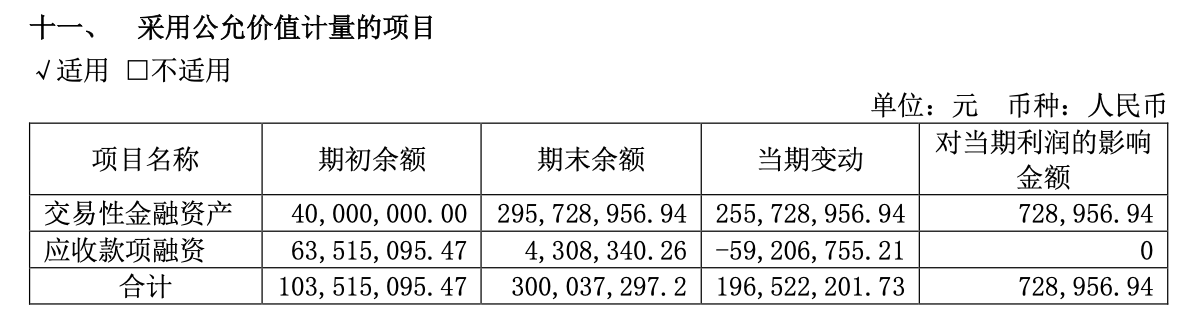
# 练习题

1. 任务：
   1. 1. 熟悉PDF文档的OCR JSON结果的结构
   2. 2.通读原pdf后，总结需要处理的表格cases（包括但不限于任务3中描述的cases）
   3. 3.写一个python程序（必要的时可以使用llm/AI，并附上必要性论述），将JSON格式存储的表格进行拼接/提取
      1. 输入：
         1. 一份年报pdf ocr结果
      2. 输出：
         1. 一个结构化的表格提取
         2. 字段
            1. 'title'
            2. 'unit'
            3. 'header'
            4. 'key\_index'
            5. 'values'
      3. 必做 Case：
         1. 1.普通表格
            1. 原表格:



* + - * 1. target output:

{'title':'十一、 采用公允价值计量的项目','unit':'单位:元 币种:人民币','header':['项目名称','期初余额','期末余额','当期变动','对当期利润的影响金额'],'key\_index':['交易性金融资产','应收款项融资','合计'],'values':[['40,000,000.00','295,728,956.94','255,728,956.94','728,956.94'],['63,515,095.47','4,308,340.26','-59,206,755.21','0'],['103,515,095.47','300,037,297.2','196,522,201.73','728,956.94']]}

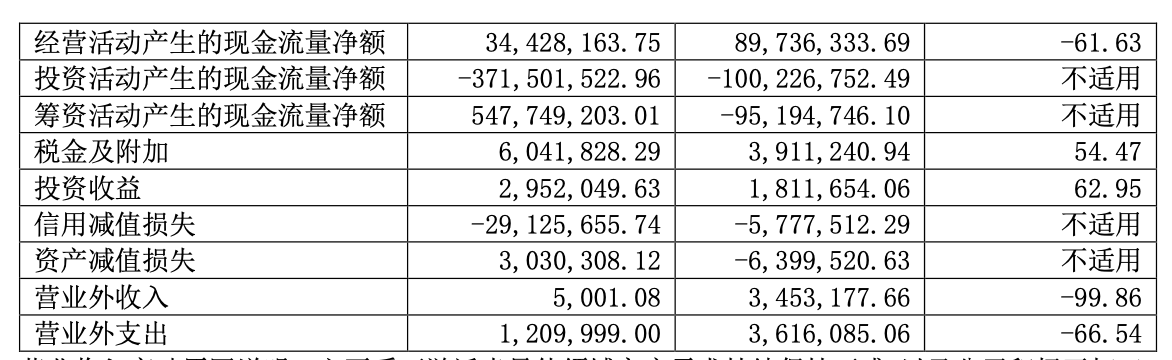
* + - 1. 2.特殊表格:
         1. 2.1 跨页表格

原表格

第一页内容



第二页内容

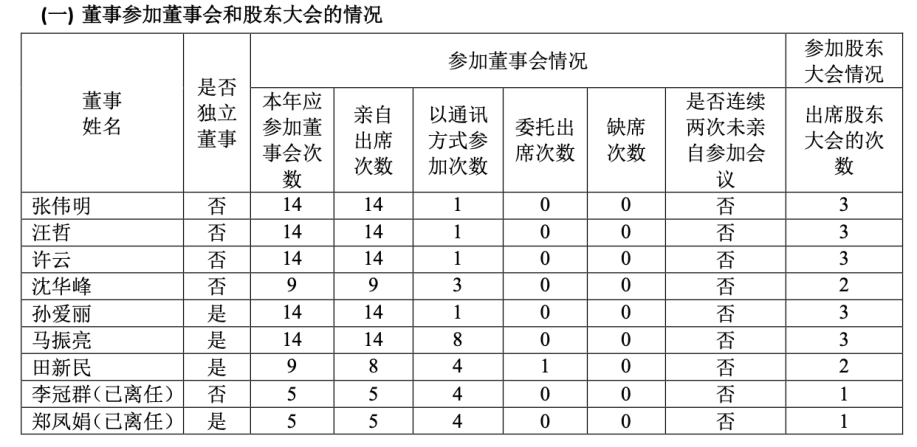


target output:

{'title':'利润表及现金流量表相关科目变动分析表','unit':'单位:元 币种:人民币','header':['科目','本期数','上年同期数变动比例（%）'],'key\_index':['营业收入','营业成本','销售费用','管理费用','财务费用','研发费用','经营活动产生的现金流量净额','投资活动产生的现金流量净额','筹资活动产生的现金流量净额','税金及附加','投资收益','信用减值损失','资产减值损失','营业外收入','营业外支出'],'values':[['1,233,029,658.53','937,742,067.28','31.49'],['897,460,584.72','676,894,628.34','32.58'],['44,390,860.54','28, 441,472.16','56.08'],['43,365,167.66','31,635,506.71','37.08'],['-3,235,120.25','9,644,278.00','不适用'],['56,186,435.18','38,921,613.29','44.36'],['34,428,163.75','89,736,333.69','-61.63'],['-371,501,522.96','-100,226,752.49','不适用'],['547,749,203.01','-95,194,746.10','不适用'],['6,041,828.29','3,911,240.94','54.47'],['2,952,049.63','1,811,654.06','62.95'],['-29,125,655.74','-5,777,512.29','不适用'],['3,030,308.12','-6,399,520.63','不适用'],['5,001.08','3,453,177.66','-99.86'],['1,209,999.00','3,616,085.06','-66.54']]}

* + - * 1. 2.2 多层级表格

原表格



target output:

{'title':'(一) 董事参加董事会和股东大会的情况','unit':'','header':[['董事姓名','是否独立董事','参加董事会情况','参加董事会情况','参加董事会情况','参加董事会情况','参加董事会情况','参加董事会情况','参加股东大会情况'],['董事姓名','是否独立董事','本年应参加董事会次数','亲自出席次数','以通讯方式参加次数','委托出席次数','缺席次数','是否连续两次未亲自参加会议','出席股东大户的次数']],'key\_index':['张伟明',',汪哲','许云','沈华峰','孙爱丽','马振亮','田新民','李冠群(已离任)','郑凤娟(已离任)'],'values':[['否','14','14','1','0','0','否','3'],['否','14','14','1','0','0','否','3'],['否','14','14','1','0','0','否','3'],['否','9','9','1','0','0','否','2'],['否','14','14','1','0','0','否','3'],['否','14','14','8','0','0','否','3'],['否','9','8','4','1','0','否','2'],['否','5','5','4','0','0','否','1'],['否','5','5','4','0','0','否','1']]}

* + 1. 其他edge case可仅做思路设计