前言

我是老王，从事数据分析工作。前期每日知乎上写一篇一页疫情，主要是将卫健委的数据通过可视化图表展现出来。

这篇文档时结合异常检测的方法、思路，来解析卫健委发布的疫情数据。不得不承认，这是比较有难度的过程。这篇文章前后花了一个星期的时间，整理思路、收集和分析数据，调整算法和编写。

1. 发现：

* 卫健委数据包含预防和疾病治疗两种数据量差异的数据。单维检测将主要把密切追踪数据标记为异常。
* 卫健委数据的变动不但和疾控力度有关，也和 《新型冠状病毒肺炎防控方案》的版本有关。 例如：2月8日后，医学观察的疾控数据被多中算法标记为异常。这和版本医学观察或许有一定关系。

1. 对我而言，获得收获是完整的梳理了一次异常检测的过程。

* **前期，数据整理和收集**。我每天早上阅读卫健委的最新报告。将文字整理成数据，录入数据仓库。整理各种维度信息，每天还在知乎写一篇 一页疫情，将卫健委报告用可视化方式展现出来。
* **中期，数据的预处理**。数据有预防，还有病例数据。数量级从几个到60万个，数值有正有负。数据预处理也比较费事，最后发现sklearn.pownertransfer 提供了支持负数的正态化工具。非常有用
* **后期。数据的异常解释。**机器学习的异常检测只管检测不管解释。因此尝试了传统统计工具，例如Mintab这样的工具，得到了一些借鉴。把Pareto图拿过来看那些因子是异常的主要原因。
* **扩展维度。**当日医学观察的解除数量，在多维数据检测中被多个算法顽强的设为异常，无论怎么调整参数，数据预处理。之前一直没有找到线索可以合理的解释这些异常。 就在我快要放弃解释这种异常时，《新型冠状病毒肺炎防控方案》中关于医学观察标准的变化引起我的关注。从2月8日第四版诊疗方案开始，医学观察标准多次修正了。 或许这就是为什么2月9日开始所有算法将解除医学观察数据标记为异常原因之一。
* **进一步分析。**需要将诊疗方案的文字数字化，维度化后，录入数据库中做相关分析。以后有空再分析。

时间有限、数据有限、精力有限，水平有限，写的来的东西更多是个人粗浅的理解和认识。过程和方法、结论更可能有诸多疏漏和差错，敬请各位知友批评指正。

目录

[什么是异常？¶ 3](#_Toc33347447)

[数据概要展示¶ 4](#_Toc33347448)

[维度模型¶ 4](#_Toc33347449)

[主题河流分析¶ 5](#_Toc33347450)

[环境准备¶ 5](#_Toc33347451)

[加载数据¶ 5](#_Toc33347452)

[单维度（从病例数据开始，事实表）¶ 6](#_Toc33347453)

[观察数据¶ 6](#_Toc33347454)

[查看10个数据样本 6](#_Toc33347455)

[统计描述¶ 6](#_Toc33347456)

[单维数据图（总体图、时滞图、自相关图）¶ 7](#_Toc33347457)

[异常检测¶ 8](#_Toc33347458)

[挑战- 寻找单维异常¶ 8](#_Toc33347459)

[算法简介¶ 8](#_Toc33347460)

[检测结果¶ 9](#_Toc33347461)

[异常解释¶ 10](#_Toc33347462)

[帕累托分析¶ 10](#_Toc33347463)

[因子分析¶ 11](#_Toc33347464)

[时间序列观察¶ 12](#_Toc33347465)

[多种维度分析¶ 14](#_Toc33347466)

[准备数据¶ 14](#_Toc33347467)

[观察数据¶ 14](#_Toc33347468)

[10个样本数据¶ 14](#_Toc33347469)

[统计描述¶ 14](#_Toc33347470)

[数据流型图(TNE)¶ 15](#_Toc33347471)

[异常检测¶ 16](#_Toc33347472)

[挑战 - 寻找长表中的异常¶ 16](#_Toc33347473)

[检测结果¶ 17](#_Toc33347474)

[异常解释¶ 17](#_Toc33347475)

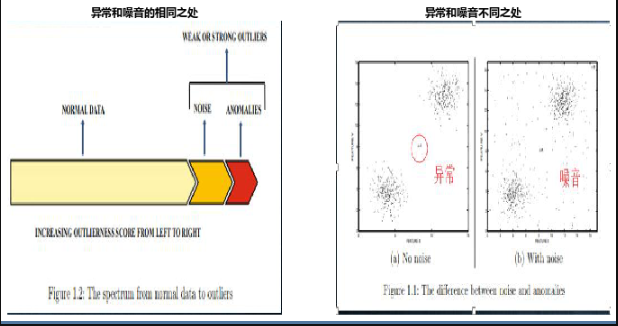
[时间序列观察¶ 19](#_Toc33347476)

[小结¶ 20](#_Toc33347477)

# 什么是异常？[¶](#什么是异常？)

所谓异常，即其明显和其它样本不同（偏差），并且该数据样本很可能是一个不同机制产生的数据。以数据噪音和异常比较可以发现

* 相同之处 - 偏离正常值，于其他样本不同
* 不同之处 - 异常是由不同机制产生的。噪音可能只是波动 异常和误差不一样。 下图展示了数据误差带来的噪音和偏差带来的数据异常的差异



# 数据概要展示[¶](#数据概要展示)

新型冠状病毒肺炎给每个人的春节带来了巨大的影响。

从过年拜年聚会到今年过年不串门，串门只串自家门，卧室门，客厅门，阳台门，厨房门，厕所门。这个疫情国家高度重视，举全国之力支援武汉，全国之人过年不外出。每一天疫情的变化，都牵动了人们的心弦。

从2020年1月21日，国家卫健委每天都会公开发布上一日的疫情数据报告。各大媒体，网络平台也都只做了各种各样的疫情动态，实时数据，方便网民了解。

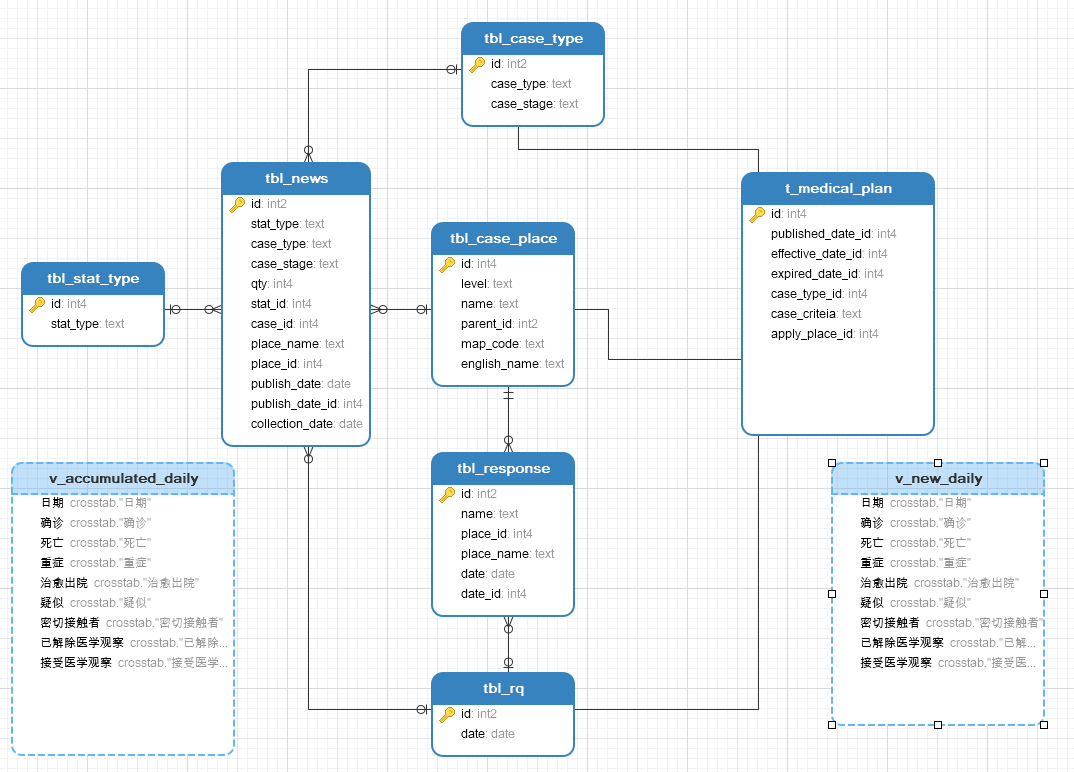
我自己每日阅读卫健委报告，并且每日都在知乎上写一页疫情进行可视化分析。

这些报告时文字形式，通过整理和归纳，就可以将其纳入数据仓库种进行后期的分析。

数据仓库是可以通过建立星型或者雪花型的数据维度表，事实表，来管理和分析数据。

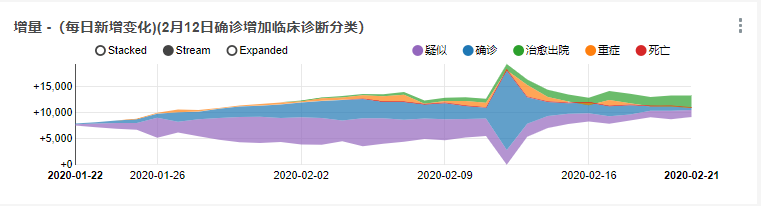
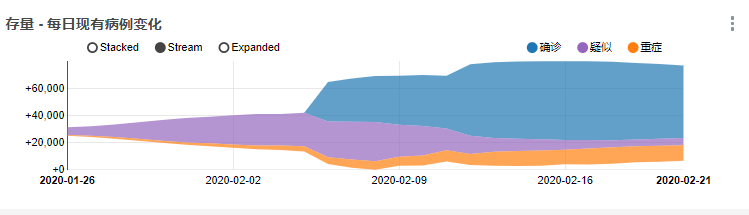
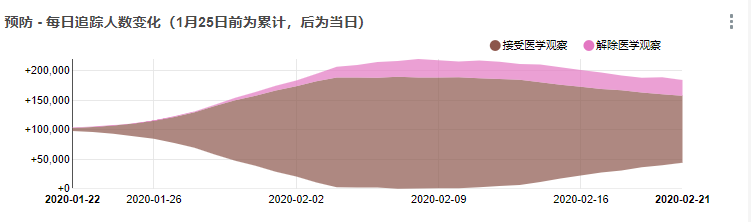
下面时比较原始和粗超的维度和事实表

