****

****

**QG中期考核详细报告书**

**题 目qg数据挖掘组中期考核详细文档**

**学   院       信息工程学院**

**专 业 电子信息**

**年级班别 20级9班**

**学 号 3120002482**

**学生姓名** 潘鑫成

**2021年 4 月15 日**

一、定义问题，数据解读

1、问题类型：分类问题

2、问题简述：您的客户是一家跨国金融公司，向消费者提供多种产品。尽管主要的贡献来自线下分销渠道，但是有多个渠道可以为消费者提供这些产品。离线渠道通过其代理商网络向消费者出售金融产品，并且根据政府规定，这些代理商必须获得销售金融产品的认证。针对不同类别的金融产品，有多种认证计划。

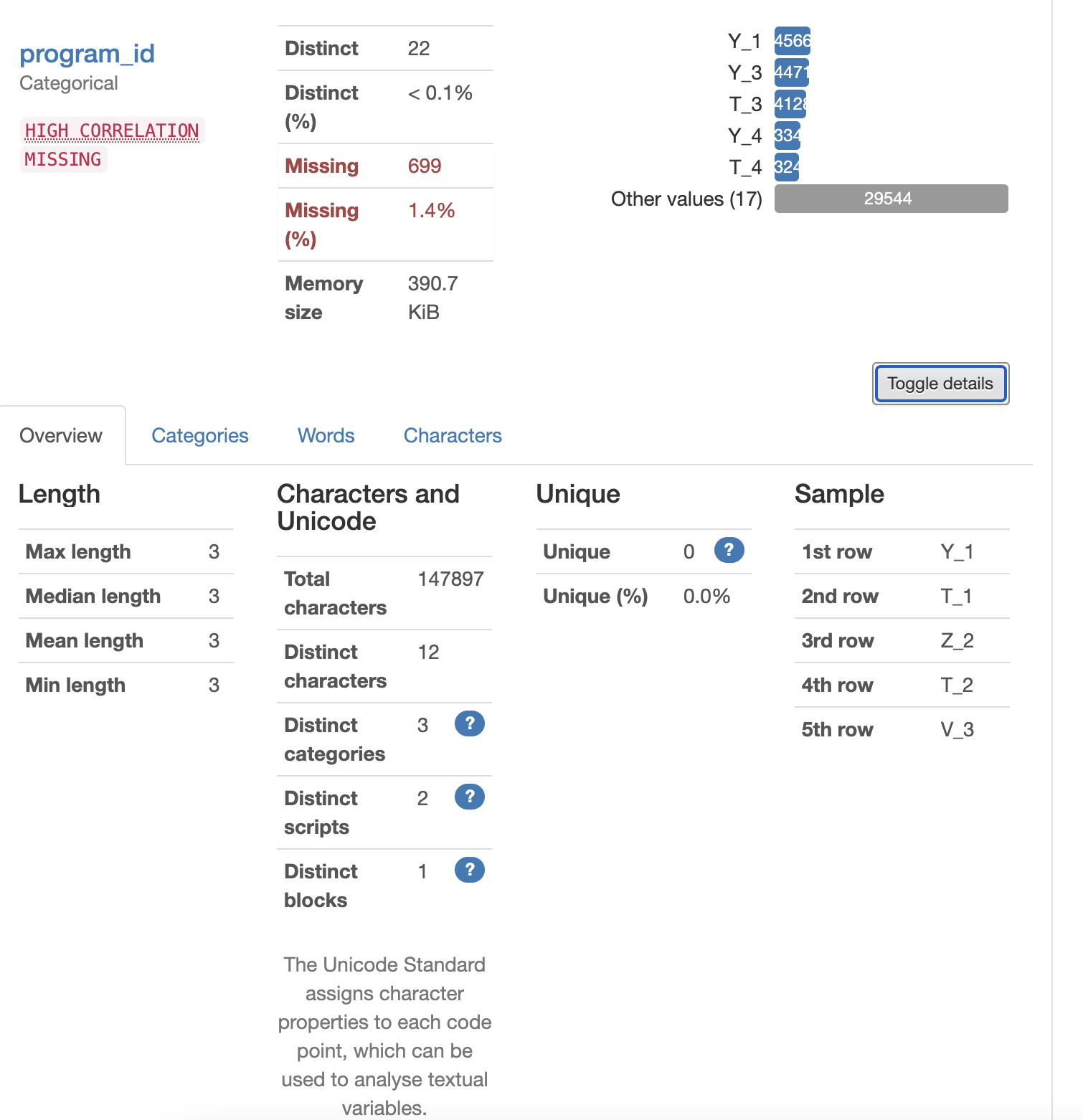
由于这个离线渠道对公司的总销售额贡献很大，因此公司着重于招聘并认证他们以建立大型代理商网络。在这里，主要的挑战是培训他们获得销售各种类型产品的认证。

在多个程序中，您将获得针对培训课程测试明智的学员绩效数据集。您的任务是根据人口统计信息和培训计划/测试详细信息来预测此类测试的性能。通过找出最重要的因素来提高受训者的参与度和表现，这将使您的客户加强其培训问题。

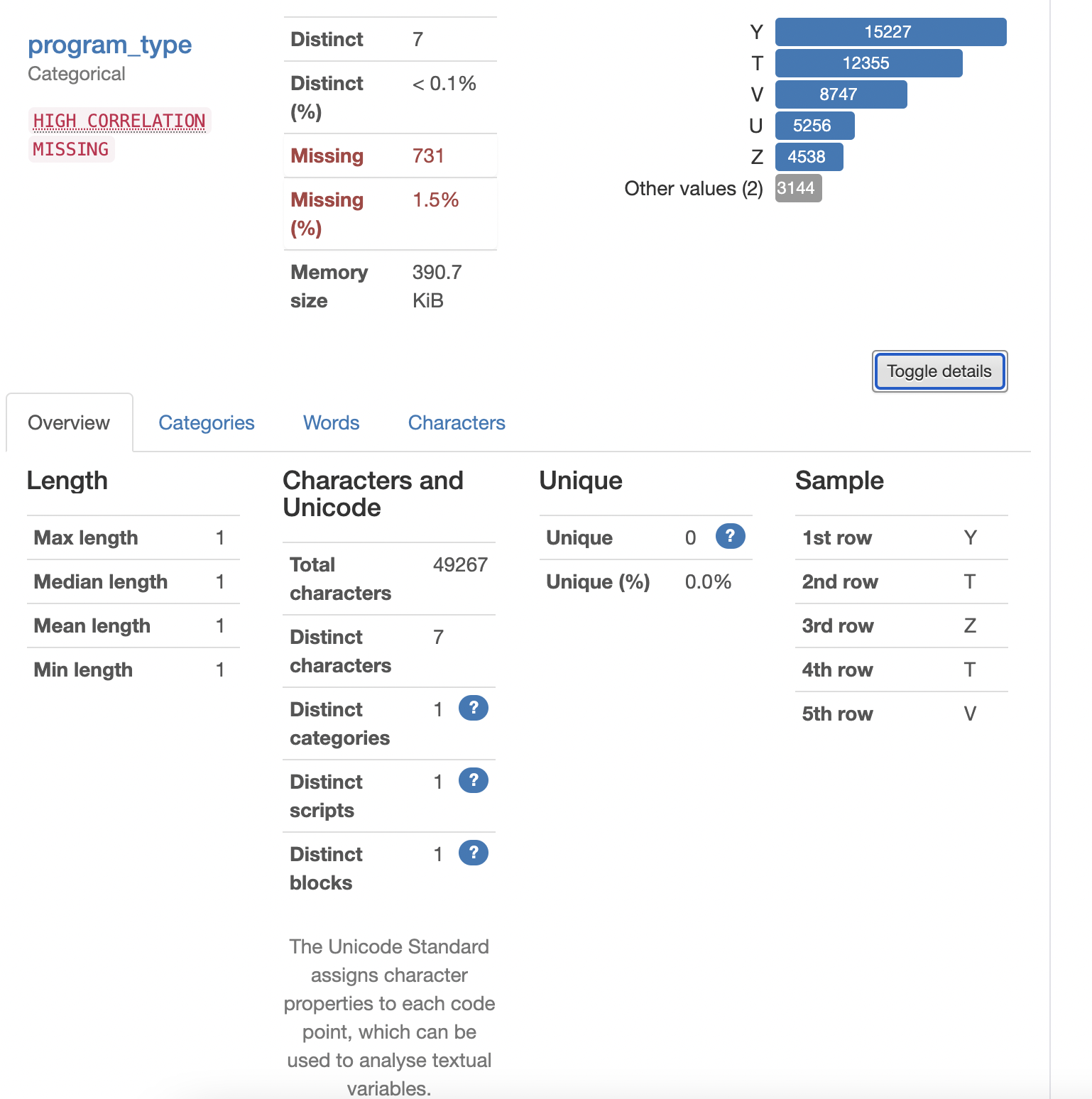
3、数据解读与预分析

id\_num-唯一ID

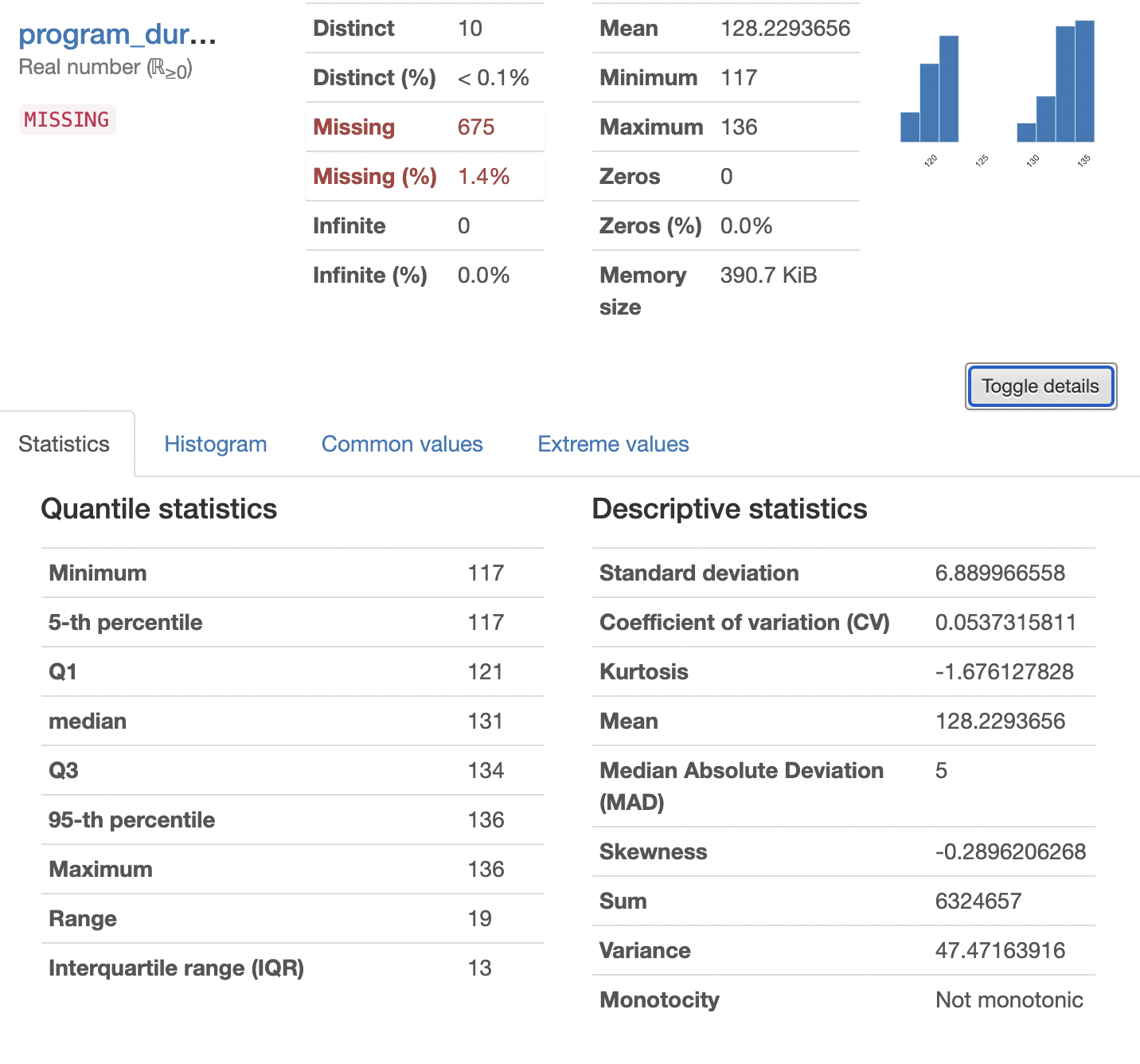
program\_id-程序的ID



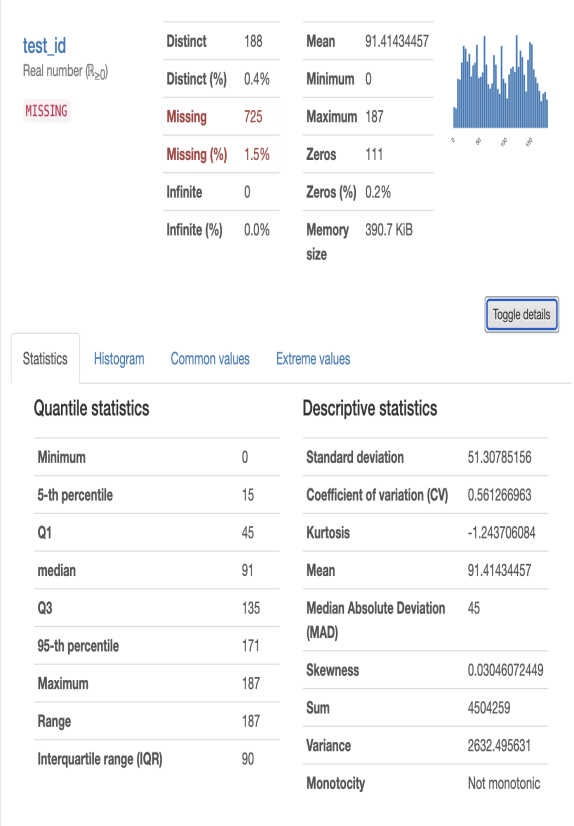
program\_type-程序类型



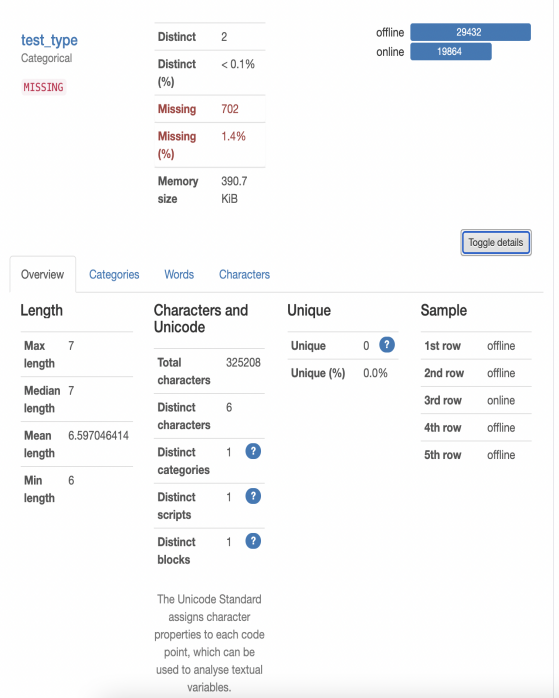
program\_duration-计划持续时间（天）



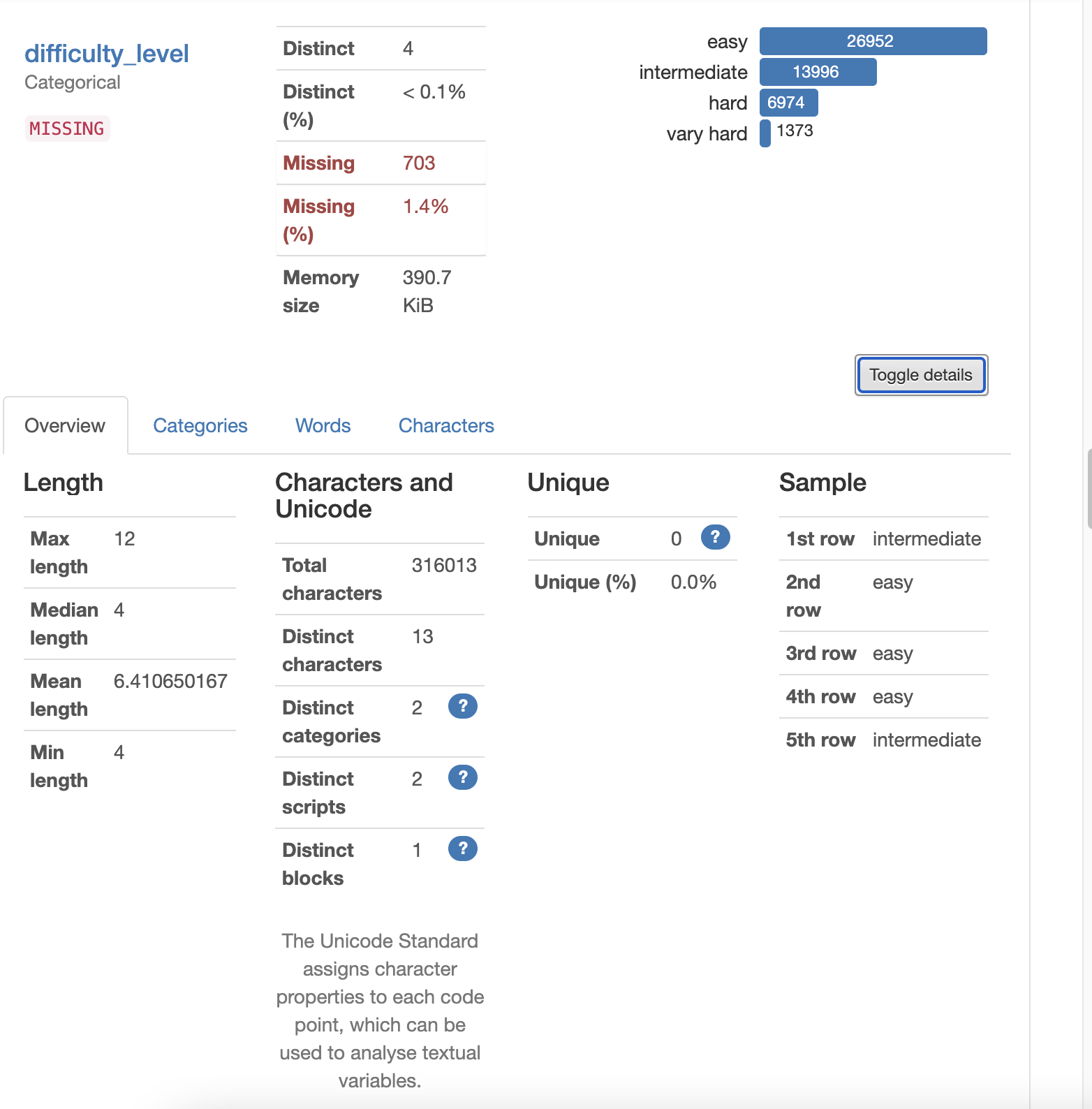
test\_id-测试ID



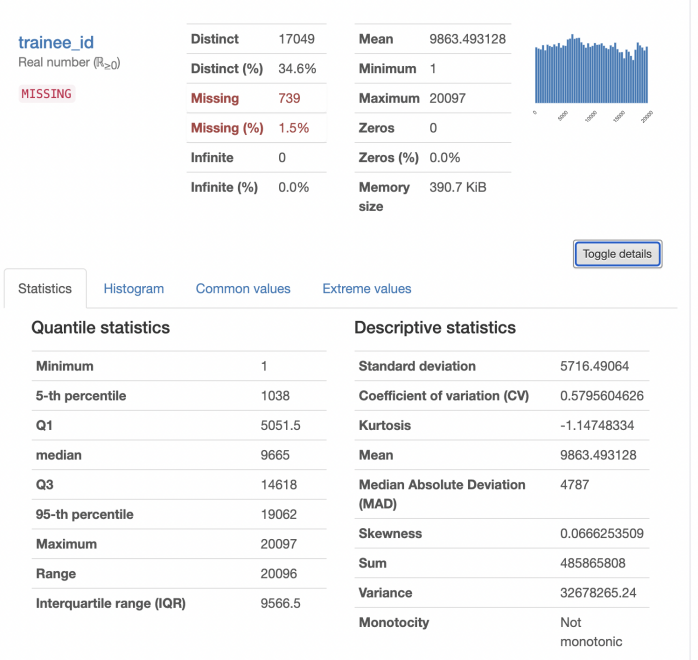
test\_type-测试类型（离线/在线）



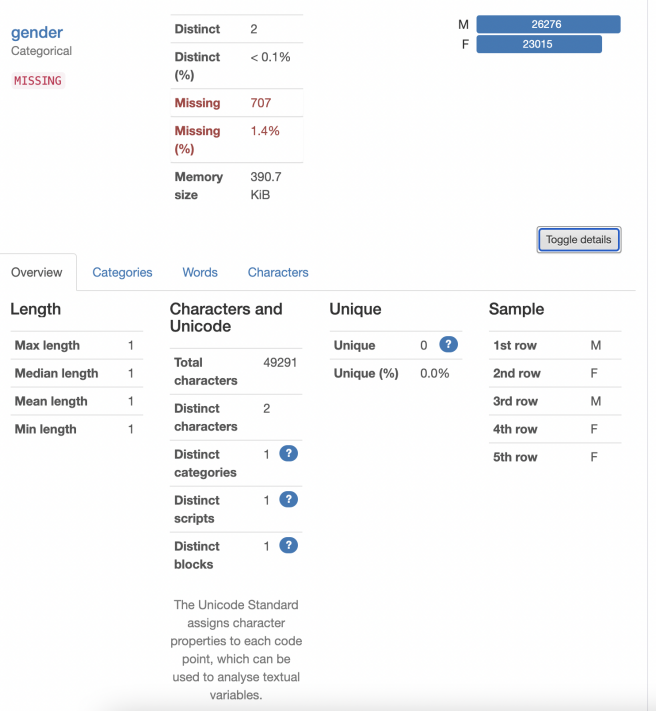
难度级别-测试难度级别



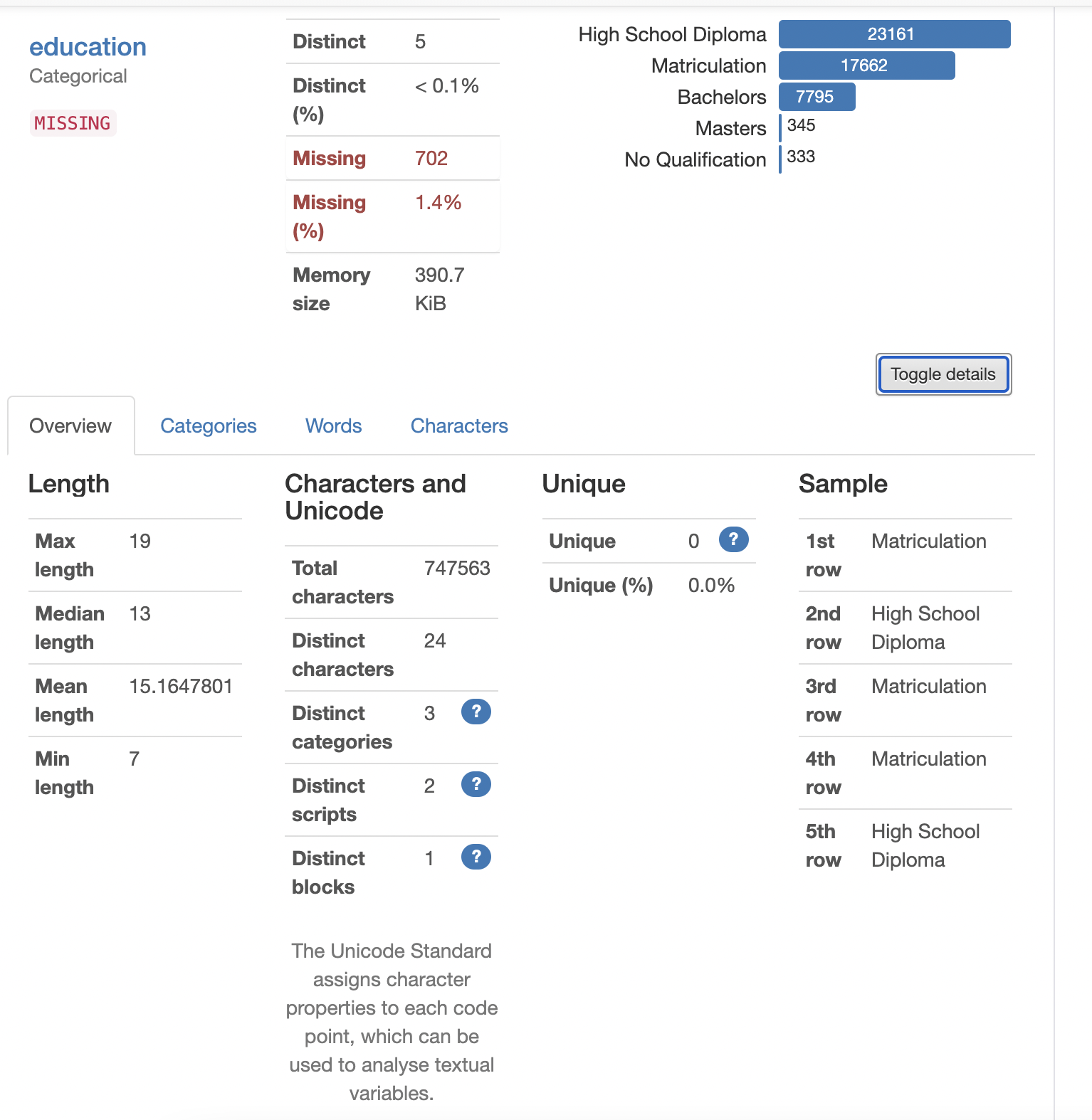
trainee\_id-学员的ID



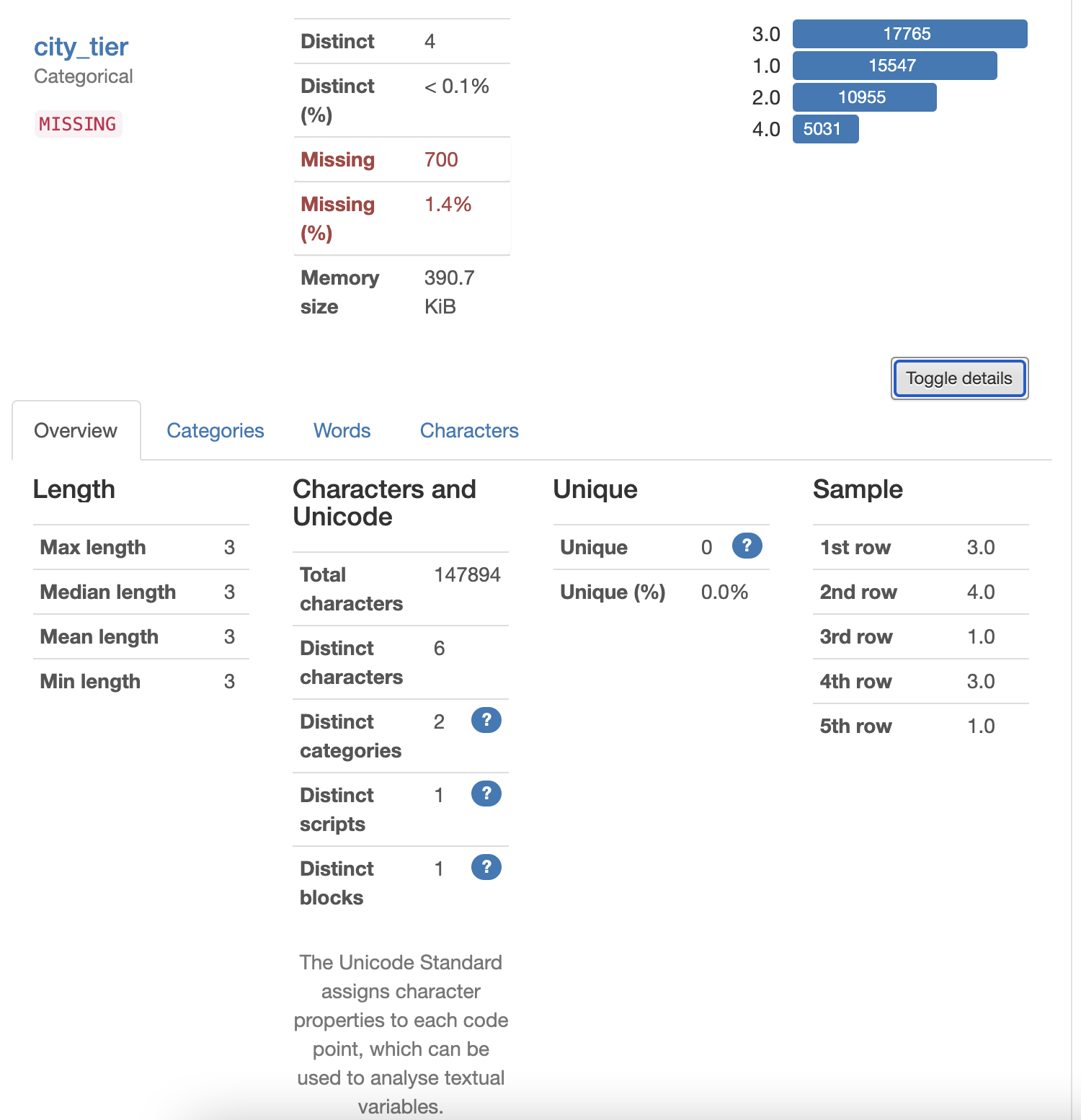
性别-受训者性别



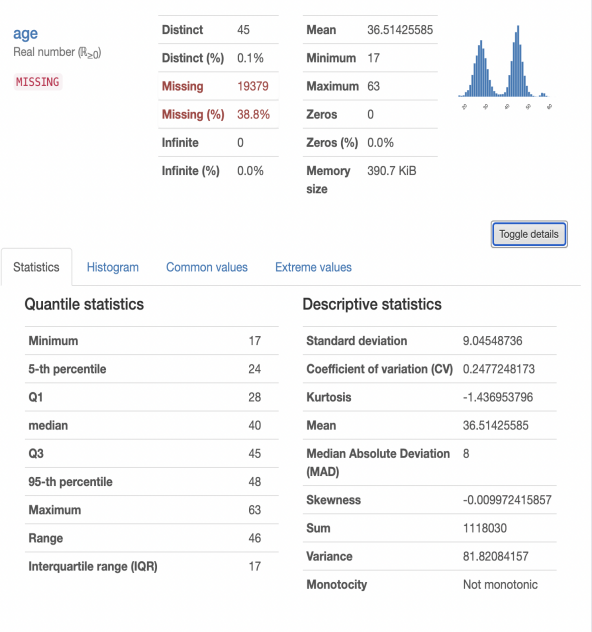
教育-学员的教育水平

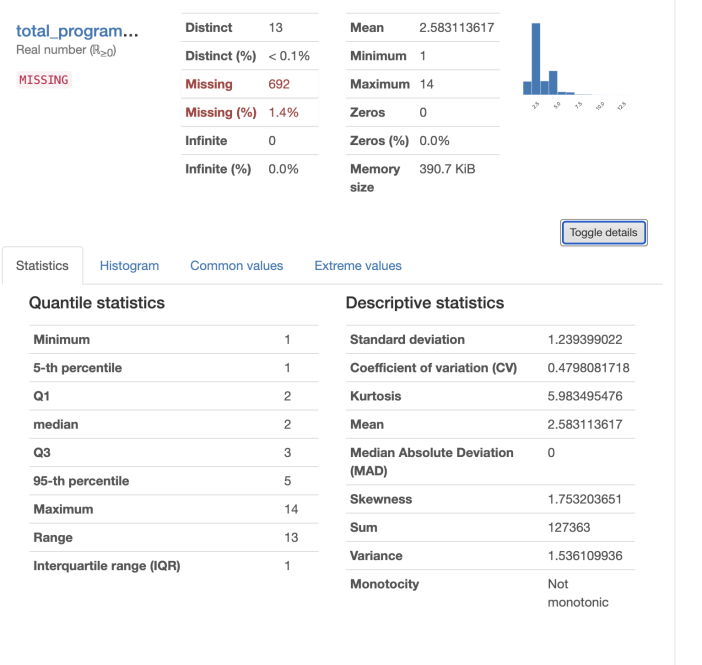


city\_tier-实习生居住城市的等级

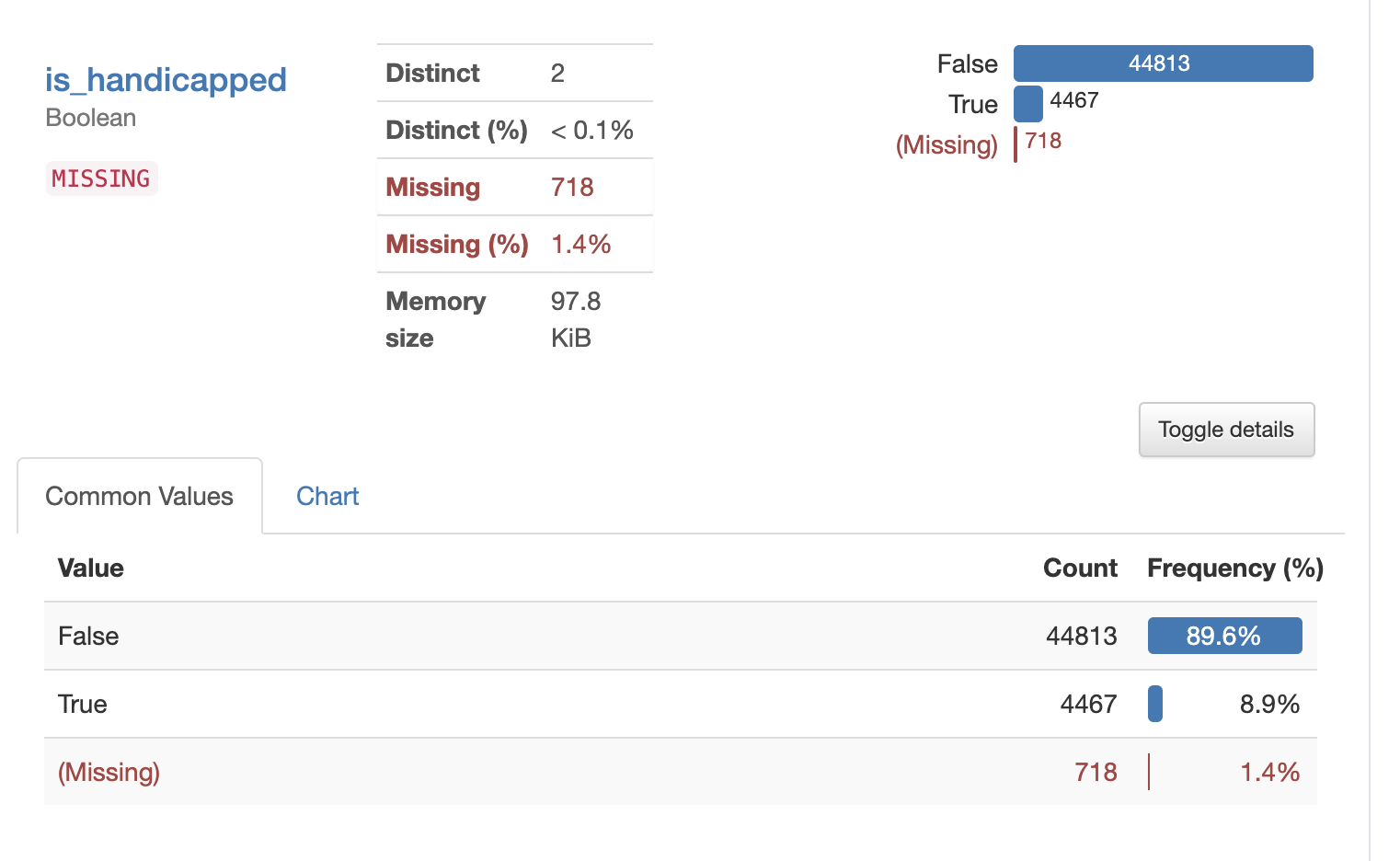


年龄-受训者年龄

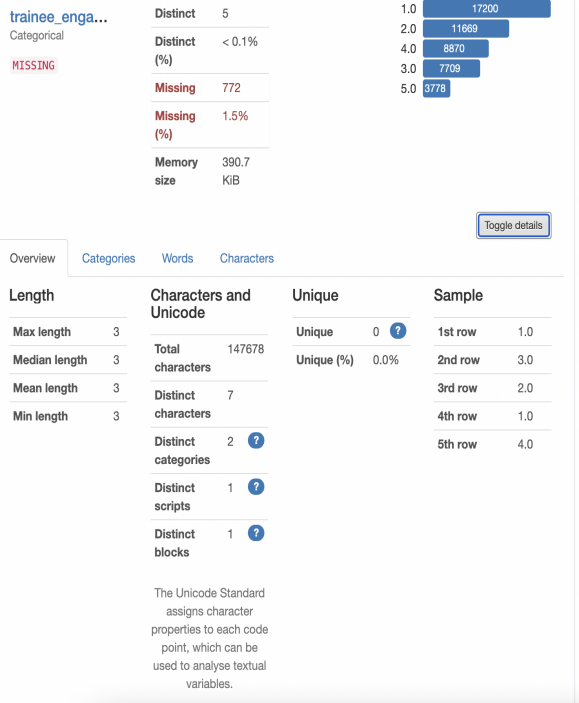


总计划招收-总课程的学生通过实习

is\_handicapped-受训者是否患有残疾



学员参与度-讲师/教学助理为课程提供学员参与度

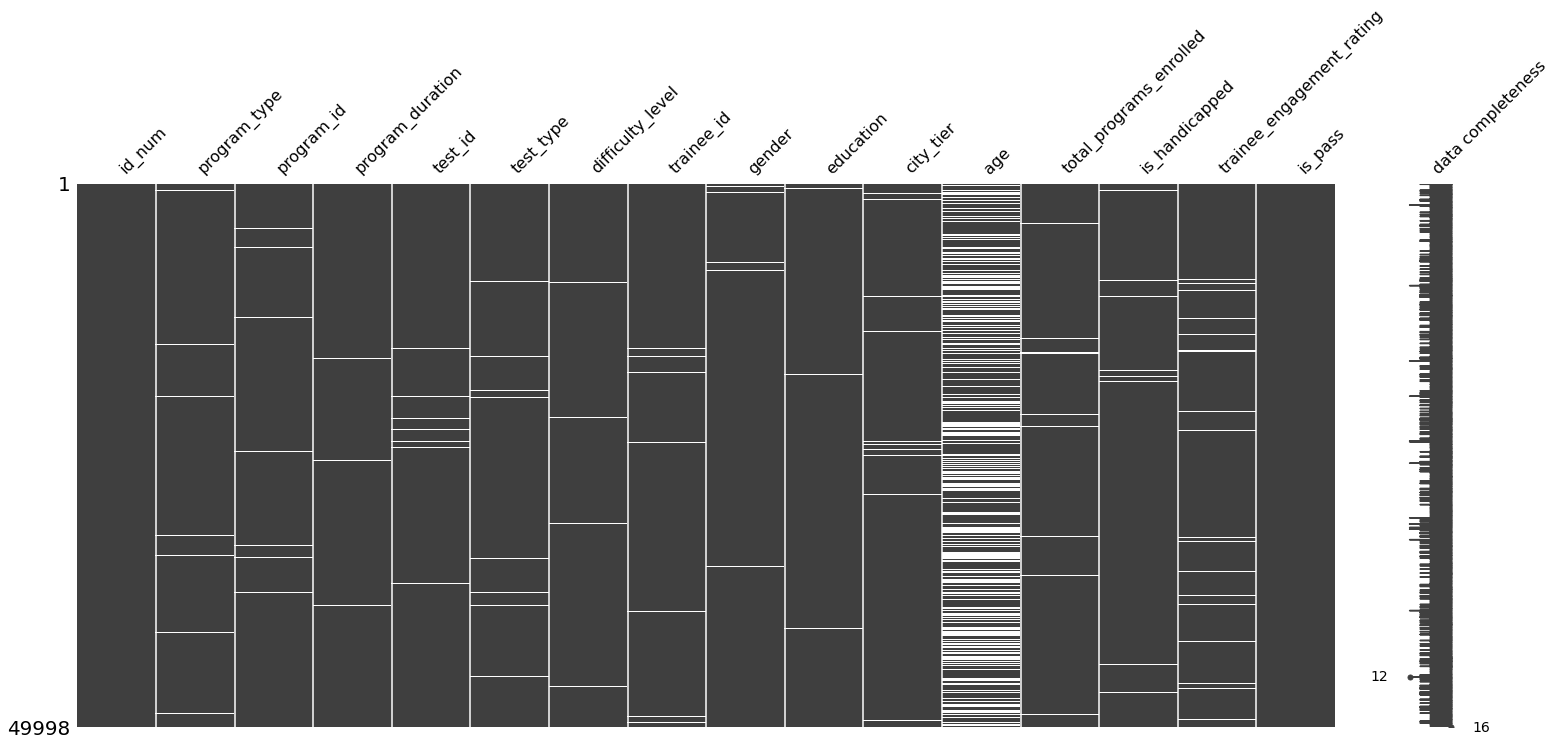


is\_pass 0-测试失败，1-测试通过

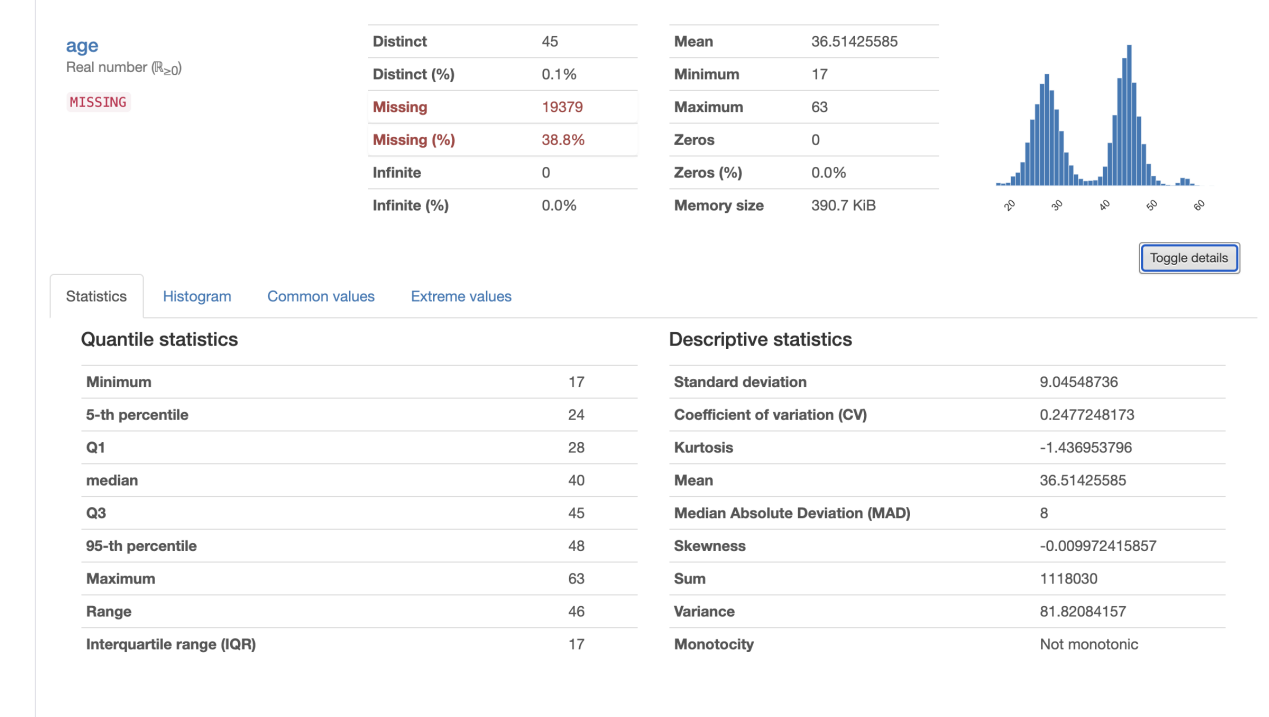
**二、数据分析**

通过pandas,pandas—profiling和missingno库，进行数据的分析。

首先，这个库的缺失值大致分布如下



其中最大的缺失是age，根据pandas—profling得到age的详细缺失值

****

**38.8%的缺失不好填补，所以考虑直接删除，但是在后面我使用了xgboost模型，其可以自动对缺失值进行预测，所以没有删除。**

**2. 编码和缺失值处理**

通过简单的分析发现age的缺失值较高，所以考虑删除，同时对一些特征进行离散化编码，将字符串类型编码成离散化数据，使其可以进入模型运算。对于某些特征，不能简单对进行标签编码，所以采用独热编码，使其可以合理的映射到欧式空间，方便使用一些分类算法。同时通过合理的独热编码可以提升数据集的维度，合理添加特征。但是，对于一些特征，分类太多，随意使用独热编码可能会造成维度灾难，使模型的训练时间过长，同时可能让模型过拟合，虽然可以使用降维算法来简化模型，但是会比较麻烦，对这些特征选择简单的标签编码就好了。

编码选择：

Test—type 标签编码

Gender 独热编码

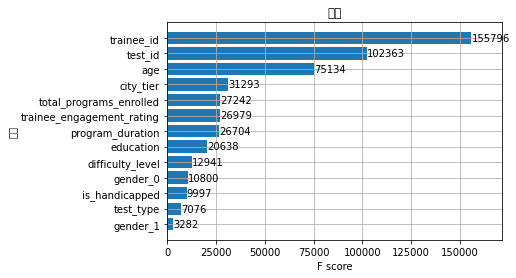
Difficulty\_level 标签编码

Is\_handicapped 标签编码

填补缺失值时就只是简单的取众数填补

1. **特征工程**

在xgboost中，有一个方法plot\_important可以将训练的特征重要性以图像的形势表现出来，通过第一次训练，得到了特征重要性图



通过此图我们可以看到，age与city\_tier重要性较高（且进行计算感觉比较有实际意义），所以我尝试构造了一个新变量为这两个值之和，在第二次训练过后，重要性图为

**四、模型的选择与训练**

由于个人能力的限制，没有考虑那些高大上的深度学习的方法，首先我想到的是用朴素贝叶斯试试，不出我所料，朴素贝叶斯由于假定特征相互独立从而使其准确率较低，之后我考虑了决策树，在决策树中，准确率只有0.6几，于是我决定用随机森林，结果大约为0.73到0.74中徘徊。最后选择了xgboost回归，因为它训练出来的模型在网站上的准确率较高。至于模型的评判指标，我选择了auc。因为在这次的数据集中，我观察到正反比分布不平均，而使用准确率评判模型会导致模型的评判不准确，于是我选择了auc来评判模型。

而由于我用的是xgboost回归，所以最后输出为概率，我设定了个0.5的阈值，当概率大于0.5时，判断其为1类。

至于模型训练，我将数据集分隔成了0.8的训练集与0.2的测试集，其效果为**结果**

1. **模型的优化**

由于我对模型的了解不深，所以只是简单的从网上找了一下各参数的意义，随便调了一下。