# 《数据挖掘导论》实验 3: 可视化分析实验

# 一、实验目的：

（1）了解matplotlib的绘图组件

（2）掌握pandas常用的绘图方法

# 二、实验环境：

（1）Anaconda2 开发环境

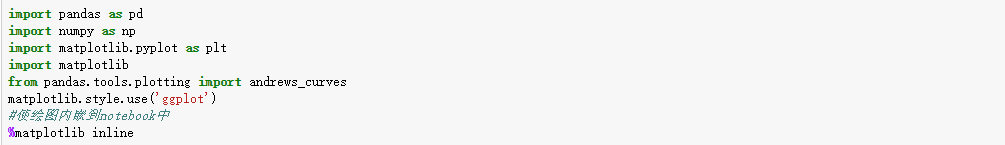
（2）IDE是ipython notebook

（3）使用的库有numpy，pandas，matplotlib

# 三、实验内容：

1. 皮马印第安人糖尿病数据可视化

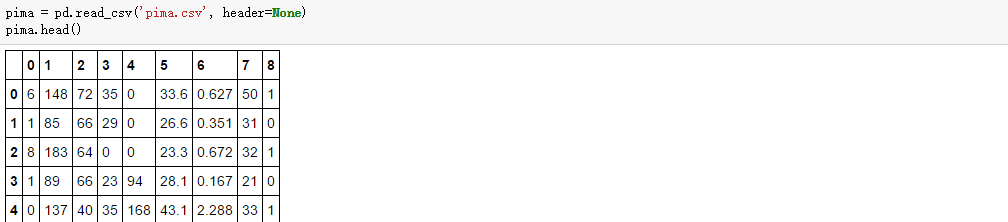
* 1. 导入本次实验需要用的库



图一

图一中引入我们本次实验需要引入的包，并导入相应的配置，如内嵌绘图至ipython notebook中，使用ggplot等。

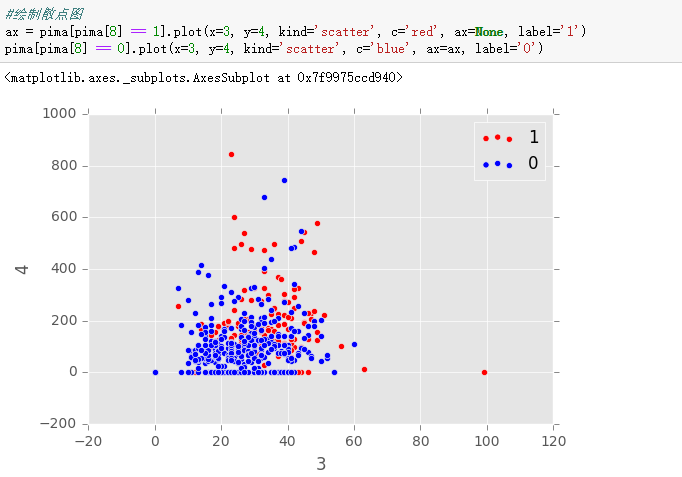
* 1. 打开数据源文件



图二

图二中，我们打开了pima.csv文件，并设置其为无头数据源（pima数据源不包含头部解释），并显示了pima.csv中的头五个字段，图二下方即为生成的内容。

* 1. 绘制散点图



图三

在这里，我们选用了第四行以及第五行作为我们的x轴和y轴，并将其分为两类进行对比，1类为红色，0类为蓝色。此处的绘制的散点图如图三所示。

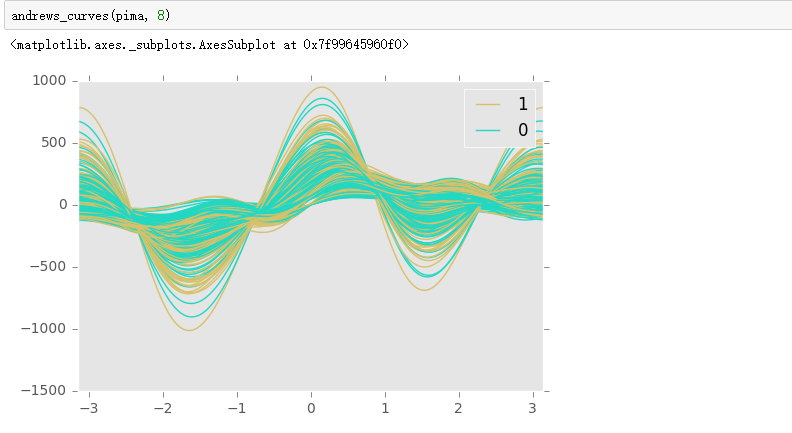
* 1. 使用全部字段绘制散布图



图四

此处先将0类换做红色，1类换做蓝色，然后调用pandas中的绘制图方法，指定绘图后绘制如图四所示的散布图，图中能显示出全部数据的散布关系。

* 1. 绘制调和曲线图



图五

这里直接调用andrews\_curves来绘制调和曲线图，指定第九行为我们存放类别的名字，图五所示即为绘制出来的调和曲线图。

# 四、实验小结：

此次实验我们演示以及熟悉了numpy、pandas以及matplotlib的常用操作，熟悉这些库的操作对我们研究数据科学有很大的帮助。