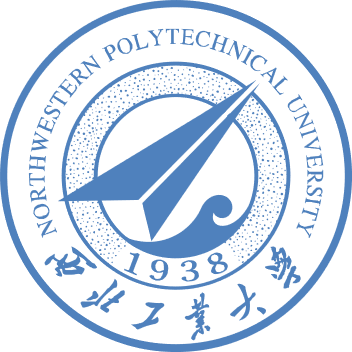
**西北工业大学**

**《数据分析与挖掘》**

**儿童养育环境风险因素分析报告**



|  |  |  |  |
| --- | --- | --- | --- |
| 学院 | 计算机学院 | | |
| 学号 | 2022302628 | 2022302626 |  |
| 姓名 | 陈浩明 | 李一晟 |  |
| 专业 | 数据科学与大数据技术 | | |
| 提交时间 | 2024/7/6 | | |

摘 要

本文旨在通过深入分析一项针对家长的问卷调查数据，系统性地识别并评估影响儿童健康成长的主要养育环境风险因素。首先进行数据预处理，包括缺失值填充、数据清洗、数据集分割。接着进行总体均值和方差的计算与可视化，以样本均值的15分位点为界限筛选风险因素。最后针对不同分组条件下的样本进行t检验，计算哪些因素会导致样本总分的均值发生差异。

基于上述发现，本文提出了一系列针对性强、具有可操作性的建议与策略，旨在帮助家庭、社区及政策制定者识别并有效干预这些风险因素。

**关键词：数据预处理、t检验**

目录

[摘 要 2](#_Toc171153316)

[一、 问题背景 4](#_Toc171153317)

[1. 题目背景及说明。 4](#_Toc171153318)

[2. 数据集构成 4](#_Toc171153319)

[3. 评分标准 4](#_Toc171153320)

[二、 分析思路概述 5](#_Toc171153321)

[1. 如何定义“主要养育风险因素”？ 5](#_Toc171153322)

[2. 数据集该如何使用？ 5](#_Toc171153323)

[三、 分析过程 5](#_Toc171153324)

[1. 数据预处理 5](#_Toc171153325)

[2. 计算并查看均值 8](#_Toc171153326)

[3. 两两进行t检验（置信度0.95） 10](#_Toc171153327)

[4. 关于能否使用机器学习模型的一点思考——以随机森林为例。 11](#_Toc171153328)

[四、得出结论 13](#_Toc171153329)

[1.总体结果解读 13](#_Toc171153330)

[2.t检验结果解读 17](#_Toc171153331)

# 问题背景

## 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。

## 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 |

## 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

## 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。

## 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

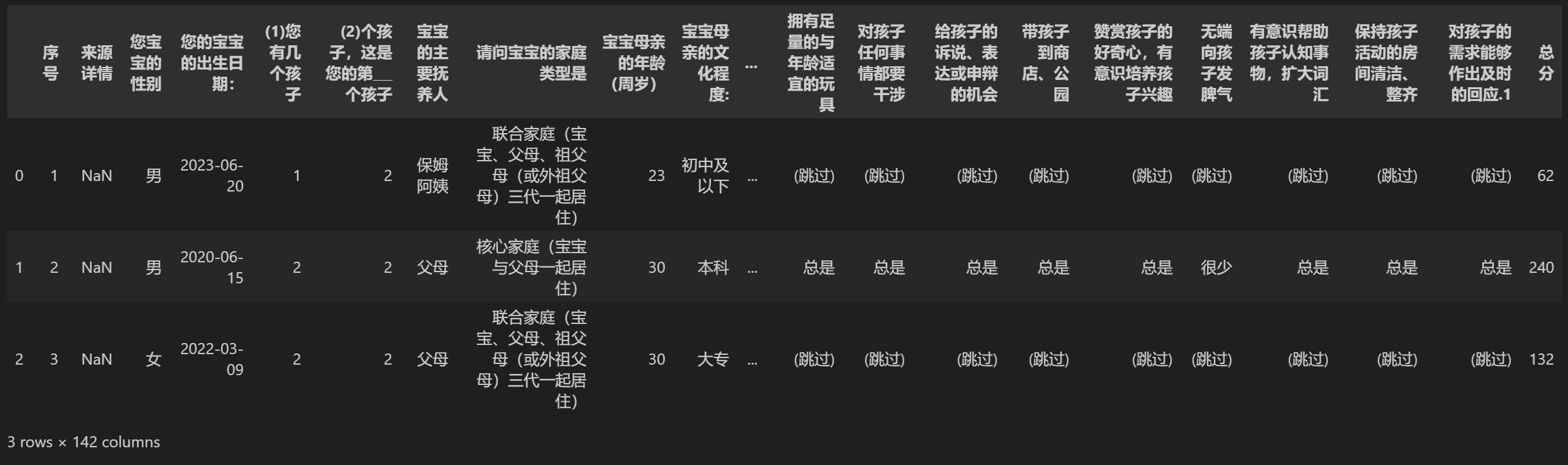
剩下三个部分，是对不同年龄段孩子的题目，可以看成是特征，这一部分也构成了最后的总分。

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。

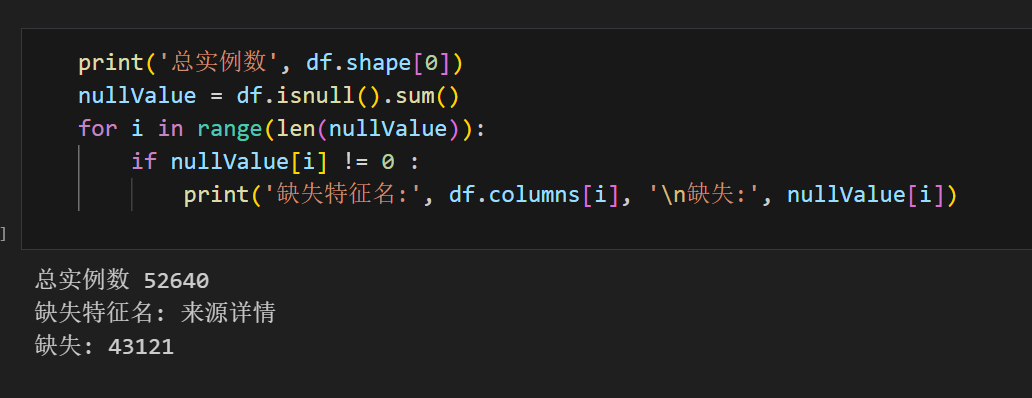
# 分析过程

## 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。

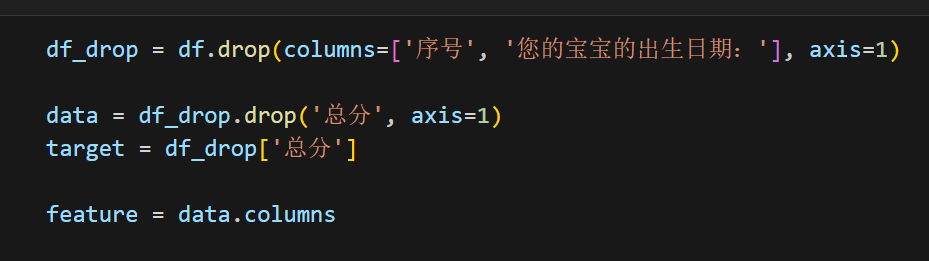


接着我们检查下其他列还有没有缺失值。可以看出只有这一列。



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。

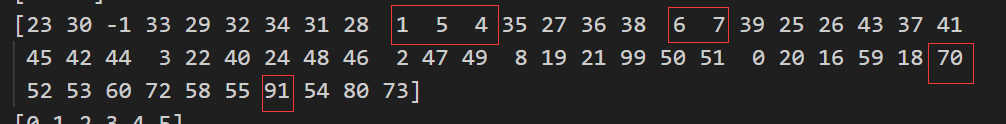


解决方案：对第一个问题，数字字符串可以直接通过转换得到int类型数据，对于其他输入，我们人工挑选出一些出现次数比较多的答案，例如“一”、“二”。第二个问题，我们将答案为“其他\*”的数据编码为同一个值。最后，将无法转换的值编码为int类型 -1，并在后续中删去包含-1的行。

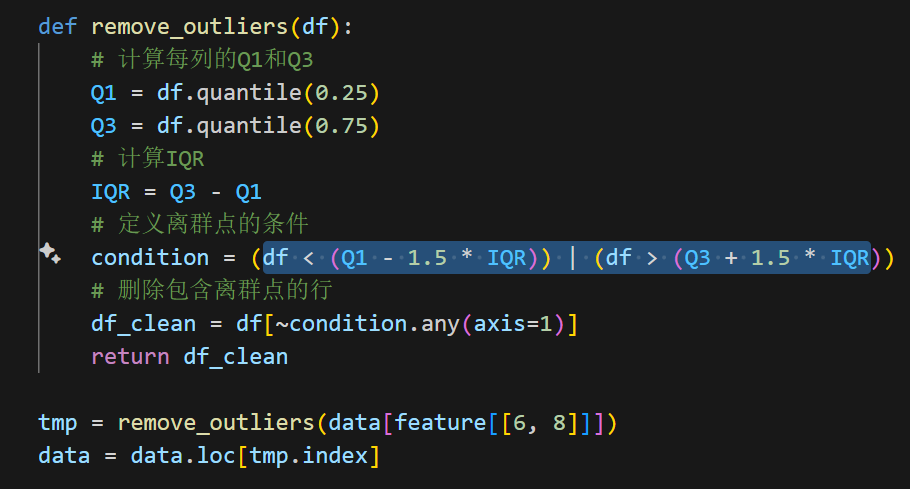
具体实现很枯燥，请查看源代码及注释。转换后的数据概览如下。



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。

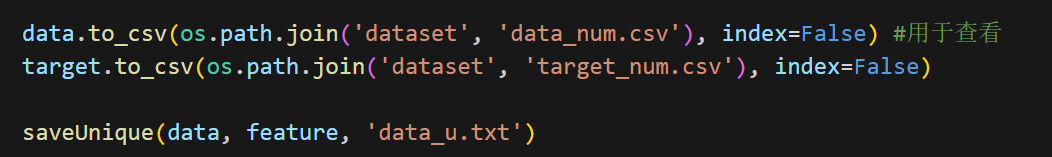


解决方法是采用离群点检测，删除在Q1 - 1.5 \* IQR 和 Q3 + 1.5 \* IQR之外的数据。

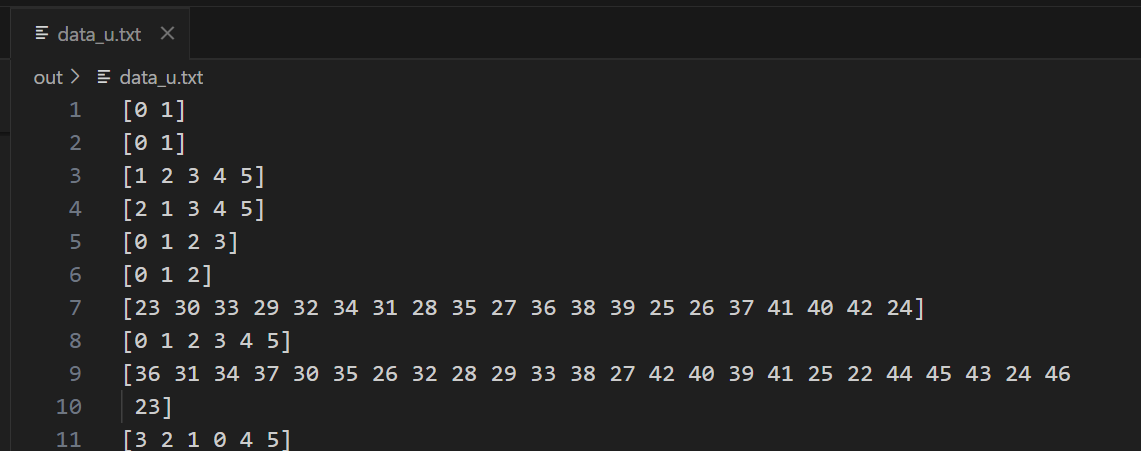


基础问题部分的处理到此结束，后面3个部分的问题都只有固定的5种答案（实际还有一个“空”答案，影响不大，编码的时候计0就行）。将5种答案转换成对应的分数即可。详细参考源代码及注释。

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。



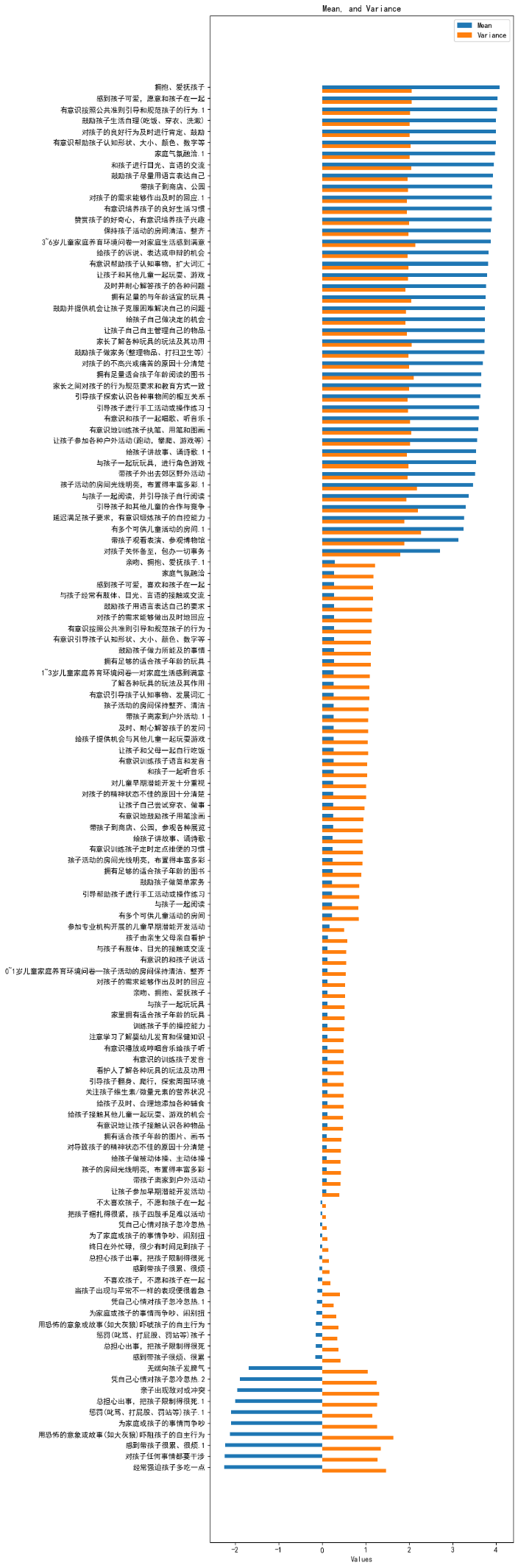
检查没有发现非法值。



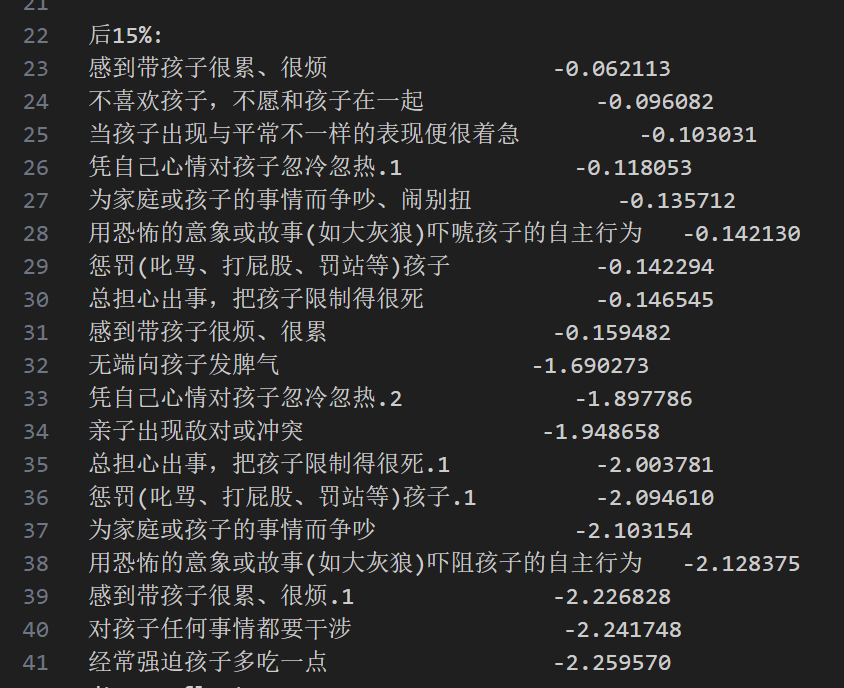
之后分析数据将直接使用 data\_num.csv 和 target\_num.csv。

## 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。

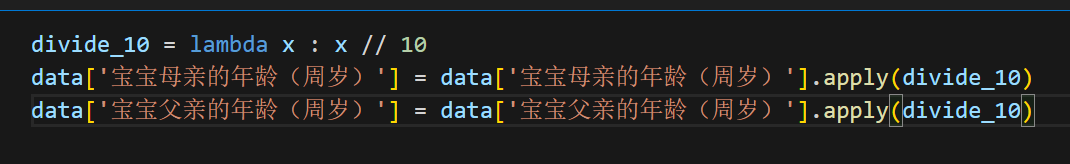
得分均值较低的列，说明普遍不重视，可以看作是养育风险因素。规定输出后15%的特征为养育风险因素。这里输出到out/risk.txt。



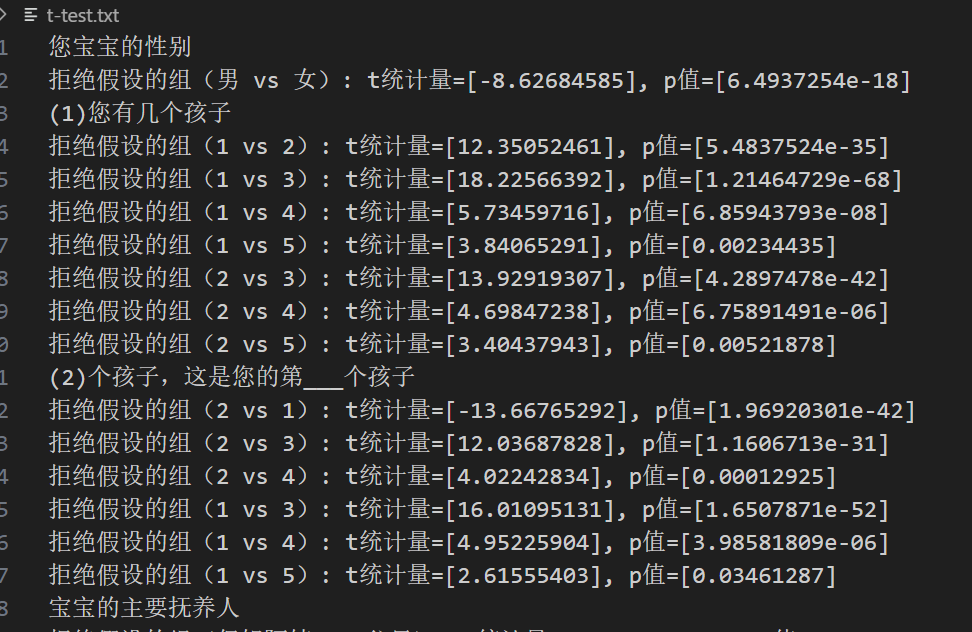
## 两两进行t检验（置信度0.95）

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。



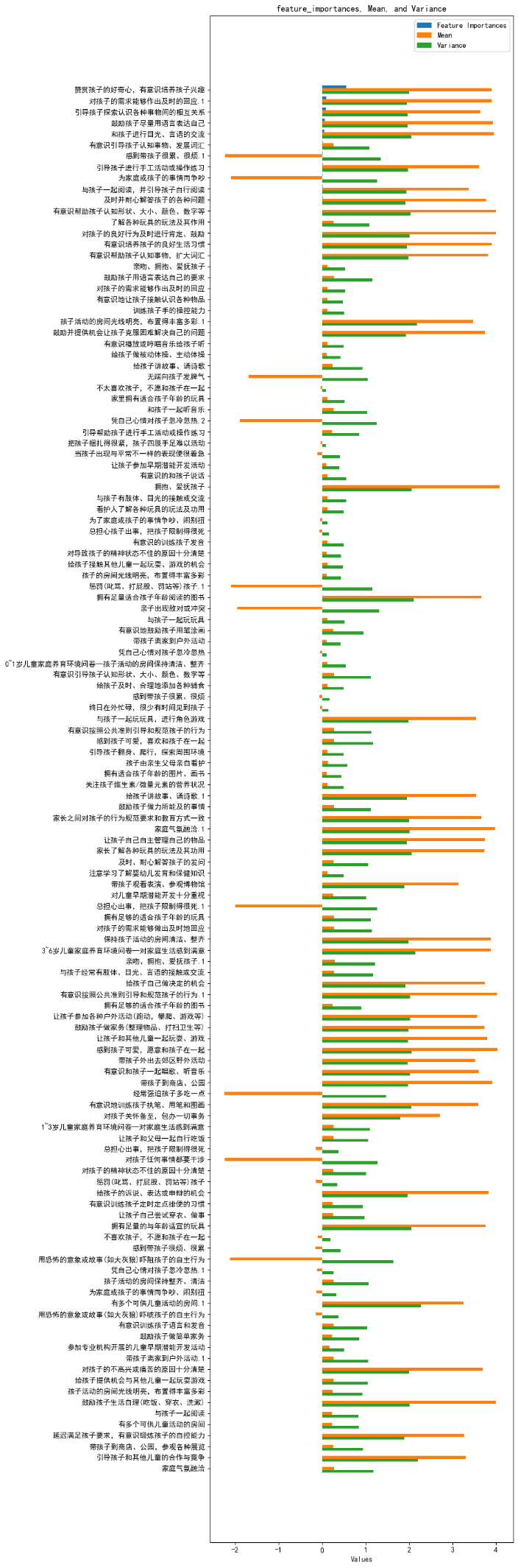
## 关于能否使用机器学习模型的一点思考——以随机森林为例。

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

1.总体结果解读

根据输出结果，我们获取了得分前15%的标签，并且认为这部分标签的得分高对孩子养育有积极作用。同时也得到了得分后15%的标签，认为这部分得分低对孩子养育有负面影响。

文本

描述已自动生成

接下来我们将这些标签按照：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。

对于前15%标签分类：

（1）感知/认知

有意识帮助孩子认知形状、大小、颜色、数字等

有意识帮助孩子认知事物，扩大词汇

总计：2次

（2）情感温暖

拥抱、爱抚孩子

感到孩子可爱，愿意和孩子在一起

和孩子进行目光、言语的交流

赞赏孩子的好奇心，有意识培养孩子兴趣

给孩子的诉说、表达或申辩的机会

总计：5次

（3）忽视/限制

无

（4）人际互动/喂养

带孩子到商店、公园

让孩子和其他儿童一起玩耍、游戏

总计：2次

（5）社会适应/自理

鼓励孩子生活自理(吃饭、穿衣、洗漱)

有意识培养孩子的良好生活习惯

总计：2次

（6）活动多样性/游戏参与

家庭气氛融洽.1

对孩子的需求能够作出及时的回应.1

对孩子的良好行为及时进行肯定、鼓励

鼓励孩子尽量用语言表达自己

有意识按照公共准则引导和规范孩子的行为.1

3~6岁儿童家庭养育环境问卷—对家庭生活感到满意

保持孩子活动的房间清洁、整齐

总计：7次

对以上信息进行汇总，得到如下结果：

感知/认知：2次

情感温暖：5次

忽视/限制：0次

人际互动/喂养：2次

社会适应/自理：2次

活动多样性/游戏参与：7次

通过上面的统计，容易得出，在前15%的标签中，活动多样性/游戏参与和情感温暖方面的条目较多，表明这些方面对于积极的养育环境有较大的贡献。

对后15%的标签进行分类：

（1）感知/认知

用恐怖的意象或故事(如大灰狼)吓唬孩子的自主行为

用恐怖的意象或故事(如大灰狼)吓阻孩子的自主行为

当孩子出现与平常不一样的表现便很着急

总计：3次

（2）情感温暖

凭自己心情对孩子忽冷忽热.1

凭自己心情对孩子忽冷忽热.2

不喜欢孩子，不愿和孩子在一起

无端向孩子发脾气

总计：4次

（3）忽视/限制

总担心出事，把孩子限制得很死

总担心出事，把孩子限制得很死.1

对孩子任何事情都要干涉

感到带孩子很累、很烦

感到带孩子很累、很烦.1

感到带孩子很烦、很累

总计：6次

（4）人际互动/喂养

经常强迫孩子多吃一点

总计：1次

（5）社会适应/自理

惩罚(叱骂、打屁股、罚站等)孩子

惩罚(叱骂、打屁股、罚站等)孩子.1

总计：2次

（6）活动多样性/游戏参与

为家庭或孩子的事情而争吵

为家庭或孩子的事情而争吵、闹别扭

亲子出现敌对或冲突

总计：3次

经过统计，得到如下结果：

感知/认知：3次

情感温暖：4次

忽视/限制：6次

人际互动/喂养：1次

社会适应/自理：2次

活动多样性/游戏参与：3次

通过上面的统计，容易得出，忽视/限制和情感温暖方面的条目较多，表明这些方面的问题是主要的养育风险因素。

综上所述，忽视/限制和情感温暖是影响养育风险的主要方面。同时，感知/认知和活动多样性/游戏参与也有较大的影响。

2.t检验结果解读

对于孩子数量、主要扶养人、父母年龄和文化程度等特征，采用t检验进行结果检验。我们假设用于t检验的组两两互无差异。通过查看t统计量与p值，我们确定是否拒绝原假设。如果拒绝原假设，意味着这两个组有差异；反之则说明无差异。

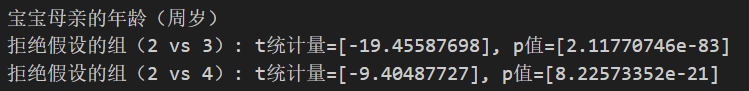
例如，我们以“(1)您有几个孩子”为例子进行分析。对各个孩子数量两两进行检验，将p值小于0.05的并输出结果，如下所示：

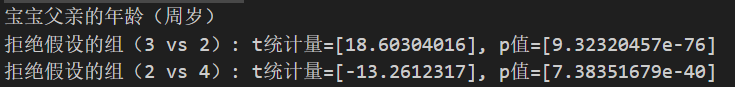
文本

中度可信度描述已自动生成

结果表明，1vs2、1vs3、...、2vs5这几组之间有差异。说明孩子个数会影响总分即养育水平。

利用同样的方法，对特征检验，并将要拒绝的组筛选出来，输出到t-test.txt文件中。结果表明几乎基本问题中的每个因素都会造成总分均值有差异（置信度0.95），但也有一些情况下没有差异，例如父亲、母亲年龄30岁和40岁，其总分没有差异不大（不拒绝假设）。





此外我们还看到，体检的社区位置也会影响总分，这属于是间接因素，因为社区位置影响家庭经济情况。

