Importing Libraries (Toolkit)

In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

In [ ]:

```

```

Importing & Inspecting Data

In [2]:

```
startups = pd.read_excel('startup-expansion.xlsx')
startups
```

Out[2]:

	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 [3]:

```
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 [6]:

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

Out[6]:

	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 [ ]:

Preprocessing Data

In [7]:

```
startups['City'].unique()
```

```
Out[7]: 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 [8]: startups['City'].value_counts()
```

```
Out[8]: City
        Rochester          2
        Midland            1
        Spokane            1
        Denton             1
        Peoria             1
                          ..
        Paterson           1
        Brownsville        1
        Rockford           1
        College Station    1
        Thousand Oaks      1
        Name: count, Length: 149, dtype: int64
```

```
In [13]: startups['City'].unique()
```

```
Out[13]: 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 [15]: startups['City'].nunique()
```

```
Out[15]: 149
```

```
In [11]: startups['State'].unique()
```

```
Out[11]: 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 [12]: startups['State'].nunique()
```

```
Out[12]: 35
```

```
In [9]: startups['State'].value_counts()
```

```
Out[9]:  State
         California        40
         Texas             17
         Florida           12
         Washington         7
         Colorado           5
         Illinois           5
         New Jersey         4
         Connecticut        4
         Georgia            4
         Alabama            4
         Arizona            3
         South Carolina     3
         Michigan           3
         Utah               3
         Iowa               3
         Tennessee          3
         Massachusetts      3
         New York           3
         Kansas             3
         Oregon             2
         North Carolina     2
         Louisiana          2
         Virginia           2
         Ohio               2
         Oklahoma           1
         New Mexico         1
         Mississippi        1
         Arkansas           1
         New Hampshire      1
         Indiana            1
         Montana            1
         South Dakota       1
         Minnesota          1
         Wisconsin          1
         Rhode Island       1
         Name: count, dtype: int64
```

```
In [17]:  startups['Sales Region'].unique()
```

```
Out[17]:  array(['Region 2', 'Region 1'], dtype=object)
```

```
In [18]:  startups['Sales Region'].nunique()
```

```
Out[18]:  2
```

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

```
Out[19]:  Sales Region
          Region 2    86
          Region 1    64
          Name: count, dtype: int64
```

```
In [23]:  startups['New Expansion'].value_counts()
```

Out[23]:

```
New Expansion
Old    140
New     10
Name: count, dtype: int64
```

In [24]:

```python
startups.isna().sum()
```

Out[24]:

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

In [25]:

```python
startups.duplicated().sum()
```

Out[25]:

```
np.int64(0)
```

In [ ]:

```python
```

Exploring & Analysing Data

In [26]:

```python
startups.sample(10)
```

Out[26]:

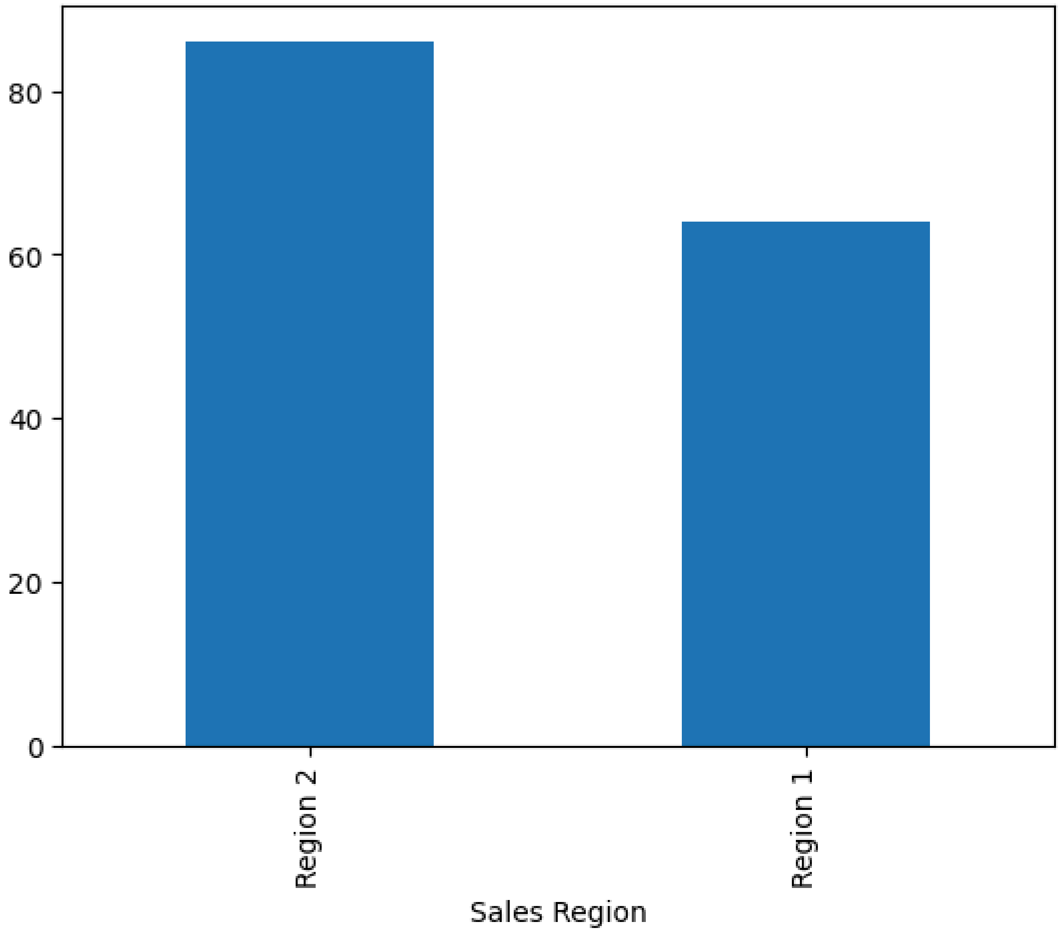
	Store ID	City	State	Sales Region	New Expansion	Marketing Spend	Revenue
76	77	West Valley City	Utah	Region 2	Old	2555	49609
30	31	Boulder	Colorado	Region 2	Old	3083	22680
65	66	Santa Clara	California	Region 2	Old	2462	29008
39	40	Cape Coral	Florida	Region 1	Old	2886	52250
44	45	San Bernardino	California	Region 2	Old	3399	59870
101	102	Coral Springs	Florida	Region 1	Old	3079	41319
66	67	Oceanside	California	Region 2	Old	3084	55684
142	143	Joliet	Illinois	Region 1	New	3279	48315
16	17	Tallahassee	Florida	Region 1	Old	2737	47729
5	6	Yonkers	New York	Region 1	Old	3080	53827

In [28]:

```python
startups['Sales Region'].value_counts().plot.bar()
```

Out[28]:

```
<Axes: xlabel='Sales Region'>
```



```
In [29]: startups.groupby('New Expansion').groups
```

```
Out[29]: {'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, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 4
         7, 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, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 8
         8, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, ...]}
```

```
In [30]: startups[startups['New Expansion'] == 'New']
```

Out[30]:

	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 [31]:

```
startups[startups['New Expansion'] == 'Old']
```

Out[31]:

	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 [32]:

```
startups[startups['New Expansion'] == 'Old'].groupby('City').max()['Revenue'].nlarg
```

Out[32]:

```
City
Little Rock         68828
Grand Rapids        65475
Rochester           64906
Oxnard              64302
Fontana             63027
Providence          62337
Birmingham          60338
Overland Park       59887
San Bernardino      59870
Worcester           59840
Name: Revenue, dtype: int64
```

In [33]:

```
startups[startups['New Expansion'] == 'New'].groupby('City').max()['Revenue'].nlarg
```

Out[33]:

```
City
Brownsville           63148
Chattanooga           55357
Rancho Cucamonga      52366
Glendale              49376
Tempe                 48954
Joliet                48315
Rockford              43377
Thousand Oaks         40141
Paterson              34603
College Station       22457
Name: Revenue, dtype: int64
```

In [61]:

```
startups['ROM'] = round((startups['Revenue'] / startups['Marketing Spend']) * 100,2
startups['ROM']
```

Out[61]:

```
0      1868.90
1      1675.43
2      1790.25
3      1387.60
4      2087.38
        ...
145    1537.23
146    1718.31
147    1638.10
148     750.07
149    1651.21
Name: ROM, Length: 150, dtype: float64
```

In [62]:

