吉林大学

开放创新实验报告

组员:

数学学院 2017 级 4 班 杨治恒 10170436

数学学院 2017 级 1 班 茅家勋 10170124

数学学院 2018 级 3 班 梁鲁旭 10180328

指导老师: 李慧君

2019 年 11 月 24 日

1 实验目的

- 根据 HTML5 前端设计，利用 Python 编写爬虫程序获取精准的网页快照，从而理解并巩固 Python 相关数据;
- 基于 PowerBI 对获取的数据进行分析，从而获得有价值的结论.

2 实验内容与方法

2.1 选题——长春市过往天气分析

通过 Python 爬虫程序，爬取长春市 2011 年 1 月 1 日至今的天气，整理成 Excel 表格加以分析，并辅以图表实现数据可视化.

2.2 环境与工具

本实验基于 Python 3.6.5 版本环境，并利用 PyCharm 2017.2 版本进行. Python 版本信息如代码 2.1.

```
Python 3.6.5 (v3.6.5:f59c0932b4, Mar 28 2018, 17:00:18) [MSC
v.1900 64 bit (AMD64)] on win32
```

代码 2.1 Python 版本信息

实验爬虫部分所调用到的库如代码 2.2.

```
import requests  # 用于获取网页 HTML 代码
import re  # 利用正则表达式筛选数据
import datetime  # 爬虫爬取时显示时间用
```

代码 2.2 调用库的列表

本次实验用到的其它工具有 Excel 和 PowerBI.

3 实验步骤

3.1 爬虫爬取天气数据的实现

3.1.1 利用 requests 库请求网页 HTML 代码

想要了解长春过往的天气，就需要找到一个能记录过往天气的网站进行爬取

数据的准备. 经过查阅、对比后发现,“天气后报”网准确、完整地记录了各省市往年每日昼夜的天气状况、气温、风向等信息, 并清晰地按日期记录在表格中. 网址如代码 3.1.

```
'http://www.tianqihoubao.com/'
```

代码 3.1 “天气后报”的网址

为了爬取长春市往年的天气情况, 我们需要了解待爬取的网页有哪些, 即该域名下什么样的链接里包含了我们想要的数据.

经过浏览, 我们不难发现, 此站点在 lishi/changchun/month/标签下包含有按月分的长春天气的表格, 且对于 yyyy 年 xx 月的天气, 记录于 yyyymm.html 的网页中. 例如想要翻阅 2011 年 1 月的天气, 可以在代码 3.2 的链接中找到.

```
'http://www.tianqihoubao.com/lishi/changchun/month/201101.html'
```

代码 3.2 “天气后报”2011 年 1 月天气的链接

因此, 要想爬取长春 2011 年 1 月至今 (2019 年 11 月) 的天气, 首先需要用一个 list 存放我们待爬取的网站. 该 list 生成过程如代码 3.3.

```
urls = []
for year in range(2011,2020):
    for month in ['01', '02', '03', '04', '05', '06', '07',
                  '08', '09', '10', '11', '12']:
       if str(year) + str(month) != '201912':
           date_url = 'http://www.tianqihoubao.com/' + \
               'lishi/changchun/month/' + str(year) + \
               str(month) + '.html'
           urls.append(date_url)
```

代码 3.3 待爬取网站的获取

有了待爬取网站以后, 我们利用 for 循环遍历所有网站即可实现爬取. 而在利用 requests 库爬取之前, 还需要做一件准备工作, 那就是配置请求头 requests headers. 请求头的配置如代码 3.4.

```
headers = {
    'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64;' + \
'x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/' + \
'77.0.3865.120 Safari/537.36',
    'Host':'www.tianqihoubao.com',
    'Accept': 'text/html,application/xhtml+xml,' + \
'application/xml;q=0.9,image/webp,image/apng,' + \
```

```
'*/*;q=0.8,application/signed-exchange;v=b3',
    'Accept-Language': 'zh-CN,zh;q=0.9',
    'Accept-Encoding': 'gzip, deflate',
    'DNT': '1',
    'Connection': 'keep-alive',
    'Upgrade-Insecure-Requests': '1',
    'Cache-Control': 'max-age=0',
'Cookie': 'Hm_lvt_f48cedd6a69101030e93d4ef60f48fd0=' + \
'1574492500; ASP.NET_SessionId=n0ljqp55g1k5yauc4pzfhx45; ' + \
'__tins__4560568=%7B%22sid%22%3A%201574518508687%2C%20%' + \
'22vd%22%3A%201%2C%20%22expires%22%3A%' + \
'201574520308687%7D; __51cke__=; __51laig__=1'
}
```

代码 3.4 请求头的配置

至此已经完成了爬取的准备工作，可以开始爬取了. 为了能够清楚地了解爬取的进度，我们可以利用 datetime 库获取当前时间并与爬取 url 一起打印出来. 爬取的代码如代码 3.5.

```
for url in urls:
    curr_time = datetime.datetime.now()
    time_str = datetime.datetime.strftime(curr_time,
                                          '%Y-%m-%d %H:%M:%S')
    print(time_str, '  ', url)
    try:
        response = requests.get(url, headers=headers,
                                verify = False)
        if response.status_code == 200:
            text = response.text
    except Exception as e:
        print(e)
```

代码 3.5 用 requests 库进行 html 网页的爬取

3.1.2 利用 re 库清洗数据

在上一节爬取到的数据为 HTML 代码，这里以 2011 年 1 月的数据为例讨论. 2011 年 1 月天气如图 3.1，其对应的代码见附件 201101.html. 要在如此繁杂的代码中清洗出我们所需要的数据，我们首先得以这个例子作为讨论来发现这些代码的规律，从而设计正则表达式，实现提取出有效数据.

图 3.1 “天气后报”中 2011 年 1 月的天气表

不难发现，我们需要提取出的有效数据即是这张表格中的数据．打开对应的 html 代码，我们发现，在<table>…</table>标签内的数据对应了整张表格，且整个网页仅这张表格含有此标签．在此标签内，每个<tr>…</tr>标签划分了行，而在每一行内，<td>…</td>标签划分了每一格．而在每一行的日期部分，含有…的链接型标签，这是需要被清洗掉的内容．

基于上述讨论，我们可以设计出如下的四条正则表达式．

```
pattern1 = re.compile(r'<table(.*?)</table>', re.S)
pattern2 = re.compile(r'<tr>(.*?)</tr>', re.S)
pattern3 = re.compile(r'<td>(.*?)</td>', re.S)
pattern4 = re.compile(r'>(.*?)</a>', re.S)
```

代码 3.6 设计清洗数据用的正则表达式

在对数据进行清洗之时，我们可以”顺带”地将所获得的数据放入一个临时的字典中．考虑到最终获得的数据可以用 csv 的格式导出并手动另存为 excel 格式，我们可以将字典里的内容转化为以逗号分隔的字符串并添加到一个统一的列表里．考虑到网页中昼、夜的数据是以”/”分隔的，所以我们可以在添加到列表的时候用 replace 方法将”/” 替换为”, ”．此分析对应的代码如代码 3.7，且是嵌套在代码 3.5 的 for 循环下面的．

```
table_data = re.findall(pattern1, text)[0].replace('\r\n', '')
day_data = re.findall(pattern2, table_data)
for day in day_data:
    if '日期' not in day:
```

```
            day = day.replace(' ', '')
            day_infos = re.findall(pattern3, day)
            for i in range(len(day_infos)):
                if '</a>' in day_infos[i]:
                    day_infos[i] = re.findall(pattern4,
                                              day_infos[i])[0]
            weathers.append({'date': day_infos[0], 'weather':
                 day_infos[1], 'temperature': day_infos[2],
                'wind': day_infos[3]})
    for weather in weathers:
        a_line = weather['date']+','+weather['weather']+','+\
                 weather['temperature']+','+weather['wind']
        w_line.append(a_line.replace('/',','))
```

代码 3.7 清洗数据

3.1.3 导出数据

在上一节中，我们得到了清洗后的数据，并做成了一个数据由逗号隔开的字符串做成的列表. 现在我们可以将其导出到一个 csv 文件中. 考虑到可能会存在重复数据，因此我们可以用一个类似 set()但不会打乱顺序的方法将列表转化为集合型列表. 最后将这个列表用换行符"\n"连接并写入文件即可. 导出数据的代码如代码 3.8.

```
new_w_line = []
for ww in w_line:
    if ww not in new_w_line:
        new_w_line.append(ww)
new_info = '\n'.join(new_w_line)
with open('weather_by_day_night.csv', 'w') as f:
    f.write(new_info)
    f.close()
```

代码 3.8 导出数据

导出以后获得文件 weather_by_day_night.csv. 用 Excel 打开并另存为即可得到一张表 weather_by_day_night.xlsx. 这两份文件见附件.

以上过程完整代码见附件 WeatherCrawler.py.

3.2 利用 PowerBI 处理数据

3.2.1 将数据导入 PowerBI

将天气数据导入 PowerBI，首先，我们先对数据进行编辑处理，将 Excel 表中的数据转化为合适的数据类型．然后在软件上方建模行列的新表里，利用 CALENDAR 函数建立智能时间表，时间从 2011 年 1 月 1 日到 2019 年 11 月 23 日，创建其所对应的季度，月份，年信息．如图 3.2.

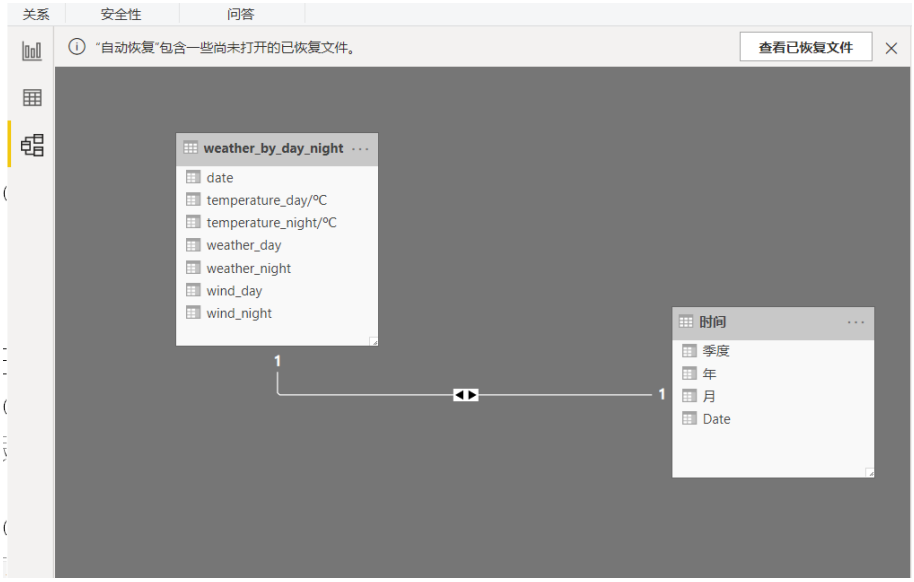

图 3.2 PowerBI 导入数据示意图

3.2.2 建立起两个表的关系

举个例子，看 11 年到 19 年来每种季度下白天温度天数分布情况，我们就可以以白天温度为轴，天数为计数单位形成的柱状图如图 3.3.

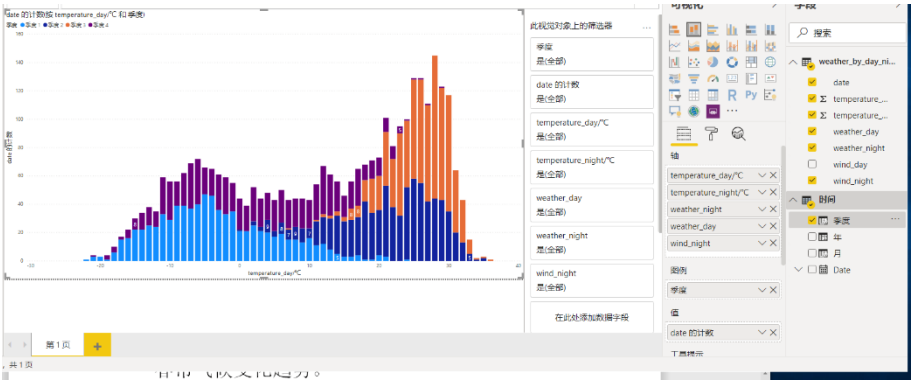

图 3.3 白天温度天数分布状况柱状图生成示意图

我们还可以把多项数据放在一个轴内，点击图像的钻取功能就能够实行数值

的切换. 如图 3.4.

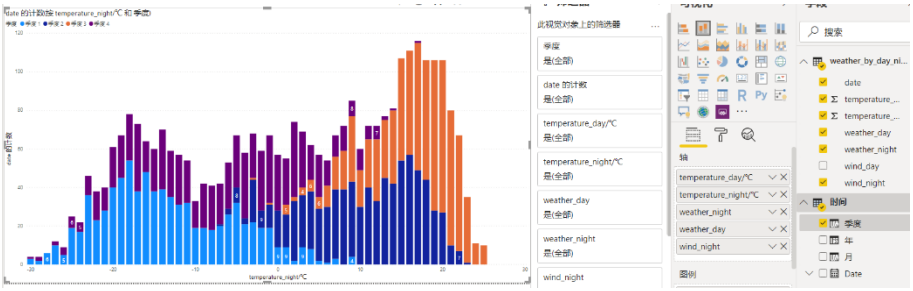

图 3.4 数值切换实现示意图

全部操作在这不再赘述，实验内容即可体现全部图像.

3.2.3 数据分析

最后根据不同图表和图像利用各种数据分析方法进行气象数据分析，并结合相关气候知识归纳长春市气候变化趋势.

4 实验结果

4.1 温度与时间

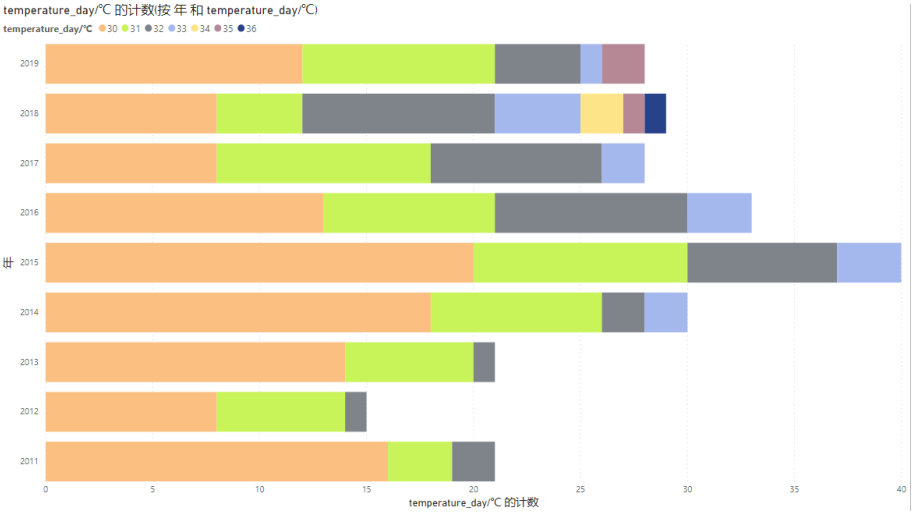

图 4.1 各年白天极端温度天数（堆积条形图）

根据图 4.1 可直接得到:

自 2011 年以来长春市几乎每年的气温峰值都在升高. 且高温天气持续的时间在 2015 年前一直增加并在 2015 年达到时长峰值，后又逐渐减少并趋于稳定. 长春市总体夏季气温呈现变暖的趋势.

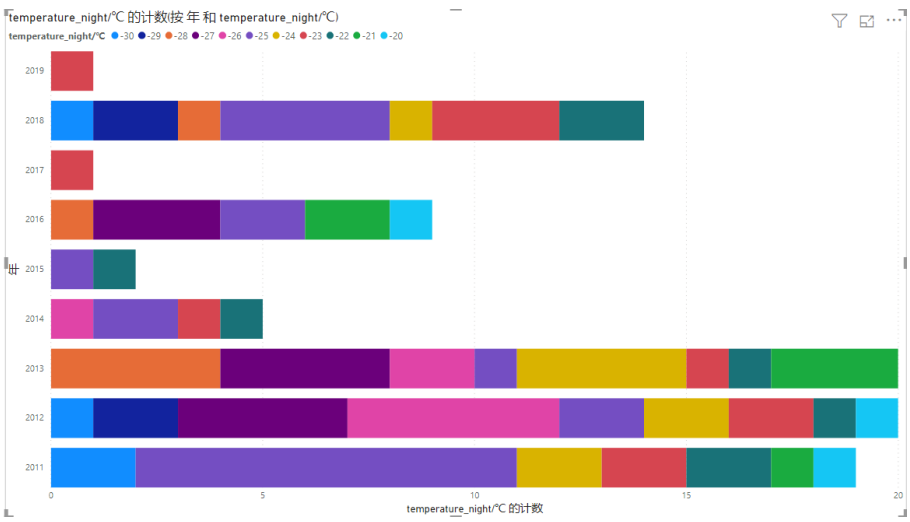

图 4.2 各年夜晚极端温度天数（堆积条形图）

根据图 4.2 直接可以得:

自 2011 年以来长春市冬季最低气温大致呈持续升高趋势，极寒时段也较快缩短，虽然 2018 年仍然有较长时间低温天气，但总体来看长春冬季在逐渐变暖.

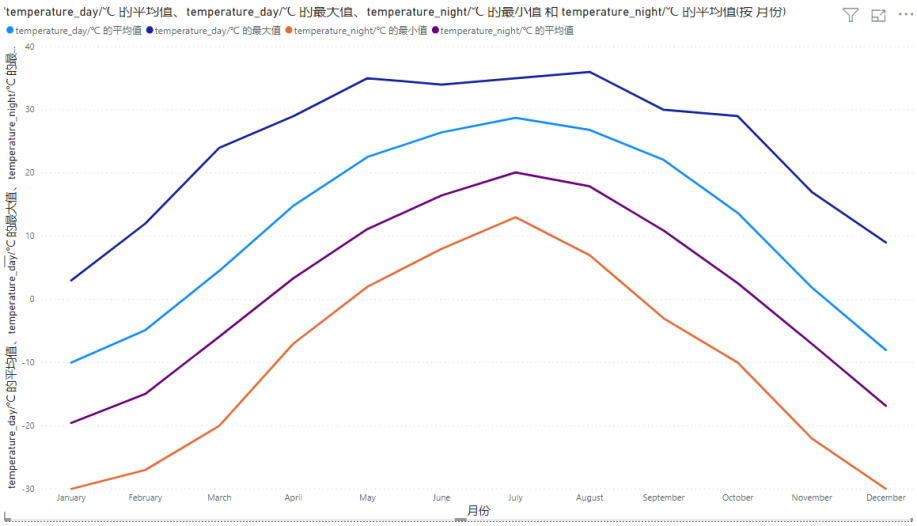

图 4.3 一年中各月的平均日夜温度与极端日夜温度（折线图）

根据图 4.3 可直接得到:

长春是一个昼夜温差较大的城市，一年四季都维持在 10 度到二十度左右. 另外作为北方城市，长春冬季较长，一年有大约四个月都在零下，最低时甚至达到零下 30 摄氏度，夏季最高气温也一般不超过 35 摄氏度，气候偏冷.

4.2 雨雪与时间

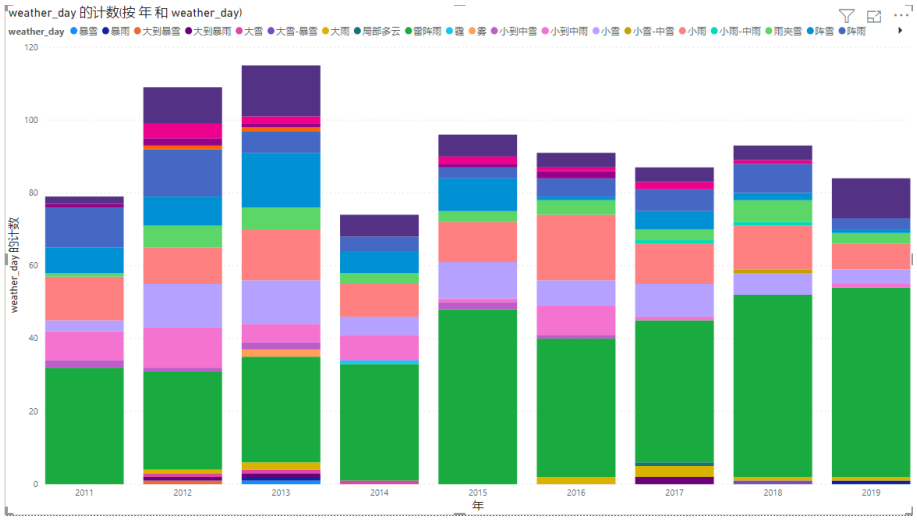

图 4.4 各年各类雨雪天数（堆积柱状图）

根据图 4.4 可直接得到:

长春市降雨量不多，且一般以小雨或雷阵雨方式出现. 长春市降雪较为充足，甚至能占到雨雪总量的一半. 长春市近九年的降水量呈现一个较为平稳的水平，未出现较明显波动.

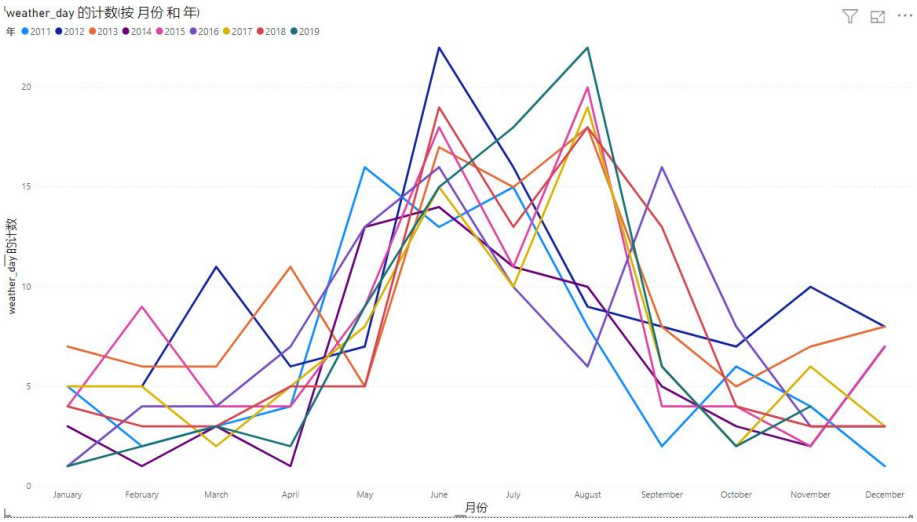

图 4.5 各年各月雨雪天数（折线图）

根据图 4.5 可直接得到:

长春的雨季峰值集中在六月到八月之间，最多时能达到每月 20 天，其次冬季

降雪量分别在上半年和下半年也会产生两个坡度较小的峰值. 总的来看, 近九年长春每年雨雪总量随月份的变化呈一个变化趋势保持稳定, 峰值在时间段内来回波动的形式.

4.3 风速、风向与时间

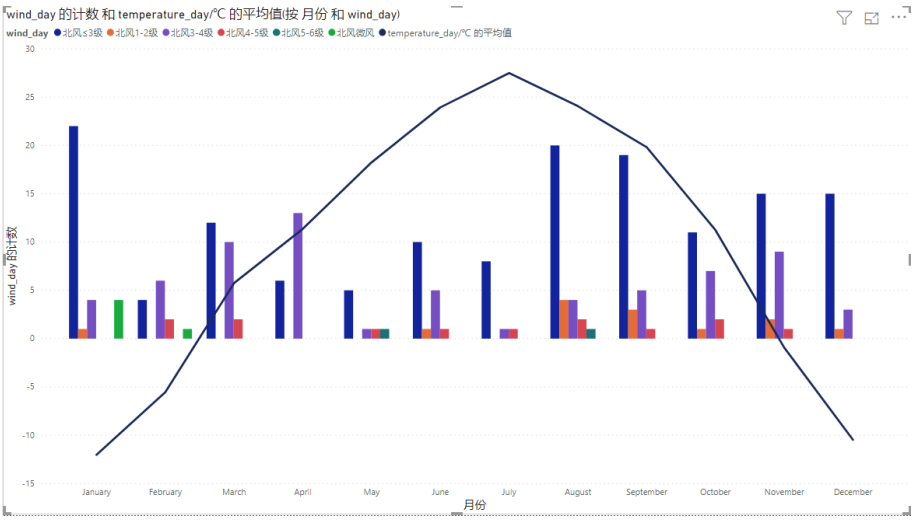

图 4.6 一年内各月份北风风速风向天数（堆积条形图）

根据图 4.6 可直接得到:

长春在一年内各月份都会刮北风, 但风速且频率较大的时间集中在秋冬两季. 但总的来说北风在各月份频率都较低且风速较小, 北风强度随季节性变化不明显.

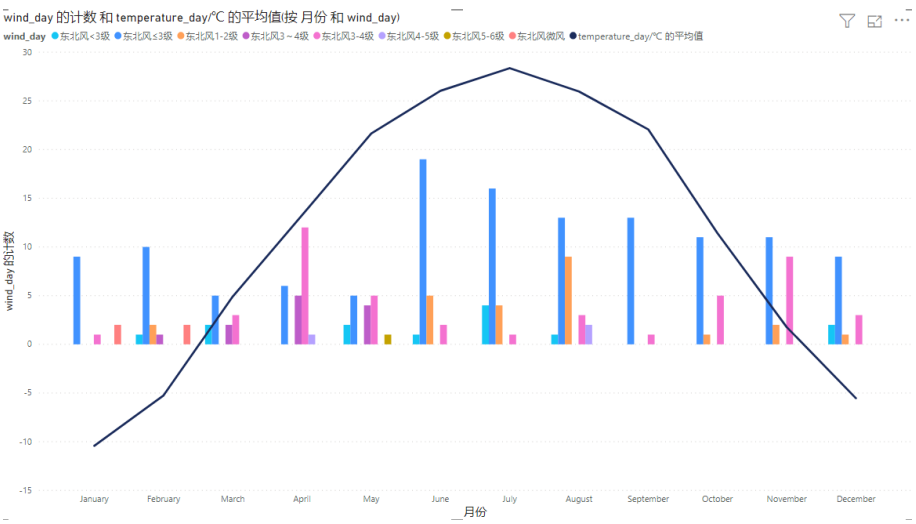

图 4.7 一年内各月份东北风风速风向天数（堆积条形图）

根据图 4.7 可直接得到:

长春市东北风的频率明显非常小且都以弱风为主，与北风类似，不是主流风向但全年都有.

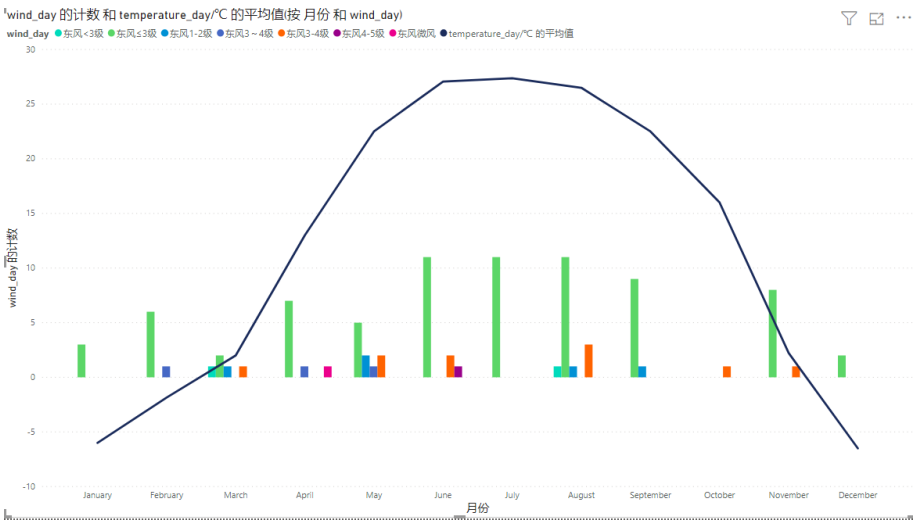

图 4.8 一年内各月份东风风速风向天数（堆积条形图）

根据图 4.8 可直接得到:

长春市刮东风的天很少且都以小于三级的弱风为主，但仍然出现在集中在春夏两季的季节特征.

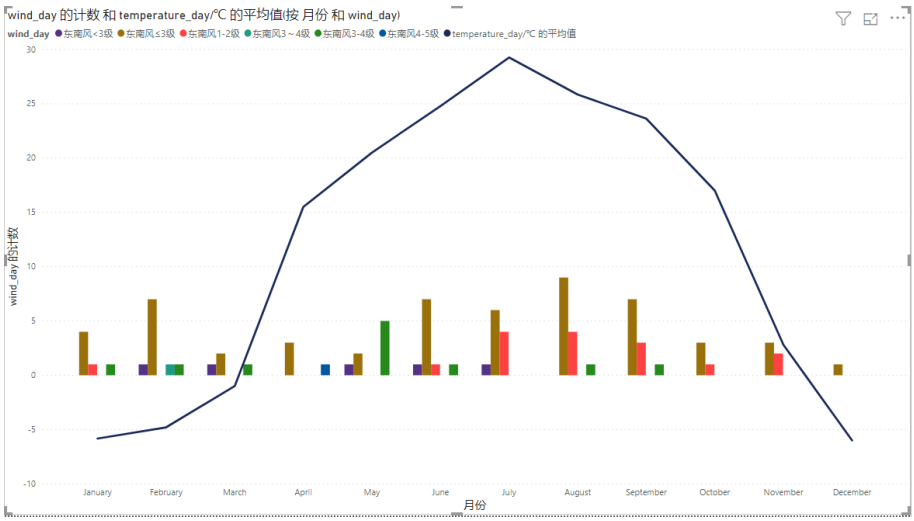

图 4.9 一年内各月份东南风风速风向天数（堆积条形图）

根据图 4.9 可直接得到:

长春市每年偶尔刮风速较小的东南风，随季节和温度变化不明显.

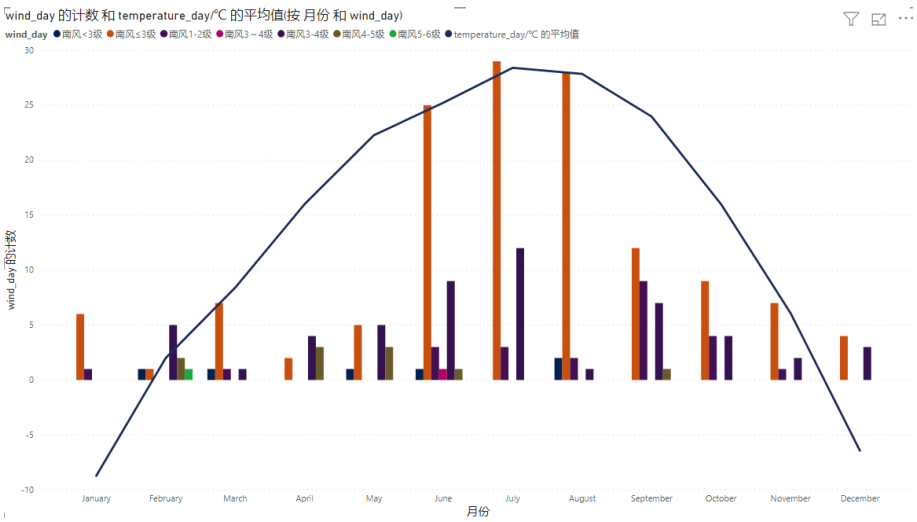

图 4.10 一年内各月份南风风速风向天数（堆积条形图）

根据图 4.10 可直接得到:

长春市刮南风的频率与风速随季节性变化非常明显，频率较高时只出现在夏季且峰值与其他三季相比差异很大. 但南风通常以小于三级的弱风为主.

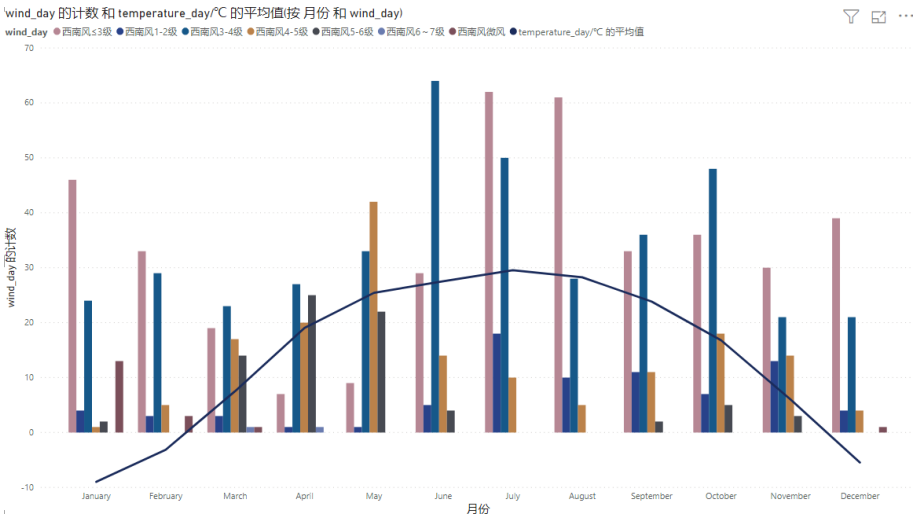

图 4.11 一年内各月份西南风风速风向天数（堆积条形图）

根据图 4.11 可直接得到:

长春市出现频率最大且强度也最大的风向当属西南风，与其他风向相比西南风一年四季频率都较高，但仍能明显看出西南风在夏季时风速与频率都达到峰值. 西南风的强度随季节性和温度变化较为明显.

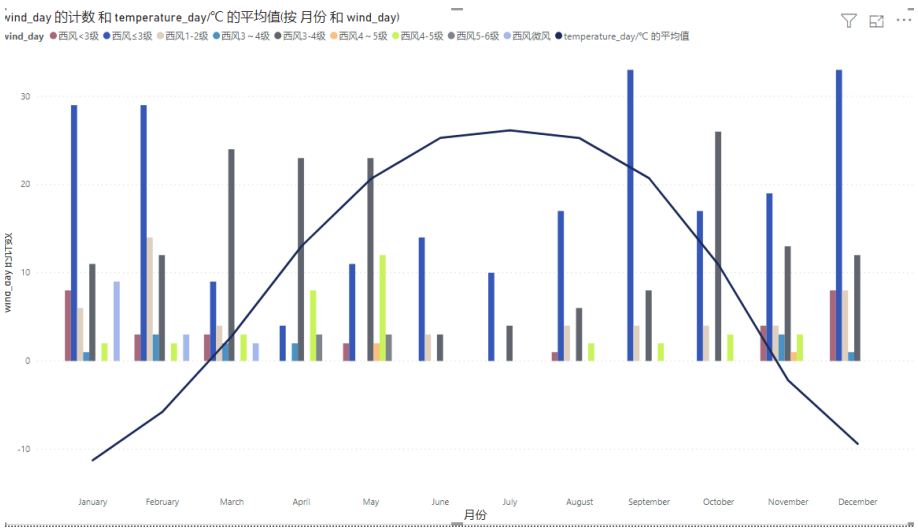

图 4.12 一年内各月份西风风速风向天数（堆积条形图）

根据图 4.12 可直接得到:

长春市一年四季都有西风出现且风速不强，且也稍集中在春秋冬三季，且除夏季外频率一直不算低但随温度变化不明显.

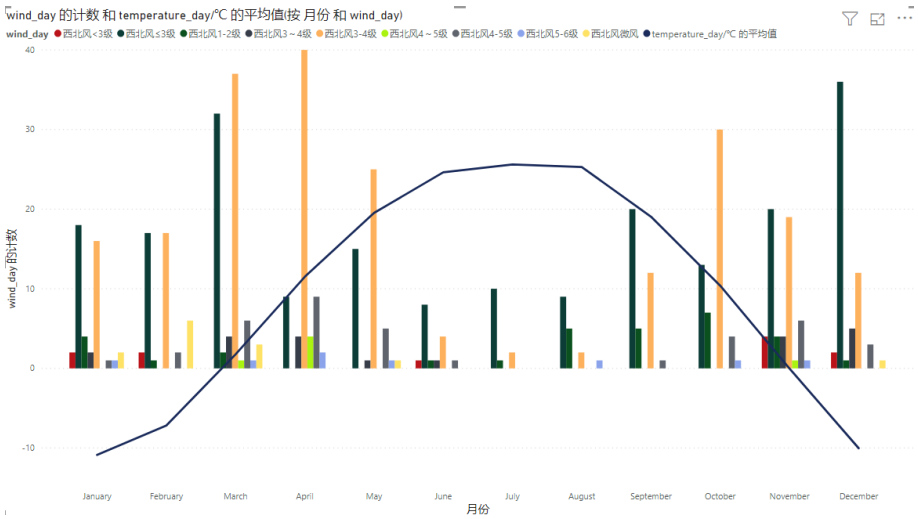

图 4.13 一年内各月份西北风风速风向天数（堆积条形图）

根据图 4.13 可直接得到:

长春市集中在春秋两季刮西北风且风速常见 3~4 级，气温处于零度左右及零下时常刮西北风. 西北风强度和频率随季节性变化较为明显.

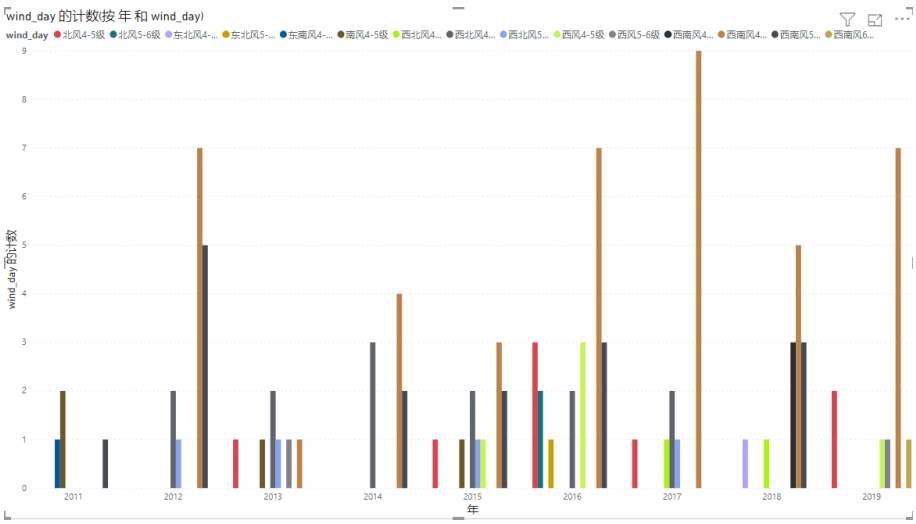

图 4.14 各年极端大风天数（柱状图）

根据图 4.14 可直接得到:

长春市每年的极端大风天气多为刮西南风时出现，其次为西北风和北风等. 长春市的极端大风天气每年都会出现但不会维持很久，且多为 4~5 级大风. 总的来看，长春市每年极端大风天气出现的频率没有一个明显的变化规律，但近些年大风天明显变多.

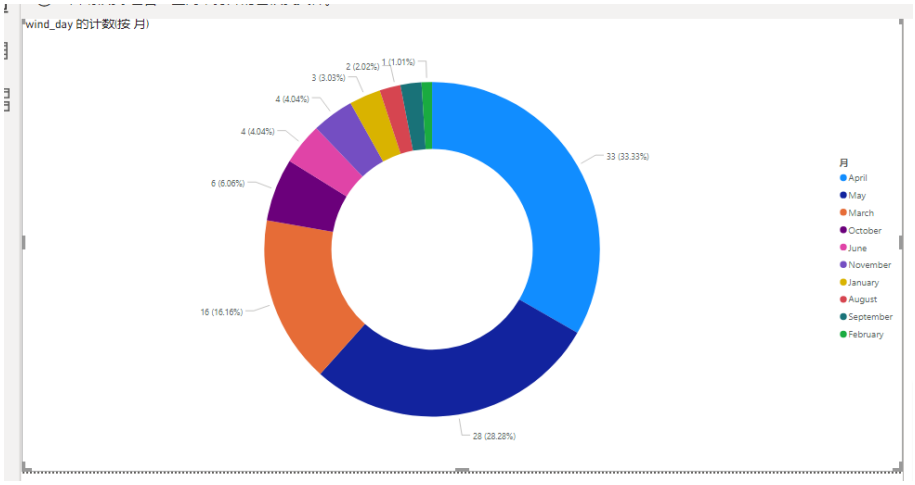

图 4.15 极端大风天在各月份对比（环形图）

根据图 4.15 可直接得到:

长春的大风天集中在三到五月，其次在十一月到一月，也就是集中在冬春两季.

4.4 结论

综合以上图像，结合相关地理知识，我们可以得出以下分析:

长春市介于东部山地湿润与西部平原半干旱区之间的过渡带，属温带大陆性湿润气候类型. 由于长白山地的阻挡，削弱了夏季风的作用，导致春季干旱多风，夏季温暖短促，西部和北部为地势平坦的松辽平原，西伯利亚极地大陆气团畅通无阻，故秋季晴朗温差大，冬季严寒漫长.

5 参考文献

- Eric Matthes: 《Python 编程 从入门到实践》人民邮电出版社 2016.7.