

使用Jupyter Notebook 撰写数据分析报告

SmartyDoc项目组

2020年8月

目录

1. 前言	
2. 规划报告的逻辑结构	
2.1 显示Python程序中输出的文本	4
2.2 使用toc2插件管理文档目录	4
3. 在报告中插入图表	
3.1 使用的工具包	6
3.2 自动添加图表编号	6
3.3 图表示例 – 在报告中插入图片	7
3.4 图表示例 – 表格	7
3.5 图表示例 – 数据图	12
3.6 图表示例 – 较复杂的复合图	17
4. 生成PDF格式的报告文档	
4.1 PDF文档转换流程	20
4.2 需要准备哪些文件	20
4.3 使用printview2插件实现PDF文档生成	20
4.4 使用命令行操作实现PDF文档生成	21

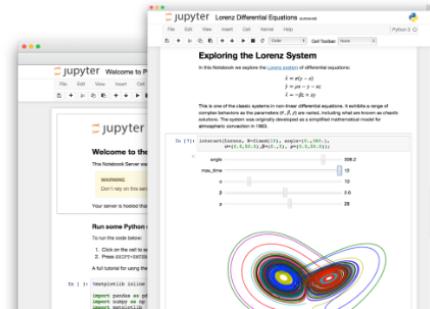
1. 前言

Jupyter Notebook是基于网页的用于交互计算的应用程序，其可被应用于全过程计算：开发、文档编写、运行代码和展示结果。

— Jupyter Notebook 官方介绍



The Jupyter Notebook is a web-based interactive computing platform that allows users to author data- and code-driven narratives that combine live code, equations, narrative text, visualizations, interactive dashboards and other media.



简而言之，Jupyter Notebook是以网页的形式存储，可以在网页页面中直接编写和运行代码，代码的运行结果也会直接在代码块下显示。如在编程过程中需要编写说明文档，可在同一个页面中直接编写，便于作及时的说明和解释。

基于Jupyter Notebook具备富文档和标准化的特性，我们也希望可以利用它生成美观的数据分析报告。**SmartyDoc** 就是为此而生。利用 **SmartyDoc** 提供的文本标准化流程，我们可以将包含图文的 Jupyter Notebook 文件转存为具备特定层级结构的html文件，配合适当的css文件以及 **weasyprint** 工具，即可产生PDF格式的报告文档。

要顺利使用 **SmartyDoc** 完成报告，要求使用者掌握Python和MarkDown两种语言。

下面将具体介绍如何利用这套工具撰写报告文档。

2. 规划报告的逻辑结构

在撰写数据分析报告时，把报告的逻辑结构规划好，将报告内容划分为几个相对独立、且逻辑连贯的章节，会让后续工作更加顺利。如整篇报告可以包含几大 章，每 章 可以包含多个 小节，每个 小节 可以进一步划分为多个 子节。

在 **SmartyDoc** 的框架内，文档的层级结构基于MarkDown语言中的 *Heading* 实现，即报告的题目为一级标题，用 `# 报告题目` 这种形式实现，各个章节的标题为二级标题，用 `## 章节标题` 的形式实现，以下可以继续出现三级、四级... 等不同等级的章节。

在撰写报告时，报告题目的样式以一段程序自动生成，用户只需提供报告封面的内容即可，具体可以参考此ipynb文件的 *Cell 4* 和 *Cell 5* 中的内容。需要用户自己写的报告章节结构，以二级标题为最高层级。

在每个章节下，可以出现文字、表格和图片等信息。为了保证用MarkDown语言撰写的报告内容可以正常显示，需要将对应 *Cell* 的语言设置为 **MarkDown**。

2.1 显示Python程序中输出的文本

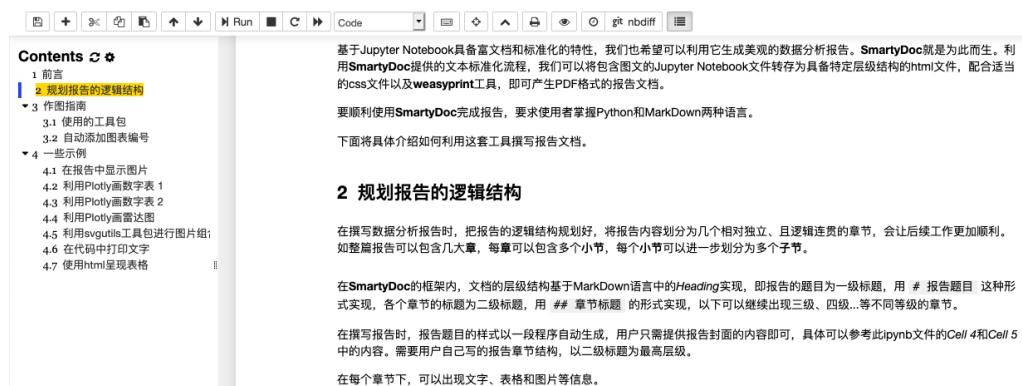
在进行数据分析时，我们经常需要根据分析结果打印对应的数据和结果描述，因此需要保证Python程序中打印的文字也能够正常显示在报告正文中。在 **SmartyDoc** 的框架下，我们可以使用如下方式显示Python程序中输出的文本。

打印Python程序中的输出文本 ... 变量的数值是 78。

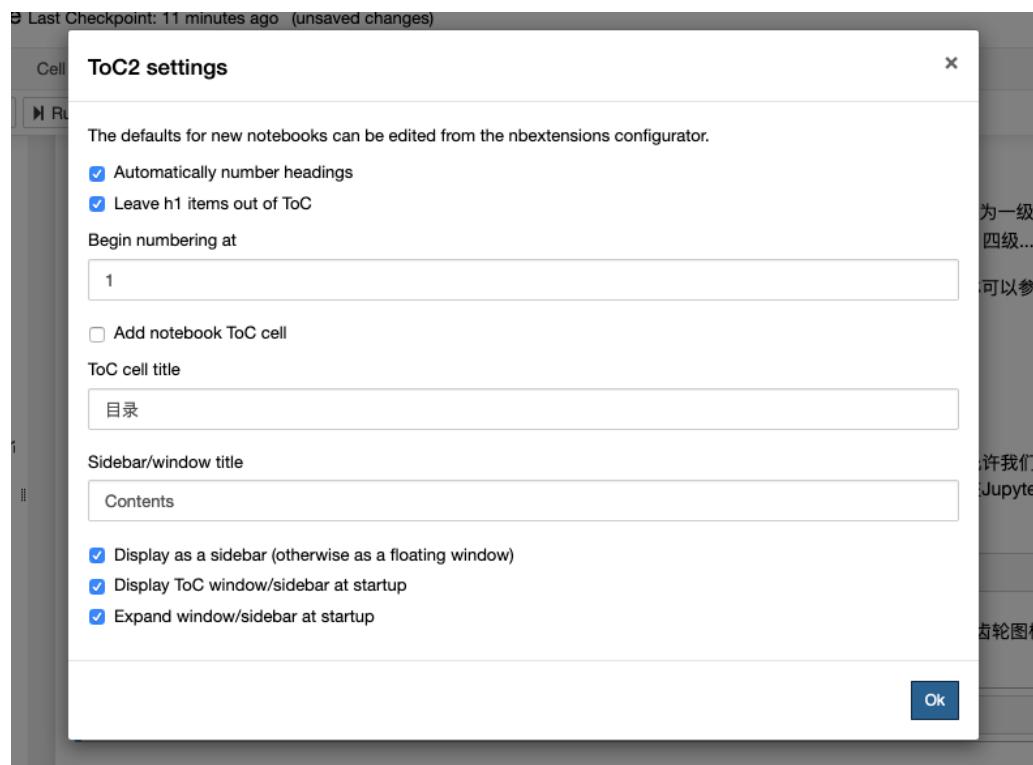
2.2 使用toc2插件管理文档目录

Jupyter Notebook是一个很棒的教学、探索和编程环境，但其功能仍存在很多不足。幸好，它允许我们使用一些插件来扩展它的功能。其中有一个插件叫Table of Contents (2) (toc2)，可以为Jupyter Notebook提供目录。可以在Jupyter的Nbextensions扩展管理页内，通过勾选框启用扩展。

点击按钮栏最后部的图标，目录列表会以左侧边栏的形式显示，toc2的使用效果见下图。



为了与 `SmartyDoc` 的框架兼容，需要对 `toc2` 扩展进行一些设置，可以通过点击目录列表上部的 齿轮图标 进行设置(如下图)，并将 `h1` 级标题排除出目录范畴（在 `SmartyDoc` 的框架下，章节结构以二级标题为最高层级）。



3. 在报告中插入图表

图表可以让报告内容清晰明了。

在Jupyter Notebook中，用户可以直接将图片或表格嵌入到文本中。但为了便于将报告文本中所包含的文字、图片和表格都顺利地转换为HTML或PDF等格式，在 `SmartyDoc` 框架下，要求用户将图片和表格保存成图片文件，并在ipynb文本中进行引用，或直接在ipynb文件中以html格式保存。

为了方便报告内容的撰写，可以使用变量 `PRODUCTION_PHASE` 来控制图片在ipynb文件中保存的形式。具体可以参考 `Cell 2` 中的变量定义方式和下面的例子。

3.1 使用的工具包

为了得到美观且样式丰富的数据图，这里建议主要使用 `plotly` 工具包作图，并将图片保存为SVG格式，以保证在不同设备和缩放尺度下的图片质量。

对于一些非常具有设计感但难以通过 `plotly` 简单实现的图片，可以使用 `svgutils` 工具包中的图片组合功能，对图片进行重组和编辑。

具体使用样例请参考下面的示例。

3.2 自动添加图表编号

在撰写报告时会产生大量的图片，要准确标注每张图的编号将会消耗很大的人力，因此我们将这项工作交给程序自动完成。

在 `SmartyDoc` 中的 `decorator` 工具包中，提供了 `FigCounter` 类用来对图表编号进行自动计数。这个类的初始化方法请见 `Cell 2`，在初始化时需要提供图表名称的起始词，如图1，图2，... 中的“图”字，以及设置文字的字号和字体等。初始化后，具体的使用方法请参考下文的示例。

3.3 图表示例 – 在报告中插入图片



图1 插入的png图片

3.4 图表示例 – 表格

3.4.1 表格1

学校名称	总人数	男生	女生	班级数
A学校	100人	60人	40人	10
B学校	90人	30人	60人	15
C学校	80人	70人	10人	6

表1 人数统计

3.4.2 表格2

平均等级	文理倾向	学科发展均衡程度	A等级学生人数	B等级学生人数	全市排名	全省排名
B	综合	较为均衡	54人	0人	9/61	22/287
A	文	均衡	100人	10人	12/61	40/287

表2 结果统计

3.4.3 可跨页的长表格

在报告中，如果需要呈现比较长的表格，如会跨越多页的表格，可以使用html形式来实现长表格，如下例。

学校名称	总人数	男生	女生	班级数
A学校	100人	60人	40人	10
B学校	90人	30人	60人	15
C学校	80人	70人	10人	6
学校1	38人	16人	127人	19
学校2	19人	31人	128人	8
学校3	171人	140人	44人	3
学校4	121人	42人	120人	9
学校5	187人	87人	151人	14
学校6	172人	136人	153人	7
学校7	56人	48人	168人	13
学校8	111人	81人	143人	10
学校9	13人	31人	200人	12
学校10	80人	109人	119人	9
学校11	110人	84人	85人	8
学校12	140人	141人	84人	11
合计	270人	160人	110人	21

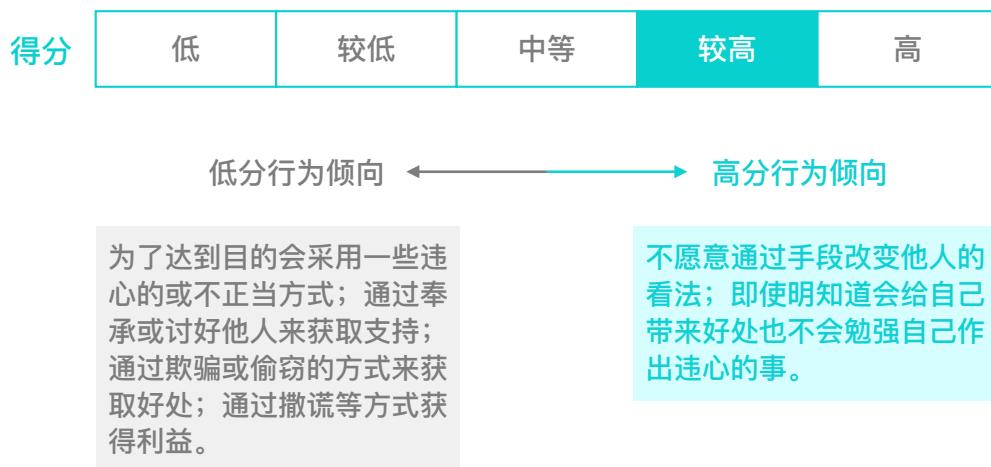
学校名称	总人数	男生	女生	班级数
学校13	144人	34人	194人	16
学校14	149人	60人	18人	4
学校15	196人	42人	42人	5
学校16	61人	93人	35人	19
学校17	86人	169人	168人	14
学校18	71人	162人	17人	12
学校19	137人	173人	75人	13
学校20	27人	141人	118人	14
学校21	197人	175人	112人	4
学校22	75人	108人	70人	16
学校23	14人	186人	165人	19
学校24	83人	115人	143人	18
学校25	177人	45人	127人	3
学校26	83人	145人	57人	15
学校27	181人	186人	176人	16
学校28	166人	33人	138人	3
学校29	15人	35人	150人	19
学校30	169人	157人	184人	9
学校31	94人	55人	120人	16
学校32	25人	199人	140人	5
学校33	32人	37人	178人	7
学校34	19人	123人	27人	3
学校35	11人	93人	158人	3
学校36	18人	173人	101人	9
学校37	90人	40人	109人	17
学校38	102人	39人	191人	4
学校39	179人	14人	138人	15
学校40	20人	12人	141人	14
学校41	115人	83人	25人	13
合计	270人	160人	110人	21

学校名称	总人数	男生	女生	班级数
学校42	173人	116人	20人	20
学校43	162人	124人	121人	19
学校44	147人	194人	160人	5
学校45	71人	95人	184人	12
学校46	116人	38人	115人	5
学校47	152人	62人	155人	17
学校48	111人	133人	170人	14
学校49	38人	52人	70人	20
学校50	199人	137人	143人	20
学校51	31人	15人	20人	6
学校52	39人	157人	33人	17
学校53	186人	40人	96人	7
学校54	186人	46人	82人	13
学校55	155人	172人	135人	18
学校56	152人	76人	165人	11
学校57	29人	183人	149人	7
学校58	33人	125人	93人	5
学校59	65人	70人	173人	5
学校60	62人	46人	65人	13
学校61	83人	123人	20人	13
学校62	20人	143人	75人	18
学校63	169人	155人	199人	4
学校64	129人	146人	113人	7
学校65	20人	185人	111人	15
学校66	29人	103人	142人	19
学校67	193人	196人	189人	20
学校68	97人	13人	17人	19
学校69	172人	125人	40人	17
学校70	46人	157人	78人	19
合计	270人	160人	110人	21

学校名称	总人数	男生	女生	班级数
学校71	49人	143人	76人	20
学校72	11人	106人	129人	8
学校73	139人	57人	190人	19
学校74	199人	10人	67人	6
学校75	190人	186人	154人	15
学校76	62人	165人	59人	16
学校77	150人	192人	148人	18
学校78	63人	146人	73人	7
学校79	186人	130人	100人	17
学校80	186人	195人	88人	11
学校81	133人	83人	195人	17
学校82	155人	88人	184人	10
学校83	112人	190人	183人	3
学校84	22人	56人	64人	19
学校85	97人	25人	165人	7
学校86	190人	45人	56人	4
学校87	40人	152人	31人	4
学校88	70人	182人	110人	19
学校89	74人	112人	137人	16
学校90	53人	196人	72人	5
学校91	119人	78人	50人	12
学校92	123人	177人	194人	13
学校93	64人	17人	31人	8
学校94	45人	137人	57人	20
学校95	30人	71人	133人	4
学校96	125人	186人	86人	15
学校97	103人	17人	48人	4
学校98	73人	137人	171人	11
学校99	30人	100人	80人	14
合计	270人	160人	110人	21

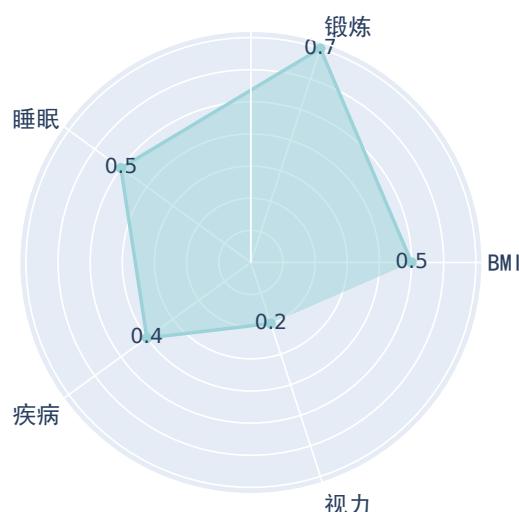
学校名称	总人数	男生	女生	班级数
学校100	55人	50人	126人	20
合计	270人	160人	110人	21

3.4.4 含有等级说明的5点量表



3.5 图表示例 – 数据图

3.5.1 雷达图



注：睡眠情况指学生的平均睡眠时间/天， ≥ 8 小时为合格。

图2 学生综合素质指标

3.5.2 呈现多组数据的雷达图

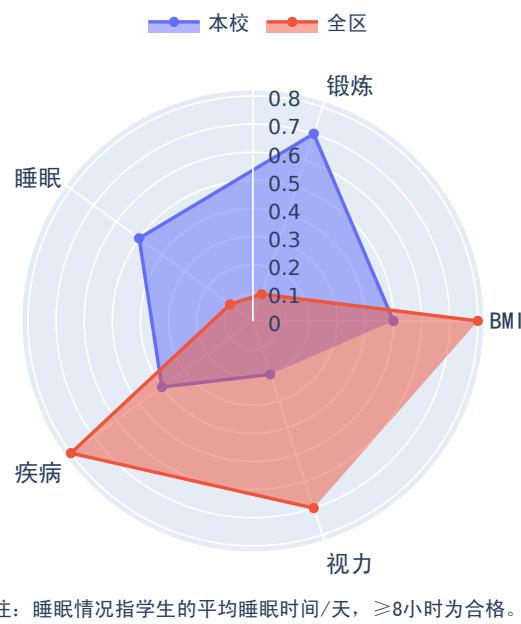


图3 学生综合素质指标比较

3.5.3 横向柱形图

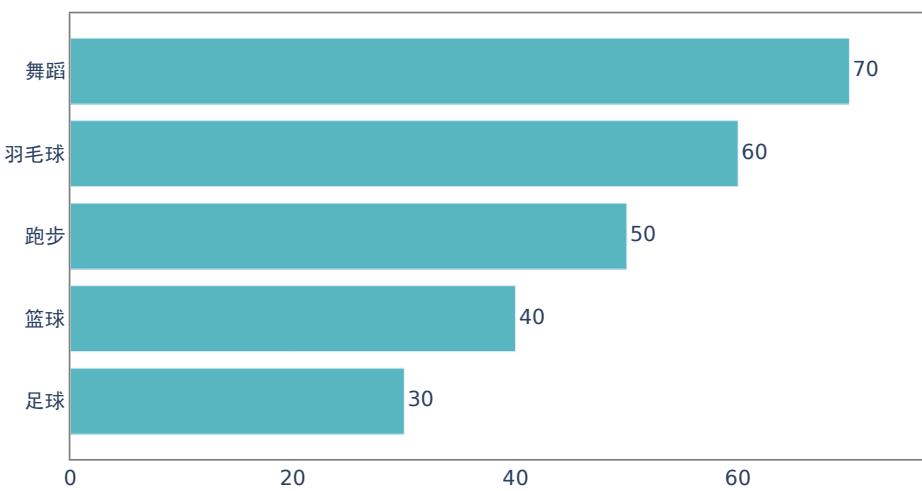
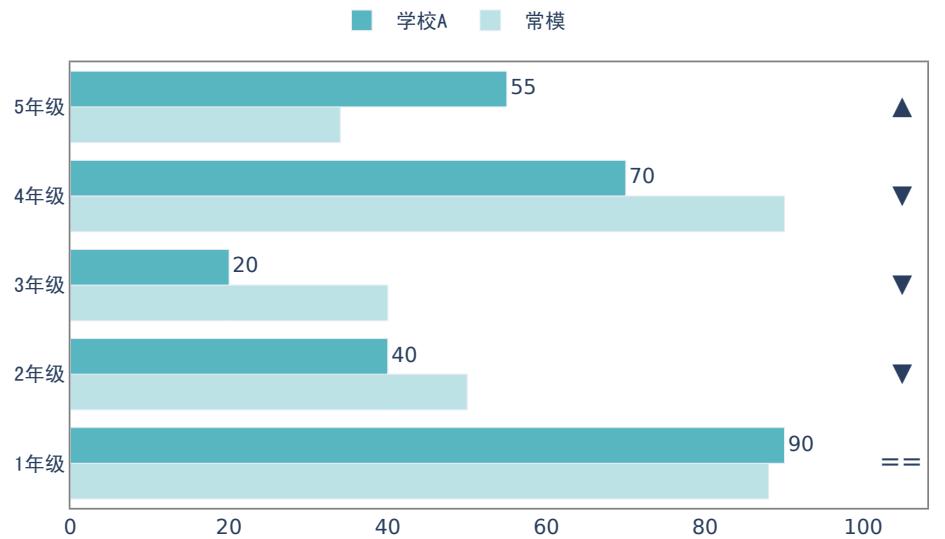


图4 本校学生选择的运动种类比例

3.5.4 横向柱形图 – 两组数据比较



说明: ▲代表本校的得分在常模以上, ▼代表本校的得分在常模以下, ==代表本校和常模之间没有差异

图5 学校A与常模的对比

3.5.5 堆叠柱形图

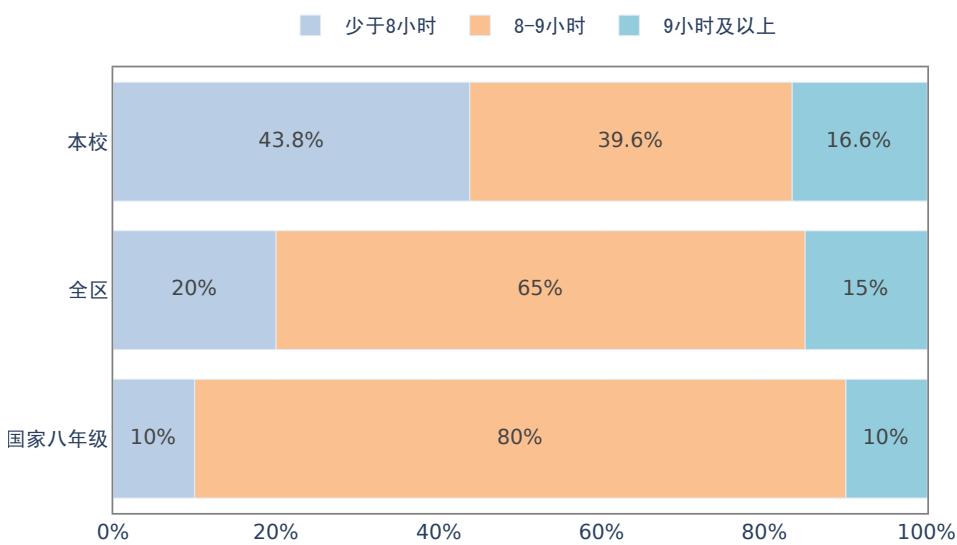


图6 本校学生睡眠情况与全区、国家八年级的比较

3.5.6 堆叠柱形图 – 划分正负性评价

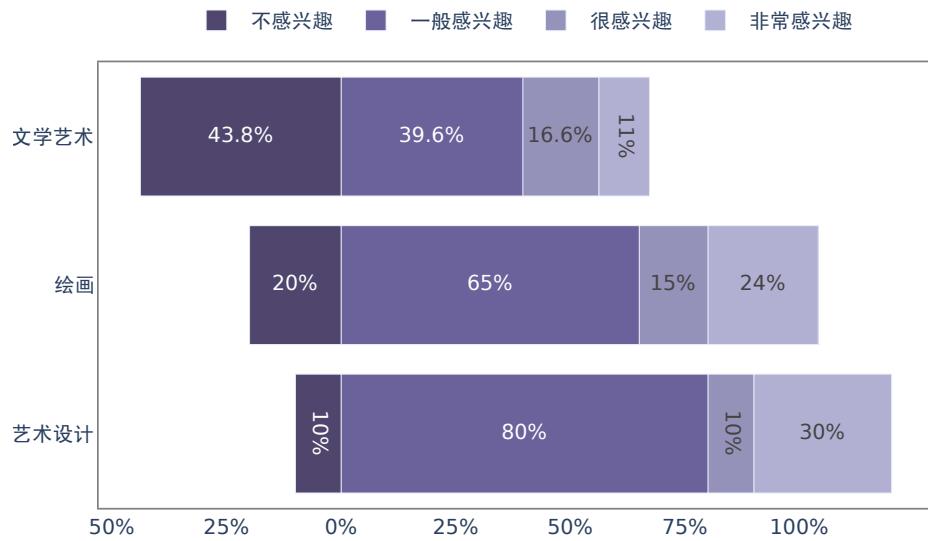


图7 艺术类上不同兴趣水平学生的人数占比

3.5.7 堆叠柱形图 – 多组数据的分布比较

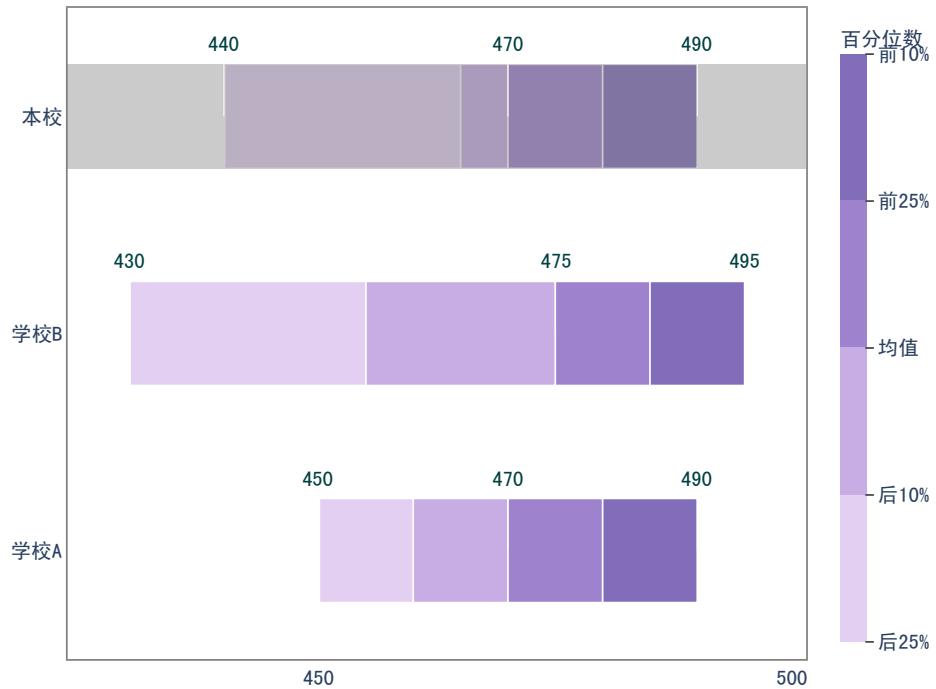


图8 本校与本区其他学校学习能力的比较

3.5.8 饼图

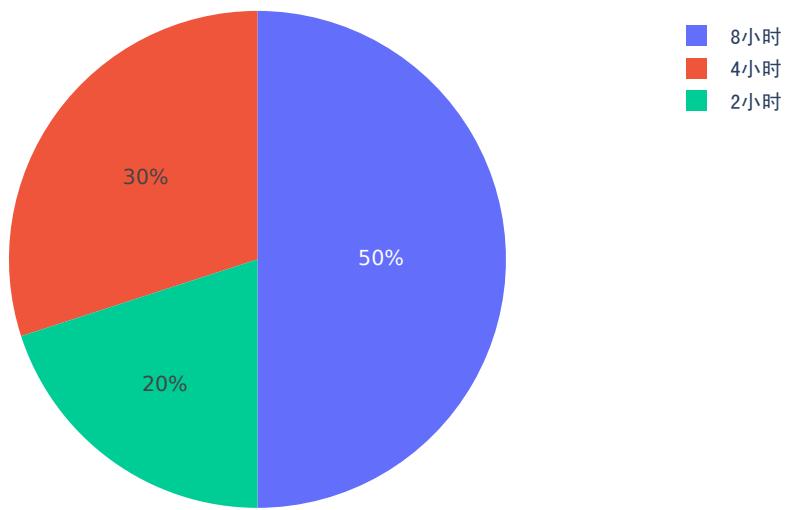


图9 本校学生锻炼时间分布情况

3.5.9 用方块面积表示百分比



3.5.10 分布位点图

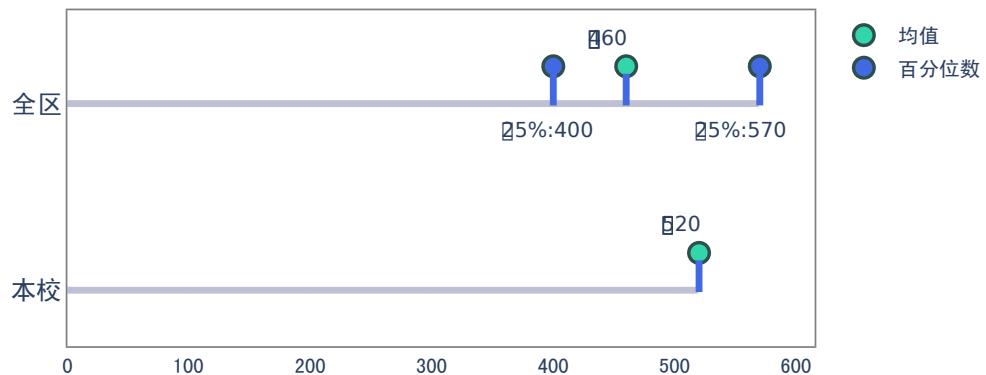


图10 本校学生体验美得分与全区比较

3.6 图表示例 – 较复杂的复合图

有时一般的数据图表无法满足直观和美观等要求，因此需要通过美术设计来制作比较复杂的图表。

这里我们可以使用 `svgutils` 工具包对多个图片进行组合，以及添加图形、文字等元素，进一步丰富图片的内容。

可参考下面这个例子制作图片。

3.6.1 综合能力得分

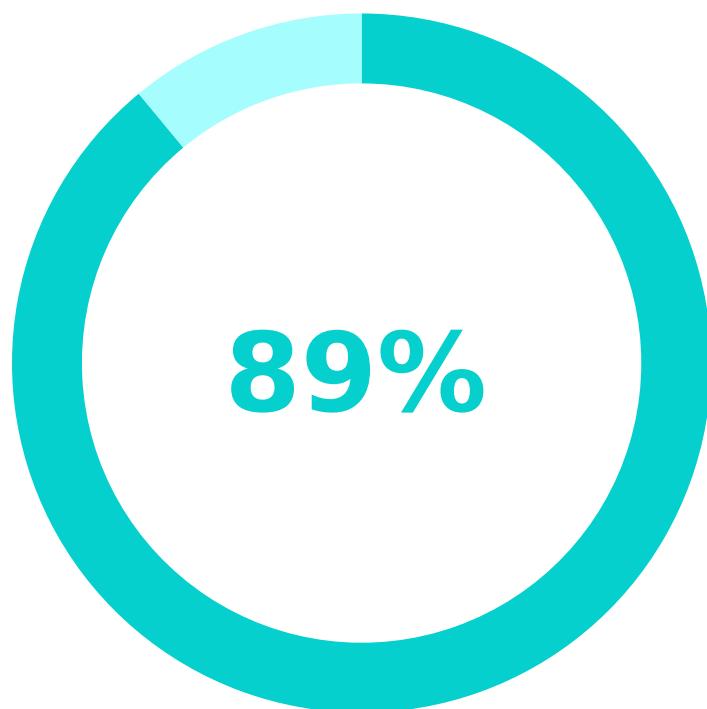


图11 一张合成的复杂图

3. 6. 2 利用\$\pi\$图展示得分

■ 得分：**598**

■ 等级：**13**



■ 超过同年级学生比例

4. 生成PDF格式的报告文档

在使用Jupyter Notebook完成报告内容后，可以使用 **SmartyDoc** 提供的处理流程，将ipynb格式的报告文本转换成PDF格式。

4. 1 PDF文档转换流程

在 **SmartyDoc** 的框架下，Jupyter Notebook的文档首先转换为html网页，之后配合css文件（即样式设计文件），转换为最终的PDF文件。

4. 2 需要准备哪些文件

编写好的Jupyter Notebook文件（ipynb文件）。

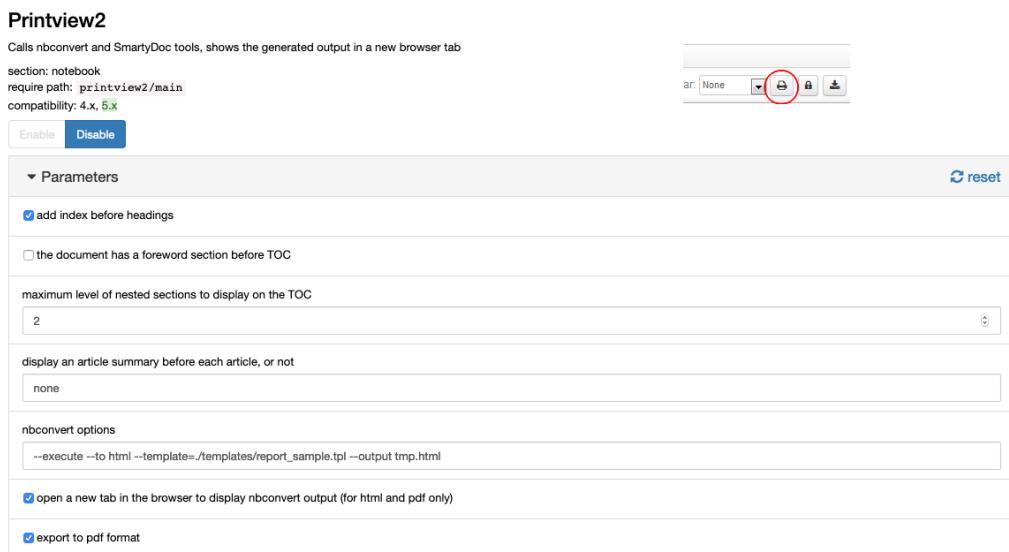
ipynb文件转换为html文件时所需的转换模版，具体地，该文件保存于 **SmartyDoc** 目录下的 **templates** 目录内，您可以将该目录直接拷贝到与ipynb文件同级的目录内。用户不需要对 **templates** 目录下的文件做任何修改。

css文件，用于描述文字、图片等内容的显示样式

文档内要插入的图片等文件

4. 3 使用printview2插件实现PDF文档生成

为了方便用户在Jupyter的操作界面下方便地完成PDF文件生成的操作，我们专门开发了一个Jupyter的扩展插件 **printview2**，为用户提供图形化的操作界面。用户可以在Jupyter的Nbextensions扩展管理页内，通过勾选框启用扩展。并在勾选框下方的页面设置PDF文件生成所需的参数。具体节目如下图。



在准备好格式转换所需的文件后，根据要生成的PDF文件的样式，对 `printview2` 的参数进行设置，具体包括：

`add index before headings`：若勾选此项，则系统会自动在章节名前面加上 1. 、 2. 、 2.1 这样的序列号；

`the document has a foreword section before TOC`：若勾选此项，则正文第一个 `h2` 所包含的内容会以 前言 的形势出现在目录之前，且该部分内容不计入页码；

`maximum level of nested sections to display on the TOC`：在目录中要呈现的目录级别，如果设置为 1，则只显示 `h2` 所表示的章名；

`display an article before each article, or not`：对各章的起始页样式进行设置，具体包括三种形式，分别为 `none` 、 `toc` 和 `intro`， `none` 表示章节首页不做特殊处理， `toc` 表示章节首页包含章名和该章内的二级标题列表， `intro` 表示章节首页包含章名和该章的内容简介。内容简介部分须在正文中以标签 括起来；

`nbconvert options`：ipynb文件转换为html文件时的参数设置，一般不需要用户做修改；

`open a new tab in the brower to display nbconvert output`：做格式转换完成后自动在浏览器打开一个新的标签页进行显示，Chrome浏览器查看PDF文件会出现问题，请直接下载文件查看转换结果；

`export to pdf format`：若勾选此项，生成PDF文件，否则只生成html网页文件。

完成以上设置后，在浏览器中切换到要进行格式转换的notebook的页面（注意：如果该页面在上述设置前已经打开，请对该页面进行刷新，以更新参数），点击工具栏中的打印图标，自动开始文件格式转换。

4.4 使用命令行操作实现PDF文档生成

待补充，敬请期待 ...