```
startups['Profit'] = startups['Revenue'] - startups['Marketing Spend']
startups['Profit']
```

```
Out[62]: 0      46009
         1      42962
         2      46786
         3      35525
         4      57018
                ...
         145    32352
         146    59473
         147    40729
         148    19463
         149    37710
         Name: Profit, Length: 150, dtype: int64
```

```
In [63]: (startups['Revenue'] - startups['Marketing Spend']) / startups['Marketing Spend']
```

```
Out[63]: 0      17.688966
         1      15.754309
         2      16.902457
         3      12.876042
         4      19.873824
                  ...
         145    14.372279
         146    16.183129
         147    15.381042
         148     6.500668
         149    15.512135
         Length: 150, dtype: float64
```

```
In [64]: startups
```

Out[64]:

	Store ID	City	State	Sales Region	New Expansion	Marketing Spend	Revenue	ROM	Pro
0	1	Peoria	Arizona	Region 2	Old	2601	48610	1868.90	460
1	2	Midland	Texas	Region 2	Old	2727	45689	1675.43	429
2	3	Spokane	Washington	Region 2	Old	2768	49554	1790.25	467
3	4	Denton	Texas	Region 2	Old	2759	38284	1387.60	355
4	5	Overland Park	Kansas	Region 2	Old	2869	59887	2087.38	570
...	...	...	...	...	...	...	...	...	
145	146	Paterson	New Jersey	Region 1	New	2251	34603	1537.23	323
146	147	Brownsville	Texas	Region 2	New	3675	63148	1718.31	594
147	148	Rockford	Illinois	Region 1	New	2648	43377	1638.10	407
148	149	College Station	Texas	Region 2	New	2994	22457	750.07	194
149	150	Thousand Oaks	California	Region 2	New	2431	40141	1651.21	377

150 rows × 9 columns

In [65]:

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

In [66]:

```
startups
```

Out[66]:

	Store ID	City	State	Sales Region	New Expansion	Marketing Spend	Revenue	ROM	Pro
0	1	Peoria	Arizona	Region 2	Old	2601	48610	1868.90	460
1	2	Midland	Texas	Region 2	Old	2727	45689	1675.43	429
2	3	Spokane	Washington	Region 2	Old	2768	49554	1790.25	467.
3	4	Denton	Texas	Region 2	Old	2759	38284	1387.60	355.
4	5	Overland Park	Kansas	Region 2	Old	2869	59887	2087.38	570
...	...	...	...	...	...	...	...	...	
145	146	Paterson	New Jersey	Region 1	New	2251	34603	1537.23	323
146	147	Brownsville	Texas	Region 2	New	3675	63148	1718.31	594
147	148	Rockford	Illinois	Region 1	New	2648	43377	1638.10	407.
148	149	College Station	Texas	Region 2	New	2994	22457	750.07	194
149	150	Thousand Oaks	California	Region 2	New	2431	40141	1651.21	377

150 rows × 10 columns

In [68]:

```
startups['ROMS%'] = startups['ROMS'] /100
```

In [69]:

```
startups
```

Out[69]:

	Store ID	City	State	Sales Region	New Expansion	Marketing Spend	Revenue	ROM	Pro
0	1	Peoria	Arizona	Region 2	Old	2601	48610	1868.90	460
1	2	Midland	Texas	Region 2	Old	2727	45689	1675.43	429
2	3	Spokane	Washington	Region 2	Old	2768	49554	1790.25	467.
3	4	Denton	Texas	Region 2	Old	2759	38284	1387.60	355.
4	5	Overland Park	Kansas	Region 2	Old	2869	59887	2087.38	570
...	...	...	...	...	...	...	...	...	
145	146	Paterson	New Jersey	Region 1	New	2251	34603	1537.23	323
146	147	Brownsville	Texas	Region 2	New	3675	63148	1718.31	594
147	148	Rockford	Illinois	Region 1	New	2648	43377	1638.10	407.
148	149	College Station	Texas	Region 2	New	2994	22457	750.07	194
149	150	Thousand Oaks	California	Region 2	New	2431	40141	1651.21	377

150 rows × 11 columns

In [70]:

```
startups.to_csv('start-expansion-modified.csv')
```

In [ ]: