Data Science Tutorial 2: Plotting

The Matplotlib can be used to plot charts. In this example we will analyse four datasets, and use then to chat trends. In this case we will use a scatter plot to show points on a chart.

1 Health in the US

We normally read in the data with a CSV file. We will first be using a US data set on health and social care. It analyses US states, and defines metrics such as Infant Mortality Rate (MR), Health Disease DR, and so on. This dataset is contained in the dataset folder or can be downloaded here:

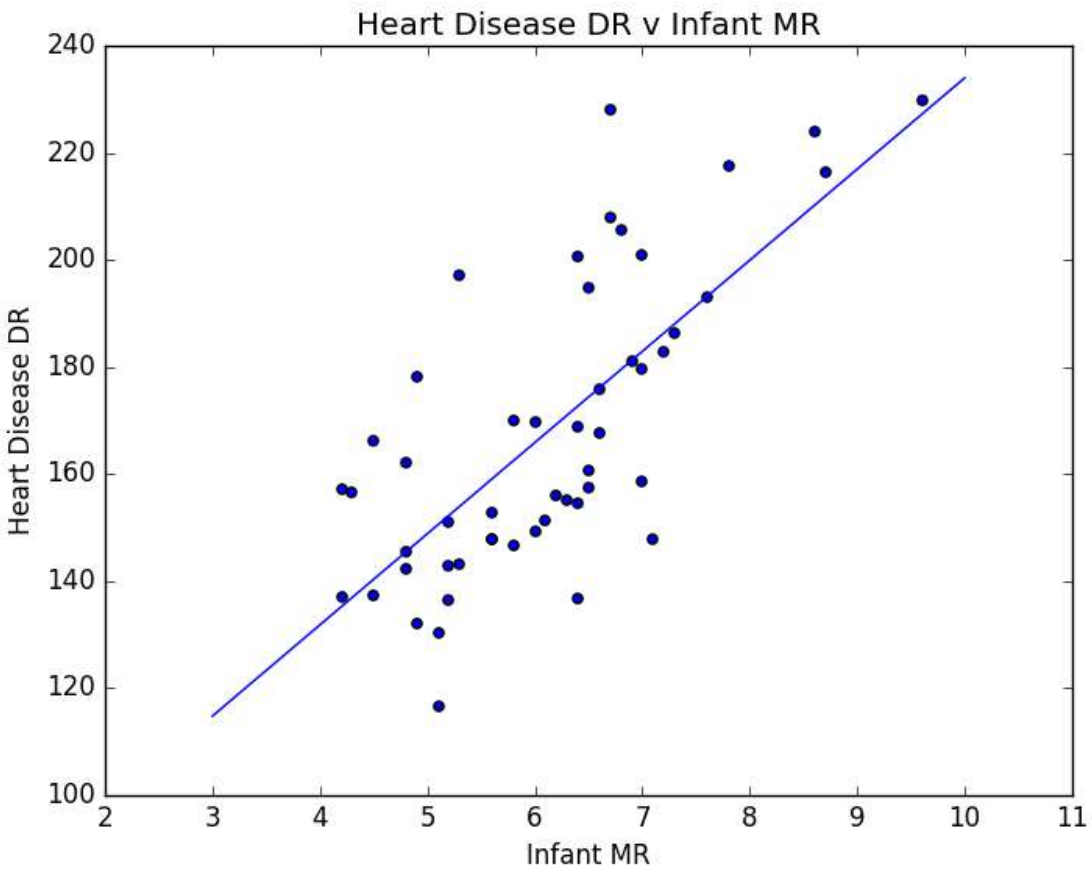
```
http://asecuritysite.com/log/df.csv
```

Example results are shown at: http://asecuritysite.com/bigdata/pandas02

First use **nano** and create a file named data02.py, and add the following lines:

```
import numpy as np
import pandas as pd
import sys
import matplotlib.pyplot as plt

xval = 'Infant MR';
yval = 'Heart Disease DR';

file='1111'

ver=pd.read_csv("datasets/df.csv")

plt.xlabel(xval)
plt.ylabel(yval)

plt.scatter(ver[xval],ver[yval])
plt.show()

f2= file+".svg"

plt.savefig(f2,format='SVG')

f2= file+".png"

plt.savefig(f2,format='PNG')
```

1. We should now have a char of Heart Disease DR and Infant DR, and created two graphics files (with .png and .svg extensions). Show that you generate the following chart:



2. Check that you can see the scatter plot. Now plot “Suicide DR” and “Cancer DR”.
3. The graph above has an estimation of the trend. For this we will use numpy and determine the slope and plot it. Add the required code:

```
import numpy as np
import pandas as pd
import sys
import matplotlib.pyplot as plt
import statsmodels.api as sm

xval = 'Infant MR';
yval = 'Heart Disease DR';

file='1111'

ver=pd.read_csv("datasets/df.csv")
```

```
plt.title(yval+' v ' + xval)
plt.xlabel(xval)
plt.ylabel(yval)

plt.scatter(ver[xval],ver[yval])
plt.show()

axes = plt.gca()
m, b = np.polyfit(ver[xval], ver[yval], 1)
X_plot = np.linspace(axes.get_xlim()[0],axes.get_xlim()[1],100)
plt.plot(X_plot, m*X_plot + b, '-')

if (b>0):
      print yval,'=',round(m,3),' x ',xval,'+',round(b,3)
else:
      print yval,'=',round(m,3),' x ',xval,round(b,3)

print sm.OLS(ver[xval], ver[yval]).fit().summary()

f2=  file+".svg"
plt.savefig(f2,format='SVG')
f2=  file+".png"
plt.savefig(f2,format='PNG')
```

4. Now plot the following and see if the line for the approximation fits:

Infant MR v Stroke DR Outline the linear equation:

By observing the chart, is there a good correlation between the two?

Is it a positive or negative correlation?

Heart Disease DR v Suicide DR Outline the linear equation:

By observing the chart, is there a good correlation between the two?

Is it a positive or negative correlation?

Homicide DR v Drug Poisoning DR Outline the linear equation:

By observing the chart, is there a good correlation between the two?

Is it a positive or negative correlation?

Motor Vech DR Cancer DR Outline the linear equation:

By observing the chart, is there a good correlation between the two?

Is it a positive or negative correlation?

2 Crime in the US Cities

This dataset is contained in the dataset folder or can be downloaded here:

```
http://asecuritysite.com/log/city.csv
```

Example results are shown at: http://asecuritysite.com/bigdata/pandas02_crime

Next analyse the data for the following:

1. Now plot the following and see if the line for the approximation fits:

Violent Crime v Murder Outline the linear equation:

By observing the chart, is there a good correlation between the two?

Is it a positive or negative correlation?

Rape v Robbery Outline the linear equation:

By observing the chart, is there a good correlation between the two?

Is it a positive or negative correlation?

Aggravated Assault v Property Crime Outline the linear equation:

By observing the chart, is there a good correlation between the two?

Is it a positive or negative correlation?

3 Education in the US

This dataset is contained in the dataset folder or can be downloaded here:

```
http://asecuritysite.com/log/edu.csv
```

Example results are shown at: http://asecuritysite.com/bigdata/pandas_edu_plot

Next analyse the data for the following:

1. Now plot the following and see if the line for the approximation fits:

High school graduate v Bachelor degree Outline the linear equation:

By observing the chart, is there a good correlation between the two?

Is it a positive or negative correlation?

Advanced degree v Household income Outline the linear equation:

By observing the chart, is there a good correlation between the two?

Is it a positive or negative correlation?

Cost of living v Unemployment Outline the linear equation:

By observing the chart, is there a good correlation between the two?

Is it a positive or negative correlation?

4 Gun Ownership in the US

This dataset is contained in the dataset folder or can be downloaded here:

```
http://asecuritysite.com/log/guns.csv
```

Example results are shown at: http://asecuritysite.com/bigdata/pandas_gun_plot

Next analyse the data for the following:

1. Now plot the following and see if the line for the approximation fits:

Gun ownership v Murders per 100K Outline the linear equation:

By observing the chart, is there a good correlation between the two?

Is it a positive or negative correlation?

Gun murders per 100K v Firearms murders as % of all murders Outline the linear equation:

By observing the chart, is there a good correlation between the two?

Is it a positive or negative correlation?

Murders per 100K v Firearms murders per 100,000 population Outline the linear equation:

By observing the chart, is there a good correlation between the two?

Is it a positive or negative correlation?