天子脚下，天气如何？

1. 数据分析目标

本问题的因变量为北京空气质量指数的因素；因为天气系统都具有一定的空间尺度和时间尺度，而且各种尺度系统之间相互交织、相互作用。

自变量为：时间、星期几、最低气温、最高气温、天气、风向、风力、空气污染程度。

分析目标：建立因变量和自变量之间的多元线性回归模型，估计模型系数、检验系数显著性以确定自变量是否对因变量有影响，并预测因变量新值。

1. 数据预处理

首先观察因变量和自变量是什么样的数据类型。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 变量类型 | | 变量名 | 详细说明 | 取值范围 | 备注 |
| 因变量 | | 空气质量指数 | 连续变量 | 15~450 |  |
| 自变量 |  | 时间 | 天气的时间 | [2016/1/1，2017/12/13] |  |
|  | 星期 | 分类型变量 | 星期一；星期二 |  |
|  | 最低气温 | 连续型变量(℃) | [-13，29] | 转变为数字 |
|  | 最高温度 | 连续型变量(℃) | [-2，41] | 转变为数字 |
|  | 天气 | 分类型变量 | 包括晴天、多云转晴、阴转多云、阴转小雨、多云转雨夹雪、小雨转多云等； | 进行分类 |
|  | 风向 | 分类型变量 | 包括东风、西风、南风、北风、东南风、东北风、西南风、西北风等； | 哑变量 |
|  | 风力 | 顺序型变量 | 取值范围[0，6]； | 等级分类 |
|  | 空气污染程度 | 顺序型变量 | 包括优、良、轻度污染、中度污染、重度污染和严重污染。 | 等级分类 |

进行初步分析可得结果：

天气水平分类变量，天气包括晴天、多云转晴以及晴转多云等，如果将其转变为哑变量，多云转晴和晴转多云就会形成两个变量，因此我们假设晴转多云和多云转晴是等价的都是由晴和多云组成，将其变为哑变量，共形成晴，多云，阴，小雨，中雨，小到中雨，雾，阵雨，雷阵雨，中到大雨，大雨，扬沙，雨夹雪，小雪，霾，雾，大雪，暴雨，大到暴雨19个定性变量；其中晴天为基础变量。

风向也属于分类型变量，参照经验，风向能够影响当天的空气质量指数的大小，将其变为哑变量，其中无持续为基础变量。

在自变量中，星期几对于空气的影响未知，描述性统计中进行详细分析。

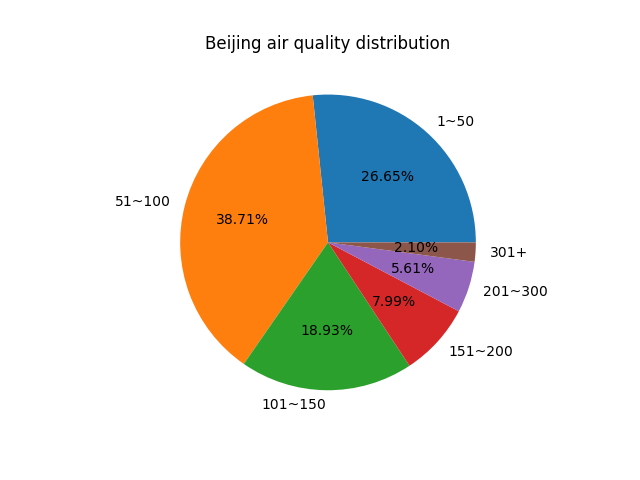
1. 描述性分析
2. 对北京近两年的空气质量数据进行总体分析

空气污染指数的取值范围定为0～500，其中0～50、51～100、101～150、151~200、201～300和大于300，分别对应国家空气质量标准中日均值的 I级、II级、III级、IV级和V级标准的污染物浓度限定数值。通过筛选得：

**2016-2017年空气质量状况统计**

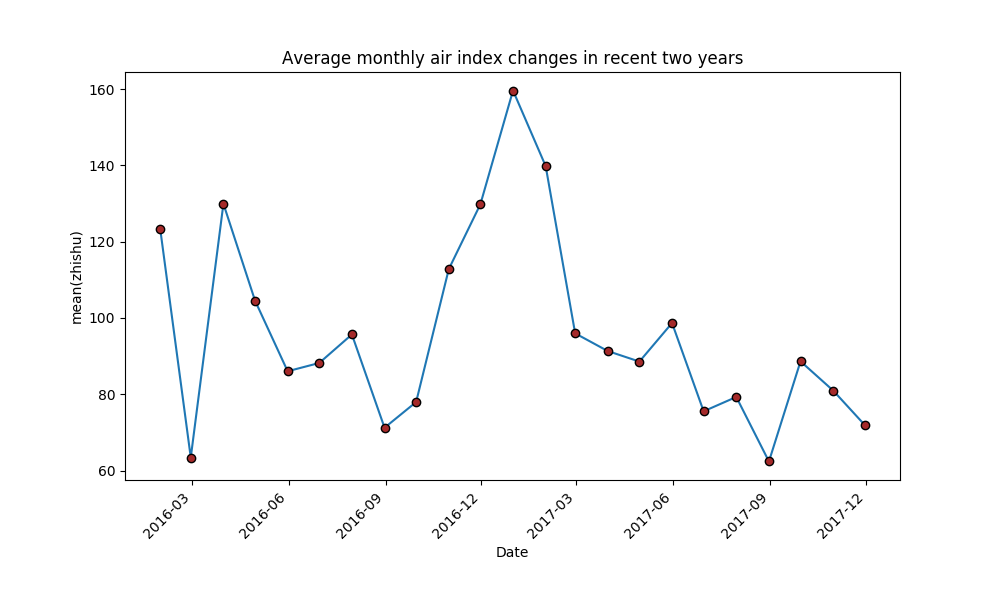
|  |  |  |
| --- | --- | --- |
| 空气质量指数范围 | 空气质量状况 | 天数 |
| 0～50 | 优（I级） | 190 |
| 51～100 | 良（II级） | 276 |
| 101～150 | 轻度污染（III级） | 135 |
| 151~200 | 中度污染（IV级） | 57 |
| 201~300 | 重度污染（V级） | 40 |
| 301以上 | 严重污染 | 15 |

画出饼状图更直观反映结果。

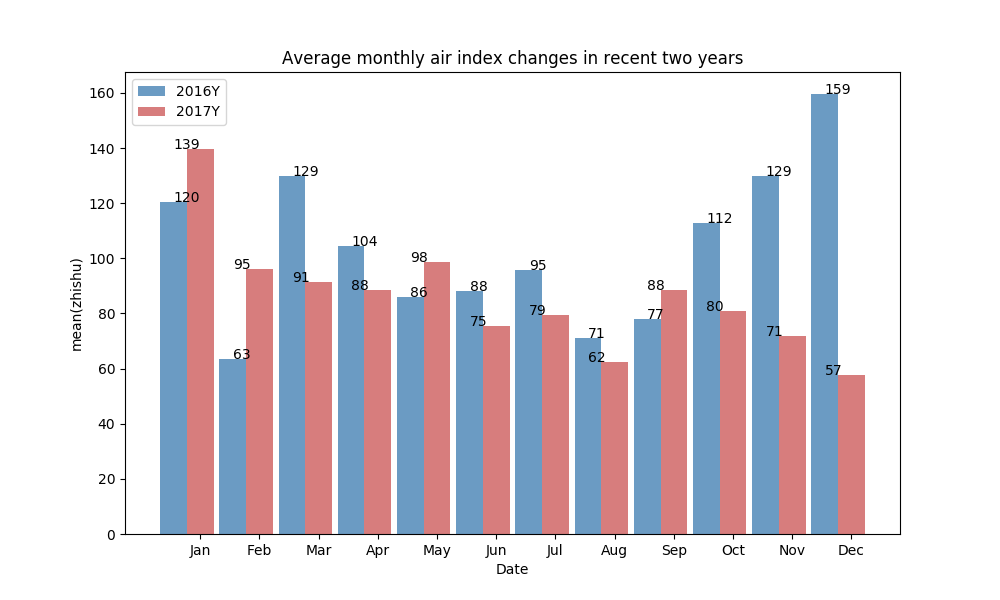


由上图可得，质量指数在100以下属于空气质量较好占比为55.36%，也就意味这有接近一半的天气属于不易运动的天气，

1. 空气质量指数和时间的关系

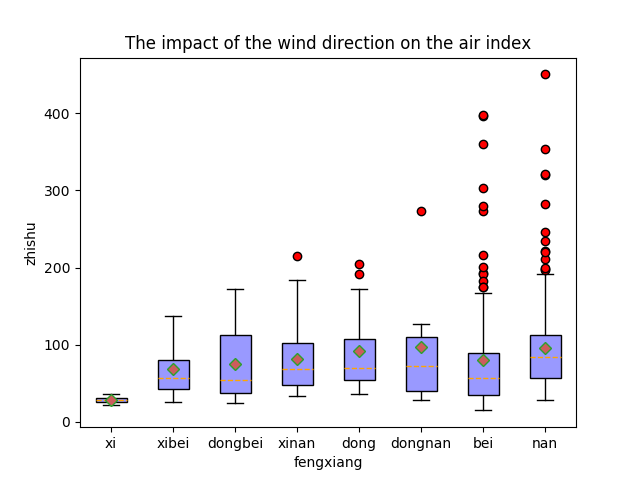


由图可以看出，每个月的平均空气质量指数波动较大，但是最高点是在2016年的12月，并且2016年的10、11、12月连续上升，考虑到主要原因可能和冬天的暖气供应造成整体空气较差。但是2017年的10、11、12月空气质量指数总体在优良的水平，可能和这一年的十月十八号十八大召开和一系列整治政策相关，是的空气质量明显改善。



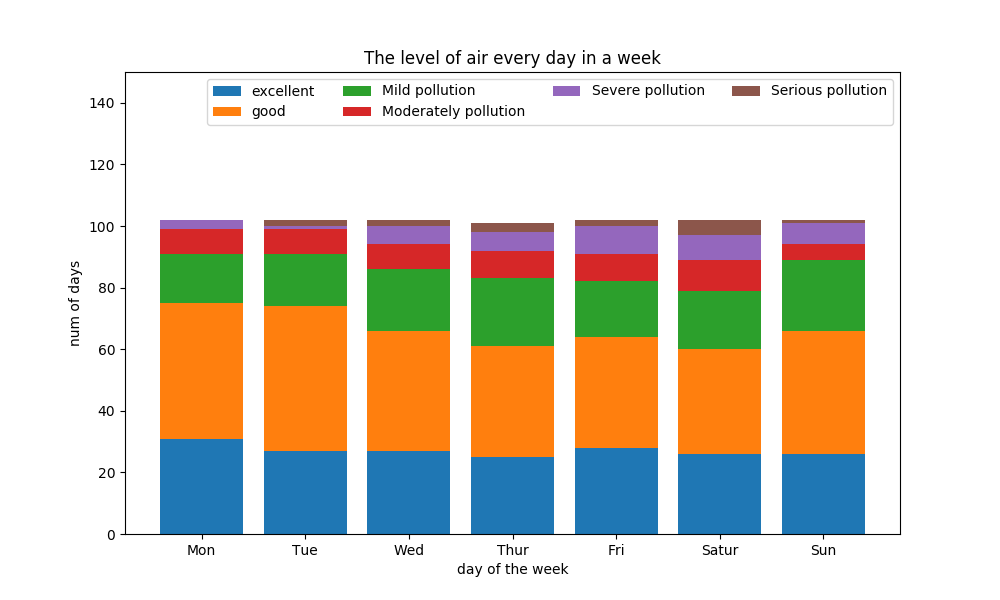
上图表示的是2017年每月均值和2016年同期进行比较，可以看出2017年每月空气质量指数的均值比2016年同期有所下降，这和一年来环境整治密切相关。10、11、12月下降明显，十九大召开和大兴火灾事件后北京的规范居住环境应该有很大关系。

1. 风向对北京空气指数的影响



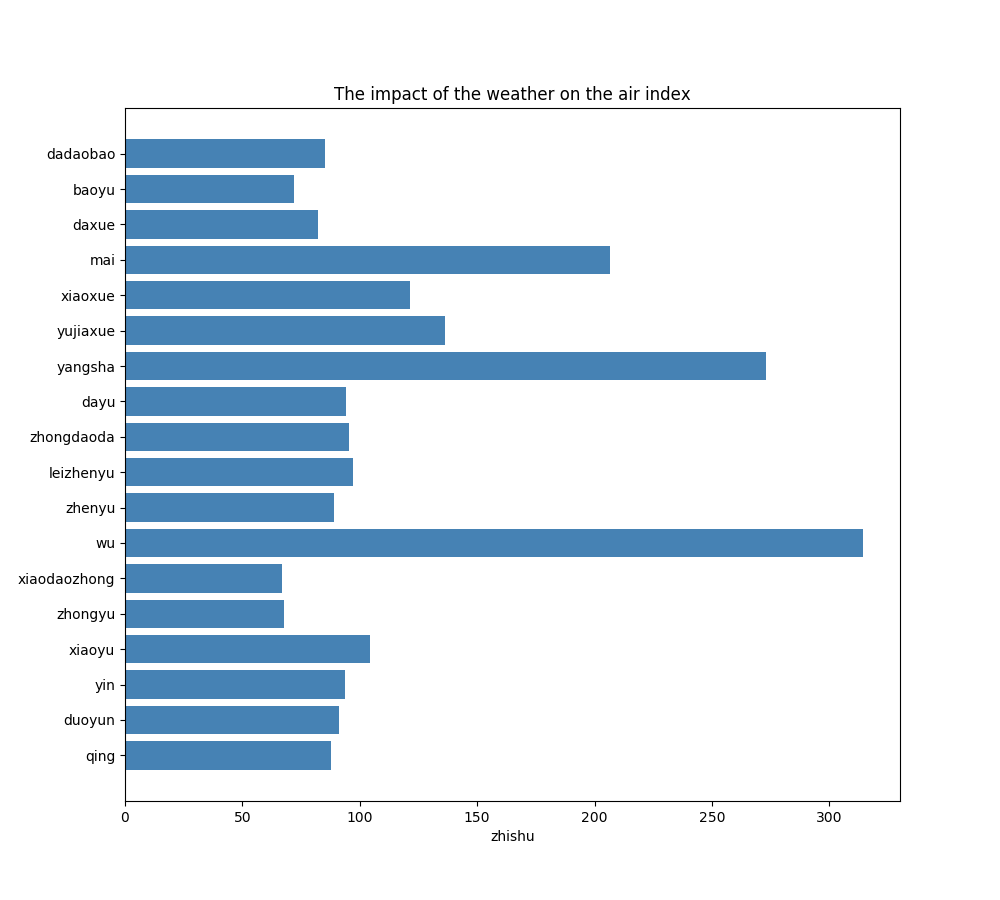
根据风向与空气质量指数数据画出箱线图如下：由箱线图可以得出，风向对空气质量指数是有影响的。各个风向天气平均水平都处于100以下，天气良好。西风的时候样本量最少（只有五天是刮西风），但是空气最好。严重污染多出现在北风、南风时。从北京市的工业总体布局上，东部南部和北部的工业较多，然而来自西面的风由于工业比较薄弱因而能够带来质量较好的空气。

1. 一周内出现不同等级空气质量的占比



上图中表示2016年至2017年一周中每一天的空气污染程度的条形图，在过去两年周一几乎没有发生过严重污染天气，周三、周五发生污染的可能性最大，周六发生严重污染的次数最多，应该大城市中人们周六最狂欢相关。但是总体比例变化不大，说明市民日常生活对空气质量指数影响不大。

1. 天气的状况和空气质量指数的关系



上图中，表示的是天气状况和该天气状况下的空气质量指数平均值之间的关系，从上图可以看出，出现雾、霾、扬沙（只出现了一天可以删除）的天气，空气质量指数比其他天气下明显高出，当出现三种天气是一定要减少外出。当出现小到中雨、中雨、暴雨、大雪等雨雪天气时，平均空气质量指数会下降。但是总体上讲，北京市的空气质量指数属于良的居多，还有很大的改善空间。

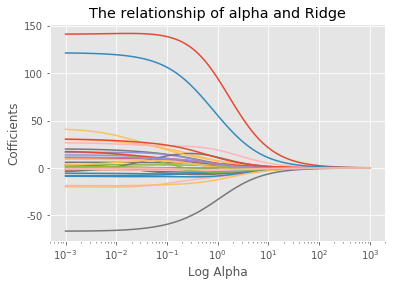
1. 模型建立

本文设立哑变量较多，利用岭回归和Lasso回归建立模型。

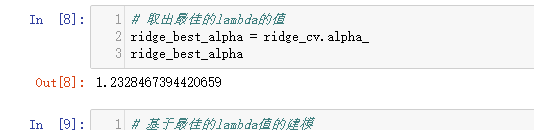
其中自变量为： 'Wednesday', 'Tuesday', 'Friday', 'Saturday', 'Thursday', 'Sunday', 'high', 'low', 'qing', 'duoyun', 'yin', 'xiaoyu', 'zhongyu','xiaodaozhong', 'wu', 'zhenyu', 'leizhenyu', 'zhongdaoda', 'dayu','yangsha', 'yujiaxue', 'xiaoxue', 'mai', 'daxue', 'baoyu', 'dadaobao','dong', 'dongbei', 'dongnan', 'bei', 'nan', 'xi', 'xibei', 'xinan', 'fengli',

因变量为： 'zhishu'

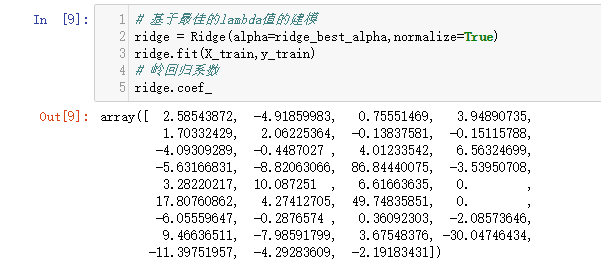
岭回归

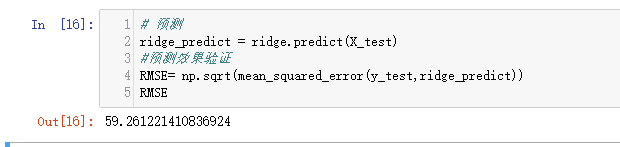


从上面个图形可以看出，alpha在1~10之间时，所有的自变量系数趋于稳定，我们采用交叉验证（CV）方法确定lambda值。



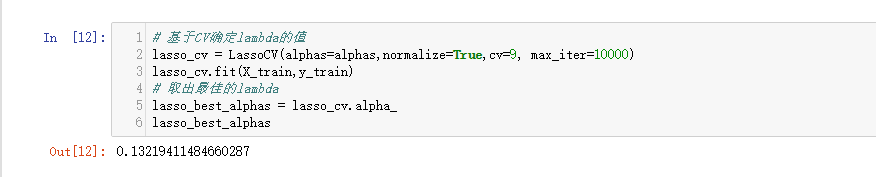
最终结果可以看出，lambda=1.2328，利用最佳lambda建立模型，最终MSE为59.26



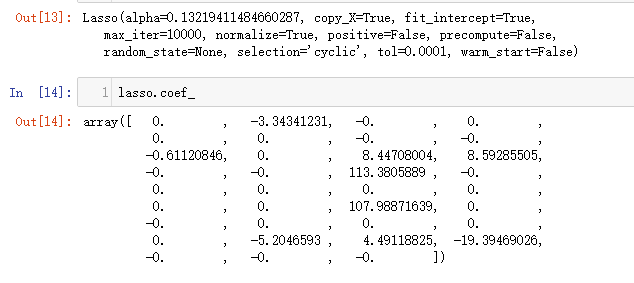


与上面的自变量一一对应，发现有雾的天气比晴天空气质量指数高86.844，有霾的天气比晴天空气质量指数高出50，但是扬沙天气由于只出现了一天，在这里影响并不明显，这和我们作出的天气的状况和空气质量指数的关系条形图不谋而合。

二、LASSO回归



基于Lambda=0.1321进行建模，并且MSE为56.344。



从上图可以看出，对空气质量指数影响最大的为：雾、霾比晴天分别高出113、107。然而刮西风却能带来好天气。这也比较符合大家的常识，雾霾天气减少外出，风能够把霾吹跑，但是北京周围工业较多，风对空气质量指数影响有限，应该进行产业升级，才能真正的带来空气质量指数的下降。其余变量都指数的影响不大。