CSDN新首页上线啦,邀请你来立即体验!(http://blog.csdn.net/)



博客 (//blog. **(#kulnywet/9def:ntxt/9lled+)**toolba**学**院 (//edu.csdn.net?ref=toolbar)

下载 (//download.csdn.net?ref=toolbar)

GitChat (//gitbook.cn/?ref=csdn)

更多心









(//write(b/logict)cosdkncnét/exastietclitat/activity?

(//my.csdn.

python机器学习

番番要吃肉 (htt

Python机器学习

立即体

Python数据分析练习:北京、广州PM2.5空气质量分析(1)ef=toollbiar_source=csdnblog

2017年08月18日 11:20:09

☆ 标签: Python (http://so.csdn.net/so/search/s.do?q=Python&t=blog) /

数据分析 (http://so.csdn.net/so/search/s.do?g=数据分析&t=blog) /

大数据 (http://so.csdn.net/so/search/s.do?q=大数据&t=blog)

\$36

由于雾霾问题,全社会都很关注空气质量,政府也花了很多钱力图改善空气质量。我们作为城市市民经常要问:我们城市 的空气质量到底怎样?这几年我们城市的空气质量是在改善还是恶化?我们城市的空气质量与其他城市相比,是更好还是 更差?

官方媒体一般都是说:我们的空气质量在改善,但有数据证明吗?官方数据可信吗?我们心存疑虑。所以,作为数据狗, 还是自己动手吧。

PM2.5是最近几年特别热议的空气质量指标,这要归功于美帝大使馆哦。美国驻华使馆和领馆自己检测PM2.5数据并且在 官网发布,这才让国人知道了这个指标,后来国内官方也开始公布这个指标了。

+ 关注 (http://blog.csdn.net/xiexf189)

> 码云 未开诵

原创 粉丝 (https://gite 4 utm sourc

▮他的最新文章

更多文章 (http://blog.csdn.net/xiexf189)

使用python进行简单的分词与词云 (htt p://blog.csdn.net/xiexf189/article/detail s/77477283)

Python数据分析练习:北京、广州PM 2.5空气质量分析(2)(http://blog.csd n.net/xiexf189/article/details/7736858 3)

 \triangle 内容举报



返回顶部

Python-sklearn 机器学习的第一个样例 (7) (http://blog.csdn.net/xiexf189/arti cle/details/72598976)

言归正传,差点忘了,这是一篇Python技术贴。回到这篇文章的主题:用Python做数据分析。我生活在广州,当然以我大广州为分析目标了。

数据源:美国驻华使馆的空气质量检测数据,http://www.stateair.net/web/historical/1/1.html (http://www.stateair.net/web/historical/1/1.html) 包含以下字段:

Table 1. Definitions and examples of terms.

r\\		
Term	Definition	Example(s)
Site	City or post where the measurements were taken.	Beijing, Shenyang
Parameter	The air quality pollutant measured; either fine particles (particulate matter less than 2.5 micrometers in diameter – $PM_{2.5}$) or ground-level ozone (O ₃).	PM _{2.5} , O ₃
Date (LST)	The date and hour of the measurement in local standard time (e.g., BJT – Beijing Time ^a). All data are reported in begin-time. ^b The date-time format follows YYYY-MM-DD HH:mm, where 00:00 is midnight, 14:00 is 2:00 p.m., etc.	2013-05-01 00:00, 2013-05-20 07:00, 2013-05-20 23:00
Year	4 digit year that corresponds to YYYY in Date	2013
Month	1 or 2 digit month (1 to 12) that corresponds to MM in Date	5, 12
Day	1 or 2 digit day (1 to 31) that corresponds to DD in Date	1, 31
Hour	1 or 2 digit hour (0 to 23) that corresponds to HH in Date	0, 18
Value	The measurement in concentration. Missing values are listed as -999.	49, 450, -999
Unit	Parameter concentration units, either micrograms per cubic meter $(\mu g/m^3)$ for PM _{2.5} or parts per billion (ppb) for ozone.	μg/m³, ppb
Duration	The averaging period of the data reported; either 1-hour (1 Hr) for $PM_{2.5}$ or ozone, 24 hours (24 Hrs) for daily $PM_{2.5}$, or 8-hr begin-time average-daily max ^c for ozone.	1 Hr, 24 Hrs, 8 Hr begin time Avg- Daily Max
QC Name	The quality control status of the data; either valid or missing (unavailable). Invalid data are not included in these files. http://blo	Valid, Missing g. csdn. net/x1exf189

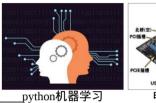
其中,PM2.5数值所代表的空气质量评价如下表1-1所示:

Python-sklearn机器学习的7 (6) (http://blog.csdn.net cle/details/72598910)

Python-sklearn机器学习的 (5) (http://blog.csdn.net cle/details/72560725)



Python机器学习



相关推荐

架构设计:系统存储(25) 性与Paxos算法(下)(http://blog.csdn.net/yinwenjie/article/details/62044199)

机器学习与深度学习(一)决策树算法 (Decision Tree) (http://blog.csdn.net/tz_z s/article/details/73796241)

python爬虫抓取全国pm2.5的空气质量(2 015.12.21版) (http://blog.csdn.net/djd12 34567/article/details/45601283)

皇后区空气质量数据分析.zip (http://download.csdn.net/download/qq_37879432/10 168409)

□ 0

⚠
内容举报



空气质量指数 (AQI)	PM2.5 健康建议
健康 (0-50)	无
中等 /(51-100)	特别敏感的人群应该考虑减少长期或沉重的负荷。
砂敏感人群不 健康 √101-150)	有心脏或肺部疾病的人、老人和小孩应该减少长期或沉重的负荷。
承健康 (151-200)	有心脏或肺部疾病的人、老人和小孩应该避免长期或沉重的负荷。其他人也应该减少长期或沉重的负荷。
。 作常不健康 (201-300)	有心脏或肺部疾病的人、老人和小孩应该避免所有户外活动。其他人也应该避免长期或沉重的负荷。
危险 (301-500)	所有人都应该避免户外活动。有心脏或肺病的人、老人和小孩应该保持在室内,减少活动。翻译: English http://blog.csdn.net/xiexf189

在此基础上,如果指数大于500,就是大家戏称为"爆表"

问题:

- 一.广州的空气质量总体如何?
- 二.广州最近两年的空气质量是否有改善?
- 三.广州北京两地的空气质量对比

这几个问题其实比较大而空,要落实到具体的指标和维度。从我们掌握的以上数据出发,可以从以下几个指标和维度分析:



他的热门文章

 \triangle

Python数据分析练习:北京、广州PM2.5 内容举报 空气质量分析(1)(http://blog.csdn.net/x iexf189/article/details/77367504)

₩ 826

返回顶部

Python-sklearn机器学习的第一个样例 (6) (http://blog.csdn.net/xiexf189/article/details/72598910)

- 1.2016年全年来看,空气质量较好(PM2.5<=100)的天数占比是多少?2015年相比2016年的对比?北京与各自的对比如何?
- 2.2016年全年来看,空气质量最严重(PM2.5>300)的天数占比是多少?2015年相比2016年的对比?北京与各自的对比如何?
- 3.2016年全年来看,空气质量与季节(月份)的变化关系是什么?

4.2016年与2015年同时间点对比,空气质量较好和较差的次数分别是多少?

0

5.2016年广州与北京同时间点对比,空气质量较好和较差的次数分别是多少?

W

下面就开始进入数据集吧。

1. 数据集下载、读取和清洗

从美驻华使馆官网下载了广州、北京两个城市,2015年和2016年的PM2.5数据。分别是这四个文件: Beijing 2015 HourlyPM25 created20160201.csv

Beijing_2016_HourlyPM25_created20170201.csv

Guangzhou_2015_HourlyPM25_created20170201.csv

 $Guangzhou_2016_HourlyPM25_created 20170201.csv$

为了简化工作,每个文件只保留这几个字段:Year、Month、Day、Hour、Value和QC name,用Excel删掉其他的字段和文字,千万注意要保留表头哦。

经过大致观察,可以发现以下几点:

1. 数据以csv格式保存,每年1个文件

37

Python-sklearn机器学习的 (3) (http://blog.csdn.net. e/details/72528755)



Python-sklearn机器学习的 (2) (http://blog.csdn.net e/details/72528667)

₩ 589

Python-sklearn 机器学习的

(1) (http://blog.csdn.net e/details/72518860)

497



Python机器学习



python机器学习

<u>小</u> 内容举报



http://blog.csdn.net/xiexf189/article/details/77367504

2. 每个小时一条检测数据

首先引入必要的Python包

In [1]:

```
import unicodecsv
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sb
```

₩ 从csv文件中读入数据

pandas是Python中常用的数据分析工具,包含了大量简便的数据分析处理方法。本文通过这个例子,展现了几个常用的pandas使用方法。

其中,使用pandas可以很方便地从csc文件中读取数据,保存在DataFrame对象中,方法如下:

In [2]:

```
# 从csv中读取北京2015年pm2.5数据,保存在DataFrame中
df_bj2015 = pd.read_csv('data/Beijing_2015_HourlyPM25_created20160201.csv')
# 重新设定各个字段的名称
df_bj2015.columns=['Year','Month','Day','Hour','Value','QC']
```

数据探索和清洗



内容举报

拿到数据集,并且导入了DataFrame对象后,我们需要对数据进行观察,看看是否有缺失和错误的情况,并且对数据集进行必要的清洗处理,以便后续得到更准确的分析结果。

In [3]:

查看DataFrame中的前几条数据 df_bj2015.head()

Qut[3]:

0	Year	Month	Day	Hour	Value	QC
0	2015		1	0	22	Valid
1	2015	1	1	1	9	Valid
2)2015	1	1	2	9	Valid
3	2015	1	1	3	13	Valid
4	2015	1	1	4	10	Valid

In [4]:

DataFrame的简要描述 df_bj2015.describe()

Out[4]:

	Year	Month	Day	Hour	Value
count	8760.0	8760.000000	8760.000000	8760.000000	8760.000000
mean	2015.0	6.526027	15.720548	11.500114	71.658904
std	0.0	3.448048	8.796749	6.922433	
min	2015.0	1.000000	1.000000	0.000000	-999.000000
25%	2015.0	4.000000	8.000000	5.750000	21.000000
50%	2015.0	7.000000	16.000000	11.500000	53.000000



⚠
内容举报

TOP

	Year	Month	Day	Hour	Value
75%	2015.0	10.000000	23.000000	17.250000	108.000000
max	2015.0	12.000000	31.000000	23.000000	722.000000

从简要描述中,可以看出Value字段(pm2.5的值)最小值是-999,正常来说,PM2.5的值不会小于零的,这说明有些观测值是缺失或者记录错误。我们可以计算一下其中有多少Value<0的记录:

I្ណា [5]:

len(df_bj2015.ix[df_bj2015.Value<0, :])

Out[5]:

97.

金年8000多条数据中,只有97条记录缺失或错误,所以把这97条记录删除,对整个分析影响不大。以下把缺失或错误的记录删除:

In [6]:

把错误值置为空值

df_bj2015.loc[df_bj2015.Value<0,'Value']=np.nan

删除空值记录

df_bj2015.dropna(inplace=True)

再用describe方法看看数据集的概况:

In [7]:

df_bj2015.describe()

Out[7]:

		Year	Month	Day	Hour	Value
C	ount	8663.0	8663.000000	8663.000000	8663.000000	8663.000000



<u>小</u>

内容举报

TOP

	Year	Month	Day	Hour	Value
mean	2015.0	6.538728	15.674939	11.494055	82.733810
std	0.0	3.448080	8.787195	6.938605	88.556186
min	2015.0	1.000000	1.000000	0.000000	0.000000
25%	2015.0	4.000000	8.000000	5.000000	21.000000
50%	2015.0	7.000000	16.000000	11.000000	54.000000
75%	2015.0	10.000000	23.000000	18.000000	109.000000
max	2015.0	12.000000	31.000000	23.000000	722.000000

可以看出,Value字段最小值不再为负数,这样的数据基本可信。

og 另外,我们还要为数据集增加一个字段,根据前面表1-1的划分方法,把pm2.5的测量值Value,转换为空 气质量等级。 首先,我们先写一个函数,通过pm2.5测量值对应的空气质量等级。

In [8]:

df_bj2015.describe()

Out[8]:

	Year	Month	Day	Hour	Value
count	8663.0	8663.000000	8663.000000	8663.000000	8663.000000
mean	2015.0	6.538728	15.674939 11.494055		82.733810
std	0.0	3.448080	8.787195	6.938605	88.556186
min	2015.0	1.000000	1.000000	0.000000	0.000000
25%	2015.0	4.000000	8.000000	5.000000	21.000000
50%	2015.0	7.000000	16.000000	11.000000	54.000000



⚠
内容举报

	Year	Month	Day	Hour	Value
75%	2015.0	10.000000	23.000000	18.000000	109.000000
max	2015.0	12.000000	31.000000	23.000000	722.000000

In [9]:



以下使用DataFrame的apply函数,增加一个字段'Grade',并且根据'Value'字段的值,调用前面的转换函数,为'Grade'字段赋值。

In [10]:

```
df_bj2015.loc[:, 'Grade'] = df_bj2015['Value'].apply(get_grade)
```

再来看看增加字段后的数据集

In [11]:

```
df_bj2015.head()
```



⚠
内容举报

Out[11]:

	Year	Month	Day	Hour	Value	QC	Grade
0	2015	1	1	0	22.0	Valid	Good
1	2015	1	1	1	9.0	Valid	Good
2	2015	1	1	2	9.0	Valid	Good
3	⁷ 2015	1	1	3	13.0	Valid	Good
4	2015	1	1	4	10.0	Valid	Good

 \vee

经过以上的处理,我们就把北京2015年的PM2.5数据集处理完成了。对于其他三个表格,我们都采取类 **(図)**的处理。

In [12]:

```
# 从csv中读取北京2016年pm2.5数据,保存在DataFrame中

df_bj2016 = pd.read_csv('data/Beijing_2016_HourlyPM25_created20170201.csv')

# 重新设定各个字段的名称

df_bj2016.columns=['Year','Month','Day','Hour','Value','QC']

# 对错误和缺失值赋为空值

df_bj2016.loc[df_bj2016.Value<0,'Value']=np.nan

# 删除空值记录
```

删除全值记录

df_bj2016.dropna(inplace=True)

#增加空气质量等级字段'Grade'

df_bj2016.loc[:, 'Grade'] = df_bj2016['Value'].apply(get_grade)

In [13]:

df_bj2016.head()

Out[13]:



⚠
内容举报

	Year	Month	Day	Hour	Value	QC	Grade
0	2016	1	1	0	231.0	Valid	Very Unhealthy
1	2016	1	1	1	239.0	Valid	Very Unhealthy
2	2016	1	1	2	205.0	Valid	Very Unhealthy
3	2016	1	1	3	167.0	Valid	Unhealthy
4	⁷ 2016	1	1	4	132.0	Valid	Unhealthy for Sensi

In [14]:

In [15]:

df_gz2016.head()

Out[15]:

	Year	Month	Day	Hour	Value	QC	Grade
0	2016	1	1	0	55.0	Valid	Moderate
1	2016	1	1	1	58.0	Valid	Moderate
2	2016	1	1	2	59.0	Valid	Moderate



⚠
内容举报

内谷伞:

		Year	Month	Day	Hour	Value	QC	Grade
	3	2016	1	1	3	58.0	Valid	Moderate
	4	2016	1	1	4	51.0	Valid	Moderate

In [16]:

In [17]:

df_gz2015.head()

Out[17]:

	Year	Month	Day	Hour	Value	QC	Grade
0	2015	1	1	0	38.0	Valid	Good
1	2015	1	1	1	40.0	Valid	Good
2	2015	1	1	2	40.0	Valid	Good
3	2015	1	1	3	34.0	Valid	Good
4	2015	1	1	4	42.0	Valid	Good



⚠
内容举报

TOP

好了,我们已经把本次分析所需的数据集全部准备完毕,接下来就是进行数据的分析

<未完待续>

ďЪ

0

发表你的评论

(http://my.csdn.net/weixin 35068028)

ಹ

相关文章推荐

架构设计:系统存储(25)——数据一致性与Paxos算法(下)(http://blog.csdn.net/yinwenj...

从本文开始,笔者将花三到四篇文章的篇幅,介绍Paxos算法。包括它的理论基础、基本实现、变种实现,其它保证最终一致 性的算法,等等。...



》yinwenjie (http://blog.csdn.net/yinwenjie) 2017年03月20日 18:34 €2538

机器学习与深度学习(一)决策树算法 (Decision Tree) (http://blog.csdn.net/tz_zs/article/de...

tz zs学习笔记 决策树算法: 决策树 (decision tree) 是一个类似于流程图的树结构:其中,每个内部结点表示在一个属 性上的测试,每个分支代表一个属性输出,而每个树叶结点代表...





 $\hat{\mathbb{A}}$ 内容举报

TOP



票选结果:Python再上天,微软要求全员学Python?

宇宙语言Python荣登年度排行榜,吴恩达,微软纷纷为它站台,Python这么牛逼的原因是....

(http://www.baidu.com/cb.php?c=IqF pvfqnHmknjnvPjc0IZ0qnfK9ujYzP1nYPH0k0Aw-

5Hc3rHnYnHb0TAq15HfLPWRznjb0T1dbPyndrHI-PvDkmyF9Pvf40AwY5HDdnHfzrHD3nWT0IqF 5y9YIZ0IOzquZR8mLPbUB48uafElAaspvnETZ-YpAa8nWadlAdxTvadThP-

5yF UvTkn0KzujYk0AFV5H00TZcqn0KdpyfqnHRLPjnvnfKEpyfqnHc4rj6kP0KWpyfqP1cvrHnz0AqLUWYs0ZK45HcsP6KWThnqnW0zrjT)

python爬虫抓取全国pm2.5的空气质量(2015.12.21版)(http://blog.csdn.net/djd1234567/a...

这 (次) 这个编码格式真的是很闹心啊,看来真的得深入学习一下编码格式,要不这各种格式错误。 这个编码还和编辑器有关系,最 开始的时候实在sublime Text里编辑的代码,运行起来卡卡的,特别顺畅,但突然发现它...





皇后区空气质量数据分析.zip (http://download.csdn.net/download/qq_378...

(http://download.c

2017年12月21日 20:04 41.87MB

Qt实现天气预报与PM2.5监测系统(9)空气质量(http://blog.csdn.net/Tony_Shen/article/de...

下载

Qt实现天气预报与PM2.5监测系统(9)空气质量 程序运行界面 程序代码 头文件 #ifndef SHOWAQI H #define SHOWAQI H #include #inclu...



Tony_Shen (http://blog.csdn.net/Tony_Shen)

2017年04月25日 11:34



3.10/只 家直销ABS防水接线 盒85x85x50mm 电线



1.00/只 家直销塑料防水接线 盒 户外塑料防水盒



1.98/只 现货供应LED 3/4/6/8/10/12口串联接



Ŵ 内容举报

TOP

用Echarts 制作全国城市空气质量(PM2.5)自己的源代码(http://blog.csdn.net/Chen_yanxi...

String path = request.getContextPath(); String basePath = request.getScheme()+"://"+request.getServ...



Chen yanxia (http://blog.csdn.net/Chen yanxia) 2014年04月24日 16:44

iQS版本PM2.5空气质量监控仪 (http://blog.csdn.net/gianguozheng/article/details/44078257)

前言鉴于柴静版《苍穹之下》的火爆,或许更多的人会关心环境,空气质量等环保因素,参考某学院的Swift版本的PM2.5的 获取,本人自己写个Objective-C版本的。基本原理从服务器获取当前的空气质量...



_qianguozheng (http://blog.csdn.net/qianguozheng) 2015年03月05日 09:47



PM2.5空气质量仪液晶版-IAP15W4K58S4单片机方案 (http://download.csdn...

(http://download

2017年03月05日 10:23 18.42MB





基于谷歌地图的Web应用开发示例——基于AOD、AQI数据的北京市空气质量...

(http://download.c

2014年12月28日 22:46 17.24MB

下载

Python爬虫实践:获取空气质量历史数据 (http://blog.csdn.net/u013337691/article/details/5...

利用BeautifulSoup库构建一个简单的网络爬虫,从天气后报网站抓取兰州空气质量历史数据。...



u013337691 (http://blog.csdn.net/u013337691) 2016年07月13日 08:38 **1034**

使用R语言实现的城市空气质量分析模型 (http://blog.csdn.net/jarry046/article/details/430543...



 \mathbb{A} 内容举报

TOP

这个题是这学期数学建模的课题,不过鉴于本人数学比较渣而且时间精力有限,实现的思路比较传统和简单,用到了多元线性回归和主成分分析来求解模型,这也是统计学习方法里面比较常见的一种,正好在上个学期学习了R语...



中国空气质量分析 (http://blog.csdn.net/qisamuelzhang/article/details/29605627)

由于不懂如何自律,中国人现在享受经济发展的同时却需要担心空气、水、食品等必需品。特别是空气问题,促成了口罩与空气滤清器等行业,却也极大影响了人们的工作生活。在这么下去最火的乐队绝对是Air Suppl...



城市空气质量可视化分析实现 (http://download.csdn.net/download/lhclust...

2017年08月22日 09:26 2.53MB

下载



Java实现BP神经网络,实现对空气质量的分析和评级(http://download.csdn..

(http://download.

2016年12月14日 20:22 266KB

下载

使用 Cloud Insight SDK 监控北京空气质量! (http://blog.csdn.net/wangpeng198688/article...

最近刚好在用一款系统监控工具 Cloud Insight,它提供的 SDK 可以把任一数据上传到他们那做展示。 灵机一动,作为一个程序员,自己动手丰衣足食,没什么不能解决的。 pip instal...





基于GIS下的空气质量分析系统方案的研究 (http://download.csdn.net/down...

(http://download / 2012年08月13日 21:55 5.5MB



⚠
内容举报

下载

就北京来说,有对 PM2.5 有用的空气净化器么? (http://blog.csdn.net/u013676568/article/de...

如果是属于已装修完住宅,上海市消改委曾经做了一个调查和测试,对空气净化器的效果进行了一个比对,由上海市环境保护产品质量监督检验总站,在专家论证的基础上依据GB 4706.45 - 2008《家用和类似用途...



科普:关于PM2.5(危险的北京空气) (http://blog.csdn.net/huhao_bupt/article/details/7043...

因为得了鼻炎,开始关注北京的环境状况。于是就上网查了下前一段时间广为流传的PM2.5事件。 搜到了美国大使馆的网站:"我到了美国使馆空气质量监测站的页面。看到了今天北京的空气质量指数(AOI)报告(能...

huhao_bupt (http://blog.csdn.net/huhao_bupt) 2011年12月05日 19:45 👊2152

Android-JSON数据解析之空气质量检测 (http://blog.csdn.net/AAAAA_Sean_M/article/detai...

一、JSON基础1.JSON优点:格式简单,占宽带小,多语言支持2.JSON缺点:编码必须是Unicode,语法过于严谨 但也因为JSON语法严谨(如四大基础规则),才能支持多语言 3.JSON四...



🦻 AAAAA_Sean_M (http://blog.csdn.net/AAAAA_Sean_M) 2016年08月13日 15:06 🔲 815

用Java代码抓取"天津环境空气质量数据" (http://blog.csdn.net/maxilbert/article/details/4464...

首先,天津市环境空气质量GIS发布平台的IP地址为:http://air.tjemc.org.cn/ 然后,用HTTP Analyzer进行HTTP包分析,确定http://air.tjemc.o...





⚠
内容举报

TOP