Data Science Tutorial 1: Using Pandas

Pandas is a library which allows data to be analysed, and is often used in data analysis. In this example we will analyse four datasets. Demonstration: https://youtu.be/TE_CJLFn4jo

The login for myfabix is:

- Username: datalabs
- Password: Foxtrot123!

The login for the VMs is:

- Username: Datalabs User
- Password: datalabs

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/pandas

First use **nano** and create a file named data01.py, and add the following lines:

```
import pandas as pd
ver=pd.read_csv("df.csv")
```

1. We now have an object named **ver** that we can analyse. First view the first three lines of the data:

```
print ver.head(3)
```

Outline the data contained:

2. We can see that our columns are defined with the US state, then Infant Mortality Rate, Heart Disease Death Rate, and so on. Let's now examine the number of rows and columns that we have. For this we can use the len(ver):

```
print len(ver)
```

How many records are there:

3. We can also view the columns used with the ver.columns property:

```
print ver.columns
```

What are the names for the columns:

4. Next we can look at the data type of the data in the columns (using the ver.dtypes property):

```
print ver.dtypes
```

5. In this case we see that most of our data columns are defined with values as floating point numbers.

Which are the data types used for each column:

6. Next, if we want to pin-out a specific column, such as the region/state:

```
print ver['RegionState']
```

Outline the first five values of this column:

7. We can the access the column data for Infant Mortality Rates with: (ver['Infant MR']):

```
print ver['Infant MR']
```

Outline the first five values of this column:

8. Where we are listing Infant Mortality Rates, and we can look at two columns together and to include the state/region:

```
print ver[['RegionState','Infant MR','Heart Disease DR']]
```

Outline the first five values of this merge:

9. If we want to see the basic stats for the data:

```
print ver.describe()
```

For Infant MR, outline the following values:

count:

mean:

std:

min:

25%:

50%:

75%:

max:

10. And so we can see that the average Infant Mortality Rate is 6.1 with a standard deviation of 1.17. The lowest Infant MR is 4,2 and the highest is 9.6. Next we want to find out the Top 3 states with the lowest Infant Mortality Rates:

```
print ver.sort(['Infant MR']).head(3)
```

Outline these three values? Which state(s) has the least problems with Infant MR:

11. And we see that Massachusetts has the lowest with 4.2. Now we can do a reserve (descending list) to analyse the highest Infant Mortality Rate:

```
print ver.sort(['Infant MR'],ascending=False).head(3)
```

Outline these three values? Which state(s) has the most problems with Infant MR:

12. In this case Massachusetts has the best mortality rate for infants, but Mississippi has the worst. You should also be able to see that the Heart Disease Death Rate is also much lower in Massachusetts than Mississippi. In fact the Top 3 in both cases seem to differ greatly.
13. So is there a correlation between the Infant Mortality Rate and Heart Disease? With correlation, we get a value of +1.0 or -1.0 when there is a strong correlation, and 0.0 when there is no correlation. We can now run:

```
print ver.corr()
```

Outline three pairs of parameters which have a strong correlation:

Outline three pairs of pameters which have a weak correlation:

14. We can now see that for Infant MR that there is a strong correlation with Heart Disease DR (Death Rate), Stroke Rate DR and Cancer DR, while there is no direct correlation between the Suicide DR and Infant MR. There is, as expected, a strong correlation between Heart Disease DR and the Stroke DR (0.664). The Suicide Rate MR does not have a strong correlation with most factors, apart from Motor Vehicle Death Rates.
15. If we want, we can just take Infant MR and Heart Disease DR:

```
print ver[['Infant MR','Heart Disease DR']].corr()
```

Outline output:

16. Create a command which analyses the correlation in the average income against the other factors:

Command used:

Which has the strongest correlation with average income:

17. You should see the strongest correlation is between the Stroke Death Rate and Average Income. In this case there is a negative correlation, so that as the Stroke Death Rate goes up the Average Income goes down.
18. We can also look at the covariance which is a measure related to how much two variables change in the same way. The larger the value, the greater the significance:

```
print ver.cov()
```

Which parameters have the strong covariance:

19. and where we can see that there is a strong covariance between the Infant MR and Heart Disease DR and Drug Poisoning DR. As we would expect Heart Disease DR has a strong covariance with Stroke DR (104.6).
20. We can also analysis for thresholds, such as for states with an Infant MR rate is greater than 4.5:

```
print ver['Infant MR']> 4.8
```

Which are the first three index values that have an Infant MR greater than 4.9

21. We now want to see the US states who have an Infant MR greater than 7.0:

```
print ver[(ver['Infant MR'] > 7.0)]
```

Which are the first three US states who have an Infant MR greater than 7.0:

22. Now we will analyse those states with an Infant MR greater than 7.0, and who have a Heart Disease DR higher than 45.0:

```
print ver[(ver['Infant MR'] > 7.0) &  (ver['Stroke DR'] > 45.0 )]
```

Outline the top three US states for this:

23. Based on this data, in the two areas analysed, we now see that it is Alabama, Arkansas, Louisiana, MIssissippi and West Virginia that have the great health problems.
24. We can now perform some linear regression on two of the parameters (Infant MR and Heart Disease) and look at the fit. For this, import the **statsmodel.api** library:

```
import statsmodels.api as sm
print sm.OLS(ver['Infant MR'], ver['Heart Disease DR']).fit().summary()
```

Outline the parameters of the result:

R-squared:

Model:

Adj. R-squared:

Method:

F-statistic:

Prob (F-statistic):

Log-Likelihood:

25. Now let's import **numpy** which fits to a straight line and calculate the gradient (m) and the point it cuts the y-axis:

```
import numpy as np
m, b = np.polyfit(ver['Infant MR'], ver['Heart Disease DR'], 1)
print 'Infant MR = '+str(round(m,3))+' * Heart Disease DR + ',str(round(b,3))
```

Outline the linear relationship:

Gradient: **Point it cuts y-axis:**

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/pandas_crime

Next analyse the data for the following:

Data columns and data types:

Which US city has the highest population:

Which US city has the highest violent crime rate:

Which US city has the highest burglary rate:

Which US city has the highest arson rate:

Which US city has the lowest population:

Which US city has the lowest violent crime rate:

Which US city has the lowest burglary rate:

Which US city state has the lowest arson rate:

Top three correlation pairs in the data:

Linear regression equation for the top correlation in the data:

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

Next analyse the data for the following:

Data columns and data types:

Which US state has the best academic attainment:

Which US state has the poorest academic attainment:

Top three correlation pairs in the data:

Linear regression equation for the top correlation in the data:

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

Next analyse the data for the following:

Data columns and data types:

Which state has the highest gun ownership:

Which state has the lowest gun ownership:

Top three correlation pairs in the data:

Linear regression equation for the top correlation in the data: