谷歌“探索”功能方案v0.0

1. 图表推荐接口 2

2. 类型识别规则 8

2.1. ID型 8

2.2. 连续型 8

2.3. 类别型 8

2.4. 日期型 8

2.5. 时间型 8

2.6. 单词型 8

2.7. 文本型 8

3. 单列分析 9

3.1. ID型 9

3.2. 连续型 9

3.3. 类别型 12

3.4. 日期型 12

3.5. 时间型 12

3.6. 单词型 12

3.7. 文本型 12

4. 双列组合分析 12

4.1. 连续型-连续型 12

# 图表推荐接口

**功能描述:**

图表推荐接口

**接口地址:**

https://192.168.10.147:8600/chart\_recommend

**请求方式:**

GET/POST

**请求参数:**

|  |  |  |  |
| --- | --- | --- | --- |
| 参数名称 | 类型 | 是否必需 | 描述 |
| taskid | string | 是 | 推荐id，在回调的时候原值返回 |
| ip | string | 是 | 数据库ip |
| port | int | 是 | 数据库端口 |
| user | string | 是 | 数据库用户名 |
| pwd | string | 是 | 数据库密码 |
| db | string | 是 | 数据库库名 |
| schema | string | 是 | 数据库schema |
| table | string | 是 | 数据库表名 |
| db\_type | string | 是 | 数据库类型:mysql, postgresql |
| showFrequencyHistogramStat | int | 否 | 连续型处理方式默认1：bit0：输出频率图 bit1：输出频率图cdf bit2：输出kde图 bit3：输出kde对应cdf |
| singleResponse | Boolean | 否 | 是否单个输出 |
| singleResponseUrl | string | 否 | 输出url |

实例：

{

tasked : “id\_123456” ,

ip : “192.168.10.147” ,

port : 5432 ,

user : “user” ,

pwd : “pwd” ,

db : “machinelearining” ,

schema : “schema” ,

table : “iris” ,

db\_type : “postgresql” ，

showFrequencyHistogramStat : 4

}

返回参数:

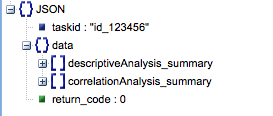
|  |  |  |
| --- | --- | --- |
| 参数名称 | 类型 | 描述 |
| return\_code | int | 0:成功; 不为0:失败 |
| return\_msg | string | return\_code为非0时的具体错误信息 |
| data | json | 返回数据,具体见下图，当增加算法分析结果后放在此处 |
| taskid | string | 推荐id，在回调的时候原值返回 |

将echarts对应图形setting直接放入变量即可。

需遍历descriptiveAnalysis\_summary元素，添加单列分析结果。

需遍历correlationAnalysis\_summary元素，添加回归分析结果。

举例如下：



data部分：

|  |  |  |
| --- | --- | --- |
| 参数名称 | 类型 | 描述 |
| descriptiveAnalysis\_summary |  | 单列分析部分 |
| correlationAnalysis\_summary |  | 回归分析部分 |

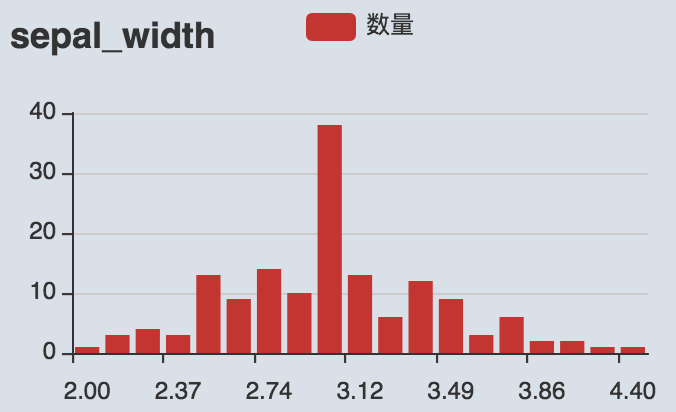
descriptiveAnalysis\_summary部分：

|  |  |  |
| --- | --- | --- |
| 参数名称 | 类型 | 描述 |
| frequencyHistogramAnalysisEchartsVo | Json | Echarts对应图形setting，请求参数中showFrequencyHistogramStat bit1 = 1时此项存在。 |
| KdeAnalysisEchartsVo | Json | Echarts对应图形setting，请求参数中showFrequencyHistogramStat bit3 = 1时此项存在。 |
| msg | String | 描述信息 |
| tbCol | Json | 列信息 |

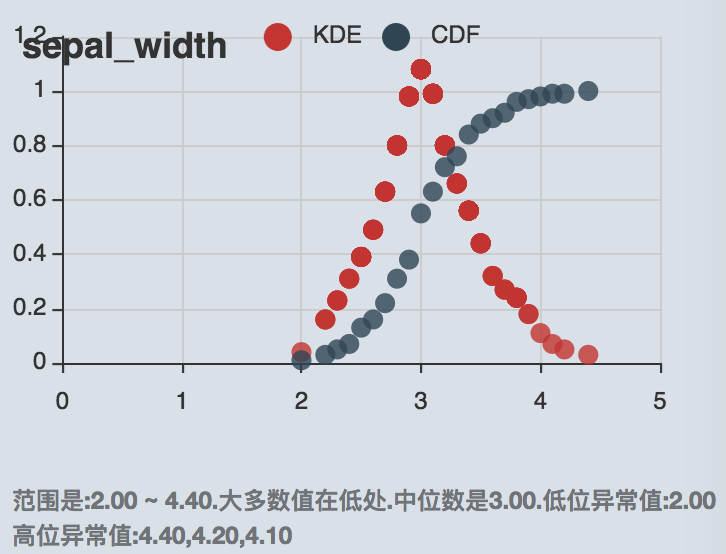
举例如下图：



频率直方图:



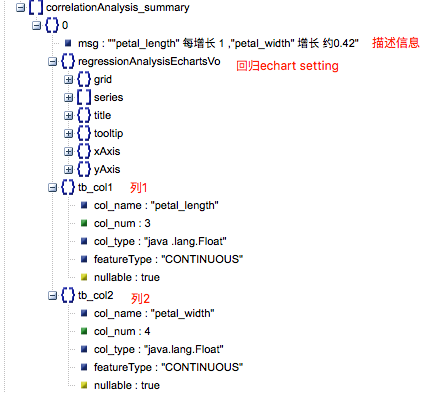
KDE CDF图：

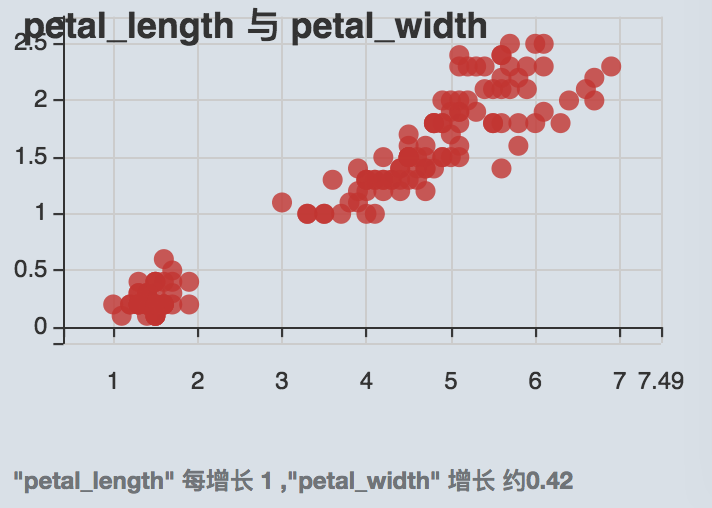


descriptiveAnalysis\_summary部分：

|  |  |  |
| --- | --- | --- |
| 参数名称 | 类型 | 描述 |
| regressionAnalysisEchartsVo | Json | Echarts对应图形setting。 |
| msg | String | 描述信息 |
| tbCol1 | Json | 列1信息 |
| tbCol1 | Json | 列2信息 |

举例如下图





当singleResponse为true时

举例返回

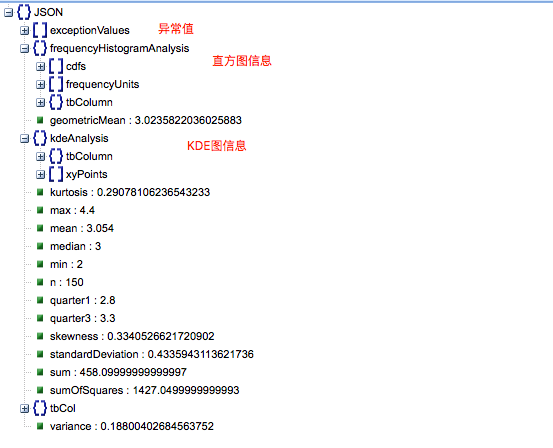
{

"taskid": "id\_123456",

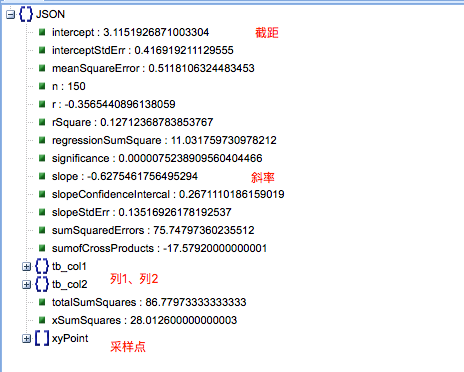
"return\_code": 0

}

单列分析返回值：



回归分析返回值：



数据分为两种类型：描述型、记录型。此方案只针对描述型。

对所有数据按照特征类型分为七种，不同特征类型数据分析方式不同：

1. ID型。
2. 连续型。如 身高、体重
3. 类别型。如性别、所在省份
4. 日期型。
5. 时间型。
6. 单词型。
7. 文本型。

# 类型识别规则

## ID型

数据库类型为int、long。处于第一列。无重复项。列名中含有id。

## 连续型

数据库类型为int、long、double、float。不同值数量超过10。不出现非数字该有的字符。

## 类别型

数据库类型为int、long、string。不同值数量低于10。

## 日期型

数据库类型 int、long、string、datetime、timestamp、date采用正则表达式匹配第一个有数据项。之后遍历此列所有数据、如全部符合为日期型，否则为文本型。

## 时间型

待定。

## 单词型

待定。

## 文本型

非其它类型即为文本型。

# 单列分析

## ID型

待定。

## 连续型

输出基本描述信息，如：

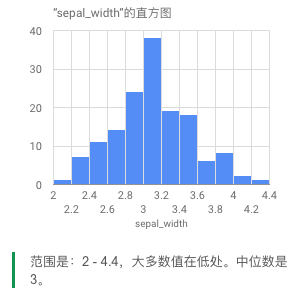


输出频率分布直方图，分组间隔取，如 5 < < 100 取interval ，否则取 5或100。（N:数据条数）直方图输出条件为，直方数量>5.

输出KDE图，输出条件：点数量超过10个。

输出CDF图，分为基于直方图和KDE图。

如果偏度偏离高斯分布正常范围，输出数据处于低处还是高处、中位数。（正常范围 （ ）），据此输出是稍微处于低处还是基本处于低处。如下



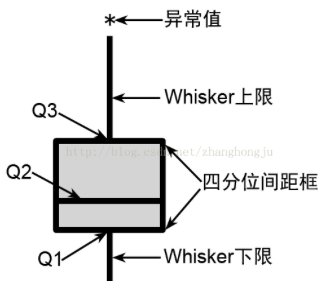
如果峰度偏离高斯分布正常范围，判断数据是否集中。当10%的横坐标范围能够表示90%以上数据时给出提示。（正常范围 （ ）），给出x轴10%范围的覆盖度。

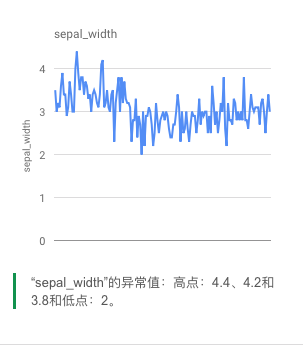
输出异常数据，异常数据判断标准参见箱形图。

第一四分位数（Q1），第二四分位数（Q2）， 第三四分位数（Q3），

第三四分位数与第一四分位数的差距又称四分位间距。

Whisker上限是延伸至距框顶部1.5倍框高范围内的最[大数据](http://lib.csdn.net/base/spark)点，Whisker下限是延伸至距框底部1.5倍框高范围内的最小数据点，超出Whisker上限或下限的数值将使用星号“\*”表示。

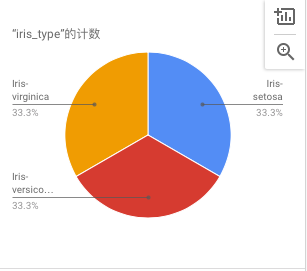




数据符合高斯分布情况下输出密度函数。

## 类别型

输出类别概率密度。如下



## 日期型

待定

## 时间型

待定

## 单词型

待定

## 文本型

待定

# 双列组合分析

## 连续型-连续型

回归分析，得到y = kx + p 。

验证拟合直线是否具有线性关系：计算，其中大于0 ~ 0.6为无线性关系，0.6~-0.8为基本有线性关系，0.8~1为很强线性关系。

做K-S检验，峰度范围（ ）偏度范围（ ）为符合高斯分布，符合高斯分布情况下输出拟合函数图。如下

