In [1]:
```
#importing Libraries
```

In [4]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [5]:
```
#Importing & Inspecting Data
```

In [9]:
```
startups =  pd.read_excel('startup-expansion.xlsx')
```

In [10]:
```
startups
```

Out[10]:

	Store ID	City	State	Sales Region	New Expansion	Marketing Spend	Revenue
0	1	Peoria	Arizona	Region 2	Old	2601	48610
1	2	Midland	Texas	Region 2	Old	2727	45689
2	3	Spokane	Washington	Region 2	Old	2768	49554
3	4	Denton	Texas	Region 2	Old	2759	38284
4	5	Overland Park	Kansas	Region 2	Old	2869	59887
...	...	...	...	...	...	...	...
145	146	Paterson	New Jersey	Region 1	New	2251	34603
146	147	Brownsville	Texas	Region 2	New	3675	63148
147	148	Rockford	Illinois	Region 1	New	2648	43377
148	149	College Station	Texas	Region 2	New	2994	22457
149	150	Thousand Oaks	California	Region 2	New	2431	40141

150 rows × 7 columns

In [13]:
```
startups.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 7 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Store ID         150 non-null    int64
 1   City             150 non-null    object
 2   State            150 non-null    object
 3   Sales Region     150 non-null    object
 4   New Expansion    150 non-null    object
 5   Marketing Spend  150 non-null    int64
 6   Revenue          150 non-null    int64
dtypes: int64(3), object(4)
memory usage: 8.3+ KB
```

In [32]:
```
startups[['Marketing Spend','Revenue']].describe().round(2)
```

Out[32]:

	Marketing Spend	Revenue
count	150.00	150.00
mean	2893.15	39301.43
std	367.86	15465.75
min	1811.00	15562.00
25%	2662.25	21113.50
50%	2898.00	42993.00
75%	3111.50	51145.50
max	3984.00	68828.00

In [17]:
```
#PreProcessing Data
```

In [20]:
```
startups['City'].unique()
```

Out[20]:
```
array(['Peoria', 'Midland', 'Spokane', 'Denton', 'Overland Park',
       'Yonkers', 'Birmingham', 'Antioch', 'Worcester', 'Rochester',
       'Rialto', 'Santa Maria', 'Las Cruces', 'Jackson', 'Hillsboro',
       'Temecula', 'Tallahassee', 'Fontana', 'Kent', 'Broken Arrow',
       'Concord', 'Modesto', 'Montgomery', 'Burbank', 'Elk Grove',
       'Port St. Lucie', 'Elizabeth', 'Salt Lake City', 'Waco', 'Edison',
       'Boulder', 'Grand Rapids', 'Tyler', 'Charleston', 'Huntsville',
       'Pearland', 'Inglewood', 'Oxnard', 'Miramar', 'Cape Coral',
       'Syracuse', 'Newport News', 'Lewisville', 'Carrollton',
       'San Bernardino', 'Pasadena', 'Roseville', 'Murrieta',
       'San Angelo', 'Olathe', 'Akron', 'Fullerton', 'Manchester',
       'Everett', 'West Covina', 'Thornton', 'Hampton', 'Waterbury',
       'Ventura', 'Davenport', 'Columbia', 'Simi Valley', 'Richmond',
       'Little Rock', 'El Cajon', 'Santa Clara', 'Oceanside', 'Davie',
       'Lakeland', 'Centennial', 'Lowell', 'Ontario', 'Palm Bay',
       'Murfreesboro', 'Vancouver', 'Topeka', 'West Valley City',
       'New Haven', 'Pueblo', 'Costa Mesa', 'Garden Grove',
       'Fort Lauderdale', 'North Charleston', 'Cambridge', 'Greeley',
       'Gresham', 'Amarillo', 'High Point', 'Vista', 'Tacoma', 'Mesquite',
       'Augusta', 'Elgin', 'Aurora', 'Gainesville', 'Dayton',
       'Wichita Falls', 'Naperville', 'Clovis', 'Billings', 'Surprise',
       'Coral Springs', 'Visalia', 'Killeen', 'Orange', 'Richardson',
       'South Bend', 'Fayetteville', 'Sioux Falls', 'Grand Prairie',
       'Stamford', 'West Palm Beach', 'Knoxville', 'Renton', 'McAllen',
       'Woodbridge', 'Shreveport', 'Bellevue', 'Huntington Beach',
       'Santa Clarita', 'Sterling Heights', 'Mobile', 'Bridgeport',
       'Daly City', 'Sandy Springs', 'Cedar Rapids', 'Columbus',
       'Moreno Valley', 'Pompano Beach', 'Savannah', 'West Jordan',
       'Des Moines', 'Green Bay', 'Santa Rosa', 'San Mateo', 'Warren',
       'Norwalk', 'Lafayette', 'Providence', 'Chattanooga', 'Tempe',
       'Joliet', 'Rancho Cucamonga', 'Glendale', 'Paterson',
       'Brownsville', 'Rockford', 'College Station', 'Thousand Oaks'],
      dtype=object)
```

In [24]:
```
startups['City'].nunique()
```

Out[24]:
```
149
```

In [18]:
```
startups['City'].value_counts()
```

Out[18]:
```
Rochester          2
College Station    1
Greeley            1
Coral Springs      1
Santa Clarita      1
                  ..
Costa Mesa         1
Bellevue           1
Thornton           1
Vancouver          1
Ontario            1
Name: City, Length: 149, dtype: int64
```

In [21]:
```
startups['State'].unique()
```

Out[21]:
```
array(['Arizona', 'Texas', 'Washington', 'Kansas', 'New York', 'Alabama',
       'California', 'Massachusetts', 'New Mexico', 'Mississippi',
       'Oregon', 'Florida', 'Oklahoma', 'New Jersey', 'Utah', 'Colorado',
       'Michigan', 'South Carolina', 'Virginia', 'Ohio', 'New Hampshire',
       'Connecticut', 'Iowa', 'Arkansas', 'Tennessee', 'North Carolina',
       'Georgia', 'Illinois', 'Montana', 'Indiana', 'South Dakota',
       'Louisiana', 'Minnesota', 'Wisconsin', 'Rhode Island'],
      dtype=object)
```

In [23]:
```
startups['State'].nunique()
```

Out[23]:
```
35
```

In [19]:
```
startups['State'].value_counts()
```

Out[19]:
```
California        40
Texas             17
Florida           12
Washington         7
Colorado           5
Illinois           5
Connecticut        4
Alabama            4
New Jersey         4
Georgia            4
Michigan           3
Massachusetts      3
Tennessee          3
Iowa               3
Arizona            3
Utah               3
South Carolina     3
New York           3
Kansas             3
Ohio               2
Louisiana          2
Virginia           2
North Carolina     2
Oregon             2
Montana            1
New Hampshire      1
Arkansas           1
Mississippi        1
Indiana            1
Oklahoma           1
Wisconsin          1
South Dakota       1
Minnesota          1
Rhode Island       1
New Mexico         1
Name: State, dtype: int64
```

In [22]:

In [27]:
```
startups['Sales Region'].unique()
```

Out[27]:
```
array(['Region 2', 'Region 1'], dtype=object)
```

In [29]:
```
startups['Sales Region'].value_counts()
```

Out[29]:
```
Region 2    86
Region 1    64
Name: Sales Region, dtype: int64
```

In [30]:
```
startups['New Expansion'].unique()
```

Out[30]:
```
array(['Old', 'New'], dtype=object)
```

In [31]:
```
startups['New Expansion'].value_counts()
```

Out[31]:
```
Old    140
New     10
Name: New Expansion, dtype: int64
```

In [33]:
```
startups.isna().sum()
```

Out[33]:
```
Store ID           0
City               0
State              0
Sales Region       0
New Expansion      0
Marketing Spend    0
Revenue            0
dtype: int64
```

In [34]:
```
Startups.duplicated().sum()
```

Out[34]:
```
0
```

In [35]:
```
#Exploring&Analysing Data
```

In [36]:
```
startups.sample(10)
```

Out[36]:

