# 数据分析

基于Pandas和Numpy两种工具，分析给定数据集包含的特征。主要包括数据探索、数据清洗等，其目的在于保持数据的“整洁”，为后续应用机器学习模型做准备。

整个数据分析流程如下：

1、加载数据：可以直接从网站下载，也可以使用numpy或者Pandas、python等本地加载

2、观察和理解数据：数据集大小，数据集的饱和度，数据集中各个属性的含义、数值分布，正负标签比例等。

3、数据清洗

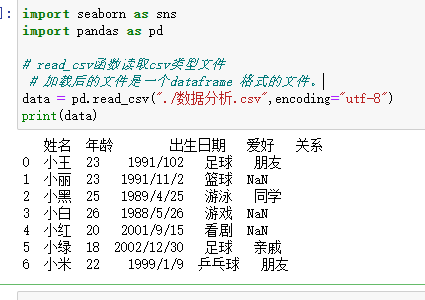
## 1. 加载数据

加载数据使用pandas 库中的read\_csv函数，read.csv 函数的意义是，读取一个csv类型的文件。

例如：csv表中内容。



使用read.csv 函数加载后默认的的输出结果

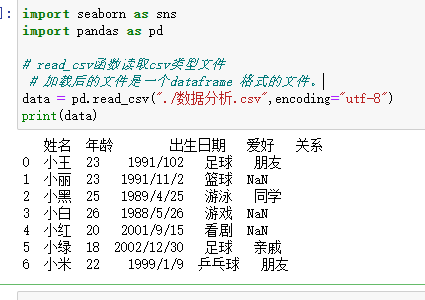


read\_csv函数中的参数

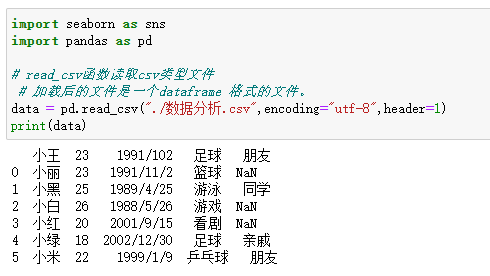
* encoding 用于指定字符编码
* header 用于设置DataFrame的列名，默认为infer。
* names 用于自定义列明

例如：

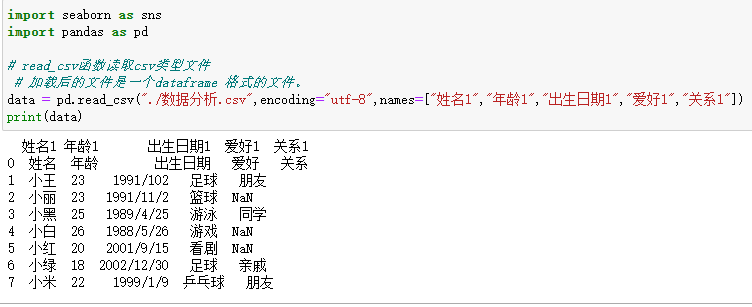
1. names 没有被赋值，header 也没赋值（默认第一列为DataFrame数据的列名）



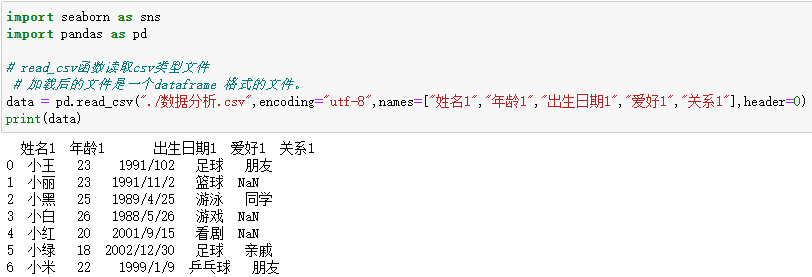
1. names 没有被赋值，header = 1 指定第一行为列名



1. names 被赋值，header 没有被赋值（用于自定义DataFrame数据的列名）



1. names 和 header 同时被赋值（其中names 用于自定义 DataFrame 数据的列名名替换，header 指定的行值）

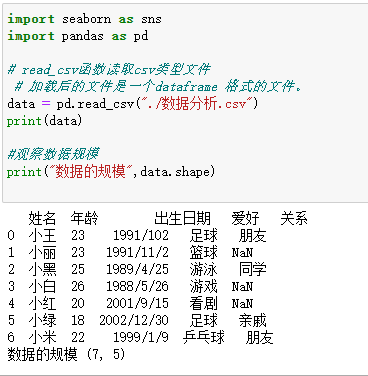


## 2. 观察数据

### 2.1 shape函数

DataFrame.shape 函数主要是用于观察数据的规模，统计数据的行列数。

例如：shape 输出行和列数，shape[0]输出第一个元素（行数），shape[1]输出第二个元素（列数）。

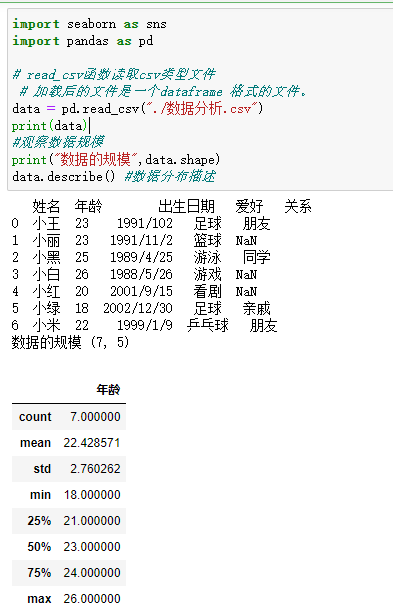


### 2.2 describe() 函数

DataFrame.describe() 参数列表以及相关意义。

DataFrame.describe(percentiles=None, include=None, exclude=None)

* include 用于指定描述数据的类型，默认不写则是对数字类型进行描述，而include= “O“ 则是描述object类型的熟悉， include= ”all“则是对所有属性的描述。

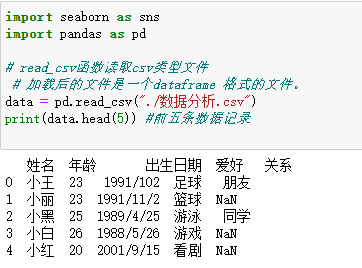


对于数值数据生成的结果集：count：总数；mean：均值；std：标准差；min：最小值；25%：四分之一分位数；50%：中位数；75%：四分之三分位数；max：最大值。

### 2.3 head()函数

DataFrame.head()函数用于显示数据的前几行

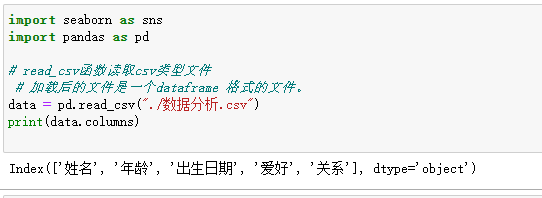
例如：显示数据的前5行



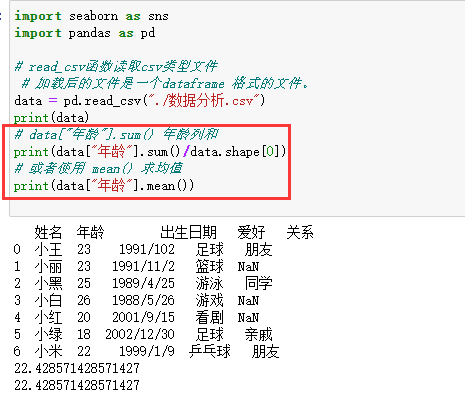
### 2.4 columns函数

DataFrame.columns 用于列出所有的字段

例如：



### 2.5 求列均值



### 2.6 value\_counts()

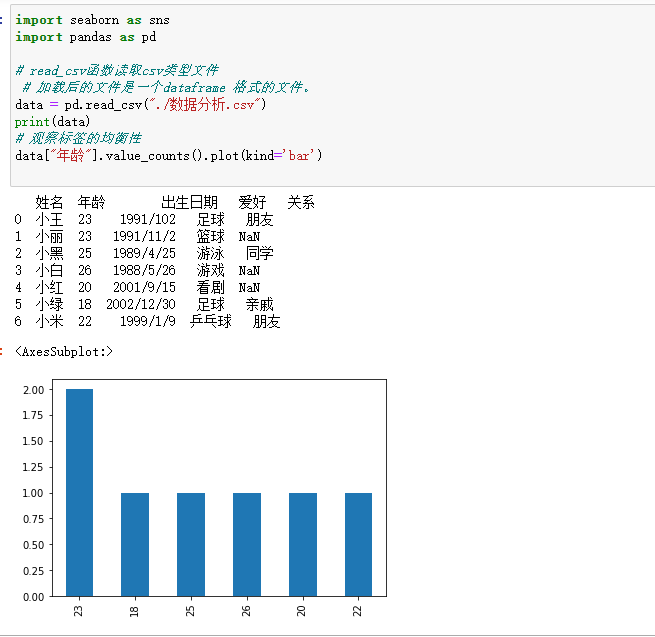
在pandas中，value\_counts常用于数据表的计数及排序，它可以用来查看数据表中，指定列里有多少个不同的数据值，并计算每个不同值有在该列中的个数，同时还能根据需要进行排序。

**函数体及主要参数：**

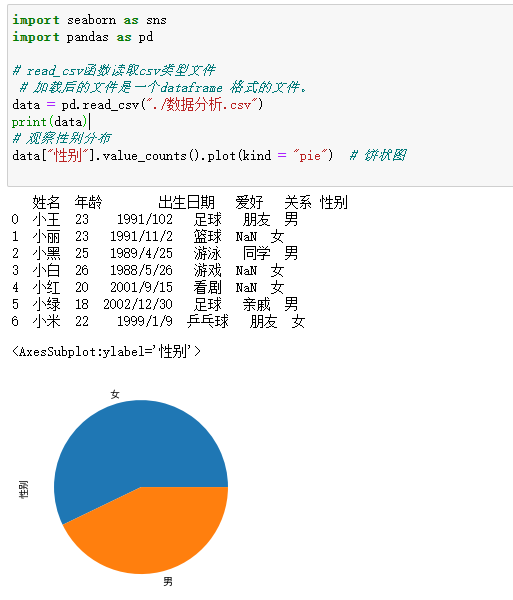
value\_counts(values,sort**=True**,ascending**=False**,normalize**=False**,bins**=None**,dropna**=True**)

**sort=True**： 是否要进行排序；默认进行排序  
**ascending=False：** 默认降序排列；  
**normalize=False**： 是否要对计算结果进行标准化并显示标准化后的结果，默认是False。  
**bins=None：** 可以自定义分组区间，默认是否；  
**dropna=True：**是否删除缺失值nan，默认删除。

例题：统计每个年龄的人数用柱状图表示



例题：统计性别，用饼状图表示



总结：plot 中kind参数

kind =

直线图：line(默认值)

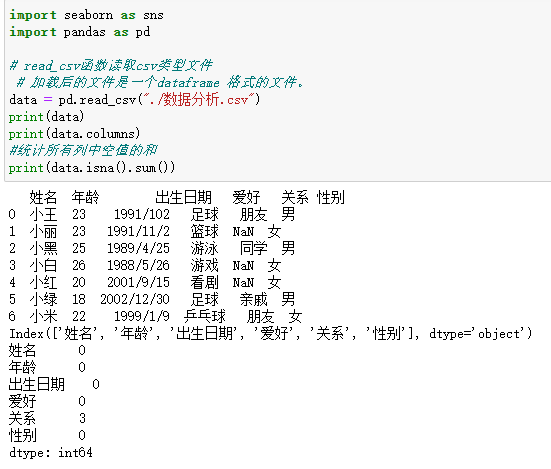
直方图：bar

饼状图：pie

### 2.7 isna（）函数

DataFrame.isna() 主要是用于找空值（缺省值）

例如

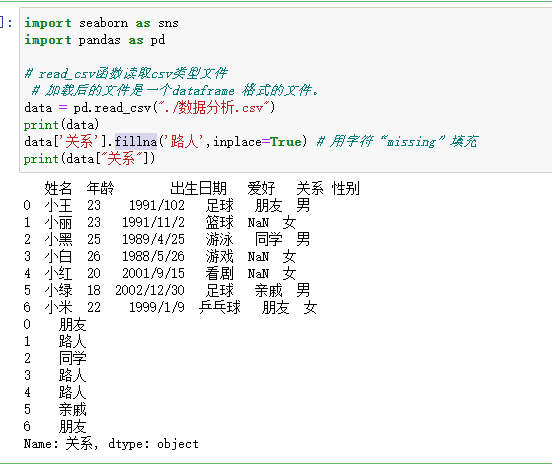


## 3. 数据的清洗

### 3.1 fillna（）函数

DataFrame.fillna()用于填充缺省值。

例题：使关系列中的缺省值用“路人”填充



inplace=True：不创建新的对象，直接对原始对象进行修改；

inplace=False：对数据进行修改，创建并返回新的对象承载其修改结果。

fillna()函数只是修改了返回值，数据源结构并没有发生改变。

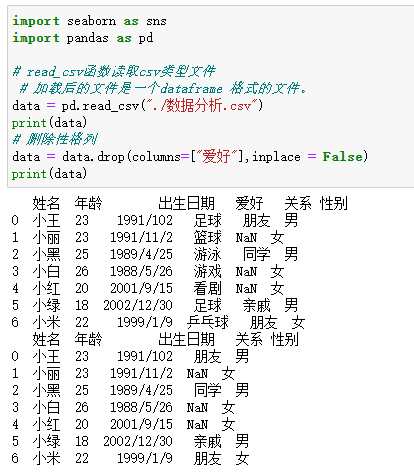
### 3.2 drop()函数

pandas中drop()函数：删除 Series 的元素或 DataFrame 的某一行（列）

格式：DataFrame.drop(labels=None, axis=0, index=None, columns=None, level=None, inplace=False, errors='raise')

**labels**:待删除的行名or列名；   
**axis**:删除时所参考的轴，0为行，1为列；   
**index**:待删除的行名   
**columns**:待删除的列名   
**level**:多级列表时使用，暂时不作说明   
**inplace**:布尔值，默认为False,这是返回的是一个copy;若为True,返回的是删除相应数据后的版本

例题：删除列



删除行

