

《数据可视化技术》期末大作业

课程名称 数据可视化技术

题目名称 北京市2010~2014空气质量可视化研究

学生学院 计算机学院

专业班级 计算机科学与技术17(7)

学 号 3117004706

学生姓名 梁俊斌

指导教师 林志毅

2020年 6 月18 日

目录

1. 题目——————————————————3
2. 数据——————————————————3
3. 可视化工具———————————————4
4. 可视化方案或可视化实现过程———————4
5. 导入数据并进行数据清洗—————————4
6. 处理数据属性并建立联系—————————5
7. 规划与整理可视化窗口——————————6
8. 调节参数————————————————7
9. 检验可视化方案—————————————7
10. 可视化结果———————————————7
11. 北京市气温、污染、风速和气压的概览———7
12. 空气污染与时间的关联分析————————10
13. 空气污染与积累风速的关联分析——————11
14. 总结——————————————————12
15. 心得体会————————————————13
16. 参考资料————————————————13
17. **题目**

从网络上下载一组数据（自行获取），选择一种可视化工具（Excel、Tableau、Matlab、Echarts等），设计一种可视化方案实现该数据的可视化，并做适当的数据分析（或挖掘）。

内容还可实现： 堆积式南丁格尔玫瑰图、电影人物关系可视化、社会网络可视化（可采用相关工具箱）、Pm2.5浓度分析等。

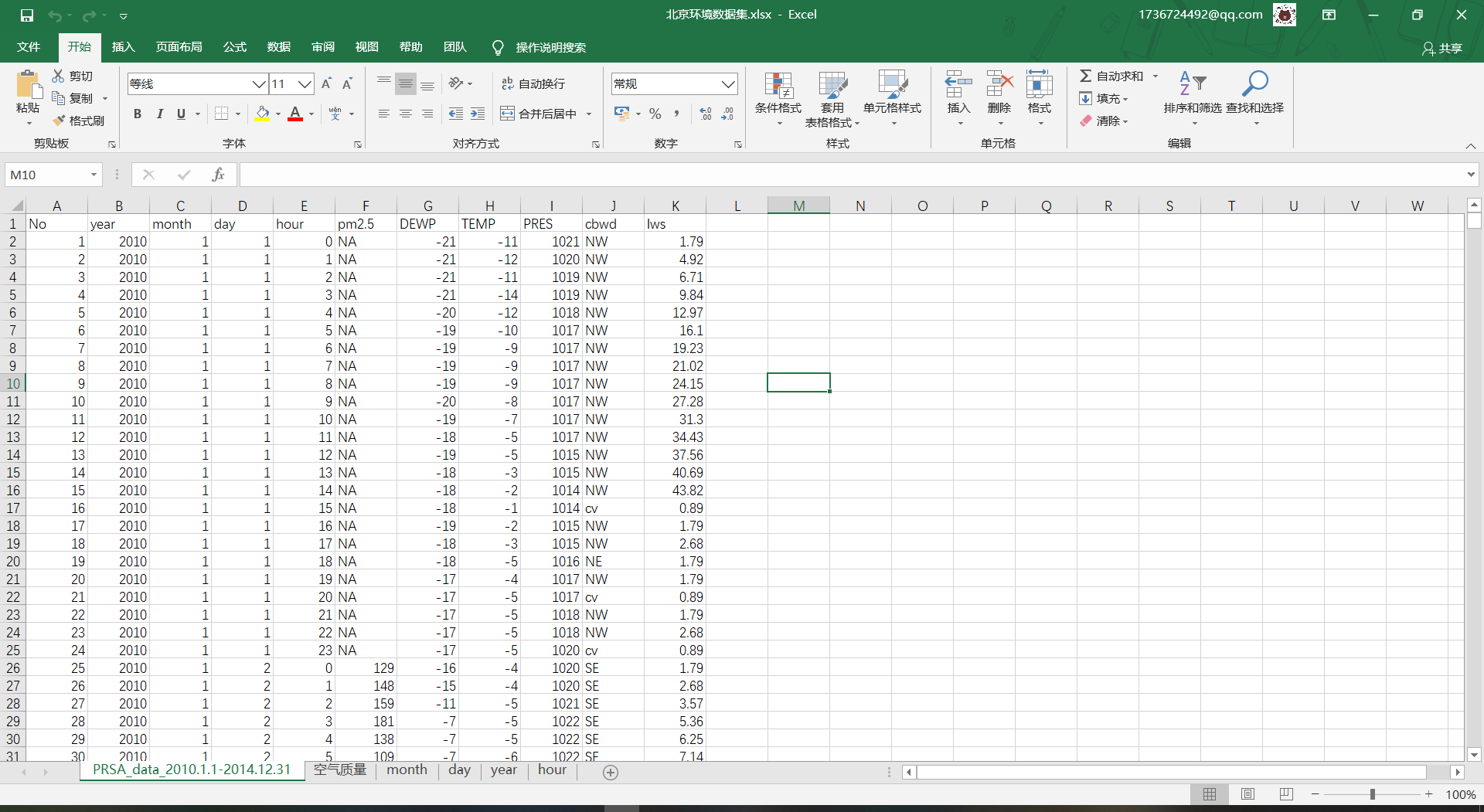
PS: 可使用数据包括(不限于以下数据)：

1. 数据来源多样化（如爬虫、其他网站下载的相关数据、自己整理的相关数据等）；
2. 报告中包含除了基本数据集外的其他相关资料（网页、期刊、报纸、中英文资料等），以作为报告的支撑材料。
3. **数据**

这次作业的主要目的是分析北京市2010~2014年的空气污染情况，所以可以从清华大学的研究机器学习的数据集中得到相应数据：

<http://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data>

下载后为CSV文件，可以利用Excel保存为常用的xlsx格式，以便于查看、修改和导入。

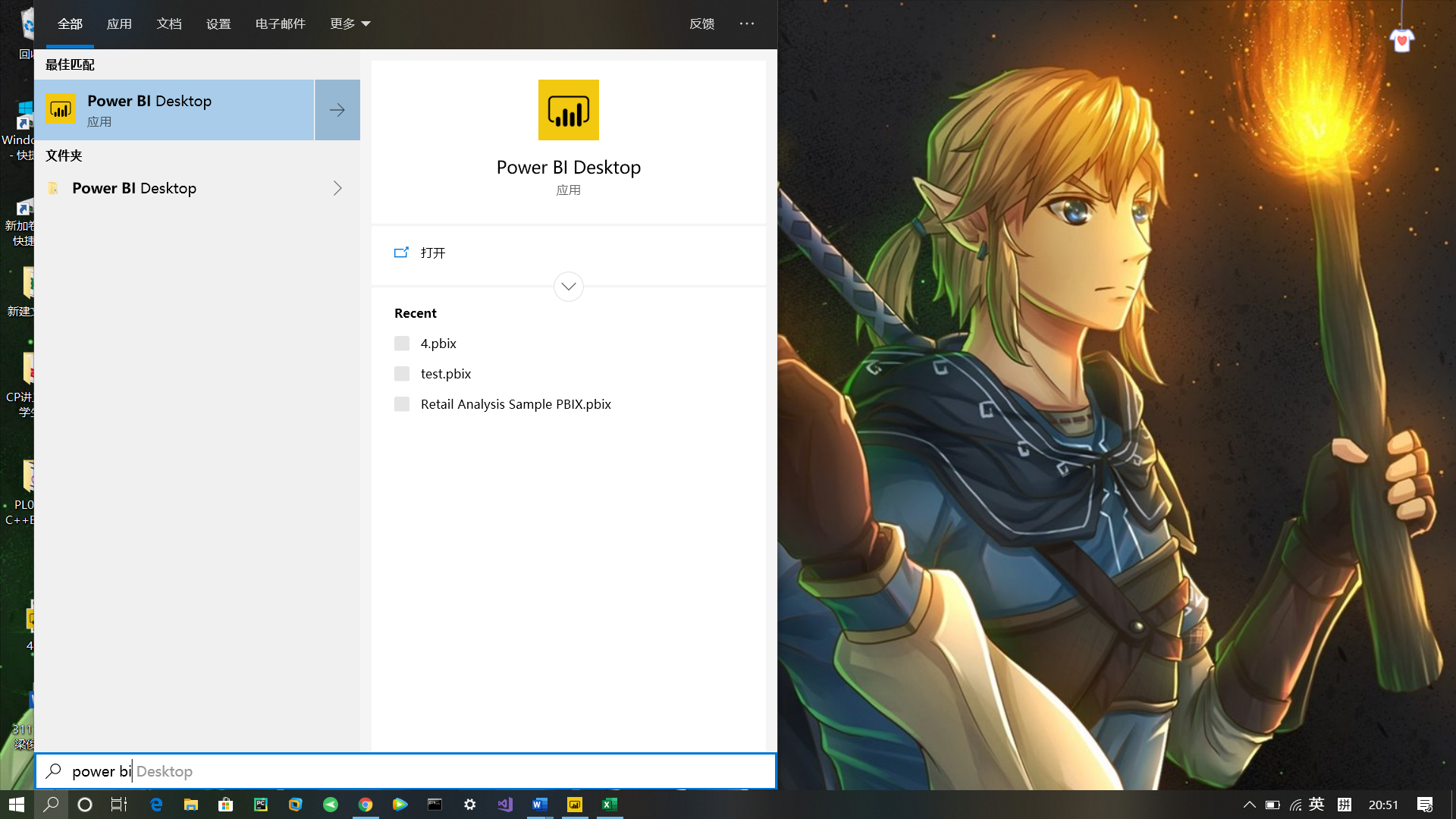


为了便于操作，在原有表格的基础上手动新增了month、year、day、hour、空气质量几个表格，对各自的值汇总了一下。

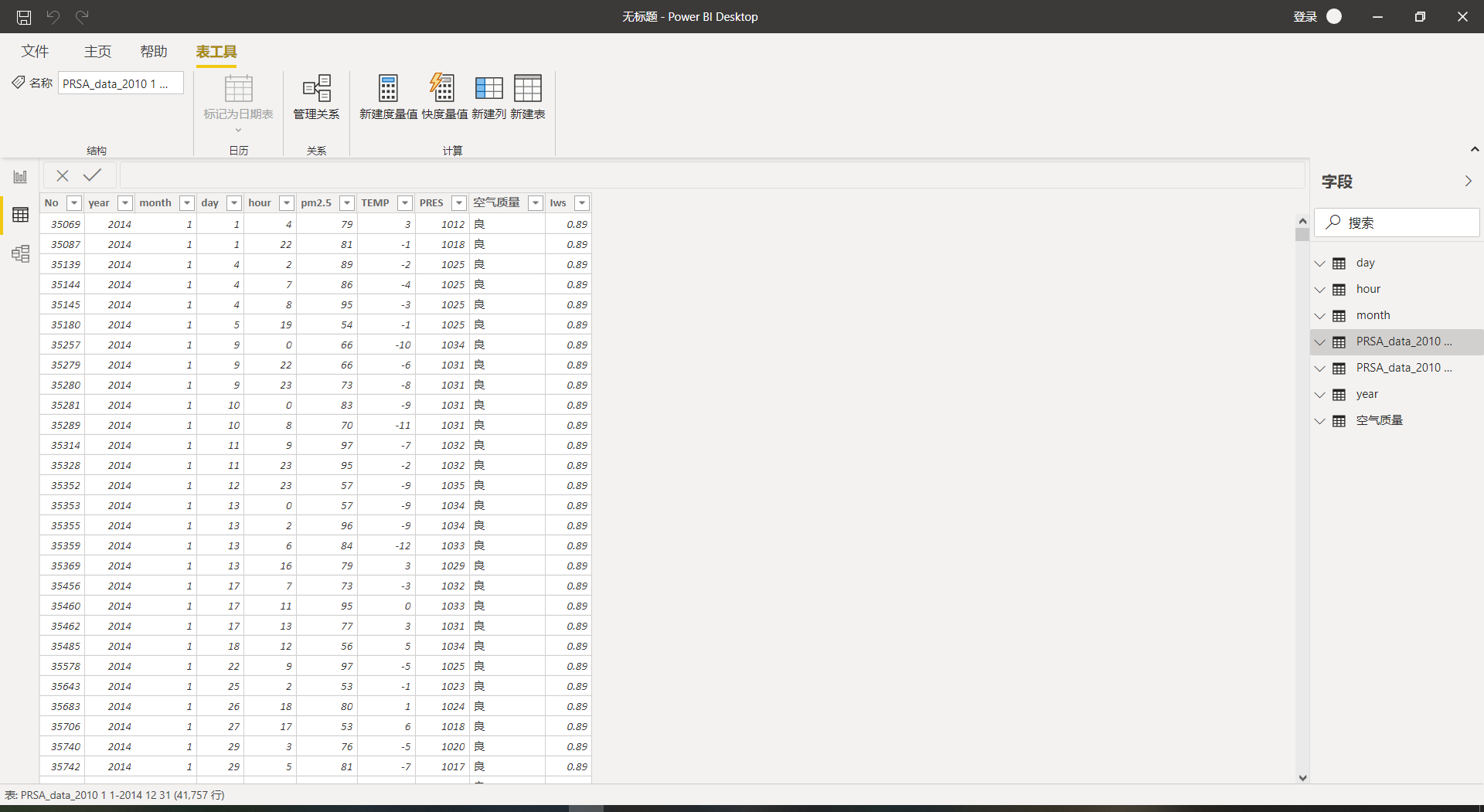
1. **可视化工具**

本次实验采取的可视化工具是Microsoft Power BI,理由是：

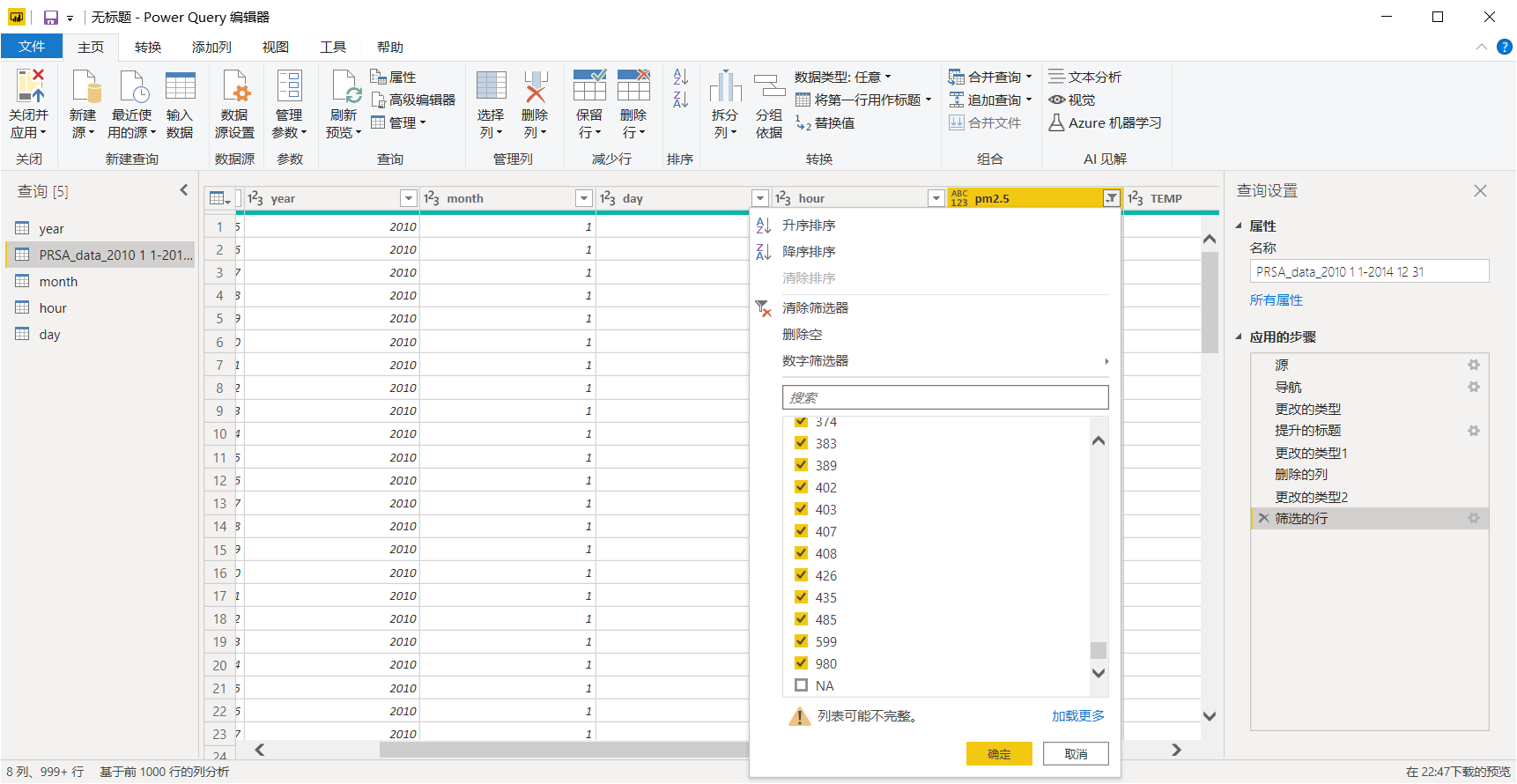
1. 发行者为微软公司，与Windows兼容性好，稳定性强。
2. 是市面上主流的数据可视化工具之一，开发者比较多，有完整的官方中文文档和开发者社区，遇到困难易于寻找解决方案。
3. 支持多种不同的数据可视化方案，易于操作，简洁美观高效。数据来源支持连接xls,csv,txt,SQL Server,Access等多种不同形式数据，Power Query与常用的Excel数据清洗方式较为相似，学习成本较低。
4. Power BI在不发布共享的情况下不收费，对学生群体比较友好。



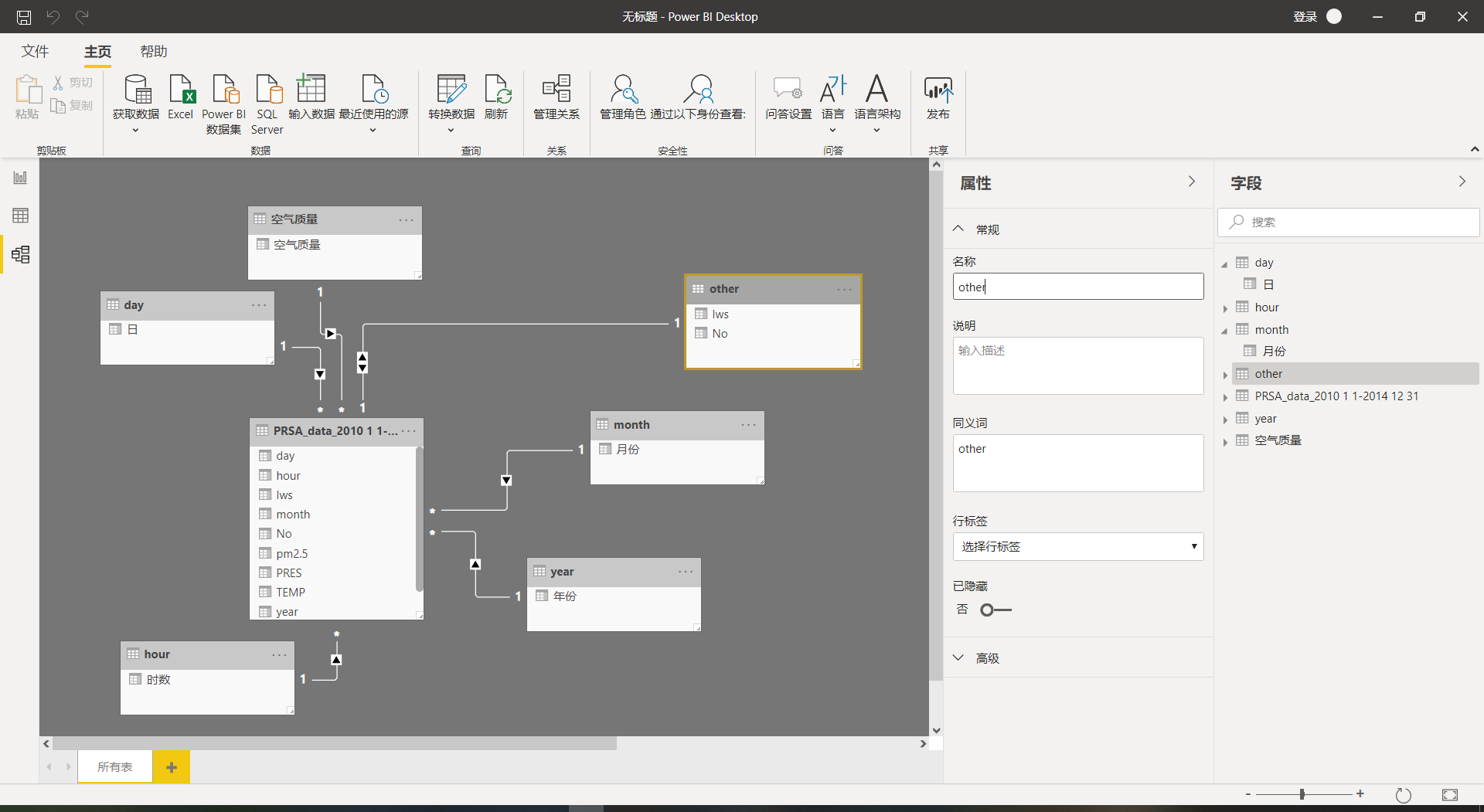
1. **可视化方案或可视化实现过程**
   1. 导入数据并进行数据清洗



删除pm2.5数据缺失的项（pm2.5=NA）



* 1. 处理数据属性并建立联系



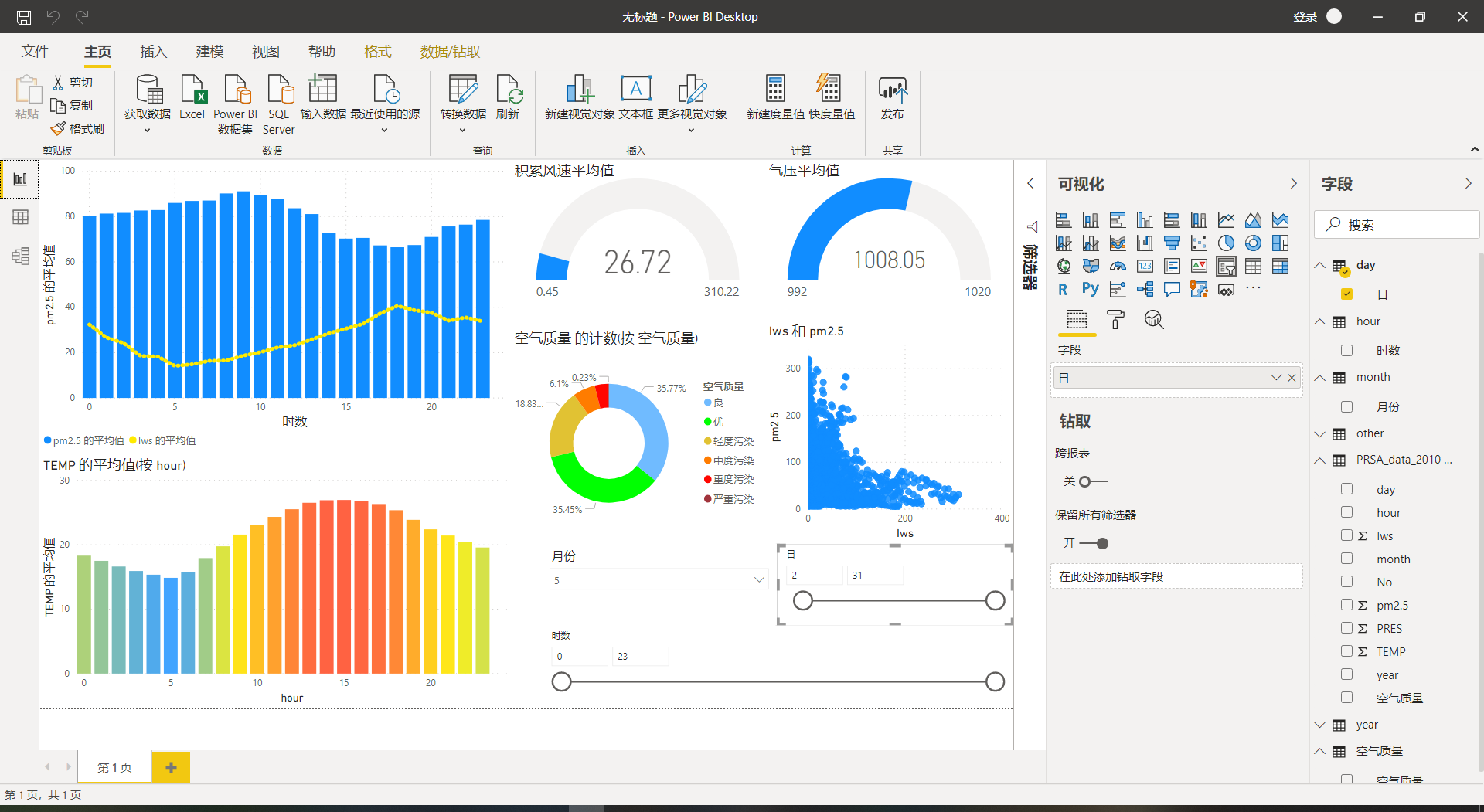
Power BI中最重要的步骤：建立属性连接，会影响到后面可视化不同仪表盘的同步变化。



将pm2.5指数按照规定条件划分等级（建一个新的列）

参考：《环境空气质量指数(AQI)技术规定(试行)》(HJ 633—2012)

* 1. 规划与整理可视化窗口



从工具栏中拖取相应的可视化结构并调整好摆放的位置

左上：pm2.5数值变化表（均值）

左下：温度变化表（均值）

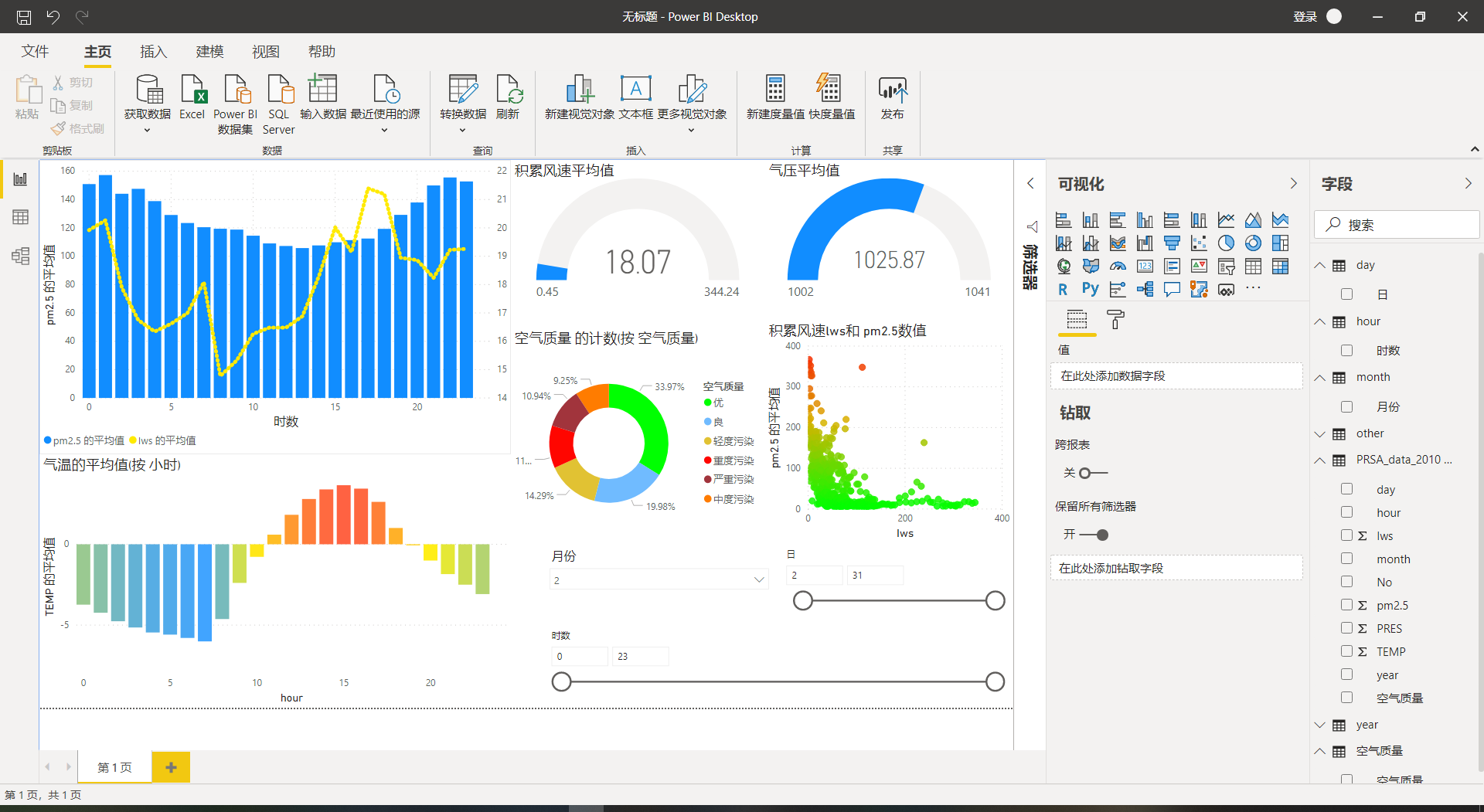
半圆环仪表1：积累风速（均值）

半圆环仪表2：气压（均值）

圆环图：空气质量评价占比（计数）

散点图：pm2.5与积累风速的关联

* 1. 调节参数

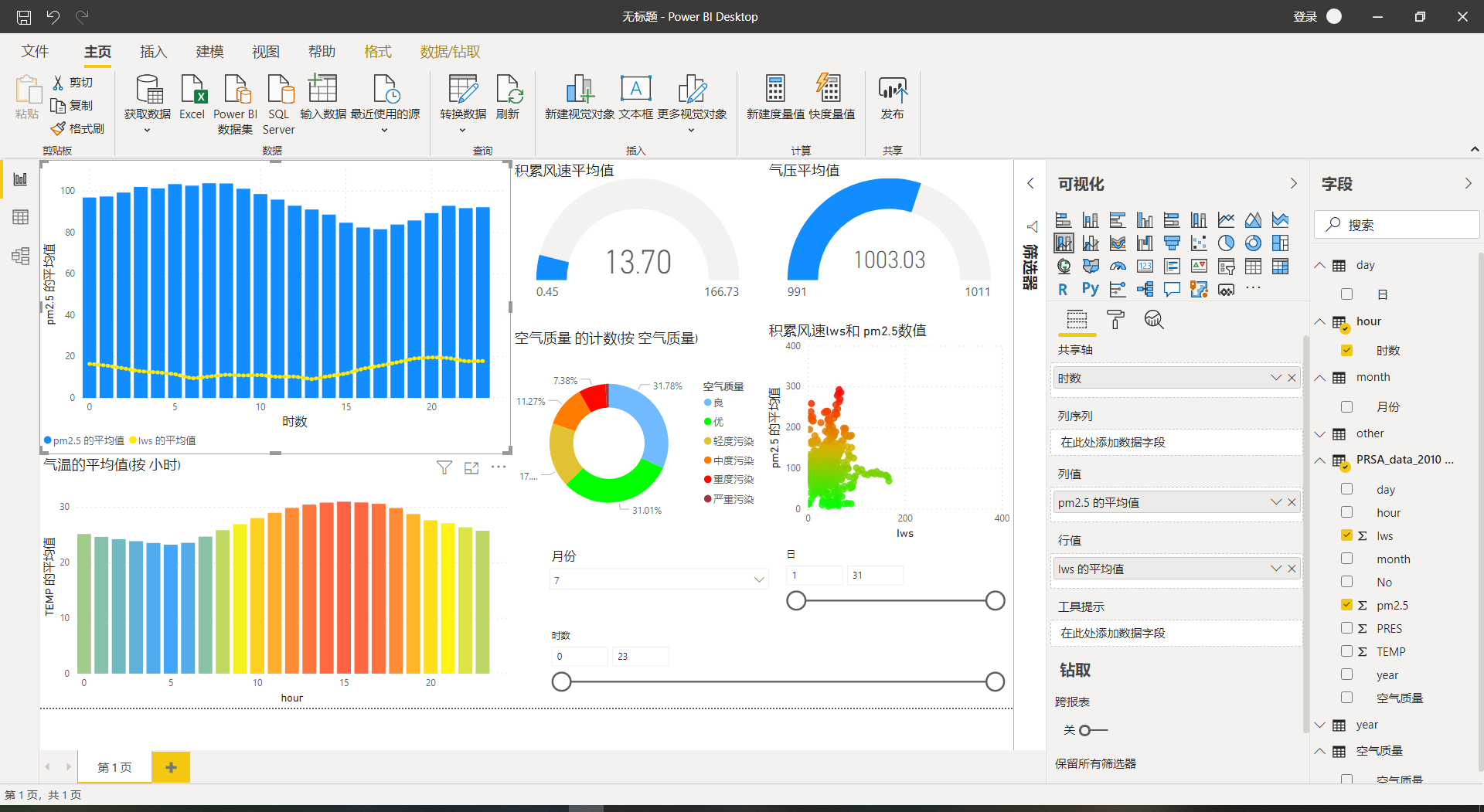


将各个部件的颜色、标题、形式设置好。（上图已设好）

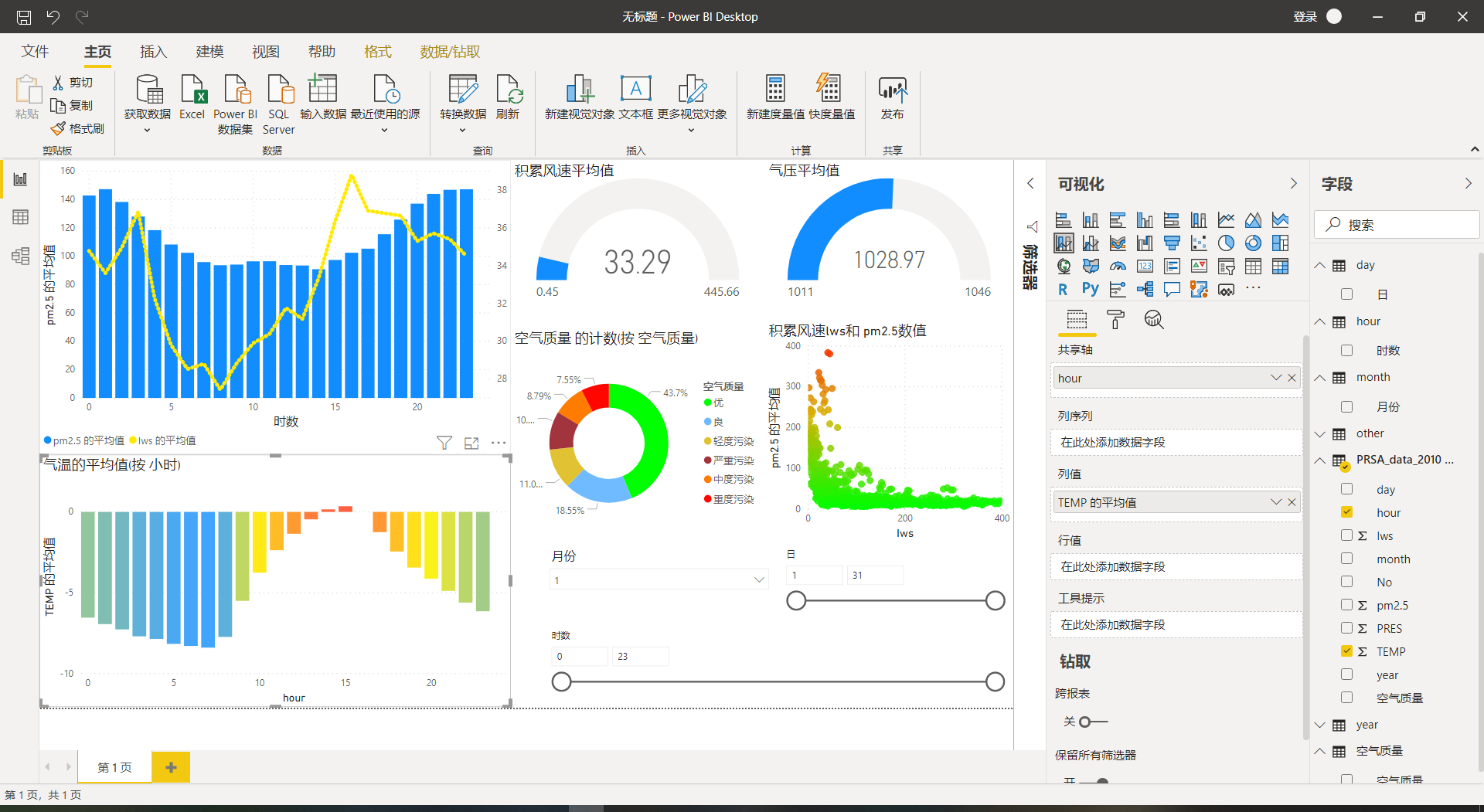
* 1. 检验可视化方案

经过测试，各个部件能够顺利运行，且与设想预期一致。

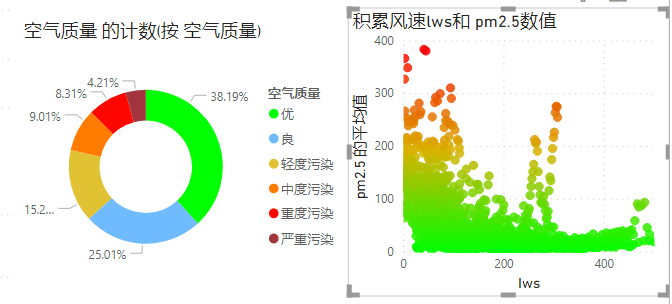
1. **可视化结果**
   1. 关于北京市气温、污染、风速、气压的概览



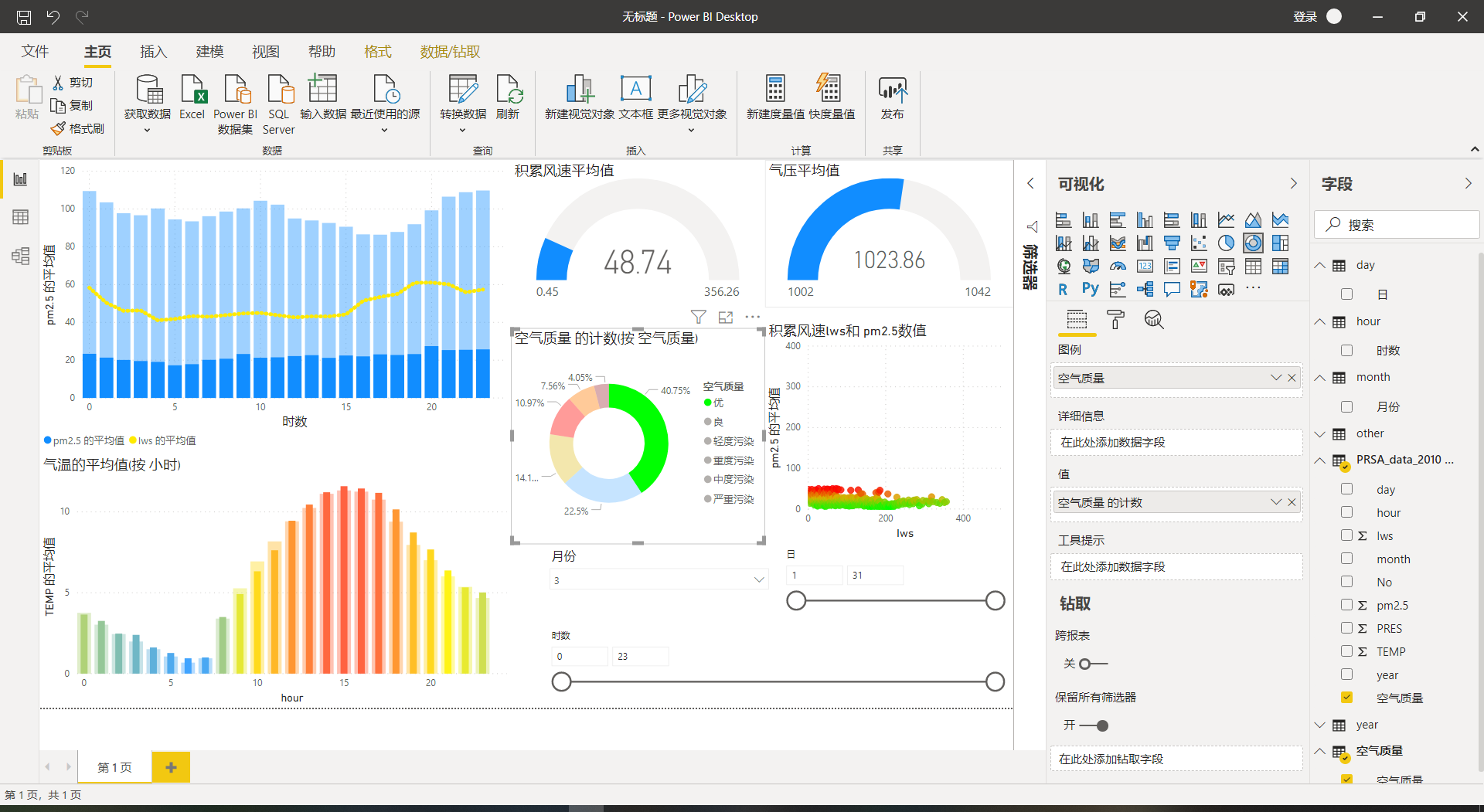
7月的概况（夏季）



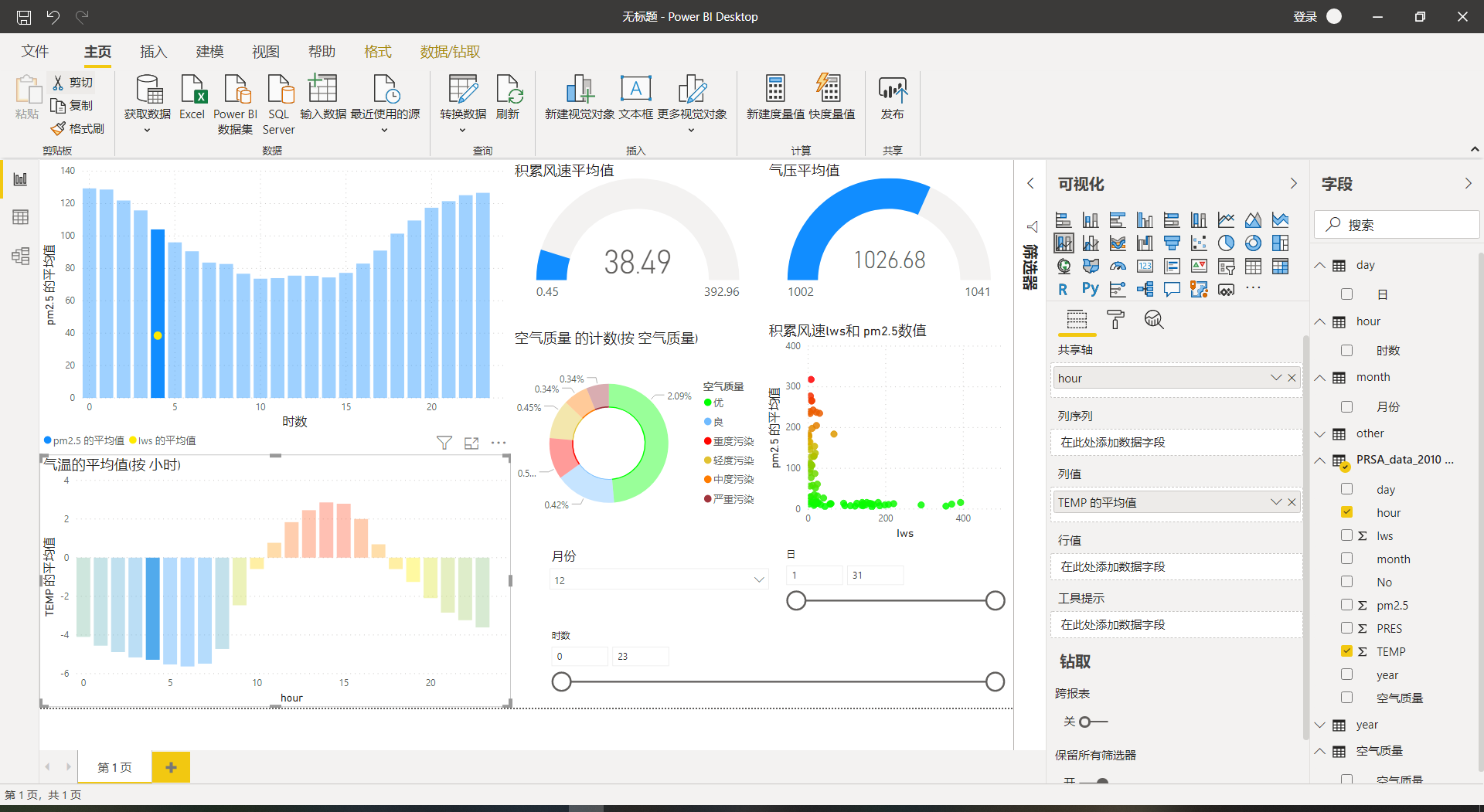
1月的概况（冬季）



空气质量情况（所有）



3月份空气质量为优的情况



12月4点的平均污染情况

概览分析得出结论：

根据所有数据的概览可以看出北京市气温平均在-6（1月）~32度（7月）之间，空气污染方面，冬季污染普遍比夏季有更大波动，冬季平均风速更强，气压方面冬季较高。

* 1. 空气污染与时间的关联分析

根据常识，城市中空气的pm2.5成分主要由汽车尾气和工业废气造成，由（图1）看出北京冬季夜间低于0度，需要实行供暖，而煤炭燃烧也会加重空气污染程度，因此冬季傍晚至夜间的污染程度大幅上升可能与此有关（图2）；在夏季中，由于不需要燃烧化石能源供暖，所以污染指数相对平稳（图3），两个峰值与早晚通勤高峰对应，体现汽车尾气对pm2.5的影响。

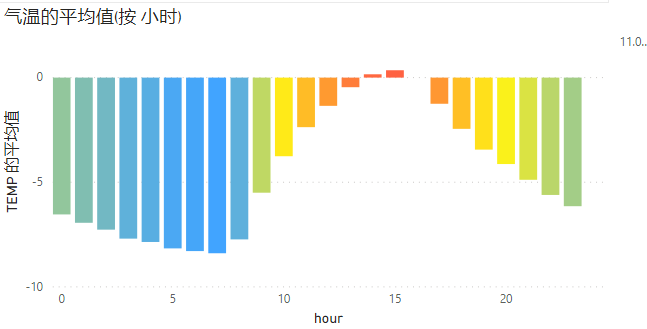


图1——北京市1月平均气温

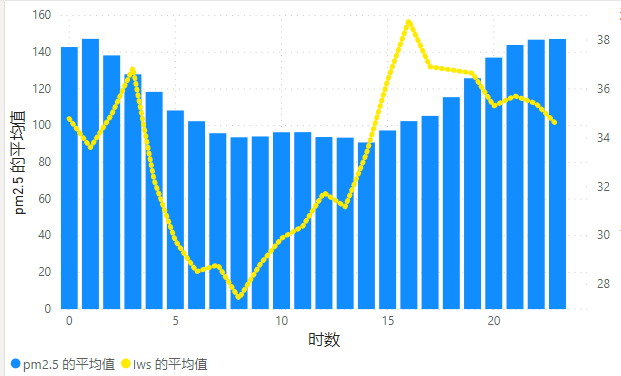
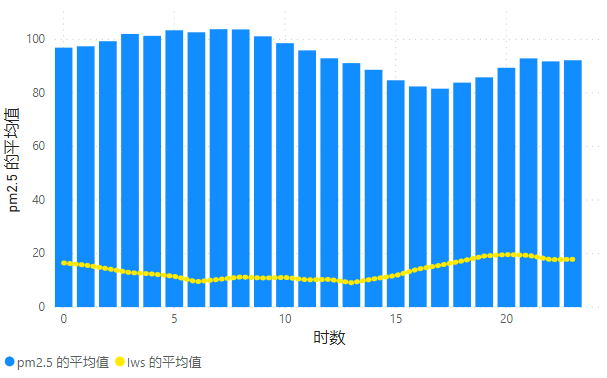


图2——北京市1月平均pm2.5污染值

图3——北京市7月平均pm2.5污染值



* 1. 空气污染与积累风速的关联分析

关于空气污染的另一个假设是：持续的吹风可以驱散雾霾，降低pm2.5的积累量。由于北京市冬季与夏季的pm2.5主要排放途径差异较大（冬季供暖集中于夜间，夏季汽车尾气排放集中于通勤高峰），而且夏季积累风速相对较小（没有“西北风），所以折中用5月的数据来观察。（图4）

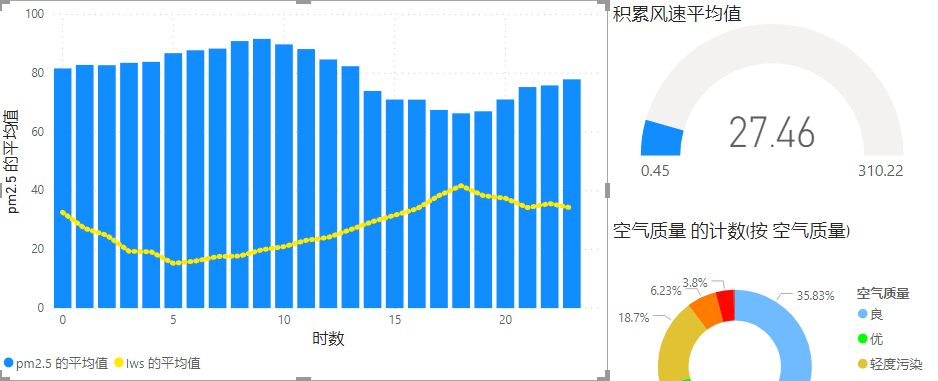


图4——北京市5月pm2.5污染值与累积风速

可以看出空气污染和积累风速之间存在明显的负相关现象。为了进一步验证，我们可以通过所有数据的积累风速与pm2.5散点图来归纳（图5）。

在散点图中，绿色部分代表空气质量优，红色则表示差，显然累计风速与空气质量存在强关联，污染严重的红色点多落在靠近y轴，代表风速较小；污染轻微的绿点多落在x轴延申，表示风速较高。对于风力较强的冬季（图6），效果更加显著

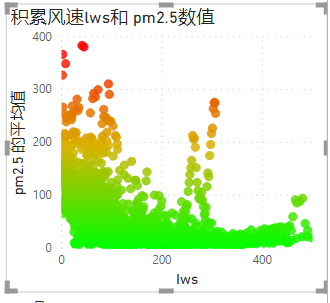


图5

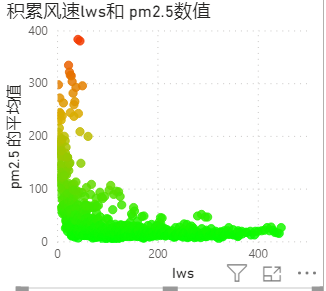


图6

4、总结

根据上文的分析，可以总结出以下结论：

1. 北京市空气质量在冬季波动加剧，原因是一方面由于供暖系统的运作和汽车尾气排放使空气质量总体上差于夏季，但另一方面北京地处蒙古——西伯利亚高压带和太平洋阿留申低压之间，冬季有较强的季风效应（“西北风”），因此在高风量的正面影响和高排放的负面影响下空气质量具有较大波动。
2. 从夏季的情况来看，与冬季相反，既不需要燃烧化石能源供暖也没有“西北风”的掠过，空气质量总体稳定，主要的波动与上下班高峰期吻合，体现了汽车尾气对空气污染。
3. 气压与气温对空气污染的直接影响不明显，而是体现在间接影响上。（气温低——>需要供暖；气压低——>气压差大——>季风猛烈）
4. **心得体会**

在这次数据可视化的大作业中我明白了：不仅要在理论知识上认真学习，更要亲自动手实践。数据可视化这门课程看上去挺简单的，理论上的知识似乎更偏向于心理学和艺术，这也是最难的地方，就是将艺术的美学、数学分析的严谨和计算机编程的灵巧相结合。数据可视化与大数据分析是密不可分的两者，也是将大数据应用到生活的重要接口，通过对心理学、数据挖掘、可视化理论和应用软件Power BI 的学习实践，我又初步掌握了一门基本技能。非常感谢老师的指导！

1. **参考资料**
2. 实验数据来源：清华大学光华管理学院，中心统计科学，机器学习与智能训练中心。<http://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data>
3. 报告格式：<https://max.book118.com/html/2019/0821/7162063111002050.shtm>

3.《环境空气质量指数(AQI)技术规定(试行)》(HJ 633—2012)

4.Power BI官方文档： <https://docs.microsoft.com/zh-cn/power-bi/>

1. 地理与气候原理：<http://blog.xxt.cn/showSingleArticle.action?artId=2964168>

6.环境污染原因参考：<http://bj.bendibao.com/news/20131231/129506.shtm>