## 维度模型[¶](#维度模型)



## 主题河流分析[¶](#主题河流分析)

通过常见的可视化工具，例如superset等，可以制作可视化数据面板，动态观察疾控数据的变化情况。之前我每日发布的一页疫情文章就这样的数据分析报告。可以定制：

* 每日新增疫情，
* 当日疫情数据，
* 预防数据等
* 

# 环境准备[¶](#环境准备)

后台采用Postgresql 作为数据库，数据分析使用IBIS,Pandas等获取和存储Postgresql的数据仓库内容

## 加载数据[¶](#加载数据)

# 单维度（从病例数据开始，事实表）[¶](#单维度（从病例数据开始，事实表）)

## 观察数据[¶](#观察数据)

首先对数据进行常规的描述性分析。，均值、标准差，最小，最大，25%，50%，75%分位的数值是多少？ 从下面可以发现，共有394个数据。最大574418，最小-628. 均值是92010.8

### 查看10个数据样本

([ 106, 76288, 72436, 11633, 4928, 170, 1230, 5794,

5248, 213, 1072, 2788, 15, 11272, 744, 9239,

3887, 600, 6973, 5072, 2110, 26724, 6242, 12552,

1323, 8969, 3996, 55748, 361, 14376, 3536, 38800,

28985, 28942, 124, 3971, 1087, 189660, 2590, 26,

1868, 47833, 1239, 189583, 252154, 29307, 314028, 109,

1309, 9692, -156, 52526, 849, 188183, 135881, 259,

31774, 393, 17205, 254, 11741, 1081, 51, 23589,

589163, 6101, 86, 18457, 187519, 3248, 2649, 3219,

3399, 4812, 871, -113, 440, 29429, 325, 126363,

34546, 186045, 444, 10844, 6188, 2102, 428438, 118478,

37, 5173, 80, 21556, 3281, 571, 89, 1367,

47, 185556, 26, 1370], dtype=int32)

从这些数据，可以发现数据有从个位数到六位数都有。有大有小，肉眼无法看出规律，也不可能发现异常情况

### 统计描述[¶](#统计描述)

count 433.000000

mean 38719.076212

std 100297.840506

min -628.000000

25% 444.000000

50% 2656.000000

75% 24324.000000

max 618915.000000

Name: qty, dtype: float64

从统计描述表中可以看到数据：

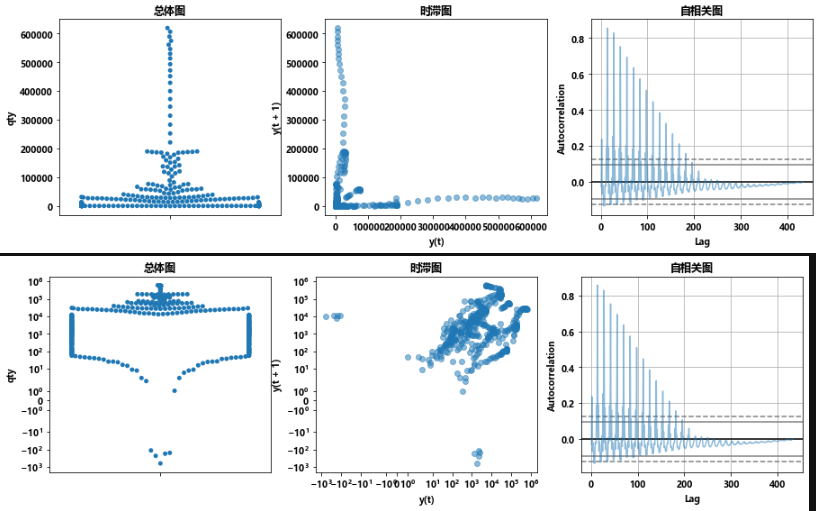
* 有419条。
* 最大超过60万，
* 最小是-628，
* 均值是37784，
* 标准差是97709。

标准差（9.7万）\*3，约为29万。根据3个西格玛经验法则，大于29万的数值非常可能是异常数据。

### 单维数据图（总体图、时滞图、自相关图）[¶](#单维数据图（总体图、时滞图、自相关图）)

接下来用图形化的方式观察单维数据。这里选用三种：

* 总体图。类似于增强的茎叶图，可以像观察大树的枝叶一样观察数据的分布情况。知道那里多，那里少。
* 时滞图。用当前数值和下一个序号数指数作为x,y轴坐标。可以发现数据产生的先后顺序和数据本身之间的关系，
* 自相关图。可以用来查看数据是否可自己的过去相关，或者说存在周期性。此外除了使用原始数据外， 还使用log10将数据进行了平滑处理。

从上面的图形可发现：

* 总体图，
  + 数据正常上限约为20万，但是个别数值非常大，分布在20-60万区间
  + 数值下限约为10，个别数值为负数从分布在-100到-100之间。此为不一样的机制产生的数据。
* 时序图，
  + 前后相邻的数据存在数量级差异，例如上一个数据是百级，下一个数据可能就是十万级。说明数据的计量单位不一样。
  + log化的时滞图中数据呈现增长态势。从左下角往右上角的斜线
  + 有负数
* 自相关图中
  + 发现明显的周期性。周期固定
  + 随着数据的增长，在第150个数据后，数据的自相关性不显著了。也就是说，10天后的数据，和10天前的关系不大了。

尽管只有单维数据，但是也挖掘出来卫健委每日疫情报告若干关键信息，即：

* 数量级差异。 应为有预防数据（10万级），也有病例数据（百，千级）
* 周期性。 每天发布核心数据约14个。即每14个数据为1个周期（1天）
* 增长性。 疫情还是阻击战。预防和疾控数据每日都在增长。尽管增速已经减慢。
* 负数。 新增重疾的病例数，有几天是负增长，即负数。（核减所致）

## 异常检测[¶](#异常检测)

### 挑战- 寻找单维异常[¶](#挑战--寻找单维异常)

虽然已经将卫健委的数据按照日期，类别、统计口径等整理好了类别维度，以及病例数量作为事实维度。 但是我们首先从事实维度开始。即假设没有维度信息，没有任何数据的说明，只有事实数据，能否发现异常？ 如果能，能发现那些类型的异常？

#### 方法：[¶](#方法：)

* 1. 数据。按照单列的形式构建事实数据。
     + 事实： 卫健委每日发布报告中的疾控病例或人群的数值
     + 维度： 异常检测时不使用维度数据。尽在后期图形化解释时候引用。
  2. 将异常点的阈值设置为0.05，即5%.即不超过5%的数据，将被标记为异常数据。
  3. 采用多种异常检测算法比较，HBOS, Isolated Forest, CBLOF， PCA
  4. 考虑到数值存在数量级差异，且有负数，采用sklearn.powertransfer提前进行标准化预处理
  5. 按行检测，检测结果包含：
     + 算法名称
     + 检测结果（0，1） 0代表正常，1代表异常
     + 检测分值 越大，约异常。注：在同一个算法内，分值比较有意义。
     + 检测批次号 。注：在同一个算法内，分值比较有意义。
     + 检测批次号

### 算法简介[¶](#算法简介)

通过4种常用的异常检测算法，展示异常检测的结果。这四种算法分布是：

* HBOS, 基于频数直方图的无监督异常点检测算法
* Isolated Forest, 孤立森林法，
* CBLOF，基于聚类的局部因子
* PCA, 主成分分析

这四种算法在大、中、小数据机上均可以使用。尽管我个人比较偏爱孤立森林法，将其他异常检测算法列出来作为参考。截至2月9日，我只收集了疫情数据394条，数据量小。在不包含类别数据是， 这些算法的异常检测能力类似。 下面是检测的结果输出

In [9]:

import dxy\_etl

imp.reload(dxy\_etl)

od\_df = dxy\_etl.MVOD\_detect(df[["qty","id"]],id\_col="id",is\_transfer=True)

df['shiftp1'] = df.qty.shift(1)

tmp\_df = df.set\_index("id").join(od\_df.set\_index("id")).sort\_index()

tmp\_df.head()

HBOS, HBOS(alpha=0.1, contamination=0.05, n\_bins=10, tol=0.5)

IForest, IForest(behaviour='old', bootstrap=False, contamination=0.05,

max\_features=1.0, max\_samples=128, n\_estimators=100, n\_jobs=1,

random\_state=None, verbose=0)

CBLOF, CBLOF(alpha=0.9, beta=5, check\_estimator=False, clustering\_estimator=None,

contamination=0.05, n\_clusters=5, n\_jobs=1, random\_state=None,

use\_weights=False)

PCA, PCA(contamination=0.05, copy=True, iterated\_power='auto', n\_components=None,

n\_selected\_components=None, random\_state=None, standardization=True,

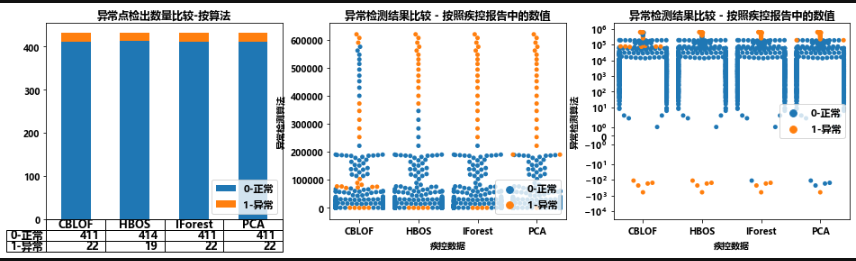
svd\_solver='auto', tol=0.0, weighted=True, whiten=False)

OD results (1732, 5)exported to databaseEngine(postgres://postgres:\*\*\*@192.168.1.200:15432/wh\_coronavirus),job\_name=0112

Out[9]:

|  | **stat\_type** | **case\_type** | **case\_stage** | **qty** | **stat\_id** | **case\_id** | **place\_name** | **place\_id** | **publish\_date** | **publish\_date\_id** | **collection\_date** | **update\_time** | **shiftp1** | **name** | **result** | **label** | **job\_name** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **id** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **3** | 累计 | 解除医学观察 | 追踪 | 817 | 2 | 7 | 中国 | 1 | 2020-01-21 | 2 | 2020-01-20 | NaT | 1739.0 | HBOS | 0.559616 | 0 | 0112 |
| **3** | 累计 | 解除医学观察 | 追踪 | 817 | 2 | 7 | 中国 | 1 | 2020-01-21 | 2 | 2020-01-20 | NaT | 1739.0 | IForest | -0.251225 | 0 | 0112 |
| **3** | 累计 | 解除医学观察 | 追踪 | 817 | 2 | 7 | 中国 | 1 | 2020-01-21 | 2 | 2020-01-20 | NaT | 1739.0 | CBLOF | 0.033236 | 0 | 0112 |
| **3** | 累计 | 解除医学观察 | 追踪 | 817 | 2 | 7 | 中国 | 1 | 2020-01-21 | 2 | 2020-01-20 | NaT | 1739.0 | PCA | 0.553819 | 0 | 0112 |
| **4** | 累计 | 确诊 | 病例 | 291 | 2 | 1 | 中国 | 1 | 2020-01-21 | 2 | 2020-01-20 | NaT | 77.0 | HBOS | 0.559616 | 0 | 0112 |

### 检测结果[¶](#检测结果)



所有算法均发现了异常数据，异常检测比较图显示

* 相同点是 大数：绝大多数40万以上的数值被各个算法标记为异常。 负数：绝大多数负数被标记成了异常。
* 不同点是
  + HBOS 检测出较少的异常（19个）。
  + 其余算法均按照比例输出了22个（433\*5%）的异常数据
    - IForest 和HBSO检测结果几乎一直。但是比HBOS多标记几个大数异常点。
    - PCA 未将所有负数标记成异常
    - CBLOF将10万左右的一些数值标记成异常

考虑到不存在标签数据，也不存在所谓的正确结果。 上面的检测结果恰恰反映了不同算法的特点。

* HBOS，Isolated Forest 擅长找极端异常。加权求分位数后，检测出来的异常点，数值一定是极值。通常发现的异常点也会被其他算法标记成异常
* CBLOF, 可以找全局极值，而且能发现局部的稀疏点。
* PCA 擅长的基于特征值方法的异常发现

各有所长

## 异常解释[¶](#异常解释)

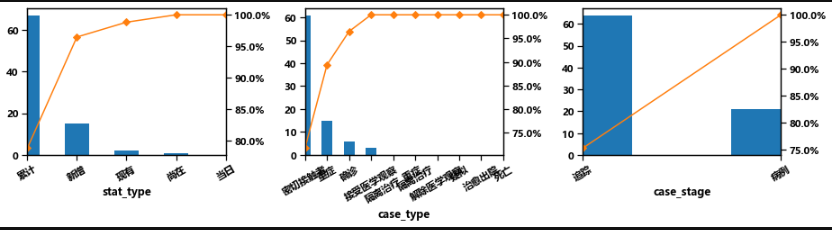
在单维数据异常检测中，尽管不了解这些数据的属性或者维度信息，例如日期，病例类型，统计口径，预防还是治疗等类别，异常检测算法还是可以将异常数据给检测出来。

为了解释这些数据，就必须通过各种各样的专业知识，数据探索，将数据进行分类。按日期，类别，统计等分类。

分类后就可以通过帕累托分析，因子分析等常规统计方法，查看异常的主要因素是什么，各个因子的贡献度等。 为后期的多维异常检测打下基础。

### 帕累托分析[¶](#帕累托分析)

我们将单维数据检测就发现的异常数据和原有的维度数据关联在一起进行帕累托分析。查看异常数据主要来源于那些类型：

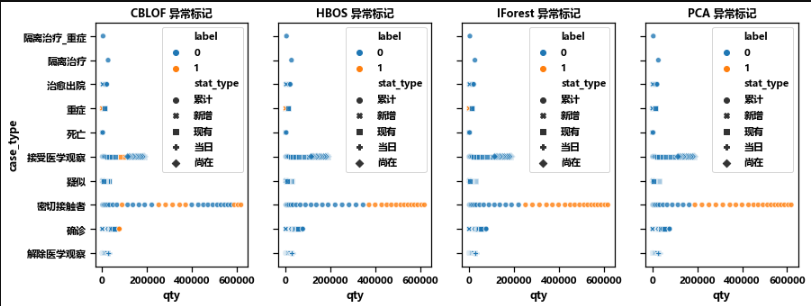


通过Pareto图可以了解到：

* 按照统计口径来看， 累计的数据贡献了80%的异常数据量。
* 按照病例类型来看， 密切接触者、重症、确诊贡献了超过95的异常数据量
* 按照病例阶段来看， 预防阶段贡献了75%的异常数据量。

### 因子分析[¶](#因子分析)

接下来，综合异常检测算法、病例类型、统计口径来看病例的异常数据量主要是发生在什么因子上？



从上面的因子分析图中，可以发现：

* 相同之处：
  + 密切接触者：几乎所有算法均将大多数密切接触者数值超过40万标记成异常。密切接触者是累计口径，并且按照万，或者10万增长。和其他数据不在一个数据级别。
  + 重症：HBOS, Isolated, CBLOF 算法将重症某些数据标记为异常。（最近新增重症为负数，即不在增加，而是减少了）
* 不同之处：
  + 确诊：CBLOF 算法将12月16日之后累计确证数据标记成异常
  + 重疾：PCA算法没有标记任何重症为异常，即使时负数
  + 接受医学观察：CBLOF将部分数据标记为异常。

综上所述，数据采集机制不同，数量级不同，符合异常检测的基本原则：与众不同，产生机制不同。

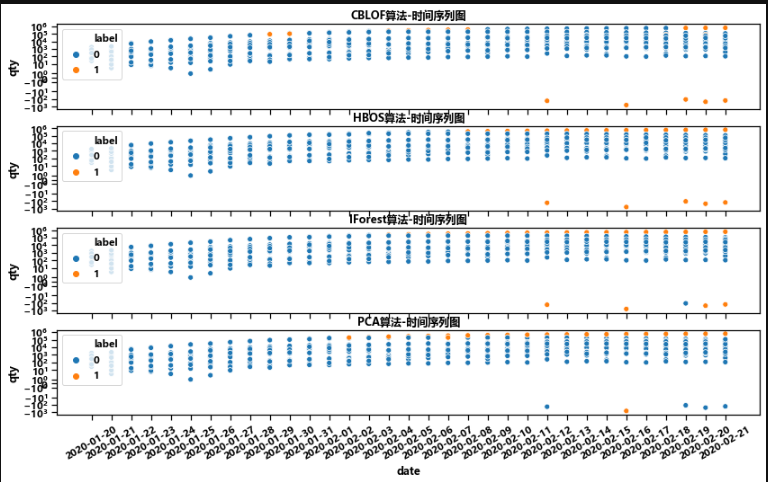
### 时间序列观察[¶](#时间序列观察)

至此，为什么疾控数据正在接受医学观察数据和重疾中有的被标记异常，还没有得到解释

我们把这些疾控数据用时间序列的方式来观察，来解释这些的异常点。

方法：

* 我们把将X轴设成每日，Y轴设置成各种疾控数据的数值。
* y轴对数化方式为(symlog)



从上面的时间序列看来，有两种异常：

* 负数。 2月12日后，出现了多次负增长的数据（新增重疾）。所有算法均发现这个新的数据特点，并且标记成异常。
* 超大数。 密切接触人数，绝大多数算法将2月6日以后的数据标记成异常
* 数值跳变异常。1月28/29日，出现累计医学观察人数的快速上升。这处是一个稀疏的跳变点，CFLOF算法将这两日的医学观察数据标记为一场检测出此处局部异常。

单维异常检测总结。 优点：

* 从数值本身的大小，稀疏来进行异常检测。
* 不具备行业知识和背景情况下检测，作为异常数据挖掘的起点。
* 可以在不具备维度信息时，可以不断通过异常数据挖掘找出数据中未知的维度。 缺点：
* 没有利用数据的维度信息，假定所有数据独立，无维度关联
* 没有利用数据的周期信息，假定所有数据独立，无周期关联
* 没有利用数据的模式信息，假定所有数据独立，无模式关联

# 多种维度分析[¶](#多种维度分析)

如果从数据仓库的角度来看，排列不同的事实数据（即可以进行有意义的加法运算的数据）。可以有两种数据视图类型。即长表和宽表。

* 所谓长表，就是每行只有一列或者少数列的事实数据。每行只有一种维度信息。例如： 时间，类型，数量格式的表格。表格列数比较少 行数比较多。
* 所谓宽表，就是每行有非常多个事实数据。维度信息已经转换成在表格列上。 例如：时间，类型1数量，类型2数量，类型3数据，。。。。类型N数量。

从数据分析角度来看。长表的行数量更大，前期数据预处理量较少。 宽表组合了多个维度的信息，已经隐含了周期，关系或者模型在内，但是会出现较多的空值，需要投入较多的精力进行数据预处理，和数据模式和关系的组合。

下面将用长表方式进行异常分析。 数据还是卫健委每日发布的全国疫情数据

多个维度列加上1个事实列。即将卫健委的数据按照日期，类别、统计口径加入作为维度信息，以及病例数量作为事实维度。

## 准备数据[¶](#准备数据)

## 观察数据[¶](#观察数据)

## 10个样本数据[¶](#10个样本数据)

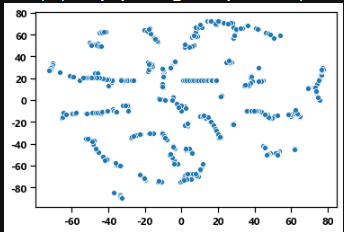
|  | **publish\_date** | **collection\_date** | **stat\_type** | **case\_type** | **case\_stage** | **qty** | **qty\_3d\_rm** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **286** | 2020-02-12 | 2020-02-11 | 新增 | 治愈出院 | 病例 | 744 | 697.0 |
| **199** | 2020-02-06 | 2020-02-05 | 现有 | 重症 | 病例 | 3859 | 3289.0 |
| **354** | 2020-02-17 | 2020-02-16 | 新增 | 重症 | 病例 | -628 | 147.0 |
| **143** | 2020-02-02 | 2020-02-01 | 累计 | 确诊 | 病例 | 14380 | 11954.0 |
| **171** | 2020-02-04 | 2020-02-03 | 累计 | 死亡 | 病例 | 425 | 363.0 |
| **80** | 2020-01-28 | 2020-01-27 | 累计 | 死亡 | 病例 | 106 | 81.0 |
| **329** | 2020-02-15 | 2020-02-14 | 累计 | 治愈出院 | 病例 | 8096 | 6910.0 |
| **288** | 2020-02-12 | 2020-02-11 | 新增 | 疑似 | 病例 | 3342 | 3629.0 |
| **21** | 2020-01-23 | 2020-01-22 | 累计 | 重症 | 病例 | 95 | 98.0 |
| **327** | 2020-02-15 | 2020-02-14 | 现有 | 重症 | 病例 | 11053 | 9762.0 |

### 统计描述[¶](#统计描述)

|  | **collection\_date** | **stat\_type** | **case\_type** | **case\_stage** | **qty** | **qty\_3d\_rm** |
| --- | --- | --- | --- | --- | --- | --- |
| **count** | 433 | 433 | 433 | 433 | 433.000000 | 433.000000 |
| **unique** | 33 | 5 | 10 | 2 | NaN | NaN |
| **top** | 2020-02-05 00:00:00 | 新增 | 确诊 | 病例 | NaN | NaN |
| **freq** | 15 | 168 | 82 | 334 | NaN | NaN |
| **first** | 2020-01-20 00:00:00 | NaN | NaN | NaN | NaN | NaN |
| **last** | 2020-02-21 00:00:00 | NaN | NaN | NaN | NaN | NaN |
| **mean** | NaN | NaN | NaN | NaN | 38719.076212 | 36695.355658 |
| **std** | NaN | NaN | NaN | NaN | 100297.840506 | 96225.492400 |
| **min** | NaN | NaN | NaN | NaN | -628.000000 | -167.000000 |
| **25%** | NaN | NaN | NaN | NaN | 444.000000 | 386.000000 |
| **50%** | NaN | NaN | NaN | NaN | 2656.000000 | 2511.000000 |
| **75%** | NaN | NaN | NaN | NaN | 24324.000000 | 23809.000000 |
| **max** | NaN | NaN | NaN | NaN | 618915.000000 | 604705.000000 |

### 数据流型图(TNE)[¶](#数据流型图(TNE))

为了把多维数据用一张图来显示， 在观察数据的稀疏情况时，可以通过tnse流型化处理，我们可以把保持原有数据行数，转换列数为2列或者更多列。



## 异常检测[¶](#异常检测)

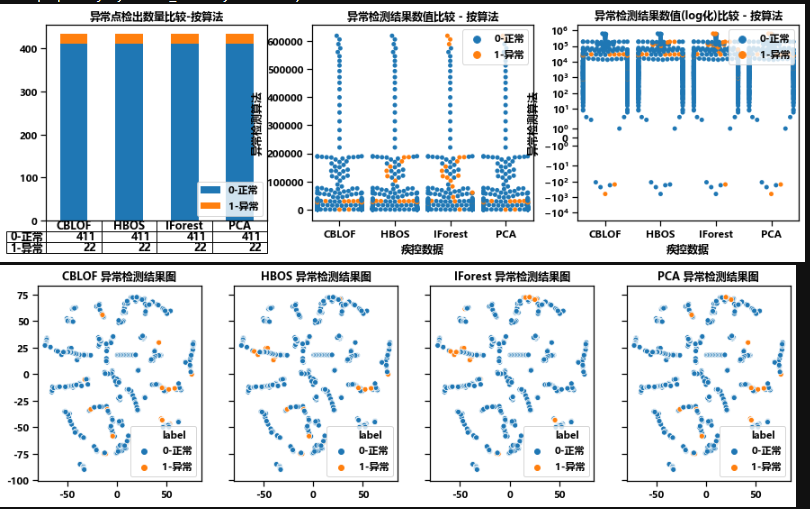
### 挑战 - 寻找长表中的异常[¶](#挑战---寻找长表中的异常)

能否发现异常？ 如果能，能发现那些类型的异常？ 和之前单一维度发现的异常有什么不同？

#### 方法[¶](#方法)

* + **数据。按照长表的形式构建维度数据和事实数据**
    - * 事实（2个）：
        + 当日数值：卫健委每日发布报告中的疾控病例或人群的数值
        + 3日平均数值：将相同维度信息的数据过去3天的值取平均值（注：默认不含当日。窗口大小为3日，最小为0）
      * 维度，将报告中的文字信息转化成维度表
        + 日期：发布日期，采集日期
        + 统计口径：新增、累计、当日、正在
        + 病例类型：疑似、确诊、重疾、死亡、治愈出院
        + 病例阶段：密切追踪人口，正在医学观察，接触医学观察
  + **阈值。**对每一行的数据进行异常检测，将异常行的阈值设置为0.05，即5%.即不超过5%的数据，将被标记为异常数据。
  + **算法。**算法和之前相同，还是CBLOF, HBOS, Isolated Forest, PCA
  + **预处理。**考虑到数值存在数量级差异，且有负数，采用sklearn.powertransfer提前进行标准化预处理
  + **检测结果**。检测结果包含：
    - * 算法名称
      * 检测结果（0，1） 0代表正常，1代表异常
      * 检测分值 越大，约异常。注：在同一个算法内，分值比较有意义。
      * 检测批次号

### 检测结果[¶](#检测结果)



所有算法均发现了多维数据的异常点，异常检测比较图显示

* 相同点是
  + 都检出了22个异常点。
  + **流型图没有给出特别有帮助的图解**
* 不同点是：
  + 数值大小，不再是判断标准。当添加维度信息后，异常检测时，数据被划分成更多的小块。检测在不同的维度组合内进行。而不仅仅依靠数值的大小和稀疏了。
  + 数值正负，也不再是主要判断标准。绝大多数的新增重疾减少数没有在被标记为异常数据

**小结：**

多维数据检测和单维数据检测结果不一样。在结合事实数据+维度信息后的异常检测结果不仅和数值相关，更与维度信息相关。

接下来，我们用维度信息来解释这些异常

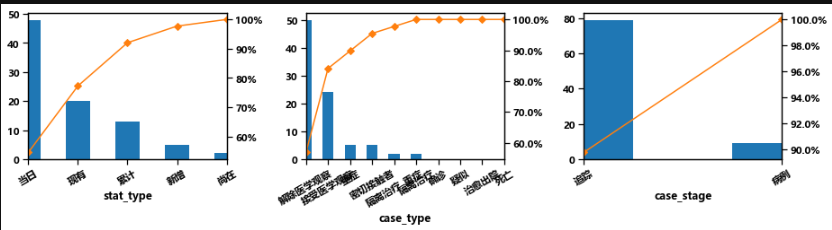
### 异常解释[¶](#异常解释)

在多维数据异常检测中，通过组合事实数据和维度数据，例如日期，病例类型，统计口径，预防还是治疗等类别，异常检测算法可以检测出更加有意义、更加有价值的异常结果。

为了解释这些数据，就需要迭代的方式。不断发现新异常，找到不能解释的异常数据，再通过各种各样的专业知识，数据探索，将数据进行分类。按日期，类别，统计等分类，建立新的维度信息表。再通过新的维度，创建新的事实数据。进行新一轮的异常检测。知道满意或者不在继续投入时间和精力。

异常检测的结果和维度信息结合，还可以通过帕累托分析，因子分析等常规统计方法，查看异常的主要因素是什么，各个因子的贡献度等。 为后期的多维异常检测打下基础。

#### 帕累托分析[¶](#帕累托分析)



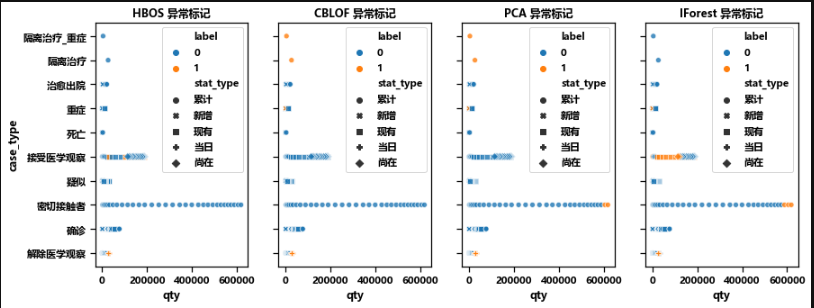
如相关性所示，异常数据标记原因主要与下列因素相关：

* 疾控的口径 当日和现有月战略80%的数据约占了80%的检出率。
* 疾控的病例类别
  + 解除医学观察
  + 接受医学观察
  + 重症

**说明**

* 在目前卫健委的报告，网上相关新闻，还有网上诸多分析中。较少见到接受医学观察和解除医学观察的描述。
* 在多次调整算法，增加3日滚动平均之后。各个算法还是依然将当日解除医学观察，现有接受医学观察和重症 列为了主要的异常数据原因。
* 原本已经要放弃对医学观察的异常原因解释，在文章发布之前，我留意到了《新型冠状病毒肺炎防控方案》的版本差异，从第四版开始（2月8日发布），对医学观察和解除医学观察进行了密集调整。
* 如果能将各个版本的诊疗方案建立参考维度信息，将有利于发现新治疗方法颁布的异常数据变化。

#### 因子分析[¶](#因子分析)



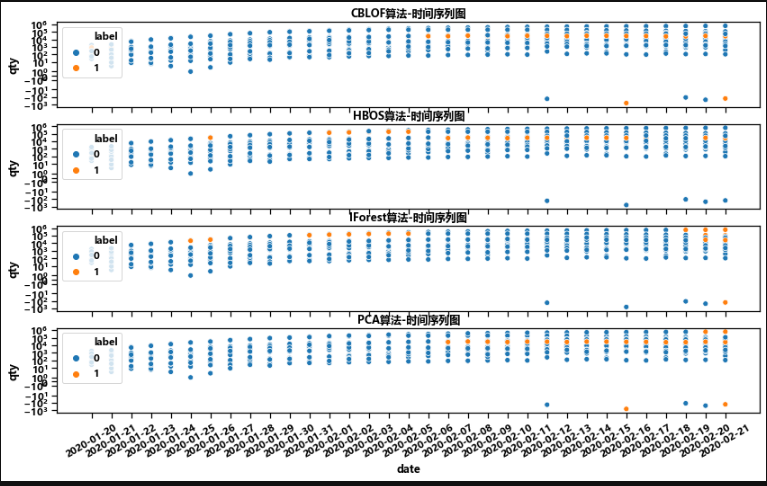
这张图中列出了多个维度，并且把正常点标记为蓝色，异常点标记为黄色。恰好其中几个因子，我观察到不一样的地方：

* 罕见口径： 隔离治疗\_重症、隔离治疗。 这两个口径下的数据被PCA, CBLOF 标记为异常。 这个口径只有2月6日发布的报告中使用过一次。
* 重症: 21日的新增重症负增长。

### 时间序列观察[¶](#时间序列观察)

把这些疾控数据用时间序列的方式来观察，可以发现类似的异常点。即：

* 在右上方的是数据最大的疾控数据，也就是前面提到的每日累计接触者（从1月20日后的数值表大）。



从上面的时间序列看来，有两种异常：

* 负数。 2月15,21日的新增重症是负数，PCA, CB\_LOF算法比较为异常，Isolated Forest仅将21日的新增重疾减少标记成异常。
* 超大数。2月21，22的密切追踪人口被三种算法标记成异常。CBLOF, IForest和 PCA.
* 有关联原因。 2月9日到2月16日期间，解除医学观察数量都被标记为异常。2月8日发布的《新型冠状病毒肺炎防控方案》第四版开始，各个版本对医学观察不断进行修正。也许这是可能因素支持。

## 小结[¶](#小结)

多维异常检测总结。

* 优点：
  + 灵活：
    - 可以组合多个事实数据。原始数据或滑窗数据等
    - 可以组合多个维度。日期、类别、地点、口径等等。
    - 可结合行业经验（例如诊疗方案的变化），添加维度信息，帮助找到有意义的异常
* 缺点：
  + 需要一些行业经验，去解释标记出来的异常数据。
  + 需要知道准备什么事实数据，
  + 异常检测的结果，不同算法结果有差异。需要多个算法综合比对

从卫健委这辆天的报告来看，2020年的疫情已经获得阶段性胜利。从狙击站变成了反击站。 也许下周就要复工，谨以此文预先纪念2020的疫情，并分享给对异常检测感兴趣的同好。

In [ ]:

In [ ]: