In [34]:

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn import metrics
```

In [3]:

```
file_path="C:\\Users\\Venkatesh\\TechnoHacks internship\\Data Files\\city_attr
df=pd.read_csv(file_path)
df
```

Out[3]:

	City	Country	Latitude	Longitude
0	Vancouver	Canada	49.249660	-123.119339
1	Portland	United States	45.523449	-122.676208
2	San Francisco	United States	37.774929	-122.419418
3	Seattle	United States	47.606209	-122.332069
4	Los Angeles	United States	34.052231	-118.243683
5	San Diego	United States	32.715328	-117.157257
6	Las Vegas	United States	36.174969	-115.137222
7	Phoenix	United States	33.448380	-112.074043
8	Albuquerque	United States	35.084492	-106.651138
9	Denver	United States	39.739151	-104.984703
10	San Antonio	United States	29.424120	-98.493629
11	Dallas	United States	32.783058	-96.806671
12	Houston	United States	29.763281	-95.363274
13	Kansas City	United States	39.099731	-94.578568
14	Minneapolis	United States	44.979969	-93.263840
15	Saint Louis	United States	38.627270	-90.197891
16	Chicago	United States	41.850029	-87.650047
17	Nashville	United States	36.165890	-86.784439
18	Indianapolis	United States	39.768379	-86.158043
19	Atlanta	United States	33.749001	-84.387978
20	Detroit	United States	42.331429	-83.045753
21	Jacksonville	United States	30.332180	-81.655647
22	Charlotte	United States	35.227089	-80.843132
23	Miami	United States	25.774269	-80.193657
24	Pittsburgh	United States	40.440620	-79.995888
25	Toronto	Canada	43.700111	-79.416298
26	Philadelphia	United States	39.952339	-75.163788
27	New York	United States	40.714272	-74.005966
28	Montreal	Canada	45.508839	-73.587807
29	Boston	United States	42.358429	-71.059769
30	Beersheba	Israel	31.251810	34.791302
31	Tel Aviv District	Israel	32.083328	34.799999
32	Eilat	Israel	29.558050	34.948212
33	Haifa	Israel	32.815559	34.989170
34	Nahariyya	Israel	33.005859	35.094090
35	Jerusalem	Israel	31.769039	35.216331

In [4]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 36 entries, 0 to 35
Data columns (total 4 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   City       36 non-null     object 
 1   Country    36 non-null     object 
 2   Latitude   36 non-null     float64
 3   Longitude  36 non-null     float64
dtypes: float64(2), object(2)
memory usage: 1.3+ KB
```

In [29]:
```
# TO get first 5 columns
df.head()
```

Out[29]:

	City	Country	Latitude	Longitude
0	Vancouver	Canada	49.249660	-123.119339
1	Portland	United States	45.523449	-122.676208
2	San Francisco	United States	37.774929	-122.419418
3	Seattle	United States	47.606209	-122.332069
4	Los Angeles	United States	34.052231	-118.243683

In [15]:
```
# TO get last 5 columns
df.tail()
```

Out[15]:

	City	Country	Latitude	Longitude
31	Tel Aviv District	Israel	32.083328	34.799999
32	Eilat	Israel	29.558050	34.948212
33	Haifa	Israel	32.815559	34.989170
34	Nahariyya	Israel	33.005859	35.094090
35	Jerusalem	Israel	31.769039	35.216331

In [23]:
```
# TO get number of rows and columns
df.shape
```

Out[23]:
```
(36, 4)
```

In [6]:
```
df.size
```

Out[6]:
```
144
```

In [27]:
```
df.columns
```

Out[27]:
```
Index(['City', 'Country', 'Latitude', 'Longitude'], dtype='object')
```

```
In [7]: df.dtypes
```

```
Out[7]: City           object
        Country        object
        Latitude      float64
        Longitude     float64
        dtype: object
```

```
In [12]: df.isnull().sum()
```

```
Out[12]: City         0
         Country      0
         Latitude     0
         Longitude    0
         dtype: int64
```

```
In [13]: df.isnull().values.any()
```

```
Out[13]: False
```

In [20]:

```
df.select_dtypes('object')
```

Out[20]:

	City	Country
0	Vancouver	Canada
1	Portland	United States
2	San Francisco	United States
3	Seattle	United States
4	Los Angeles	United States
5	San Diego	United States
6	Las Vegas	United States
7	Phoenix	United States
8	Albuquerque	United States
9	Denver	United States
10	San Antonio	United States
11	Dallas	United States
12	Houston	United States
13	Kansas City	United States
14	Minneapolis	United States
15	Saint Louis	United States
16	Chicago	United States
17	Nashville	United States
18	Indianapolis	United States
19	Atlanta	United States
20	Detroit	United States
21	Jacksonville	United States
22	Charlotte	United States
23	Miami	United States
24	Pittsburgh	United States
25	Toronto	Canada
26	Philadelphia	United States
27	New York	United States
28	Montreal	Canada
29	Boston	United States
30	Beersheba	Israel
31	Tel Aviv District	Israel
32	Eilat	Israel
33	Haifa	Israel
34	Nahariyya	Israel
35	Jerusalem	Israel

In [25]:

```
df.describe()
```

Out[25]:

	Latitude	Longitude
count	36.000000	36.000000
mean	37.066743	-73.544668
std	5.815514	51.612349
min	25.774269	-123.119339
25%	32.766126	-105.401312
50%	36.170429	-86.471241
75%	40.998211	-74.874332
max	49.249660	35.216331

In [41]:

```
df['Country'].value_counts()
```

Out[41]:

```
Country
United States    27
Israel            6
Canada            3
Name: count, dtype: int64
```

In [45]:

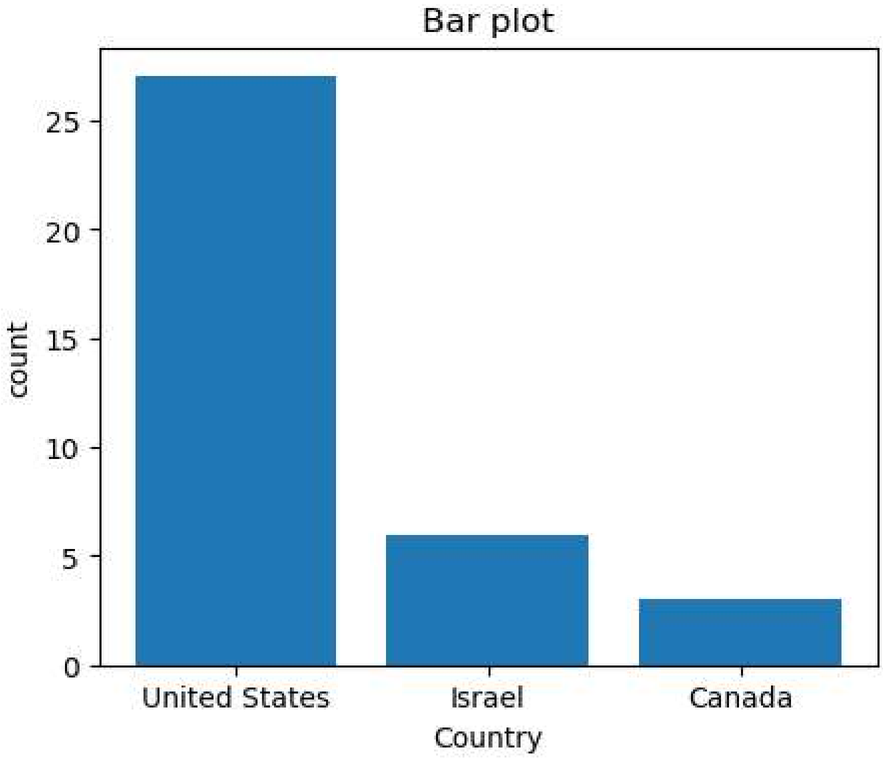
```
count=df['Country'].value_counts().keys().tolist()
values=df['Country'].value_counts().values.tolist()
Country_df=pd.DataFrame(zip(count,values),columns=['Country','count'])
Country_df
```

Out[45]:

	Country	count
0	United States	27
1	Israel	6
2	Canada	3

```
In [46]: plt.figure(figsize=(5,4))
         plt.title('Bar plot')
         plt.xlabel('Country')
         plt.ylabel('count')
         plt.bar('Country','count',data=Country_df)

         plt.show()
```



In [67]:
```
data1=pd.read_csv('C:\\Users\\Venkatesh\\TechnoHacks internship\\Data Files\\w
data1
```

