BMIS2542-数据编程要点与Python

会议-第2个和第3个挑战选项：与熊猫和EPA数据集的数据整理

使用[EPA数据集发布在这里](https://pitt-my.sharepoint.com/:f:/g/personal/narayanr_pitt_edu/EqOZpO62G2xOu9DcAl5x3mEBbp0kt6sD8vNZt6i1d2fsUQ?e=OkZyo5)，EPA-data-2015.csv和EPA-data-2016.csv，为这个练习。将数据集读入Python作为熊猫数据帧，并回答以下问题。您可以使用Python生态系统中可用的任何库：

1. 2015年数据与2016年数据的比较：
   1. 在2015年至2016年期间，美国环保署数据中的总体缺失值是否有所增加或减少？
   2. 列出数据中缺失值增加的状态和数据中缺失值减少的状态。
   3. 使用以下列创建一个新的Pandas数据帧，填充该数据帧，并将其写入一个CSV文件：
      * 年
      * 月
      * 国家名称
      * 县名
      * 缺失比例：这是缺失观测值与该县在相应月份和年份报告的总观测值的比率。
2. 使用“AQI”和“算术平均数”（PM2.5）作为空气质量的指标，以回答以下问题：
   1. 平均而言，2015年至2016年，美国存在污染。S.增加或减少？
   2. 在2015年和2016年的数据帧中插入一列，并称之为“AQI类别”。填充AQI类别的值如下：

|  |  |
| --- | --- |
| AQI类别 | AQI值 |
| 好的 | 0-50 |
| 中度 | 51-100 |
| 不健康敏感 | 101-150 |
| 不健康 | 151-200 |
| 非常不健康 | 201-300 |
| 冒险的 | >300 |

* 1. 当AQI水平在0-50范围内时，空气质量被认为是“良好的”。“匹兹堡在2015年有多少天没有达到良好的空气质量标准？”2016年怎么样？
  2. 比较和评论匹兹堡和费城之间的空气质量水平。
  3. 在2015年和2016年，在匹兹堡没有很好的空气质量的日子里，费城的空气质量好吗？
  4. 如果你对U进行排序。S.根据空气质量，数据集中的城市，2015年和2016年匹兹堡在哪里？提供绝对排名和百分位数项中的排名。

1. 使用数据集中的“事件类型”变量来回答以下问题：
   1. 在数据集中创建一个指示器（或虚拟）变量，并将其命名为“事件虚拟”；如果事件类型=为“无”，则事件虚拟应设置为0，否则应设置为1。
   2. 使用事件虚拟变量，找出在任何给定的“当地现场”的空气质量是否受到现场所经历的事件的影响。将这些州划分为美国的知名地区。[[1]](#footnote-0)，看看你在不同地区的发现是否有差异。
2. 评论数据中是否存在异常值。
3. 合并这两个数据集，然后创建以下内容：
   1. 一个新的变量，它是“州代码”和“县名”的组合；称之为“州-县”。
   2. 创建一个以“州-县”、“月”和“年”为索引的数据透视表。AQI的值必须在数据透视表中取平均值。然后，查询您创建的包含短语“新”值的城市透视表（纽约、新港等）。

1. 例如，这个维基百科页面列出了美国人口普查中考虑的不同地区：<https://en.wikipedia.org/wiki/List_of_regions_of_the_United_States> [↑](#footnote-ref-0)