

1. Markdown 简介及应用

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。

优点：

(1)、因为是纯文本，所以只要支持 Markdown 的地方都能获得一样的编辑效果，可以让作者摆脱排版的困扰，专心写作。

(2)、操作简单。比如:WYSIWYG 编辑时标记个标题，先选中内容，再点击导航栏的标题按钮，选择几级标题。要三个步骤。而 Markdown 只需要在标题内容前加#即可

缺点：

(1)、需要记一些语法（当然，是很简单。五分钟学会）。

(2)、有些平台不支持 Markdown 编辑模式。

由于我们有了 RStudio 这样的神级编辑器，我们还可以快速将 Markdown 转化为演讲 PPT、Word 产品文档、LaTeX 论文甚至是用非常少量的代码完成最小可用原型。

1.1 Rmarkdown

在 Windows 10 系统下配置 RStudio 的 R Markown，导出 PDF，Word 和 HTML。

(1)、安装 Pandoc 和 Miktex

Pandoc: <http://www.pandoc.org/installing.html>

Miktex: <https://miktex.org/2.9/setup> （最好选择 Net Installer 安装完整版本）

(2) RStudio 环境设置

安装两个包：

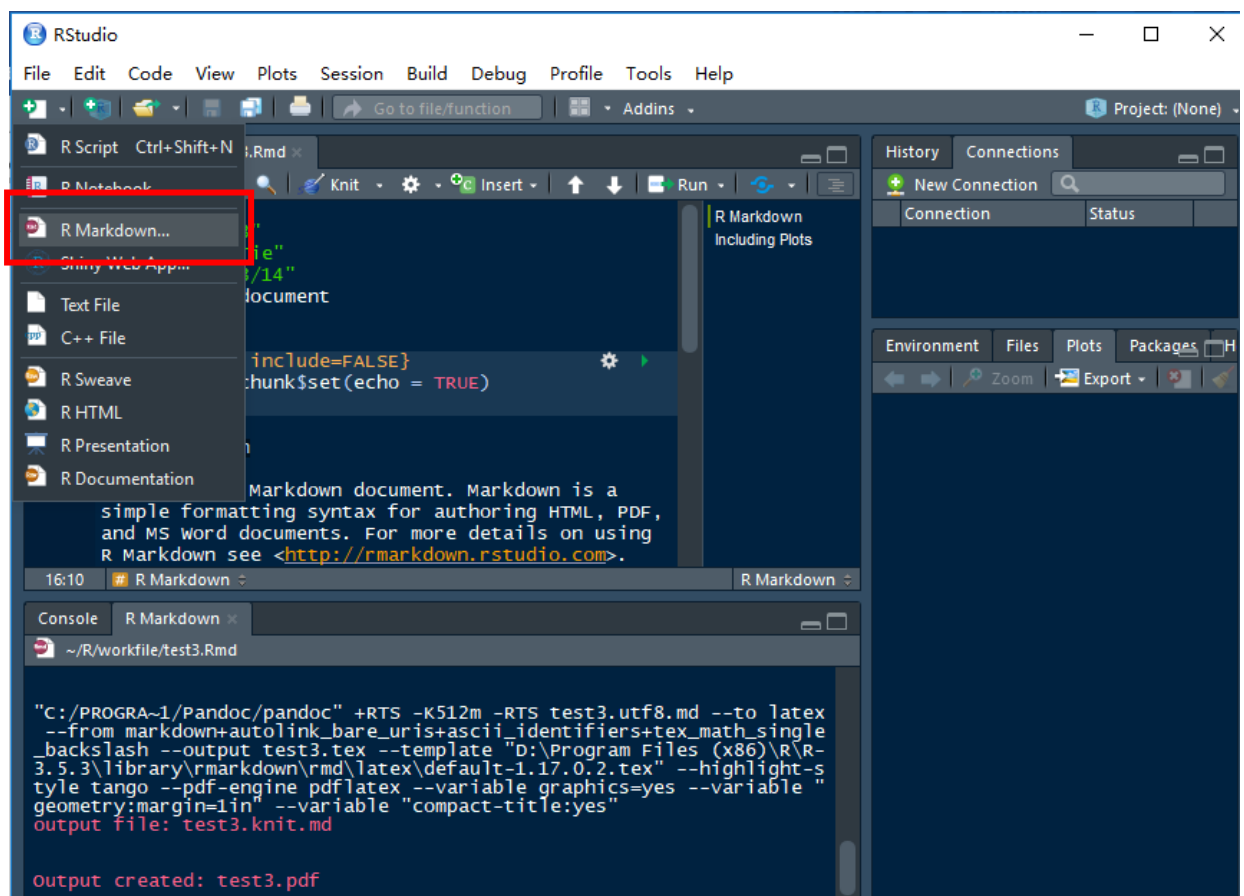
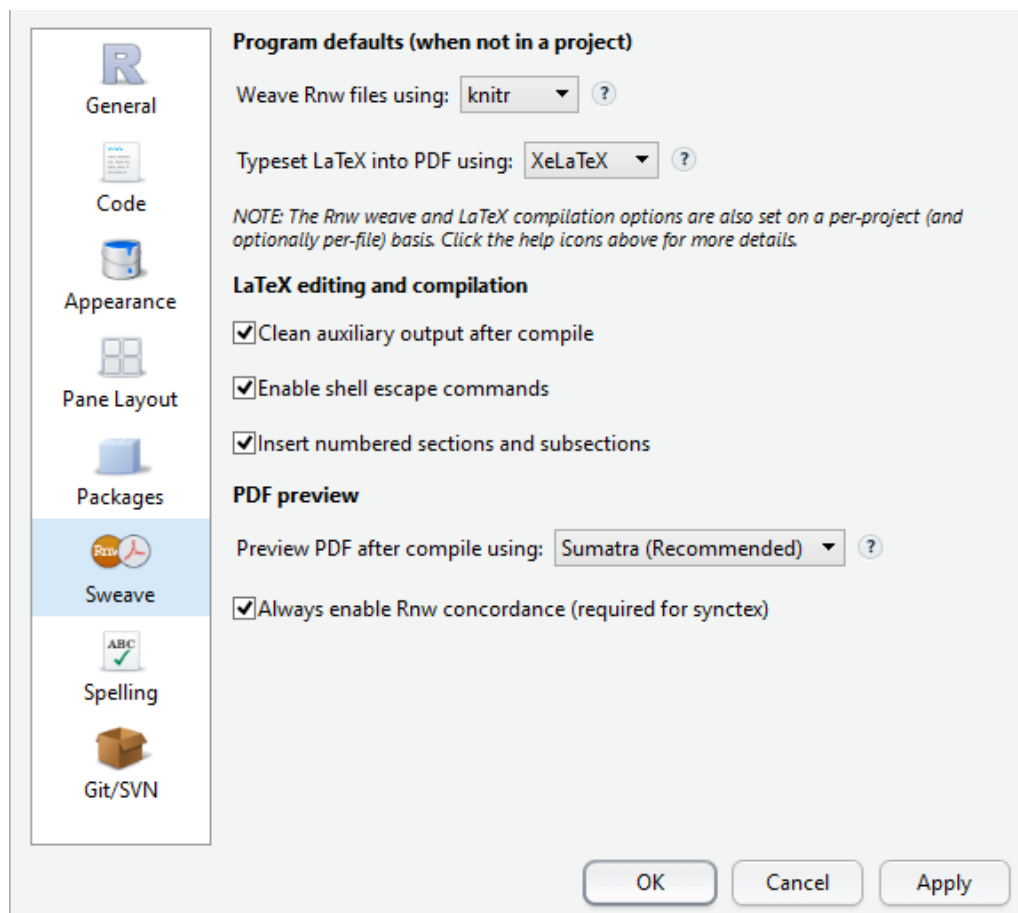
```
install.packages("knitr")
```

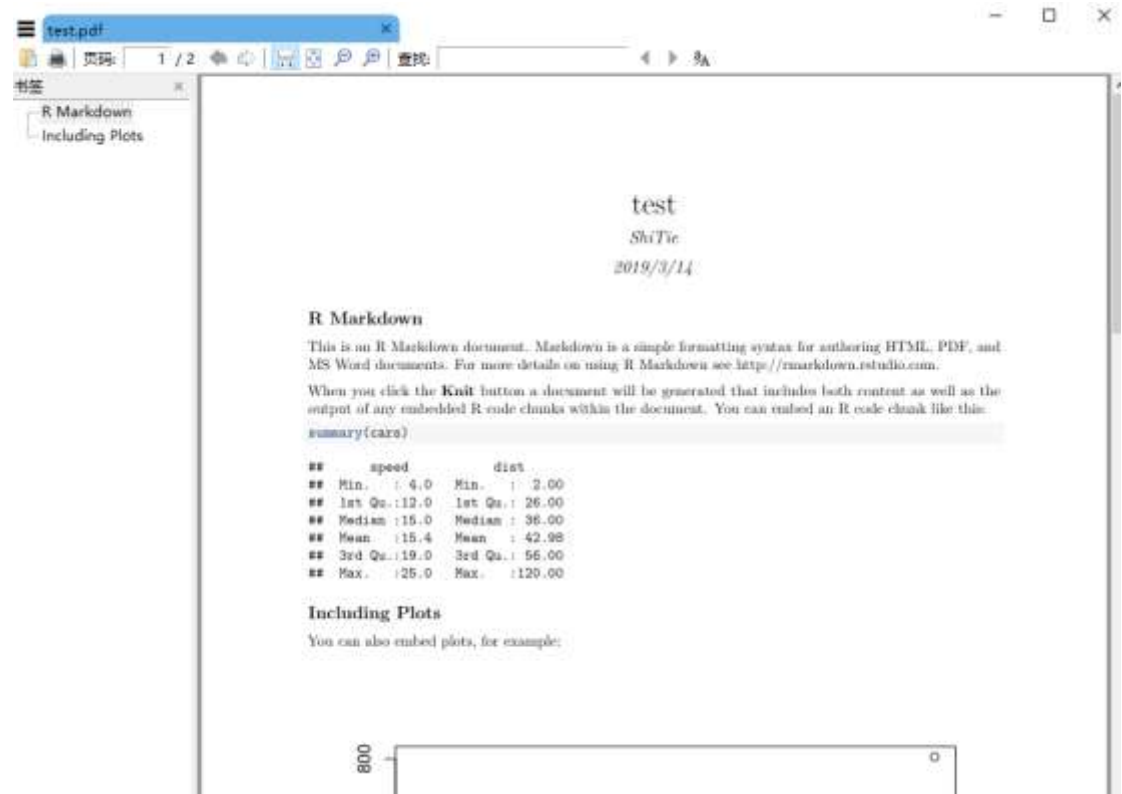
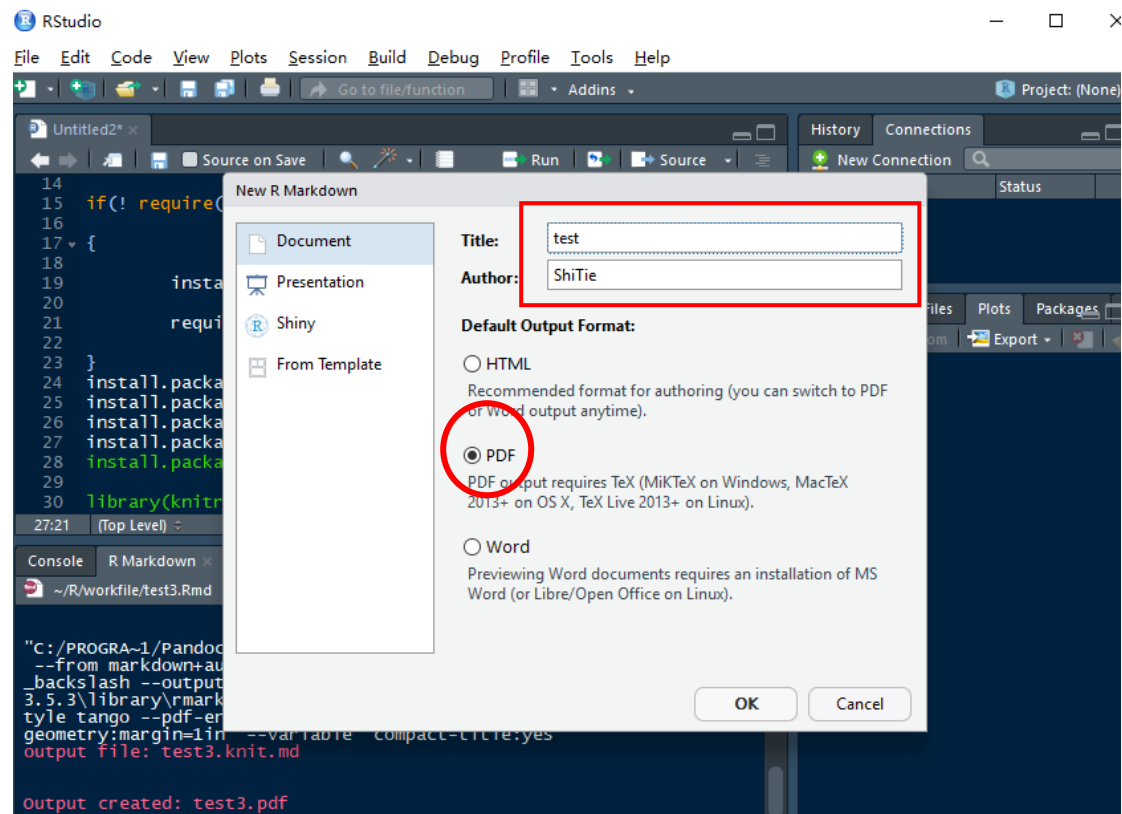
```
install.packages("rmarkdown")
```