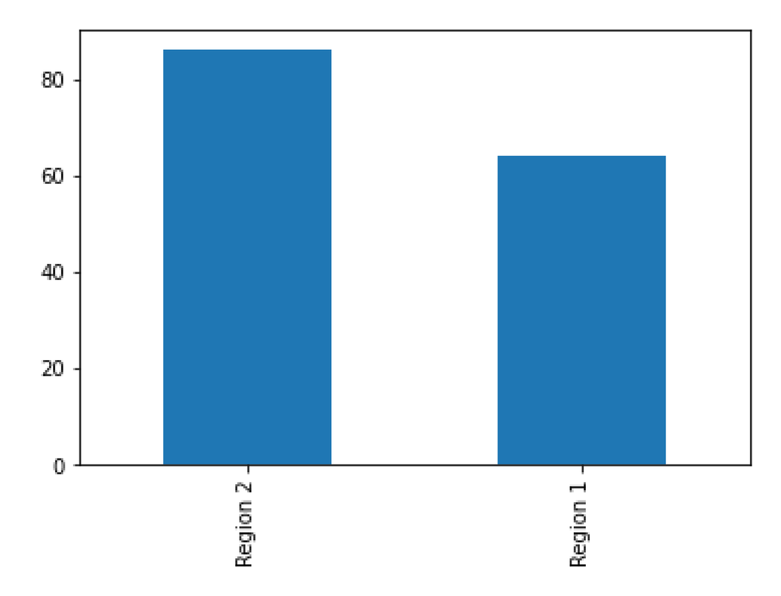
	Store ID	City	State	Sales Region	New Expansion	Marketing Spend	Revenue
112	113	Knoxville	Tennessee	Region 2	Old	3086	56504
122	123	Bridgeport	Connecticut	Region 1	Old	2914	47108
1	2	Midland	Texas	Region 2	Old	2727	45689
116	117	Shreveport	Louisiana	Region 1	Old	3081	56140
132	133	Des Moines	Iowa	Region 1	Old	2995	57432
33	34	Charleston	South Carolina	Region 1	Old	2484	34829
92	93	Elgin	Illinois	Region 1	Old	2553	18215
79	80	Costa Mesa	California	Region 2	Old	2275	41361
22	23	Montgomery	Alabama	Region 1	Old	3287	52114
113	114	Renton	Washington	Region 2	Old	2754	44635

In [ ]:

In [42]:

In [51]:
```
startups['Sales Region'].value_counts().plot.bar()
```

Out[51]:
```
<AxesSubplot:>
```



In [53]:
```
startups.groupby('New Expansion').groups
```

Out[53]:
```
{'New': [140, 141, 142, 143, 144, 145, 146, 147, 148, 149], 'Old': [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 3
0, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 7
5, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, ...]}
```

In [55]:
```
startups[startups['New Expansion'] == 'Old']
```

Out[55]:

	Store ID	City	State	Sales Region	New Expansion	Marketing Spend	Revenue
0	1	Peoria	Arizona	Region 2	Old	2601	48610
1	2	Midland	Texas	Region 2	Old	2727	45689
2	3	Spokane	Washington	Region 2	Old	2768	49554
3	4	Denton	Texas	Region 2	Old	2759	38284
4	5	Overland Park	Kansas	Region 2	Old	2869	59887
...	...	...	...	...	...	...	...
135	136	San Mateo	California	Region 2	Old	1811	19426
136	137	Warren	Michigan	Region 1	Old	2736	47262
137	138	Norwalk	California	Region 2	Old	3112	19703
138	139	Lafayette	Louisiana	Region 1	Old	2603	40255
139	140	Providence	Rhode Island	Region 1	Old	3191	62337

140 rows × 7 columns

In [56]:
```
startups[startups['New Expansion'] == 'New']
```

Out[56]:

	Store ID	City	State	Sales Region	New Expansion	Marketing Spend	Revenue
140	141	Chattanooga	Tennessee	Region 2	New	3587	55357
141	142	Tempe	Arizona	Region 2	New	2911	48954
142	143	Joliet	Illinois	Region 1	New	3279	48315
143	144	Rancho Cucamonga	California	Region 2	New	2945	52366
144	145	Glendale	California	Region 2	New	2363	49376
145	146	Paterson	New Jersey	Region 1	New	2251	34603
146	147	Brownsville	Texas	Region 2	New	3675	63148
147	148	Rockford	Illinois	Region 1	New	2648	43377
148	149	College Station	Texas	Region 2	New	2994	22457
149	150	Thousand Oaks	California	Region 2	New	2431	40141

In [61]:
```
startups[startups['New Expansion'] == 'Old'].groupby('State').sum()['Revenue'].nlargest(10)
```

Out[61]:
```
State
California     1362468
Texas           554964
Florida         479023
Washington      298013
Alabama         221025
New York        160046
Connecticut     158511
Georgia         157656
Colorado        156495
Michigan        147759
Name: Revenue, dtype: int64
```

In [62]:
```
startups[startups['New Expansion'] == 'New'].groupby('State').sum()['Revenue'].nlargest(10)
```

Out[62]:
```
State
California    141883
Illinois       91692
Texas          85605
Tennessee      55357
Arizona        48954
New Jersey     34603
Name: Revenue, dtype: int64
```

In [66]:
```
startups['Profit'] = startups['Revenue'] - startups['Marketing Spend']#lina zene colun, okhraa fiiha el profit
```

In [67]:
```
startups
```

Out[67]:

	Store ID	City	State	Sales Region	New Expansion	Marketing Spend	Revenue	Profit
0	1	Peoria	Arizona	Region 2	Old	2601	48610	46009
1	2	Midland	Texas	Region 2	Old	2727	45689	42962
2	3	Spokane	Washington	Region 2	Old	2768	49554	46786
3	4	Denton	Texas	Region 2	Old	2759	38284	35525
4	5	Overland Park	Kansas	Region 2	Old	2869	59887	57018
...	...	...	...	...	...	...	...	...
145	146	Paterson	New Jersey	Region 1	New	2251	34603	32352
146	147	Brownsville	Texas	Region 2	New	3675	63148	59473
147	148	Rockford	Illinois	Region 1	New	2648	43377	40729
148	149	College Station	Texas	Region 2	New	2994	22457	19463
149	150	Thousand Oaks	California	Region 2	New	2431	40141	37710

150 rows × 8 columns

In [70]:
```
round((startups['Profit'] / startups['Marketing Spend'])*100,2)
```

Out[70]:
```
0      1768.90
1      1575.43
2      1690.25
3      1287.60
4      1987.38
        ...
145    1437.23
146    1618.31
147    1538.10
148     650.07
149    1551.21
Length: 150, dtype: float64
```

In [72]:
```
startups['ROMS'] = round((startups['Profit'] / startups['Marketing Spend'])*100,2)
```

In [74]:
```
startups['%ROMS']=startups['ROMS']/100
```

In [75]:
```
startups
```

Out[75]:

	Store ID	City	State	Sales Region	New Expansion	Marketing Spend	Revenue	Profit	ROMS	%ROMS
0	1	Peoria	Arizona	Region 2	Old	2601	48610	46009	1768.90	17.6890
1	2	Midland	Texas	Region 2	Old	2727	45689	42962	1575.43	15.7543
2	3	Spokane	Washington	Region 2	Old	2768	49554	46786	1690.25	16.9025
3	4	Denton	Texas	Region 2	Old	2759	38284	35525	1287.60	12.8760
4	5	Overland Park	Kansas	Region 2	Old	2869	59887	57018	1987.38	19.8738
...	...	...	...	...	...	...	...	...	...	...
145	146	Paterson	New Jersey	Region 1	New	2251	34603	32352	1437.23	14.3723
146	147	Brownsville	Texas	Region 2	New	3675	63148	59473	1618.31	16.1831
147	148	Rockford	Illinois	Region 1	New	2648	43377	40729	1538.10	15.3810
148	149	College Station	Texas	Region 2	New	2994	22457	19463	650.07	6.5007
149	150	Thousand Oaks	California	Region 2	New	2431	40141	37710	1551.21	15.5121

150 rows × 10 columns

In [76]:
```
startups.to_csv('startup-expansion-modified.csv')
```

In [ ]: