数据分析过程（以水资源数据为例）

作者：Loyio （李斯特）

写于2018年5月

pdf离线文档下载

下载地址

目录：

> 本篇文章只做大数据分析比赛（第一组）参考用，如果其中有错误的分析言论，请及时联系我做出更改，水资源的数据太少了，但比较容易展示数据分析过程

一：数据预处理

> 现实世界中数据大体上都是不完整，不一致的脏数据，无法直接进行数据分析，或分析结果差强人意。数据预处理有多种方法：数据清理，数据集成，数据变换，数据归约等。把这些影响分析的数据处理好，才能获得更加精确地分析结果。

1.查看表名

首先拿到数据，先查看有几张表，如水资源分析数据，共三张表

分别是

1. 五年水资源
2. 五年供水用水
3. 五年废水主要污染物排放

2.查看字段名

然后，分别对每个表进行预观察

- 五年水资源
  - 横向数据类型（行）
    - 2014年
    - 2013年
    - 2012年
    - 2011年
    - 2010年
  - 纵向数据类型（列）
    - 人均水资源量（立方米/人）
    - 地下水资源量（亿立方米）
    - 地表水与地下水资源重复量（亿立方米）
    - 地表水资源量（亿立方米）
    - 水资源总量（亿立方米）

> 其余两张表分类相似

通过以上的横向与纵向相互分析，我们就可以得到，这个数据他能体现什么

3.特殊情况

在数据预处理中，我们要注意以下几点：

1. 去除重复无用的数据
2. 对于有残缺的数据可以尝试使用以下两种方法处理
   1. 去掉数据
   2. 通过临近值补全数据
3. 数据的分组要基于实际预分析

比如我在第三张表，也就是 五年废水主要污染物排放 发现了一个问题

指标	2014年	2013年	2012年	2011年	2010年
六价铬排放量(千克)	34925.33	58291.45	70533.6	106395.37	0.0
化学需氧量排放量(万吨)	2294.59	2352.7	2424.0	2499.86	1238.1
废水排放总量(万吨)	7161750.53	6954432.7	6847612.14	6591922.44	6172562.0
总氮排放量(万吨)	456.14	448.1	451.37	447.08	0.0
总磷排放量(万吨)	53.45	48.73	48.88	55.37	0.0
总铬排放量(千克)	132797.43	163117.68	190079.08	293166.34	0.0
挥发酚排放量(吨)	1378.43	1277.33	1501.31	2430.57	0.0
氨氮排放量(万吨)	238.53	245.66	253.59	260.44	120.29
汞排放量(千克)	745.91	916.52	1223.44	2829.15	0.0
石油类排放量(吨)	16203.64	18385.35	17493.88	21012.09	0.0
砷排放量(千克)	109729.85	112230.03	128493.75	146615.97	0.0
铅排放量(千克)	73184.74	76111.97	99358.81	155242.0	0.0
镉排放量(千克)	17251.1	18435.72	27249.89	35898.98	0.0

2010年的数据指标，与其他年份相差太大，这时就应该考虑是统计误差问题（当然误差这个词说的有点不准确），还是真实的情况体现（如果数据是真的，那么我就想是不是2010年与2011年以间是不是有什么很大的新闻）

二：发现数据特征

> 通过第一步对数据的预处理，我们应该能够大概清楚这个数据集的简单的线性关系，当然没有讨论到数据本身

1.拆分表名

例如水资源这个数据集，我们从表名就能大概清楚一些特征（分词法）

表名：`五年水资源`

五年：包含五年的数据，这也是我们第一步得出的行特征，从2010年到2014年的数据

水资源：数据是关于水资源的，如第一步得出的列特征，包含人均水资源量，地下水资源量等数据

然后，我们再次来分析表名：`五年水资源`

很显然，水资源这个词相比五年更容易成为一个关键词

2.获取问题

下一步，我们必须要了解水资源相关的新闻，资料，这样才能对后续的数据分析有帮助

> 如果分析一个数据集，连这个数据集里面讲的是什么都不知道，那真的就是一点数据分析的思路都没有，比如，我们示例数据集中的索赔率分析

日期	保险单号	服务中心	性别	区域	省级	市级	响应状态	源代码	年
2/11/11	1.00127E+11	Beijing	男	西北	甘肃省	兰州	保持中	网络	3(
2/12/11	1.00432E+11	Beijing	女	中南	海南省	海口	保持中	网络	3(
2/13/11	1.00831E+11	Shanghai	女	东北	辽宁省	大连	保持中	手机	2:
3/1/11	1.00427E+11	Beijing	男	华东	浙江省	衢州	保持中	网络	61
3/2/11	10,089,611,922	Beijing	女	华东	山东省	东营	保持中	医生	4(
3/4/11	1.00833E+11	Beijing	女	华北	北京市	北京	保持中	手机	2(
3/13/11	100,420,053	Beijing	男	华东	上海市	上海	保持中	医生	41
3/14/11	1.0012E+11	Shanghai	女	西南	重庆市	重庆	保持中	手机	4(
3/15/11	1.00719E+11	Chongqing	女	中南	广东省	深圳	保持中	网络	3:
3/18/11	10,013,203,868	Shanghai	男	中南	广西省	河池	保持中	手机	47
2/3/11	1.00133E+11	Chongqing	男	中南	湖北省	武汉	等待表单	网络	51
2/4/11	1.0013E+11	Beijing	女	华东	浙江省	杭州	等待表单	网络	5:
2/5/11	10,079,791,545	Beijing	女	华北	内蒙古	呼和浩特	等待表单	网络	2(
2/10/11	1.00133E+11	Chongqing	男	中南	广东省	深圳	等待表单	手机	5(

> 刚开始看到这个数据集时，我是懵逼的，什么说明信息都没有，只是给了我们这么一张excel表，其中有些字段我根本就看不懂（俺只是个敲代码的人）
>
> 比如响应状态，源代码（难道不是程序源代码吗），索赔率，赔付率，完全不懂

响应状态	源代码	年龄	费用	索赔额	赔付额
保持中	网络	36	¥0.76	¥1,159.20	¥693.00
保持中	网络	36	¥17.51	¥837.90	¥630.00

> 当然不能因为有障碍就放弃啊，在接下来的时间，我不断的查阅资料，百度、google，慢慢对这张表有了兴趣，了解如何对数据进行分析

通过在网上搜索，我了解到了我国目前水资源主要问题

1、水资源缺乏；

2、用水浪费严重；

3、水质污染得不到改善；

4、水土流失严重得不到控制；

5、水资源的开发利用不合理；

6、人民节约用水意识不够；

7、政府部门对水资源的管理不是太强。

当然网上查到的，远远不止这些，我只是在这做个例子

3.逐表分析

有了问题，我们分析数据的兴趣自然也就来了

比如第一张表，我们可以简单概况为"从2010年到2014年水资源的资源量情况（减少了还是增多了）"

指标	2014年	2013年	2012年	2011年	2010年
人均水资源量(立方米/人)	1998.64	2059.69	2186.05	1730.2	2310.41
地下水资源量(亿立方米)	7745.03	8081.11	8416.12	7214.5	8417.05
地表水与地下水资源重复量(亿立方米)	7182.75	6962.75	7260.64	6171.4	7308.25
地表水资源量(亿立方米)	26263.91	26839.47	28371.35	22213.6	29797.62
水资源总量(亿立方米)	27266.9	27957.86	29526.88	23256.7	30906.41

第二张表，可以概括为“从2010年到2014年用水量与供水量指标增长与减少情况表’（当然，特征远远不止这些，比如换个角度，只看供水总量和用水总量，可以通俗概括为供了多少水，就用了多少水）

指标	2014年	2013年	2012年	2011年	2010年
人均用水量(立方米/人)	446.75	455.54	454.71	454.4	450.17
供水总量(亿立方米)	6094.88	6183.45	6141.8	6107.2	6021.99
其他供水总量(亿立方米)	57.46	49.94	44.55	44.8	33.12
农业用水总量(亿立方米)	3868.98	3921.52	3880.3	3743.6	3689.14
地下水供水总量(亿立方米)	1116.94	1126.22	1134.22	1109.1	1107.31
地表水供水总量(亿立方米)	4920.46	5007.29	4963.02	4953.3	4881.57
工业用水总量(亿立方米)	1356.1	1406.4	1423.88	1461.8	1447.3
生态用水总量(亿立方米)	103.2	105.38	108.77	111.9	119.77
生活用水总量(亿立方米)	766.58	750.1	728.82	789.9	765.83
用水总量(亿立方米)	6094.86	6183.45	6141.8	6107.2	6021.99

第三张表，也就是那张看起来有点问题的表，我们可以简单概括为“五年间（2010-2014）各污染排放量之间的数量情况”

指标	2014年	2013年	2012年	2011年	2010年
六价铬排放量(千克)	34925.33	58291.45	70533.6	106395.37	0.0
化学需氧量排放量(万吨)	2294.59	2352.7	2424.0	2499.86	1238.1
废水排放总量(万吨)	7161750.53	6954432.7	6847612.14	6591922.44	6172562.0
总氮排放量(万吨)	456.14	448.1	451.37	447.08	0.0
总磷排放量(万吨)	53.45	48.73	48.88	55.37	0.0
总铬排放量(千克)	132797.43	163117.68	190079.08	293166.34	0.0
挥发酚排放量(吨)	1378.43	1277.33	1501.31	2430.57	0.0
氨氮排放量(万吨)	238.53	245.66	253.59	260.44	120.29
汞排放量(千克)	745.91	916.52	1223.44	2829.15	0.0
石油类排放量(吨)	16203.64	18385.35	17493.88	21012.09	0.0
砷排放量(千克)	109729.85	112230.03	128493.75	146615.97	0.0
铅排放量(千克)	73184.74	76111.97	99358.81	155242.0	0.0
镉排放量(千克)	17251.1	18435.72	27249.89	35898.98	0.0

> 刚开始看到这张表的时候，我在想是不是要把化学老师请出来讲讲 😂

对于这些表的数据，我们无需死扣字眼，不要因为里面的一些信息自己不太懂，就不断的要弄懂他们，比如我们无需清楚这些元素哪个更污染水资源，我们只需要知道他们都是污染排放物即可，根据数据做出分析即可

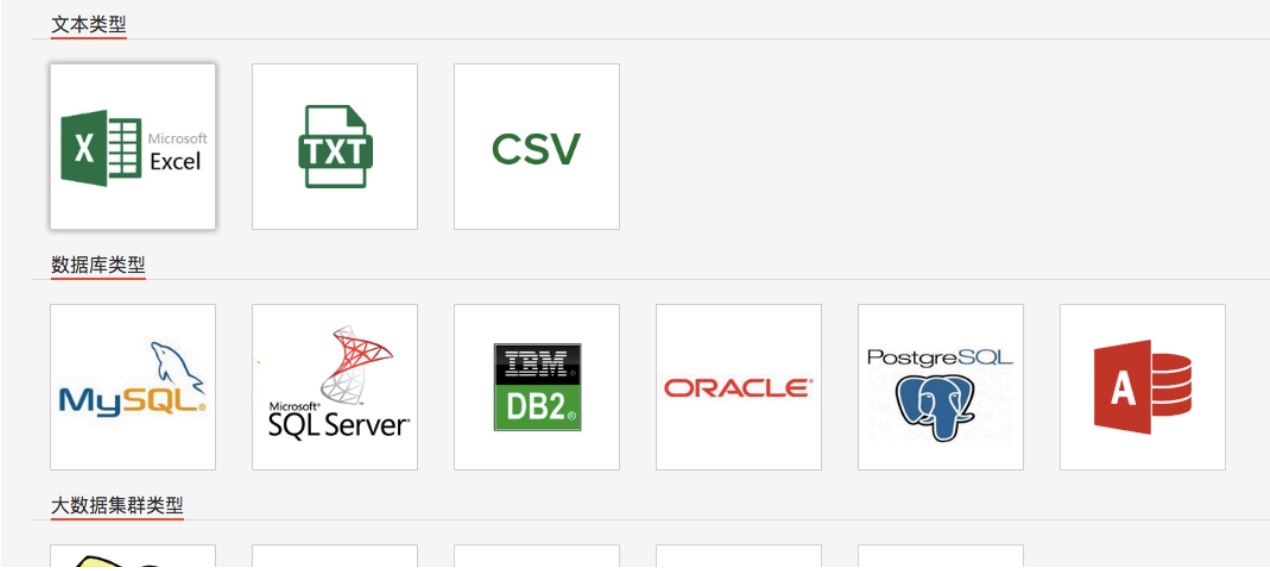
有了问题，和数据的特征，我们就可以动用工具更直观的对数据进行分析

三：使用魔镜进行数据分析

> 本来以为要自己写程序，自己搭建大数据分析平台，自己画图.....
>
> 竟然有了这个可以免费使用的工具，那当然不能错过了

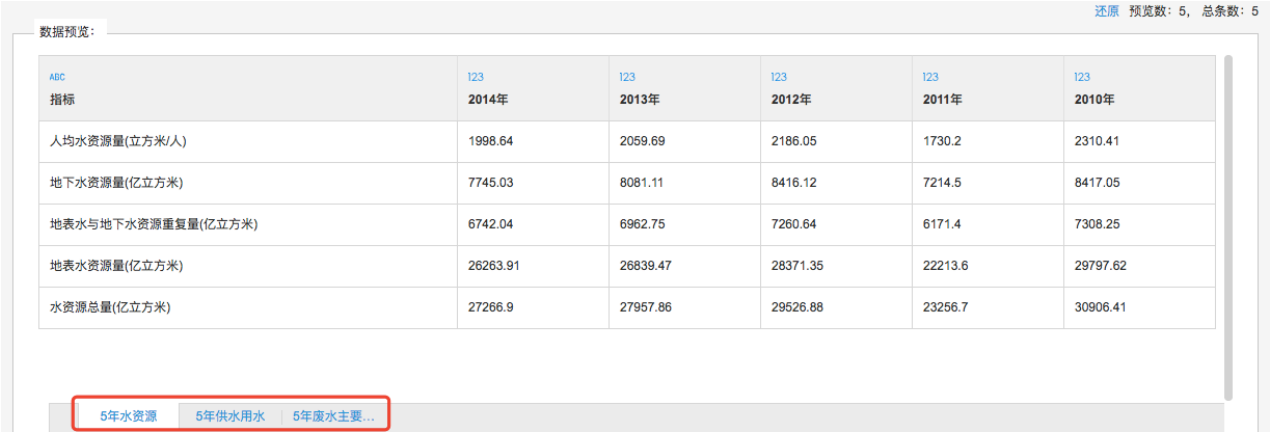
首先打开网站，登录账号，选择新建应用

然后选择数据类型



我们这个数据是xlsx表格，所以我们就选择第一个Excel表格，然后点击下一步

选择文件上传，上传成功后会显示数据预览，查看数据是否齐全，是不是每张表都有



检查完后，输了数据源名称就可以点击保存了

然后，呈现的就是这样一张图

第一次，看到这个的时候，内心😭,(这些都是什么啊，都看不懂啊)

于是想到群里不是有说明文档吗，拿来看看，打开后

- **创建字段**

创建维度字段，点击维度的下拉框，选择创建字段即在下方打开填写字段名的填写框，同时在右侧打开该字段的属性设置框，其中包含关联设置；属性i
等设置。如图12、图13所示。

图12

图13

创建度量字段，在度量下拉菜单中选择创建字段，填写新建度量字段名称，填写完成后打开该度量的属性设置菜单，包括汇总和计数属性设置，

图14

图15

- **创建计算字段**

创建计算字段：通过自定义计算形成新的字段；新的字段在维度、度量列表呈现；新的计算字段可进行删除和编辑操作。

图16

- **创建参数**

点击创建参数菜单，弹出【编辑参数】框。参数类似于维度集，可以切换不同的维度，通过创建参数字段，当在“行/列/标记/筛选器”中时，可以快速切换

图17

- **创建组字段**

点击“创建组”，在页面中央弹出创建组弹框，分组名称：默认名称为【原字段名+（组）】
添加至：若当前没有分组内容，至灰不可操作。选中所需创建的字段值后，点击【分组】，即可将选中的归类到一个新得分组里，新的分组名称默认为【分组N】。如图18所

图18

- **创建分层结构**

拖动字段至另一个字段上会生成创建分层结构弹框，在弹框中可以编辑分成结构名称，点击确认即完成了一个对分层结构的创建，创建完成后在业务对

图19

- **隐藏业务对象和字段**

点击隐藏业务对象或隐藏字段后，该隐藏的业务对象或字段会置灰显示，并且在数据分析中不会显示。如图20所示。

图20

完全对操作没有一点帮助啊，不懂的还是不懂啊

抱着试一试的心态，自己捣鼓了一下这个系统，发现其实还好

1.添加分组

在数据处理，面板，最左边有添加业务分组，和快速分组

这两个选项都可以点，我们在这点击上面的快速分组，然后选择快速分组

然后弹出这样一个面板

注意到，旁边有我们刚刚导入的excel数据表，我前面命名的是五年水资源数据集，还有一个关联表，暂时先不用管

我们双击五年水资源数据集

可以看到，里面正是我们刚开始预览数据的三张表，我们首先拖动**五年水资源**到虚线内

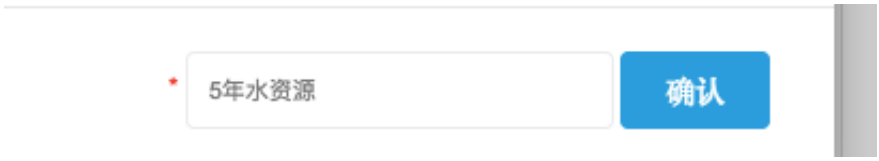
然后下面就会出现表字段等数据

可以看到，他是以第一行的名称做特征字段的

指标	2014年	2013年	2012年	2011年	2010年

注意你在拖动的时候，如果有错误，可以将其脱出，或者还可以点击清空按钮

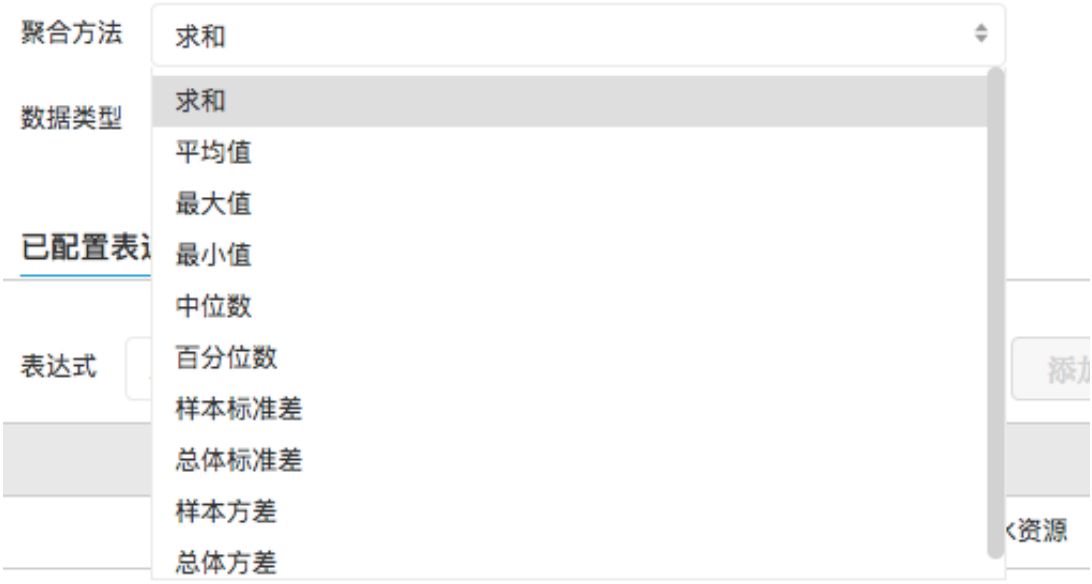
下一步，点击输入名称，确认即可



这个时候在数据处理面板的左侧，就会看到我们添加的业务对象了

我们可以在这对某些属性进行更改，以便数据分析

例如，我们可以更改聚合方法



当然，在这里我们暂时不对这个做出讨论更改

为了简便的展示操作，数据处理页面操作暂时不做过多讨论

2.简单表格

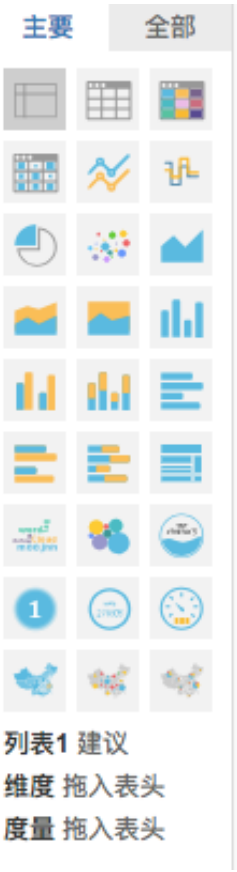
下一步，我们开始正式运用这个系统

首先点击导航栏上的**数据分析**

然后，我们进入了一个全新的界面

最左边的区域，相信都很熟悉，就是我们刚刚添加的业务对象，在这被作为要处理的数据

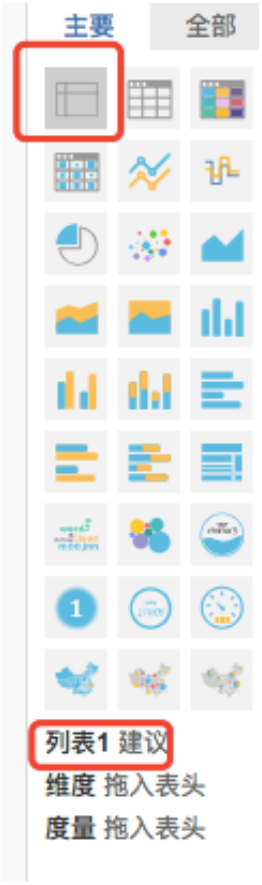
中间我们，暂时不做讨论，最右边可以看到有很多图表类型，分为主要和全部


主要
全部
列表1 建议
维度 拖入表头
度量 拖入表头

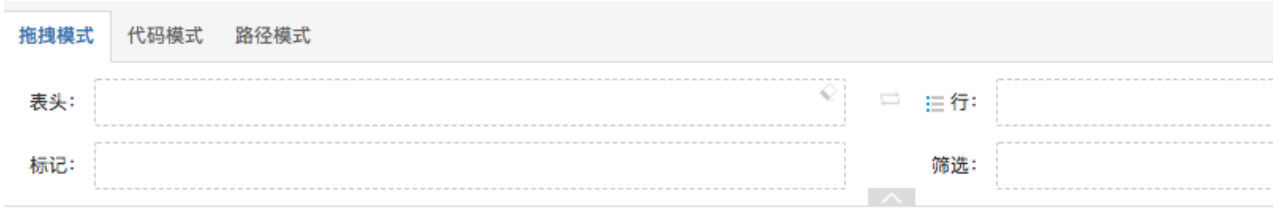

主要
全部
表格
饼图
线图
面积图
条柱图
散点图
树状图
气泡图
标签云
地图
数字图
仪表
更多

区别很明显了，当鼠标悬置在全部中的某个标签时，会有更多详细情况

下一步，我们正式开始来绘制图表，点击主要中的第一个标签，也就是**列表1**



然后在中间的区域，我们可以看到有三个选项



默认拖拽模式，我们从左边的数据中，按照顺序依次拖拽进表头

也就是最后，表头中会有所有的特征，如下图所示

然后我们可以看到，下面的显示区域，变成了表格的形式(而且是不是和之前的excel一模一样)

指标	2014年(汇总)	2013年(汇总)	2012年(汇总)	2011年(汇总)	2010年(汇总)
人均水资源量(立方米/人)	1998.640	2059.690	2186.050	1730.200	2310.410
地下水资源量(亿立方米)	7745.030	8081.110	8416.120	7214.500	8417.050
地表水与地下水资源重复量(亿立方米)	6742.040	6962.750	7260.640	6171.400	7308.250
地表水资源量(亿立方米)	26263.910	26839.470	28371.350	22213.600	29797.620
水资源总量(亿立方米)	27266.900	27957.860	29526.880	23256.700	30906.410

鼠标悬停在表格的单元格上会显示详细信息

指标	2014年(汇总)	2013年(汇总)	2012年(汇总)
人均水资源量(立方米/人)	1998.640	20	
地下水资源量(亿立方米)	7745.030	80	
地表水与地下水资源重复量(亿立方米)	6742.040	6962.750	7260.640
地表水资源量(亿立方米)	26263.910	26839.470	28371.350
水资源总量(亿立方米)	27266.900	27957.860	29526.880

指标：地表水与地下水资源重复量(亿立方米)
2013年(汇总)：**6962.750**

3 选择图表

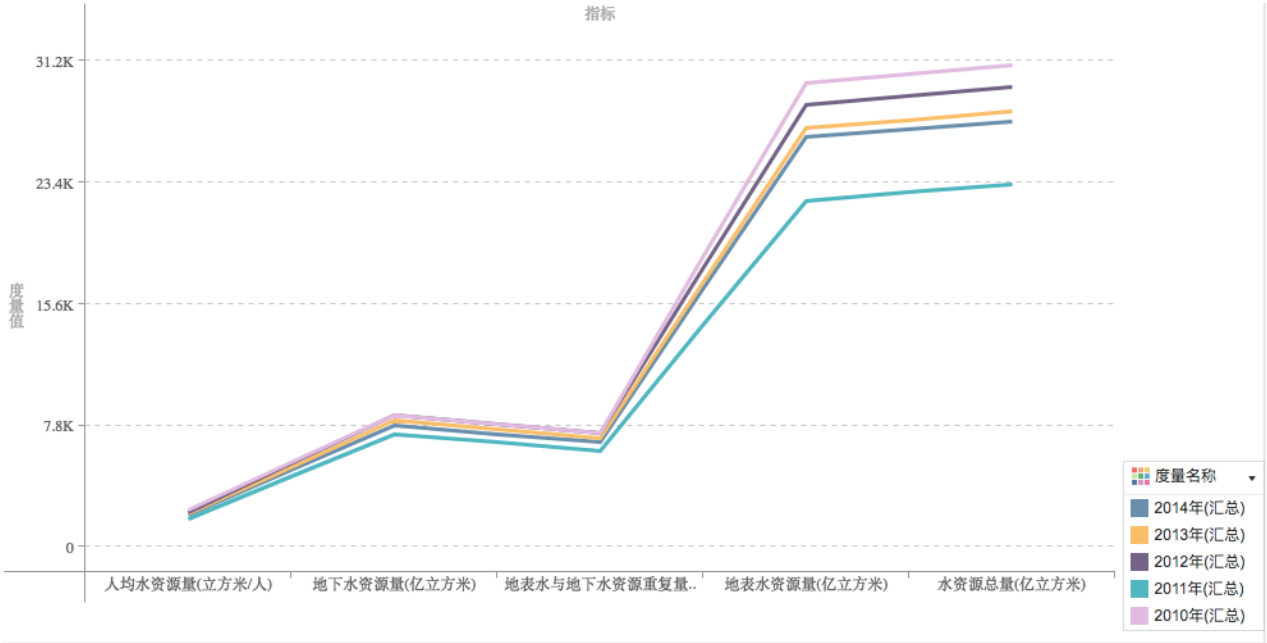
> 基本统计图形的绘制、数据取值的转换、数据的正态化处理等，能够帮助我们掌握数据的分布特征，是进一步深入分析和建模的基础。

完成以上步骤后，就可以进行进一步的分析了，选择合适的图表，以达到可视化数据的目的

回顾我们之前的预处理分析，这个跟年份有关，所以我们要显示出一种梯度的数据图表

以获得这五年的水资源变化情况

比如我们制作的这一张图



实际曲线并无多大用处，尽管我们可以在其中找到一些突破口

通过不断的尝试，最终还是不能正常获得，这个反应时间变换的曲线图

于是我就想到通过对excel表做转置(当然，首次使用这个魔镜平台，难免会有一些问题，之后我们我们就可以根据情况提前做转置)

刚开始的excel表

指标	2014年	2013年	2012年	2011年	2010年
人均水资源量(立方米/人)	1998.64	2059.69	2186.05	1730.2	2310.41
地下水资源量(亿立方米)	7745.03	8081.11	8416.12	7214.5	8417.05
地表水与地下水资源重复量(亿立方米)	6742.04	6962.75	7260.64	6171.4	7308.25
地表水资源量(亿立方米)	26263.91	26839.47	28371.35	22213.6	29797.62
水资源总量(亿立方米)	27266.9	27957.86	29526.88	23256.7	30906.41

转置后的excel表（转置就是切换行和列，数据没有发生变化）（如何转置，可以自行上网搜索）

指标	人均水资源	地下水资源	地表水与地	地表水资源	水资源总量(亿立方
2014年	1998.64	7745.03	6742.04	26263.91	27266.9
2013年	2059.69	8081.11	6962.75	26839.47	27957.86
2012年	2186.05	8416.12	7260.64	28371.35	29526.88
2011年	1730.2	7214.5	6171.4	22213.6	23256.7
2010年	2310.41	8417.05	7308.25	29797.62	30906.41

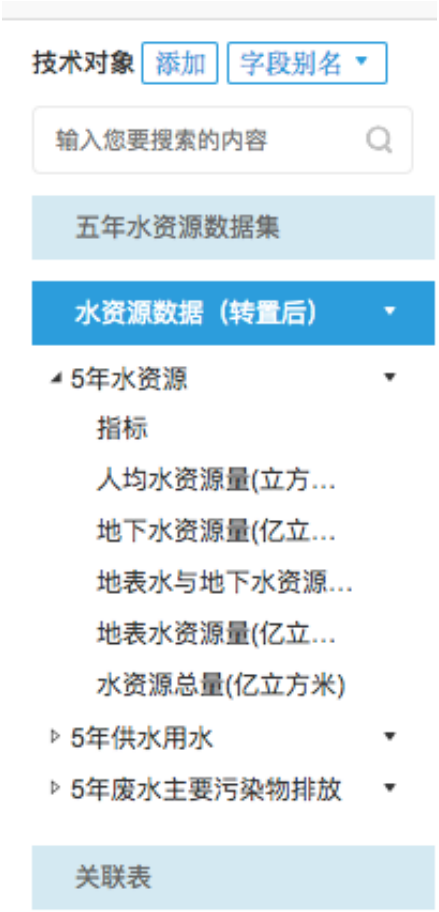
接下来，我们重新上传数据文件(可以看到和之前完全不同)

因为接下来的操作跟之前一模一样，我就不详细附图了

填入名称，点击保存后

我们新建了一个分组

然后从最左边的技术对象中拖入

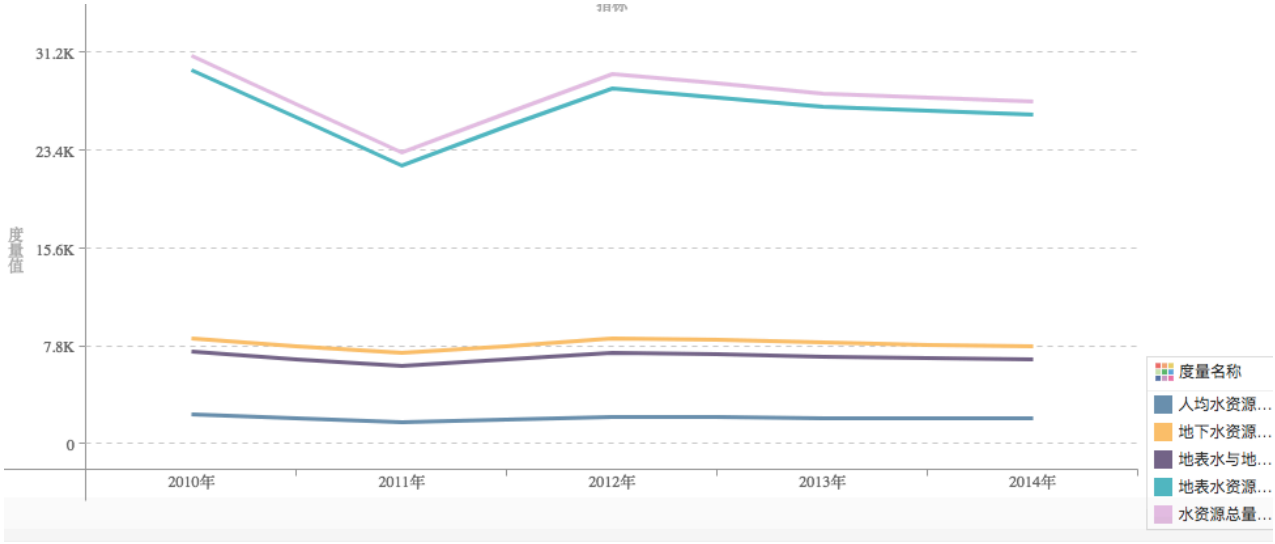


拖入后，显示如下图，（注意每次拖入要点击小勾确认）

拖动数据到列和行等选框中，再选择图表模型，即可得到想要的图表

指标	人均水资源量(立方米/人)(汇总)	地下水资源量(亿立方米)(汇总)	地表水与地下水资源重复量(亿立方米)(汇总)	地表水资源量(亿立方米)(汇总)	水资源总量(亿立方米)(汇总)
2010年	2310	8417	7308	29798	30906
2011年	1730	7215	6171	22214	23257
2012年	2186	8416	7261	28371	29527
2013年	2060	8081	6963	26839	27958
2014年	1999	7745	6742	26264	27267

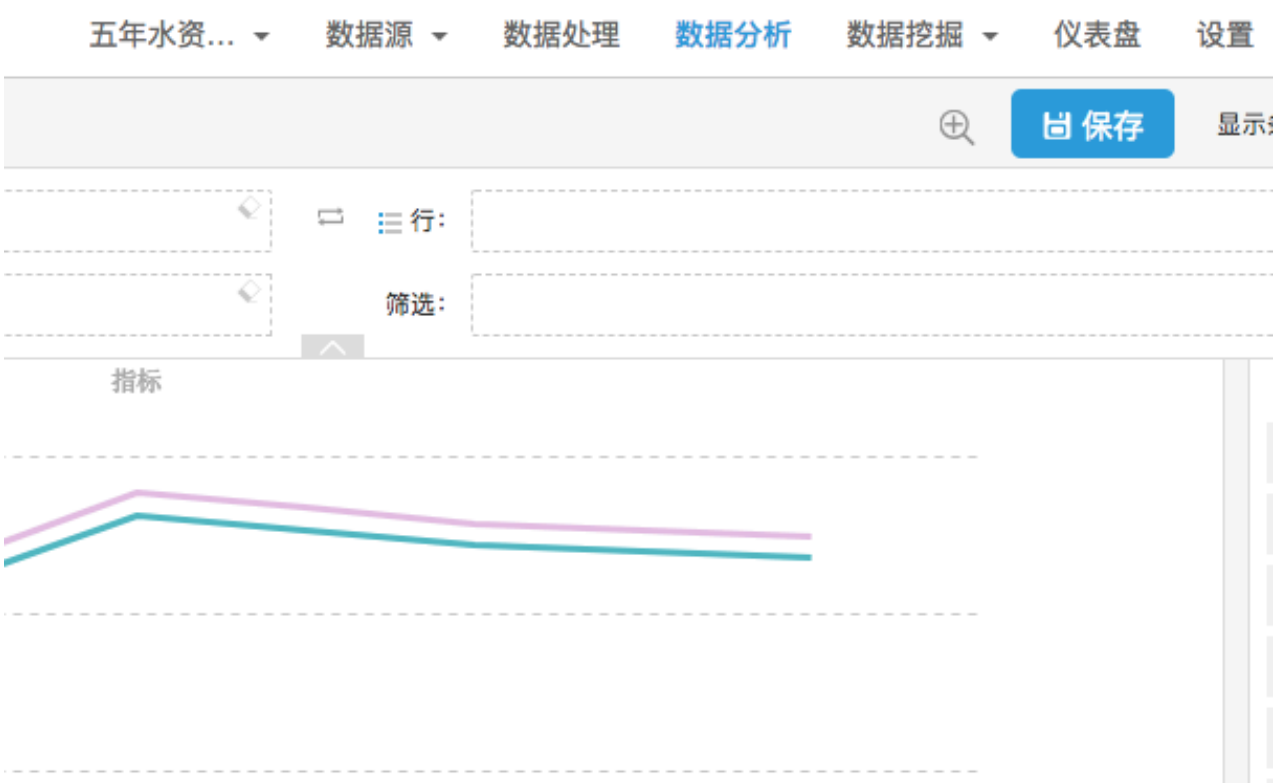
最需要的曲线图



其他图表的绘制我不在这做过多的解释了，多试就会有好结果

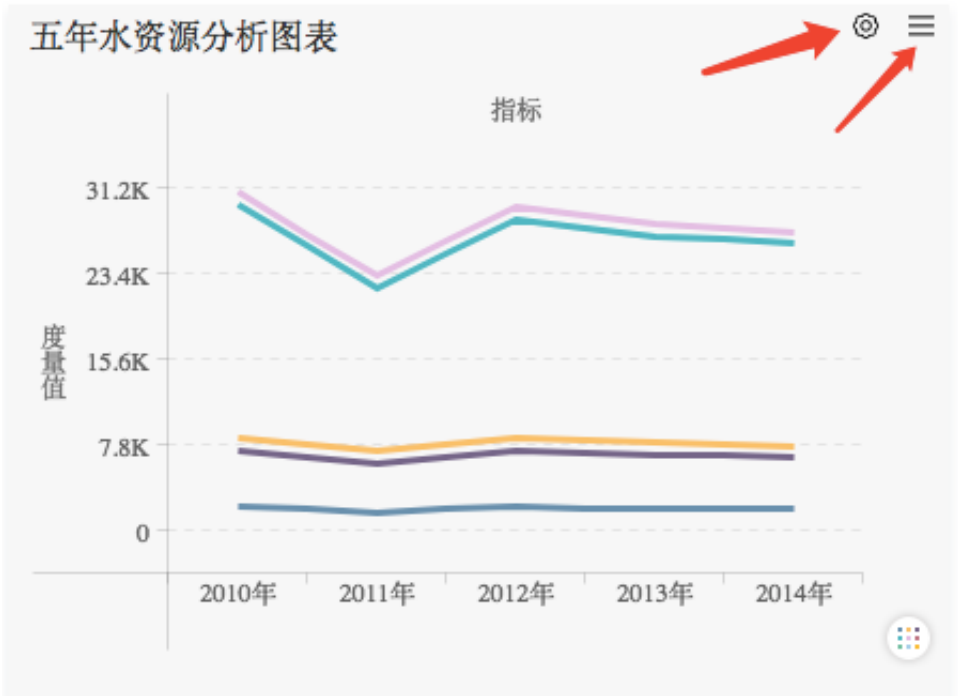
4. 保存图表

对图表满意后，便可以保存图表，以便以后查阅

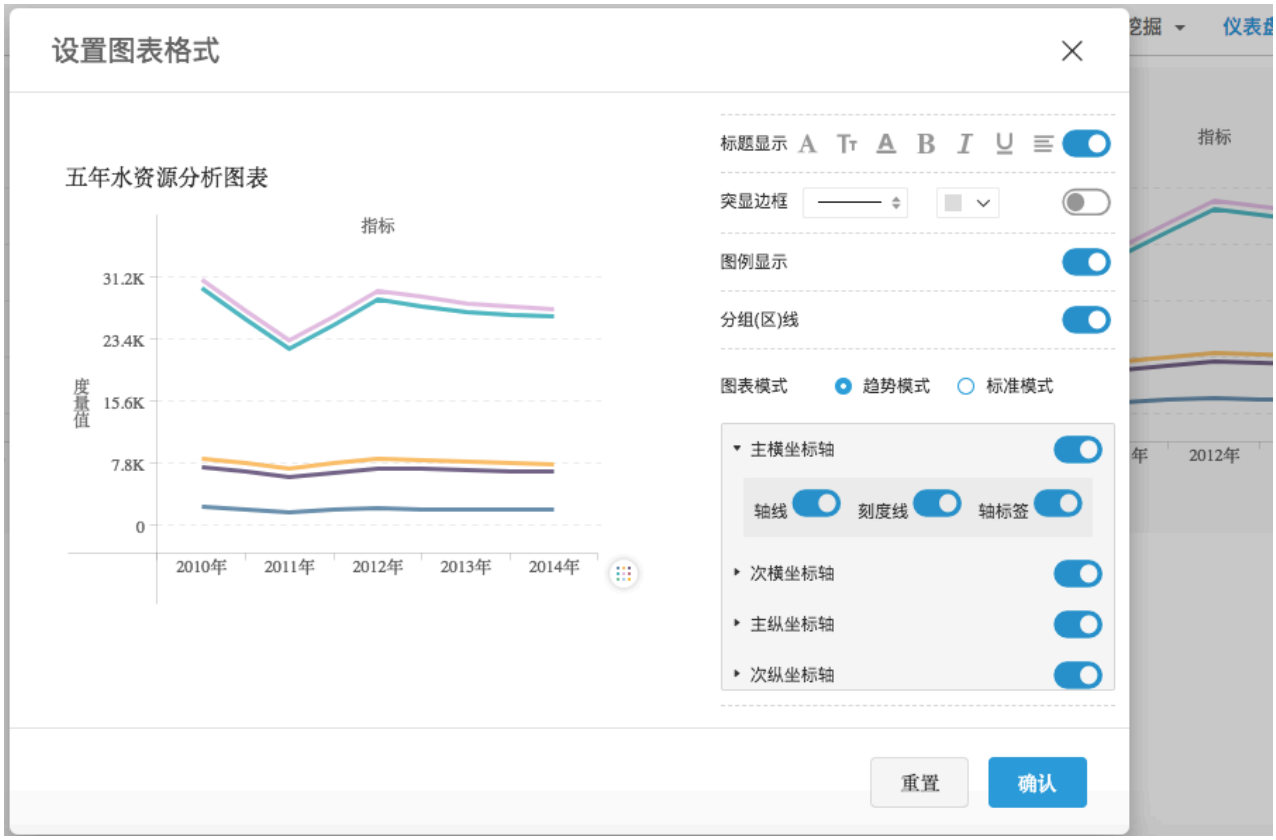


点击保存后，你就会跳到仪表盘，这里面存储的都是你保存好的图表

当鼠标悬置到某个图表上，就会出现两个按钮



第一个是设置图表的一些格式，如下图



第二个，就是一些常用的操作

至此，魔镜的主要操作大约就是这些，接下来就是根据图表进行分析了

四：根据图表进行文字表述

未完待续