**一些著名的数据科学公开数据集与数据源**

### 藏在Python库里的数据集

一些可视化库和机器学习库有着内置数据集的传统。因为库的文档和案例通常会使用一些数据集来举例、内置数据集后方便用户学习该库的可视化语法，方便复现效果。

Python里内置数据集的库有： plotnine、ggplot、Altair、seaborn、bokeh、plotly、sklearn等等。

**plotnine**里：通过import plotnine.data as pnd; pnd.\_\_all\_\_语句列出可以直接调用的数据集：

\_\_all\_\_ = ['diamonds', 'economics', 'economics\_long', 'midwest', 'mpg', 'msleep', 'presidential', 'seals', 'txhousing', 'luv\_colours', 'faithful', 'faithfuld', 'huron', 'meat', 'mtcars', 'pageviews']

通过 df=pnd.diamonds 语句载入数据。 数据表实际保存在 *site-packages\plotnine\data* 文件夹下，都是csv格式存在本地。

**ggplot**提供的数据集如下：

chopsticks, diamonds, mtcars, meat, pageviews, pigeons, movies, mpg, salmon, load\_world

通过 import ggplot.datasets as gds;df=gds.diamonds 可以导入对应数据集。 除了world世界地图数据是需要从github下载之外，其他数据集都存在本地，site-packages\ggplot\datasets目录下，也都是csv文件。

**Altair**：数据集在vega\_data里，

from vega\_datasets import data

data.list\_datasets()

#['7zip', 'airports', 'annual-precip', 'anscombe', 'barley', 'birdstrikes', 'budget', 'budgets', 'burtin', 'cars', 'climate', 'co2-concentration', 'countries', ...]

vega\_data里的数据集内容挺多，有统计数据也有地理数据，还有不同数据量的版本，例如flights数据集包含了2k、5k、200k、3m等多个版本。 调用是写：df = data('iris')或者df = data.iris()，数据存在在*Anaconda3/Lib/site-packages/vega\_datasets*目录下，存本地的在local\_datasets.json里有描述。本地存的有csv格式的也有json格式的。

**seaborn**库：seaborn.get\_dataset\_names()列出可以调用的数据集：

import seaborn as sns

sns.get\_dataset\_names()

['anagrams', 'anscombe', 'attention', 'brain\_networks', 'car\_crashes', 'diamonds', 'dots', 'exercise', 'flights', 'fmri', 'gammas', 'geyser', 'iris', 'mpg', 'penguins', 'planets', 'tips', 'titanic']

通过iris=seaborn.load\_dataset('iris')载入数据，不过值得说明的是seaborn库本地初始时是不存着这些数据集的，这个和其他库不同，seaborn调用 load\_dataset() 的时候是从GitHub下载到本地，所以有时候会下载失败，使用时可以从[seaborn-data](https://link.zhihu.com/?target=https%3A//github.com/mwaskom/seaborn-data) 自行下载文件再导入（或者考虑用其他库内置的数据集）。

**plotly**：默认的数据集不多，有7个。

import plotly.express as px

px.data.\_\_all\_\_

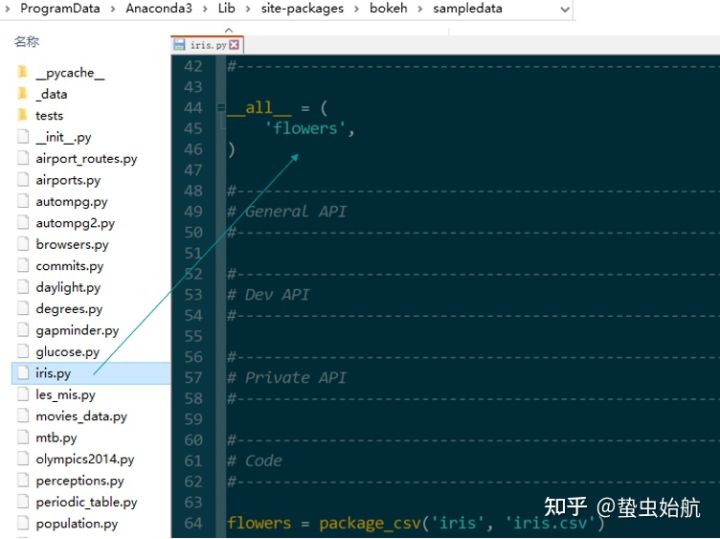
#['carshare', 'election', 'election\_geojson', 'gapminder', 'iris', 'tips', 'wind']

通过df=px.data.gapminder()调用。

**bokeh**：数据集相关文件在site-packages/bokeh/sampledata路径下，它把一些通用的数据集都封装为py文件进行调用，例如iris数据集经过了一层封装叫flowers，载入数据是用from bokeh.sampledata.iris import flowers。 没有语句可以列出有哪些数据集，从代码中拆出来它内置的数据集有：

#files=

['CGM.csv', 'US\_Counties.zip', 'us\_cities.json', 'unemployment09.csv', 'AAPL.csv', 'FB.csv', 'GOOG.csv', 'IBM.csv', 'MSFT.csv', 'WPP2012\_SA\_DB03\_POPULATION\_QUINQUENNIAL.zip', 'gapminder\_fertility.csv', 'gapminder\_population.csv', 'gapminder\_life\_expectancy.csv', 'gapminder\_regions.csv', 'world\_cities.zip', 'airports.json', 'movies.db.zip', 'airports.csv', 'routes.csv']



**sklearn**，内置了一些经典的机器学习练手数据集。

import sklearn.datasets as dts

[i for i in dts.\_\_all\_\_ if 'load' in i]

#输出：

['load\_boston','load\_diabetes','load\_digits','load\_files','load\_iris','load\_breast\_cancer','load\_linnerud', 'load\_mlcomp','load\_sample\_image', 'load\_sample\_images','load\_svmlight\_file','load\_svmlight\_files']

因此其调用方式：df=dts.load\_iris()。

另外，sklearn的数据集有很多种：

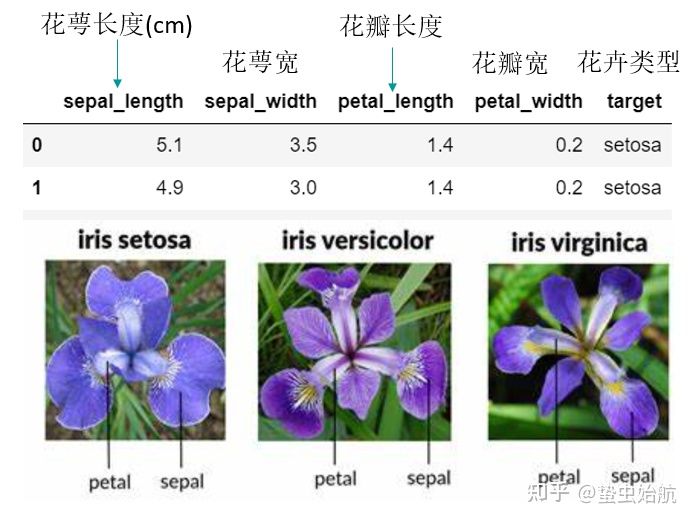
* 自带的小数据集，也就是我们上面列出的可以通过dts.load\_name()直接导入的；
* 需在线下载的数据集，通过dts.fetch\_name()导入；
* 模拟生成的数据集，dts.make\_name()；
* svmlight/libsvm格式的数据集，导入语句：dts.load\_svmlight\_file(name)；
* 从买了[http://data.org](https://link.zhihu.com/?target=http%3A//data.org)在线下载获取的数据集：dts.fetch\_mldata(name)；；

### 著名公开数据集

一些数据科学领域广泛使用（在教程文章、课程练习等场景也广泛使用）的数据集简介。

### iris

鸢尾花数据集。首次出现在著名的英国统计学家和生物学家Ronald Fisher 1936年的论文《The use of multiple measurements in taxonomic problems》中，被用来介绍**线性判别式分析**。 数据集中，包括了三类不同的鸢尾属植物：Iris Setosa，Iris Versicolour，Iris Virginica。 每类50个样本、共150个样本。属性包括花萼长度、宽、花瓣长、宽，单位都是cm。



、如果安装了seaborn、plotly、bokeh这些可视化库，可以通过上一章提到的语句获取iris数据。 一些在线的数据集下载地址如下：

* [https://scikit-learn.org/stable/auto\_examples/datasets/plot\_iris\_dataset.html](https://link.zhihu.com/?target=https%3A//scikit-learn.org/stable/auto_examples/datasets/plot_iris_dataset.html)
* [https://www.kaggle.com/arshid/iris-flower-dataset](https://link.zhihu.com/?target=https%3A//www.kaggle.com/arshid/iris-flower-dataset)

### titanic

声名远扬的泰坦尼克号船员数据集，数据集描述的是船员的性别、年龄、所在船仓等级等特征及最后是否存活。网络上大量的文章是关于泰坦尼克号船员生存率分析的，通过分类或回归算法拟合船员的基本特征与获救情况的关系，甚至一些文章鼓吹将这个项目写入简历。



在seaborn库里可以通过sns.load\_dataset('titanic')获取泰坦尼克数据集。 在线数据源：

* [https://www.kaggle.com/hesh97/titanicdataset-traincsv](https://link.zhihu.com/?target=https%3A//www.kaggle.com/hesh97/titanicdataset-traincsv)

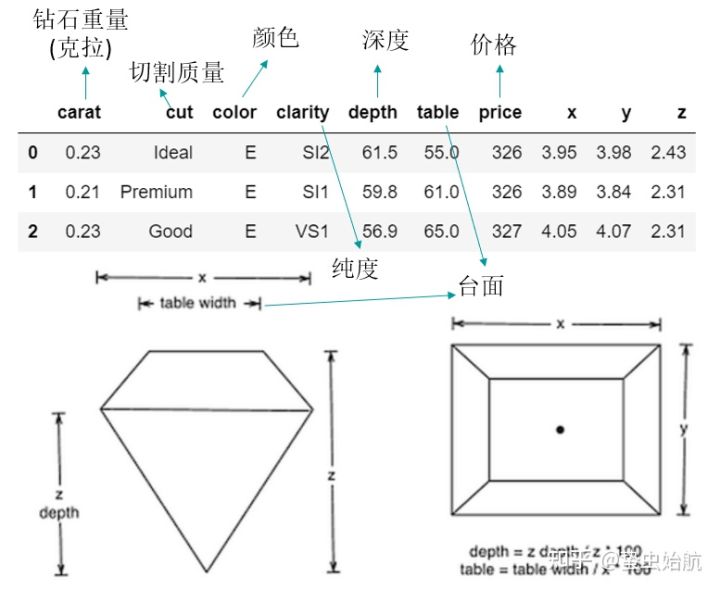
### boston

波士顿房价数据集。经典的用于回归任务的数据集，机器学习类文章大家喜欢写的是波士顿房价预测，热度和泰坦尼克号船员生存率分析接近了。boston数据集有13个特征，包括各类用地比例、师生比例、每居民房子数、可达性指数等，除土地是否在查尔斯河旁边是一个二值变量，其余特征为数值类型。

该数据集在sklearn库里可以直接导入。

### diamonds

钻石数据集。该数据集有10列，csv文件约2.64MB，包含了近54000颗钻石的价格及其他属性。Diamonds数据集在各种R语言数据分析文章里有很高的出镜率，大家普遍用来做探索性数据分析和可视化案例。



具体各特征含义如下：

* 钻石价格：以美元计价；
* 克拉：钻石重量（0.2-5.01）；
* 切割质量：分为公平，良好，非常好，高级，理想 5个等级；
* 台面：钻石顶部宽度相对于最宽点（43-95）；
* 钻石颜色，从J（最差）到D（最好）；
* 纯度：测量钻石的纯净度（I1（最差），SI2，SI1，VS2，VS1，VVS2，VVS1，IF（最佳））；
* X：长度mm（0-10.74）；
* Y：宽度mm（0-58.9）；
* Z：深度mm（0-31.8）；
* 深度：总深度百分比，根据X,Y,Z算出：=z/平均值(x，y)= 2 \* z /(x + y)(43-79)；

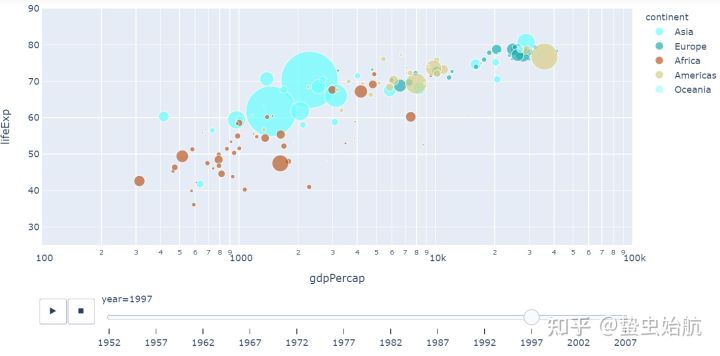
diamonds数据集详细描述： [https://ggplot2.tidyverse.org/reference/diamonds.html](https://link.zhihu.com/?target=https%3A//ggplot2.tidyverse.org/reference/diamonds.html)

包含该数据集的库包括seaborn、plotnine及ggplot。

### gapminder

[Gapminder](https://link.zhihu.com/?target=https%3A//www.gapminder.org/about/) 本身是一家位于瑞典斯德哥尔摩的非盈利机构，其收集和公开了关于世界的经济、教育、环境、健康等专题数据。Hans Rosling的TED演讲用了几个生动的例子展现了gapminder统计数据的魅力。

plotly包含了一个gapminder样本数据集， 在线数据集下载： - [https://www.gapminder.org/data/](https://link.zhihu.com/?target=https%3A//www.gapminder.org/data/)



### wine

葡萄酒数据集，该数据集描述来自意大利同一地区三个不同品种的葡萄酒进行化学分析的结果，每种所含的13种成分的数量。特征包括Alcohol、Ash、Total phenols、Hue等13种。

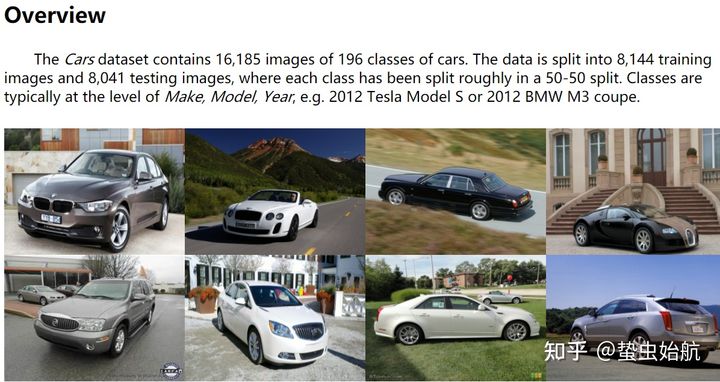
### cars

cars数据集现在有两种可能性，一种是在1983年美国统计协会(ASA)公布的，由Ramos和Donoho收集的406辆汽车数据，包括车重(weight)、马力(horsepower)、没加仑油行驶(MPG)、汽车产地(origin)等特征。



该数据集在Altair库里有，Altair的教程文档里有关于cars数据集的可视化案例： [https://altair-viz.github.io/altair-tutorial/notebooks/01-Cars-Demo.html](https://link.zhihu.com/?target=https%3A//altair-viz.github.io/altair-tutorial/notebooks/01-Cars-Demo.html)

另一种是斯坦福大学整理的16185张汽车图片数据集。

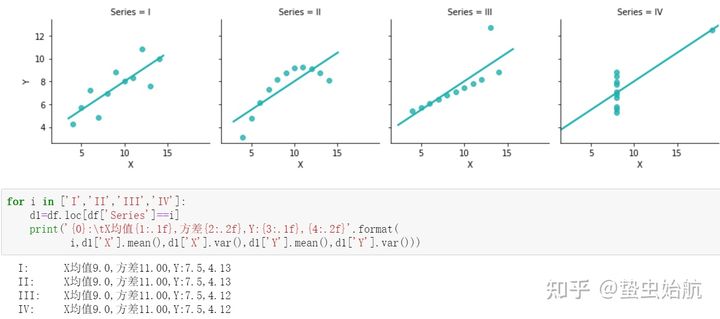


汽车图片数据集下载源：

* [http://ai.stanford.edu/~jkrause/cars/car\_dataset.html](https://link.zhihu.com/?target=http%3A//ai.stanford.edu/~jkrause/cars/car_dataset.html)

### anscombe

安斯库姆四重奏(Anscombe's Quartet) ，是一个用于展现当数据序列差别非常大时一些统计值却相等的著名数据集。Anscombe数据集由统计学家Francis Anscombe在1973年构造出来，一共包含了4组数据，每组11个X-Y散点。这四组数据均值、方差、相关系数及线性回归曲线(mean, variance, correlation, and linear regression lines)都相同，X的平均数都是9.0，Y值的平均数都是7.5；X值的方差都是11.0，Y值的方差都是4.12；X、Y之间的相关系数皆为0.816，线性回归线都是y=3.0+0.5x。从统计值来看，这四个数据集似乎非常接近，但具体分布差别巨大。 这个数据集体现了数据实际分布的可视化的重要性以及用对拟合方式的重要性。文字还是苍白了，可以看图去体会：



I是最“正常”的一组数据，也是这几个统计值擅长勾画的内容，II所反映的事实上是一个精确的二次函数关系，只是在错误地应用了线性模型后，各项统计数字与第一组数据恰好都相同；III描述的是一个精确的线性关系，只是这里面有一个异常值，它导致了上述各个统计数字，尤其是相关度值的偏差；IV则是一个更极端的例子，数据集中在(8,7)附近，但异常值(19,12.5)使得平均数、方差、相关度、线性回归线等所有统计数字全部发生偏差。

*注，上图绘制代码：*

import seaborn as sns

g = sns.FacetGrid(df, col="Series")

g.map(sns.regplot, "X", "Y",color='#1EAFAE',ci=None)

g.add\_legend()

### airports

地理数据，美国机场的坐标点。包含5列，分别为airport code, city, state, latitude, and longitude。

在Altair、bokeh库里能直接导入使用。 数据集公开地址： - [http://ourairports.com/data/](https://link.zhihu.com/?target=http%3A//ourairports.com/data/)

### 热门整合数据源

* **国家统计局数据**： [http://data.stats.gov.cn](https://link.zhihu.com/?target=http%3A//data.stats.gov.cn/) 数据来源中华人民共和国国家统计局，包含了我国经济民生等多个方面的数据，并且在月度、季度、年度都有覆盖，全面又权威。
* Kaggle：一个数据竞赛网站，收集了很多著名和实用的数据集用于训练和竞赛。[https://www.kaggle.com/](https://link.zhihu.com/?target=https%3A//www.kaggle.com/)
* **UCI**：机器学习数据集集合，挺老牌的， [http://archive.ics.uci.edu/ml/index.php](https://link.zhihu.com/?target=http%3A//archive.ics.uci.edu/ml/index.php), 包括经典的Iris、波尔多葡萄酒等数据集；
* 卡内基·梅隆大学(CMU)统计学院网站的整合数据集页面： [http://lib.stat.cmu.edu/datasets/](https://link.zhihu.com/?target=http%3A//lib.stat.cmu.edu/datasets/)
* 机器学习数据集整合：
* [Machine learning datasets](https://link.zhihu.com/?target=https%3A//www.datasetlist.com/)
* [https://archive.ics.uci.edu/ml/machine-learning-databases/](https://link.zhihu.com/?target=https%3A//archive.ics.uci.edu/ml/machine-learning-databases/)。
* 伯克利自动驾驶数据集：[http://bdd-data.berkeley.edu/](https://link.zhihu.com/?target=http%3A//bdd-data.berkeley.edu/)
* UCI 垃圾邮件数据集：[https://archive.ics.uci.edu/ml/datasets/Spambase](https://link.zhihu.com/?target=https%3A//archive.ics.uci.edu/ml/datasets/Spambase)
* 北京城市实验室地理数据：[https://www.beijingcitylab.com/data-released/data1-20/](https://link.zhihu.com/?target=https%3A//www.beijingcitylab.com/data-released/data1-20/)
* 体育大数据:[http://www.sportsdt.com/](https://link.zhihu.com/?target=http%3A//www.sportsdt.com/)