# pandas_toGo-slides

Pandas tutorial for the Master in Neurosciences.

Florencia Noriega July 2014

```
In [55]:  import pandas as pd
          import matplotlib.pyplot as plt
          import numpy as np
          from IPython.display import Image # for the showing of images

          #http://en.wikipedia.org/wiki/Subtropics#mediaviewer/File:World_map_indicating_tropics
          #pandas = Image('./panditaBambu.jpg', width=(100,100))
          bye = Image('./images/panditaBye.png', width=(100,100))
          #longitude = Image('./longitude.png', width=(100,100)) # wikipedia

          %matplotlib inline
```

# Part I

# Pandas

**References:**

- http://pandas.pydata.org/
- Cookbook — pandas 0.14.0 documentation

## 1 Reading CSV files

```
In [21]:  D = pd.read_csv("amazonianBirds.csv")#, sep = ',') # data obtaied from http://www.xeno
```

## 2 What is in this csv file?

Number of lines, names of the columns?

```
In [22]:  colNames = D.columns.values # names of the columns
          print "* length of the file:", len(D), "\n* names of the columns (%d):\n"%len(colNames
```

```
* length of the file: 8588
* names of the columns (7):
['recordist' 'date' 'time' 'location' 'longitude' 'latitude'
'elevation']
```

Get a taste of the file

```
D.head()  # shows the first 5 lines, try D.tail()
```

In [23]:

Out [23]:
```
     recordist        date   time  \
0   Daniel Lane  2011-02-24  05:55
1   Daniel Lane  2011-02-24  06:05
2  Eric DeFonso  2011-09-03  18:00
3  Eric DeFonso  2011-09-04  06:00
4  Eric DeFonso  2011-09-04  06:05


                                      location  longitude  latitude
\
0  10 km S Pocone on Transpantaneira, Mato Grosso   -56.6480  -16.3620
1  10 km S Pocone on Transpantaneira, Mato Grosso   -56.6480  -16.3620
2                     Pantanal Wildlife Center, MT   -56.8764  -16.7581
3                     Pantanal Wildlife Center, MT   -56.8764  -16.7581
4                     Pantanal Wildlife Center, MT   -56.8764  -16.7581

   elevation
0        115
1        115
2        110
3        110
4        110

[5 rows x 7 columns]
```

# Part II

# Selecting and plotting a column

## 2.1 Selecting a column from the data frame

```
print len(D.longitude)
D.longitude[:3]
```
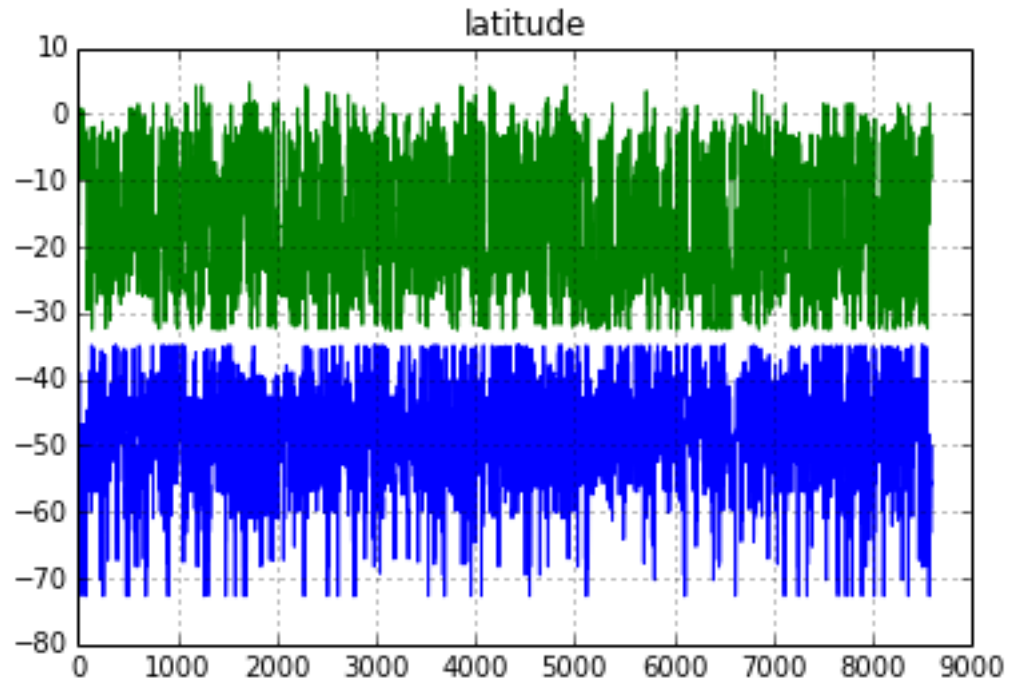
In [24]:

```
8588
0   -56.6480
```

