
Kentucky            55.25
Louisiana          152.67
Maryland            10.25
Massachusetts      -15.75
Michigan           209.67
Minnesota           93.67
Mississippi        268.50
Missouri           128.25
Nebraska             5.00
Nevada             125.75
New Hampshire      -48.50
New Jersey          30.50
New Mexico           2.00
New York          -435.22
North Carolina    -476.25
Ohio              -519.75
Oklahoma           128.20
Oregon            -208.62
Pennsylvania      -771.44
Rhode Island       -13.20
South Dakota        23.50
Tennessee         -443.67
Texas             -604.56
Utah               147.50
Vermont           1013.00
Virginia           370.62
Washington         610.33
West Virginia      -77.00
Wisconsin          160.40
dtype: float64
```

In [24]:

```
table1.mean().plot(kind="bar", figsize=(16,8), title="Overall Profits")
plt.show()
```



In [25]:

```
table2 = table1.mean()
```

In [26]:

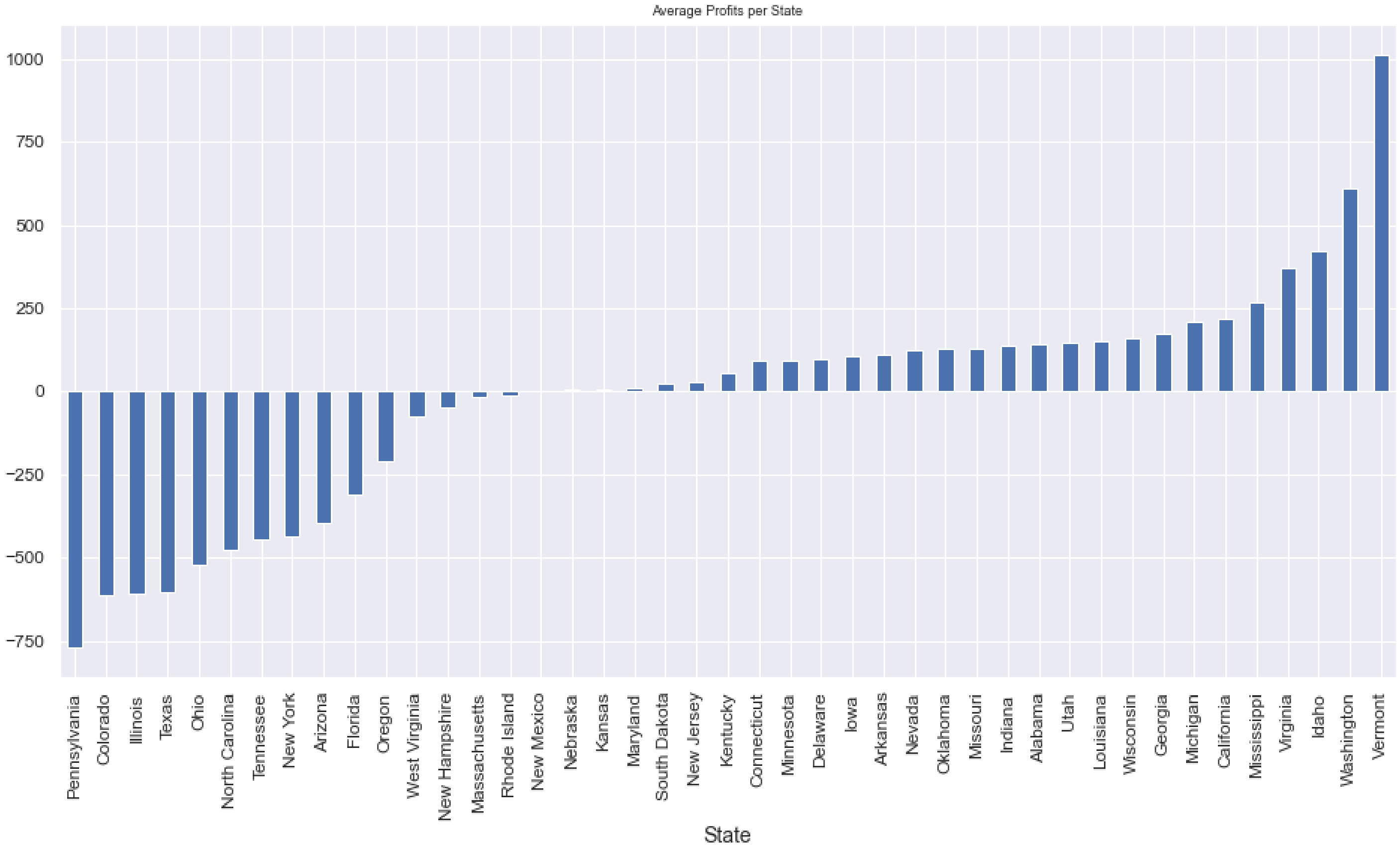
```
table2.sort_values().head()
```

Out[26]:

```
State
Pennsylvania   -771.44
Colorado       -612.83
Illinois       -606.38
Texas          -604.56
Ohio           -519.75
dtype: float64
```

In [27]:

```
table2.sort_values().plot(kind="bar", figsize=(16,8), title="Average Profits per State")
plt.show()
```



Answer Your Managers Questions

How does your visualization leverage at least one “pop-out effect” or “pre-attentive attribute?” Which one(s) was (were) chosen and why?

Using bar chart horizontally to highlight the profits obtained

How does your visualization utilize at least one Gestalt principle? Which principle(s) is (are) being reflected, and how?

Principle of Enclosure, used to differentiate sub-categories, states, segments. The bar plots showed how.

How does your design reflect an understanding of cognitive load and clutter?

I simplify the axis description, big fonts and clear distinct coloring.

Is your visualization static or interactive? Why did you choose that format?

I used static as the figures are unchanging. Interactivity requires a different software.

What need does this visualization address that words or numbers alone cannot fill?

The dataset is big, 9000 plus observations which make it hard for readers to make and figure out. With proper visualizations, we can break into small parts to tell the data story.

Python code done by Dennis Lam