城市天气数据的分析与预测项目报告

钟阅，罗庆典

1、问题背景

在当今社会，天气数据在城市规划、农业生产、能源管理等领域发挥着重要作用。通过深入分析天气变化规律并构建有效的预测模型，可以帮助人们更好地应对气候变化、优化资源配置，并支持多领域的科学决策。本项目选取了中国四个具有代表性的城市——北京、上海、海口和拉萨，基于其 2011 至 2021 年的历史天气数据，尝试揭示城市天气特征，探索潜在规律，并预测未来天气趋势。这不仅为理解不同地域气候特性提供了数据支持，也为应对日益复杂的气候变化提供了科学依据。

选择北京、上海、海口和拉萨这四个城市作为研究对象，主要基于以下几个方面的考虑。首先，这四个城市分别位于中国的不同地理区域，具有显著不同的气候特点，能够全面展示中国的气候多样性。其次，各城市的天气数据特征（如温度、降水、风力等）存在显著差异。例如，北京的四季分明，冬季寒冷干燥，夏季多雨；而海口常年高温高湿，拉萨则以高海拔气候为特征。这些差异性使得对比分析更具价值。此外，这些城市有较完整的历史天气记录，并且在全国具有较高的关注度，对相关研究与实际应用（如气候适应性规划）具有重要意义。

这种选择使得分析涵盖了从温带到高原气候的多样性，为研究不同气候类型下的天气模式提供了丰富的数据支持，同时也为模型预测提供了更广泛的适用性场景。以下表格总结了四个城市的气候类型和主要天气特点：

城市	**气候类型**	**主要特点**
北京	温带季风气候	四季分明，春季多风沙，夏季炎热多雨，秋季凉爽晴朗，冬季寒冷干燥。
上海	亚热带季风气候	夏季高温多雨，冬季温和湿润，春秋季短促；气候湿润，降水较多，但极端天气较少。
海口	热带季风气候	常年高温，夏秋季多台风和强降雨；冬季气候温暖舒适，降水较少，湿度较高。
拉萨	高原高寒气候	气温偏低，日照充足，昼夜温差大；降雨稀少，天气以

		晴天为主，空气稀薄，风速较低。

2、研究方法

2.1 数据获取

在本项目中，天气数据的来源是“天气后报网”（http://www.tianqihoubao.com）。该网站提供了过去若干年间多个城市的历史天气数据，包括天气状况、温度、风向及风力等信息。为获取这些数据，我们采用了基于 R 的爬虫技术，主要使用 httr 和 rvest 两个包：httr 包负责发送 HTTP 请求，获取网页 HTML 源代码，rvest 包用于解析 HTML 内容，提取目标数据。此外，为确保网络请求能够成功并避免被网站反爬机制限制，我们通过自定义 HTTP 请求头的方式模拟浏览器访问。

首先我们确定了目标网址与数据结构，每个城市的历史天气数据以年度划分，分布在不同的月份页面中。通过观察网页结构，可以发现每个城市的主页面（如 http://www.tianqihoubao.com/lishi/beijing.html）包含了指向各月份数据页面的链接。然后使用 httr::GET() 函数向目标网址发送请求，同时添加 User-Agent 请求头，模拟浏览器访问。使用 rvest::read_html() 函数加载网页内容，并利用 CSS 选择器提取月份链接。遍历所有月份链接，逐一访问对应页面，提取每日天气数据表格。数据表格中的字段包括日期、天气状况、最高气温、最低气温、风力等级等。

通过上述方法，我们成功地获取了北京、上海、海口和拉萨四个城市从 2011 年至 2021 年的天气数据，数据内容包括天气状况、日/夜温度、风向以及风力等关键维度，为后续分析奠定了数据基础。爬取的原始数据存在格式不一致、噪声较多的问题，因此对其进行了全面的数据清洗与预处理。清洗过程中，我们提取并转换白天和夜间气温、天气、风向及风力等信息，去除了无关字符，将天气状况转化为数值编码，风力等级取平均值，并规范了日期与气温字段的格式，以确保数据质量和后续分析的可操作性。

2.2 数据分析

本研究首先从天气后报网站获取了上述四个城市的历史天气数据，在数据清理完成后，采用探索性数据分析（EDA）深入挖掘数据特征。首先对温度数据进行了统计和可视化分析，揭示城市间日/夜温差和季节性变化的规律；随后统计了各类天气状况的分布特征，比较不同城市的天气类型及其出现频率；最后考察了风向、风力的分布及其与温度之

间的相关性。此外，为增强数据的直观性和交互性，开发了基于 Rshiny 的可视化界面，用于动态展示气温、风力及天气变化趋势。

2.3 预测模型

在预测部分，我们采用了随机森林和 R 中的 Prophet 包对气温和天气状况数据进行预测。

- **随机森林**

随机森林是一种基于分类树的算法。它利用 bootstrap 重抽样方法从原始样本中抽取多个样本，对每个样本进行决策树建模，然后组合多棵决策树的预测，通过投票得出最终预测结果。大量的理论和实证研究证明，随机森林具有很高的预测准确率，对异常值和噪声都具有很好的容忍度，且不容易出现过拟合。

- **Prophet**

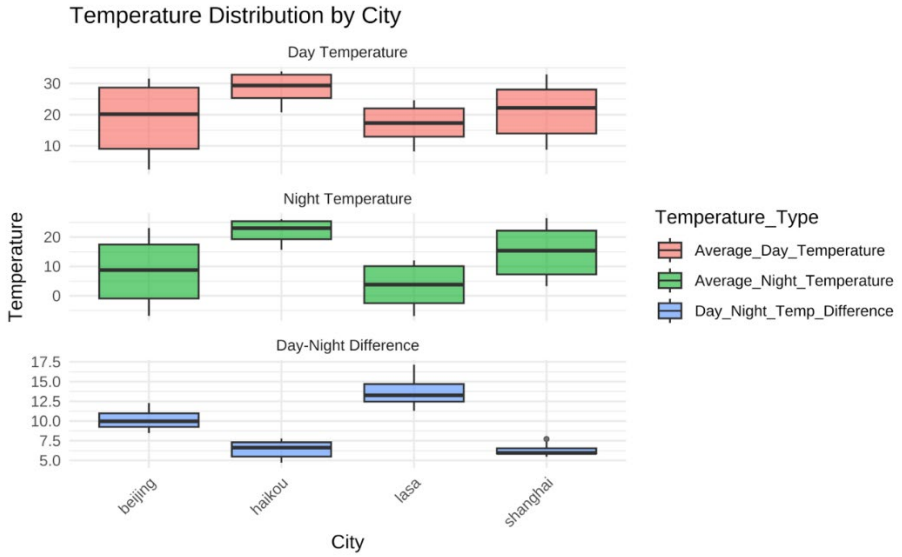
Prophet 是由 Facebook 开发的时间序列预测工具，适用于包含趋势、季节性和假期影响的数据。它使用加法模型将趋势和周期分解，可快速处理缺失值或异常值，对气温等数据的长期预测尤为高效。

3、实验结果

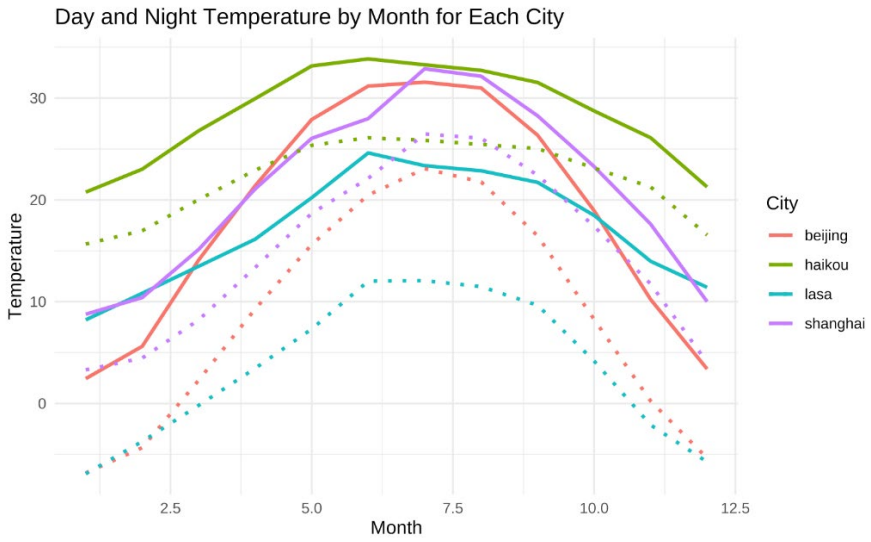
3.1 EDA 结果

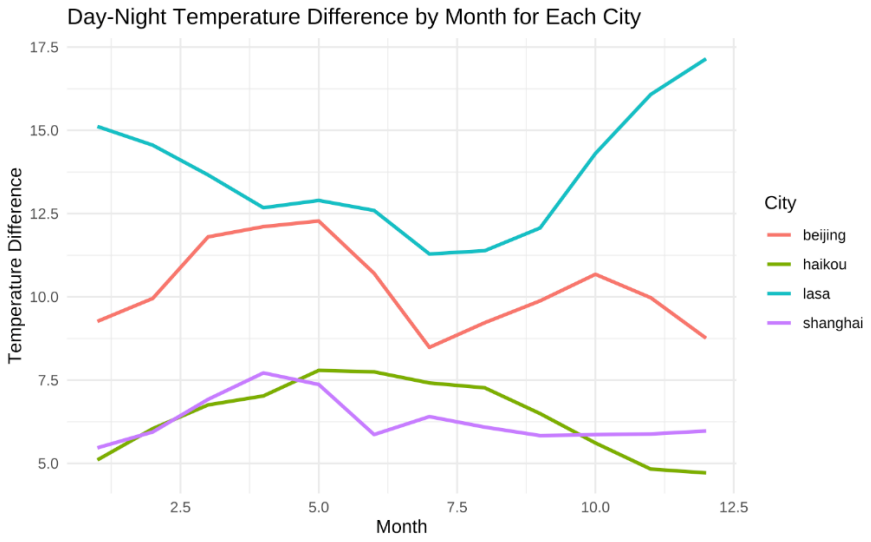
3.1.1 温度特征

实验结果表明，北京、上海、海口和拉萨的气温分布存在显著差异，展现了强烈的地域性和季节性特征。海口的年平均温度最高，拉萨最低，而北京和上海则位于中间。



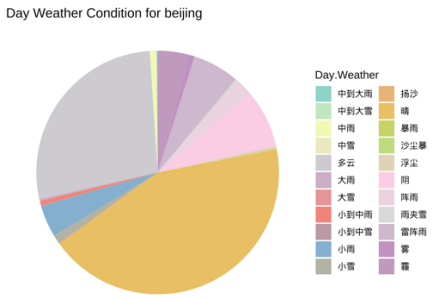
温差方面，冬季的日/夜温差普遍大于夏季，尤其在北京和拉萨表现尤为突出。此外，温度在年份间也表现出一定的波动性，但总体呈现稳定的周期性变化。

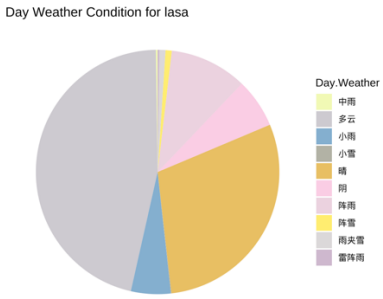


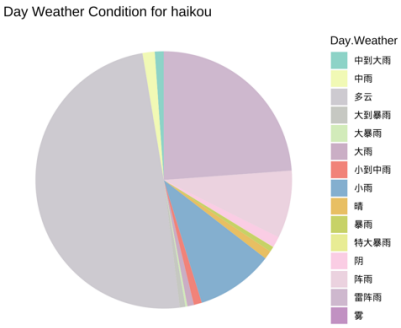


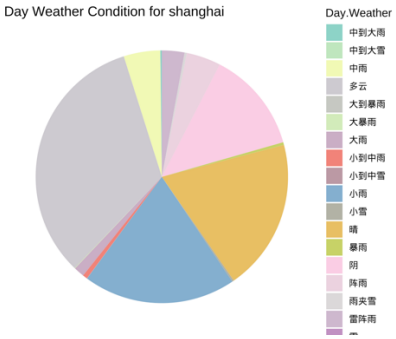
3.1.2 天气特征

四个城市的天气状况分布具有很大差异。北京的天气类型最为丰富，春夏季多雨，秋冬季常见晴天和霾天；上海的天气类型与北京类似，但沙尘天气较少；海口因地处热带，降雨频率最高，且多见强降雨天气；拉萨的天气类型则较为简单，多以晴天为主，降雨天稀少。这样的差异反映了不同城市气候背景的多样性。

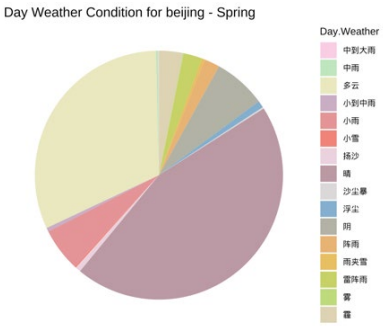


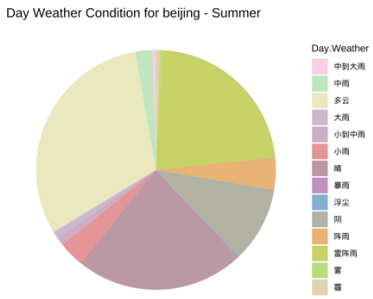


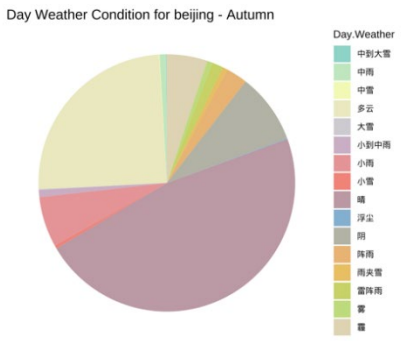


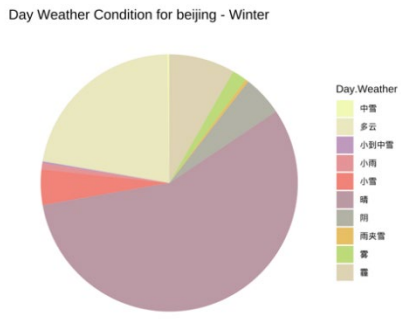


基于获得的天气数据，我们也可以深入研究不同城市在不同季节的天气特征。以北京为例，以下为四季北京的天气频率饼状图。

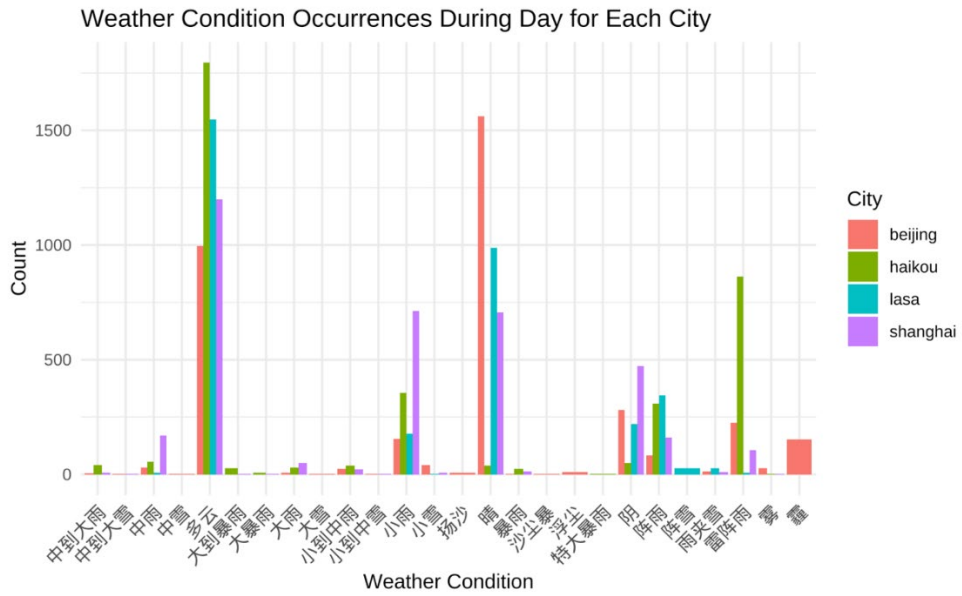






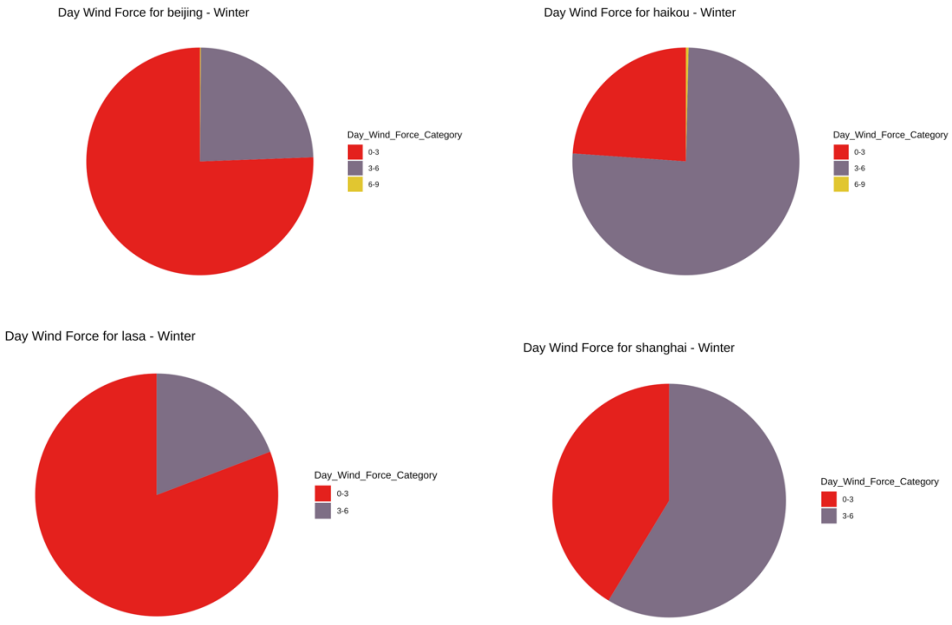


以下柱状图综合反映了四个城市白日天气的出现频率，可以发现不同城市的差异显著。如海口天气多云居多，这与其四面环海，空气湿度大，已形成云有关。而北京的天气则为晴天和多云居多。

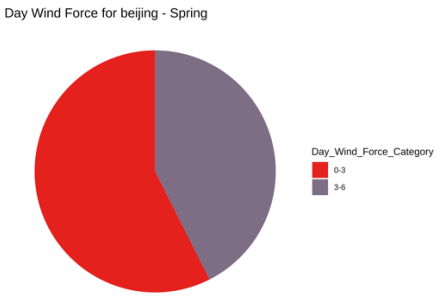


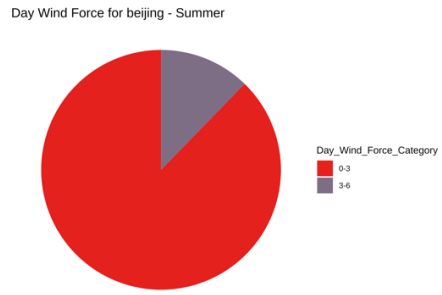
3.1.3 风力特征

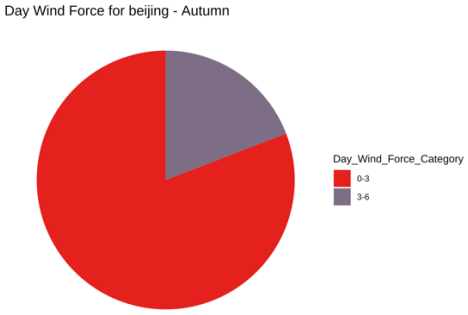
风力分析显示，北京冬季风力较强，而其他季节则相对平缓。海口的风力全年较为稳定，尤其在台风季节期间，风力强度显著高于其他城市。拉萨的风力较为温和，上海的风力分布则受季风影响明显。

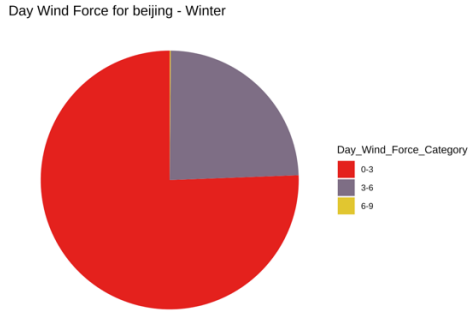


基于获得的风力风向数据，我们也可以深入研究不同城市在不同季节的风力风向特征。以北京为例，以下为四季北京的天气频率饼状图。

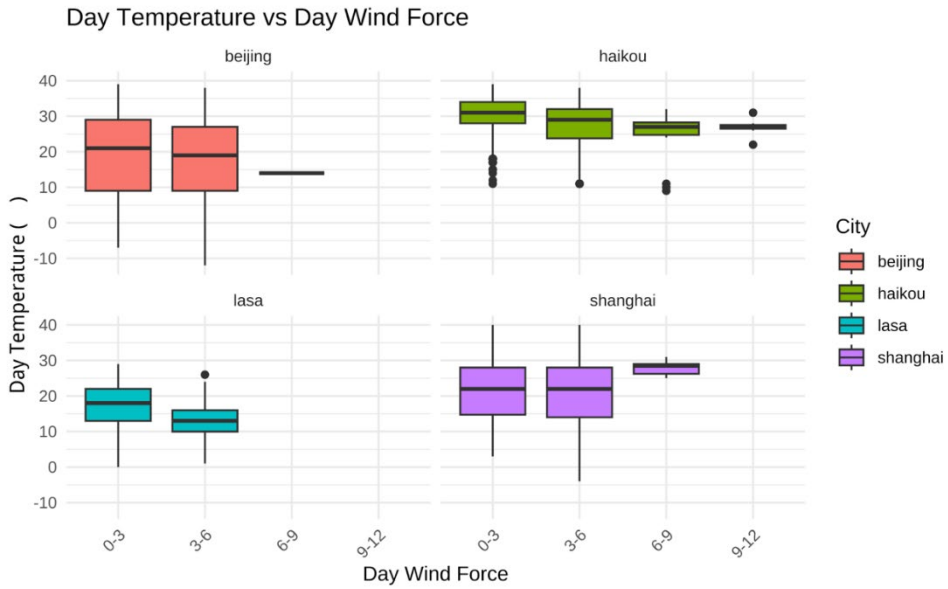




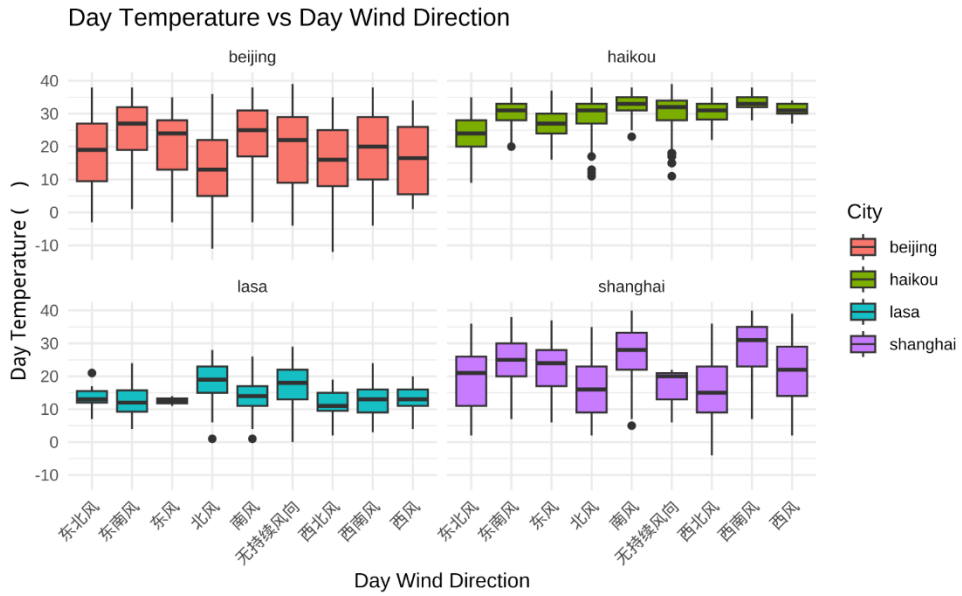




进一步分析显示，温度与风力在不同城市间的相关性存在显著差异，其中海口的温度与风力呈现较强的负相关关系，而其他城市的相关性较弱。



而温度与风向在不同城市间的相关性存在共同点。如通常在南风和东南风下温度较高，而在北风和西北风下温度则较低。



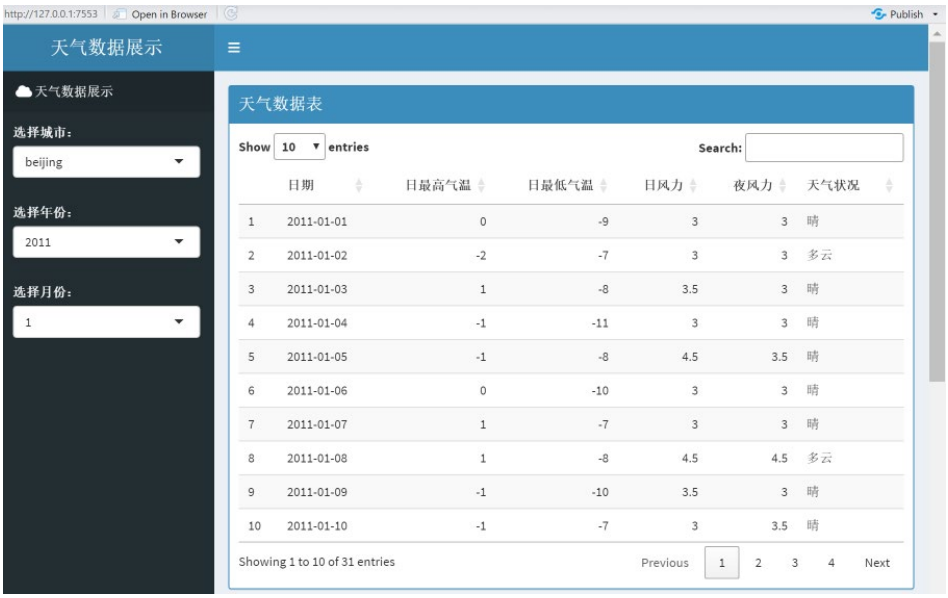
最后我们建立了温度关于风力风向的混合模型并进行了 ANOVA 检验。检验结果与直观观察相符，西南风且风力等级较高与温度表现出显著的正相关，而北风且风力较高则与温度呈现负相关。

```
Fixed effects:
                                                Estimate Std. Error t value
(Intercept)                                      18.2469     2.5893   7.047
Day.Wind.Direction东南风                          5.6591     0.5339  10.600
Day.Wind.Direction东风                            3.1252     0.5757   5.429
Day.Wind.Direction北风                            2.8175     0.4788   5.884
Day.Wind.Direction南风                            5.4251     0.5140  10.554
Day.Wind.Direction西北风                         -2.2409     0.6395  -3.504
Day.Wind.Direction西南风                          3.1201     0.5581   5.591
Day.Wind.Direction西风                            1.4951     0.8466   1.766
Day_Wind_Force_Category3-6                       -0.7342     0.4451  -1.649
Day.Wind.Direction东南风:Day_Wind_Force_Category3-6  0.9363     0.6186   1.513
Day.Wind.Direction东风:Day_Wind_Force_Category3-6    0.2575     0.6869   0.375
Day.Wind.Direction北风:Day_Wind_Force_Category3-6   -4.9550     0.6107  -8.114
Day.Wind.Direction南风:Day_Wind_Force_Category3-6    4.0013     0.6287   6.365
Day.Wind.Direction西北风:Day_Wind_Force_Category3-6  1.5402     0.7710   1.998
Day.Wind.Direction西南风:Day_Wind_Force_Category3-6  5.3447     0.6670   8.013
Day.Wind.Direction西风:Day_Wind_Force_Category3-6    0.7977     1.0377   0.769
```

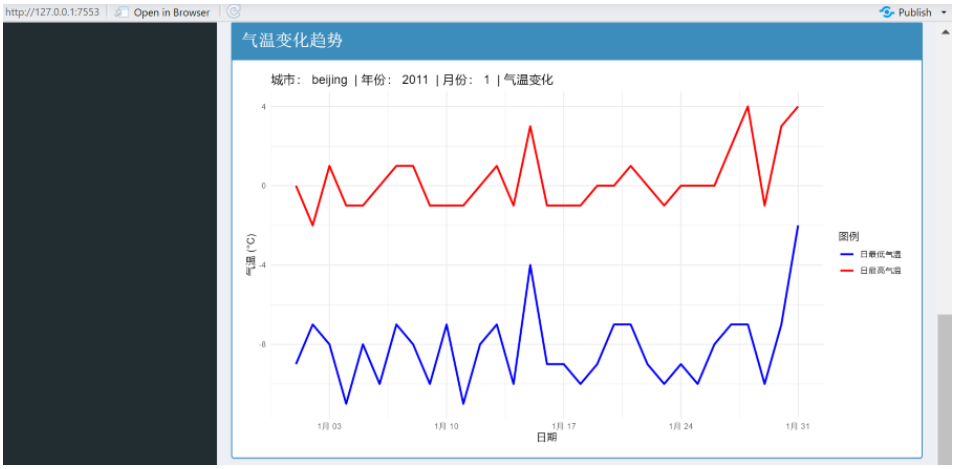
3.2 Rshiny

根据清洗的数据，我们实现了一个可交互的 Rshiny APP，可以用于动态展示气温、风力及天气变化趋势。用户可以通过左侧的交互式筛选面板选择特定的城市、年份和月份。筛选后的结果会动态更新右侧的表格和图表，便于用户分析特定时间段的天气数据。

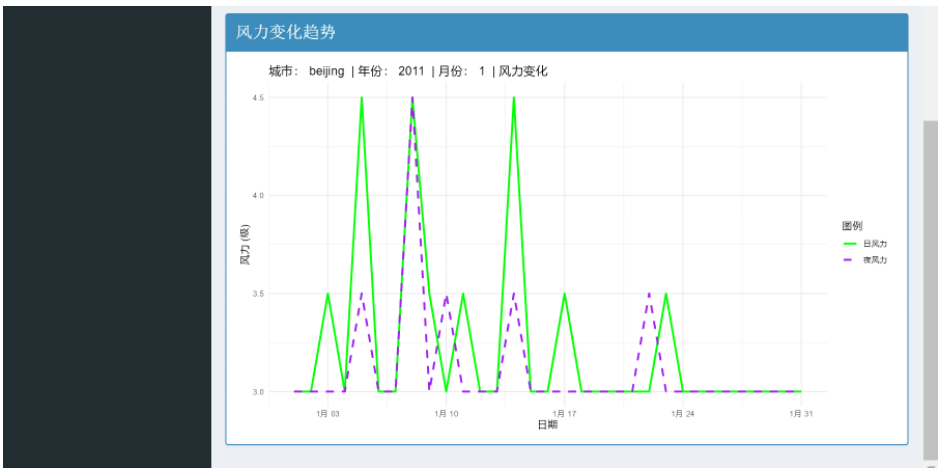
界面上方展示了一张动态更新的天气数据表格，包括日期、日最高气温、日最低气温、日风力、夜风力及天气状况。表格支持分页显示，并提供搜索功能，方便用户快速定位感兴趣的数据。



下方的“气温变化趋势”图直观展示了选择城市和时间范围内的日最高气温和最低气温变化情况。



界面还包括“风力变化趋势”图，通过绿色和紫色虚线分别表示日风力和夜风力的变化。

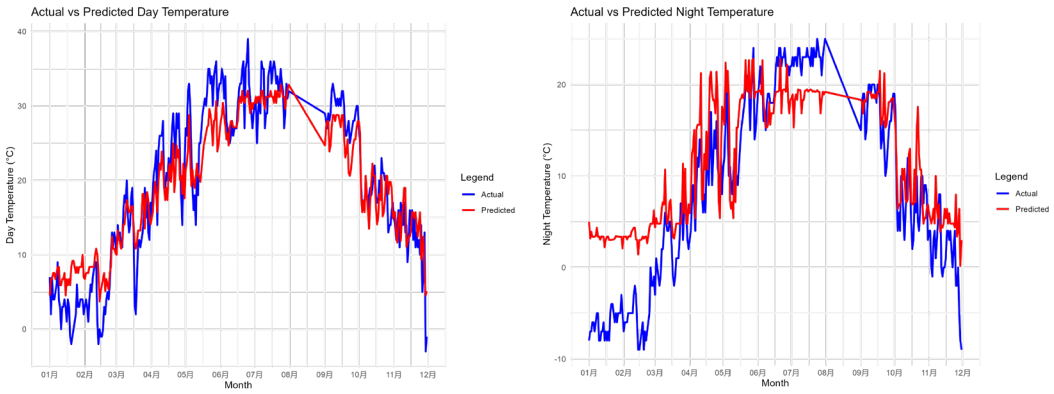


3.3 模型预测

基于上一部分得到初步清洗的数据，我们进行进一步清洗。我们将不同的天气状况编为整数编码，在用随机森林进行预测天气状况时，预测结果和整数编码的顺序无关。随机森林通过树的分裂点对数据进行划分，不依赖变量之间的顺序或线性关系。因此随机森林不会将这些数字视为有序的，而仅将其视为不同的类别。

3.3.1 日最高、最低气温预测

我们按照 9:1 的比例，将文本划分为训练集和测试集，并用 set.seed 设置随机种子，以保证结果的可重复性。在调参部分，我们首先通过循环，挑选出了二叉树上最佳的变量个数。接着，我们通过绘图找出了最佳的树的个数。n_tree 在 1000 附近时，模型误差较低，且较为稳定，因此，将树的个数设为 1000。



日最高（左）、最低（右）气温预测 vs 实际

随机森林模型对最高气温和最低气温的预测结果表现良好，从趋势上看，预测值较好地捕捉了实际温度的变化趋势，符合气温随季节变化的特征。从波动情况来看，预测值能够反映气温的日常波动，较为贴近实际观测数据。在误差分布方面，预测误差在 1 月至 3 月以及 6 月至 9 月期间稍大。因此考虑结合 Prophet 进行预测。

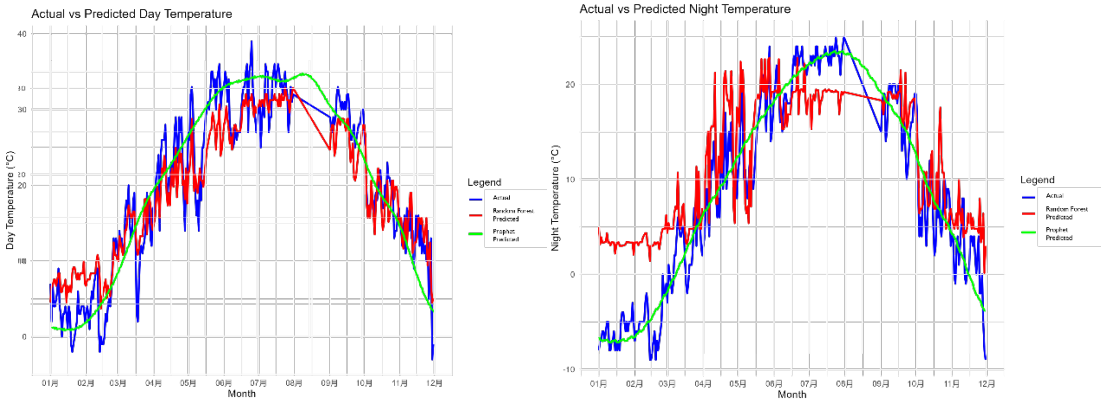
Prophet 模型的特点是，它基于分段线性趋势和傅里叶级数的时间序列模型，能够自动捕捉趋势变化点和复杂季节性波动。

它将时间序列分解为以下三部分：

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

$y(t)$是时间 t 的预测值。$g(t)$长期趋势函数，用于捕捉时间序列的长期变化。$s(t)$周期性成分，用于描述季节性变化。$h(t)$特殊事件的影响（例如假期），本研究采用简化模型，不考虑假期效应等特殊事件的影响。ϵ_t是残差项，用于捕捉未建模的噪声。

Prophet 模型自动化程度高，易于实现，并且对缺失值、异常值具有较强的稳健性。通过对比，我们发现 Prophet 更适合建模非线性趋势和复杂季节性波动的气温数据。



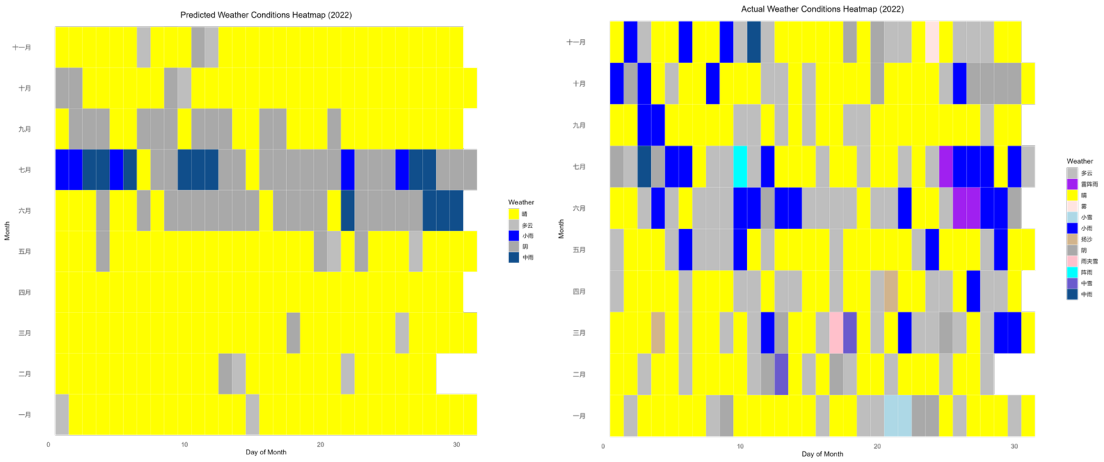
日最高（左）、最低（右）气温预测 vs 实际

从趋势上看，Prophet 模型的预测曲线（绿色）很好地捕捉了全年气温的总体变化趋势。从波动情况来看，预测曲线较为平滑，难以反映气温的日常波动。这是因为 Prophet 偏向于捕捉长期趋势和季节性变化，而对短期波动的敏感度较低。从误差分布来看，预测值与实际值之间的误差较小，尤其在较大时间跨度上表现出较高的准确性。

因此，Prophet 非常适合预测日最高气温和最低气温的整体变化值。若需要短期内的气温波动预测（如几天或几周内的高低起伏），可以结合随机森林模型的结果。

3.3.2 天气状况预测

用随机森林预测天气状况结果如下。



天气状况预测（左）vs 实际（右）

模型对“晴”、“多云”等高频天气状况的预测表现较好。从图中可以看出，这些天气状况的预测结果与实际分布基本一致，能够较好地捕捉高频类别的规律。

模型没有预测出“小雪”、“中雪”、“雨夹雪”等的原因分析如下。由于“小雪”、“中雪”、“雨夹雪”、“扬沙”、“雷阵雨”、“雾”、“霾”、“阵雨”、“大雨”在训练集中出现的频数非常低，而随机森林在处理类别的频数不均匀的问题时倾向于优化整体的准确率，因此更容易预测出高频类别，而忽略低频类别。

四、总结与展望

本研究基于 2011 至 2021 年的历史天气数据，系统分析了北京、上海、海口和拉萨四个城市的气温、天气和风力特征，并通过回归和分类模型对未来天气状况进行了预测。研究结果揭示了不同城市气候特性的差异性和复杂性，为理解气候变化规律及其潜在影响提供了有力的数据支撑。

然而，本研究仍有一定局限性。一是数据范围较为局限，仅选取了四个城市，无法全面反映全国范围内的天气特点；二是在预测极端天气时，模型性能仍有改进空间。未来研究可以扩展数据范围，纳入更多城市和更长时间跨度的数据，并探索更为复杂的建模方法，如深度学习和集成学习，以提高预测的精度和鲁棒性。此外，还可尝试引入气象因子（如气压、湿度）等外部变量，进一步提升模型对复杂天气现象的解释力和预测力。

通过不断完善，天气数据分析与预测研究将为气候适应性规划和资源管理提供更加科学的支持，为社会经济发展贡献更多价值。

五、大作业分工

罗庆典：网络爬虫、数据清洗、数据分析(EDA)、结果可视化与解读、总结与展望

钟阅：数据清洗、Rshiny 搭建、随机森林预测、Prophet 预测、结果可视化与解读

六、代码

https://github.com/GoToB3d/Weather-data-analysis