**儿童养育环境风险因素分析**

摘 要

// 我写

目录

[摘 要 1](#_Toc165805136)

[一、 问题重述 3](#_Toc165805137)

[二、 问题分析 3](#_Toc165805138)

[三、 模型的建立与求解 4](#_Toc165805139)

[3.1数据探索 4](#_Toc165805140)

[3.2数据预处理 5](#_Toc165805141)

[3.3模型探索 6](#_Toc165805142)

[3.3.1逻辑回归 6](#_Toc165805143)

[3.3.2随机森林 7](#_Toc165805144)

[3.4模型选择与参数调优 7](#_Toc165805145)

[3.5模型评估 8](#_Toc165805146)

[四、 问题解答 9](#_Toc165805147)

[五、 参考文献 12](#_Toc165805148)

# 问题背景

1. 题目背景及说明。

家庭是儿童生存的基本环境，尤其在儿童早期，家庭几乎是儿童发展的全部环境。为了对儿童家庭养育环境进行评价和分析，且鉴于0~6岁儿童发展十分迅速，各阶段发育内容变化很大，与之相对应的养育要求和家庭环境也有诸多不同，因此将0~6岁细分为3个年龄阶段分别编制家庭养育环境问卷进行调查。附件的数据即为某地区收集的5万余条儿童养育招呼问卷数据，希望通过对数据的分析，能提炼儿童养育风险因素的主要条目，加强儿保人员的专业培训，提高养育指导能力；针对问题儿童进行跟踪随访，主要随访体格发育和心理行为发育问题；加强养育人的健康教育指导；制定儿童养育环境问卷在基层推广应用；为政府提供养育调查报告，同时针对养育主要问题提出建议供决策参考。

1. 数据集构成

“家庭养育照护调查问卷数据”中，分为4个部分：第1部分为儿童基本情况的问卷题目，第2~4部分分别为0~1岁、1~3岁、3~6岁的问卷题目。

|  |  |
| --- | --- |
| 基本情况的问卷题目 | 1~13 |
| 0~1岁问卷题目 | 14~45 |
| 1~3岁问卷题目 | 46~86 |
| 3~6岁问卷题目 | 87~139 |

1. 评分标准

各个问题计分规则：从不-1分、很少-2分、有时-3分、经常-4分、总是-5分。正面问题做反向计分。总分是各个问题得分的代数和。

总分<P15的可认为养育环境较差，得分>P85可认为养育环境良好，中间的认为养育环境一般。

# 分析思路概述

1. **如何定义“主要养育风险因素”？**

问卷给出了许多问题（特征）和一个总分（标签值），每条问题可以看成一个“风险因素”，一个很自然的思路是，去找到哪些问题对总分的影响是占主导的，这些问题是“主要风险因素”，具体做法是以问题得分为特征，以总分为目标，做一个线性回归去拟合，找出权重大的特征。

但很遗憾不能这样做，这个数据集的问题在于，总分那一项实际上是冗余的，它由各个问题的得分直接代数和计算而来，也就是说，总分和数据集的特征不是互相独立的，并且每个特征同等重要（总分是无权重的和），这样做回归是没有意义的。同理，题目规定上下15分位点为分界线，划分养育环境好坏为三个等级，这样做的分类和数据集的特征仍然是相关的，也无法通过分类算法来分析。

因此，我们还是通过直接比较样本均值来判断吧。我对“主要风险因素”的理解是，样本总分的均值较低的就是主要风险因素。如何衡量低？计算15分位数，若小于，说明这个值较低。

1. **数据集该如何使用？**

从先前的介绍可以看出，数据集分为了四个部分。“基本情况的问卷题目”可以看成对总体做了许多不同角度的划分，例如：妈妈年龄，爸爸学历，孩子性别等

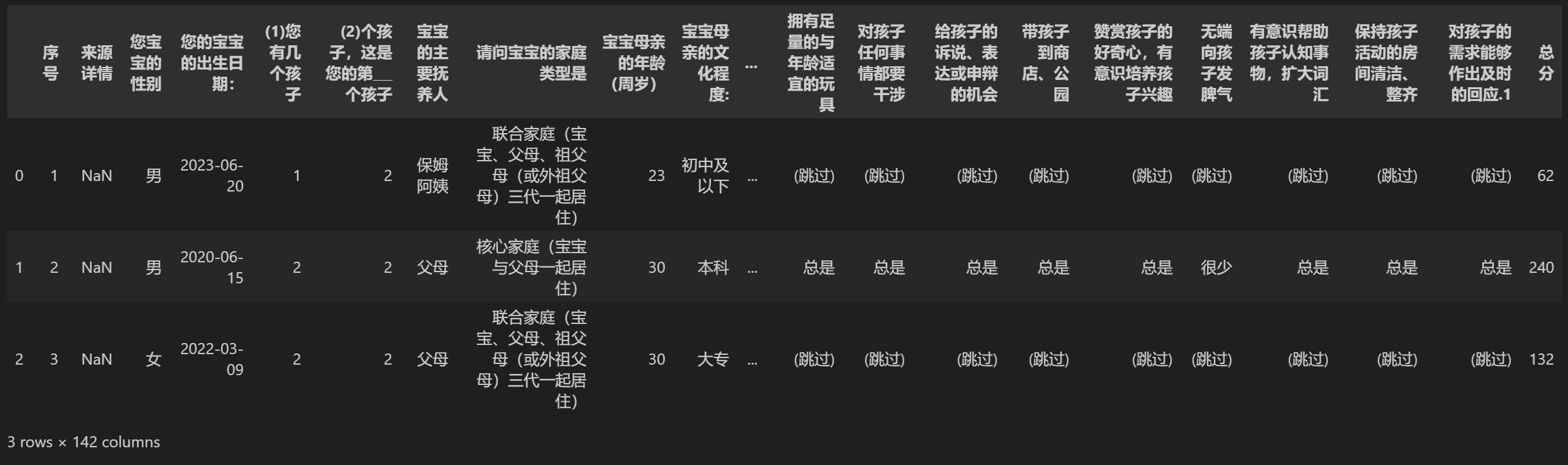
剩下三个部分，是对不同年龄段孩子的题目，可以看成是特征，这一部分也构成了最后的总分。

因此，我们分析数据集，依据第一部分对样本进行划分，将样本分成若杠组找到主要风险因素。具体方法可以采用比较均值，对两组样本做t检验等方法实现。

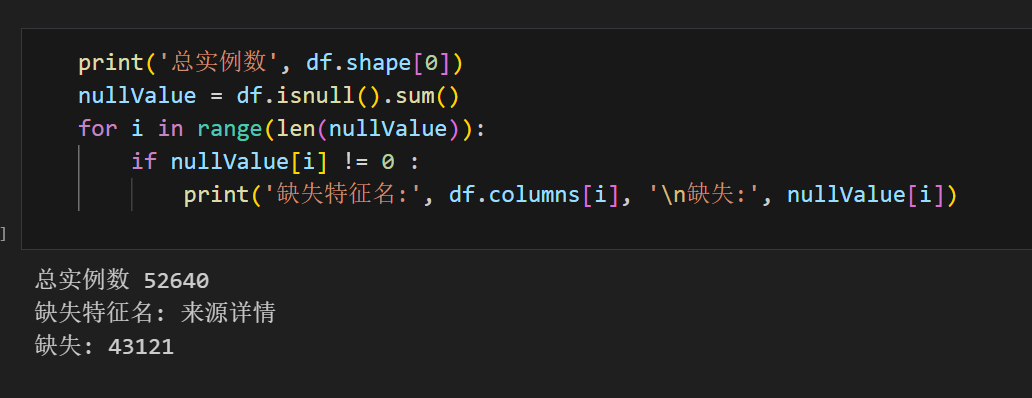
# 分析过程

1. **数据预处理**

初步查看数据前3行。可以看到“来源详情”这一列有缺失值。

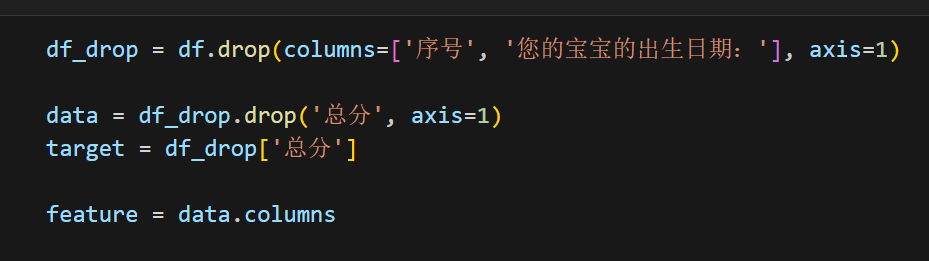


接着我们检查下其他列还有没有缺失值。可以看出只有这一列。



对数据进行缺失值填充。因为来源详情是一个离散变量，只有{NAN，“直接调查”}两种值，因此“NAN”的用0填充。

为了简化，将**序号**和**出生日期**删去，因为后面已经有年龄段数据了，出生日期不会使用。



接下来要检查每一列的值是否合法，这是预处理中工作量最大的。将每一列的值的种类输出到文件中，以value\*.txt命名。可以看出，主要有两个问题：第一，与数字有关的答案格式非常混乱，有数字字符串，汉字，甚至Unicode字符。第二，部分固定答案个数非常多，例如“其他xxx”，如果直接编码会造成很大的空间浪费。

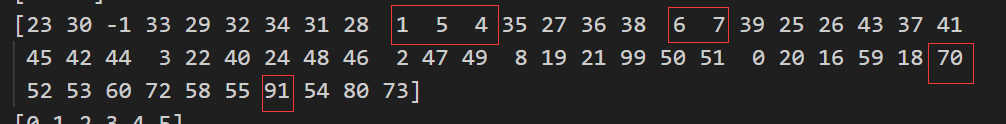


解决方案：对第一个问题，数字字符串可以直接通过转换得到int类型数据，对于其他输入，我们人工挑选出一些出现次数比较多的答案，例如“一”、“二”。第二个问题，我们将答案为“其他\*”的数据编码为同一个值。最后，将无法转换的值编码为int类型 -1，并在后续中删去包含-1的行。

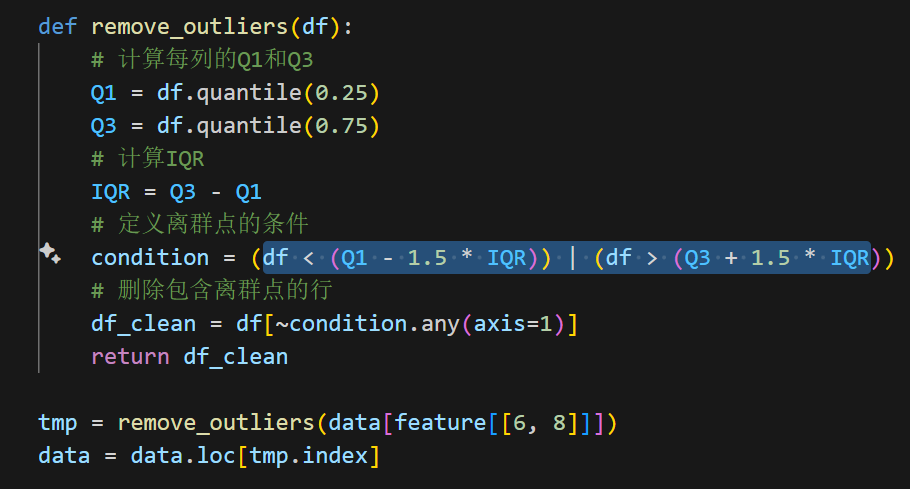
具体实现很枯燥，请查看源代码及注释。转换后的数据概览如下。



接着我们还发现有一些值存在明显错误，例如“母亲年龄”这一列，居然有1, 5, 7, 70, 91 这样的数据，这显然是不合乎认知的。

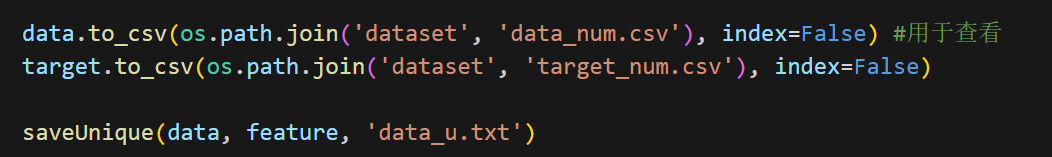


解决方法是采用离群点检测，删除在Q1 - 1.5 \* IQR 和 Q3 + 1.5 \* IQR之外的数据。

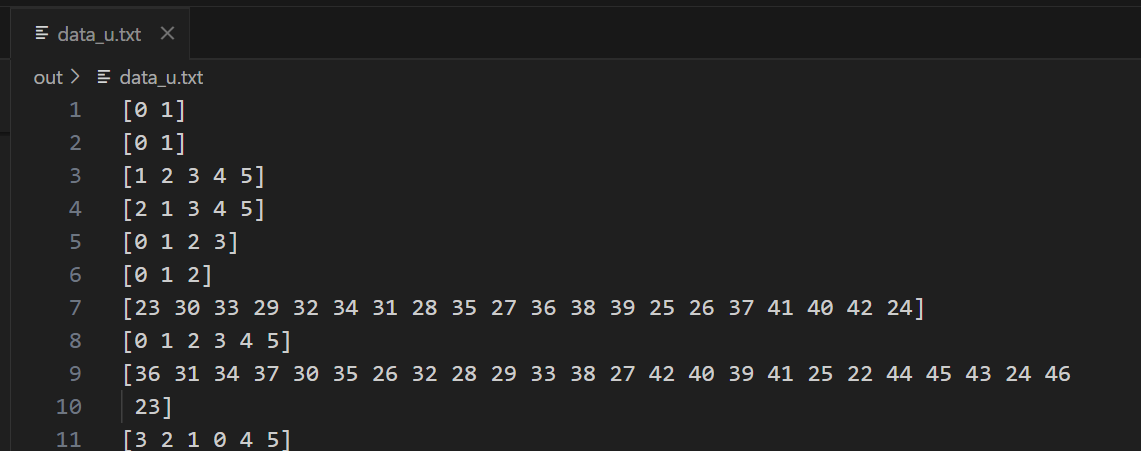


基础问题部分的处理到此结束，后面3个部分的问题都只有固定的5种答案（实际还有一个“空”答案，影响不大，编码的时候计0就行）。将5种答案转换成对应的分数即可。详细参考源代码及注释。

最后将处理好的数据保存到文件，此处再输出每一列的值的种类，检查是否无误。



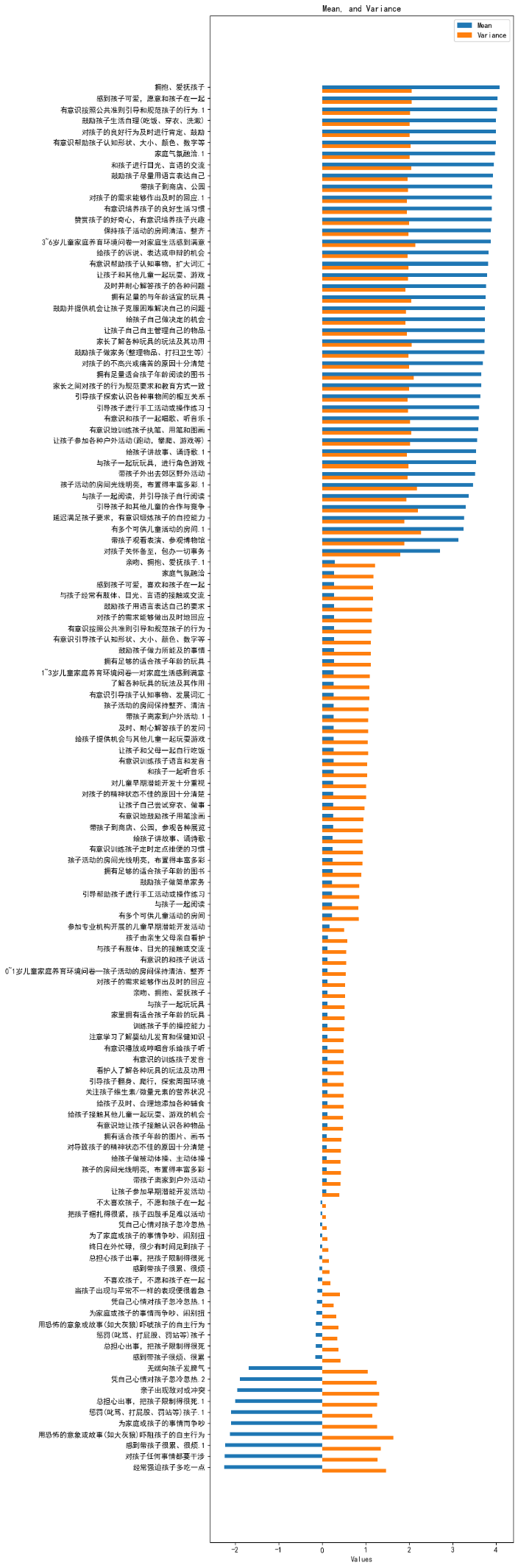
检查没有发现非法值。



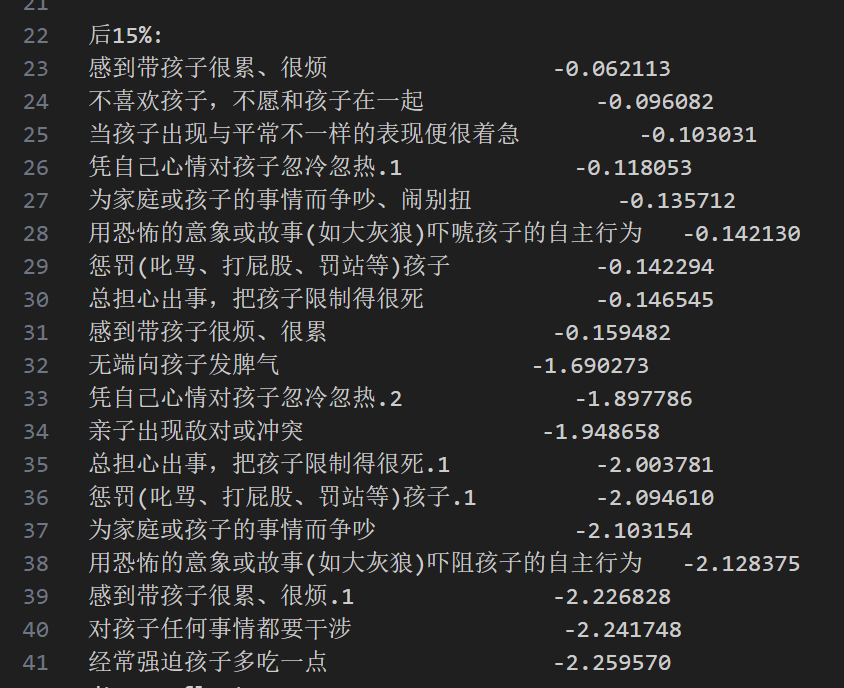
之后分析数据将直接使用 data\_num.csv 和 target\_num.csv。

1. **计算并查看均值**

首先我们看下总体的分布，画出每个问题的均值和方差条形图。



样本均值的排名保存在out/risk-full.txt中。



接下来探究不同组的均值情况。依据基本问题，将数据划分为多个组。例如，按照孩子年龄段不同，将数据按照3个年龄段划分为3个组。然后分别对每个组进行求列平均的操作。我们可以了解所有家庭在这些因素上的总得分情况。

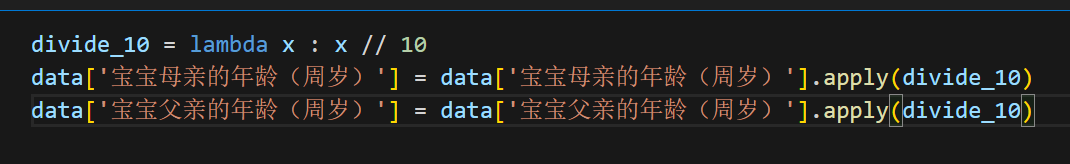
得分均值较低的列，说明普遍不重视，可以看作是养育风险因素。规定输出后15%的特征为养育风险因素。这里输出到out/risk.txt。



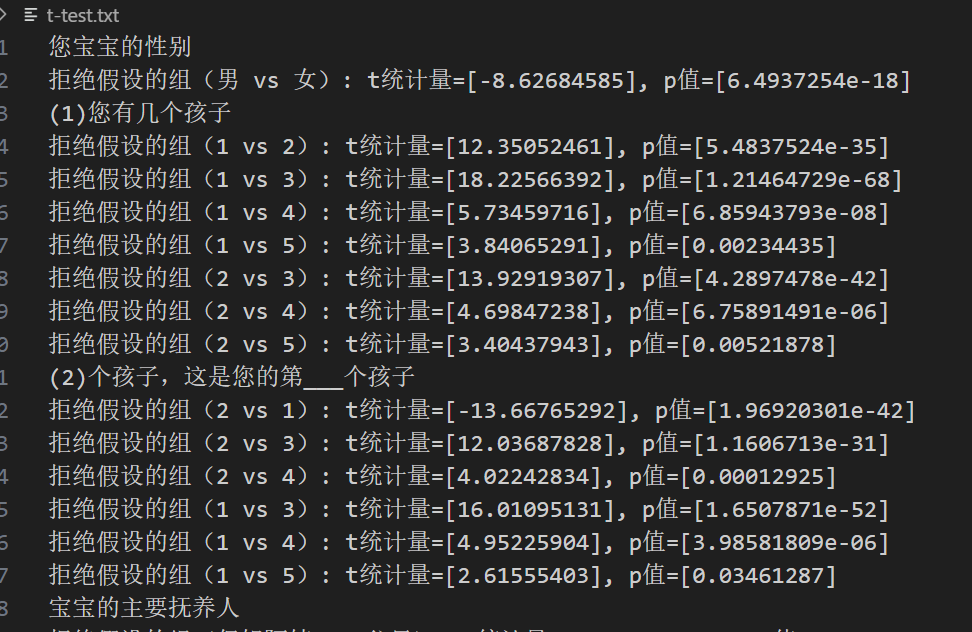
1. **两两进行t检验（置信度0.95）**

先前我们通过分组输出均值，找到了较低的得分问题，这个部分我们探究哪些因素对总分有影响。使用t检验，对两组样本做u1 – u2 即均值差检验。

此处做了一个小改动，父母亲年龄的值均除以10，也就是按照10岁一个年龄段分组，因为如果每一岁一个组，组数太多了，每个组样本也很少。



接下来开始t检验，置信度设置0.95。详细过程参考源代码及注释。最终我们将结果输出到out/t-test.txt文件中。



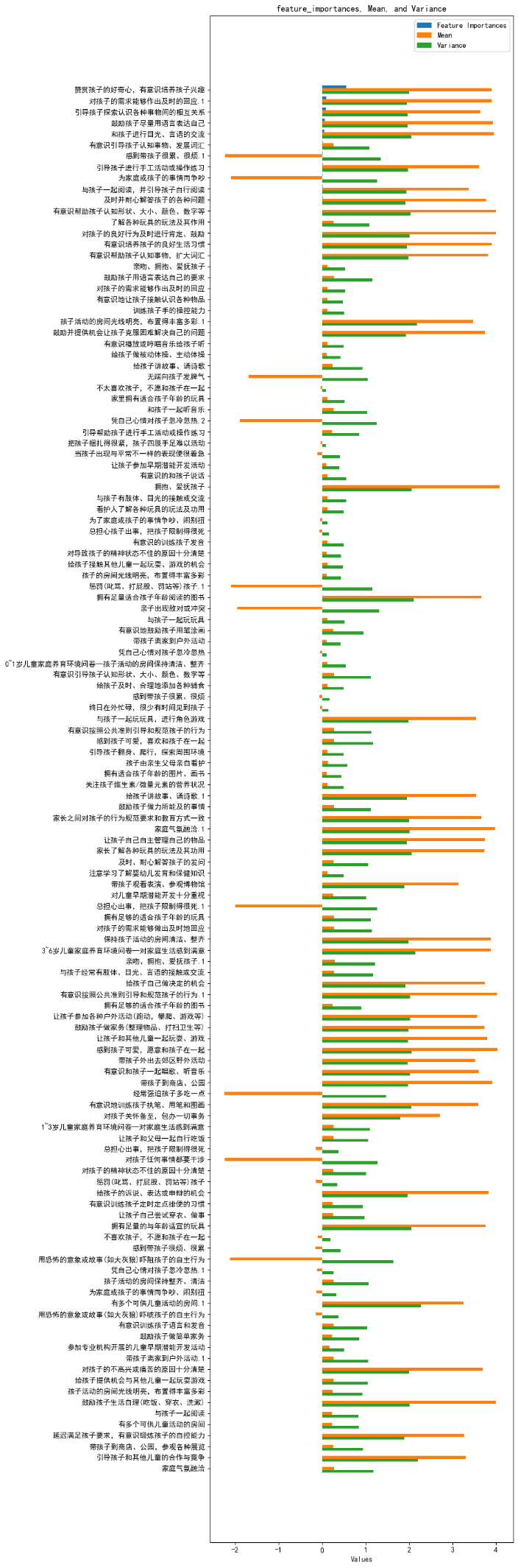
1. **关于能否使用机器学习模型的一点思考——以随机森林为例。**

在一开始我就提到，由于总分和特征相关，总分实际上如同没有一样，因此无法训练机器学习模型给出特征重要性。在我与其他同学交流过程中，有的人直接将数据输入给随机森林模型，输出特征重要性，我觉得这样不对，因此接下来强行套用机器学习模型来看下效果。

对于线性回归模型，特征重要性就是其系数coef。由于总分是特征直接加和而来，用线性回归去拟合数据大概率得到一个全接近1的系数，此处不做验证。

这里我们测试一个常用的非线性模型：随机森林。随机森林输出的特征重要性简单来说，是跟据每个节点分裂所减少的不纯度来算的。我的猜想是均值更大的特征可能影响更大，随机森林会以为这些特征更加重要。

下面训练了一个随机森林回归模型，将输出的特征重要性按降序排列，并画出对应的均值和方差。



可以看出，feature\_importance具有这样的特点：

* 较大重要性的特征，均值和方差都比较大。
* 排名靠后的特征，feature\_importance的值相差很小且十分接近0
* 排名靠后的特征，均值和方差无明显规律

因此我认为，随机森林以及其他机器学习模型应该不能解释特征重要性，归根结底是标签(label)和特征(feature)不独立造成的。

# 四、得出结论

在risk.txt文件中，保存了满足分析条件的特征。我们将这些特征按照：感知/认知、情感温暖、忽视/限制、人际互动/喂养、社会适应/自理、活动多样性/游戏参与这几个风险因素进行分类。统计每个分类下的特征条目数量。由于每个特征的权重相同，因此可以认为底下特征条目多的风险因素是该分类依据下的主要养育风险因素。

在年龄为分类依据的情况下，对风险因素包含的特征数量统计结果为：感知/认知：6，情感温暖：8，忽视/限制：12，人际互动/喂养：6，社会适应/自理：4，活动多样性/游戏参与：2。因此我们认为，忽视/限制是该分类依据下的主要养育风险因素。

根据如上方法，可以得出其他分类依据下的养育风险因素。

// 上面这一段请你重写一下，分123点这样写，加上图片，整多一点

// 思路：1. risk-full.txt 分析编故事

// 2. t检验找几个说说，然后说剩下的去查txt文件