Out [24]:
```
1   -56.6480
2   -56.8764
Name: longitude, dtype: float64
```

# 3 Plotting the data

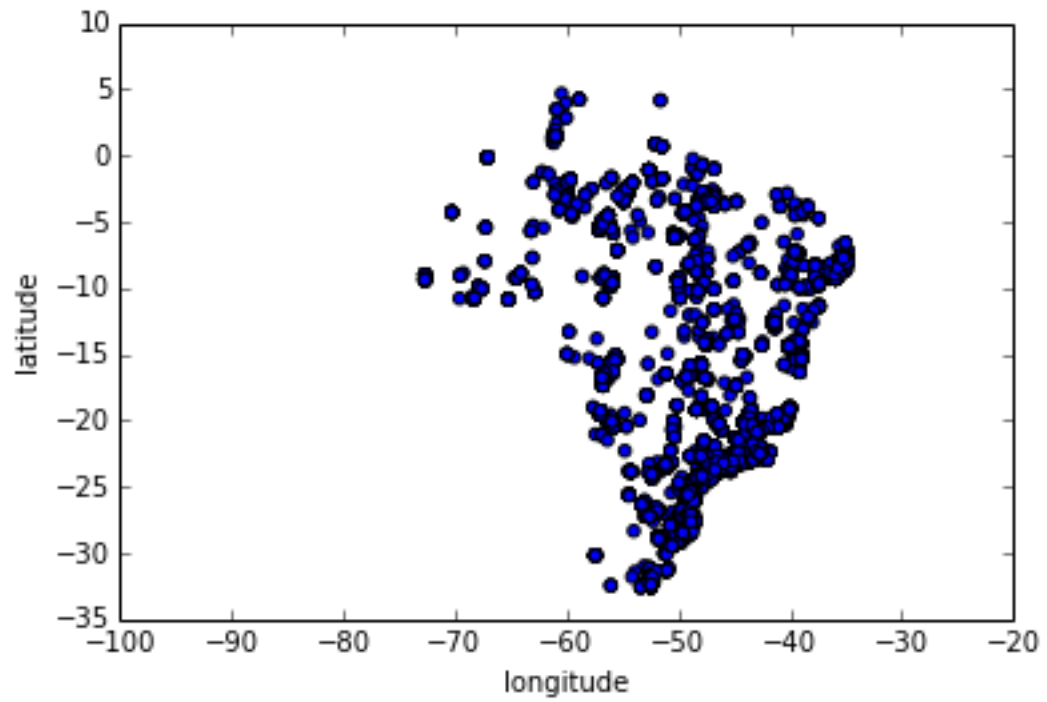## 3.1 Plotting a column

```
In [25]:  D['longitude'].plot( title = 'Longitude')
          D['latitude'].plot( title = 'latitude')
          plt.savefig('images/series.png')
```



