数据分析 7 天入门项目

编写：小乐

审核：优达数据分析助教团队

作为一名新手小白，如何着手项目中的数据分析工作呢？具体步骤有哪些？

如何展现分析成果呢？

阅读下面的文档，了解数据分析项目（下简称“**项目**”）的基本工作流程。

一． 数据背景

介绍数据采集的背景（调查收集单位和事由），说明数据来源（文献名称，下载链接等），解释数据组成（变量及其含义等），注明数据的采集时间或时间跨度。

项目：数据包含有中国五个城市(北京、上海、成都、广州、沈阳)从 2010/1/1-2015/12/31 的空气和气象数据。数据包含 PM，year，month，season 等 15 个属性列（见下图），数据中的缺失值被标记为 NA。
数据来源：https://archive.ics.uci.edu/ml/datasets/PM2.5+Data+of+Five+Chinese+Cities#

```
No: 行号
year: 年份
month: 月份
day: 日期
hour: 小时
season: 季节
PM: PM2.5浓度 (ug/m^3)
DEWP: 露点 (摄氏温度) 指在固定气压之下，空气中所含的气态水达到饱和而凝结成液态水所需要降至的温度。
TEMP: Temperature (摄氏温度)
HUMI: 湿度 (%)
PRES: 气压 (hPa)
cbwd: 组合风向
Iws: 累计风速 (m/s)
precipitation: 降水量/时 (mm)
Iprec: 累计降水量 (mm)
```

问题 1：至少写下一个你感兴趣的问题，请确保这些问题能够由现有的数据进行回答。
提示：后续的数据分析步骤应基于提出的问题内容展开，注意项目内在逻辑的一致性。此外，考虑到项目的难度设定，请暂时不要提出与气象数据（DEWP，TEMP，HUMI，PRES，cbwd，Iws，precipitation，Iprec）相关的问题。

二． 项目目的

阐述本次项目目标，明确任务点。

项目：本项目中的数据分析流程和分析中使用的函数已经给出，重在培养以数据分析师的身

份执行数据的探索性分析，了解数据分析过程的基本流程。

三． 库与数据导入

使用 Python 进行数据分析常依赖第三方库，常用的有 numpy, pandas, matplotlib 等，使用时需用使用 “import” 语句导入。

Python 支持多类型数据的导入，支持的格式包括 csv, excel, sas, json, html 等，使用时需要依赖不同的模块导入数据。例如：pandas 的 read_csv 函数用于读取 csv，返回数据类型为 DataFrame。

项目：使用 import 语句导入 csv，numpy，pandas，matplotlib.pyplot，seaborn。学习学习导入时 as 的用法，注意%matplotlib inline 功能是可以内嵌绘图。此外，”from helper_functions import filter_data, reading_stats, univariate_plot”语句导入本项目中定义的三个函数 filter_data, reading_stats 和 univariate_plot。

四． 数据评估

使用 head、info、describe 方法快速查看和了解数据集的基本信息，包括数据类型，有效记录个数，变量名称，统计信息（平均值、标准差、最大值、最小值、上下四分位等）。此外，需对错误数据、缺失值（NaN 表示）和离群值进行评估。

项目：使用 read_csv 方法读取数据集，以上述方法查看数据集 Shanghai_data 的基本信息，对数据集做简单处理（列名中的空格用下划线替代，season 列处理为分类变量），以及使用 Shanghai_data.isnull().sum()查看数据缺失情况。

参照上述方法，进行北京数据的评估，包括：数据读取，缺失值评估等（这部分内容在习题部分已完成）。

五． 数据整理

依据上述项目目的，将原始数据数据整理为便于分析的格式。

项目：构建数据集 df_all_cities，数据列包含序号、日期、季节、PM_US Post 站点 PM2.5 监测值和城市名称。只选取 PM_US Post 站点的原因是该站点监测的缺失值相对较少，利于分析。

注意：后续的数据筛选基于数据集 df_all_cities 完成，如同学对其它气象参数感兴趣可自行分析。

	No	year	month	day	hour	season	PM_US Post	city
0	1	2010	1	1	0	Winter	NaN	Beijing
1	2	2010	1	1	1	Winter	NaN	Beijing
2	3	2010	1	1	2	Winter	NaN	Beijing
3	4	2010	1	1	3	Winter	NaN	Beijing
4	5	2010	1	1	4	Winter	NaN	Beijing

六． 数据筛选

针对不同的分析目标，定义数据筛选条件，获取可以直接用于分析及可视化展示的数据。

项目：通过定义 filter_data 函数筛选、提取感兴趣的数据，实现类似于 Excel 中的筛选功能；利用 reading_stats 函数对筛选出来的数据进行统计分析，包括平均值、中位值、上下四分位值。需要关注每个函数的输入值和返回值：filter_data 函数包括数据集 data 和筛选条件 condition 两个输入，返回满足筛选条件的数据集；reading_stats 函数包括数据集 data，数据过滤器 filters 和 verbose 三个输入，返回满足筛选条件的数据集，同时在 verbose=True（默认条件）时，打印统计分析结果。由于 filter_data 函数已经内嵌于 reading_stats 函数中，在实际分析时只需调用 reading_stats 函数。因此，需注意 reading_stats 函数输入值的定义要求，具体如下：

第一个参数（必须）：需要被加载的 dataframe，数据将从这里开始分析。
第二个参数（可选）：数据过滤器，可以根据这些条件来过滤将要被分析的数据点。过滤器应作为一系列条件提供，每个条件之间使用逗号进行分割，并在外侧使用""将其定义为字符串格式，所有的条件使用[]包裹。每个单独的条件应该为包含三个元素的一个字符串：'<field> <op> <value>'（元素与元素之间需要有一个空格字符来作为间隔），<op>可以使用以下任意一个运算符：>、<、>=、<=、==、!=。数据点必须满足所有条件才能计算在内。例如，["city == 'Beijing'", "season == 'Spring'"] 仅保留北京市，季节为春天的数据。在第一个条件中，<field>是 city，<op>是 ==，<value>是'Beijing',因为北京为字符串，所以加了单引号，它们三个元素之间分别添加一个空格。最后，这个条件需要使用双引号引用起来。这个例子中使用了两个条件，条件与条件之间使用逗号进行分割，这两个条件最后被放在[]之中。
第三个参数（可选）：详细数据，该参数决定我们是否打印被选择的数据的详细统计信息。如果 verbose = True，会自动打印数据的条数，以及四分位点，并绘制箱线图。如果 verbose = False，则只会返回筛选后的 dataframe，不进行打印。

问题 2a: 要回答你前面的提出的问题，你需要分别筛选哪部分的数据？请具体说明。
提示：基于问题 1 中提出的问题，进行数据筛选（最好可以基于数据集 df_all_cities 通过 filter_data 函数实现筛选，如若不能，需自行分析）。
问题 2b：请使用上面给出的 reading_stats 函数来调用你所需要的数据，请在下面填写合适的条件（conditions）。
提示：参考上述“第二个参数（可选）”的要求定义，注意每个条件分三个元素，每个元素需一个空格字符间隔，每个条件定义为字符串格式，多个条件之间用逗号隔开。此外，所定义的条件需能回答前文提出的问题 1，注意保持项目内在逻辑的一致性。
问题 2c：你获取的数据分别包含多少条记录，统计特征如何？
提示：使用问题 2b 中定义的条件补充下面的代码，获取统计特征：

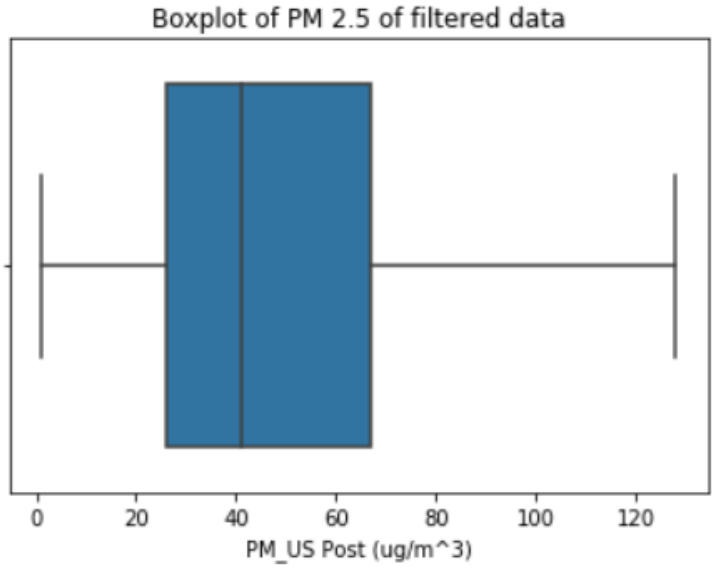
```
# TO DO: First question
df1 = reading_stats(df_all_cities, _____________)
```

七． 数据探索

探索数据中单变量的分布（箱线图，柱状图等），双变量的相关性，多变量之间的关系，建立数据模型。

项目：reading_stats 函数的结果之一（verbose=True），绘制满足筛选条件的数据的箱线

图，展示数据的分布特征，例如上海 PM_US Post 站点 2012 至 2015 年 PM2.5 监测值的分布特征如下：



八． 数据可视化

依据上述项目目的，将整理好的数据进行可视化展示，有助于沟通交流。

项目：定义 univariate_plot 作图函数，实现 PM 2.5 的观测平均值的柱形图可视化展示，关于该函数的输入项详细说明如下：

1. 第一个参数（必须）：筛选后数据的 dataframe，将从这里分析数据。
2. 第二个参数（必须）：数据分析进行的维度，在这里可以填入一个 column_name，比如'season', 'month', 'hour'等，对数据进行分组分析。
3. 第三个参数（可选）：可视化中柱形的颜色，默认为蓝色，你也可以选择你喜爱的其他颜色，比如 red，blue，green 等。但是请尽量保证一份可视化报告中图表颜色的一致和整洁性。

TO DO:
please use univariate_plot to visualize your data
提示：调用 univariate_plot 函数，探索数据可视化，注意该可视化需以前文提出的问题 1 为目标进行。
问题 3：上述可视化有何有趣的趋势？是否能够回答你的第一个问题？（如果不能，请说明你需要什么信息来帮助你来回答问题）
提示：基于可视化结果，分析 PM2.5 数据随时间、季节等的变化规律。

九． 总结与讨论

总结项目的发现，讨论项目的局限性和有待完善的方面。

项目：以本项目为例，展示了数据从导入、评估、整理到可视化的流程。受项目代码的限制，在这里我们无法直接完成多维度的灵活分析。在接下来的正式课程中，随着大家学习

的深入，可以观察其他气象数据与 PM2.5 的相关关系，不同站点 PM2.5 数据的横向比较等，甚至可以建立数学模型进行模拟预测。

问题 4：根据目前你对数据分析的了解，请思考一个可以应用数据科学技术的话题或兴趣领域。你希望使用什么样的数据，来得到什么样的信息？

提示：请大家结合自己的工作学习实际发散思考，畅所欲言！