In [1]:
```
!mamba install pandas==3.3.1
!mamba install numpy==1.21.2
!mamba install scipy==1.7.1-y
!mamba install seaborn==0.9.0-y
!mamba install sklearn==0.20.1-y
```

```
'mamba' is not recognized as an internal or external command,
operable program or batch file.
'mamba' is not recognized as an internal or external command,
operable program or batch file.
'mamba' is not recognized as an internal or external command,
operable program or batch file.
'mamba' is not recognized as an internal or external command,
operable program or batch file.
'mamba' is not recognized as an internal or external command,
operable program or batch file.
```

In [6]:
```
import pandas as pd
import numpy as np
import seaborn as sns
```

In [7]:
```
import matplotlib.pyplot as plt
%matplotlib inline
```

In [8]:
```
from scipy import stats
from sklearn.linear_model import LinearRegression, Ridge
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import PolynomialFeatures,StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score, mean_squared_error
```

In [9]:
```
store = pd.read_csv("SampleSuperstore.csv")
```

In [10]:
```
store.head()
```

Out[10]:

	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 [11]:
```
store.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 [13]:
```
store.shape
```

Out[13]:
```
(9994, 13)
```

In [15]:
```
store.describe()
```

Out[15]:

	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 [24]:
```
store.isna().sum()
```

Out[24]:
```
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 [27]:
```
store.duplicated()
```

Out[27]:
```
0       False
1       False
2       False
3       False
4       False
        ...
9989    False
9990    False
9991    False
9992    False
9993    False
Length: 9994, dtype: bool
```

In [58]:
```
store['Quantity'].value_counts()
```

Out[58]:
```
3     2409
2     2402
5     1230
4     1191
1      899
7      606
6      572
9      258
8      257
10      57
11      34
14      29
```

```
13      27
12      23
Name: Quantity, dtype: int64
```

In [41]:
```
store['City'].nunique()
```

Out[41]:
```
531
```

In [36]:
```
store['State'].value_counts()
```

Out[36]:
```
California              2001
New York                1128
Texas                    985
Pennsylvania             587
Washington               506
Illinois                 492
Ohio                     469
Florida                  383
Michigan                 255
North Carolina           249
Arizona                  224
Virginia                 224
Georgia                  184
Tennessee                183
Colorado                 182
Indiana                  149
Kentucky                 139
Massachusetts            135
New Jersey               130
Oregon                   124
Wisconsin                110
Maryland                 105
Delaware                  96
Minnesota                 89
Connecticut               82
Oklahoma                  66
Missouri                  66
Alabama                   61
Arkansas                  60
Rhode Island              56
Utah                      53
Mississippi               53
Louisiana                 42
South Carolina            42
Nevada                    39
Nebraska                  38
New Mexico                37
Iowa                      30
```

```
New Hampshire             27
Kansas                    24
Idaho                     21
Montana                   15
South Dakota              12
Vermont                   11
District of Columbia      10
Maine                      8
North Dakota               7
West Virginia              4
Wyoming                    1
Name: State, dtype: int64
```

In [37]:
```
store['Region'].value_counts()
```

Out[37]:
```
West       3203
East       2848
Central    2323
South      1620
Name: Region, dtype: int64
```

In [19]:
```
store['Sales'].describe()
```

Out[19]:
```
count     9994.000000
mean       229.858001
std        623.245101
min          0.444000
25%         17.280000
50%         54.490000
75%        209.940000
max      22638.480000
Name: Sales, dtype: float64
```

In [20]:
```
store['Sales'].mean()
```

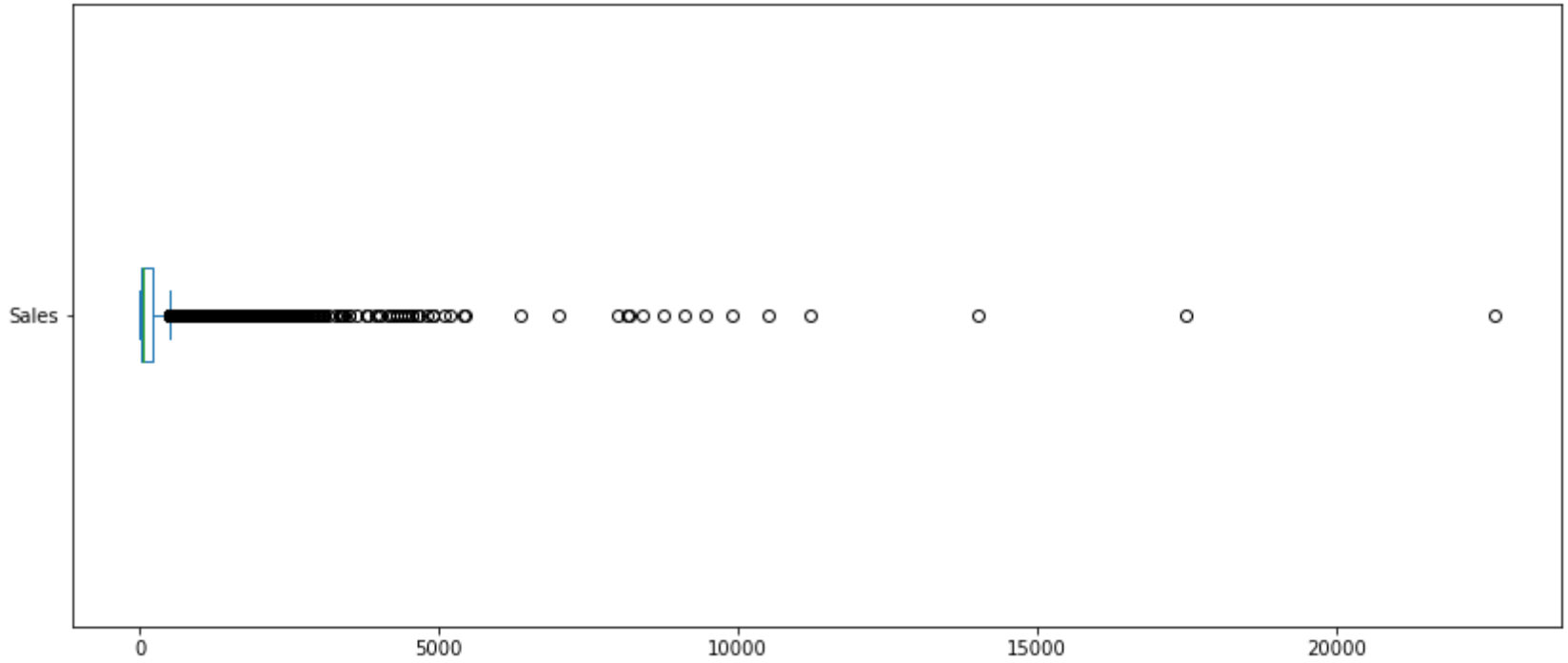
Out[20]:
```
229.8580008304938
```

In [21]:
```
store['Sales'].median()
```

Out[21]:
```
54.489999999999995
```

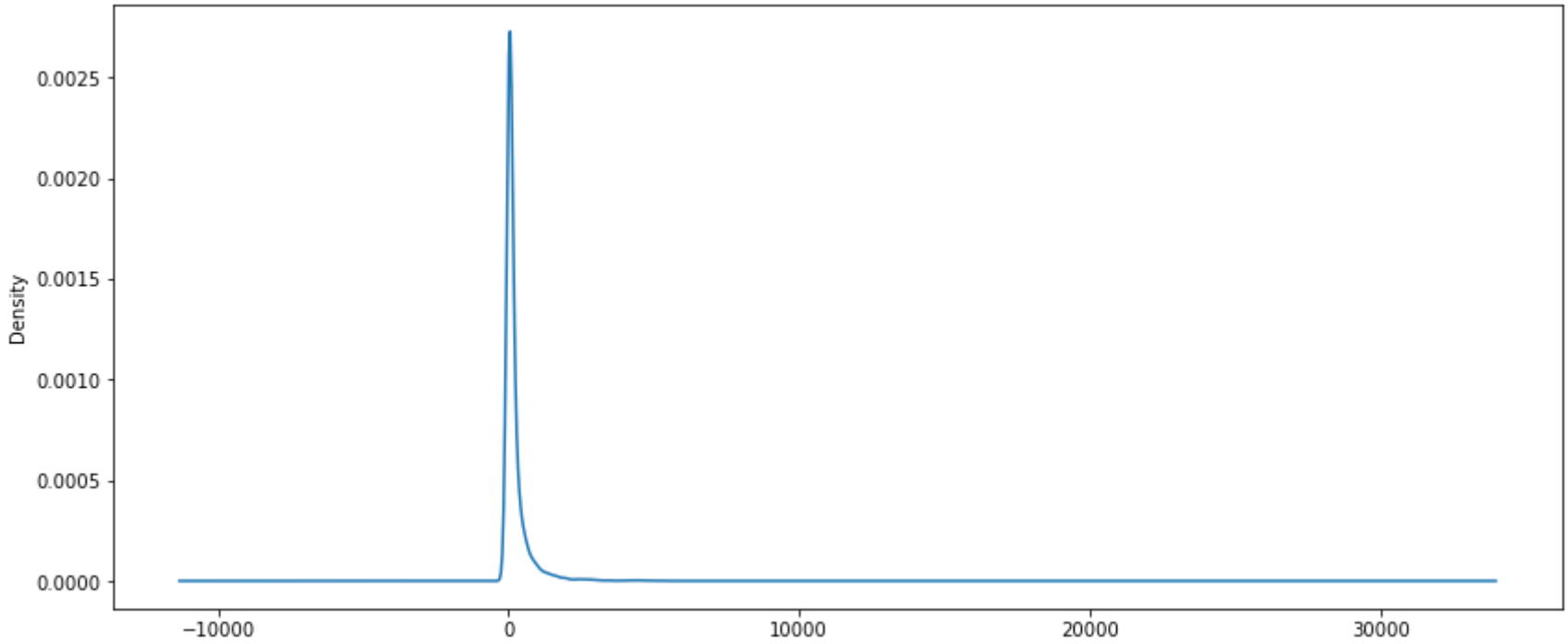
In [22]:
```
store['Sales'].plot(kind='box', vert=False, figsize=(14,6))
```

Out[22]:
```
<AxesSubplot:>
```



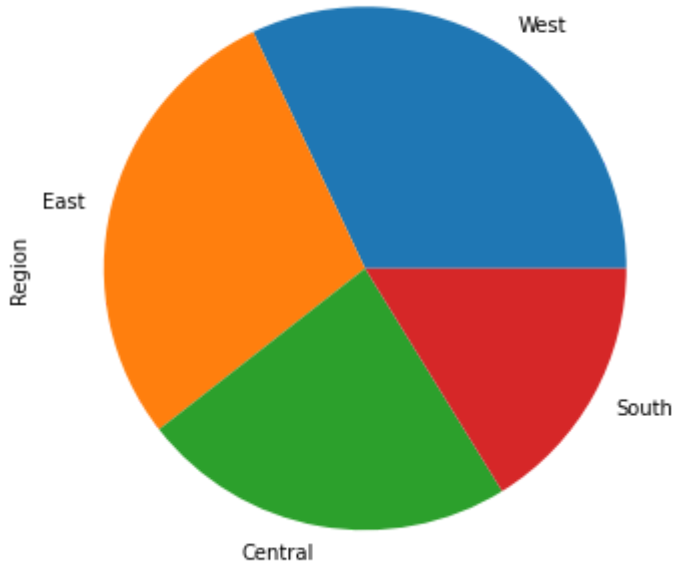
In [23]:
```
store['Sales'].plot(kind='density', figsize=(14,6))
```

Out[23]:
```
<AxesSubplot:ylabel='Density'>
```

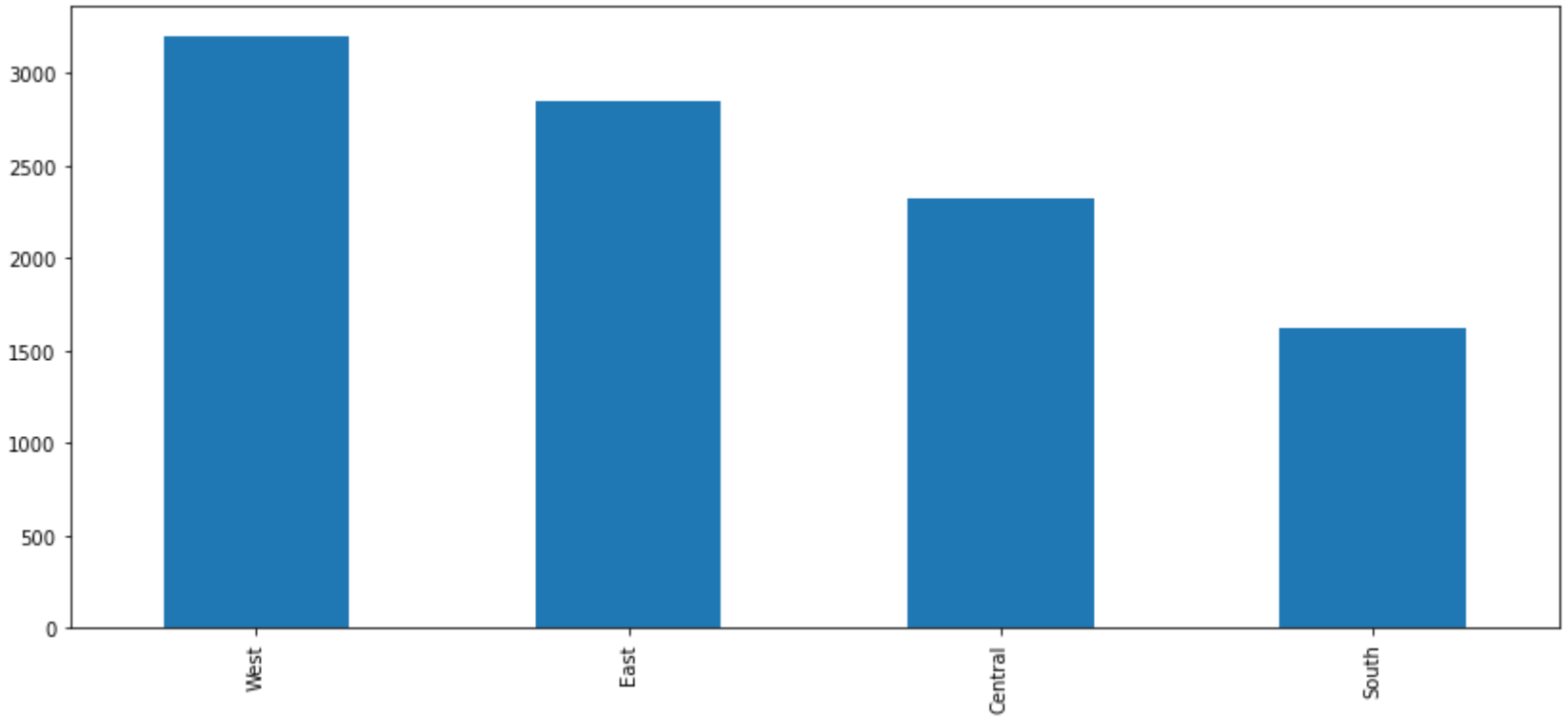


In [53]:
```
store['Region'].value_counts().plot(kind='pie', figsize=(6,6))
```

Out[53]:
```
<AxesSubplot:ylabel='Region'>
```



In [55]:
```
ax = store['Region'].value_counts().plot(kind='bar', figsize=(14,6))
```



In [59]:
```
store['Category'].value_counts()
```

Out[59]:
```
Office Supplies    6026
Furniture          2121
Technology         1847
Name: Category, dtype: int64
```

