Data Analysis CAPESTONE

Import Librairies

In [9]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

import & Inspect DATA

In [10]:
```
startups = pd.read_excel('startup-expansion.xlsx')
```

In [11]:
```
startups
```

Out[11]:

	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 [12]:
```
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 [14]:
```
startups.describe().round(2)
```

Out[14]:

	Store ID	Marketing Spend	Revenue
count	150.00	150.00	150.00
mean	75.50	2893.15	39301.43
std	43.45	367.86	15465.75
min	1.00	1811.00	15562.00
25%	38.25	2662.25	21113.50
50%	75.50	2898.00	42993.00
75%	112.75	3111.50	51145.50
max	150.00	3984.00	68828.00

In [ ]:

Preprocessing DATA

In [17]:
```
startups['City'].value_counts()
```

Out[17]:
```
Rochester        2
Killeen          1
Wichita Falls    1
Naperville       1
Clovis           1
                ..
Akron            1
Fullerton        1
Manchester       1
Everett          1
Thousand Oaks    1
Name: City, Length: 149, dtype: int64
```

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

Out[19]:
```
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 [20]:
```
startups['State'].value_counts()
```

Out[20]:
```
California        40
Texas             17
Florida           12
Washington         7
Colorado           5
Illinois           5
Georgia            4
Alabama            4
Connecticut        4
New Jersey         4
Arizona            3
Tennessee          3
Iowa               3
Michigan           3
South Carolina     3
Utah               3
Massachusetts      3
Kansas             3
New York           3
Louisiana          2
North Carolina     2
Ohio               2
Virginia           2
Oregon             2
Mississippi        1
New Mexico         1
Arkansas           1
New Hampshire      1
Oklahoma           1
Montana            1
Indiana            1
South Dakota       1
Minnesota          1
Wisconsin          1
Rhode Island       1
Name: State, dtype: int64
```

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

Out[21]:
```
array(['Region 2', 'Region 1'], dtype=object)
```

In [24]:
```
startups['Sales Region'].value_counts()
```

Out[24]:
```
Region 2    86
Region 1    64
Name: Sales Region, dtype: int64
```

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

Out[23]:
```
array(['Old', 'New'], dtype=object)
```

In [25]:
```
startups['New Expansion'].value_counts()
```

Out[25]:
```
Old    140
New     10
Name: New Expansion, dtype: int64
```

In [26]:
```
startups.isna().sum()
```

Out[26]:
```
Store ID           0
City               0
State              0
Sales Region       0
New Expansion      0
Marketing Spend    0
Revenue            0
dtype: int64
```

In [27]:
```
startups.duplicated().sum()
```

Out[27]:
```
0
```

In [28]:
```
startups['Store ID'].duplicated().sum()
```

Out[28]:
```
0
```

Exploring & Analysing Data

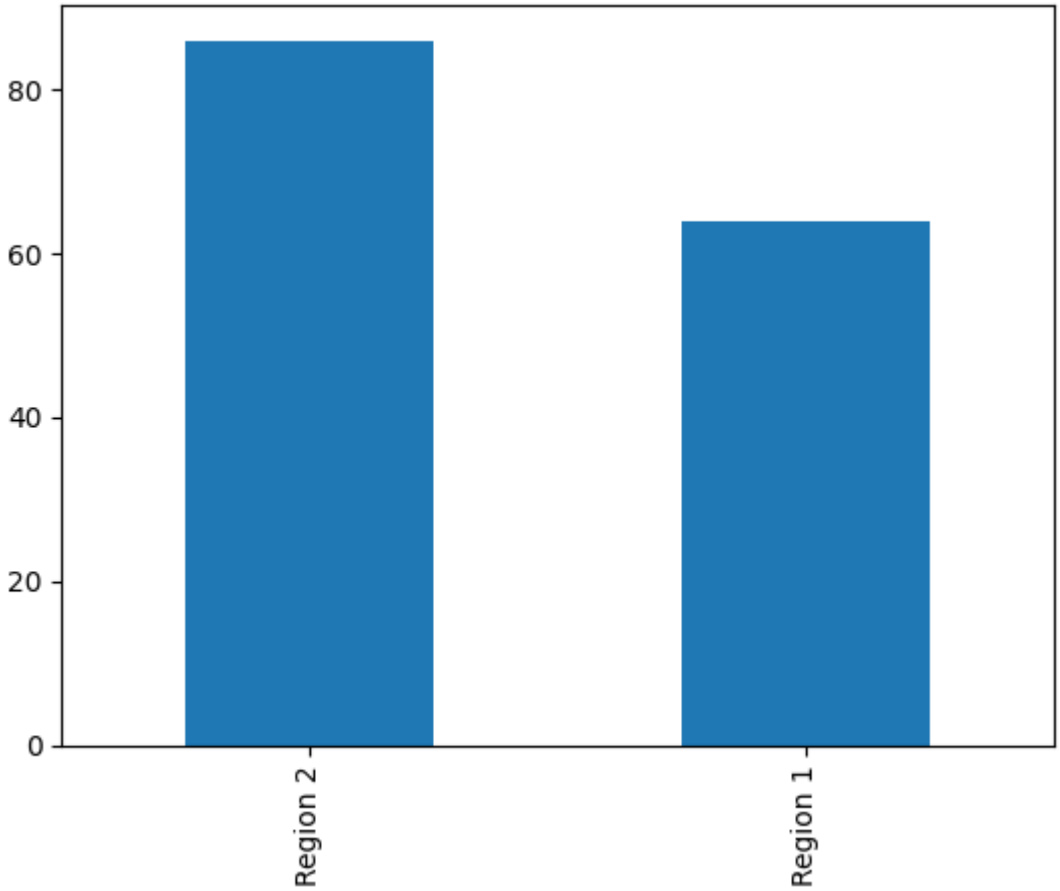
In [29]:
```
startups.sample(10)
```

Out[29]:

	Store ID	City	State	Sales Region	New Expansion	Marketing Spend	Revenue
112	113	Knoxville	Tennessee	Region 2	Old	3086	56504
140	141	Chattanooga	Tennessee	Region 2	New	3587	55357
137	138	Norwalk	California	Region 2	Old	3112	19703
147	148	Rockford	Illinois	Region 1	New	2648	43377
134	135	Santa Rosa	California	Region 2	Old	3067	59060
44	45	San Bernardino	California	Region 2	Old	3399	59870
41	42	Newport News	Virginia	Region 1	Old	2758	57625
77	78	New Haven	Connecticut	Region 1	Old	3162	45550
55	56	Thornton	Colorado	Region 2	Old	2642	46490
89	90	Tacoma	Washington	Region 2	Old	2552	45666

In [35]:
```
startups['Sales Region'].value_counts().plot.bar()
```

Out[35]:
```
<Axes: >
```



In [37]:
```
startups.groupby("New Expansion").groups
```

Out[37]:
```
{'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, 2
8, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 4
9, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 7
0, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 9
1, 92, 93, 94, 95, 96, 97, 98, 99, ...]}
```

In [38]:
```
startups[startups['New Expansion']=='New']
```

Out[38]:

	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 [44]:
```
startups[startups['New Expansion']=='New'].groupby('City').max()['Revenue']
```

Out[44]:
```
City
Brownsville          63148
Chattanooga          55357
College Station      22457
Glendale             49376
Joliet               48315
Paterson             34603
Rancho Cucamonga     52366
Rockford             43377
Tempe                48954
Thousand Oaks        40141
Name: Revenue, dtype: int64
```

In [46]:
```
startups[startups['New Expansion']=='Old'].groupby('City').max()['Revenue'].nlargest(
```

Out[46]:
```
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 [47]:
```
startups['Profit'] = startups['Revenue'] - startups['Marketing Spend']
```

In [48]:
```
startups
```

Out[48]:

	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 [49]:
```
startups['ROMS']= round((startups['Profit'] / startups['Marketing Spend'])*100,2)
```

In [52]:
```
startups['ROMS']
```

Out[52]:
```
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
Name: ROMS, Length: 150, dtype: float64
```

In [53]:
```
startups['ROMS%']= startups['ROMS']/100
```

In [54]:
```
startups['ROMS%']
```

Out[54]:
```
0      17.6890
1      15.7543
2      16.9025
3      12.8760
4      19.8738
        ...
145    14.3723
146    16.1831
147    15.3810
148     6.5007
149    15.5121
Name: ROMS%, Length: 150, dtype: float64
```

In [55]:
```
startups
```

Out[55]:

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

150 rows × 10 columns

In [56]:
```
startups.to_csv('startups_modified.csv')
```

In [ ]:
```

```

Startup Expansion

432K
Sum of Marketing Spend

5M
Sum of Profit

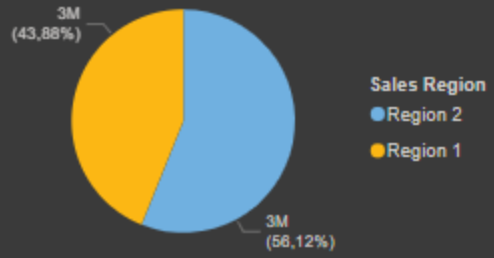
6M
Sum of Revenue

New Expansion
- New
- Old

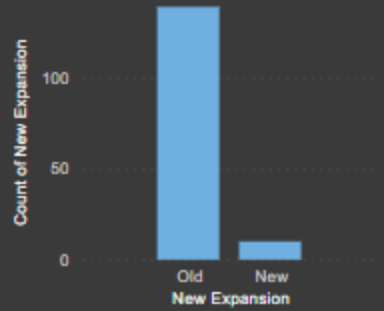
Sales Region
- Region 1
- Region 2

Sum of Profit by State

Sum of Revenue by Sales Region

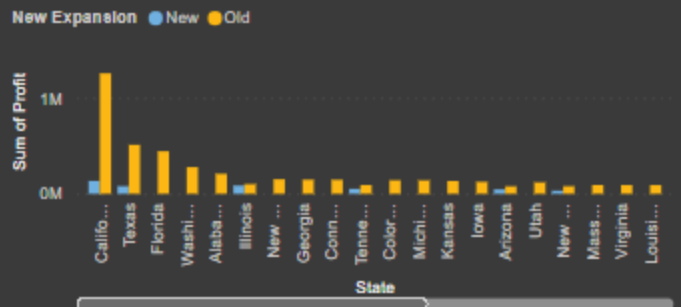


Count of New Expansion by New Expansion



City and City

Sum of Profit by State and New Expansion



Sum of Profit by Store ID