

# 纯小白需要看的问题

2018年5月28日星期一 10:09

每次开启  
workspace都需要  
从前往后运  
行一遍代码

## 问题1：

```
import pandas
import matplotlib inline

以上海数据为例，我们先使用Pandas的read_csv函数导入第一个数据集，并使用head、info、describe方法来查看数据中的基本信息。

In [ ]: Beijing_data = pd.read_csv('BeijingPM20100101_20151231.csv')

In [1]: Beijing_data.head()

NameError                                Traceback (most recent call last)
<ipython-input-1-91eced043bia> in <module>()
----> 1 Beijing_data.head()

NameError: name 'Beijing_data' is not defined
```

从运行结果可以看出，除了上面提到的数据列之外，上海数据中还包含有 PM\_10main 和 PM\_10sub1 两个观测站点的监测数据。并且数据中PM2.5的这三列包含有缺失值“NaN”。

在项目工作区改，In[]后面没有数字了?咋回事。。

每次重新开workspace要从到到尾都运行一遍哦

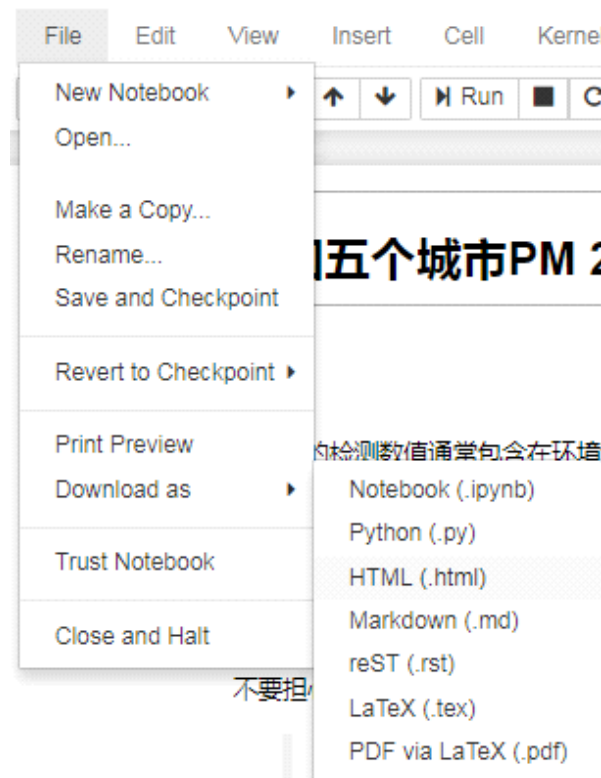
每段代码都shit+回车就行啦

## 问题2：

怎么去保存成html，我是直接在网页的项目工作区写的

File里面哦，有download as，可以选html哒，可以参考下面这个流程哈：

如何保存html  
和ipynb



ipynb也按照上面这个方式下载就可以啦~

## 问题3：

文本要改成  
markdown

```
In [46]: **问题 4b**: 上述可视化有何有趣的趋势? 是否能够回答你的第二个问题? (如果不能, 请说明你需要什么信息来帮助你来回答问题)

**答案**: 通过分析北京2012年到2015年的PM2.5和HUMI的数据, 通过图中的散点图可以看出, 随着温度是上升, 平均PM2.5的值也在上升

File "<ipython-input-46-b08b492eca13>", line 1
**问题 4b**: 上述可视化有何有趣的趋势? 是否能够回答你的第二个问题? (如果不能, 请说明你需要什么信息来帮助你来回答问题)
~
SyntaxError: invalid syntax
```

这个问题怎么解决呀

要换成markdown才能写哦, 你现在code模式哒, 所有文字部分的撰写都要在这个地方改成  
Markdown才能运行正确哦



问题4 :

老师, 我要试着运行一下代码, 不需要下载python吗?

练习直接在代  
码框里敲就可  
以了

### 练习: 平均电费

现在我们尝试一下在 Python 中进行数学运算吧!

我在过去三个月的电费分别是 \$23、\$32 和 \$64。那么我这三个月的平均月电费是多少? 请编写一个表达式来计算平均值, 并使用 `print()` 查看结果, 最后提交答案。

```
1 # Write an expression that calculates the average of 23, 32 and 64.
2 # Place the expression in this print statement.
3 print()
```

还是在这里就可以的?

不用下载的哈, 直接在里面输入你想输入就行啦

问题5 :

老师是不是把这个问题替换成我之前的问题?

由于项目难度设定的问题, 在后面的分析中我们暂时没有对气象数据的处理和分析, 如果同学感兴趣的可以自行探索。如果你有足够的能力, 我们也欢迎你不用项目模板中的代码, 对数据自行进行分析~

问题 1: 至少写下两个你感兴趣的问题, 请确保这些问题能够由现有的数据进行回答。

(问题示例: 1. 2012年-2015年上海市PM 2.5的数据在不同的月份有什么变化趋势? 2. 哪个城市的PM 2.5的含量较低?)

答案:

第一个问题: 将此文本替换为你的回答!

第二个问题: 将此文本替换为你的回答!

双击修改  
Markdown中的  
内容

对哒~你可以参考我发的html是怎么写的呢，可以仿照这写一下  
可是我找不到编辑的按钮呢  
双击就可以啦，不要在Html中修改哦，要去优达的工作区修改哈

### 问题6：

python不识别  
中文标点

```
In [2]: # print the length of data
print("The number of row in this dataset is ",len(Shanghai_data.index))

# calculating the number of records in column "PM_Jingan"
print("There number of missing data records in PM_Jingan is: ",
      len(Shanghai_data.index) - len(Shanghai_data["PM_Jingan"].dropna()))

NameError                                Traceback (most recent call last)
<ipython-input-2-227b47495a40> in <module>()
      1 # print the length of data
--> 2 print("The number of row in this dataset is ",len(Shanghai_data.index))
      3
      4 # calculating the number of records in column "PM_Jingan"
      5 print("There number of missing data records in PM_Jingan is: ",

NameError: name 'Shanghai_data' is not defined
```

老师，这个出错了是什么意思？

注意不能用中文的标点哦，要用英文的标点

### 问题7：

python中要注意  
大小写是否  
正确

```
[23]: # TO DO: Second question
df2 = reading_stats(df_all_cities, ["city == 'shanghai'"])

There are 0 readings (0.00%) matching the filter criteria.
The average readings of PM 2.5 is nan ug/m^3.
The median readings of PM 2.5 is nan ug/m^3.
25% of readings of PM 2.5 are smaller than nan ug/m^3.
```

为什么是0呢？

上海的s要大写哦



```
In [27]: # TO DO: Second question
df2 = reading_stats(df_all_cities, "city == 'Shanghai'")

ValueError                                Traceback (most recent call last)
<ipython-input-27-c5d0f54422ee> in <module>()
      1 # TO DO: Second question
--> 2 df2 = reading_stats(df_all_cities, "city == 'Shanghai'")

<ipython-input-19-4358100368de> in reading_stats(data, filters, verbose)
      9 # Apply filters to data
     10 for condition in filters:
--> 11     data = filter_data(data, condition)
     12
     13 # Compute number of data points that met the filter criteria.

<ipython-input-18-be6143987e19> in filter_data(data, condition)
     12 # Only want to select on first two values connecting field name operator and
```

这个运行又错误了

```
, ["city == 'Shanghai'", "year >= 2012"])
```

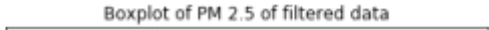
少了中括号哦

### 问题8：

注意年份选择  
是否会影响  
数据的提取


```
In [34]: # TO DO: Second question
df2 = reading_stats(df_all_cities, ["city == 'Shanghai'"])

There are 52584 readings (20.00%) matching the filter criteria.
The average readings of PM 2.5 is 52.91 ug/m^3.
The median readings of PM 2.5 is 41.00 ug/m^3.
25% of readings of PM 2.5 are smaller than 26.00 ug/m^3.
25% of readings of PM 2.5 are larger than 67.00 ug/m^3.
```



```
In [24]: # TO DO: First question
df1 = reading_stats(df_all_cities, ["city == 'Shanghai'", "year >= 2010"])

There are 52584 readings (20.00%) matching the filter criteria.
The average readings of PM 2.5 is 52.91 ug/m^3.
The median readings of PM 2.5 is 41.00 ug/m^3.
25% of readings of PM 2.5 are smaller than 26.00 ug/m^3.
25% of readings of PM 2.5 are larger than 67.00 ug/m^3.
```



我发现我这两条数据运行后的数据是一样的

因为数据是从2010开始的，那么第二个year的限制条件相当于没用上哦，所以数据是一样多的

## 问题9：

注意大小写哦

老师，这个散点图，怎么表达

```
In [37]: # TO DO:
# please use univariate_plot to visualize your data
df_wz.plot(x='PM_US Post', y='Temp', kind='scatter')

NameError                                Traceback (most recent call last)
<ipython-input-37-f5accf693602> in <module>()
      1 # TO DO:
      2 # please use univariate_plot to visualize your data
----> 3 df_wz.plot(x='PM_US Post', y='Temp', kind='scatter')

NameError: name 'df_wz' is not defined
```

PM\_US Post和Temp没写对哦，应该跟列名称一样：PM\_US\_Post和TEMP这样子哈

## 问题10：

项目中没有让  
修改的部分最  
好不要动哦

讲解部分，有举例的分析里，我把不需要改的代码改了有问题么？

项目没让我们修改的地方最好不要改，改了可能会运行报错，倒是没有报错。

嗯嗯看你改的哪部分，可能内部取数据的机制已经改变了，但是你不知道，所以有可能取出来的数据可能是不准的，所以项目没有让你修改的地方最好不要动，除了加了列那一块结论可以有否定结论么？比如一天的变化趋势，变化不明显，需要考虑其他气象因素可以的，数据分析就是这样，只要提取出来的数据没问题，结果是什么就说什么，照着可视化结果如实说就可以了。

照可视化的结  
果如实分析