自杀率可视化分析

**摘要----自杀长期被视作一个严重的公共安全问题，世界卫生组织曾统计，全球每年约有80万人死于自杀，而企图自杀的人数则要更多。影响自杀率的因素越来越成为社会关注的热点问题。本论文面向自杀率的数据挖掘及其结果的可视化模式开展研究。首先分析自杀率数据集的特点，构建数据模型；其次探讨与自杀率有关的因素；然后设计自杀率与发现结果的可视化表达模式；最后利用可视化软件echart进行相关的可视化和后续的分析。**

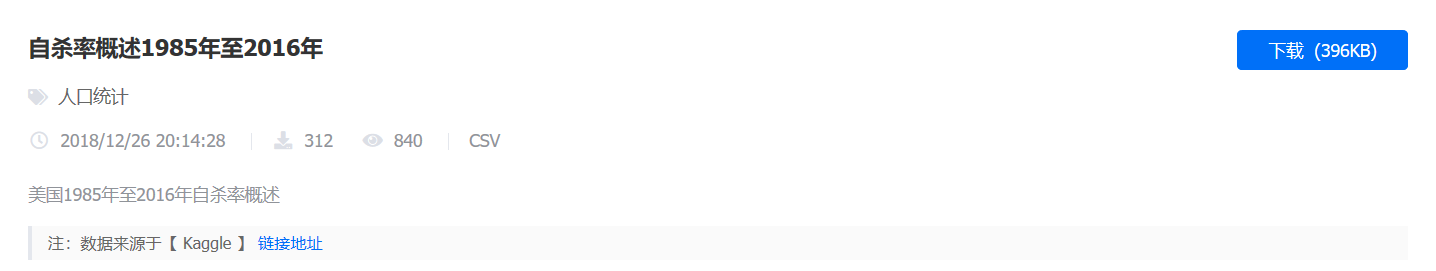
**关键字----数据挖掘；数据预处理；挖掘算法；可视化**

I．工作概述

## A．获取数据集

点击链接

<https://www.datafountain.cn/datasets/36>进入DataFountain界面，即可看到所需数据集“自杀率概述1985年至2016年”。下载完成后可以看到数据集中包含如下信息：城市、年份、性别、年龄、自杀率、人均GDP等信息。



## B．数据预处理

将原始数据集的格式规范化，去掉GDP for year（年GDP）的数据双引号。

对所需数据集进行相应的预处理。在原始数据集中筛选出实验所需的相关数据信息，并对其进行后续的相关操作。

*C．可视化*

使用Echart进行可视化。确定需要绘制的图，根据图所需要的数据格式，用Python生成相应的格式，再写.html文件进行可视化。

II．前人工作

基于地理信息系统（GIS）的人口数据挖掘，是建立在人口地理信息系统的基础上的，它是为了满足人口地理信息系统发展过程中的实际需要而产生的，所以其在国内外的发展研究与人口地理信息系统的发展密不可分。

*A．国外研究概况*

人口地理信息系统在国外的应用可以追溯到1970年，其发展是与人口普查密切相关的，美国在1970年的人口普查后，采用了自动化技术，发展了一套名为地理基础文件/独立坐标地图编码(GBF／DIME)的系统，可以说是人口地理信息系统的雏形。1990在GBF/DlME的基础上研制了功能更强大的TIGER人口地理信息系统，该系统提供1980-1990年全美的街区数字化地图及相应的人口普佥信息，以及人都市市区内地址、街道的定位，以及自动化地图绘制。整个TIGER系统文件不仪囊括了全国的街道、铁路、公路、水域，全国345个最大城市的街道地址，还包括统计地理区域边界的数据，基于此建立了地理空问数据与统计数掘的关联，继而进行空间分析。现在TIGER的数据已按邮政区域、县、市等级做成各种类型的普有数据光盘并进入应用市场，这些不同级别的人口数据在市场分析、投资决策、医疗保险、市场管理以及商业网点的选址中有着广泛的用途。

英国在1971年时就已经着于基于格网的统计，1991年提供普查区统计资料，即GAPS系统(GeographyArea Planning System)，该系统先后开发研制了HSMO、SASPAC等人口GIS，以满足人口制图和分析的需要。

日本在1973年制定了地域网络标准，以其为展现小弟去普查信息的基本空间单位，开发了普查绘图系统。从1990年开始进行普查地图的相关数字化工作。

*B．国内研究概况*

我国人口地理信息系统始建于1982年。在第三次人口普查之后，利用ARC/INOF软件输入了我国29个省份的分省地图以及相关的普查数据。1990年第四次人口普查之后，利用联合国开发的POPMAP软件和自动数字化仪，输入相关的普查信息，图形化的界面，可以直观的看到我国不同省份的出生率、死亡率、人口分布以及相关经济发展水平。

随着我国经济与科技水平的发展，目前也建立了自己的人口地理信息系统并在多地投入使用，集人口与地理要素于一体，具有很强的参考价值。

iii. 所使用的方法介绍

机器学习PAI：利用阿里提供的实验平台进行数据的预处理，数据归一化，拆分和相关的预测。其中的SQL语句是进行数据预处理的关键，利用SQL语句对自己所需信息进行过滤选择。

Python：对数据集进行处理，选取想要的属性，对数据进行统计，根据作图的要求，将数据排版成Echart可直接使用的格式。

Echart：可视化工具，借助Echart写.html文件来对数据进行动态可视化。

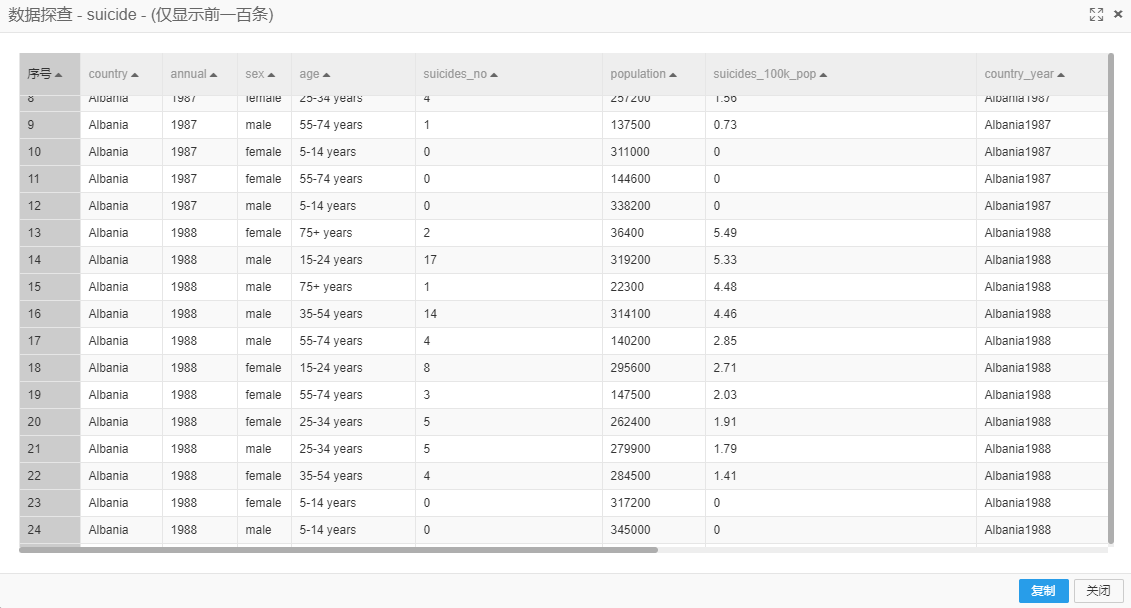
iv. 实验

1. *基于阿里实验平台的数据挖掘与分析*

利用阿里的实验平台建立相关表项及对应的描述信息。



对下载的原始数据集进行处理，使其格式符合导入规律，导入27820条数据。如下图所示：



1. *利用机器学习PAI进行相关操作，插入SQL脚本。*

*select \**

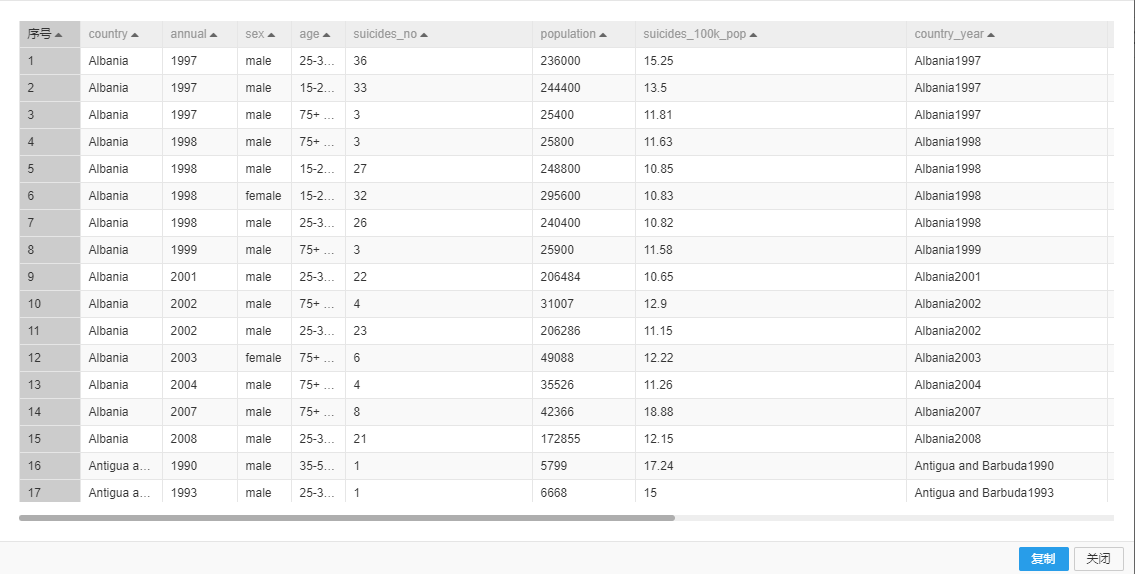
*from ${t1}*

*where suicides\_100k\_pop > 10*

*;*

*删选自杀率大于10%的相关数据项，得出如下数据：*

*分析数据可知，在自杀率大于10的区间内，没有5-14年龄段人群，75+年龄段人群所占比例最大。*



*2) 利用机器学习PAI进行相关操作，插入SQL脚本。*

*SELECT country*

*,annual*

*,sex*

*,age*

*,suicides\_no*

*,population*

*,country\_year*

*,hdi\_for\_year*

*,gdp\_for\_year*

*,gdp\_per\_capita*

*,generation*

*,(CASE WHEN suicides\_100k\_pop > 10 THEN 1 ELSE 0 END) AS suicides\_100k\_pop*

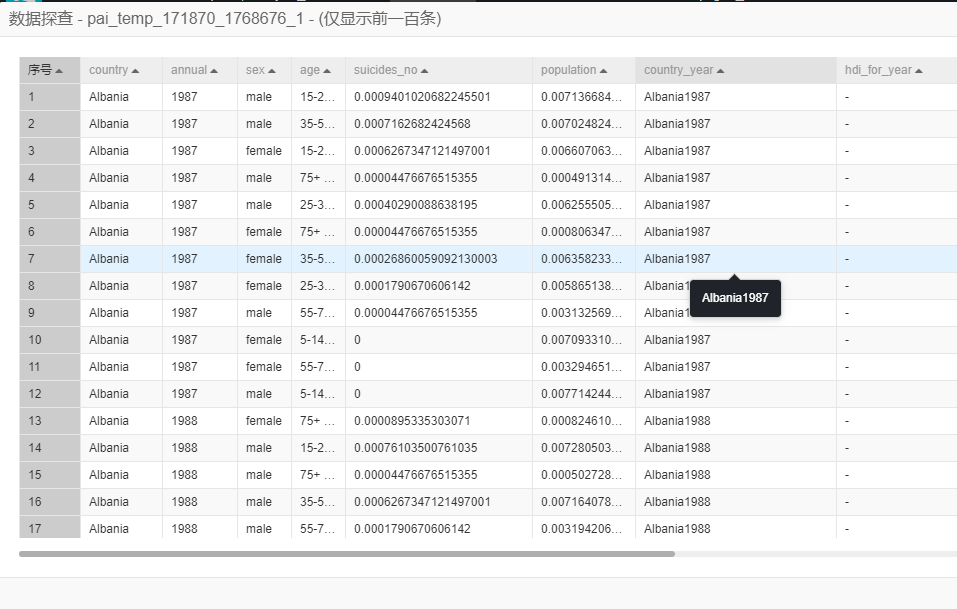
*FROM ${t1}*

*;*

*处理后suicides\_100k\_pop字段信息变为0或1。*

**

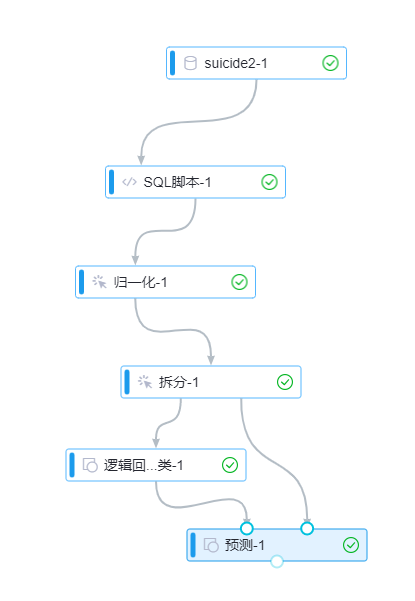
*3) 将数据进行归一化，执行完成后，可以看到数据已经完成去量纲。*

**

*4) 进行数据拆分，按比例切分，切分比例为0.8。*

*将逻辑二分类与拆分连接起来，目标列设置为suicides\_100k\_pop。*

*加入预测组件，进行相关设置。*



*5) 预测结果如下：*



1. *基于Echarts的数据挖掘与分析*

**5 参考文献**

[1]周怡.基于GIS的实有人口数据挖掘与可视化分析研究[D].上海:华东师范大学,2010. DOI:10.7666/d.y1741908

.