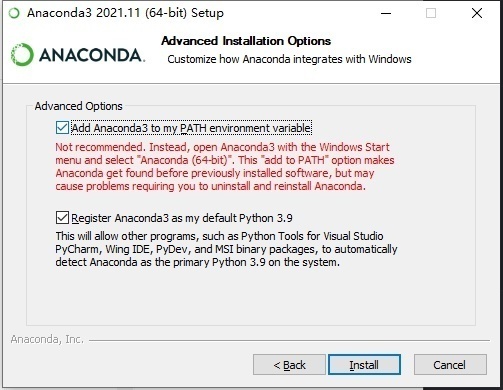
**Week 1：Python与Pandas基础**

**任务1**

搭建好基于Anaconda Individual Edition（或者Miniconda）和VSCode的Jupyter Notebook环境，更改镜像源，并且安装好本课程需要接触的几个相关依赖包

1. Anaconda3 (x64)下载地址：<https://mirrors.tuna.tsinghua.edu.cn/anaconda/archive/Anaconda3-2021.11-Windows-x86_64.exe>

安装的时候记得勾选Add Anaconda3 to my PATH environment variable：



1. VSCode (x64)下载地址：<https://code.visualstudio.com/sha/download?build=stable&os=win32-x64-user>

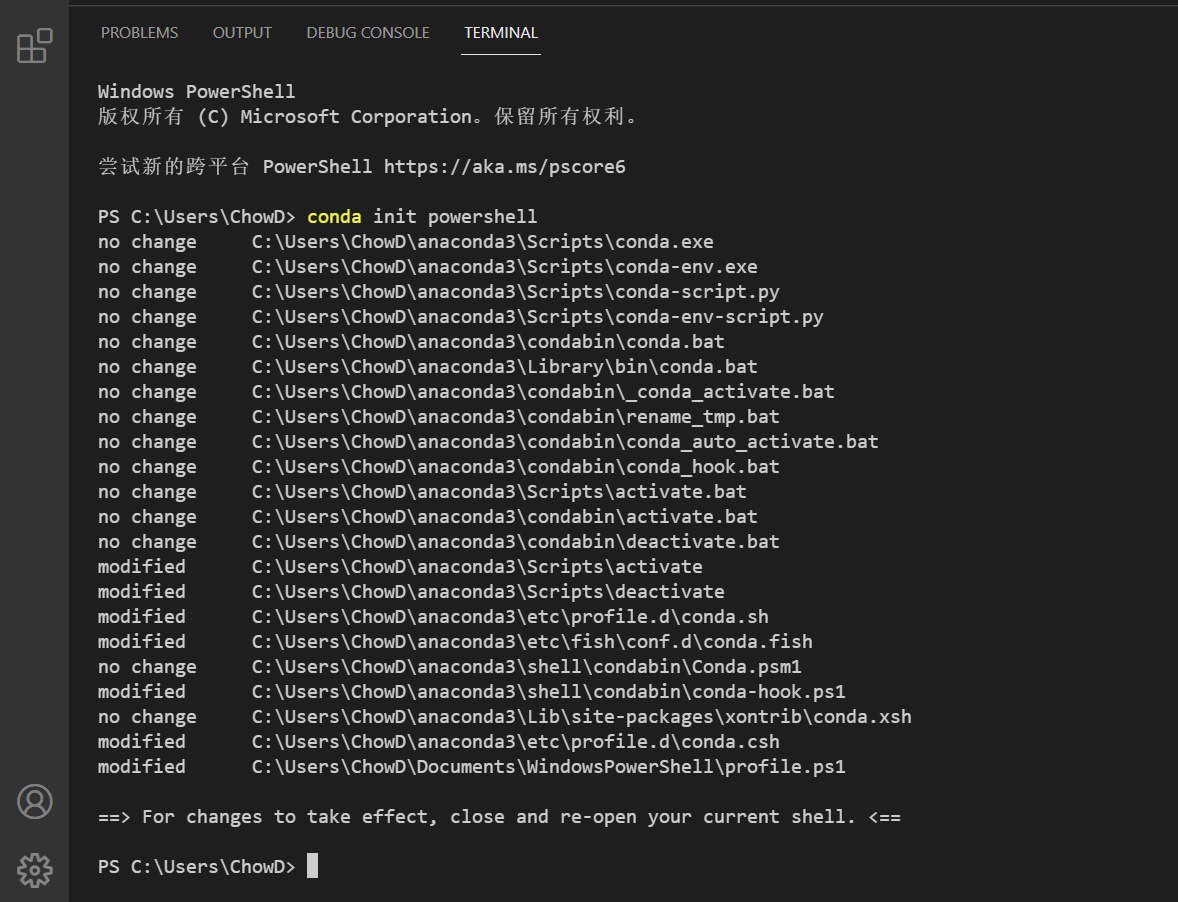
安装完成后尝试启动Code，学生可选安装中文插件包，首先在Extensions中安装Python：



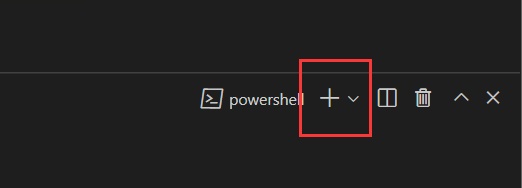
然后在Terminal - New Terminal中，尝试输入：

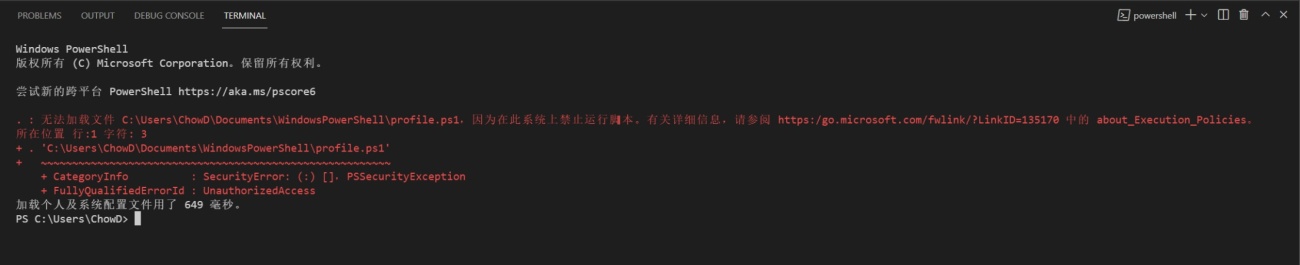
conda init powershell

没有出错的情况下，会显示以下字样：



通过右边的加号新建另一个Terminal，会发现终端窗口短暂等待后报红色的错误：

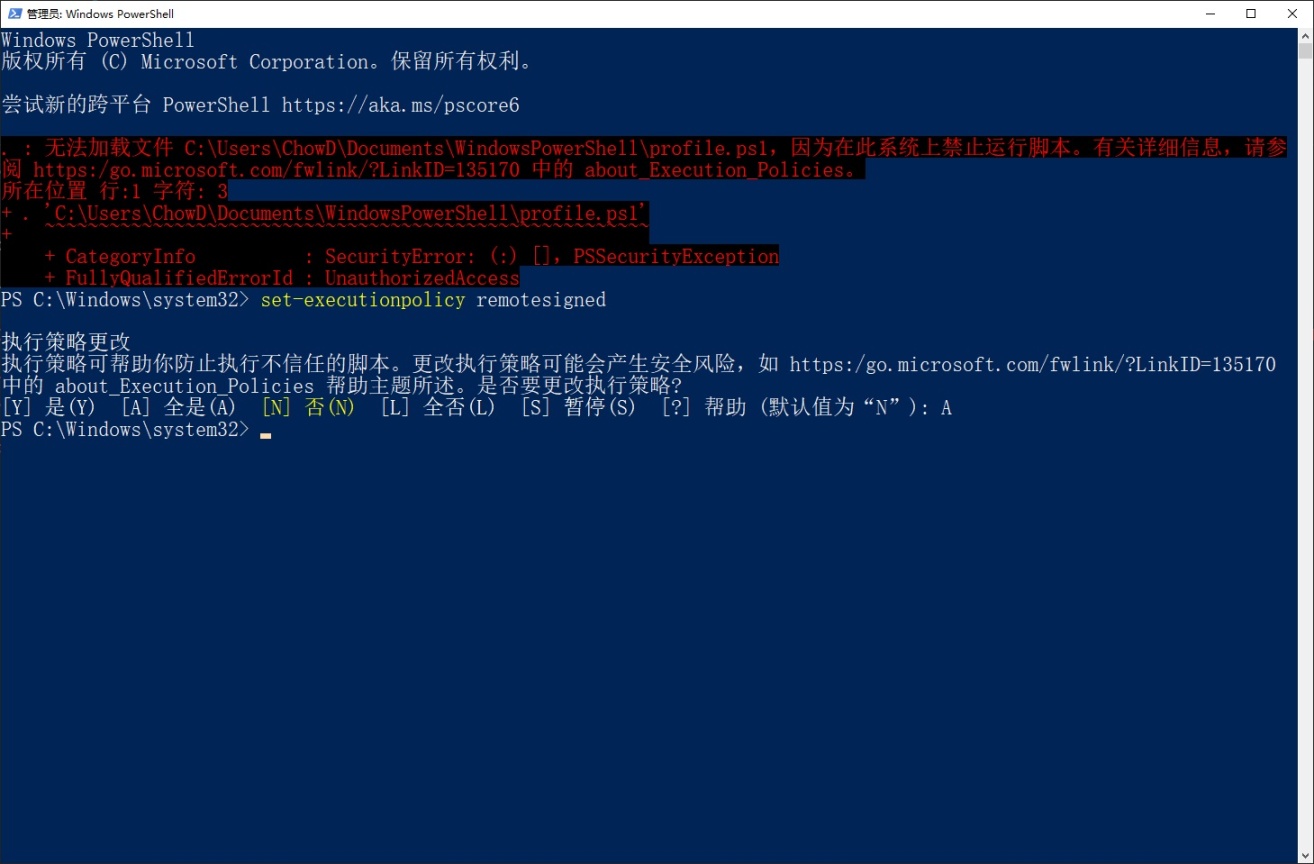




为了解决这个问题，我们需要在开始菜单右键 → Windows Powershell（管理员），并且输入

set-executionpolicy remotesigned

提示你确认信息后，输入A并且回车：