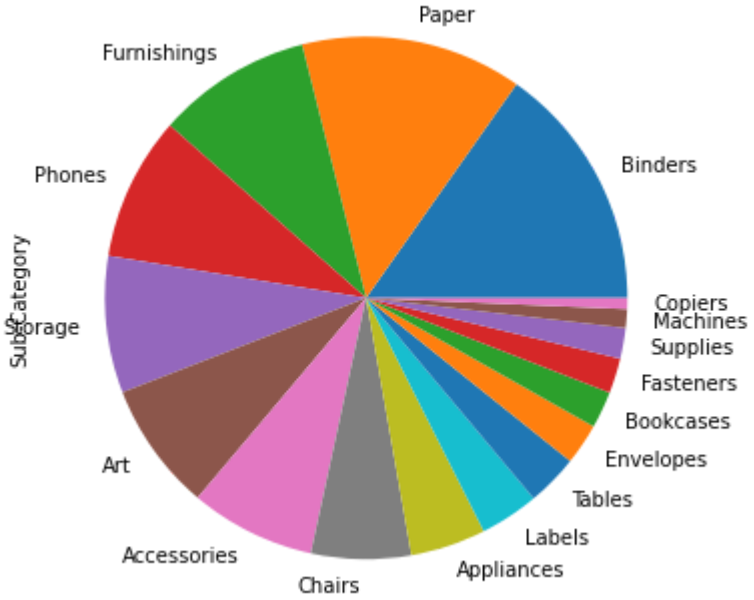
In [61]:
```
store['Sub-Category'].value_counts()
```