Out[67]:

	datetime	Vancouver	Portland	San Francisco	Seattle	Los Angeles	San Diego	Las Vegas	Phoenix	All
0	2012-10-01 12:00:00	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1	2012-10-01 13:00:00	mist	scattered clouds	light rain	sky is clear	mist	sky is clear	sky is clear	sky is clear	
2	2012-10-01 14:00:00	broken clouds	scattered clouds	sky is clear	sky is clear	sky is clear	sky is clear	sky is clear	sky is clear	
3	2012-10-01 15:00:00	broken clouds	scattered clouds	sky is clear	sky is clear	sky is clear	sky is clear	sky is clear	sky is clear	
4	2012-10-01 16:00:00	broken clouds	scattered clouds	sky is clear	sky is clear	sky is clear	sky is clear	sky is clear	sky is clear	
...	...	...	...	...	...	...	...	...	...	
45248	2017-11-29 20:00:00	NaN	broken clouds	NaN	light rain	sky is clear	broken clouds	sky is clear	sky is clear	
45249	2017-11-29 21:00:00	NaN	broken clouds	NaN	overcast clouds	sky is clear	broken clouds	sky is clear	sky is clear	
45250	2017-11-29 22:00:00	NaN	broken clouds	NaN	broken clouds	sky is clear	broken clouds	sky is clear	sky is clear	
45251	2017-11-29 23:00:00	NaN	broken clouds	NaN	broken clouds	sky is clear	broken clouds	sky is clear	broken clouds	
45252	2017-11-30 00:00:00	NaN	broken clouds	NaN	few clouds	sky is clear	broken clouds	sky is clear	broken clouds	

45253 rows × 37 columns

In [68]:

```
data1.isnull().sum()
```

Out[68]:

```
datetime               0
Vancouver            793
Portland               1
San Francisco        793
Seattle                1
Los Angeles            1
San Diego              1
Las Vegas              1
Phoenix                1
Albuquerque            1
Denver                 1
San Antonio            1
Dallas                 1
Houston                1
Kansas City            1
Minneapolis            1
Saint Louis            1
Chicago                1
Nashville              1
Indianapolis           1
Atlanta                1
Detroit                1
Jacksonville           1
Charlotte              1
Miami                793
Pittsburgh             1
Toronto                1
Philadelphia           1
New York             793
Montreal               1
Boston                 1
Beersheba            793
Tel Aviv District    793
Eilat                792
Haifa                793
Nahariyya            793
Jerusalem            793
dtype: int64
```

In [71]:
```
data1.isna().any()
```

Out[71]:
```
datetime             False
Vancouver             True
Portland              True
San Francisco         True
Seattle               True
Los Angeles           True
San Diego             True
Las Vegas             True
Phoenix               True
Albuquerque           True
Denver                True
San Antonio           True
Dallas                True
Houston               True
Kansas City           True
Minneapolis           True
Saint Louis           True
Chicago               True
Nashville             True
Indianapolis          True
Atlanta               True
Detroit               True
Jacksonville          True
Charlotte             True
Miami                 True
Pittsburgh            True
Toronto               True
Philadelphia          True
New York              True
Montreal              True
Boston                True
Beersheba             True
Tel Aviv District     True
Eilat                 True
Haifa                 True
Nahariyya             True
Jerusalem             True
dtype: bool
```

In [72]:

```
data1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45253 entries, 0 to 45252
Data columns (total 37 columns):
 #   Column             Non-Null Count  Dtype 
---  ------             --------------  ----- 
 0   datetime           45253 non-null  object
 1   Vancouver          44460 non-null  object
 2   Portland           45252 non-null  object
 3   San Francisco      44460 non-null  object
 4   Seattle            45252 non-null  object
 5   Los Angeles        45252 non-null  object
 6   San Diego          45252 non-null  object
 7   Las Vegas          45252 non-null  object
 8   Phoenix            45252 non-null  object
 9   Albuquerque        45252 non-null  object
 10  Denver             45252 non-null  object
 11  San Antonio        45252 non-null  object
 12  Dallas             45252 non-null  object
 13  Houston            45252 non-null  object
 14  Kansas City        45252 non-null  object
 15  Minneapolis        45252 non-null  object
 16  Saint Louis        45252 non-null  object
 17  Chicago            45252 non-null  object
 18  Nashville          45252 non-null  object
 19  Indianapolis       45252 non-null  object
 20  Atlanta            45252 non-null  object
 21  Detroit            45252 non-null  object
 22  Jacksonville       45252 non-null  object
 23  Charlotte          45252 non-null  object
 24  Miami              44460 non-null  object
 25  Pittsburgh         45252 non-null  object
 26  Toronto            45252 non-null  object
 27  Philadelphia       45252 non-null  object
 28  New York           44460 non-null  object
 29  Montreal           45252 non-null  object
 30  Boston             45252 non-null  object
 31  Beersheba          44460 non-null  object
 32  Tel Aviv District  44460 non-null  object
 33  Eilat              44461 non-null  object
 34  Haifa              44460 non-null  object
 35  Nahariyya          44460 non-null  object
 36  Jerusalem          44460 non-null  object
dtypes: object(37)
memory usage: 12.8+ MB
```

In [76]:

```
data1.describe()
```

Out[76]:

	datetime	Vancouver	Portland	San Francisco	Seattle	Los Angeles	San Diego	Las Vegas	Phoenix	Alb
count	45253	44460	45252	44460	45252	45252	45252	45252	45252	
unique	45253	37	24	28	29	25	22	23	26	
top	2012-10-01 12:00:00	sky is clear	sky is clear	sky is clear	sky is clear	sky is clear	sky is clear	sky is clear	sky is clear	s
freq	1	12805	11725	12654	12801	26136	14829	35090	30303	

4 rows × 37 columns

In [73]:
```
data2=pd.read_csv('C:\\Users\\Venkatesh\\TechnoHacks internship\\Data Files\\t
data2
```

Out[73]:

	datetime	Vancouver	Portland	San Francisco	Seattle	Los Angeles	San Diego	Las V
0	2012-10-01 12:00:00	NaN	NaN	NaN	NaN	NaN	NaN	
1	2012-10-01 13:00:00	284.630000	282.080000	289.480000	281.800000	291.870000	291.530000	293.41
2	2012-10-01 14:00:00	284.629041	282.083252	289.474993	281.797217	291.868186	291.533501	293.40
3	2012-10-01 15:00:00	284.626998	282.091866	289.460618	281.789833	291.862844	291.543355	293.39
4	2012-10-01 16:00:00	284.624955	282.100481	289.446243	281.782449	291.857503	291.553209	293.38
...	...	...	...	...	...	...	...	
45248	2017-11-29 20:00:00	NaN	282.000000	NaN	280.820000	293.550000	292.150000	289.54
45249	2017-11-29 21:00:00	NaN	282.890000	NaN	281.650000	295.680000	292.740000	290.61
45250	2017-11-29 22:00:00	NaN	283.390000	NaN	282.750000	295.960000	292.580000	291.34
45251	2017-11-29 23:00:00	NaN	283.020000	NaN	282.960000	295.650000	292.610000	292.15
45252	2017-11-30 00:00:00	NaN	282.280000	NaN	283.040000	294.930000	291.400000	291.64

45253 rows × 37 columns

In [74]:
```
data2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45253 entries, 0 to 45252
Data columns (total 37 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   datetime           45253 non-null  object 
 1   Vancouver          44458 non-null  float64
 2   Portland           45252 non-null  float64
 3   San Francisco      44460 non-null  float64
 4   Seattle            45250 non-null  float64
 5   Los Angeles        45250 non-null  float64
 6   San Diego          45252 non-null  float64
 7   Las Vegas          45252 non-null  float64
 8   Phoenix            45250 non-null  float64
 9   Albuquerque        45252 non-null  float64
 10  Denver             45252 non-null  float64
 11  San Antonio        45252 non-null  float64
 12  Dallas             45249 non-null  float64
 13  Houston            45250 non-null  float64
 14  Kansas City        45252 non-null  float64
 15  Minneapolis        45240 non-null  float64
 16  Saint Louis        45252 non-null  float64
 17  Chicago            45250 non-null  float64
 18  Nashville          45251 non-null  float64
 19  Indianapolis       45246 non-null  float64
 20  Atlanta            45247 non-null  float64
 21  Detroit            45252 non-null  float64
 22  Jacksonville       45252 non-null  float64
 23  Charlotte          45250 non-null  float64
 24  Miami              44448 non-null  float64
 25  Pittsburgh         45250 non-null  float64
 26  Toronto            45252 non-null  float64
 27  Philadelphia       45250 non-null  float64
 28  New York           44460 non-null  float64
 29  Montreal           45250 non-null  float64
 30  Boston             45250 non-null  float64
 31  Beersheba          44455 non-null  float64
 32  Tel Aviv District  44460 non-null  float64
 33  Eilat              44461 non-null  float64
 34  Haifa              44455 non-null  float64
 35  Nahariyya          44456 non-null  float64
 36  Jerusalem          44460 non-null  float64
dtypes: float64(36), object(1)
memory usage: 12.8+ MB
```

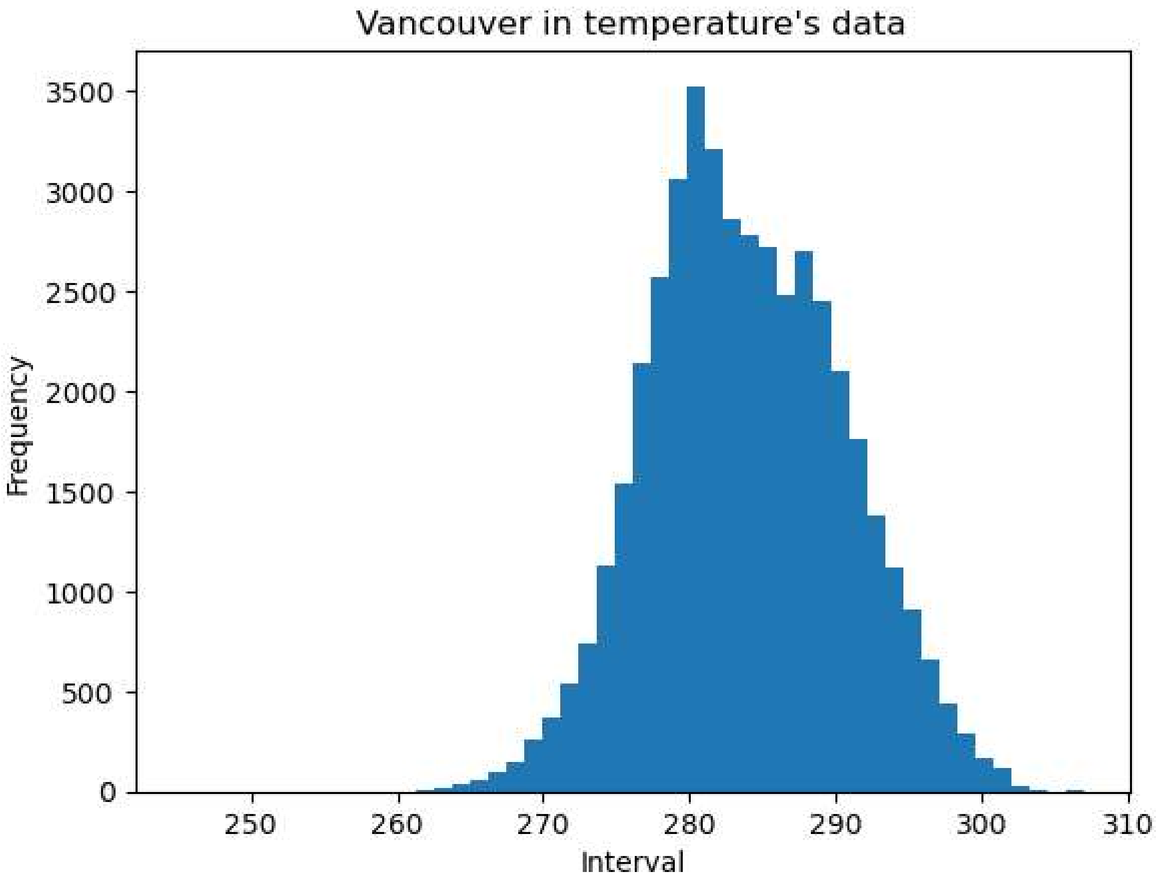
In [75]:
```
data2.describe()
```

Out[75]:

	Vancouver	Portland	San Francisco	Seattle	Los Angeles	San Diego	
count	44458.000000	45252.000000	44460.000000	45250.000000	45250.000000	45252.000000	452
mean	283.862654	284.992929	288.155821	284.409626	290.846116	290.215044	2
std	6.640131	7.452438	5.332862	6.547986	6.460823	5.889992	
min	245.150000	262.370000	272.300000	263.780000	266.503667	265.783333	2
25%	279.160000	279.850000	284.670000	279.830000	286.380000	286.254750	2
50%	283.450000	284.320000	287.610000	283.940000	290.530000	290.118750	2
75%	288.600785	289.451750	291.015167	288.530000	295.080000	294.107542	3
max	307.000000	312.520000	313.620000	307.300000	315.470000	313.360000	3

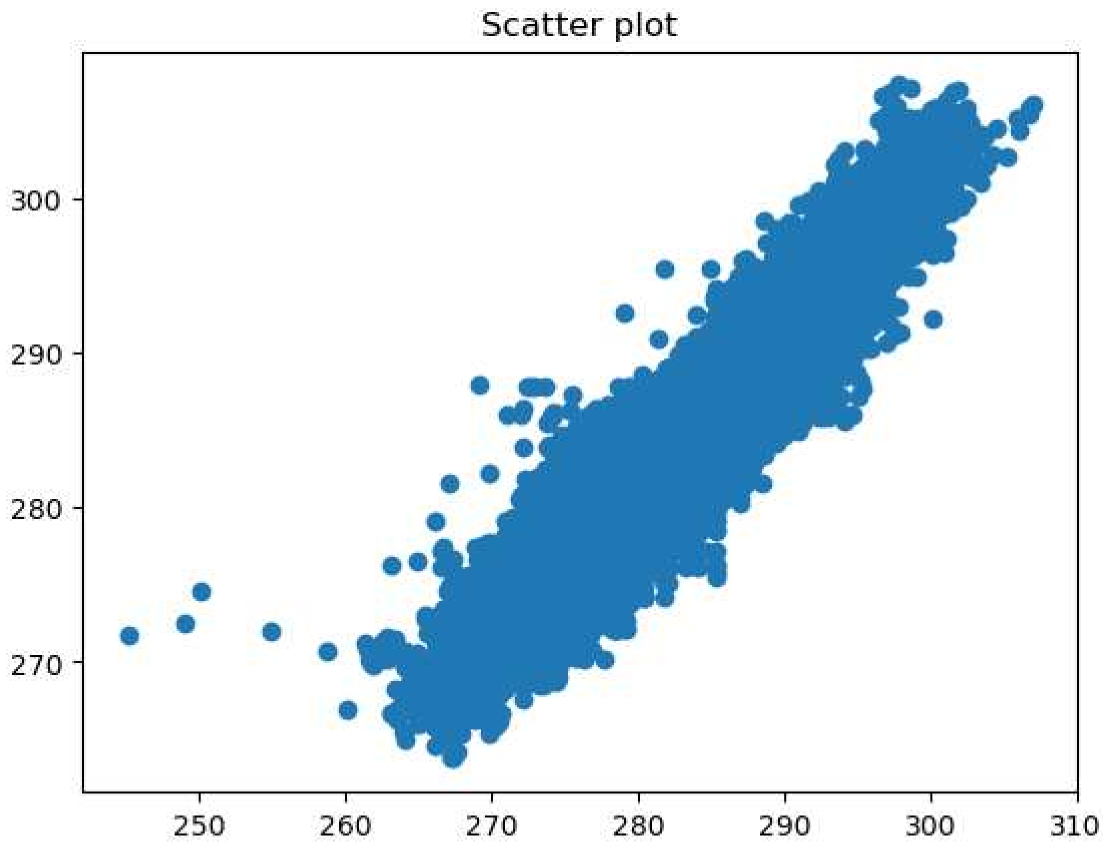
8 rows × 36 columns

In [153]:
```
plt.hist(data2['Vancouver'],bins=50)
plt.title("Vancouver in temperature's data")
plt.xlabel("Interval")
plt.ylabel('Frequency')
plt.show()
```



In [168]:

```
plt.scatter(data2['Vancouver'],data2['Seattle'])
plt.title('Scatter plot')
plt.show()
```



In [96]:

```
mean_num=np.mean(df['Latitude'])
median_num=np.median(df['Latitude'])
min_num=np.min(df['Latitude'])
max_num=np.max(df['Latitude'])
std_num=np.std(df['Latitude'])

list1=[mean_num,median_num,min_num,max_num,std_num]
index=['Mean','Median','Min','Max','Std']
pd.DataFrame(list1,columns=['Latitude'],index=index)
```

Out[96]:

	Latitude
Mean	37.066743
Median	36.170429
Min	25.774269
Max	49.249660
Std	5.734174

In [97]:
```
df.describe()
```

Out[97]:

	Latitude	Longitude
count	36.000000	36.000000
mean	37.066743	-73.544668
std	5.815514	51.612349
min	25.774269	-123.119339
25%	32.766126	-105.401312
50%	36.170429	-86.471241
75%	40.998211	-74.874332
max	49.249660	35.216331

In [98]:
```
per_25=np.percentile(df['Latitude'],25)
per_50=np.percentile(df['Latitude'],50)
per_75=np.percentile(df['Latitude'],75)
print(per_25,per_50,per_75)
```

```
32.7661255 36.1704295 40.99821125
```

In [99]:
```
mean_num=np.mean(df['Latitude'])
median_num=np.median(df['Latitude'])
min_num=np.min(df['Latitude'])
max_num=np.max(df['Latitude'])
std_num=np.std(df['Latitude'])

list1=[mean_num,median_num,min_num,max_num,std_num,per_25,per_50,per_75]
index=['Mean','Median','Min','Max','Std','25%','50%','75%']
pd.DataFrame(list1,columns=['Latitude'],index=index)
```

Out[99]:

	Latitude
Mean	37.066743
Median	36.170429
Min	25.774269
Max	49.249660
Std	5.734174
25%	32.766126
50%	36.170429
75%	40.998211

In [109]:

```
################### u-1*sigma to u+1*sigma #################
val_minus_1_sigma=mean_num-1*std_num
val_plus_1_sigma=mean_num+1*std_num

################### u-2*sigma to u+2*sigma #################
val_minus_2_sigma=mean_num-2*std_num
val_plus_2_sigma=mean_num+2*std_num

################### u-3*sigma to u+3*sigma #################
val_minus_3_sigma=mean_num-3*std_num
val_plus_3_sigma=mean_num+3*std_num
```

In [110]:

```
df['Latitude']
```

Out[110]:

```
0     49.249660
1     45.523449
2     37.774929
3     47.606209
4     34.052231
5     32.715328
6     36.174969
7     33.448380
8     35.084492
9     39.739151
10    29.424120
11    32.783058
12    29.763281
13    39.099731
14    44.979969
15    38.627270
16    41.850029
17    36.165890
18    39.768379
19    33.749001
20    42.331429
21    30.332180
22    35.227089
23    25.774269
24    40.440620
25    43.700111
26    39.952339
27    40.714272
28    45.508839
29    42.358429
30    31.251810
31    32.083328
32    29.558050
33    32.815559
34    33.005859
35    31.769039
Name: Latitude, dtype: float64
```

```
In [111]: df['Latitude']<32.7
```

```
Out[111]: 0     False
          1     False
          2     False
          3     False
          4     False
          5     False
          6     False
          7     False
          8     False
          9     False
          10     True
          11    False
          12     True
          13    False
          14    False
          15    False
          16    False
          17    False
          18    False
          19    False
          20    False
          21     True
          22    False
          23     True
          24    False
          25    False
          26    False
          27    False
          28    False
          29    False
          30     True
          31     True
          32     True
          33    False
          34    False
          35     True
          Name: Latitude, dtype: bool
```

In [112]:
```
cond=df['Latitude']<32.7
df[cond]
```

Out[112]:

	City	Country	Latitude	Longitude
10	San Antonio	United States	29.424120	-98.493629
12	Houston	United States	29.763281	-95.363274
21	Jacksonville	United States	30.332180	-81.655647
23	Miami	United States	25.774269	-80.193657
30	Beersheba	Israel	31.251810	34.791302
31	Tel Aviv District	Israel	32.083328	34.799999
32	Eilat	Israel	29.558050	34.948212
35	Jerusalem	Israel	31.769039	35.216331

In [115]:
```
cond1=df['Latitude']<val_plus_1_sigma
cond2=df['Latitude']>val_minus_1_sigma
len(df[cond1&cond2])
```

Out[115]:
```
24
```

In [116]:
```
24 / 35

# 68% of data are between u-1*sigma to u+1*sigma
```

Out[116]:
```
0.6857142857142857
```

In [117]:
```
cond1=df['Latitude']<val_plus_2_sigma
cond2=df['Latitude']>val_minus_2_sigma
len(df[cond1&cond2])
```

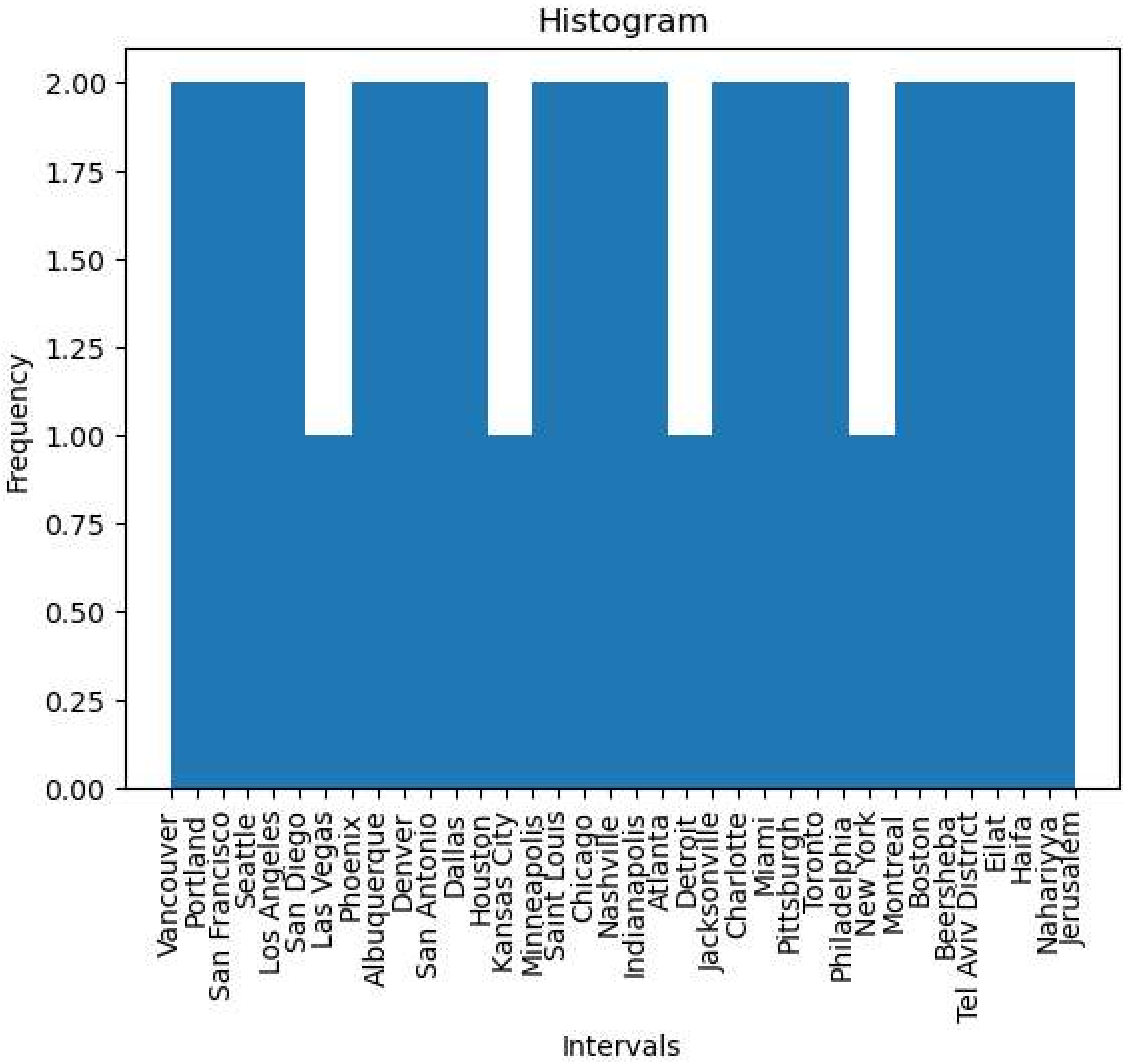
Out[117]:
```
35
```

In [118]:
```
35 / 35
```

Out[118]:
```
1.0
```

In [139]:

```
data=df['City']
plt.hist(data,bins=20)
plt.title('Histogram')
plt.xlabel('Intervals')
plt.ylabel('Frequency')
plt.xticks(rotation=90)
plt.show()
```



In [ ]: