In [1]:

```
#Store and read the data
import pandas as pd
location_df = pd.read_csv( "Assignment1_Dataset.csv" )
location_df.head( 5 )
```

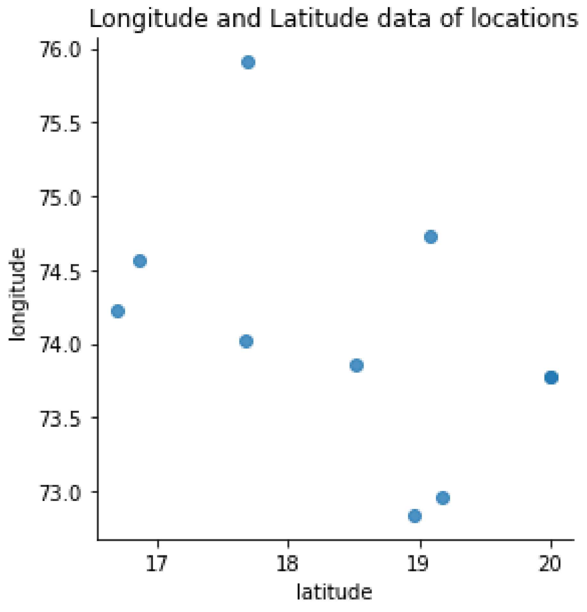
Out[1]:

	Slno	Location	latitude	longitude
0	1	Mumbai	18.9667	72.8333
1	2	Pune	18.5196	73.8553
2	3	Nashik	20.0000	73.7833
3	4	Nagpur	20.0000	73.7833
4	5	Thane	19.1800	72.9633

In [2]:

```
#plot the data
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sn
sn.lmplot( "latitude", "longitude", data=location_df, fit_reg = False, size = 4 );
plt.title( "Longitude and Latitude data of locations");
```

```
D:\Anaconda\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass
the following variables as keyword args: x, y. From version 0.12, the only v
alid positional argument will be `data`, and passing other arguments without
an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
D:\Anaconda\lib\site-packages\seaborn\regression.py:580: UserWarning: The `s
ize` parameter has been renamed to `height`; please update your code.
  warnings.warn(msg, UserWarning)
```



In [3]:

```
#Selecting the features
new_location_df = location_df[["latitude", "longitude"]]
new_location_df[0:5]
```

Out[3]:

	latitude	longitude
0	18.9667	72.8333
1	18.5196	73.8553
2	20.0000	73.7833
3	20.0000	73.7833
4	19.1800	72.9633

In [5]:

```
#K-means Clustering
from sklearn.cluster import KMeans
clusters_new = KMeans( 3, random_state=7 )
clusters_new.fit( new_location_df )
new_location_df["clusterid"] = clusters_new.labels_
new_location_df[0:9]
```

```
<ipython-input-5-c3736858f3e9>:5: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/
stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pand
as.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-v
ersus-a-copy)
  new_location_df["clusterid"] = clusters_new.labels_
```

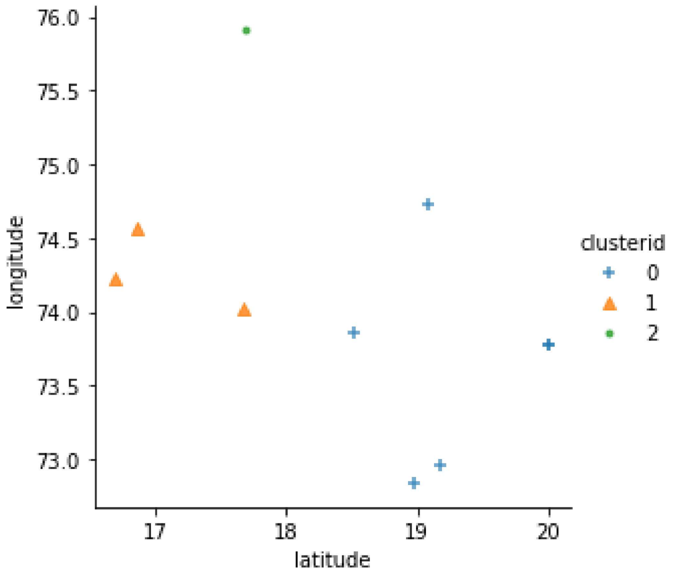
Out[5]:

	latitude	longitude	clusterid
0	18.9667	72.8333	0
1	18.5196	73.8553	0
2	20.0000	73.7833	0
3	20.0000	73.7833	0
4	19.1800	72.9633	0
5	19.0833	74.7333	0
6	17.6805	74.0183	1
7	16.7000	74.2333	1
8	17.6833	75.9167	2

In [6]:

```
#Plot the clusters
import seaborn as sn
markers = ['+','^','.']
sn.lmplot( "latitude", "longitude",data=new_location_df,hue = "clusterid", fit_reg=False,ma
```

```
D:\Anaconda\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass
the following variables as keyword args: x, y. From version 0.12, the only v
alid positional argument will be `data`, and passing other arguments without
an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
D:\Anaconda\lib\site-packages\seaborn\regression.py:580: UserWarning: The `s
ize` parameter has been renamed to `height`; please update your code.
  warnings.warn(msg, UserWarning)
```



In [7]:

```
#Centroid of the clusters
centers = np.array(clusters_new.cluster_centers_)
centers
```

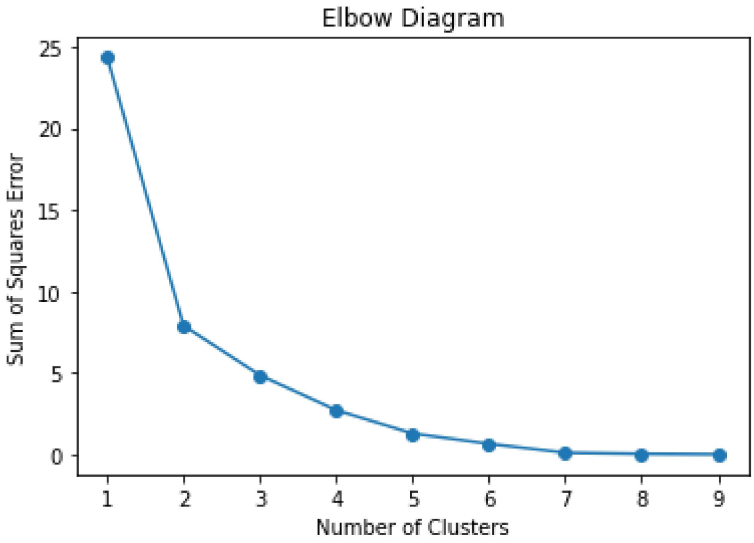
Out[7]:

```
array([[19.2916    , 73.65863333,  0.        ],
       [17.0824    , 74.27276667,  1.        ],
       [17.6833    , 75.9167    ,  2.        ]])
```

In [8]:

```
#Determining number of clusters
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
cluster_range = range( 1, 10 )
cluster_errors = []
for num_clusters in cluster_range:
    clusters = KMeans( num_clusters )
    clusters.fit( new_location_df )
    cluster_errors.append( clusters.inertia_ )
plt.figure(figsize=(6,4))
plt.plot( cluster_range, cluster_errors, marker = "o" )
plt.title('Elbow Diagram')
plt.xlabel('Number of Clusters')
plt.ylabel('Sum of Squares Error');
```

```
D:\Anaconda\lib\site-packages\sklearn\cluster\_kmeans.py:881: UserWarning: K
Means is known to have a memory leak on Windows with MKL, when there are les
s chunks than available threads. You can avoid it by setting the environment
variable OMP_NUM_THREADS=1.
  warnings.warn(
```



In [ ]: