Roll No: AA.SC.P2MCA2107423

Date: 15-04-2021

MCA AI- Business Analytics (S1)

Assignment 4

1. Plot the latitudes and longitudes of any 5 cities in India (include your hometown as well), on the map.

In [1]:

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
plt.rcParams['axes.facecolor'] = 'white'
import seaborn as sns
sns.set_style('whitegrid')
from shapely.geometry import Point
import plotly.express as px
import chart_studio.plotly as pl
import plotly.graph_objs as gobj
from plotly.offline import download_plotlyjs,init_notebook_mode,plot,iplot
init_notebook_mode(connected=True)
#import geopandas as gpd
#from geopandas import GeoDataFrame
```

In [2]:
```
list_city_location = [["Delhi", "Delhi", 28.679079, 77.069710],
                      ["Mumbai", "Maharastra", 19.076090, 72.877426],
                      ["Lucknow", "Uttar Pradesh", 26.850000, 80.949997],
                      ["Kolkata", "West Bengal", 22.572645, 88.363892],
                      ["Raipur", "Chhattisgarh", 21.250000, 81.629997],
                      ["Kollam", "Kerala", 8.893212, 76.614143]]
df_cities = pd.DataFrame(list_city_location, columns = ["City", "State", "Latitude", "Longitude"])
df_cities
```

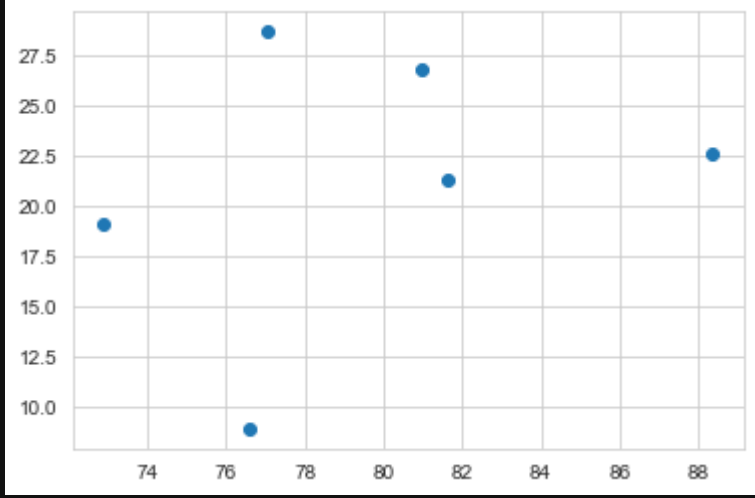
Out[2]:

	City	State	Latitude	Longitude
0	Delhi	Delhi	28.679079	77.069710
1	Mumbai	Maharastra	19.076090	72.877426
2	Lucknow	Uttar Pradesh	26.850000	80.949997
3	Kolkata	West Bengal	22.572645	88.363892
4	Raipur	Chhattisgarh	21.250000	81.629997
5	Kollam	Kerala	8.893212	76.614143

2. Illustrate by few example plots (minimum 5 plots/tool):

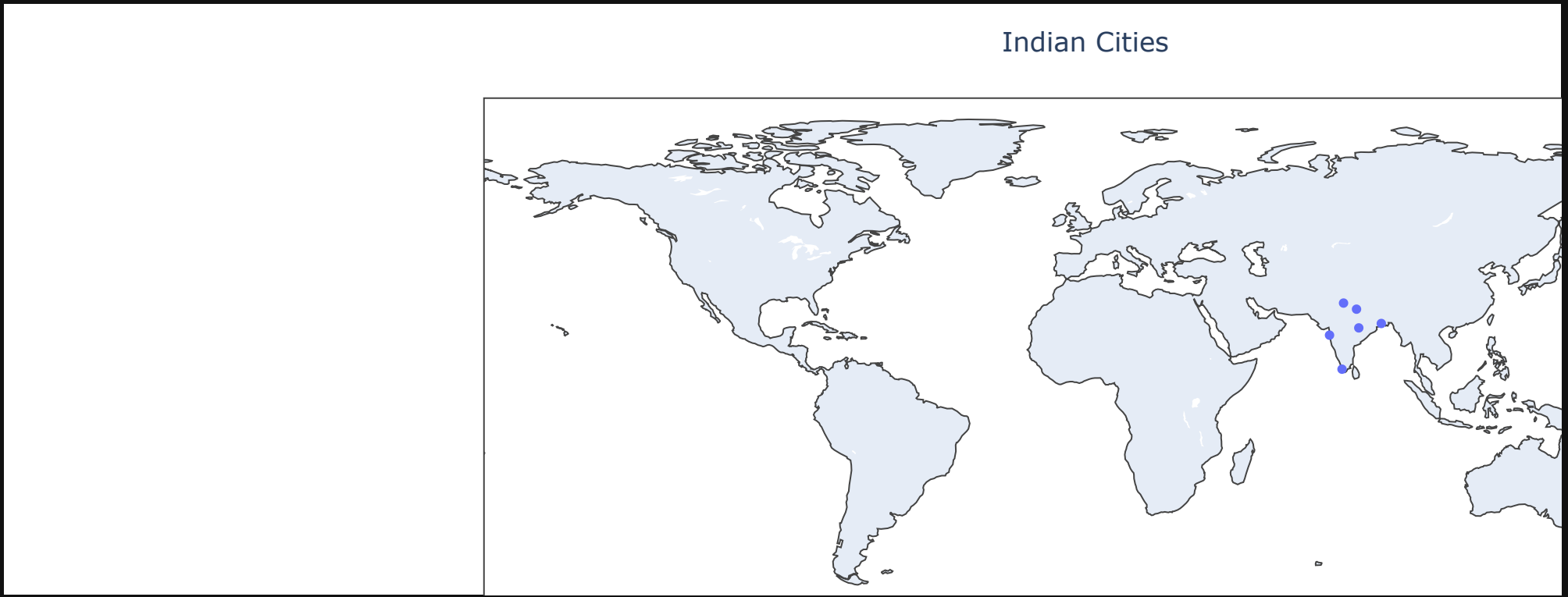
a. Plotly
b. Bokeh
c. Tableau

In [3]:
```
plt.scatter(x=df_cities['Longitude'], y=df_cities['Latitude'])
plt.show()
```



In [4]:

```
fig = px.scatter_geo(df_cities,lat='Latitude',lon='Longitude')
fig.update_layout(title = 'Indian Cities', title_x=0.5)
fig.show()
```



In [5]:

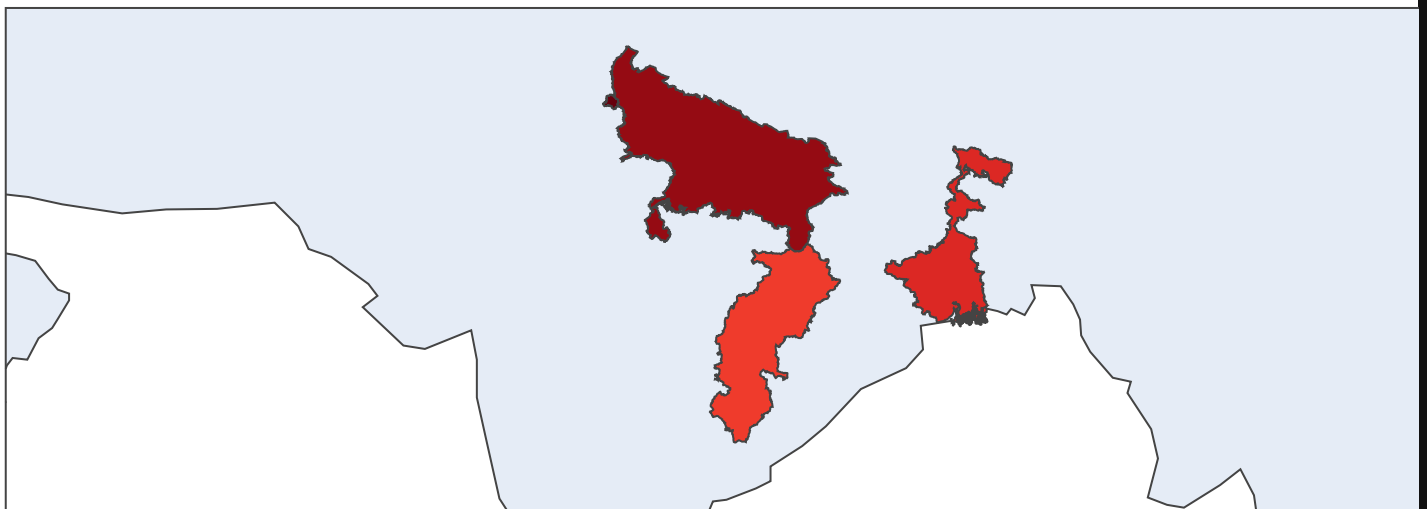
```
fig = px.choropleth(
    df_cities,

geojson="https://gist.githubusercontent.com/jbrobst/56c13bbbf9d97d187fea01ca62ea5112/raw/e388c4cae20aa53cb5090210a42ebb9b765c

    featureidkey='properties.ST_NM',
    locations='State',
    color='Latitude',
    color_continuous_scale='Reds'
)

fig.update_geos(fitbounds="locations", visible=True)

fig.show()
```



3. Table shows the distribution of various ethnic groups in the population of a particular state based on a decennial US. census

Ethnicity	White	Black	Amer.-Indian	Hispanic	Asian	Others
Proportion	0.743	0.216	0.012	0.012	0.008	0.009

Five years later a random sample of 2,500 residents of the state was taken, with the results given in Table. Test, at the 1% level of significance, whether there is sufficient evidence in the sample to conclude that the distribution of ethnic groups in this state five years after the census had changed from that in the census year

Ethnicity	Observed Frequency
White	1732
Black	538
American-Indian	32
Hispanic	42
Asian	133
Others	23

In [6]:
```
df_Ethnicity = pd.read_csv("C:\spark\MCA\Semester1\E2_BA\Assignment 4\Ethnicity.csv")
df_Ethnicity
```

Out[6]:

	Ethnicity	Assumed Distribution	Observed Frequency
0	White	0.743	1732
1	Black	0.216	538
2	American-Indian	0.012	32
3	Hispanic	0.012	42
4	Asian	0.008	133
5	Others	0.009	23

Step 1. The hypotheses of interest in this case can be expressed as

H0: The distribution of ethnic groups has not changed

Ha: The distribution of ethnic groups has changed

Step 2. The distribution is chi-square.

Step 3. To compute the value of the test statistic we must first compute the expected number for each row of Table

In [7]:
```
df_Ethnicity['Expected Frequency'] = 2500 * df_Ethnicity['Assumed Distribution']
df_Ethnicity
```

Out[7]:

	Ethnicity	Assumed Distribution	Observed Frequency	Expected Frequency
0	White	0.743	1732	1857.5
1	Black	0.216	538	540.0
2	American-Indian	0.012	32	30.0
3	Hispanic	0.012	42	30.0
4	Asian	0.008	133	20.0
5	Others	0.009	23	22.5

In [8]:
```
df_Ethnicity.info()
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
RangeIndex: 6 entries, 0 to 5
Data columns (total 4 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Ethnicity             6 non-null      object
 1   Assumed Distribution  6 non-null      float64
 2   Observed Frequency    6 non-null      int64
 3   Expected Frequency    6 non-null      float64
dtypes: float64(2), int64(1), object(1)
memory usage: 320.0+ bytes
```

In [9]:
```
sum = 0
for i in df_Ethnicity.index:
    sq = np.power((df_Ethnicity.loc[i, 'Observed Frequency'] - df_Ethnicity.loc[i, 'Expected Frequency']), 2) /
df_Ethnicity.loc[i, 'Expected Frequency']
    sum = sum + sq
```

In [10]:
```
sum
```

Out[10]:
```
651.881125068541
```

Since the random variable takes six values, I = 6. Thus the test statistic follows the chi-square distribution with df=6−1=5 degrees of freedom.

Since the test is right-tailed, the critical value is X^2(0.01). Reading from below figure, X^2(0.01)=15.086, so the rejection region is [15.086,∞).

Critical Values of Chi-Square Distributions

	x^2 Right-Tail Area									
df	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267

17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
31	14.458	15.655	17.539	19.281	21.434	41.422	44.985	48.232	52.191	55.003
32	15.134	16.362	18.291	20.072	22.271	42.585	46.194	49.480	53.486	56.328
33	15.815	17.074	19.047	20.867	23.110	43.745	47.400	50.725	54.776	57.648
34	16.501	17.789	19.806	21.664	23.952	44.903	48.602	51.966	56.061	58.964
35	17.192	18.509	20.569	22.465	24.797	46.059	49.802	53.203	57.342	60.275
36	17.887	19.233	21.336	23.269	25.643	47.212	50.998	54.437	58.619	61.581
37	18.586	19.96	22.106	24.075	26.492	48.363	52.192	55.668	59.893	62.883
38	19.289	20.691	22.878	24.884	27.343	49.513	53.384	56.896	61.162	64.181
39	19.996	21.426	23.654	25.695	28.196	50.660	54.572	58.120	62.428	65.476
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
41	21.421	22.906	25.215	27.326	29.907	52.949	56.942	60.561	64.950	68.053
42	22.138	23.650	25.999	28.144	30.765	54.090	58.124	61.777	66.206	69.336
43	22.859	24.398	26.785	28.965	31.625	55.230	59.304	62.990	67.459	70.616
44	23.584	25.148	27.575	29.787	32.487	56.369	60.481	64.201	68.710	71.893
45	24.311	25.901	28.366	30.612	33.350	57.505	61.656	65.410	69.957	73.166
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Since 651.881 > 15.086 the decision is to reject the null hypothesis.

The data provide sufficient evidence, at the 1% level of significance, to conclude that the ethnic distribution in this state has changed in the five years since the U.S. census.