接下来，把这个窗口和在VSCode中打开的Terminal全部关掉（加号旁边的垃圾桶图标），再重新Terminal - New Terminal，成功的话应当会在每一行的行首出现(base)字样。



1. 更换软件源

Conda：<https://mirrors.tuna.tsinghua.edu.cn/help/anaconda/>

*附：在Windows上快速打开.condarc的技巧*

在VSCode的Terminal中输入以下指令

code .\.condarc

Pypi：<https://mirrors.bfsu.edu.cn/help/pypi/>

1. 新建一个开发环境

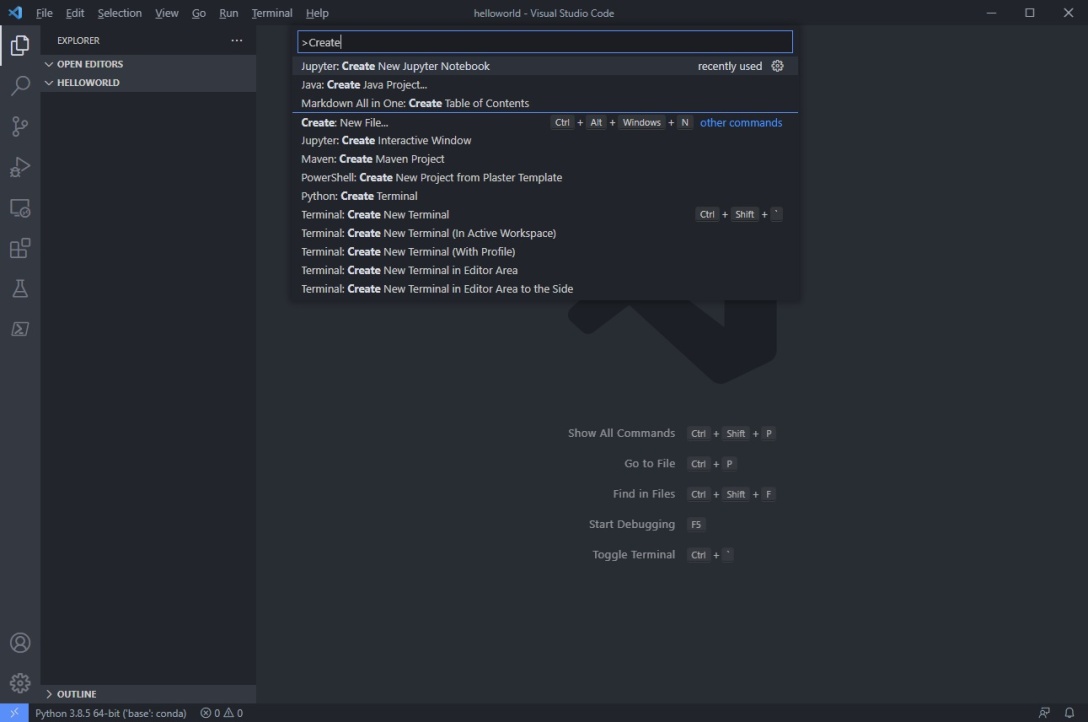
conda create --name dd python=3.7  
conda activate dd

1. 安装依赖包

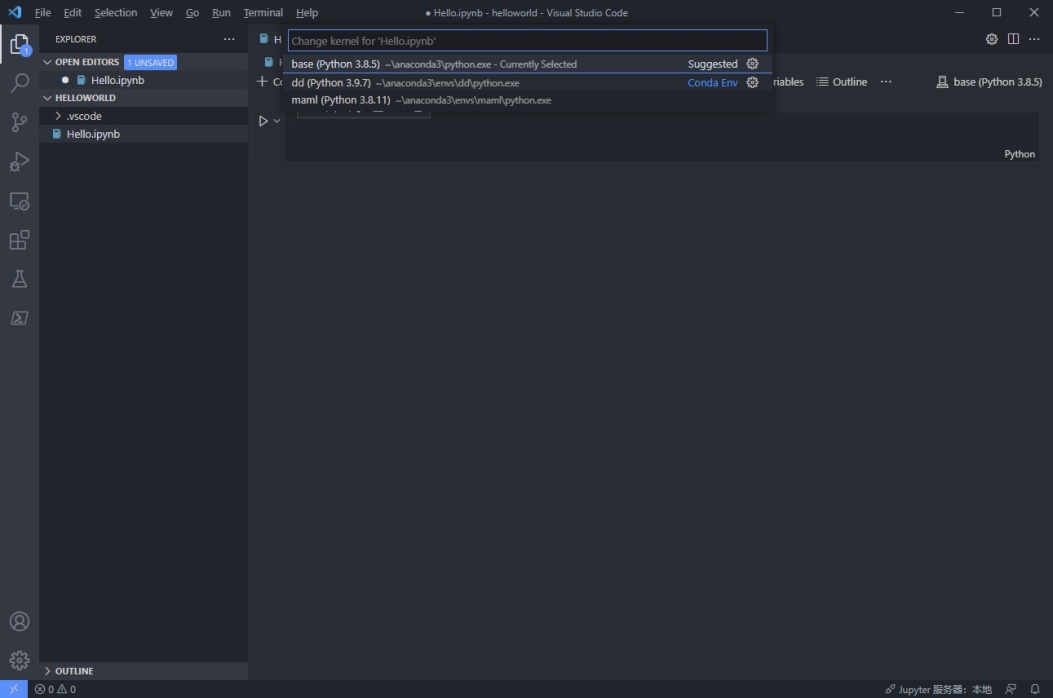
conda install numpy pandas matplotlib

1. 新建一个笔记本

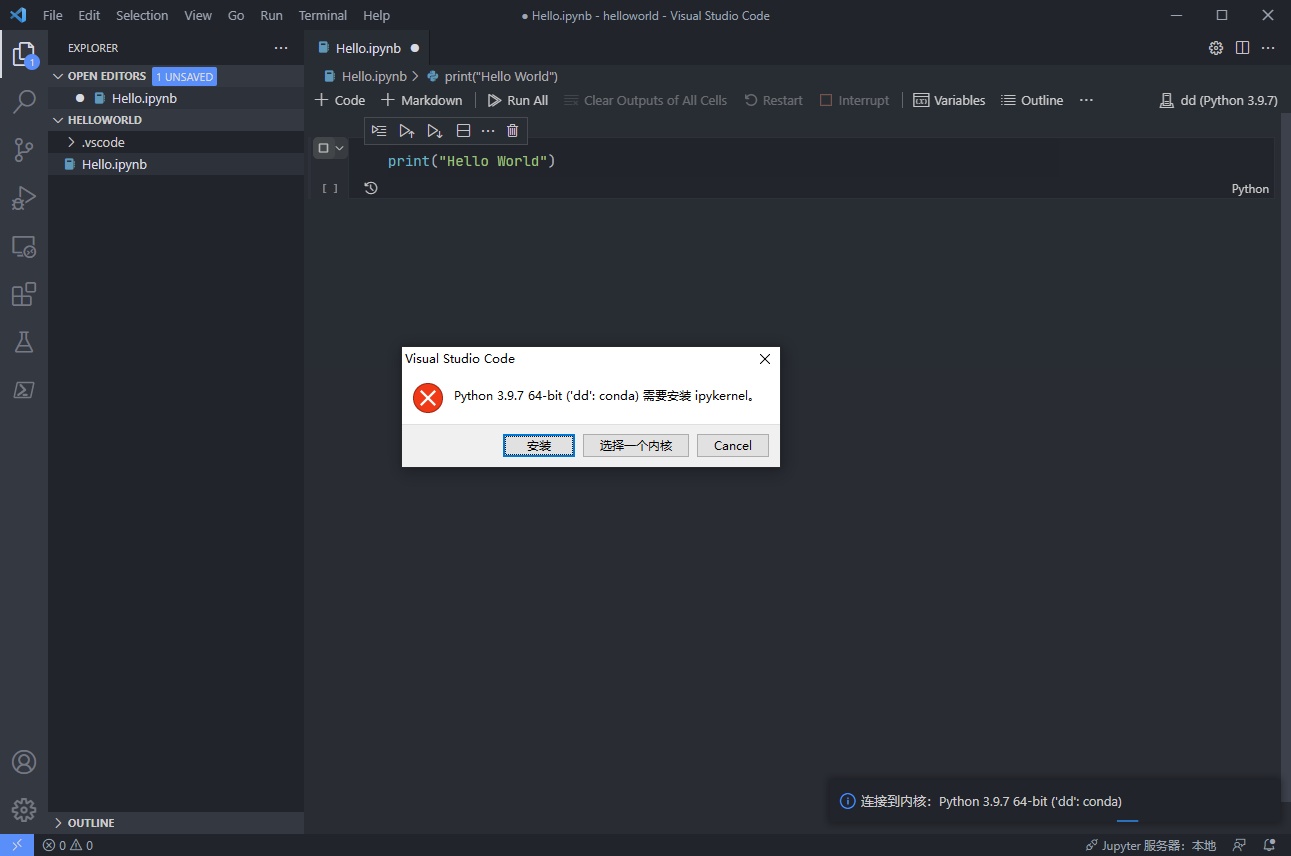
在Code中打开一个固定的Folder，使用Ctrl + Shift + P快捷键，搜索到Jupyter: Create New Jupyter Notebook，以新建一个笔记本



右上角会显示你正在使用base环境，更改为你新建的dd：



第一次运行代码时，Code会要求你安装IPykernel，选择安装：



在Code中新建一个笔记本，并且在代码格中输入以下内容，期望看到如图的结果，即为合格。

import pandas as pd  
import numpy as np  
from IPython.display import display  
​  
df = pd.DataFrame({ 'A': 1.,  
                   'B': pd.Timestamp('20130102'),  
                   'C': pd.Series(1, index=list(range(4)), dtype='float32'),  
                   'D': np.array([3] \* 4, dtype='int32'),  
                   'E': pd.Categorical(["test", "train", "test", "train"]),  
                   'F': 'foo'})  
display(df)

**任务2**

参考：<https://www.runoob.com/python3/python3-tutorial.html>

参考：<https://www.bilibili.com/video/BV1ZM4y1u7uF?p=5>

同学在这个环节中，无论是否有基础都必须要掌握Python语法。

掌握如下内容：菜鸟教程目录中**「基础语法」~「错误和异常」**，并且额外要掌握**「正则表达式」基础**

**基础题：**

1. 有四个数字：1、2、3、4，能组成多少个互不相同且无重复数字的三位数？各是多少？
2. 输入三个整数x,y,z，请把这三个数由小到大输出。
3. 输出 9\*9 乘法口诀表
4. 打印出所有的"水仙花数"，所谓"水仙花数"是指一个三位数，其各位数字立方和等于该数本身。例如：153是一个"水仙花数"，因为153=1的三次方＋5的三次方＋3的三次方。
5. 将一个正整数分解质因数。例如：输入90,打印出90=2\*3\*3\*5。

**任务3**

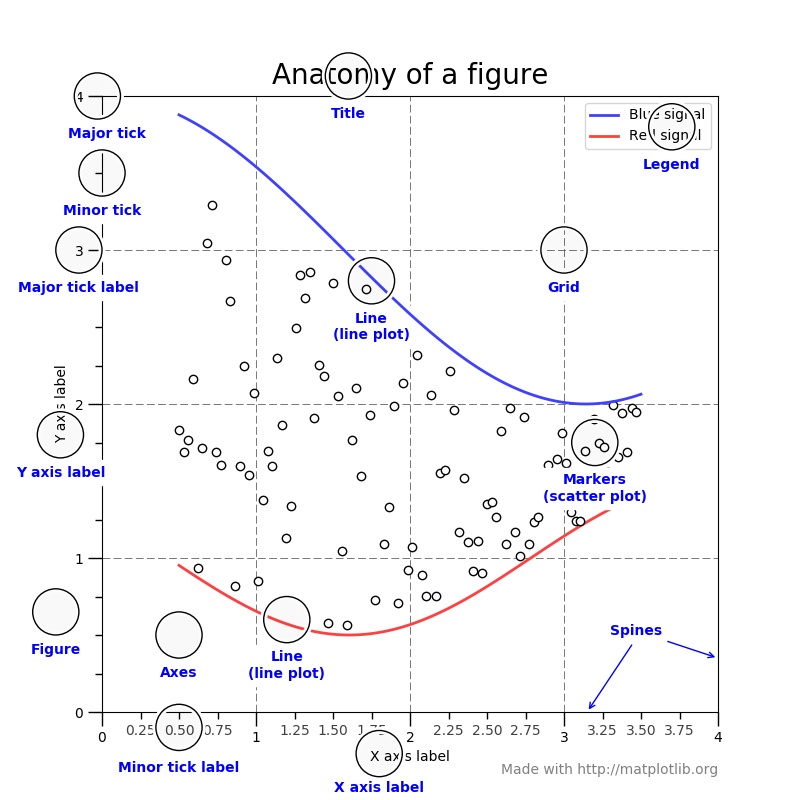
Pandas就是Python中的Excel，切不可因为是纯代码而怯场。

参考：<https://www.bilibili.com/video/BV1ZM4y1u7uF?p=5>

参考：<https://www.pypandas.cn/docs/getting_started/10min.html>

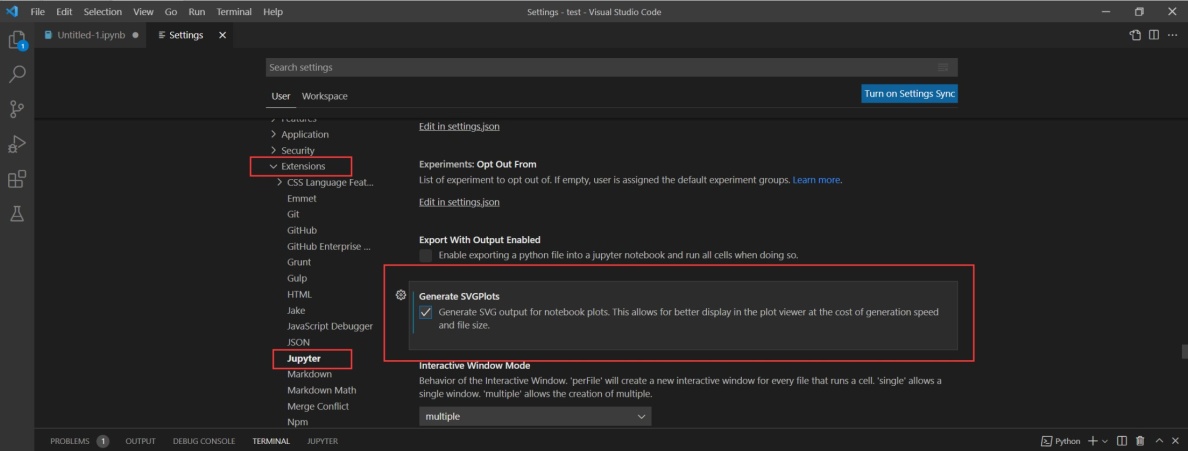
需要掌握：

* Numpy array以及围绕array的一些基本操作方法，比如shape、reshape、where、通过[:]的切片
* Pandas中Dataframe和Series的基础概念、赋值与枚举
* Pandas的文件读取，尤其是CSV
* Pandas缺失值处理
* Pandas中的DataFrame Merge与Grouping
* Pandas与Matplotlib.pyplot的联动，使用DataFrame数据绘图
* Matplotlib.pyplot需要学员掌握plot, bar, scatter, figure & subplot, text, labels, axis & grid, legend等的用法
* Matplotlib需要额外掌握image(imshow), pie



Matplotlib和Numpy基础操作不需要上完课之后烂熟于心，但是等到需要的时候查询教程和帮助文档能够快速上手。

*附：提高生成图像清晰度的小技巧*



在VSCode的设置 - Extensions - Jupyter中，勾选Generate SVGPlots。

**验收要求**

**Pandas基础练习**

**## 01.Chipotle快餐数据**

1. 导入数据

2. 查看前10行内容

3. 打印出该数据的列数

4. 打印出全部列的名称

5. 找出下单数最多的item

6. 打印出item\_name的数量

7. 找出在choice\_description中，下单次数最多的商品

8. 打印出商品被下单的总数

9. 打印出在该数据集对应的时期内的收入(revenue)数额

10. 打印出在该数据集对应的时期内的订单(order)数

11. 打印出每一笔订单的平均总价

**## 02. drinks数据**

1. 导入数据，查看头

2. 计算得到：哪个大陆(continent)平均消耗的啤酒(beer)更多？

3. 打印出每个大陆(continent)的红酒消耗(wine\_servings)的描述性统计值

4. 打印出每个大陆每种酒类别的消耗平均值

5. 打印出每个大陆对spirit饮品消耗的平均值，最大值和最小值

**## 03. Titanic**

1. 导入数据

2. 将PassengerId设置为索引

3. 绘制一个展示男女乘客比例的扇形图

4. 绘制一个展示船票Fare, 与乘客年龄和性别的散点图

5. 绘制一个展示船票价格的直方图

6. 有多少人生还？

**任务4**

参考：<https://liaocy.net/2020/05/02-statistics-formula/>

参考：<http://www.360doc.com/content/20/0225/07/6348482_894631222.shtml>

Pandas中Series或DataFrame自带的所有计算函数，具体可以查询：

<https://pandas.pydata.org/pandas-docs/stable/reference/series.html#computations-descriptive-stats>

用Pandas实现以下统计算法，并且要求必须从公式层面掌握：

* 均值
* 中位数
* 众数和异众比率
* 极差
* 方差和标准差
* 协方差
* 一维离散数据概率分布直方图
* 正态分布的拟合

**验收要求**

Pandas统计学练习

**## 01. Wind数据集练习**

1. 导入Wind数据，并展示表头

2. 设法解决违规日期（2016年）的问题，将其改为1961年

3. 将日期设为datetime64索引

4. 统计缺失值个数

5. 计算每个地区的风速平均值

6. 创建一个名为loc\_stats的数据框去计算并存储每个location的风速最小值，最大值，平均值和标准差

7. 创建一个名为day\_stats的数据框去计算并存储每天的风速最小值，最大值，平均值和标准差

8. 对于每一个location，计算一月份的平均风速

9. 对于数据，分别以年、月为频率取样

**## 02. Iris数据集练习**

1. 导入Iris数据

2. 计算Iris-setosa鸢尾花数据的：平均值、中位数、众数、极差、标准差、方差

3. 绘制Iris-virginica的四个数据直方图

**任务5**

* python中的lambda，Pandas在进行排序或筛选时会经常用到
* Pandas中的Where查询
* Pandas中的透视表
* Pandas中的字符串处理
* Pandas中的时序表
* apply()方法
* 一维线性回归

**验收要求**

使用second\_cars\_info.csv制作一个查询引擎，可以通过输入指定的参数获取对应条目的结果。

使用US\_Crime\_Rates\_1960\_2014.csv，根据1991-2014年的数据制作一条回归曲线，预测2019年的各个犯罪条目的犯罪率。