Tedu.cn 达内教育 | 达内教育研究院

数据分析与可视化

python数据分析与可视化

Tedu.cn 达内教育

1

pandas数据分析

基础知识

统计分析基础

Jupyter notebook介绍

数据预处理

达内教育研究院

1. ipython

2. 掌握 Jupyter Notebook

ipython

什么是ipython？

IPython——科学计算标准工具集的组成部分

IPython是一个免费、开源的项目，支持Linux、Unix、Mac OS X和Windows平台，其官方网址是http://ipython.org/。IPython的作者只要求你在用到IPython的科技著作中注明引用即可。

IPython中包括各种组件，其中的两个主要组件是：

基于终端方式和基于Qt的交互式Python shell

支持多媒体和绘图功能的基于Web的notebook（版本号为0.12以上的IPython支持此功能）

ipython

以理服人

IPython项目起初是Fernando Pérez在2001年的一个用以加强和Python交互的子项目。在随后的16年中，它成为了Python数据栈最重要的工具之一。虽然IPython本身没有提供计算和数据分析的工具，它却可以大大提高交互式计算和软件开发的生产率。IPython鼓励“执行-探索”的工作流，区别于其它编程软件的“编辑-编译-运行”的工作流。它还可以方便地访问系统的shell和文件系统。因为大部分的数据分析代码包括探索、试错和重复，IPython可以使工作更快。

2014年，Fernando和IPython团队宣布了Jupyter项目，一个更宽泛的多语言交互计算工具的计划。IPython web notebook变成了Jupyter notebook，现在支持40种编程语言。IPython现在可以作为Jupyter使用Python的内核（一种编程语言模式）。

IPython变成了Jupyter庞大开源项目（一个交互和探索式计算的高效环境）中的一个组件。它最老也是最简单的模式，现在是一个用于编写、测试、调试Python代码的强化shell。你还可以使用通过Jupyter Notebook，一个支持多种语言的交互式网络代码“笔记本”，来使用IPython。IPython shell 和Jupyter notebooks特别适合进行数据探索和可视化。

Jupyter notebooks还可以编写Markdown和HTML内容，提供了一种创建代码和文本的富文本方法。其它编程语言也在Jupyter中植入了内核，好让在Jupyter中可以使用Python另外的语言。

ipython

ipython安装

windows： 前提是有numpy，matplotlib pandas

更新pip python -m pip install --upgrade pip

采用pip安装 pip install ipython

在Mac OS X中安装IPython：

如有必要，请先安装苹果开发工具Xcode，可以在Mac电脑附带的OSX DVD光盘中或者苹果应用商店中找到Xcode。使用easy_install或pip安装IPython，或者从源文件安装。

ipython

以理服人

ipython壳的主要特点

提供一个更友好的界面，是一个增强的Python shell，IPython是Python的加强型交互式解释器。

目的是提高编写、测试、调试Python代码的速度。

提供了代码补全，对象检查，系统调用，获取历史数据等实用的功能

庞大的ipython社区努力使其成为一个高效的python科学计算环境

主要用于交互式数据并行处理，高效的交互式处理、呈现数据。

是分布式计算的基础架构。

提供了一个非常灵活的框架，可以作为其他应用的基础

ipython壳的主要内容

自动补全	Tab键
检查	？查看对象基本信息　　？？查看构造函数基本信息　　？*匹配对象
%run命令	使用%run调用外部Python脚本的能力　%run E:\pycharme\python_study
魔法方法	%magic来查找所有的魔法命令
异常和错误信息	使用%run命令行的方式运行，如果出现错误，ipython会打印错误的的路径和异常
和操作系统交互	可以输入shell命令，cd　pwd　env等
目标标签系统	创建同名目录　%bookmark TI C:\Users\user　　cd TI

以理服人

ipython其他技巧

获取历史数据　　IPython 中也可以通过 _ 和 __ 访问上一次和上上一次的输出

计时功能　　%time可以进行计时　　%timeit可以进行计时平均值

Jupyter Notebook安装

windows 更新pip python -m pip install --upgrade pip

采用pip安装 pip install Jupyter

以理服人

Jupyter Notebook（此前被称为 IPython notebook）是一个交互式笔记本，支持运行 40 多种编程语言。使用浏览器作为界面，向后台的IPython服务器发送请求，并显示结果。在浏览器的界面中使用单元(Cell)保存各种信息。Cell有多种类型，经常使用的有表示格式化文本的Markdown单元，和表示代码的Code单元.

Jupyter Notebook 的本质是一个 Web 应用程序，便于创建和共享文学化程序文档，支持实时代码，数学方程，可视化和 markdown。 已迅速成为处理数据的必备工具，用途包括：数据清理和转换，数值模拟，统计建模，机器学习等等

Jupyter 优势

可选择语言：支持超过40种编程语言，包括Python、R、Java等。

分享笔记本：可以使用电子邮件、GitHub和Jupyter Notebook Viewer与他人共享。

交互式输出：代码可以生成丰富的交互式输出，包括HTML、图像、视频、LaTeX等等。

大数据整合：通过Python、R编程语言使用Apache Spark等大数据框架工具。支持使用pandas、scikit-learn、ggplot2、TensorFlow来探索同一份数据。

以理服人

Notebook有两个部分组成：

网络应用：基于网络的文档处理环境，主要包括文本编辑，数学计算，及丰富的输出

Notebook文档：网络应用中所有可见的内容，主要包括文本，图片等

notebook 是 Donald Knuth 在 1984 年提出的文字表达化编程的一种形式。在文字表达化编程中，直接在代码旁写出叙述性文档，而不是另外编写单独的文档。用 Donald Knuth 的话来说：

让我们集中精力向人们解释我们希望计算机做什么，而不是指示计算机做什么。

归根到底，代码是写给人看到，不是写给计算机看的。notebook 恰恰提供了这种能力。能够直接在代码旁写出叙述性文档。这不仅对阅读 notebook 的人很有用，而且将来回头分析代码也很有用

以理服人

主要特性有：

在浏览器编辑代码，会自动进行语法加亮，缩进，补齐，检查等

通过浏览器运行代码，运行结果直接输出到代码的后面

有多种输出方式，包括PNG,SVG,HTML,LaTeX

在浏览器中，使用Markdown标记语言为代码提供注释，不再仅限于普通文本的注释方式

可以方便的标记数学公式

导出数据和数据分析过程

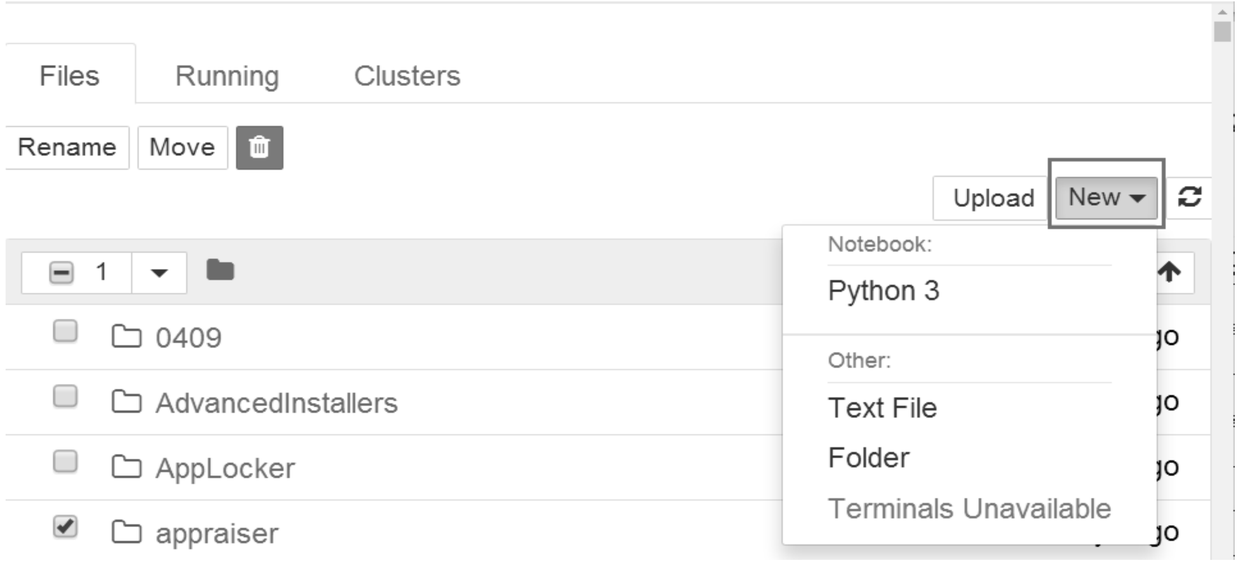
Jupyter Notebook简介与安装

打开并新建一个Notebook

以理服人

打开 Jupyter Notebook

- “Text File”为纯文本型
- “Folder” 为文件夹
- “Python 3” 表示 Python 运行脚本



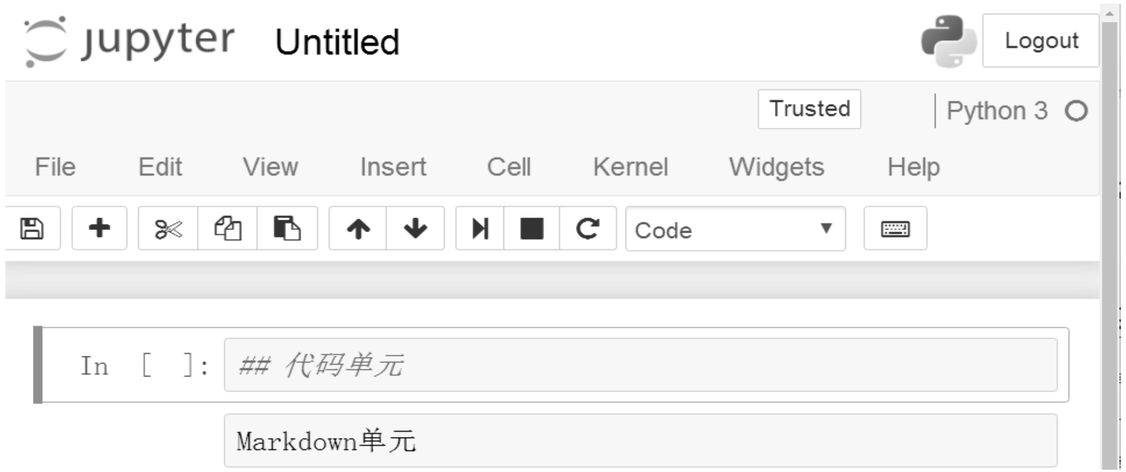
Jupyter Notebook简介与安装

Jupyter Notebook 的界面及其构成

选择”Python 3”选项，进入 Python 脚本编辑界面，Notebook 文档由一系列单元（Cell）构成，主要有两种形式的单元。

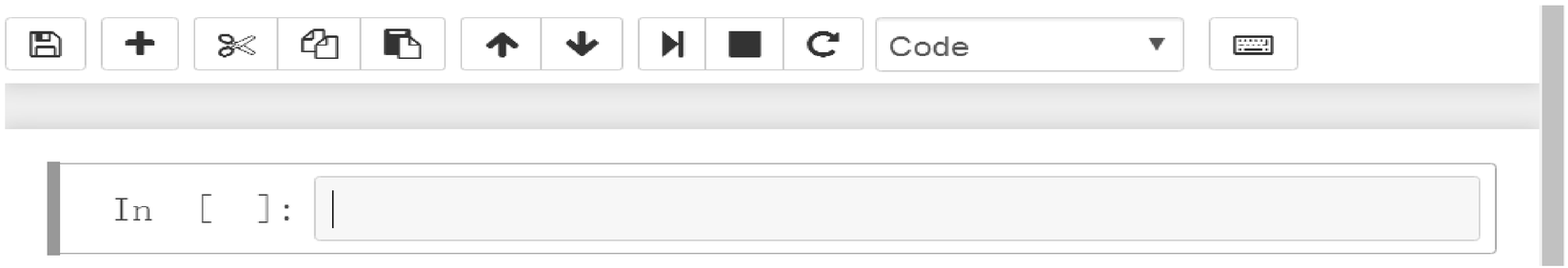
代码单元。这里是读者编写代码的地方。

Markdown 单元。在这里对文本进行编辑。

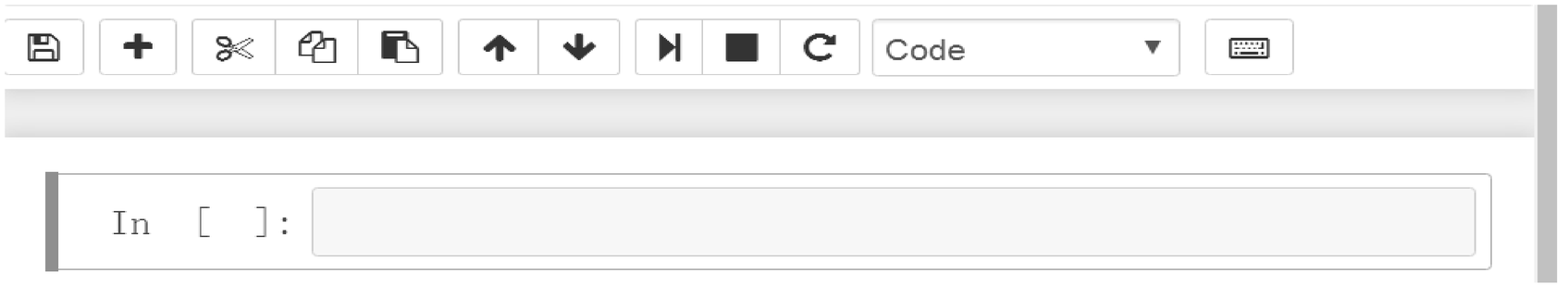


Jupyter Notebook简介与安装

编辑界面：用于编辑文本和代码



命令模式：用于执行键盘输入的快捷命令。



Jupyter Notebook简介与安装

快捷键

“Esc”键：进入命令模式

“Y”键：切换到代码单元

“M”键：切换到 Markdown 单元

“B”键：在本单元的下方增加一单元

“H”键：查看所有快捷命令

“Shift + Enter”组合键：运行代码

Markdown 使用

Markdown 是一种可以使用普通文本编辑器编写的标记语言，通过简单的标记语法，它可以使普通文本内容具有一定的格式。

- 标题：标题是标明文章和作品等内容的简短语句。一个“#”字符代表一级标题，以此类推。

```
# 一级标题
## 二级标题
### 三级标题
#### 四级标题
##### 五级标题
###### 六级标题
```

Markdown 使用

列表：列表是一种由数据项构成的有限序列，即按照一定的线性顺序排列而成的数据项的集合。

对于无序列表，使用星号、加号或者减号作为列表标记

对于有序列表，则是使用数字“，”“（一个空格）”。


```
* Python
+ Python2
- Python3


1. Python
2. Python2
3. Python3
```

- Python
- Python2
- Python3

1. Python
2. Python2
3. Python3

Markdown 使用

加粗 / 斜体：前后有两个星号或下划线表示加粗，前后有 3 个星号或下划线表示斜体。

以理服人

```
Python数据分析

**Python数据分析**

***Python数据分析***

__Python数据分析__

___python数据分析___
```

Python数据分析

Python数据分析

Python数据分析

Python数据分析

python数据分析

Markdown 使用

表格：代码的第一行表示表头，第二行分隔表头和主体部分，从第三行开始，每一行代表一个表格行；列与列之间用符号“|”隔开

以理服人

```
Python | R| MATLAB|
------------ | ------------|----|
接口统一，学习曲线平缓 | 接口众多，学习曲线陡峭 |自由度大，学习曲线较为平缓 |
开源免费 | 开源免费 | 商业收费 |
```

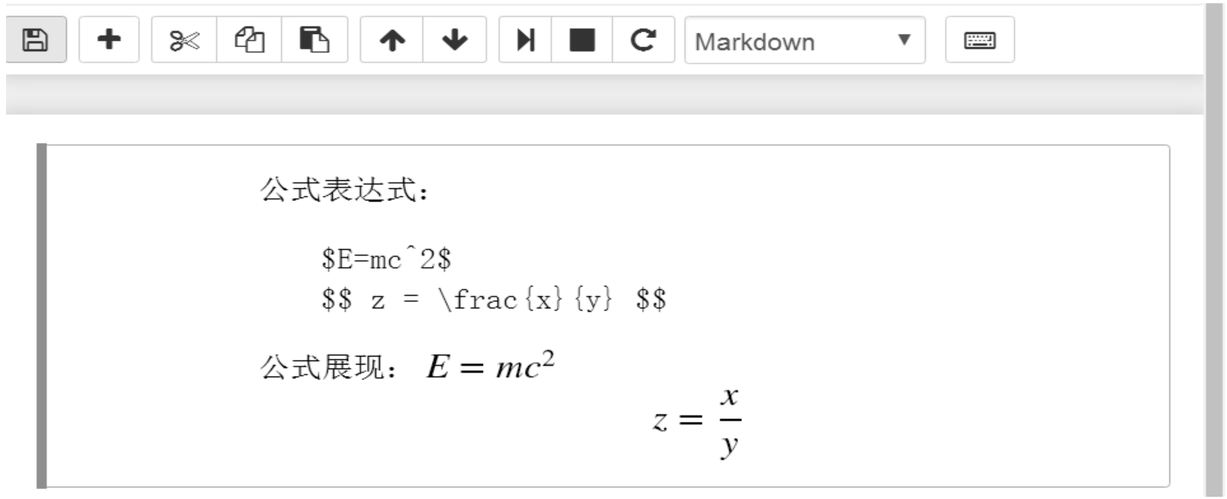
Python	R	MATLAB
接口统一，学习曲线平缓	接口众多，学习曲线陡峭	自由度大，学习曲线较为平缓
开源免费	开源免费	商业收费

Markdown 使用

数学公式编辑：LaTeX 是写科研论文的必备工具， Markdown 单元中也可以使用 LaTeX 来插入数学公式。

在文本行中插入数学公式，应在公式前后分别加上一个“$” 符号

如果要插入一个数学区块，则在公式前后分别加上两个“$$”符号。



Markdown 使用

LaTeX语法表示数学符号示例


```
\begin{displaymath}
\sum_{i=1}^{n} \qquad
\int_{0}^{\frac{\pi}{2}} \qquad
\prod_\epsilon
\end{displaymath}
```

$$\sum_{i=1}^{n} \qquad \int_{0}^{\frac{\pi}{2}} \qquad \prod_{\epsilon}$$

```
$a_{1}$ \qquad $x^{2}$ \qquad
$e^{-\alpha t}$ \qquad
$a^{3}_{ij}$\\
$e^{x^2} \neq {e^x}^2$
```

$a_1 \qquad x^2 \qquad e^{-\alpha t} \qquad a^3_{ij}$
$e^{x^2} \neq {e^x}^2$

```
$\sqrt{x}$ \qquad
$\sqrt{ x^{2}+\sqrt{y} }$
\qquad $\sqrt[3]{2}$\\[3pt]
$\surd[x^2 + y^2]$
```

$\sqrt{x} \qquad \sqrt{x^{2}+\sqrt{y}} \qquad \sqrt[3]{2}$
$\surd[x^2 + y^2]$

```
$1\frac{1}{2}$~hours
\begin{displaymath}
\frac{ x^{2} }{ k+1 }\qquad
x^{ \frac{2}{k+1} }\qquad
x^{ 1/2 }
\end{displaymath}
```

$1\frac{1}{2}$ hours

$$\frac{x^2}{k+1} \qquad x^{\frac{2}{k+1}} \qquad x^{1/2}$$

Markdown 使用

LaTeX语法集合

表 3.1: 数学模式重音符

$\hat{a}$	\hat{a}	$\check{a}$	\check{a}	$\tilde{a}$	\tilde{a}	$\acute{a}$	\acute{a}
$\grave{a}$	\grave{a}	$\dot{a}$	\dot{a}	$\ddot{a}$	\ddot{a}	$\breve{a}$	\breve{a}
$\bar{a}$	\bar{a}	$\vec{a}$	\vec{a}	$\widehat{A}$	\widehat{A}	$\widetilde{A}$	\widetilde{A}

表 3.2: 小写希腊字母

α	\alpha	θ	\theta	o	o	υ	\upsilon
β	\beta	ϑ	\vartheta	π	\pi	ϕ	\phi
γ	\gamma	ι	\iota	ϖ	\varpi	φ	\varphi
δ	\delta	κ	\kappa	ρ	\rho	χ	\chi
ϵ	\epsilon	λ	\lambda	ϱ	\varrho	ψ	\psi
ε	\varepsilon	μ	\mu	σ	\sigma	ω	\omega
ζ	\zeta	ν	\nu	ς	\varsigma		
η	\eta	ξ	\xi	τ	\tau		

表 3.3: 大写希腊字母

Γ	\Gamma	Λ	\Lambda	Σ	\Sigma	Ψ	\Psi
Δ	\Delta	Ξ	\Xi	Υ	\Upsilon	Ω	\Omega
Θ	\Theta	Π	\Pi	Φ	\Phi		

以理服人

Markdown 使用

LaTeX语法集合

以理服人

表 3.4: 二元关系符

下述命令的前面加上 \not 来得到其否定形式。

<	<	>	>	=	=
≤	\leq or \le	≥	\geq or \ge	≡	\equiv
≪	\ll	≫	\gg	≐	\doteq
≺	\prec	≻	\succ	∼	\sim
⪯	\preceq	⪰	\succeq	≃	\simeq
⊂	\subset	⊃	\supset	≈	\approx
⊆	\subseteq	⊇	\supseteq	≅	\cong
⊏	\sqsubset [a]	⊐	\sqsupset [a]	⋈	\Join [a]
⊑	\sqsubseteq	⊒	\sqsupseteq	⋈	\bowtie
∈	\in	∋	\ni , \owns	∝	\propto
⊢	\vdash	⊣	\dashv	⊨	\models
∣	\mid	∥	\parallel	⊥	\perp
⌣	\smile	⌢	\frown	≍	\asymp
:	:	∉	\notin	≠	\neq or \ne

表 3.5: 二元运算符

+	+	−	-		
±	\pm	∓	\mp	◁	\triangleleft
·	\cdot	÷	\div	▷	\triangleright
×	\times	\	\setminus	⋆	\star
∪	\cup	∩	\cap	∗	\ast
⊔	\sqcup	⊓	\sqcap	∘	\circ
∨	\vee , \lor	∧	\wedge , \land	•	\bullet
⊕	\oplus	⊖	\ominus	⋄	\diamond
⊙	\odot	⊘	\oslash	⊎	\uplus
⊗	\otimes	◯	\bigcirc	⨿	\amalg
△	\bigtriangleup	▽	\bigtriangledown	†	\dagger
◁	\lhd [a]	▷	\rhd [a]	‡	\ddagger
⊴	\unlhd [a]	⊵	\unrhd [a]	≀	\wr

Markdown 使用

LaTeX语法集合

以理服人

表 3.8: 定界符

(	(	)	)	↑	\uparrow	⇑	\Uparrow
[	[or \lbrack	]	] or \rbrack	↓	\downarrow	⇓	\Downarrow
{	\{ or \lbrace	}	\} or \rbrace	↕	\updownarrow	⇕	\Updownarrow
⟨	\langle	⟩	\rangle	|	| or \vert	‖	\\| or \Vert
⌊	\lfloor	⌋	\rfloor	⌈	\lceil	⌉	\rceil
/	/	\	\backslash		. (dual. empty)		

表 3.9: 大尺寸定界符

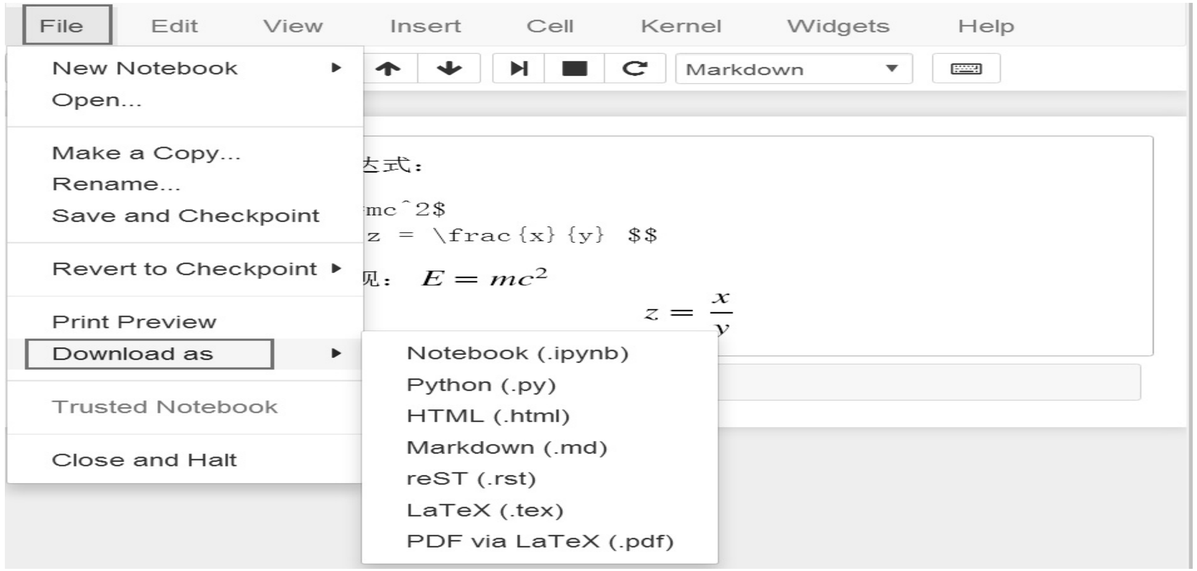
⎛	\lgroup	⎞	\rgroup	⎰	\lmoustache	⎱	\rmoustache
⎜	\arrowvert	⎟	\Arrowvert	⎪	\bracevert		

Markdown 使用

导出功能

Notebook 还有一个强大的特性，就是导出功能。可以将 Notebook 导出为多种格式，如HTML、Markdown、reST、PDF（通过 LaTeX）等格式。

导出功能可通过选择“File”→“Download as”级联菜单中的命令实现。



以理服人

Tedu.cn
达内教育

谢谢