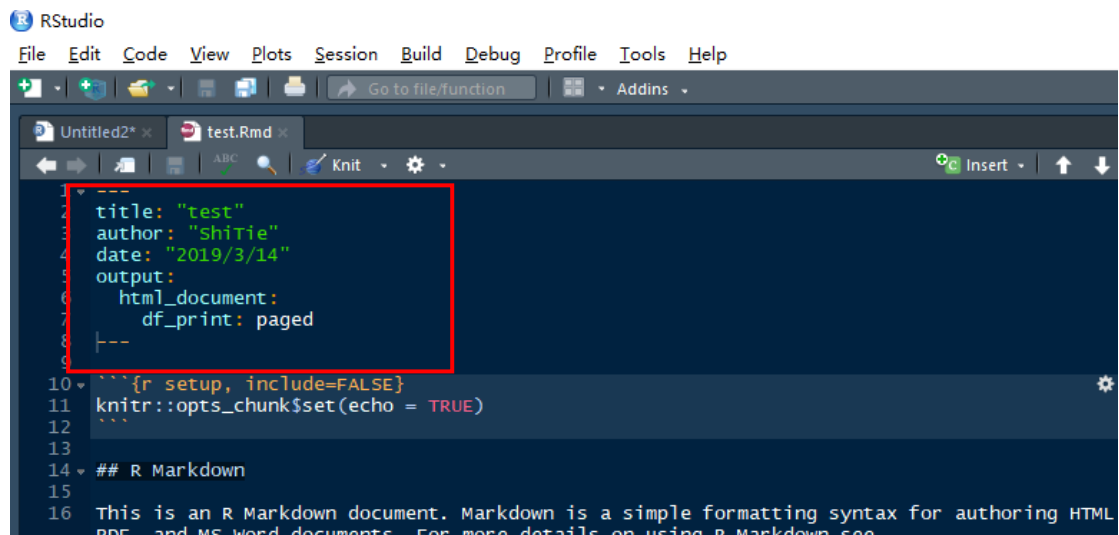
Rstudio 主菜单栏：

Tools -- Global Options -- General，设置缺省的编码格式为 UTF-8 (Default text coding: UTF-8)。当然，如果是默认的[Ask]状态，Rstudio 弹框框出来的时候再选 UTF-8 也是可以的。

Tools -- Global Options -- Sweave，将编译器设置为 pdfTeX (XeLaTeX)



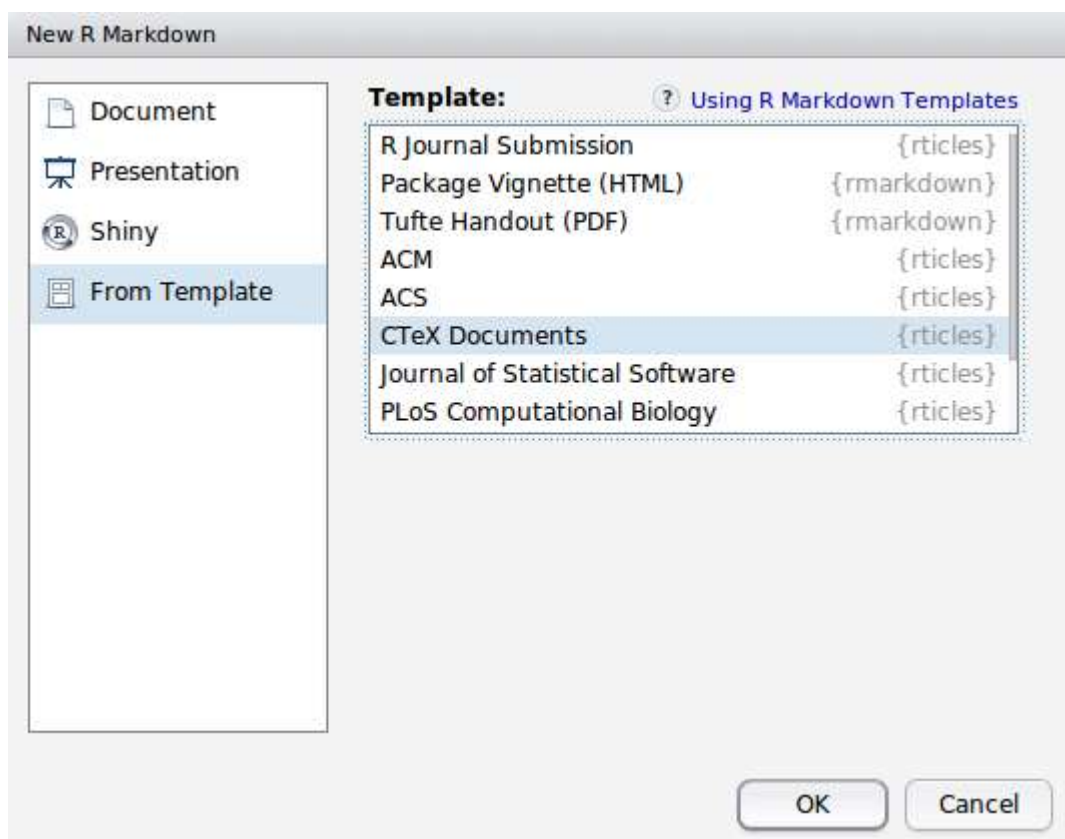




yaml 栏设置

中文环境设置

(1) `install.packages("rticles")`



(2) 在 yaml 栏添加一个 `header.tex` 文件

<https://www.cnblogs.com/loca/p/4541679.html>

https://www.cnblogs.com/Xeonilian/p/7142379.html#_caption_8

1.2 Statamarkdown

Rodríguez, G., Stata Tutorial, Stata Markdown

<https://data.princeton.edu/stata>

<https://data.princeton.edu/stata/markdown>

1.3 markdown 基本语法

标题

一个#是一级标题，二个#是二级标题，以此类推。支持六级标题

注：标准语法一般在#后跟个空格再写文字

这是一级标题

这是二级标题

这是三级标题

这是四级标题

这是五级标题

这是六级标题

字体

****这是加粗的文字****

这是倾斜的文字

******这是斜体加粗的文字******

~~~~这是加删除线的文字~~~~

效果：

**这是加粗的文字**

*这是倾斜的文字*

***这是斜体加粗的文字***

~~这是加删除线的文字~~

## 引用

在引用的文字前加>即可。引用也可以嵌套，如加两个>>三个>>>n个...

>这是引用的内容

>>这是引用的内容

>>>>>>>>>这是引用的内容

## 分割线

三个或者三个以上的 - 或者 \* 都可以。

---

----

\*\*\*

\*\*\*\*\*

## 图片

![图片 alt](图片地址 "图片 title")

图片 alt 就是显示在图片下面的文字，相当于对图片内容的解释。

图片 title 是图片的标题，当鼠标移到图片上时显示的内容。title 可加可不加

实例：

![blockchain](https://ss0.bdstatic.com/70cFvHSh\_Q1YnxGkpoWK1HF6hhy/it/u=702257389,1274025419&fm=27&gp=0.jpg "区块链")

效果：



Blockchain

更多介绍:

Markdown 基本语法: <https://www.jianshu.com/p/191d1e21f7ed>

rmarkdown-cheatsheet:

<https://www.rstudio.com/wp-content/uploads/2016/03/rmarkdown-cheatsheet-2.0.pdf>

## 2. Jupyter Notebook

### 2.1 Jupyter Notebook

#### (1) Anaconda

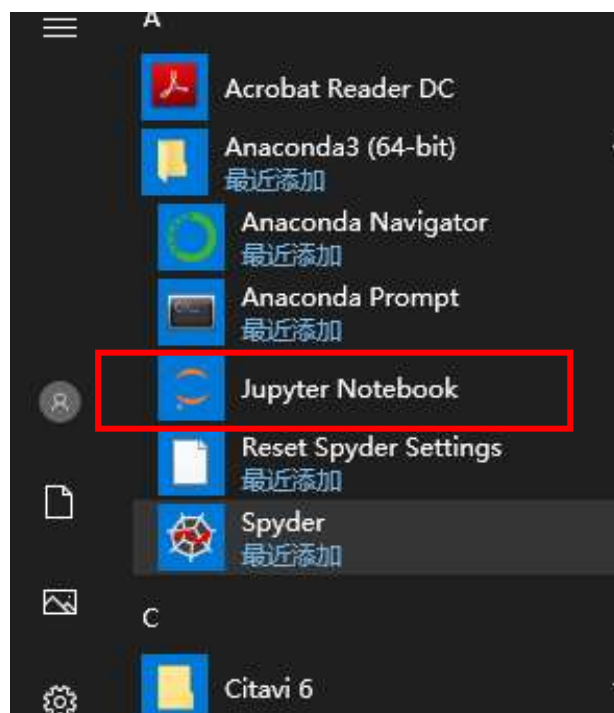
新用户可以使用 Anaconda 发行版来同时安装 Python 和 Jupyter Notebooks。Anaconda 安装了这两种工具，并包含了数据科学和机器学习社区中常用的很多软件包。你可以从这里下载最新版本的 Anaconda。

<https://www.anaconda.com/distribution/>

<https://jupyter.readthedocs.io/en/latest/install.html>



#### Get Started with Anaconda Distribution





## (2) pip 安装方法

如果出于某种原因，你决定不使用 Anaconda，那么你需要确保你的机器正在运行最新版本的 pip。怎么做？如果你已经安装了 Python，那么 pip 已经安装好了。要升级到最新的版本，请参照下面的代码：

```
#Linux and OSX
pip install -U pip setuptools
#Windows
python -m pip install -U pip setuptools
```

一旦 pip 安装完毕，你可以继续安装 Jupyter：

```
#For Python2
```

```
pip install jupyter
```

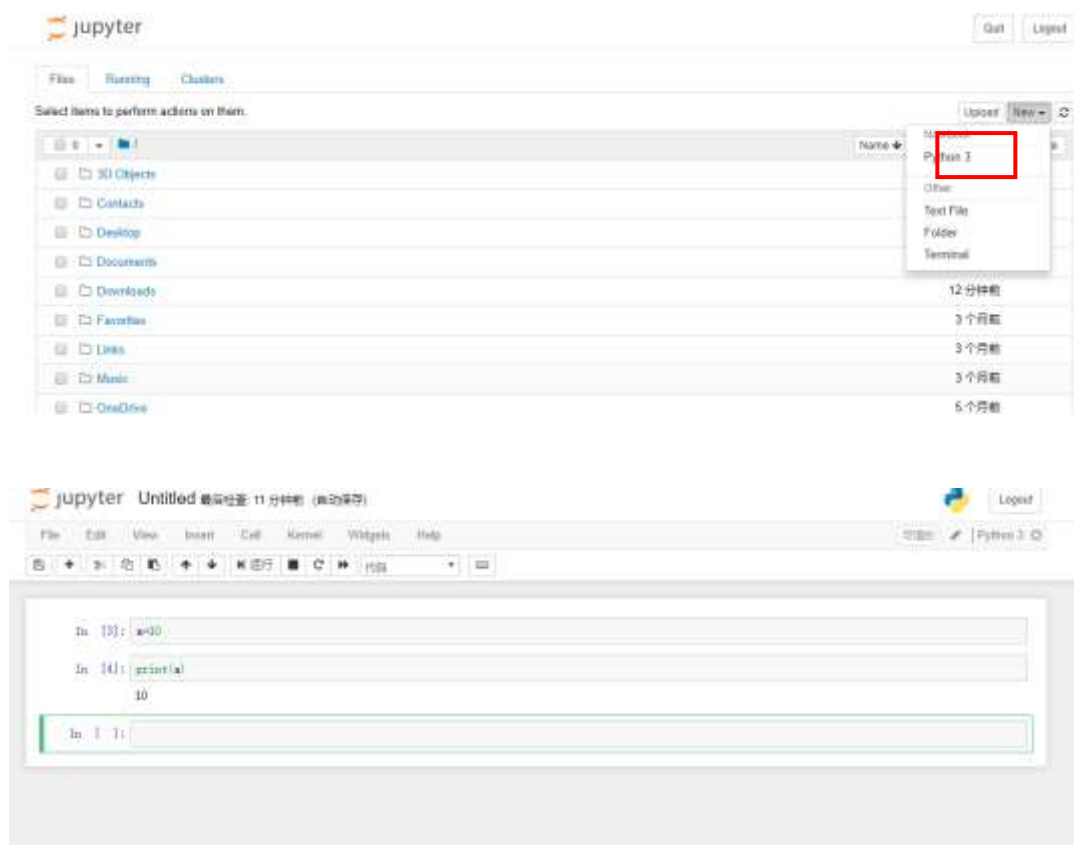
```
#For Python3
```

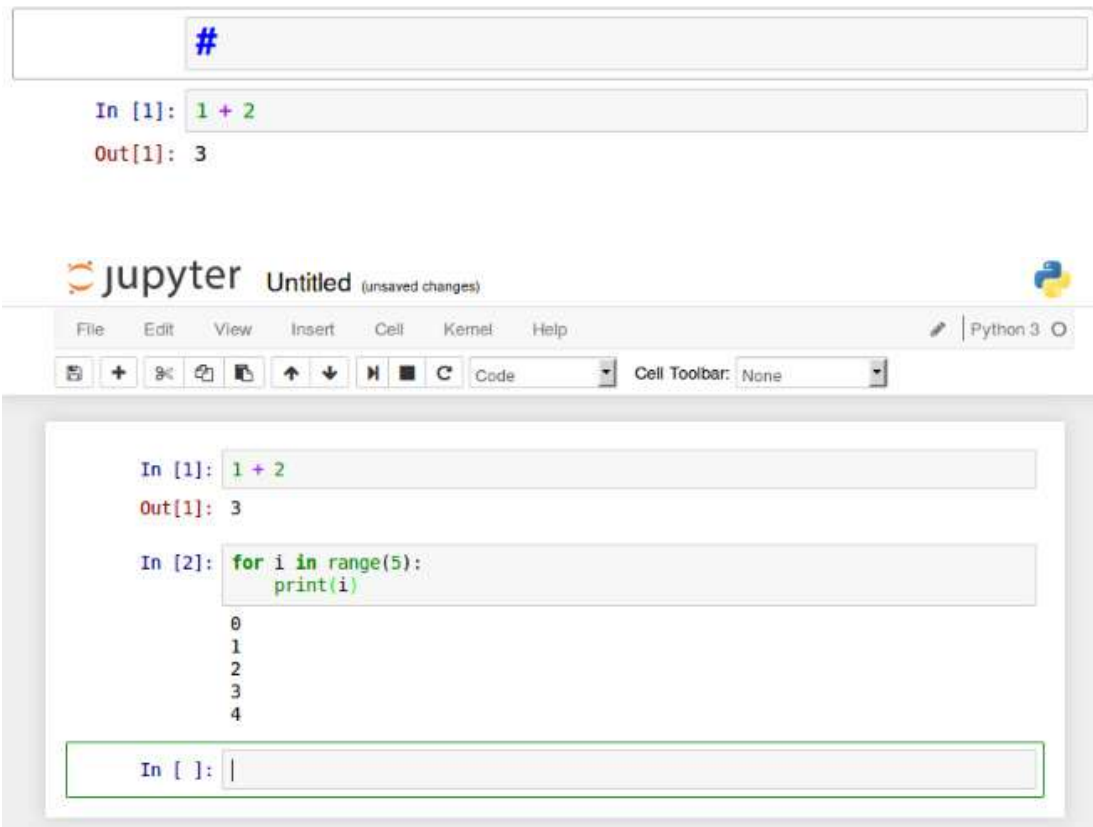
```
pip3 install jupyter
```

你可以在这里查看官方的 Jupyter 安装文档

<https://jupyter.readthedocs.io/en/latest/install.html>

## 2.2 Running Code





## My first title in Jupyter

### A very simple operation

Let's **add** two numbers:

```
In [1]: 1 + 2
Out[1]: 3
```

### Counter

Let's *count* from 0 to 4:

```
In [2]: for i in range(5):
        print(i)

0
1
2
3
4
```

```
In [ ]:
```


### 3. Github 中搭建自己的主页

(1) 打开 Github 首页，登陆后新建一个仓库，这里再次提醒要注意仓库的名称，比如我的帐号是 stranieroshitie，那么仓库名称应该是：stranieroshitie.github.io。

## Create a new repository


A repository contains all project files, including the revision history.

Owner

 stranieroshitie ▾

/


Repository name \*

stranieroshitie.github.io 


Great repository names are short, simple, and unique.

The repository stranieroshitie.github.io already exists on this account

Description (optional)

☒  **Public**

Anyone can see this repository. You choose who can commit.


☐  **Private**

You choose who can see and commit to this repository.

☒ **Initialize this repository with a README**

This will let you immediately clone the repository to your computer. Skip this step if you're importing an existing repository.

Add .gitignore: None ▾

Add a license: None ▾ 

Create repository

(2) 新建一个 index.html，这个其实就是个人主页中的“结构框架”，这个文件中选择不同的代码，可以设置不同的外观结构。本次的展示，我是直接 follow 课件中 JONATHAN MCGLONE 的 Creating and Hosting a Personal Site on GitHub。

注：其他方式 1：跳转到新建库界面,然后选择 Settings，然后点击 Choose a theme 选择一个博客主题。

## GitHub Pages

GitHub Pages is designed to host your personal, organization, or project pages from a GitHub repository.

✓ Your site is published at <https://stranieroshitie.github.io/>

### Source

Your GitHub Pages site is currently being built from the `master` branch. [Learn more.](#)

User pages must be built from the `master` branch.

### Theme Chooser

Select a theme to publish your site with a Jekyll theme. [Learn more.](#)

Choose a theme

### Custom domain

Custom domains allow you to serve your site from a domain other than `stranieroshitie.github.io`. [Learn more.](#)

Save

其他方式 2：如果你认为某人的 github 主页很漂亮，进入他 github 库网页，下载他的 code，找到其中的 `index.html`，然后将其内容复制到你的文件中。

> 0 releases    1 environment    1 contributor

Create new file   Upload files   Find File   Clone or download ▾

Clone with HTTPS ⓘ    Use SSH

Use Git or checkout with SVN using the web URL.

<https://github.com/stranieroshitie/stranieroshitie>

Open in Desktop    Download ZIP

10 hours ago

10 hours ago

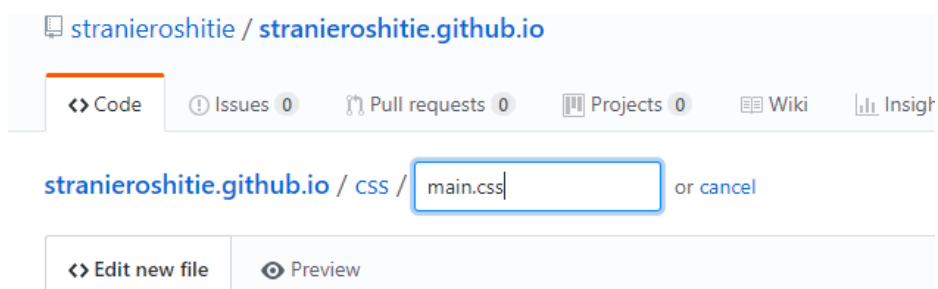
回到 Jonathan McGlone 给出的 index.html

```
<html>
  <head>
    <title>Hank Quinlan, Horrible Cop</title>
  </head>
  <body>
    <nav>
      <ul>
        <li><a href="/">Home</a></li>
        <li><a href="/about">About</a></li>
        <li><a href="/cv">CV</a></li>
        <li><a href="/blog">Blog</a></li>
      </ul>
    </nav>
    <div class="container">
      <div class="blurb">
        <h1>Hi there, I'm Hank Quinlan!</h1>
        <p>I'm best known as the horrible cop from <em>A
        Touch of Evil</em> Don't trust me. <a href="/about">Read more about my life...</a></p>
      </div><!-- /.blurb -->
    </div><!-- /.container -->
    <footer>
      <ul>
        <li><a href="mailto:hankquinlanhub@gmail.com">email</a>
        </li>
        <li><a href="https://github.com/hankquinlan">github.com/hankquinlan</a></li>
      </ul>
    </footer>
  </body>
</html>
```

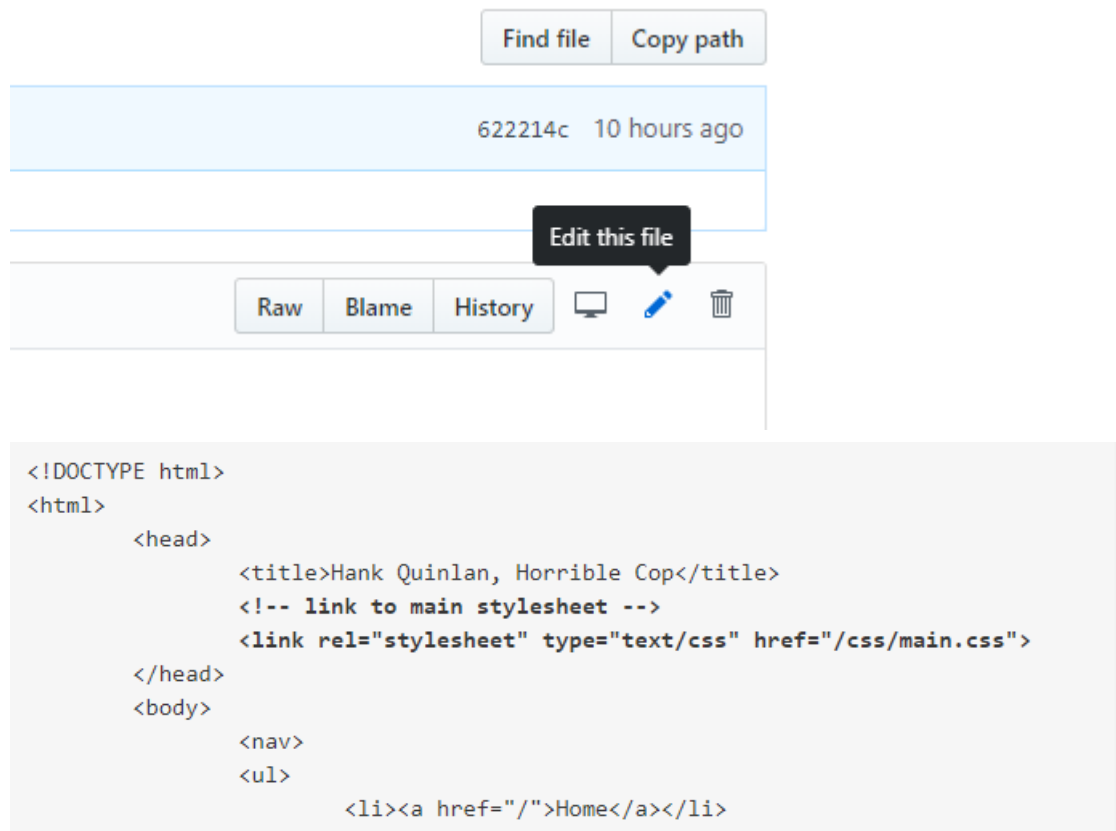
图中红色标注部分，改成自己的名称即可。而第二个红框选中部分为主页的四个主要菜单“home”“about”“CV”“blog”，大家可以按照自己的喜好选择不同的主题。

Commit 之后，你就已经建立起 github 个人主页了：http://xxxxxxx.github.io.

(3) 返回库，新建一个 `css/main.css` 文件，代码不变，直接 copy 即可（主要是版式的一些优化）。



(4) 在 index 中增加一个链接至将 main.css 文件。回到 index.html 选择 "编辑" 按钮。



原文中关于设置 jekyll 的步骤从略

(5) 发布 blog（或者作业等各类需要的“展示”）

新建一个 \_post/ 文件夹，这个文件夹用来存储每次的发布的“展示”

在 \_layout 中新建一个文件 post.html 用来设置 \_post 中的一些布局。

```
---
layout: default
---
<h1>{{ page.title }}</h1>
<p class="meta">{{ page.date | date_to_string }}</p>

<div class="post">
  {{ content }}
</div>
```

(6) 现在可以来发布一个“post”了

在 \_post/ 文件夹中新建一个 .md 文件，如下图所示 2019-03-14-test2.md

stranieroshitie.github.io / \_posts / 2019-03-14-test2.md  or cancel

```
<> Edit file    Preview changes

1  ---
2  layout: post
3  title: "HW1 Submission"
4  date: 2019-03-14
5  ---
6  why my own pdf document can't be published.
7  https://stranieroshitie.github.io/\_posts/EventStudy.pdf
8
```

到这一步，已经可以在网页中展示了

<http://stranieroshitie.github.io/2019/03/14/test2>

(7) 库中新建一个 `blog` 文件夹，并在其中新建一个 `index.htm` 文件，将所有 post“链接”在 `blog` 的目录下。

```
---
layout: default
title: Hank Quinlan's Blog
---

<h1>{{ page.title }}</h1>
<ul class="posts">

    {% for post in site.posts %}
        <li><span>{{ post.date | date_to_string }}</span> » <a href="{{ post.url }}" title="{{ post.title }}">{{ post.title }}</a></li>
    {% endfor %}
</ul>
```

(8) 按照原文指导，编辑 `_config.yml`，在 `blog/` 中新建 `atom.xml`，即可结束（如果需要继续美化网页，可以找相应的代码，将其 copy 在相应的文件中）。

最终效果

[Home](#) [About](#) [CV](#) [Blog](#)

# Hi there, I'm Shi Tie!

I'm best known as the lo straniero from *A Touch of Evil* Don about my life...

---

email [github.com/stranieroshitie](https://github.com/stranieroshitie)

[Home](#) [About](#) [CV](#) [Blog](#)

## Shi Tie's Blog

17 Mar 2019 » [TA Review Session 2](#)

15 Mar 2019 » [Shi Tie, Lo Straniero, Launches Site](#)

---

email [github.com/stranieroshitie](https://github.com/stranieroshitie)

[Home](#) [About](#) [CV](#) [Blog](#)

## TA Review Session 2

17 Mar 2019


For the 2nd review session document, see the following link.

[click here](#)

---

email [github.com/stranieroshitie](https://github.com/stranieroshitie)



stranieroshitie.github.io / \_posts / 2019-03-17-TA-Review-Si  or cancel

<> Edit file

Preview changes

```
1 ---
2 layout: post
3 title: "TA Review Session 2"
4 date: 2019-03-17
5 ---
6
7 For the 2nd review session document, see the following link.
8
9 \[click here\]{{(site.baseurl)}/TA Review Session/TA_Review_Session_2.pdf)
10
```

当你要提交作业或其他“展示”，只要在\_post/文件夹中新建一个新的.md 文档，编辑好其中的内容，然后 commit，它就会出现在个人主页 blog 目录下，

其链接：<https://stranieroshitie.github.io/blog/2019/03/17/TA-Review-Session-2>

直接打开，就是你所要 post 的内容，各位提交作业时，即可提交类似的链接。

## 4. 机器学习常用数据库介绍

### 4.1 UCI <http://archive.ics.uci.edu/ml/index.php>

UCI 数据库是加州大学欧文分校(University of CaliforniaIrvine)提出的用于机器学习的数据库, 这个数据库目前共有 468 个数据集, 其数目还在不断增加, UCI 数据集是一个常用的标准测试数据集。

每个数据文件(\*.data) 包含以“属性-值”对形式描述的很多个体样本的记录。对应的\*.info 文件包含的大量的文档资料。(有些文件\_generate\_databases; 他们不包含\*.data 文件。) 作为数据集和领域知识的补充, 在 utilities 目录里包含了一些在使用这一数据集时的有用资料。



UCI Machine Learning Repository

Welcome to the UC Irvine Machine Learning Repository!

We currently maintain 468 data sets as a service to the machine learning community. You may [view all data sets](#) through our searchable interface. For a general overview of the Repository, please visit our [about page](#). For information about citing data sets in publications, please read our [citation policy](#). For any other questions, feel free to [contact the Repository maintainers](#).

Supported By: In Collaboration With: [Read Info](#)

**Latest News:**

- 08.24.2018: Welcome to the new Repository admins Dhruv Dax and ER Karte Tarekhdoy!
- 04.04.2018: Welcome to the new Repository admins Kevin Becker and Moab Lichner!
- 03.01.2018: Note from donor regarding Netflix data
- 10.16.2008: Ten new data sets have been added
- 09.14.2008: Several data sets have been added
- 03.24.2008: New data sets have been added
- 06.25.2007: Ten new data sets have been added: UCI Pen Characters, MADAG Gamma Telescope

**Featured Data Set:** [Laryngeal Cancer](#)

**Newest Data Sets:**

- 01.07.2018: [UCI EHR data for sepsis](#)
- 01.02.2018: [UCI Parking Birmingham](#)
- 12.19.2018: [UCI Thermal Reaction Ratings](#)
- 12.19.2018: [UCI Travel Reviews](#)

**Most Popular Data Sets (hits since 2007):**

- 2865403: [k9s](#)
- 1422222: [Adult](#)
- 1092238: [Sfnet](#)
- 935418: [Car Evaluation](#)

Browse Through: 468 Data Sets

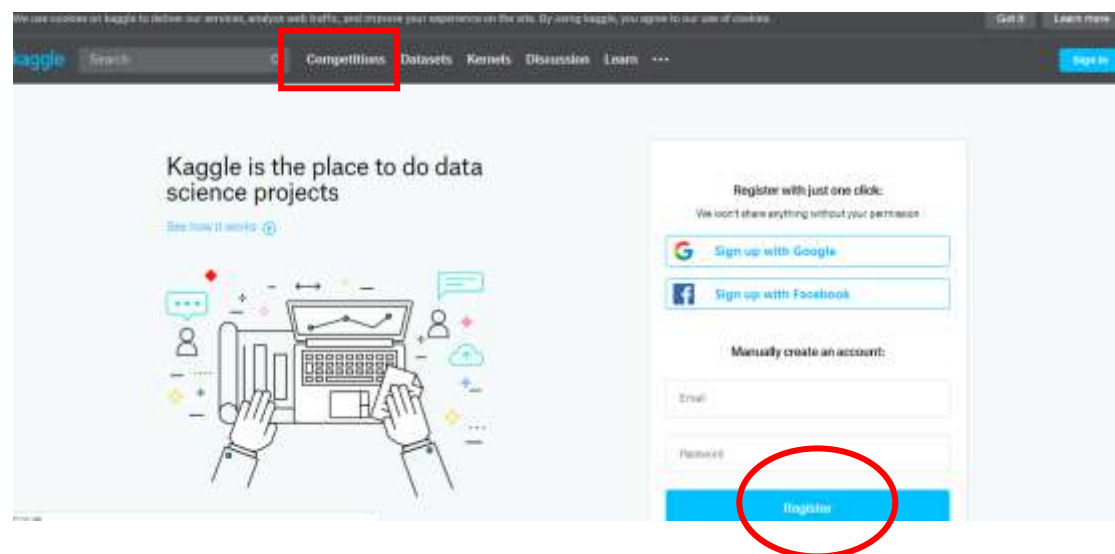
Table View: [List View](#)

| Default Task                         | Name                         | Data Types   | Default Task           | Attribute Types            | # Instances | # Attributes | Year |
|--------------------------------------|------------------------------|--------------|------------------------|----------------------------|-------------|--------------|------|
| <a href="#">Classification (345)</a> | Abalone                      | Multivariate | Classification         | Categorical, Integer, Real | 4177        | 8            | 1991 |
| <a href="#">Classification (79)</a>  | Adult                        | Multivariate | Classification         | Categorical, Integer       | 48842       | 14           | 1992 |
| <a href="#">Classification (84)</a>  | Ancestry                     | Multivariate | Classification         | Categorical, Integer, Real | 798         | 38           |      |
| <a href="#">Classification (105)</a> | Anonymous Microsoft Web Data |              | Recommendation-Systems | Categorical                | 37711       | 284          | 1998 |
| <a href="#">Classification (38)</a>  | Arrhythmia                   | Multivariate | Classification         | Categorical, Integer, Real | 452         | 278          | 1992 |
| <a href="#">Classification (398)</a> | Artificial Characters        | Multivariate | Classification         | Categorical, Integer, Real | 6000        | 7            | 1990 |
| <a href="#">Classification (105)</a> | Audiology (Otolaryngology)   | Multivariate | Classification         | Categorical                | 226         |              | 1981 |

#### 4.2 Kaggle <https://www.kaggle.com/>

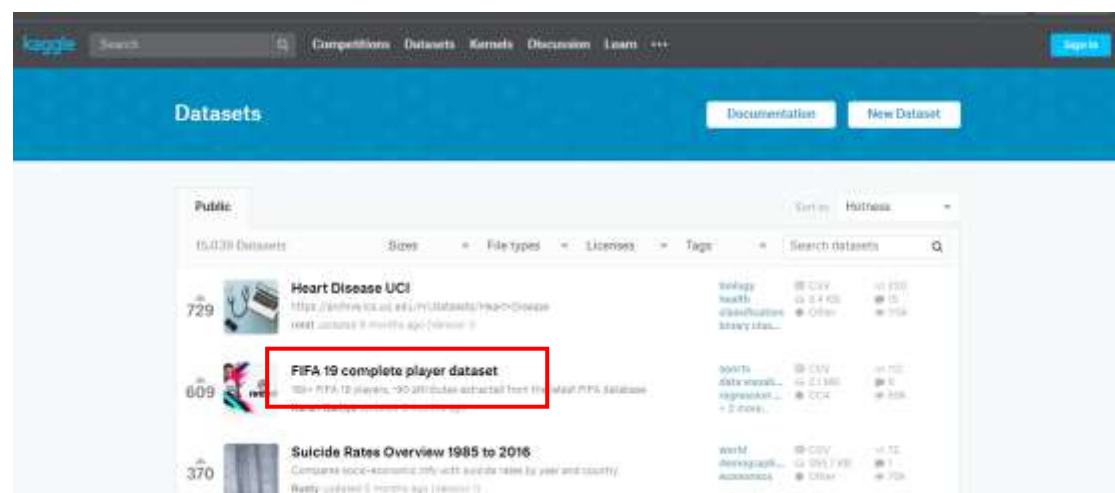
Kaggle 公司是由联合创始人兼首席执行官 AnthonyGoldbloom2010 年在墨尔本创立的，主要是为开发商和数据科学家提供举办机器学习竞赛、托管数据库、编写和分享代码的平台。

从公司角度讲，Kaggle 可以提供一些数据，进而提出一个实际需要解决的问题；从参赛者角度讲，他们组队参与项目，针对其中一个问题提出解决方案，最终由公司选出的最佳方案可以获得 5K-10K 美金的奖金。除此之外，Kaggle 官方每年还会举办一次大规模的竞赛，奖金高达一百万美元，吸引了广大数据科学爱好者参与其中。



Kaggle 可以分为 Competitions 竞赛、Datasets 数据集以及 Kernel 内核三个子平台、配套的 Forum 论坛模块以及供各类公司或组织招聘人才的 Jobs 模块。

Kaggle 从 2016 年 1 月开始上线了 Datasets 数据集服务，收集了许多公共的数据集，提供数据下载、介绍、相关脚本 Scripts 以及独立的论坛等服务。



The screenshot shows the Kaggle interface for a kernel. At the top, there are tabs for 'Data', 'Kernels (112)', 'Discussion (0)', and 'Activity'. Below these, there's a 'Data (2 MB)' section. On the left, under 'Data Sources', there's a file named 'data.csv' (10.2% k-BG). In the center, 'About this file' describes it as a 2019 FIFA player attributes dataset. On the right, 'Columns' lists fields like row number, ID, Name, Age, Photo, Nationality, Flag, Overall, and Potential. Below this is a table preview for 'data.csv (0.72 MB)' showing 20 of 80 columns. The table has columns: # (row number), ID (unique id), Name, Age, Photo (url), and Nationality. The first row shows values: 0, 17194, 17194, 18, 18207, and England.

| #          | ID                         | Name  | Age | Photo                     | Nationality |
|------------|----------------------------|-------|-----|---------------------------|-------------|
| row number | unique id for every player | name  | age | url to the player's photo | nationality |
| 0          | 17194                      | 17194 | 18  | 18207                     | England     |

Kernels 提供了数据分析所需的环境、数据集、代码和输出样式 (比如 Python Notebook), 将这些功能聚合在一起可以使得 Kernels 可以很方便的复现和分享。

The screenshot shows the 'Public' Kernels page on Kaggle. It features a list of kernels with their titles, authors, and creation times. Each kernel entry includes a 'View' icon, a 'Copy' icon, and a 'Py' icon. The kernels are sorted by 'Hotness'.

| Rank | Kernel Title                                  | Author   | Created | Views | Copy | Py | Other |
|------|-----------------------------------------------|----------|---------|-------|------|----|-------|
| 2    | Dream Team and Young Team                     | 15h ago  | 1       |       |      |    |       |
| 12   | FIFA in depth analysis with Linear Regression | 1d ago   | 7       |       |      |    |       |
| 0    | Classification of player positions            | 10h ago  | 0       |       |      |    |       |
| 59   | Clustering to Help Club Managers              | 15m ago  | 30      |       |      |    |       |
| 3    | FIFA19 analysis of players                    | 2d ago   | 0       |       |      |    |       |
| 23   | Data ScienceTutorial for Beginners            | 2hrs ago | 14      |       |      |    |       |

关于 Kaggle 详细的介绍

Kaggle 入门: [http://www.360doc.com/content/18/0106/16/44422250\\_719580875.shtml](http://www.360doc.com/content/18/0106/16/44422250_719580875.shtml)

Kaggle 数据建模分析与竞赛平台介绍: <https://www.jianshu.com/p/eb0b37500229>