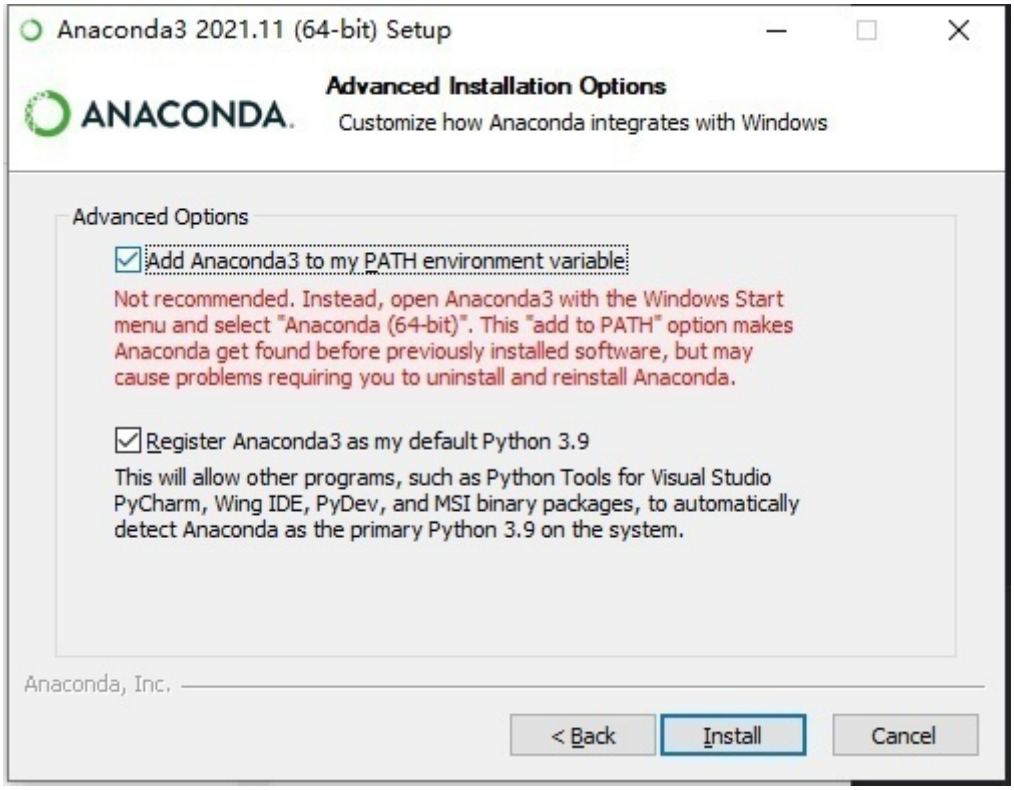
Week 1：Python与Pandas

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：



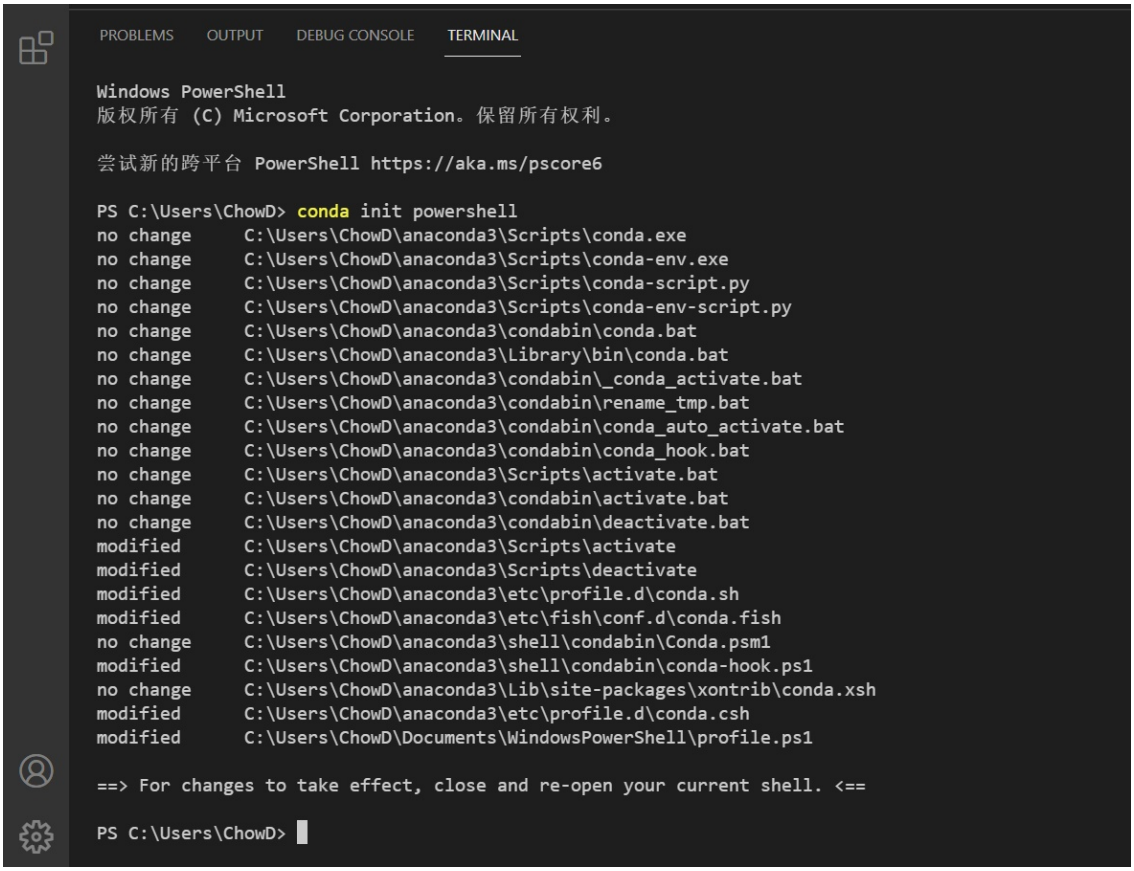
1. VSCode (x64)下载地址：https://code.visualstudio.com/sha/download?build=stable&os=win32-x64-user

   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：

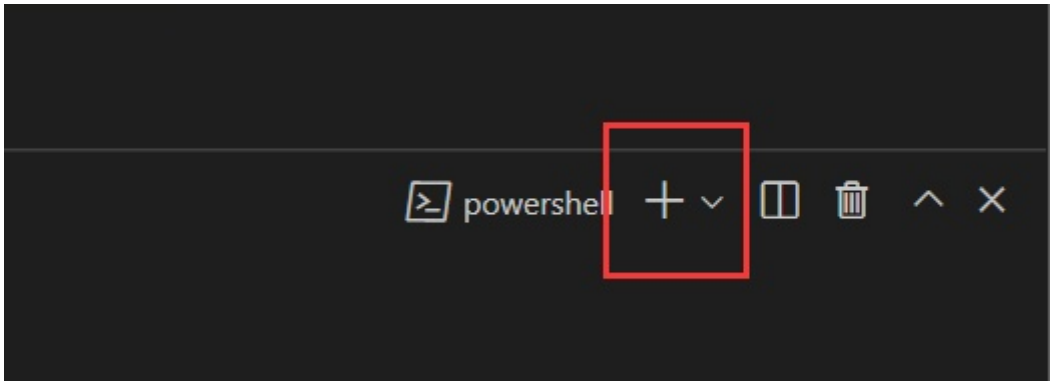
然后在Terminal - New Terminal中，尝试输入：

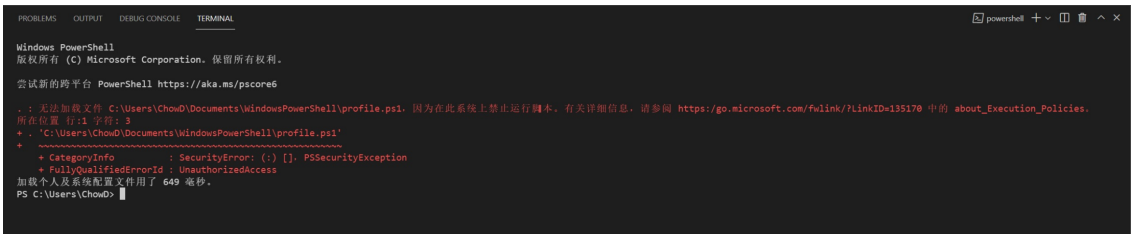
```
conda init powershell
```

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：

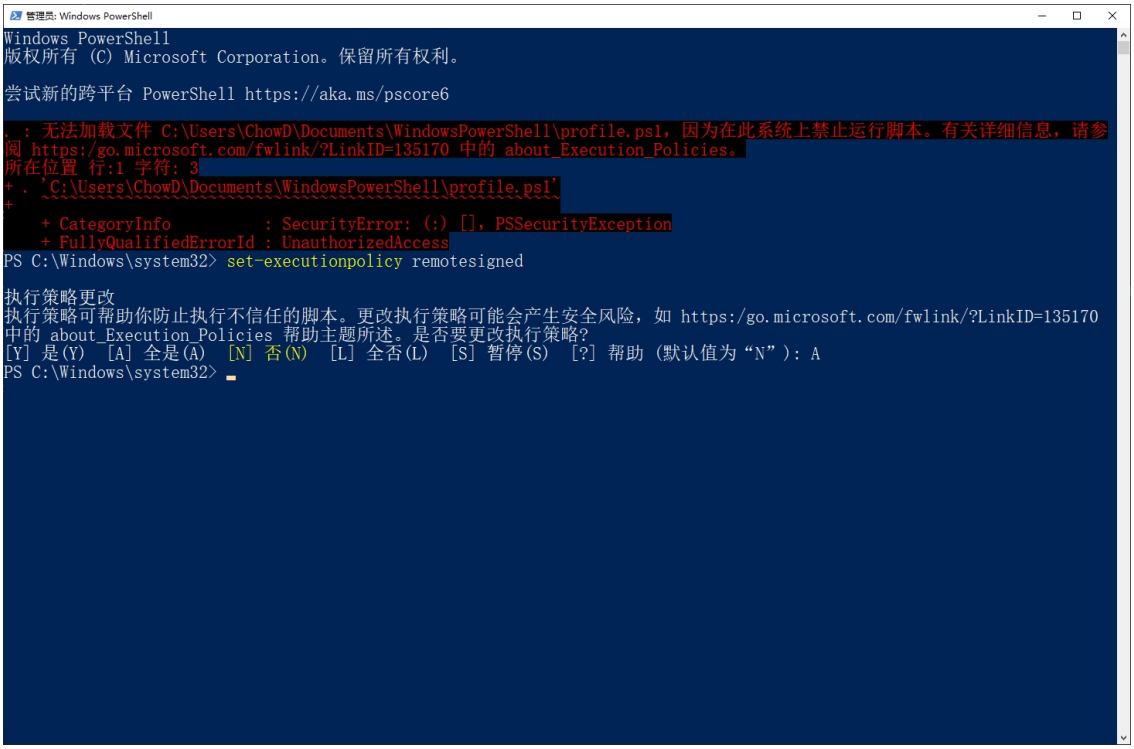




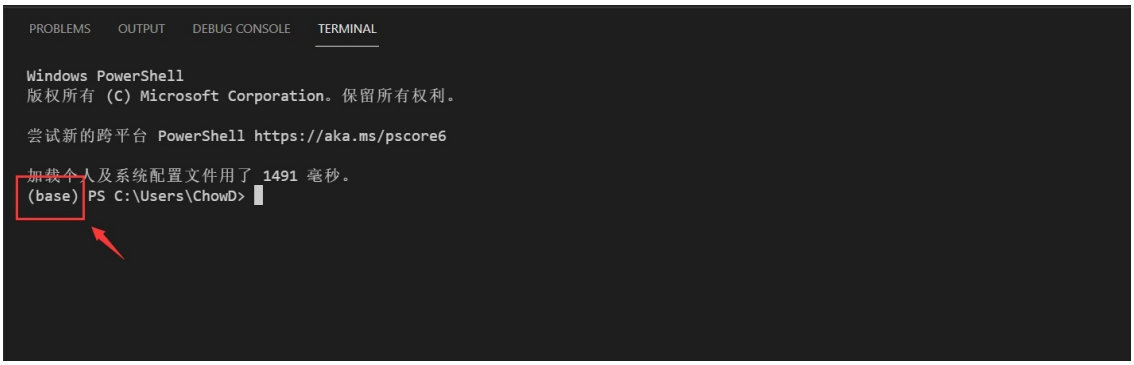
为了解决这个问题，我们需要在开始菜单右键 → Windows Powershell（管理员），并且输入

```
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。



2. 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

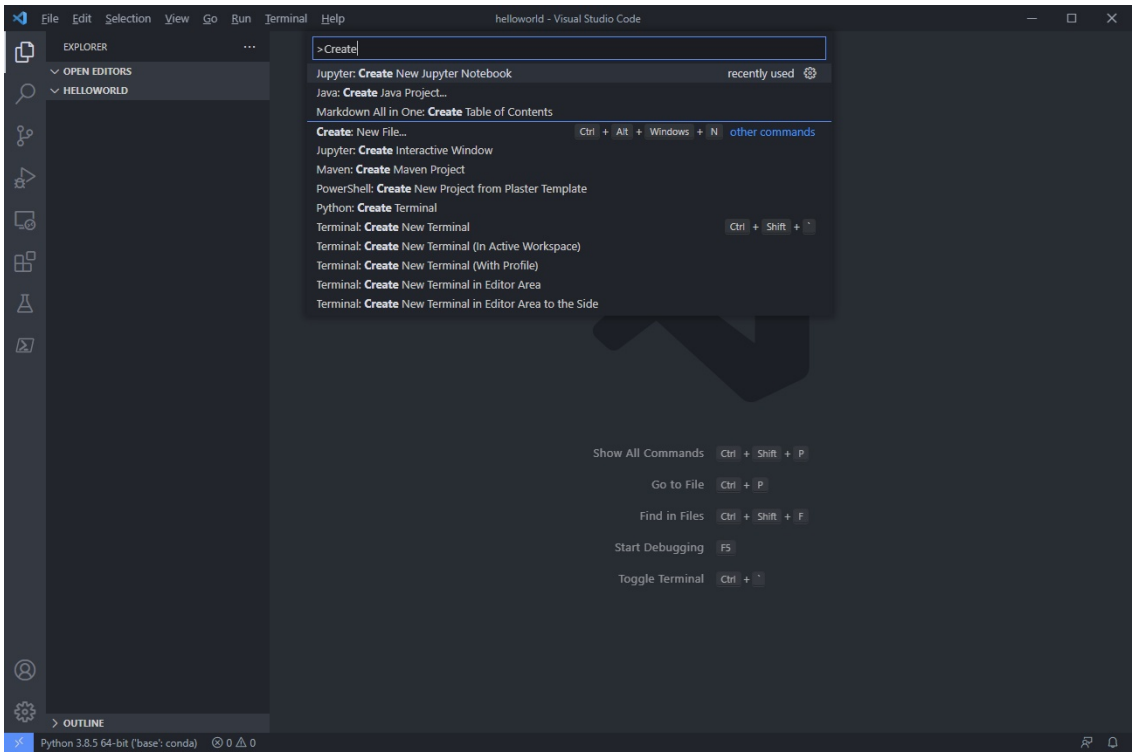
   ```
   conda create --name dd python=3.7
   conda activate dd
   ```

2. 安装依赖包

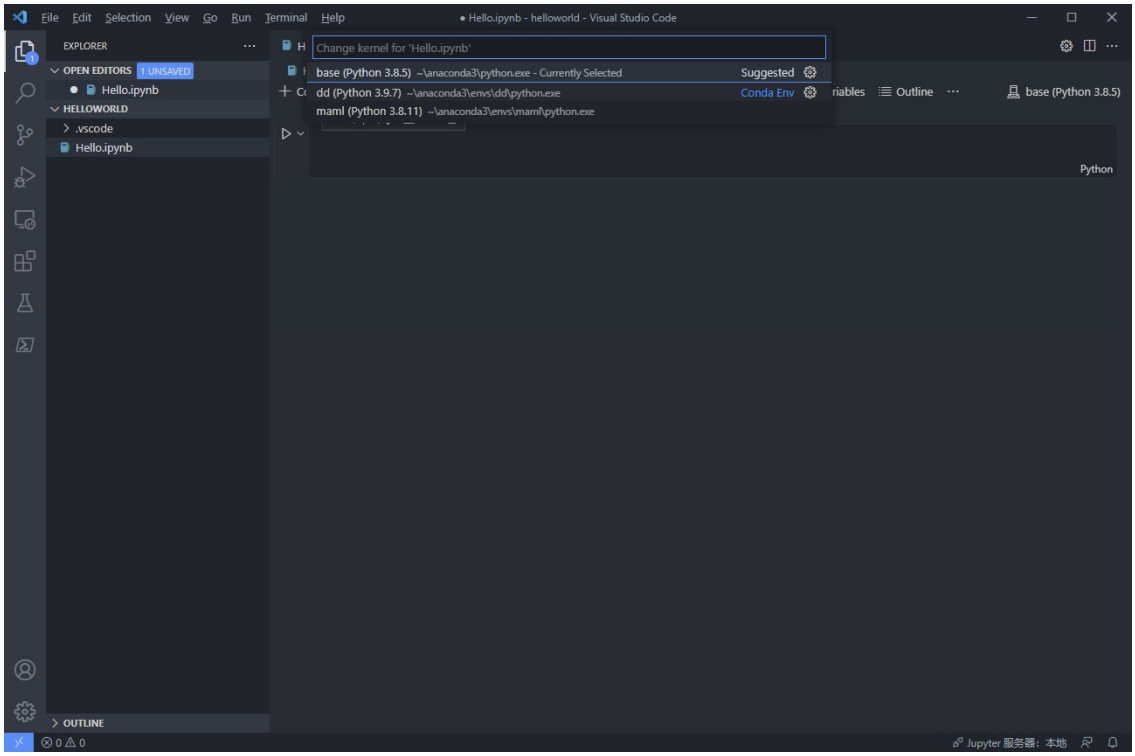
   ```
   conda install numpy pandas matplotlib
   ```

3. 新建一个笔记本

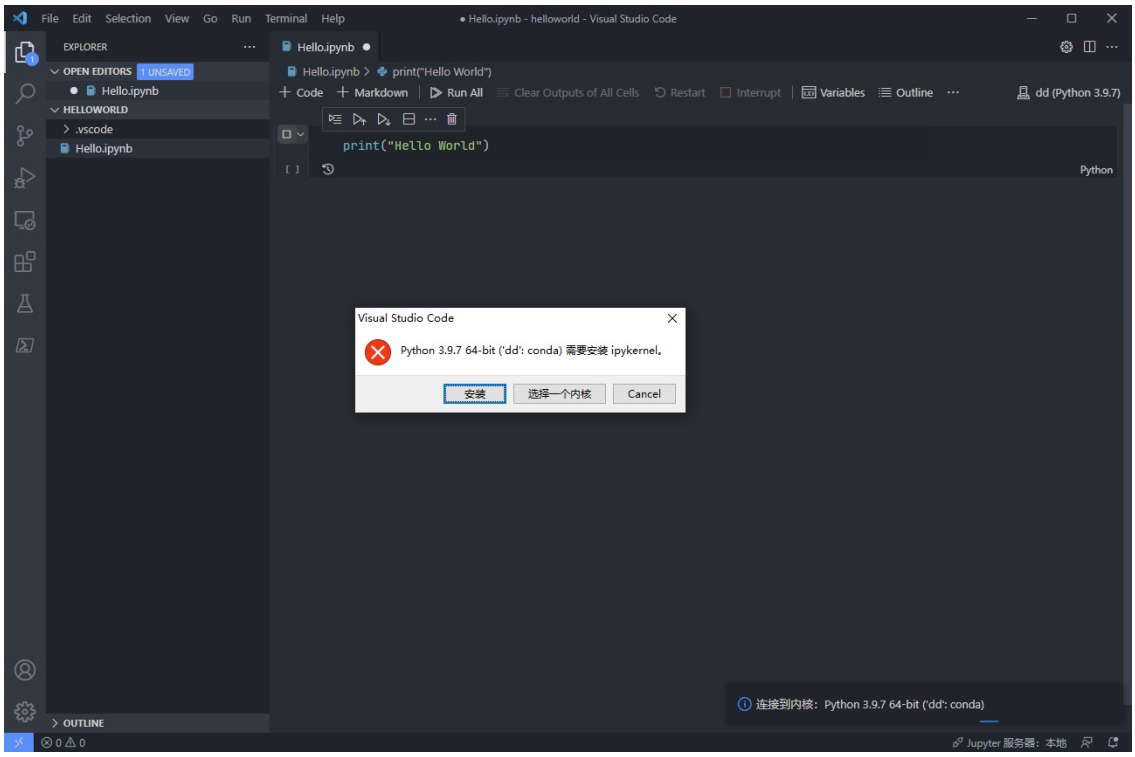
   在Code中打开一个固定的Folder，使用`Ctrl + Shift + P`快捷键，搜索到Jupyter: Create New Jupyter Notebook，以新建一个笔记本



   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

争取能够在这个环节中，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」，并且额外要求掌握「正则表达式」基础，面向对象及以后的内容不做硬性要求。

完成如下任务：

1. 理解（写成文档）

1. 简述 Python 中的几种数据类型
2. 数据类型的可变与不可变分别有哪些?
3. 元祖，列表，字典有没有长度的限制?
4. 集合有那些特性
5. 分别解释"=","==","+="的含义
6. 解释'and','or','not'
7. 深浅 copy-引用和 copy(),deepcopy()的区别

2. 编程

1. 求 1-100 之间能被 7 整除，但不能同时被 5 整除的所有整数。

2. 输出 1000 以内所有的"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

3. 判断 101-200 之间有多少个素数，并输出所有素数

4. 设 m=1*2…*n,求 m 为不大于 20000 时最大的 n

5. 利用 if 语句写出猜大小的游戏。预设一个 0-9 的随机整数，让用户输入所猜的数。如果大于预设的数，提示“bigger”；如果小于预设的数，显示“smaller”。如此循环，直至猜中该数，显示“right！”，并统计第几次猜中。

6. 读入一串字符，判断是否是回文串。“回文串”是一个正读和反读都一样的 字符串，比如“level”或者“noon”等等就是回文串。 输入 一行字符串，长度不超过 255。 输出如果是回文串，输出“YES”，否则输出“NO”。

7. 一球从 100 米高度自由落下，每次落地后反跳回原高度的一半；再落下，求它在第10 次落地时，共经过多少米？第 10 次反弹多高？

8. 使用 while,完成以下图形的输出

```
*
* *
* * *
* * * *
* * * * *
* * * *
* * *
* *
*
```

9. 输入某年某月某日，判断这一天是这一年的第几天？

```
输入：
year:
2015
month:
6
day:
7 .
输出：
it is the 158th day.
```

10. 定义一个列表名为 nameList，里面包含 6 个元素；

```
（1）.查看一下列表的长度
（2）.实现列表的增（增加一个元素）、删（删除第二个元素）、改（修改第一个
元素）、查（查询第一个元素）
```

11. 使用计数器方式（统计一下‘飘飘乎如遗世独立，羽化而登仙。’这一句话中的每个字符出现的个数），以及出现最多的 2 个数。

任务3

（注：Pandas 就是 Python 中的 Excel，切不可以因为纯代码而怯场）

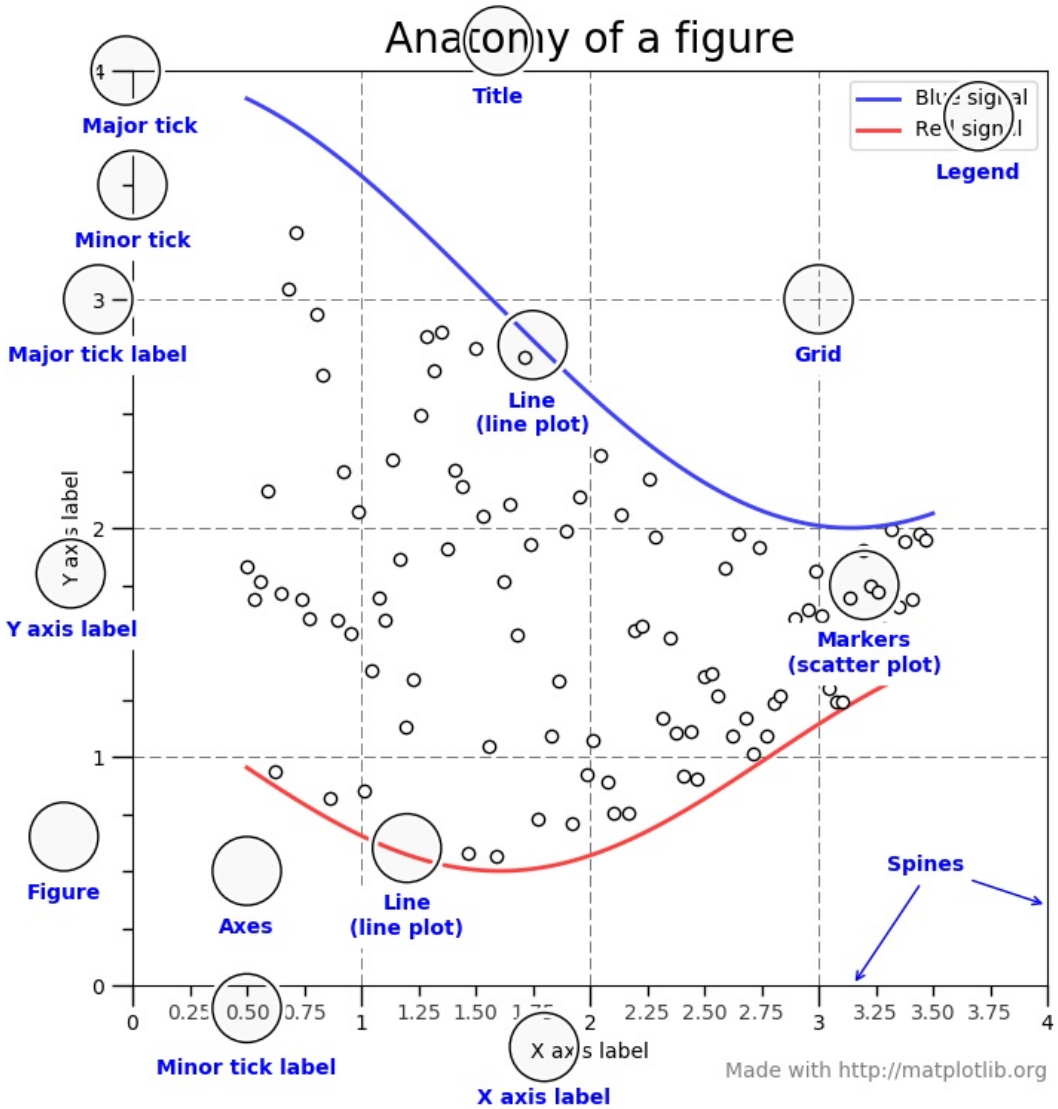
参考：https://www.bilibili.com/video/BV1ZM4y1u7uF?p=5

Numpy 参考：https://www.yiibai.com/numpy/numpy_data_types.html

Matplotlib 参考：https://www.yiibai.com/matplotlib/lifecycle.html

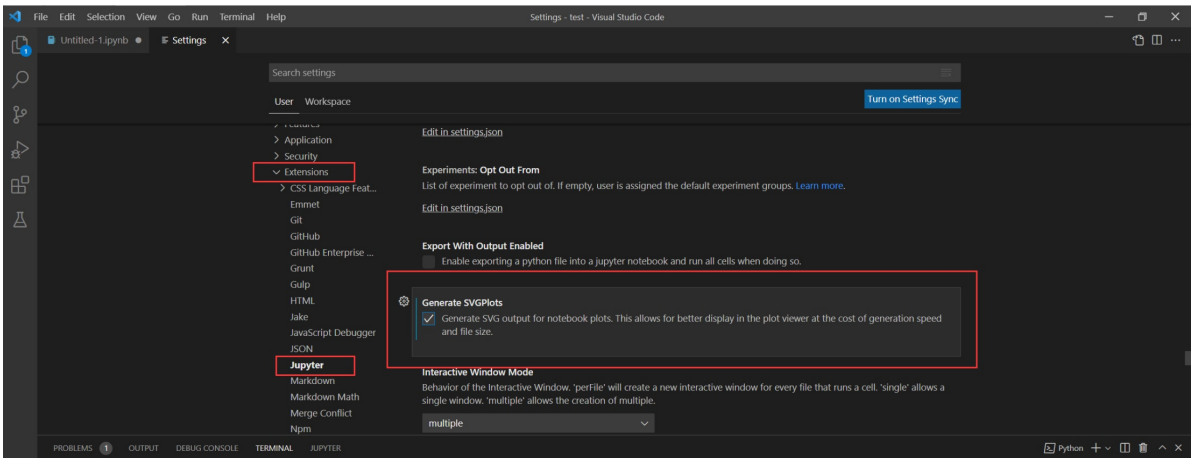
Pandas 参考：https://www.yiibai.com/pandas

- 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基础操作不需要一下子就烂熟于心，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。

任务要求：

一．根据 chipotle.tsv 的数据完成如下任务：

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

二、 根据 Titanic 数据集完成如下任务

12. 导入数据
13. 将 PassengerId 设置为索引
14. 绘制一个展示男女乘客比例的扇形图
15. 绘制一个展示船票 Fare, 与乘客年龄和性别的散点图
16. 绘制一个展示船票价格的直方图
17. 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

完成任务：

根据 Wind 数据集完成下面的内容：

1. 导入 Wind 数据，并展示表头
2. 设法解决违规日期（2061 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

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

完成任务：

1. 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。
2. 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。