## 3.2  Scatter plot with two columns

```
In [26]:  plt.scatter(D.longitude, D.latitude)
          plt.xlabel('longitude')
          plt.ylabel('latitude')
          plt.xlim(-100, -20)

          plt.savefig("scattpl.png")
```
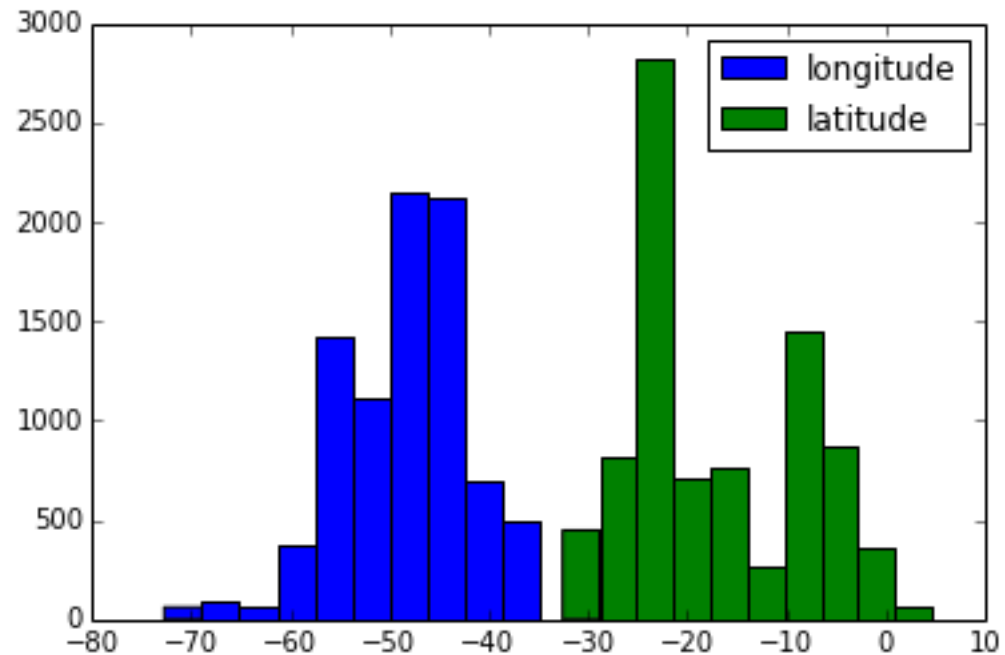
### 3.3 Plotting histograms

In [27]:
```python
x1 = 'longitude'
x2 = 'latitude'
plt.hist(D[x1], label = x1)
plt.hist(D[x2], label = x2)
plt.legend()

plt.savefig("histp.png")
```

# Part III

# Selecting data

### Symmetry

```
In [28]:  D['date'][:5]  # first five elements
          D[:5]['date']
          D.date[:5]
          0    2011-02-24
Out [28]: 1    2011-02-24
          2    2011-09-03
          3    2011-09-04
          4    2011-09-04
          Name: date, dtype: object
```

### Selection of multiple columns

```
In [29]:  D[['date', 'time', 'latitude']][:3]  # first three elements
Out [29]:        date    time  latitude
          0  2011-02-24  05:55  -16.3620
          1  2011-02-24  06:05  -16.3620
          2  2011-09-03  18:00  -16.7581

          [3 rows x 3 columns]
```

**Conditional selection**

```
In [30]:  D_nor = D.loc[D.latitude > 0]
          print len(D_nor)
          D_nor.tail(n=3)
```
```
          90
                             recordist        date   time  \
Out [30]: 8268              Jeremy Minns  2000-12-03  10:51
          8364  Thiago Orsi Laranjeiras  2012-11-07  18:00
          8574  Thiago Orsi Laranjeiras  2009-08-26  08:30

                                                   location  longitude
          latitude  \
          8268                 Porto Grande, AP. Hotel Sonho Meu   -51.4834
          0.6834
          8364  Estrada Perdida, Parque Nacional do Viruá, Car...   -60.9869
          1.4022
          8574  Trilha do Buritizal - Serra do Viruá - Parque ...   -60.9891
          1.4829

                elevation
          8268         60
          8364         50
          8574         90

          [3 rows x 7 columns]
```

# 4  Combination of conditions to select data

**Which recordings were done by Jeremy Minns during the year 2002?**

```
In [31]:  # condition 1: recordings from Jeremy Minns
          JM_recs = D['recordist'] == 'Jeremy Minns'

          # condition 2: recordings done in 2002
          data2002 = D.date.str.contains('2002') # string methods"

          # select data
          JM2002 = D[JM_recs & data2002]

          JM2002.head(n=3)
```
```
                recordist        date   time                 location
Out [31]: longitude  \
          8    Jeremy Minns  2002-09-26  17:30  Anavilhanas Archipelago
          -60.7501
          9    Jeremy Minns  2002-09-27  15:00          Rio Caurés, AM
          -62.2167
          10   Jeremy Minns  2002-09-27  15:50          Rio Caurés, AM
          -62.2167

               latitude elevation
          8     -2.6834        21
          9     -1.2667        21
          10    -1.2667        21
```

```
[3 rows x 7 columns]
```

# 5 Visualize the distribution of discrete values

## 5.1 Who is the recordist with the largest number of observations?

'value_counts()' returns the values of the given series sorted from the most frequent to the least

In [32]:
```python
cts = D['recordist'].value_counts()
print "Number of recordists", len(cts)
cts[:10]   # the top ten recorders
```

Out [32]:
```
Number of recordists 232
Jeremy Minns          1522
GABRIEL LEITE          571
Antonio Silveira       326
Ricardo Gagliardi      324
Noé Eiterer            279
Marcos Melo            273
Adrian Eisen Rupp      265
Nick Athanas           241
Eric DeFonso           239
Evair Legal            201
dtype: int64
```

In [33]:
```python
# plot
cts[:10].plot(kind='bar')
```

Out [33]:
```
<matplotlib.axes.AxesSubplot at 0x7f716f0d3bd0>
```
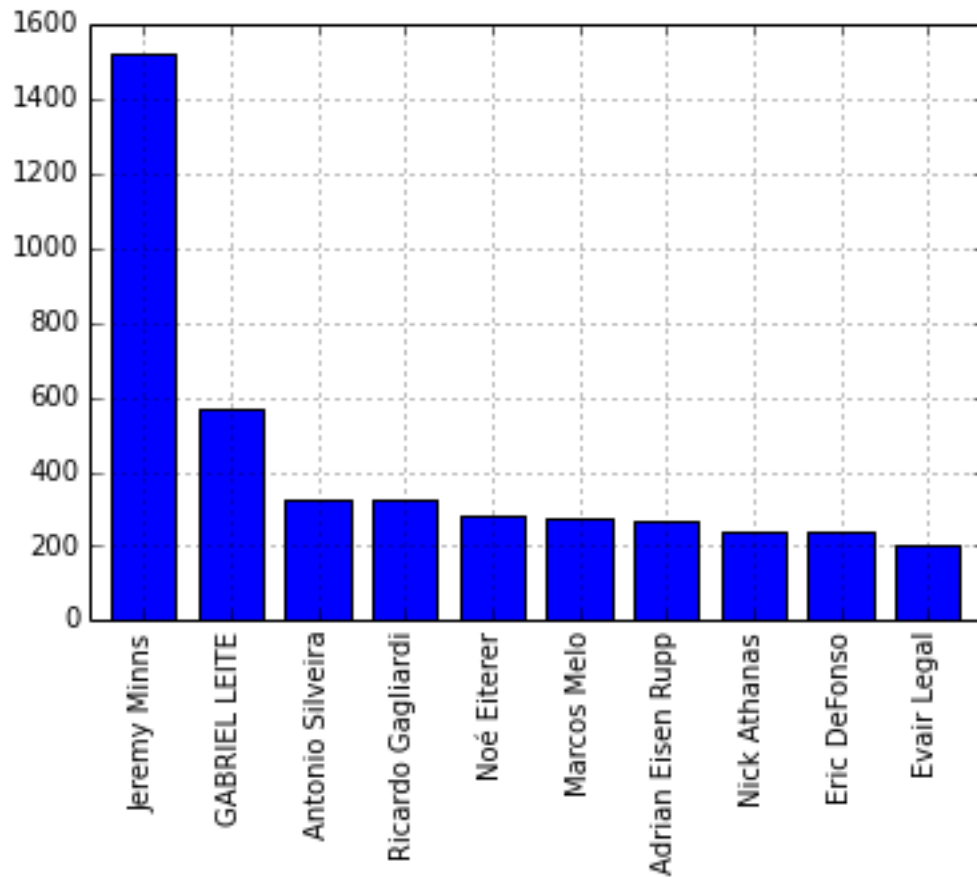
**Part IV**

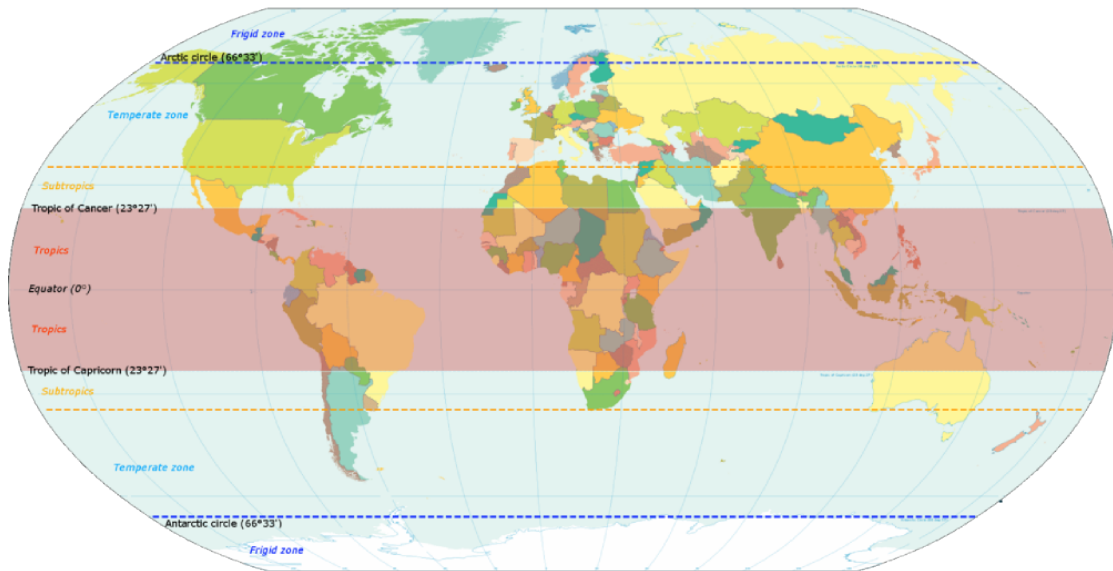# Add new columns

## 6 Determine the climate region of the birds

```
tropics = Image('./images/world.png', width=(300,300))
#http://en.wikipedia.org/wiki/Subtropics#mediaviewer/File:World_map_indicating_tropics
```
In [48]:

```
tropics
```
In [49]:

Out [49]:

# 7 Adding columns

```
In [36]: tropT = -23.4378 # latitude of the tropic of Capricorn
         # generate an array with the values of the new column
         climateRegion = ['tropical' if item > tropT else 'subtropical' for item in D.latitude]
         # define que new column
         D['climate'] = climateRegion
         D.head()
```

```
Out [36]:        recordist         date    time  \
         0   Daniel Lane   2011-02-24  05:55
         1   Daniel Lane   2011-02-24  06:05
         2  Eric DeFonso   2011-09-03  18:00
         3  Eric DeFonso   2011-09-04  06:00
         4  Eric DeFonso   2011-09-04  06:05


                                             location  longitude  latitude
         \
         0  10 km S Pocone on Transpantaneira, Mato Grosso   -56.6480  -16.3620
         1  10 km S Pocone on Transpantaneira, Mato Grosso   -56.6480  -16.3620
         2                    Pantanal Wildlife Center, MT   -56.8764  -16.7581
         3                    Pantanal Wildlife Center, MT   -56.8764  -16.7581
         4                    Pantanal Wildlife Center, MT   -56.8764  -16.7581


            elevation   climate
         0        115  tropical
         1        115  tropical
         2        110  tropical
         3        110  tropical
         4        110  tropical

         [5 rows x 8 columns]
```
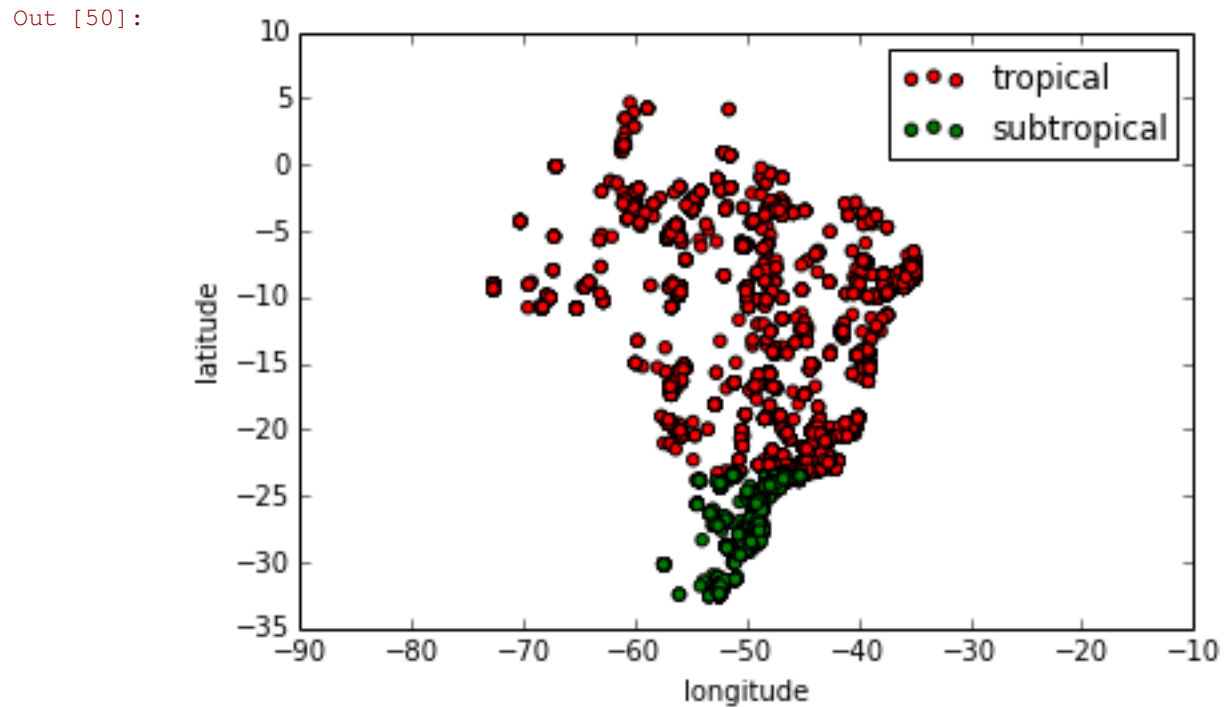
```
        DsubT = D[D.climate == 'subtropical']
In [50]: DT = D[D.climate == 'tropical']

        # plot
        plt.scatter( DT.longitude, DT.latitude, c = 'r', label = 'tropical')
        plt.scatter( DsubT.longitude, DsubT.latitude, c = 'g', label = 'subtropical')
        plt.xlabel('longitude')
        plt.ylabel('latitude')
        plt.legend()
        plt.xlim(-90, -10)
        (-90, -10)
```

Out [50]:

# 8 Save as csv

```
        D.to_csv('amazonianBirds_climate.csv', index = False)
In [38]:
```

```
        # read what we just saved
In [39]: D_new = pd.read_csv('amazonianBirds_climate.csv')
        D_new.head()
```

```
            recordist        date   time  \
Out [39]: 0  Daniel Lane   2011-02-24  05:55
        1  Daniel Lane   2011-02-24  06:05
        2  Eric DeFonso  2011-09-03  18:00
        3  Eric DeFonso  2011-09-04  06:00
        4  Eric DeFonso  2011-09-04  06:05


                                      location  longitude  latitude
        \
        0  10 km S Pocone on Transpantaneira, Mato Grosso   -56.6480   -16.3620
        1  10 km S Pocone on Transpantaneira, Mato Grosso   -56.6480   -16.3620
        2                   Pantanal Wildlife Center, MT   -56.8764   -16.7581
        3                   Pantanal Wildlife Center, MT   -56.8764   -16.7581
```

```
4                       Pantanal Wildlife Center, MT   -56.8764  -16.7581

   elevation   climate
0       115  tropical
1       115  tropical
2       110  tropical
3       110  tropical
4       110  tropical

[5 rows x 8 columns]
```

# 9 Group

```
engCounts = D.groupby('climate')
engCounts.head(n=3)
```

```
                                recordist        date    time  \
climate
tropical    0            Daniel Lane  2011-02-24  05:55
            1            Daniel Lane  2011-02-24  06:05
            2           Eric DeFonso  2011-09-03  18:00
subtropical 86    Adrian Eisen Rupp  2009-09-30       ?
            87  Bernabe Lopez-Lanus  2007-07-07     ?:?
            88        Frank Lambert  1997-12-17     ?:?


                                                  location
longitude  \
climate
tropical    0   10 km S Pocone on Transpantaneira, Mato Grosso
-56.648000
            1   10 km S Pocone on Transpantaneira, Mato Grosso
-56.648000
            2                    Pantanal Wildlife Center, MT
-56.876400
subtropical 86      FLONA de Chapecó, Guatambu, Sana Catarina
-52.778000
            87                 SF de Paula, Rio Grande do Sul
-50.583889
            88                    Santa Catarina, near Urubici
-49.580000


              latitude elevation      climate
climate
tropical    0 -16.362000       115     tropical
            1 -16.362000       115     tropical
            2 -16.758100       110     tropical
subtropical 86 -27.100000       573  subtropical
            87 -29.447778         ?  subtropical
            88 -27.830000      1000  subtropical

[6 rows x 8 columns]
```

```
bye
```

In [56]:

Out [56]:



In [ ]: