数据挖掘分析综合实训指导手册

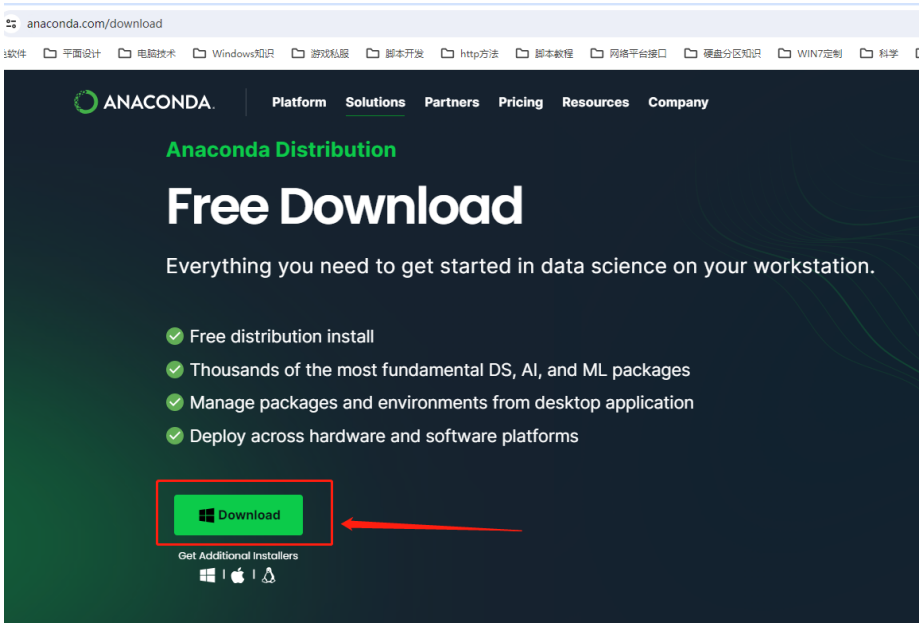
1.实训环境搭建及基本操作

1.1Anaconda 软件安装

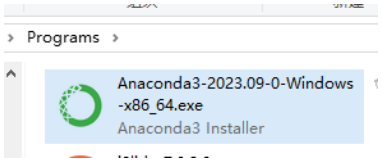
Anaconda 可以便捷获取包且对包能够进行管理，同时对环境可以统一管理的发行版本。Anaconda 包含了 conda、Python 在内的超过 180 个科学包及其依赖项。

官方网站下载地址：https://www.anaconda.com/download

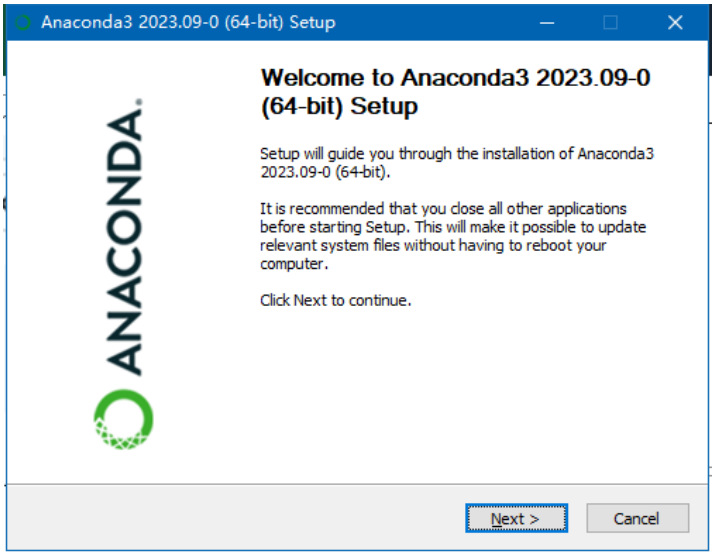
打开上面的网址，点击 download 按钮进行下载。



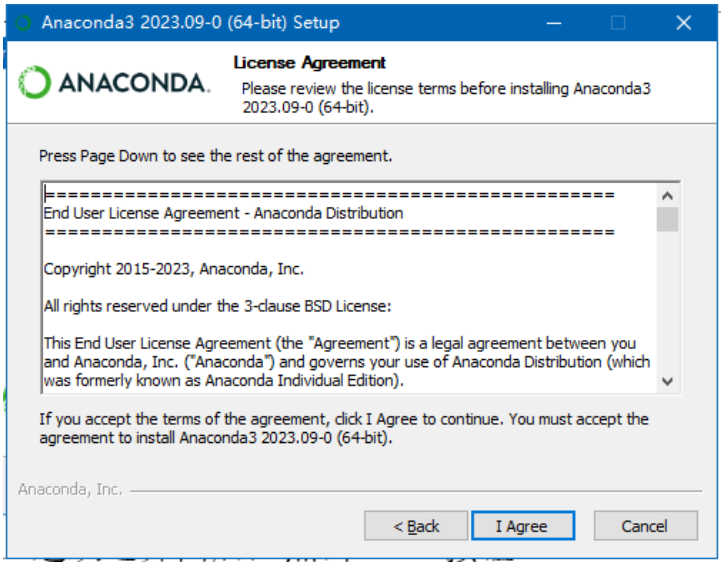
下载文件名为：Anaconda3-2023.09-0-Windows-x86_64.exe



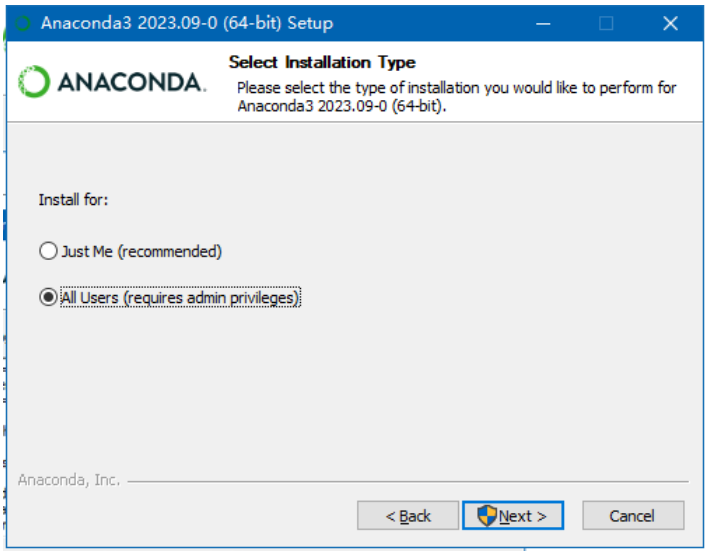
双击 Anaconda3-2023.09-0-Windows-x86_64.exe 文件进行安装。



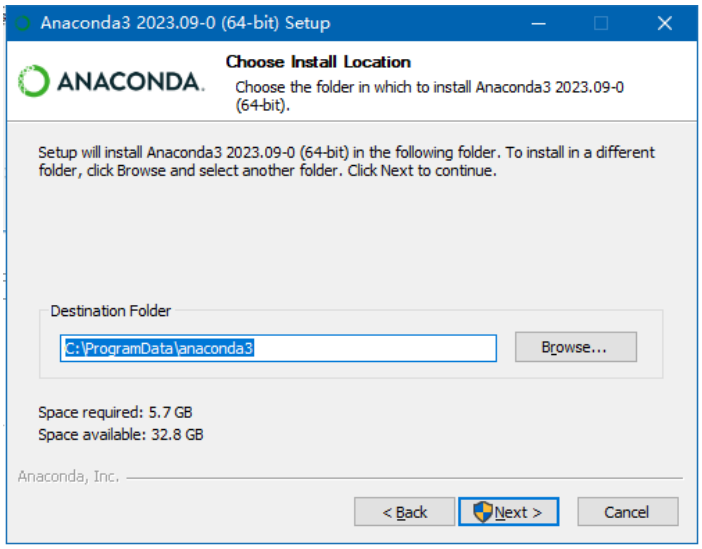
出现上述欢迎界面后，点击 Next 按钮。



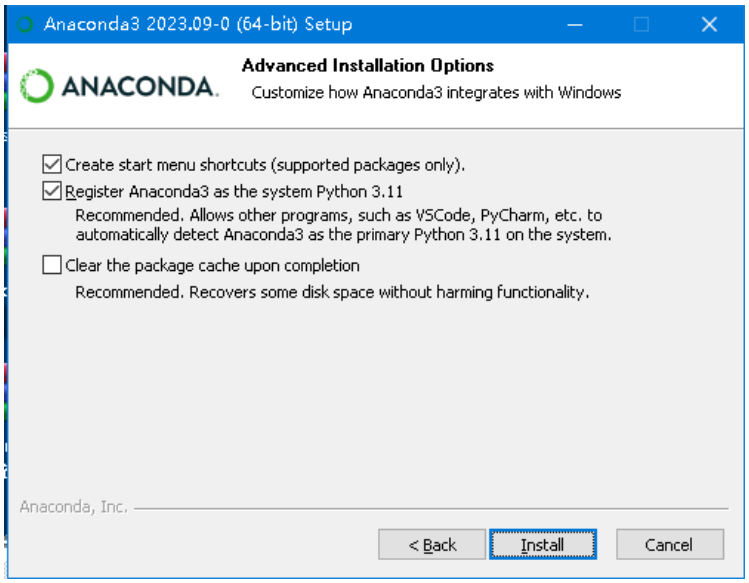
在上述使用协议界面，点击 I Agree 按钮。



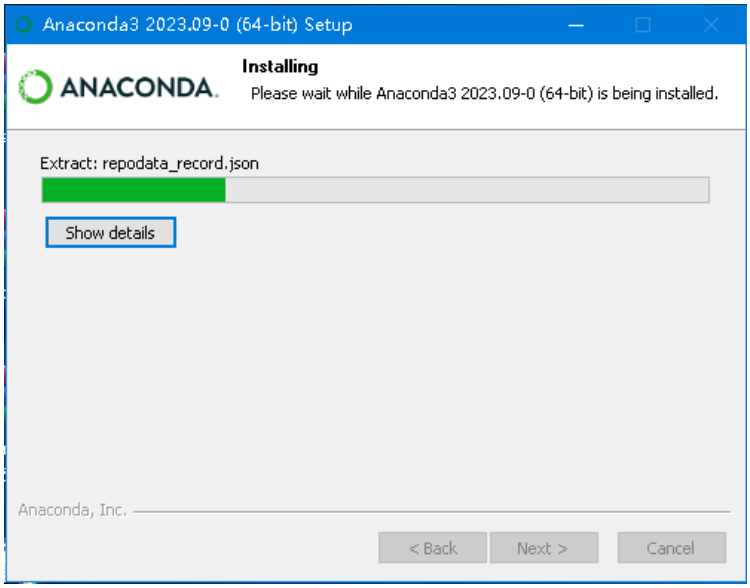
在上面的安装类型界面中，选择 All Users 单选框，点击 Next 按钮。



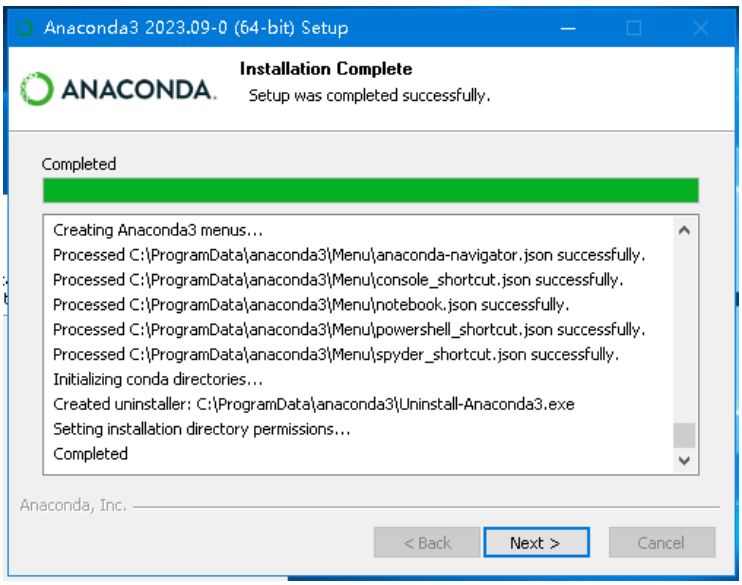
在上面的安装路径界面中，使用默认路径：C:\ProgramData\anaconda3，点击 Next 按钮。



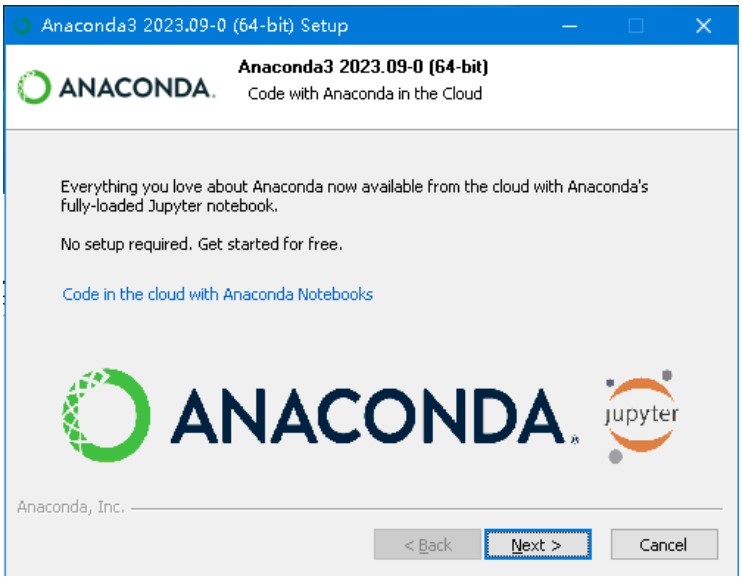
上面的高级安装选项界面中，保持默认选择项目，点击 Install 按钮。



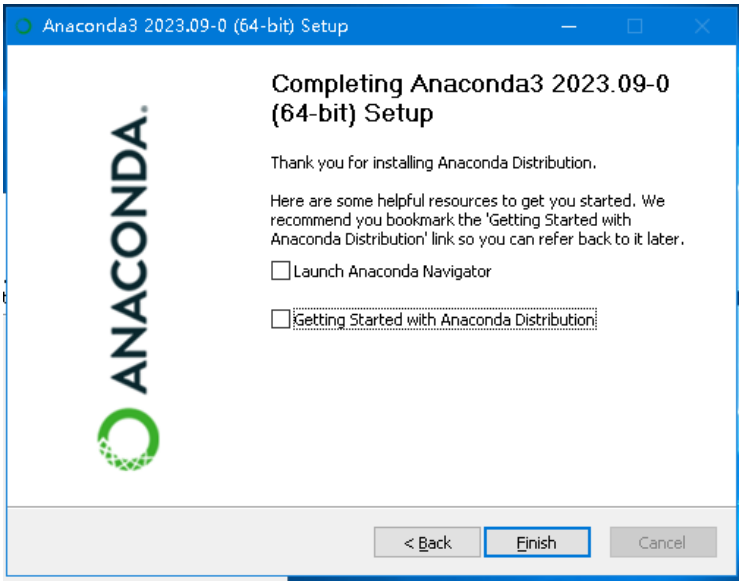
等待安装完成。



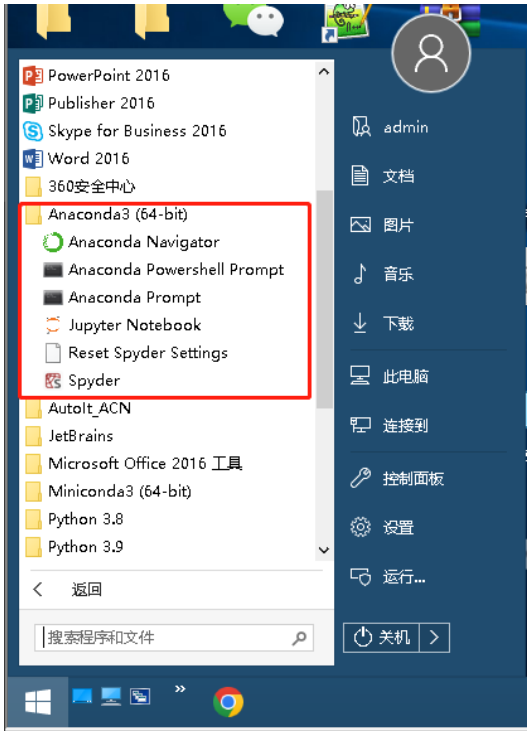
安装完成后，点击 Next 按钮。



继续点击 Next 按钮。



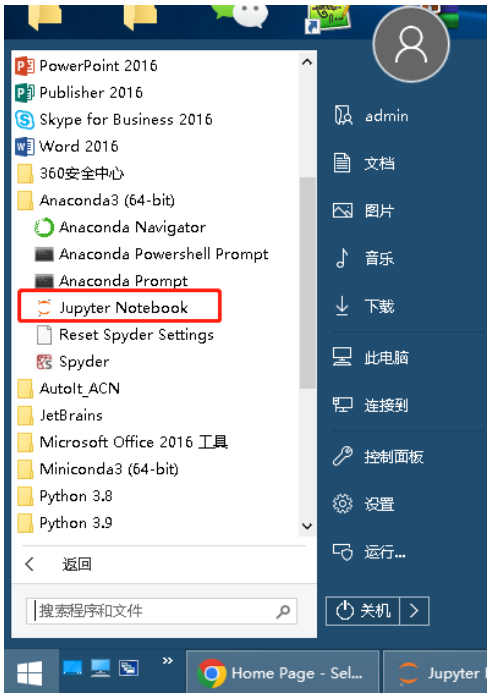
在上面的安装完成界面，取消两个复选框的选择，点击 Finish 按钮。
在开始菜单中，可以看到已经安装完成的 Anaconda。



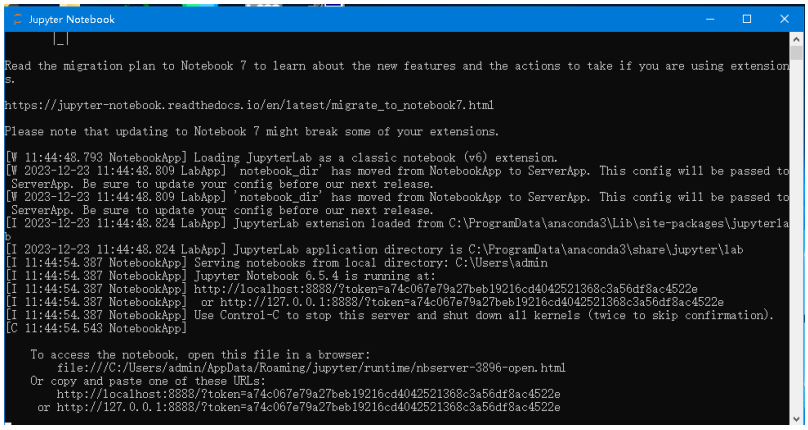
1.2Jupyter 软件使用

Jupyter notebook 是一种 Web 应用，能让用户将说明文本、数学方程、代码和可视化内容全部组合到一个易于共享的文档中。它可以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。也就是它可以能将代码、文档等这一切集中到一处，让用户一目了然。

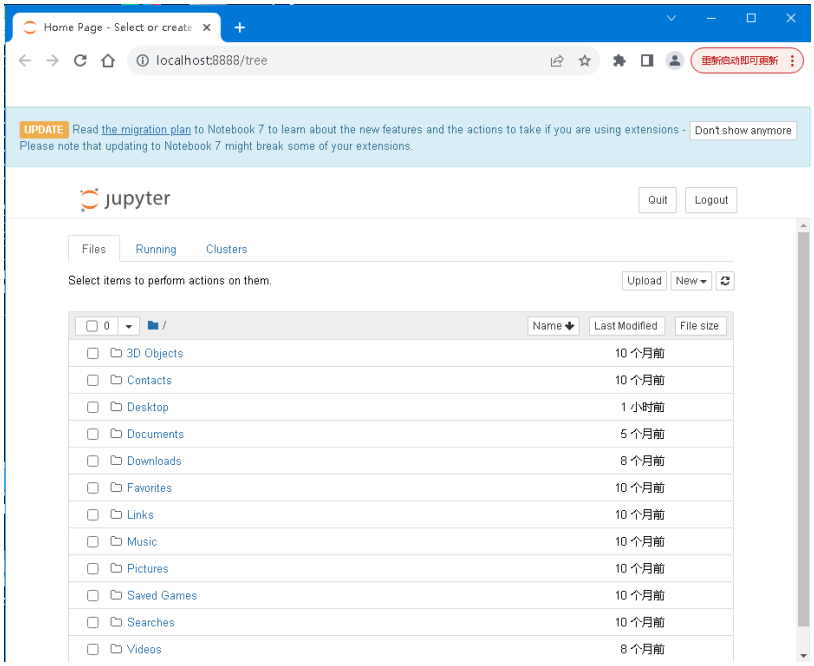
在开始菜单中，在 Anaconda3 目录下,点击 Jupyter Notebook，如下图:



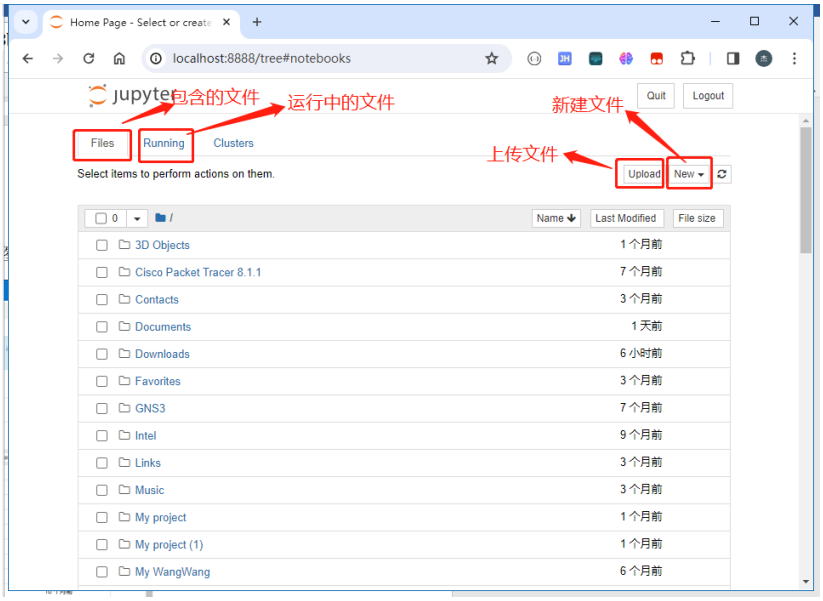
点击之后，会启动 Jypyter 命令窗，如下图：



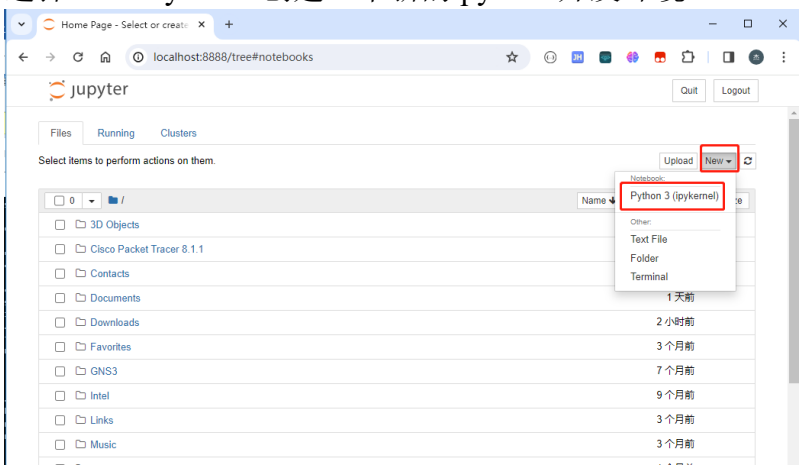
然后会启动 jupyter 的 web 窗口，如下图：



其中，Files 是 jupyter 工作目录下包含的文件，Running 是运行中的文件，Upload 是上传文件，New 是新建文件。



选择 New-Python3 创建一个新的 python 开发环境

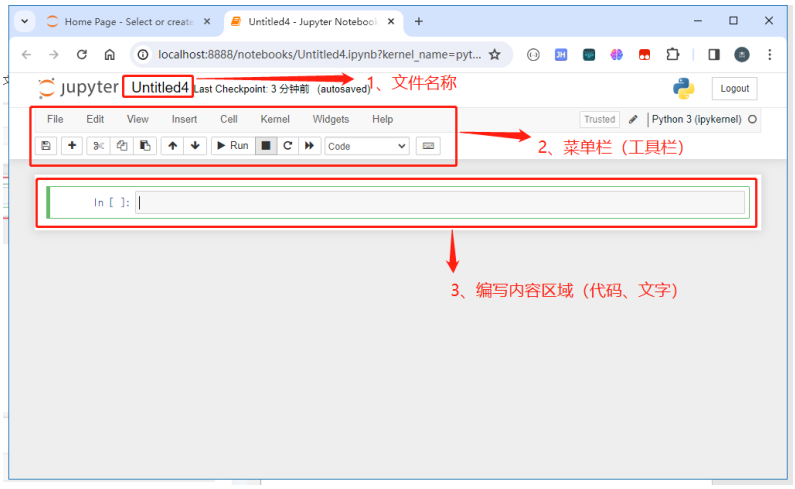


下图是新创建的 python 开发环境,Notebook 主要包含三个区域：

1、文件名

2、菜单栏（工具栏）

3、内容编辑

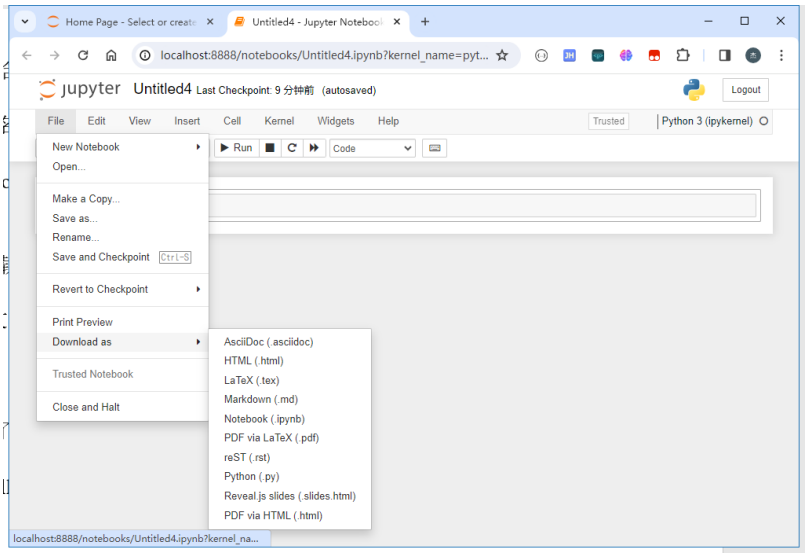


修改文件名:点击文件名，可以重命名当前 Notebook 的文件名。

熟悉菜单栏，这里介绍下常用的几个菜单栏的作用。

（1）File

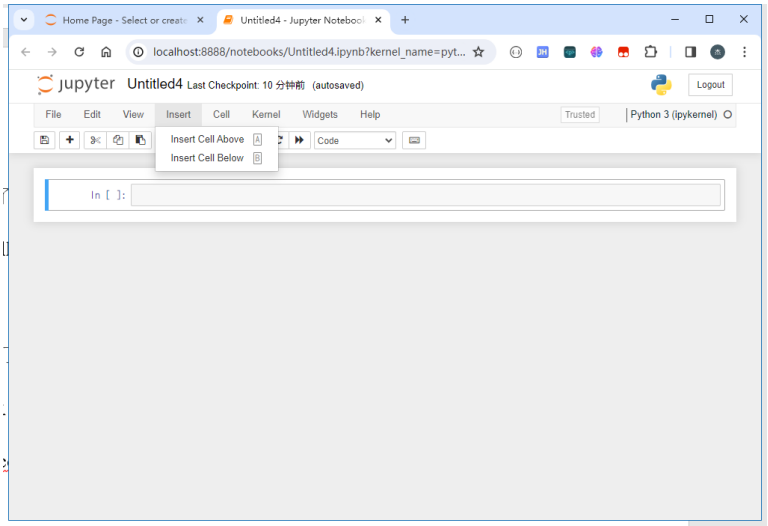
File 菜单中主要包含了以下功能：创建新的 Notebook、打开新的界面、拷贝当前 Notebook、重命名 Notebook、保存还原点、恢复到指定还原点、查看 Notebook 预览、下载 Notebook、关闭 Notebook。



这里重点强调下下载 Notebook 选项，它可以将当前 Notebook 转为 py 文件、html 文件、markdown 文件、rest 文件、latex 文件、pdf 文件。

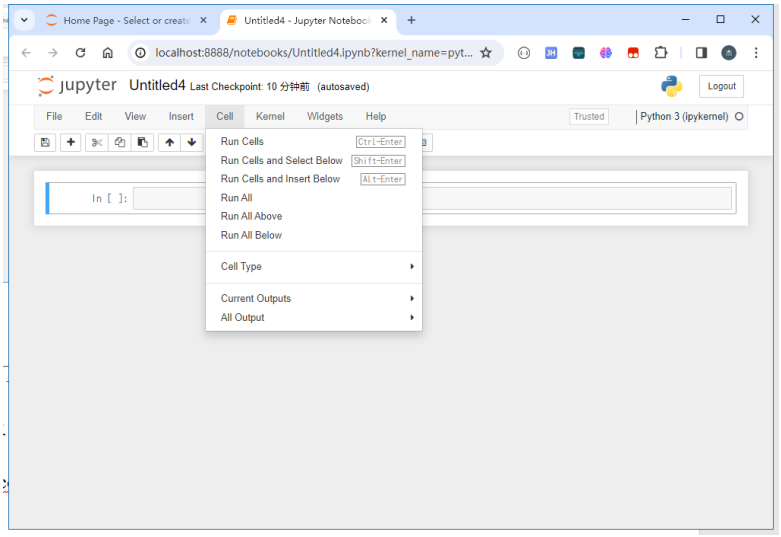
（2）Insert

Insert 菜单中包含了在当前位置之下插入一个新的 cell（单元格）、在当前位置之上插入一个新的 cell（单元格）。



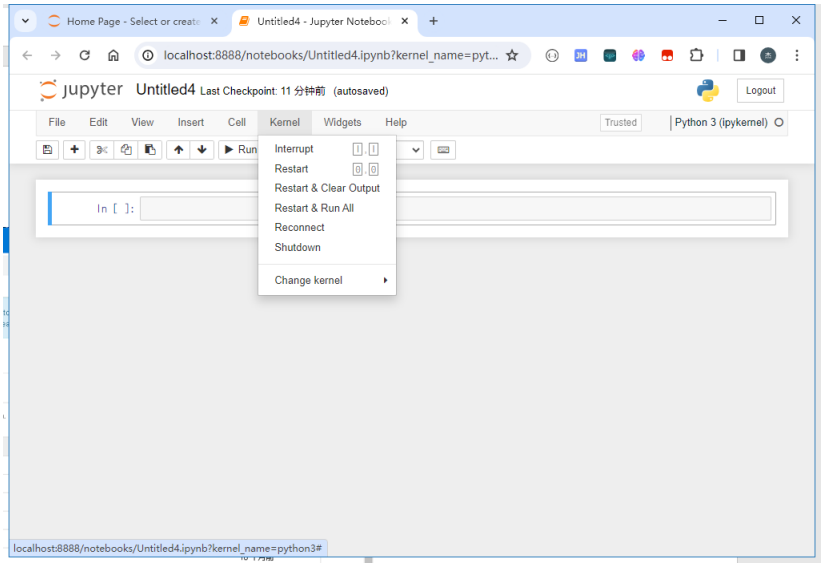
（3）Cell

Cell 菜单主要包含了运行 cells、运行 cells 后并在之后插入新的 cell、运行所有 cells、运行当前之上的所有 cell、运行当前之下的所有 cell、改变 cell 类型（code、markdown、raw nbconvert）等。

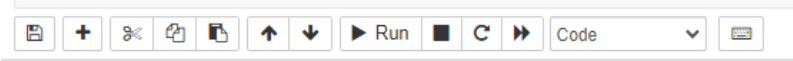


（4）Kernel

Kernel 菜单主要包含了中断 kernel、重启 kernel、重启 kernel 并清除输出、重启 kernel 并运行所有 cell、重连 kernel、关闭 kernel、改变 kernel 类型。

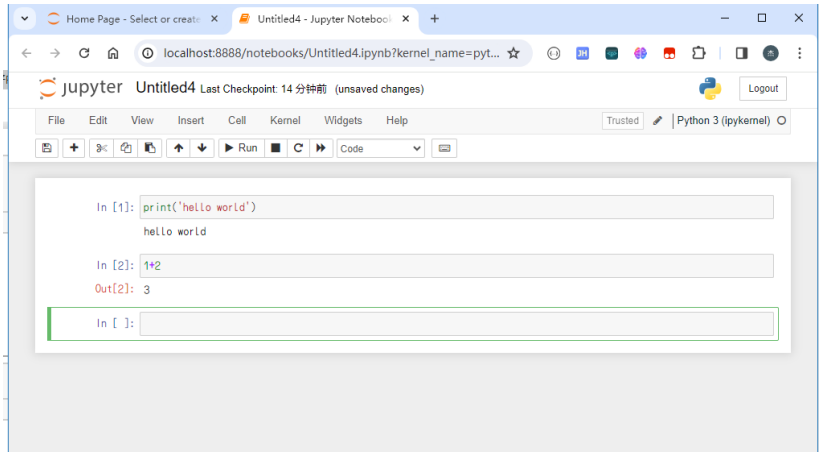


熟悉工具栏，工具栏上的内容都在下图中：很明显，工具栏中的功能大多都是菜单栏中的一部分功能的体现，主要是为了方便寻找。

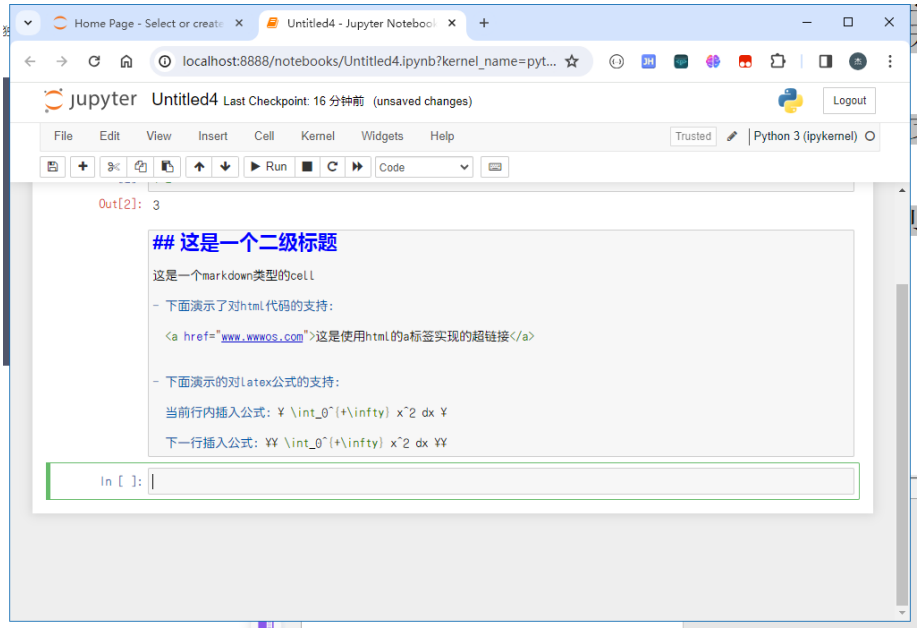


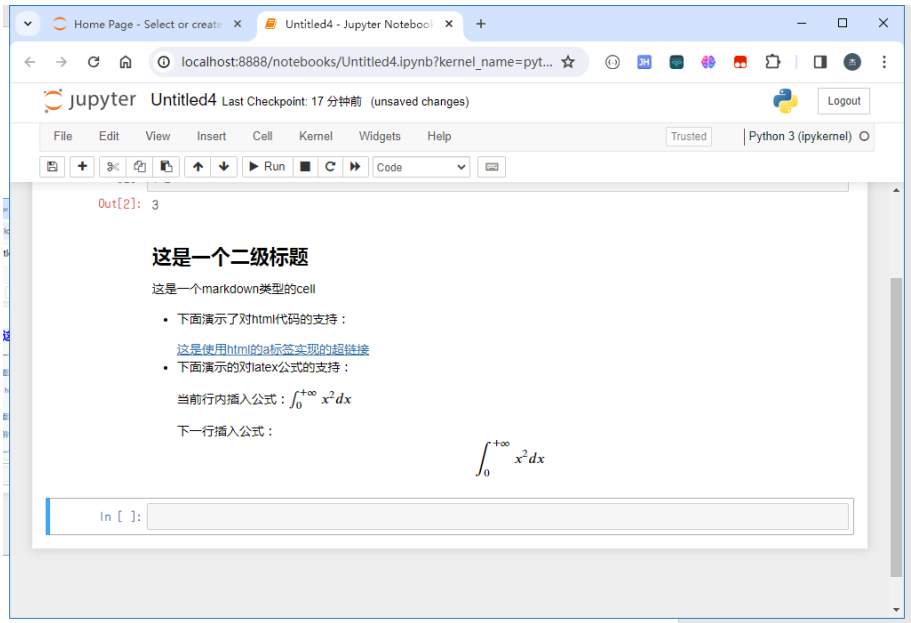
运行 Python 代码

想要运行 Python 代码，其实很简单，因为 Python 代码最后都在 Cell 中编写的。首先在 cell 中编写好 Python 代码，然后点击运行，可以直接在下面看到结果。



编写 Markdown，Notebook 最友好的一个功能就是可以在 cell 中通过 Markdown 来编写文本。我们首先创建一个 cell，然后更改类型为 markdown，更改成功后，cell 开头没有“In[?]:”的提示符。然后点击 cell，按照 markdown 语法来输入文本。





2. 电商交易数据的分析

2.1 明确需求和目标

对电商交易数据进行分析，包括商品价格、销量、销售额、地区、销售渠道、下单时间、利润等具体情况分析。根据分析结果，给出相应的销售策略。

2.2 数据选择

数据来自于网络，下载地址：https://wwjd.lanzout.com/iarOr2i7wxud　密码 2024

2.3 数据清洗

- 2.3.1 重复值的处理
- 2.3.2 异常值的处理
- 2.3.2 空值的处理
- 2.3.2 综合处理

2.4 数据分析

- 2.4.1 查看数据总体情况
- 2.4.2 价格分析
- 2.4.3 商品销量和销售额分析
- 2.4.4 城市的分析
- 2.4.5 渠道分析
- 2.4.6 下单时间分析
- 2.4.7 利润分析

2.5 总结

通过对数据分析，给出合理化的销售建议和营销策略建议。

3. 超市零售数据分析

3.1 明确需求和目标

对超市四年（2019-2022）的销售数据进行“人、货、场”分析，并给出提升销量的针对性建议。

场：整体运营情况分析，包括销售额、销量、利润、客单价、市场布局等具体情况分析。

货：商品结构、优势/畅销商品、劣势/待优化商品等情况分析。
人：客户数量、新老客户、RFM 模型、复购率、回购率等用户行为分析。
3.2 数据选择

从 Kaggle 平台，

https://www.kaggle.com/datasets/apoorvaappz/global-super-store-dataset 下载数据集

Global Super Store Dataset 数据集，解压后得到 Global_Superstore.csv 文件。

共 51290 条数据记录，每条记录共 24 个特征。
3.3 数据预处理
3.3.1 数据类型的处理
3.3.2 缺失值处理
3.3.3 重复值处理
3.4 数据分析
3.4.1 整体销售情况
3.4.1.1 销售额分析
3.4.1.2 销量分析
3.4.1.3 利润分析
3.4.1.4 客单价分析
3.4.1.5 市场占有率分析
3.4.2 商品情况
3.4.2.1 畅销商品分析
3.4.2.2 商品情况分析
3.4.3 用户情况
3.4.3.1 不同类型的用户占比分析
3.4.3.2 客户下单行为分析
3.4.3.3 RFM 模型分析
3.4.3.4 用户价值分析
3.5 总结

通过“场、货、人”三个不同的角度去分析了一家全球超市的销售、商品、用户情况，并根据分析结果给出一些利于拓展用户、提高销量和利润的方法建议。

4. 心脏病数据集挖掘分析

4.1 明确需求和目标
通过对心脏病数据集进行挖掘，进行探索性分析，对已有心脏病数据进行分析创建一个二分类模型，把已知是否有心脏病的数据未到模型中去训练，让模型自己挖掘特征学习，得到能够识别新未知数据的能力。找出对心脏病影响的各种因素，做出有针对性的对疾病的预防手段。
4.2 数据选择

从 Kaggle 平台，https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

下载数据集 Heart Disease Dataset 数据集，解压后得到 heart.csv 文件。

共 1025 条数据记录，每条记录共 14 个特征。

4.3 数据预处理

4.3.1 对缺失值进行处理

4.3.2 数据类型的处理

4.3.3 数据字段的处理

4.3.4 数据内容的处理

4.4 数据分析

4.4.1 数据探索性分析及可视化

4.4.1.1 数据缺失值处理

4.4.1.2 特征两两相关性分析

4.4.1.3 单个特征统计分布分析

4.4.1.4 单列特征与标签的关系分析

4.4.2 数据预处理

4.4.2.1 字段名修改完整特征

4.4.2.2 将定类特征由整数编码转为实际对应的字符串

4.4.2.3 将离散的定类和定序特征列转为独热编码

4.4.3 数据分析

4.4.3.1 对性别特征进行分析并数据可视化

4.4.3.2 对地中海贫血症特征进行分析并数据可视化

4.4.3.3 对最大心率特征进行分析并数据可视化

4.4.3.4 特征两两交互影响分析

4.4.4 构建随机森林分类模型

4.4.4.1 划分训练集和测试集

4.4.4.2 构建随机森林分类模型，在训练集上训练

4.4.4.3 可视化随机森林中的一棵决策树

4.4.4.4 特征重要性分析

4.4.4.5 从测试集中筛选出未知样本

4.4.4.6 预测测试集上全部数据

4.4.4.7 可解释性分析，绘制 PDP 图和 ICE 图

4.4.4.8 特征之间交互关系分析

4.4.5 构建随机森林分类模型

4.4.5.1 计算测试集每个样本的每个特征对两类预测结果的 shap 值

4.4.5.2 测试集所有样本，预测为“不患病”和“患病”各自的平均概率

4.5.总结

通过个个特征进行分析、建立分析模型，用测试集去测试特格特征的数据，得出每个特征患心脏病的概率，从而达到预测具有某个特征时，是否具有患心脏病较高特征的概率，能及早预防疾病的发生。

4、实训报告

(1)总结实训过程中遇到的问题及解决方案

(2)按给定的格式要求，撰写实训报告