Out[61]:
```
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 [63]:
```
store['Sub-Category'].value_counts().plot(kind='pie', figsize=(14,6))
```

Out[63]:
```
<AxesSubplot:ylabel='Sub-Category'>
```



In [64]:
```
store.sample(10)
```

Out[64]:

	Ship Mode	Segment	Country	City	State	Postal Code	Region	Category	Sub-Category	Sales	Quantity	Discount	Profit
6203	Standard Class	Consumer	United States	Columbus	Ohio	43229	East	Technology	Phones	235.152	8	0.4	-47.0304
1098	First Class	Home Office	United States	San Francisco	California	94122	West	Office Supplies	Art	7.040	4	0.0	2.0416
9377	Standard Class	Corporate	United States	Eau Claire	Wisconsin	54703	Central	Office Supplies	Storage	32.560	2	0.0	8.4656
422	Standard Class	Corporate	United States	Lawrence	Massachusetts	1841	East	Furniture	Furnishings	56.560	4	0.0	14.7056
4980	Second Class	Corporate	United States	Chicago	Illinois	60610	Central	Office Supplies	Appliances	4.356	2	0.8	-11.7612

1501	Standard Class	Consumer	United States	Austin	Texas	78745	Central	Office Supplies	Storage	540.048	3	0.2	-47.2542
3446	First Class	Corporate	United States	New York City	New York	10024	East	Office Supplies	Paper	70.880	2	0.0	33.3136
4852	Standard Class	Consumer	United States	Pasco	Washington	99301	West	Office Supplies	Storage	485.880	6	0.0	19.4352
2763	Standard Class	Home Office	United States	Philadelphia	Pennsylvania	19140	East	Office Supplies	Binders	10.332	3	0.7	-7.5768
1163	Standard Class	Home Office	United States	New York City	New York	10035	East	Office Supplies	Binders	125.760	3	0.2	40.8720

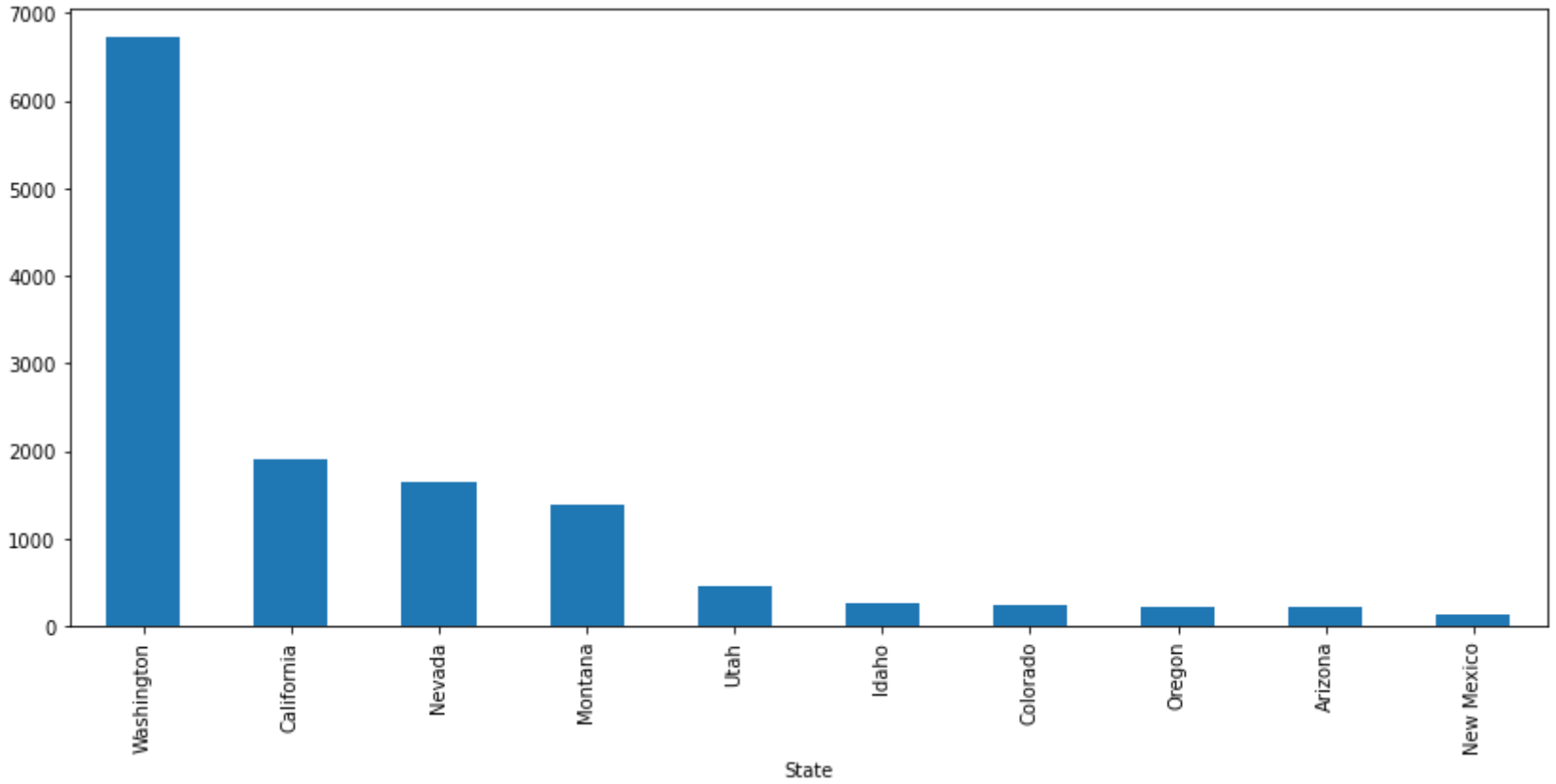
In [57]:
```
store.corr()
```

Out[57]:

	Postal Code	Sales	Quantity	Discount	Profit
Postal Code	1.000000	-0.023854	0.012761	0.058443	-0.029961
Sales	-0.023854	1.000000	0.200795	-0.028190	0.479064
Quantity	0.012761	0.200795	1.000000	0.008623	0.066253
Discount	0.058443	-0.028190	0.008623	1.000000	-0.219487
Profit	-0.029961	0.479064	0.066253	-0.219487	1.000000

In [97]:
```
store[store['Region'] == 'West'].groupby(['State']).max()['Profit'].nlargest(10).plot(kind='bar', figsize=(14,6))
```

Out[97]:
```
<AxesSubplot:xlabel='State'>
```



In [100…
```
store.to_csv('new store.csv')
```

In [ ]: