# 新型冠状病毒疫情实时分析

小组成员：张仲浩（2019110456，组长）

张智超（2019110455）

张 涛（2019110454）

杨 鑫（2019110448）

曾 瑞（2019110449）

赵 鹏（2019110459）

邹义林（2019110463）

指导教师： 卢 宇

2021年12月

1. **系统简介（此部分篇幅限制在2页内，超过2页倒扣分！）**
2. 系统是什么？

新型冠状病毒疫情发展情况实时分析。

1. 系统主要功能？
2. 功能1：爬取数据

爬取百度新型冠状病毒肺炎实时大数据报告，存为json（**https://voice.baidu.com/act/newpneumonia/newpneumonia/）**

1. 功能2：数据转换

分析json，转换为xls文件，再将xls文件导出为csv文件

1. 功能3：生成网页（图片）

利用pyechart库将数据转化为疫情地图

1. 功能4：数据分析

利用pyspark库对数据进行分析得出结论。Rdd 编程和dataframe的使用

得出结论

1. 数据留存

将数据留存，供分析使用

1. 系统主要亮点
2. 亮点1：爬取数据国内按照34个省级行政划分。国外按照所属大洲进行国家情况统计
3. 亮点2：生成疫情地图形象展示
4. 亮点3：在生成xls表的时候就以及划分出热点疫情地区作为单独一栏
5. 亮点4：数据分析
6. 亮点5：学习dataframe模块，另一种方式展示数据
7. 项目组成员及对应的工作

|  |  |  |  |
| --- | --- | --- | --- |
| 学号 | 姓名 | 完成的模块（与设计方案对应） | 对应功能（与第2点对应） |
| 2019110456 | 张仲浩 | 爬虫模块  数据分析  项目构思 | 功能1  功能4 |
| 2019110448 | 杨鑫 | 项目构思  逻辑设计 | 功能1  功能4 |
| 2019110449 | 曾瑞 | 数据转换  数据管理 | 功能2  功能5 |
| 2019110454 | 张涛 | 显示制作  数据分析 | 功能3（echarts）  功能4 |
| 2019110455 | 张智超 | 显示制作  数据分析 | 功能3  功能4 |
| 2019110459 | 赵鹏 | 项目构思  Dataframe展示 | 功能4 |
| 2019110463 | 邹义林 | 项目构思  数据分析 | 模块5 |

（排第1位为组长，原则上按贡献大小排序；做好了，组长最高分，做砸了，组长最低分）

1. **设计方案**
2. 系统组成（由哪些模块构成）
   1. 模块1

爬取网站数据并进行初步分析

* 1. 模块2

数据转换，生成可使用的数据

* 1. 模块3

数据显示，生成echarts图

* 1. 模块4

数据存储，存入hive

* 1. 模块5

数据分析预测rdd编程

* 1. 模块6

另一种方式展示数据dataframe。

1. 各模块之间的关系

模块1得到的数据提供给模块2和模块3。模块2生成的数据提供给模块4和模块5。模块5可分析来至模块2和模块4的数据。

1. **数据处理关键步骤/模块及说明**
2. 数据爬取。Python爬虫。存为json。
3. 数据初步处理。将json提取存为xls，再转为csv
4. Hive建库建表。建的库为MybigData 建表为data

表结构：area，confirmed，died，crued，curConfirm，confirmedRelative，

依次代表：地区，累计确诊，死亡，治愈，现有确诊，累计确诊增量。第一个为字符型，其余为数值型。初始值为0.主键为area。

1. rdd编程。

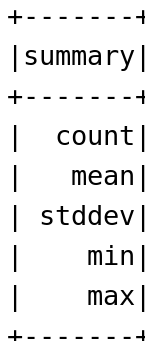
将数据存入rdd，分组排列。

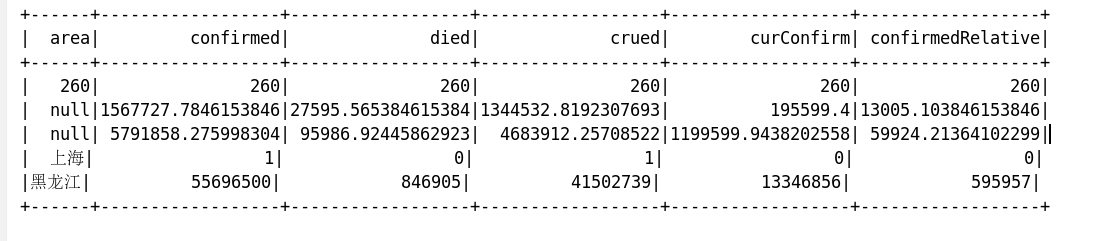
1. dataframe设计

通过dataframe展示，以类表的形式展示数据。能够看出总值（260个地区），平均数，方差，最小值。最大值。



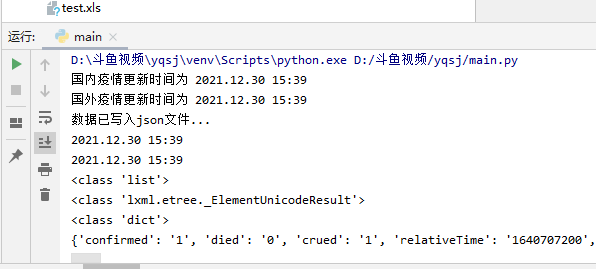
C0表示地区，c1表示累计确诊人数，c2表示死亡人数，c3表示治愈人数，c4表示现有确诊人数，c5表示累计确诊增量。



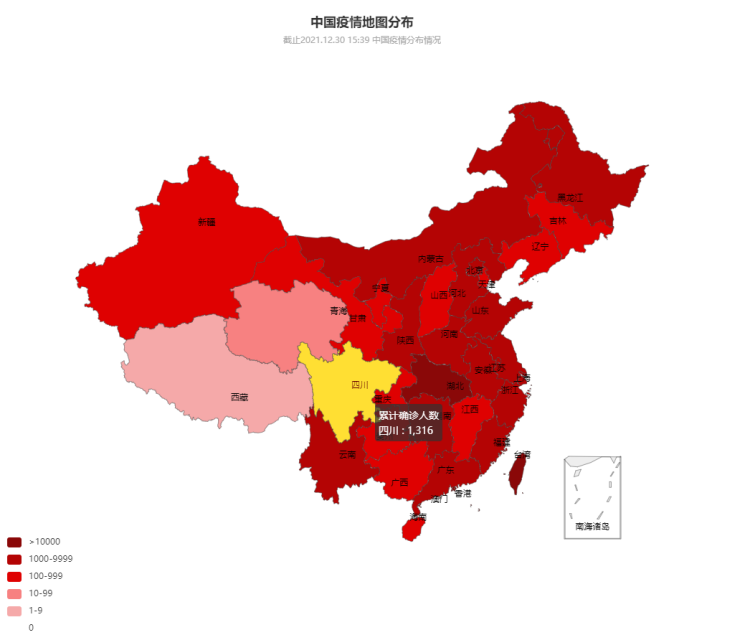
这五行依次表示统计的地区数值（260个地区），平均数，方差，最小值。最大值。

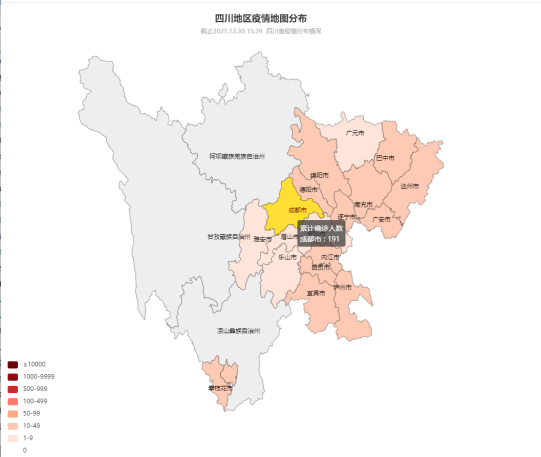
1. **运行效果展示**

数据获取：

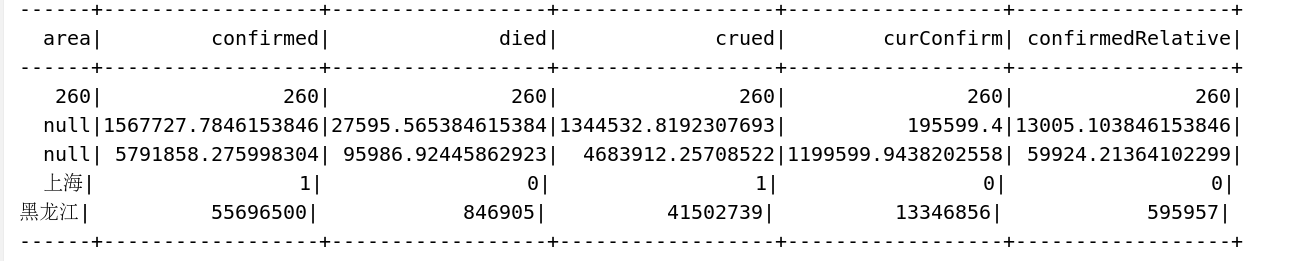


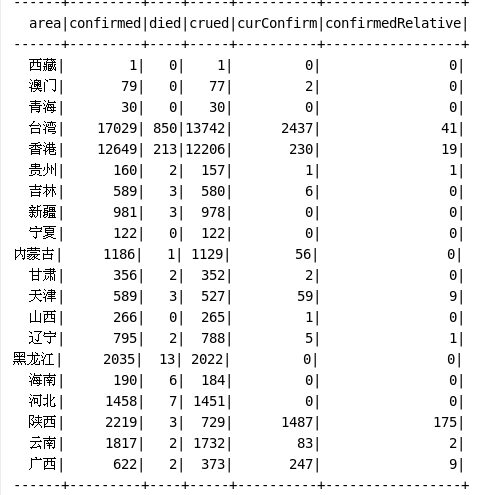
Echarts图片：





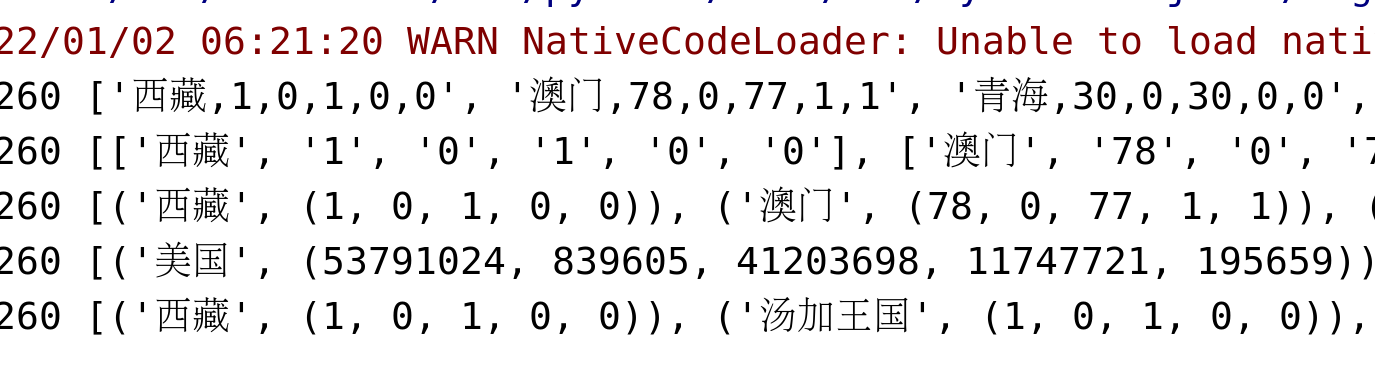
Dataframe：





这五行依次表示统计的地区数值（260个地区），平均数，方差，最小值。最大值。

Rdd编程结果

结论分析：综合分析的数据，得出结论西藏，汤加拉国，密克罗尼西亚三个地区疫情最轻松。美国，印度，巴西，英国，俄罗斯，疫情最为严重，疫情最为严重，累计确诊人数超过千万。美国治愈人数也是最多，反映了其医疗水平高，但其现有确诊也是超千万是最多的国家或地区，每天也是20万以上的确诊增量。中国三十四个省级行政区累计也比这些国家少。所以得出结论，在中国是最安全的，中国的疫情应对优于其他大国。

1. **心得体会总结（有话则长，无话则短）**
2. 解决工具问题可以多样。问题总是奇奇怪怪，但是方法也是多种多样。
3. 不够熟悉python语言，得加强
4. 平时不努力，期末徒伤悲。
5. 想法很大，手很笨。

（第二三四五部分总和超过5页倒扣分！）