Translation Script

Exploratory data analysis and Data visualization

Translators:Ruoxi Li & Jiaxi Liu

Original Author:Chong-ho Yu, Ph.Ds.

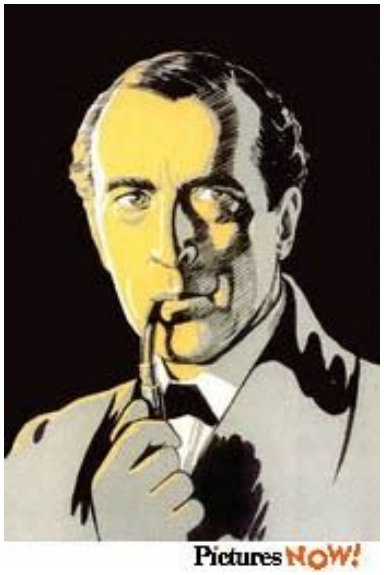
28th February , 2021

Contents

1.EDA:Detective Work

原文:

This is a brief introduction to exploratory data analysis (EDA) and data visualization. You will come across several unfamiliar terms and graphs, but you don't have to fully understand them at this moment. The purpose of this write-up is to let you be aware what tools are available and what can be done. The philosophy and specific techniques of EDA will be introduced in further readings.



When some people claim that their methodology is exploratory, actually what they mean is that they don't know what they are doing. Unfortunately, poor research is often implemented in the name of EDA. In data collection researchers flood the subjects with hundred pages of surveys since research questions are not clearly defined and variables are not identified. It is true that EDA does not require a pre-determined hypothesis to be tested, but it doesn't justify the absence of research questions and ill-defined variables or trying every test until obtaining a significant p value (p-hacking) (Jebb, Parrigon, & Woo, 2017).

EDA techniques are abundant and well-structured. Exploratory data analysis, as a supplement to confirmatory data analysis (CDA), was founded by John Tukey (1977, 1980). Tukey often related EDA to detective work. In EDA, the role of the researcher is to explore the data in as many ways as possible until a plausible "story" of the data emerges. A detective does not collect just any information. Instead he collects evidence and clues related to the central question of the case. So, from now on you can call me "Detective Yu."

1.EDA: 探索侦察型工作

译文:

这是对探索性数据分析（EDA）和数据可视化的简要介绍。您将遇到几个不熟悉的术语和图表，但是此时您不必完全理解它们。本文的目的是让您知道哪些工具可用以及可以做什么。 EDA的原理和特定技术将在进一步的阅读中介绍。

当有人声称他们的方法是探索性的时，实际上他们的意思是他们不知道自己在做什么。不幸的是，经常以EDA的名义进行拙劣的研究。在数据收集中，由于没有明确定义研究问题并且没有识别出变量，研究人员用数百页的调查资料充斥了调查对象。确实，EDA不需要预先确定的假设就可以进行测试，但是它并不能证明没有研究问题和定义不明确的变量，也无法证明每次尝试直到获得显着的p值（p-hacking）（Jebb），Parrigon和Woo，2017年）。

EDA这门技术是丰富且结构合理的。探索性数据分析是对确认数据分析（CDA）的补充，由约翰·图基（John Tukey）（1977，1980）建立。 Tukey通常将EDA与侦探工作联系起来。在EDA中，研究人员的作用是以尽可能多的方式探索数据，直到出现合理的数据“故事”为止。侦探不会收集任何信息。相反，他收集了与案件核心问题有关的证据和线索。因此，从现在开始，您可以称我为“侦探俞”。

2.Elements of EDA

原文:

Velleman and Hoaglin (1981) outlined four basic elements of exploratory data analysis as the following:

- Data visualization
- Residual analysis
- Data transformation or re-expression
- Resistance procedures

2.EDA的要素

译文:

Velleman和Hoaglin（1981）概述了探索性数据分析的四个基本要素，如下所示：

- 数据可视化
- 残差分析
- 数据转换或重新表达
- 稳定性分析过程

3.Data Visualization

原文:

The rationale for data visualization is: "A picture is worth a thousand words." It is easier to detect a data pattern from a picture than from a numeric output.

Generally speaking, there are six major categories of research goals. All of them can utilize graphing techniques for deepening our understanding of the data :

- Spotting outliers
- Discriminating clusters
- Checking distributional and other assumptions
- Examining relationships
- Comparing mean differences
- Observing a time-based process

The following are some examples. The interpretation of these graphs are very involved. Just get the idea of visualization and don't be absorbed into the detail.

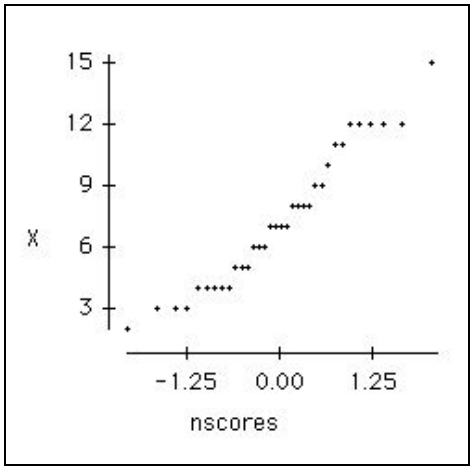
Spotting outliers

It is easy to spot univariate outliers in one-dimensional charts such as a histogram. In a multivariate case rotation plot is helpful. Please view this animated demo.

Discriminating clusters

By visualization we can cluster either variables or subjects. This example shows how brushing is used to cluster subjects in helping regression analysis. Please view this animated demo.

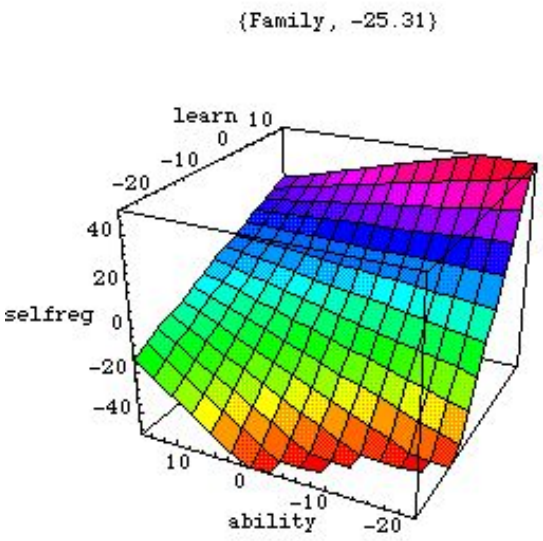
Checking distributional and other assumptions



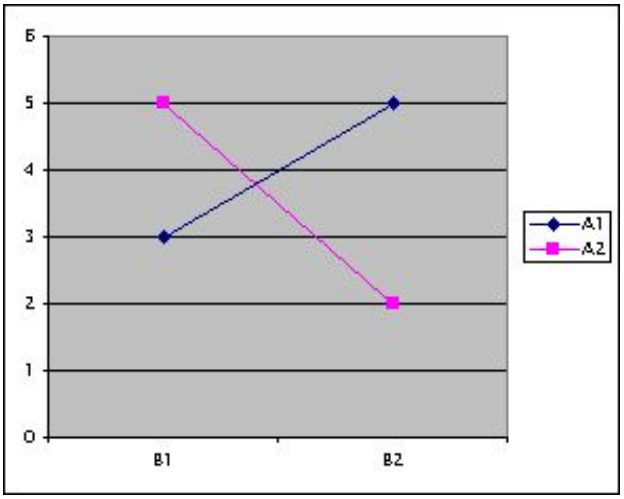
Data normality is required by many parametric tests. The researcher can use a simple histogram to examine the distribution. A more sophisticated way is to check the data with a **normality probability plot**. If the data are perfectly normal, the graph should show a diagonal straight line. The deviation from the straight line indicates the degree of non-normality.

Examining relationships

When interaction effects are present, regression lines are not consistent across all levels of other variables. The moving mesh surface depicts this change. If the animation annoys you, please press the stop button on your browser to freeze the animation. This type of data visualization can be performed in Mathematica and DataDesk.



Comparing group differences



Comparing mean differences is usually conduced using parametric tests such as a t-test or a F-test. Nonetheless, graphs can be used to supplement test statistics. A typical example is to use cell-mean plot to examine the main effects and the interaction effect. Examples of advanced graphs for comparing differences are diamond plot and leverage plot.

3. 数据可视化

译文:

数据可视化的基本原理是：“一张图片值一千字。”从图片检测数据模式比从数字输出检测数据模式容易。一般而言，研究目标有六大类。所有这些人都可以利用制图技术来加深我们对数据的理解：

- 发现异常值
- 区分集群
- 检查分配和其他假设
- 检查关系
- 比较均值差异
- 观察基于时间的过程

以下是一些示例。这些图的解释非常复杂。只是获得可视化的想法，不要过多的在意细节。

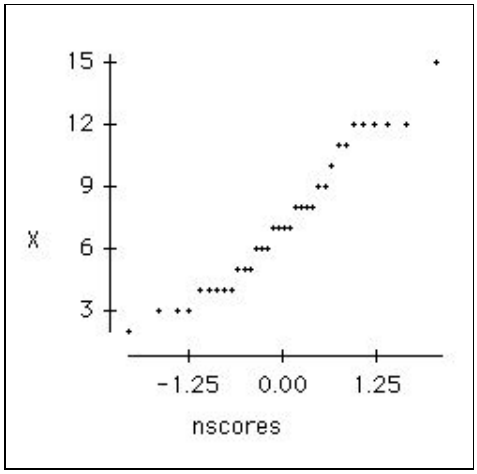
发现异常值

在直方图等一维图表中很容易发现单变量离群值。在多变量情况下，旋转图很有用。请观看此动画演示。

区分集群

通过可视化，我们可以将变量或主题聚类。此示例显示了如何使用笔刷对主题进行聚类以帮助进行回归分析。请观看此动画演示。

检查分布和其他假设

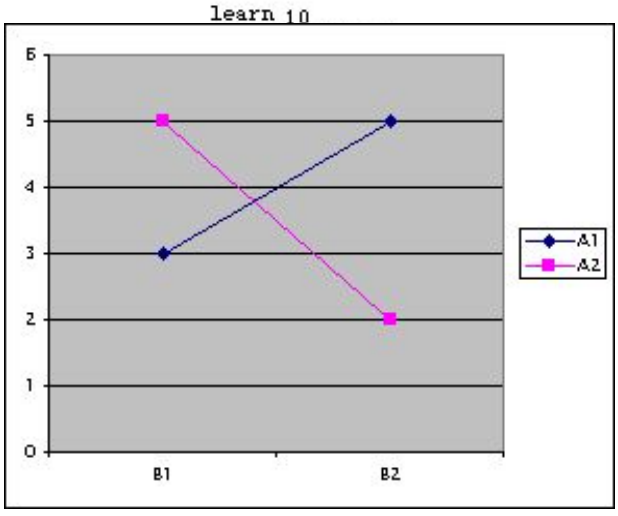


许多参数测试都需要数据规范性。研究人员可以使用简单的直方图来检查分布。一种更复杂的方法是使用正态概率图检查数据。如果数据完全正常，则图形应显示对角直线。与直线的偏离表示不正常的程度。

检查关系

当存在交互作用时，回归线在其他变量的所有级别上均不一致。移动的网格表面描述了这种变化。如果动画使您烦恼，请按浏览器上的“停止”按钮以冻结动画。这种类型的数据可视化可以在Mathematica和DataDesk中执行。

{Family, -25.31}



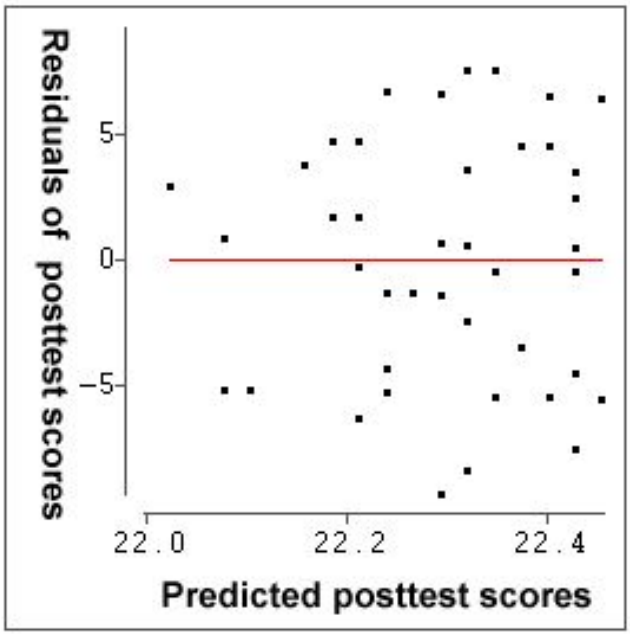
比较小组差异

通常使用参数检验（例如t检验或F检验）来比较均值差异。但是，可以使用图形来补充测试统计信息。一个典型的例子是使用单元均值图来检查主要作用和相互作用作用。用于比较差异的高级图形示例包括菱形图和杠杆图。

4.Residual Analysis

原文:

EDA follows the model that data = fit + residual or data = model + error. The fit or the model is the expected values of the data. The residual or the error is the values that deviate from that expected value. By examining the residuals, the researcher can assess the model adequacy. A simple example can be found in regression analysis. The scatterplot on the left shows the residuals in a regression model.



Today it is not difficult to see why we should examine residuals to check how well the data fits the model. Nonetheless, "residual" is a modern concept. A few centuries ago even very well-trained scientists had a weak sense of residual. * Unfortunately, at the present time this problem still exists among several researchers who tend to take modeling for granted and ignore residuals.

In the past this iterative process was performed manually by the analyst, such as the 2-way fit approach. Today machine learning algorithms automates this process. Boosting, also known as the boosted tree, is a good example of an automated iteration.

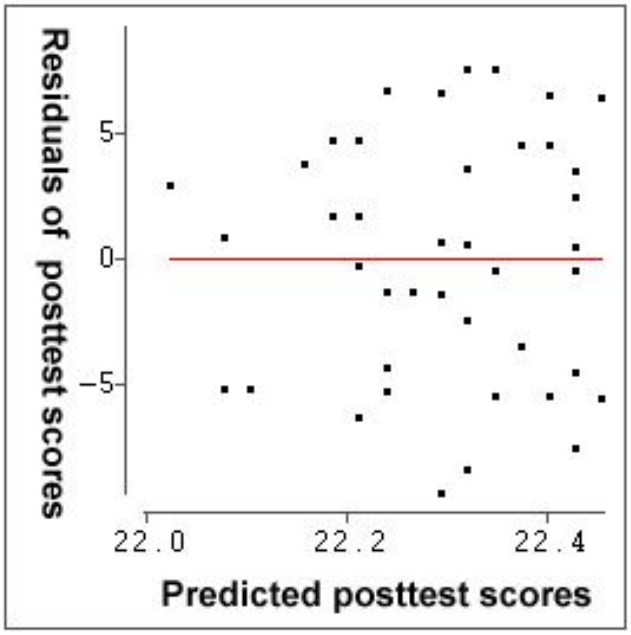
4. 残差分析

译文:

EDA遵循数据=拟合+残差 或者 数据=模型+误差的模型。拟合或模型是数据的期望值。残差或误差是偏离该预期值的值。通过检查残差，研究人员可以评估模型的适当性。一个简单的例子可以在回归分析中找到。左侧的散点图显示了回归模型中的残差。

今天，不难理解为什么我们应该检查残差以检查数据对模型的拟合程度。但是，“剩余”是一个现代概念。几个世纪前，即使是训练有素的科学家对”残差”这一概念的感受也很弱。*不幸的是，目前，这个问题仍然存在于一些倾向于将建模视为理所当然而忽略残差的研究人员中。

过去，这种迭代过程是由分析人员手动执行的，例如2-way fit方法。今天，机器学习算法可自动执行此过程。 Boosting，也称为boosted树，是自动迭代的一个很好的例子。

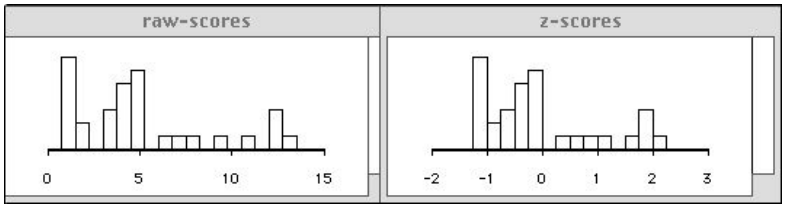


5.Data transformation or re-expression

原文: Data transformation happens in our everyday life: Converting US dollars into Canadian dollars, converting a GPA of 5-point scale to a GPA of 4-point-scale. However, these examples belong to the **linear transformation,** by which the distribution of the data are not affected. In EDA, usually the **non-linear transformation** is used and thereby it changes the data pattern. Data re-expression is exploratory in nature because prior to the transformation, the researcher never knows which re-expression approach can achieve desirable results.

There are four major objectives of transforming data:

- **Normalize the distribution:** Non-normal data violate the assumption of parametric test and thus a transformation is advisable. It is a common misconception that converting raw scores to z-scores yields a normal distribution. Actually, the raw-to-z-transformation is a linear transformation. The following figure shows that after a raw-to-z transformation, the distribution shape of the z scores is still resemble to that of raw scores. The appropriate procedure should be **natural log transformation** or **inverse probability transformation.**



- **Stabilize the variances:** Data with unequal variances are also detrimental to parametric tests. A typical example of variance stabilizing transformation is **square root transformation**: y* = sqrt(y).
- **Linearize the trend:**Regression analysis requires the assumption of linearity. When the data show a curvilinear relationship, the researcher can either apply non-linear regression analysis or straighten the data by linearizing transformation. A **logarithmic transformation** is a typical example of the latter.
- **Orthogonalize collinear variables:** In multiple regression lack of independence between predictors could make the model unstable. In terms of hyper-space, the vectors representing these variable are non-orthogonal. To rectify the situation the variables can be orthogonalized by **centering the scores**, using the **Gram–Schmidt process**, or other transformation techniques.

Nonetheless, every statistical procedure has limitations and should be used with caution. Data transformation is not an exception. Osborne (2002) advised that data transformation should be

used appropriately; many transformations reduce non-normality by changing the spacing between data points, but it raises issues in the interpretation of data. If transformations are done correctly, all data points should remain the same relative order as prior to transformation and this does not affect researchers to interpret the scores. But it might be problematic if the original variables were meant to be interpreted in a straight-forwarded fashion, such as annual income, and years of age). After the transformations, the new variables might become much more complex to interpret.

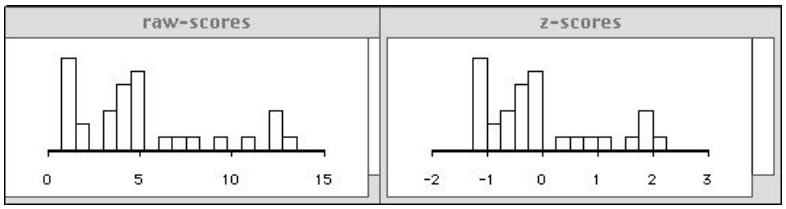
5.数据转换和重新表达

译文:

数据转换发生在我们的日常生活中：将美元转换为加元，将5分制的GPA转换为4分制的GPA。但是，这些示例属于线性变换，因此不影响数据的分布。在EDA中，通常使用非线性变换，从而改变数据模式。数据重新表达本质上是探索性的，因为在进行转换之前，研究人员永远不知道哪种重新表达方法可以达到理想的结果。

数据转换有四个主要目标：

- **标准化分布**：非常规数据违反了参数测试的假设，因此建议进行转换。常见的误解是将原始分数转换为z分数会产生正态分布。实际上，原始到z的转换是线性转换。下图显示了从原始到z的转换后，z得分的分布形状仍然类似于原始得分。适当的过程应该是自然对数变换或逆概率变换。



- **稳定差异**：方差不相等的数据也不利于参数测试。方差稳定化变换的一个典型示例是平方根变换：y * = sqrt（y）。
- **线性化趋势**：回归分析需要线性假设。当数据显示曲线关系时，研究人员可以应用非线性回归分析，也可以通过线性化变换对数据进行拉直。对数转换是后者的典型示例。
- **正交化共线变量**：在多元回归中，预测变量之间缺乏独立性会使模型不稳定。就超空间而言，代表这些变量的向量是非正交的。为了纠正这种情况，可以使用Gram–Schmidt过程或其他转换技术，通过将分数居中来使变量正交。

但是，每种统计程序都有其局限性，应谨慎使用。数据转换也不例外。 Osborne（2002）建议应适当使用数据转换。许多转换通过更改数据点之间的间距来减少非正态性，但是这在数据解释中

提出了问题。如果正确完成了转换，则所有数据点应保持与转换之前相同的相对顺序，并且这不会影响研究人员解释分数。但是，对于那些需要以直截了当的方式（例如年收入和年龄）来解释的原始变量，数据转换就可能会出现问题。因为转换后，新变量的解释可能会变得更加复杂。

6.Resistance procedures

原文: Parametric tests are based on the mean estimation, which is sensitive to outliers or skewed distributions. In EDA, robust estimators are usually used. For example:

- **Median**: The middle point of the data.
- **Trimean:** A measure of central tendency based on the arithmetic average of the values of the first quartile, the third quartile, and the median counted twice.
- **Winsorized mean:** A robust version of the mean in which extreme scores are pulled back to the majority of the data.
- **Trimmed mean:** A mean without outliers

In your first stat course you learned that the mode is more resistant against outliers than the median. You may ask why the median, instead of the mode, is used. Indeed, in most situations the median and the mode are equally robust against outliers. Please view this animated demo.

It is important to point out that there is a subtle difference between "resistance" and "robustness" though two terms are usually used interchangeably. EDA is more concerned with resistance while hypothesis testing pays more attention to robustness. Resistance is about being immune to outliers while robustness is about being immune to assumption violations. In the former, the goal is to obtain a data summary while in the latter the goal is to make a probabilistic inference.

6. 抵抗程序

译文:

参数测试基于均值估计，均值对异常值或偏斜分布敏感。在EDA中，通常使用鲁棒的估计器。例如：

- **中位数**：数据的中间点。
- **三均值**：根据第一四分位数，第三四分位数和中位数两次的算术平均值计算的集中趋势的度量。

- **Winsorized均值**：均值的可靠版本，其中极端得分被拉回到大部分数据中。
- **切尾均值**：没有异常值的均值

在您的第一节统计课程中，您就能了解到众数对异常值的抵抗力比中值大。您可能会问为什么使用中位数而不是众数。确实，在大多数情况下，中位数和众数对异常值具有同样的鲁棒性。请观看此动画演示。

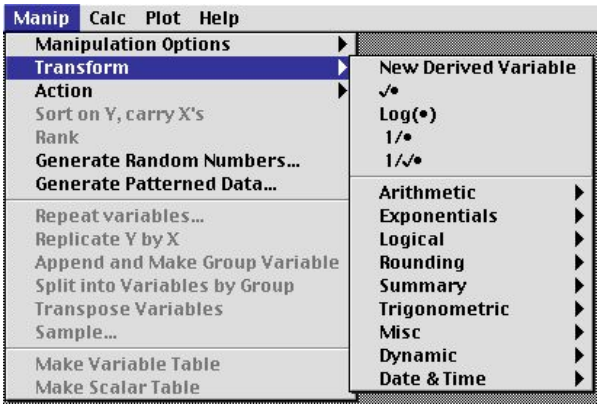
重要的是要指出，尽管“抵抗性”和“鲁棒性”通常可以互换使用，但它们之间存在细微的差别。EDA更加关注抵抗力，而假设检验则更加关注鲁棒性。抵抗力是关于不受异常值影响的，而鲁棒性是关于不受违反原假设带来的影响。在前者中，目标是获得数据摘要，而在后者中，目标是进行概率推断。

7.Recommended Software for EDA

原文:

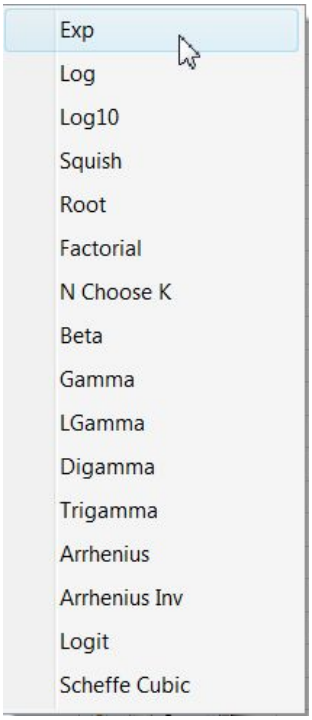
- **DataDesk**

DataDesk (Data Description, Inc., 2008) is developed by Paul Velleman, a student of John Tukey. DataDesk is the ideal tool for beginners in exploratory data analysis. It is feature-rich and flexible enough for manipulation, but yet requires little prior knowledge of computer operation. For instance, data re-expression described above can be performed using a wide variety of transformation functions in DataDesk. DataDesk has a richer version called Data Desk Plus, which incorporates a multimedia-based statistics tutorial entitled ActivStat.

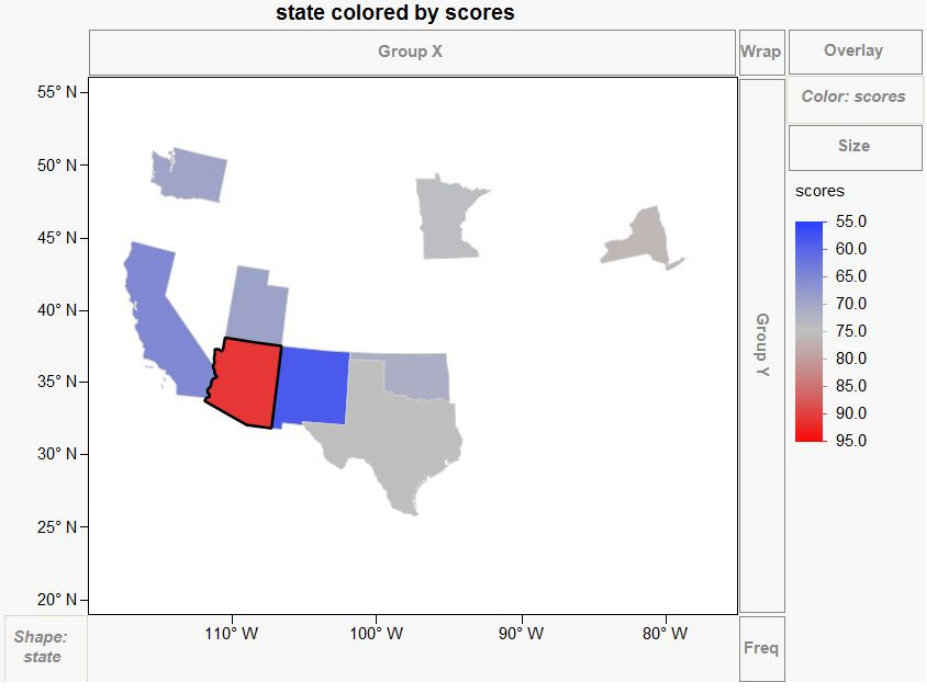


- **JMP**

JMP (SAS Institute, 2016) is a very versatile statistical program. There are two variants of JMP, namely, JMP and JMP Pro. As the name implies, JMP Pro is a professional version that includes many powerful procedures. But for most users JMP is sufficient for EDA. Like DataDesk, JMP has built-in data transformation options as shown below.



The design philosophy of JMP is similar to that of Apple's iPod. Upon installation you can start exploring your data without reading the manual. In addition to common graphing features, such as histogram and boxplot, Graph builder in JMP also provide the users with Geographical Information System (GIS).



- **XLISP-STAT**

If you like to gain a complete control by programming, XLISP-STAT should be considered. For example, in data visualization it involves data smoothing. Through programming you can view the data in different levels of detail.

LISP stands for List Processing. Someone calls it "Lots of idiotic and silly parenthesis." LISP was created during 1956-62 by John McCarthy in MIT for non-numerical computation. Later it is used specifically for the development of artificial intelligence. There are many different versions of LISP e.g. Common Lisp, Franz LISP...etc. XLISP is one of many dialects, which was developed by David Betz. Later Luke Tierney (1990) developed XLISP-STAT for statistical visualization. This package has many built-in statistical graphing functions. Based on XLISP-STAT, Cook and Weisberg (1994) developed a set of regression graphing tools called R-code. Another comprehensive EDA package named ViSta (Young, 1999) is also written in XLISP-STAT.

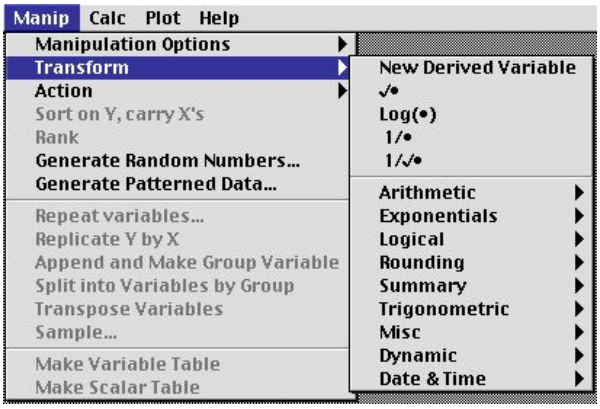
XLISP-STAT is cross-platform. However, it is an interpreted rather than a complied language, and therefore, you must load the written program into XLISP-STAT to run it.

7.推荐用于EDA的软件

译文:

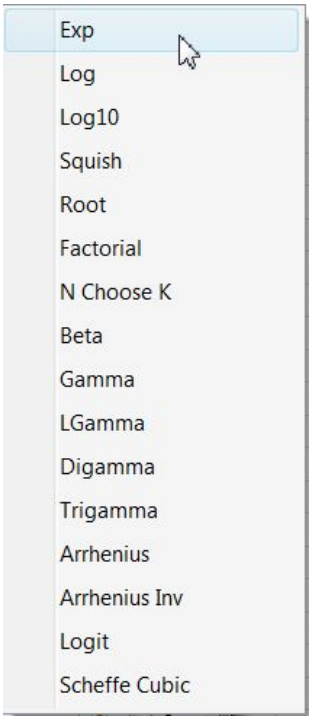
- **DataDesk**

DataDesk（数据描述公司，2008年）由John Tukey的学生Paul Velleman开发。 DataDesk是探索性数据分析初学者的理想工具。 它具有丰富的功能和足够的灵活性，可以进行操作，但是对计算机操作的了解很少。 例如，可以使用DataDesk中的各种转换函数来执行上述数据重新表达。 DataDesk具有一个称为Data Desk Plus的更丰富的版本，其中包含名为ActivStat的基于多媒体的统计教程。

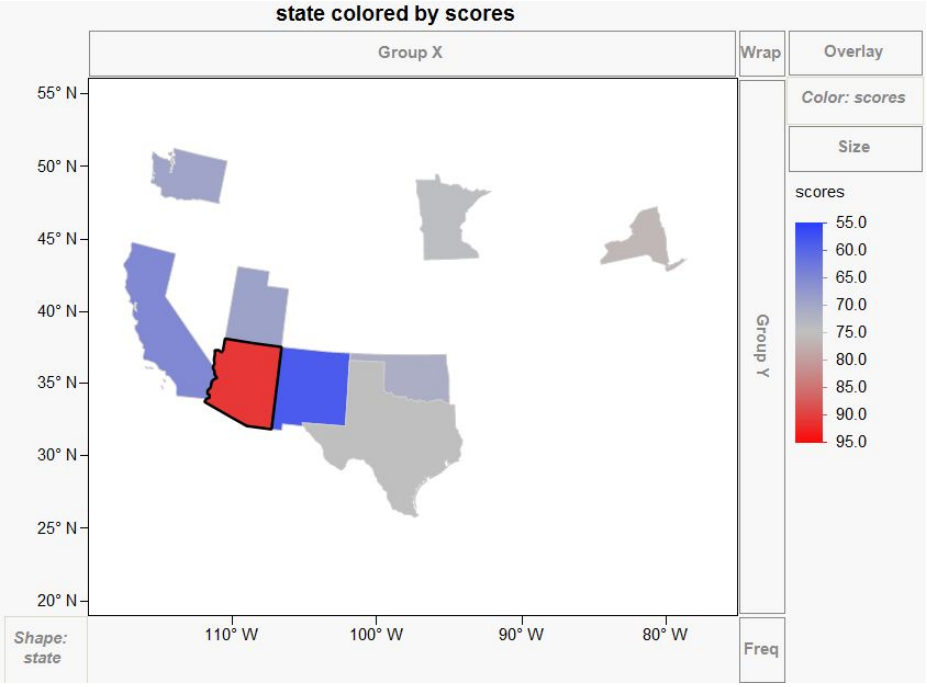


- **JMP**

JMP（SAS研究所，2016）是一个非常通用的统计程序。 JMP有两种变体，即JMP和JMP Pro。顾名思义，JMP Pro是一个专业版本，其中包含许多强大的过程。 但是对于大多数用户而言，JMP足以满足EDA的要求。 像DataDesk一样，JMP具有内置的数据转换选项，如下所示。



JMP的设计原理类似于苹果iPod的设计原理。 安装后，您无需阅读手册即可开始浏览数据。 除了直方图和箱线图之类的常用图形功能外，JMP中的“图形生成器”还为用户提供了地理信息系统（GIS）。



- **XLISP-STAT**

如果您希望通过编程获得完全的控制权，则应考虑XLISP-STAT。例如，在数据可视化中，它涉及数据平滑。通过编程，您可以查看不同详细程度的数据。

LISP代表列表处理。有人称其为“很多白痴和愚蠢的括号”。 LISP由John McCarthy在MIT于1956-62年创建，用于非数值计算。后来，它专门用于人工智能的开发。 LISP有许多不同的版本，例如普通Lisp，Franz LISP等XLISP是David Betz开发的许多方言之一。后来，Luke Tierney（1990）开发了用于统计可视化的XLISP-STAT。该软件包具有许多内置的统计图形功能。 Cook and Weisberg（1994）基于XLISP-STAT，开发了一套称为R-code的回归绘图工具。另一个名为ViSta的综合EDA软件包（Young，1999）也用XLISP-STAT编写。

XLISP-STAT是跨平台的。但是，它是一种解释性语言，而不是一种编译语言，因此，必须将编写的程序加载到XLISP-STAT中才能运行它。

8.Further readings

原文:

Tukey (1977)'s book is considered a classic in EDA. In his time computer resources were not easily accessible, but today most of his suggested graphing techniques are available in many software packages.

Behrens (1997) and Behrens & Yu (2003) are essential for both beginners and intermediate learners. Both chapters cover the detail of visualization, data transformation, residual analysis, and resistance procedures, which are briefly mentioned in this lesson.

For a quick overview of EDA, visit NIST Engineering Statistics Handbook. Although this site gives many examples of graphing techniques, it does not tell you what specific software packages can generate those graphs.

For the philosophical foundation of EDA, please consult Yu (1994 April, 2006). EDA is a philosophy/attitude rather than a collection of techniques.

To acquire a deeper understanding of data visualization, please read Yu and Behrens (1995) and Yu (2010, 2014).

8. 更多参考

译文:

Tukey（1977）的书被认为是EDA中的经典著作。 在他的时代，计算机资源不容易获得，但是如今，他建议的大多数图形技术可在许多软件包中使用。

Behrens（1997）和Behrens＆Yu（2003）对于初学者和中级学习者都是必不可少的。 这两章都涵盖了可视化，数据转换，残差分析和抵抗过程的详细信息，本课对此进行了简要介绍。

有关EDA的快速概述，请访问NIST工程统计手册。 尽管此站点提供了许多图形技术示例，但并未告诉您哪些特定的软件包可以生成这些图形。

有关EDA的哲学基础，请咨询Yu（1994年4月，2006年）。 EDA是一种哲学/态度，而不是技术的集合。

要更深入地了解数据可视化，请阅读Yu和Behrens（1995）和Yu（2010，2014）。

9.Note

原文:

* For example, Gregor Mendel (1824-1884), who is considered the founder of modern genetics, established through his scientific findings, the notion that physical properties of species are subject to heredity. Mendel conducted a fertilization experiment to confirm his belief. In his experiment, he followed up several generations of plants to observe how specific genes carried from one generation to another. While the reported data largely conform to the inheritance hypothesis, R. A. Fisher (1936) questioned the validity of Mendel's study. Fisher pointed out that Mendel's data seemed "too good to be true." Using Chi-square tests, Fisher found that Mendel's results were so close to what would be expected that such agreement could happen by chance less than once in 10,000 times.

Another example can be found in the story of Johnannes Kepler (1571-1630), the first astronomer who proposed that the earth and other planets orbit around the sun in an elliptical fashion, rather than in circle as Galileo believed. Kepler worked under another well-known astronomer, Brahe, who collected a huge database of planetary orbits. Using Brahe's data, Kepler found data to fit into the elliptical hypothesis, rather than the circular hypothesis. However, almost 400 years later when William Donahue redid Kepler's calculation, he found that the orbit data and the elliptical model do not fit each other as claimed.

Further, there is a widespread urban legend that British physicist Arthur Eddington substantiated Einstein’s theory of general relativity by observing the positions of stars during the 1919 solar eclipse. However, in the 1980s scholars found that Eddington did collect sufficient data to reach a conclusion. Rather, he distorted the result to make it fit the theory (Swayer, 2012).

Kepler, Mendel, and Eddington are not the only three scientists who failed to accept the residuals between the data and the model. William Harvey, Isaac Newton, and Charles Darwin also had the same problem; the list goes on and on. While reviewing this phenomenon in the history of science, some scholars denounced those scientists as committing fraud. In a milder tone, Press and Tanur (2001) said that the problem was caused by "the subjectivity of scientists."

My view is that those scientists had a weak sense of residuals. They conducted science in a confirmatory mode, in which only a dichotomous answer could result. Even if residuals existed, they tended to embrace the model because by admitting any inconsistency, the entire model would be rejected. In other words, they accepted the notion that DATA = MODEL.

Last revision: May, 2017

9.注解

译文:

*例如，被认为是现代遗传学创始人的格雷戈尔·孟德尔（Gregor Mendel，1824-1884年），通过他的科学发现确立了物种的物理特性受遗传影响的观念。孟德尔进行了一次受精实验，以证实他的信念。在他的实验中，他跟踪了几代植物，观察特定基因如何从一代传到另一一代。尽管报告的数据在很大程度上符合遗传假说，但R. A. Fisher（1936）质疑孟德尔研究的有效性。费舍尔指出，孟德尔的数据“实在太好了”。通过卡方检验，费舍尔发现孟德尔的结果是如此接近预期，以至于这种协议偶然发生的可能性少于万分之一。

另一个例子可以在约翰内斯·开普勒（Johnannes Kepler，1571-1630）的故事中找到，他是第一位提出将地球和其他行星以椭圆形绕太阳公转的提议的天文学家，而不是像伽利略所相信的那样绕太阳公转。开普勒在另一个著名的天文学家布拉赫（Brahe）的指导下工作，布拉赫收集了庞大的行星轨道数据库。利用Brahe的数据，开普勒发现数据适合椭圆假设，而不是圆形假设。然而，将近400年后，当威廉·多纳休（William Donahue）重新提出开普勒的计算结果时，他发现轨道数据和椭圆模型并不完全吻合。

此外，有一个广泛的城市传说，英国物理学家亚瑟·爱丁顿（Arthur Eddington）通过观察1919年日食期间恒星的位置，证实了爱因斯坦的广义相对论。但是，在1980年代，学者发现爱丁顿确实收集了足够的数据来得出结论。相反，他扭曲了结果以使其符合理论（Swayer，2012）。

开普勒，孟德尔和爱丁顿不是仅有的三位未能接受数据和模型之间残差的科学家。威廉·哈维（William Harvey），艾萨克·牛顿（Isaac Newton）和查尔斯·达尔文（Charles Darwin）也有同样的问题。这个清单不胜枚举。在回顾科学史上的这一现象时，一些学者谴责那些科学家犯有欺诈罪。 Press and Tanur（2001）用温和的语气说，问题是由“科学家的主观性”引起的。

我的观点是，这些科学家对残留物的意识较弱。他们以确认性方式进行科学，其中只能得出二分法的答案。即使存在残差，它们也倾向于包含模型，因为通过承认任何不一致之处，整个模型都会被拒绝。换句话说，他们接受了DATA = MODEL的概念。

修订日期：2017年5月

10.References

原文:

- Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. Psychological Methods, 2, 131-160.

- Behrens, J. T., & Yu, C. H. (2003). Exploratory data analysis. In J. A. Schinka & W. F. Velicer, (Eds.). Handbook of psychology Volume 2: Research methods in Psychology (pp. 33-64). New Jersey: John Wiley & Sons, Inc.
- Cook, D. R. & Weisberg, S. (1994). An introduction to regression graphics. New York : Wiley.
- Data Description, Inc. (2008). DataDesk. [On-line] Available: http://www.datadesk.com
- Fisher, R. A. (1936). Has Mendel's work been rediscovered? Annals of Science, 1, 115-137.
- Jebb, A., Parrigon, S. & Woo, S. E. (2017). Exploratory data analysis as a foundation of inductive research. Human Resource Management Review, 27, 265–276.
- Osborne, J. W. (2002). Notes on the use of data transformations. Practical Assessment, Research, and Evaluation, 8(6) Retrieved from: http://pareonline.net/getvn.asp?v=8&n=6
- Press, S. J., & Tanur, J. M. (2001). The subjectivity of scientists and the Bayesian approach. New York: John Wiley & Sons.
- SAS Institute. (2016). JMP [Computer Software]. Cary, NC: Author.
- Swayer, R. K. (2012). Explaining creativity: The science of human innovation (2nd ed.). New York, NY: Oxford University Press.
- Tierney, L. (1990). Lisp-Stat : an object-oriented environment for statistical computing and dynamic graphics. New York : Wiley.
- Tukey, J. W. (1977). Exploratory data analysis. Reading, MA: Addison-Wesley Publishing Company.
- Tukey, J. W. (1980). We need both exploratory and confirmatory. The American Statistician, 34, 23-25.
- Velleman, P. F. & Hoaglin, D. C. (1981). Applications, basics, and computing of exploratory data analysis. Boston, MA : Duxbury Press.
- Young, F. (1999). ViSta [Computer Software]. Retrieved from http://www.uv.es/prodat/ViSta/
- Yu, C. H. (1994, April). Induction? Deduction? Abduction? Is there a logic of EDA? Paper presented at the Annual Meeting of American Educational Researcher Association, New Orleans, Louisiana. (ERIC Document Reproduction Service No. ED 376 173)
- Yu, C. H., & Behrens, J. T. (1995). Applications of scientific multivariate visualization to behavioral sciences. Behavior Research Methods, Instruments, and Computers, 2, 264-271.
- Yu, C. H. (2010). Exploratory data analysis in the context of data mining and resampling. International Journal of Psychological Research, 3(1), 9-22. Retrieved from http://mvint.usbmed.edu.co:8002/ojs/index.php/web/article/download/455/460 [mirror].
- Yu, C. H. (2006). Philosophical foundations of quantitative research methodology. Lanham, MD: University Press of America.

- Yu, C. H. (2014). Dancing with the data: The art and science of data visualization. Saarbrucken, Germany: LAP.

10. 参考文献

译文:

- Behrens，J.T。（1997）。探索性数据分析的原则和程序。心理方法，第2卷，第131-160页。
- Behrens，J.T.，&Yu，C.H.（2003）。探索性数据分析。在J. A. Schinka和W. F. Velicer（编辑）中。心理学手册第2卷：心理学研究方法（第33-64页）。新泽西州：John Wiley &Sons，Inc.
- Cook，D. R.和Weisberg，S.（1994）。回归图形简介。纽约：威利。
- 数据描述公司（2008）。 DataDesk。 [在线]可用：http：//www.datadesk.com
- Fisher，R.A.（1936年）。孟德尔的作品被重新发现了吗？科学年鉴，第1卷，第115-137页。
- Jebb，A.，Parrigon，S.&Woo，S.E.（2017年）。探索性数据分析是归纳研究的基础。人力资源管理评论，第27期，第265-276页。
- Osborne，J。W.（2002）。有关使用数据转换的注意事项。实用评估，研究与评估，8（6）取自：http://pareonline.net/getvn.asp?v=8&n=6
- Press，S.J。，和Tanur，J.M。（2001）。科学家的主观性和贝叶斯方法。纽约：约翰·威利父子（John Wiley&Sons）。
- SAS研究所。（2016）。 JMP [计算机软件]。卡里，北卡罗来纳州：作者。
- Swayer，R.K.（2012年）。解释创造力：人类创新的科学（第二版）。纽约，纽约：牛津大学出版社。
- Tierney，L.（1990）。 Lisp-Stat：用于统计计算和动态图形的面向对象的环境。纽约：威利。
- Tukey，J.W。（1977）。探索性数据分析。马萨诸塞州雷丁：Addison-Wesley出版公司。
- Tukey，J.W。（1980）。我们需要探索性和确认性。美国统计学家，第34页，第23-25页。

- Velleman，P.F.和Hoaglin，D.C.（1981）。探索性数据分析的应用程序，基础知识和计算。马萨诸塞州波士顿：达克斯伯里出版社。
- Young，F.（1999）。 ViSta [计算机软件]。取自http://www.uv.es/prodat/ViSta/
- Yu，C.H.（1994年4月）。就职？扣除？绑架？ EDA有逻辑吗？该论文在路易斯安那州新奥尔良举行的美国教育研究者协会年会上发表。 （ERIC文件复制服务，编号ED 376 173）
- Yu，C.H。和Behrens，J.T。（1995）。科学多元可视化在行为科学中的应用。行为研究方法，仪器和计算机，第2卷，第264-271页。
- Yu，C.H.（2010年）。在数据挖掘和重新采样的背景下进行探索性数据分析。国际心理研究杂志，3（1），9-22。取自 http://mvint.usbmed.edu.co:8002/ojs/index.php/web/article/download/455/460 [镜像]。
- Yu，C.H.（2006年）。定量研究方法论的哲学基础。医学博士兰纳姆：美国大学出版社。
- Yu，C.H.（2014年）。与数据共舞：数据可视化的艺术和科学。德国萨尔布吕肯（Saarbrucken）：LAP。

补充：关于数据可视化软件的更多介绍

Tableau:

Tableau的真正目的是一个简单的数据可视化工具。开发该工具旨在洞察无法通过盯着电子表格来快速回答的问题。自成立以来，它已发展成为地球上最流行的数据可视化和报告工具之一。使用Tableau，用户可以以惊人的速度开发交互式仪表板和可视化文件。相较于其他可视化软件，Tableau有以下几点突出优势：

(1) 快速创建交互式可视化：使用Tableau的拖放功能，用户可以在数分钟内创建非常互动的视觉效果。该界面可以处理无尽的变化，同时还限制您创建违反数据可视化最佳做法的图表。您可以查看Tableau Gallery上创建的一些惊人的视觉效果。

(2) 易于实施：Tableau中提供了许多不同类型的可视化选项，可以增强用户体验。而且，与Python，Business Objects和Domo相比，Tableau非常易于学习，任何不具备编码知识的人都可以轻松学习Tableau

(3) 处理大量数据：Tableau可以轻松处理数百万行数据。可以使用大量数据创建不同类型的可视化文件，而不会影响仪表板的性能。另外，Tableau中有一个选项，用户可以使它“实时”建立到不同数据源（如SQL等）的连接。

(4) 使用其他脚本语言：为了避免性能问题并在Tableau中进行复杂的表计算，用户可以合并使用Python或R。使用Python脚本可以通过对数据包执行数据清除任务来减轻软件的负担。但是，Python不是Tableau接受的本机脚本语言。因此，您可以导入一些视觉效果或包装。但是，您可以看到使用Power BI的Python如何解决此问题。

R:

使用R提供的各种数据包，仅需几行代码就可以创建具有视觉吸引力的数据可视化。常用的10大数据可视化数据包有plotly, ggplot2, tidyquant, taucharts，ggiraph, geofacets, googleVis, RColorBrewer, dygraphs, shiny。

(1) Plotly软件包提供了在线互动图和质量图。该软件包扩展了JavaScript库

(2) ggplot2以其优雅和高质量的图形而闻名，这使其与其他可视化程序包区分开来。

(3) Tidyquant是用于执行定量财务分析的财务软件包。该软件包在tidyverse Universe下添加为财务软件包，用于导入，分析和可视化数据。

(4) Taucharts提供了一个声明性接口，用于将数据字段快速映射到视觉属性。

(5) Ggiraph是允许我们创建动态ggplot图的工具。该软件包使我们可以在图形中添加工具提示，JavaScript操作和动画。

(6) Geofacets软件包为“ ggplot2”提供了地理标注功能。 Geofaceting将针对不同地理实体的一系列绘图安排到保留某些地理方位的网格中。

(7) GoogleVis在R和Google的图表工具之间提供了一个界面。借助此软件包，我们可以基于R数据框创建具有交互式图表的网页。

(8) RColorBrewer提供了由Cynthia Brewer设计的地图和其他图形的配色方案

(9) Dygraphs包是dygraphs JavaScript图表库的R接口。它提供了丰富的功能来绘制R中的时间序列数据。

(10) Shiny使我们能够通过提供闪亮的程序包来开发交互式且美观的Web应用程序。该软件包提供了HTML窗口小部件，CSS和JavaScript的各种扩展

Python:

Python具有一些最具交互性的数据可视化工具。最基本的绘图类型在多个库之间共享，但是其他类型仅在某些库中可用。数据科学家经常使用的数据可视化库是Matplotlib, Seaborn 和 Plotly。

(1) Matplotlib是最受欢迎的Python数据可视化库。它用于生成简单而强大的可视化。从初学者到经验丰富的数据科学专业人士，Matplotlib是最广泛使用的绘图库。

(2) Seaborn提供了多种可视化模式。与matplotlib相比，它与Pandas数据框的集成度更高。Seaborn被广泛用于统计可视化，因为它具有一些内置的最佳统计任务。

(3) Plotly主要用于处理地理，科学，统计和财务数据。