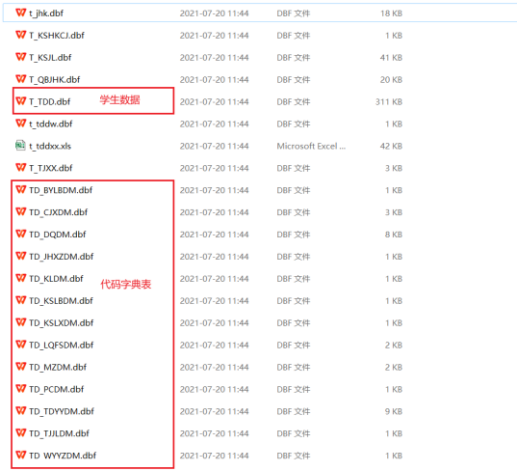
新生数据库结构说明

前言

新生数据库50余个字段,这些字段的数据都来自每个省份录取系统导出的DBF文件中，有些字段可以直接与我们的框架结构对应，但是有些字段需要进行一定的转义才能和我们的框架进行对应，本文档从省份的录取数据开始，对这些字段进行说明。

一、 各省份录取数据说明

当一个省的某一个批次投档结束并确定录取结果时，会导出该省份该批次的录取数据，这些是录取总表数据库制作的数据来源。这些录取数据的形式为多个 DBF，即 DBF 文件夹，这些 DBF 中包含的是学生的基本信息，以及基本信息中某些字段的字典表，如下图，下面将对每一个文件进行说明。



图一：录取数据源总览

DBF 文件夹中的数据的类型与是否为新高考省份有很大关系，下面分别对其进行解释。

1. 非新高考省份

2021 年非高考省份中有 27 个文件，经过分析后，可以得到分别表一的涵义。其中最核心的是 T_TDD.dbf 文件，该文件记录了考生的几乎所有信息，这些文件中包含了很多字典库，大部分对应的是 T_TDD 中的代码，并将其诠释为名称，**这些代码库对于每个省是不同的，所以如果需要对每个省进行代码映射，需要从每个省下手，不可以将一个省的代码直接对应所有省份。**除了 T_TDD.dbf 之外，另一个核心是 t_jhk.dbf 文件，该表记录了该省该批次招生的计划，对应的是 T_TDD 表中 LQZY 字段（记录专业代码），即可以得到对应的专业名称。

表一：非新高考省份 DBF 文件结构

文件名	文件涵义
T_TDD.dbf	考生信息核心库，包含了我们需要的数据库结构中几乎所有的字段，几乎所有的数据都从这张表来
t_jhk.dbf	记录该省份该批次的招生计划，包含专业代码，专业名称以及招生计划数
T_KSHKCJ.dbf	每个考生选课的成绩项，在非高考省份表为空
T_KSJL.dbf	每一个考生的经历，包含其高中时期的学校和担任职务及其起始日期和终止日期，考生在年级上升时，会导致证明人不同，所以一个考生会有多条记录
T_QBJHK.dbf	记录该省份所有批次的招生计划，包含专业代码、专业名称及招生计划数
t_tddw.dbf	投档单位代码库，将 T_TDD 中投档单位代码（TDDWDM）诠释为名称，其中也包含了当前录取结果的批次以及科类，不同身份有不用的表述，如杭州电子

	科技大学或 xxx 专业
T_TJXX.dbf	该省该批次考生的体检信息
t_tddxx.xls	桌面投档单，今年将弃用
TD_BYLBDM.dbf	毕业类别库，将 T_TDD 中的毕业类别代码（BYLBDM）诠释为名称
TD_CJXDM.dbf	**成绩项代码**，该表对 T_TDD 中的只有标号的成绩项进行解释，比如 GKCJX01 对应语文，该表将在下面进行详细说明
TD_DQDM.dbf	地区代码字典库，将 T_TDD 中的地区代码（DQDM）诠释为名称
TD_JHXZDM.dbf	计划性质代码库，将 T_TDD 中计划性质代码（JHXZ）诠释为名称
TD_KLDM.dbf	科类代码库，将 T_TDD 中科类代码（KLDM）诠释为名称
TD_KSLBDM.dbf	考生类别代码库，将 T_TDD 中考生类别代码（KSLBDM）诠释为名称
TD_KSLXDM.dbf	考试类型代码库，将 T_TDD 中考试类型代码（KSLXDM）诠释为名称
TD_LQFSDM.dbf	录取方式代码库，将 T_TDD 中录取方式代码（LQFS）诠释为名称
TD_MZDM.dbf	名族代码库，将 T_TDD 中名族代码（MZDM）诠释为名称
TD_PCDM.dbf	批次代码库，将 T_TDD 中批次代码（PCDM）诠释为名称
TD_TDYYDM.dbf	退档原因代码库，将 T_TDD 中退档原因代码（TDYYDM）诠释为名称，对于录取的考生来说，该字段为空
TD_TJJLDM.dbf	体检记录代码库，将 T_TJXX 中体检记录代码（TJJLDM）诠释为名称
TD_WYYZDM.dbf	外语语种代码库，将 T_TDD 中外语语种代码（WYYZDM）诠释为名称
TD_XBDM.dbf	性别代码库，将 T_TDD 中性别代码（XBDM）诠释为名称
TD_XTDWDM.dbf	考生系统单位库，
TD_ZCDM.dbf	政策代码库，目的不明
TD_ZYTZDM.dbf	志愿特征代码库，将 T_TDD 中志愿特征代码（ZYTZ）诠释为名称，可以理解为是否调剂
TD_ZZMMDM.dbf	政治面貌代码库，将 T_TDD 中政治面貌代码（ZZMMDM）诠释为名称

2. 新高考省份

2021 年新高考省份总计 44 个文件，其中大部分与非新高考省份类似。新高考省份将学生信息进行了分离，分为 T_BMK.dbf 和 T_TDD.dbf，其中 T_BMK.dbf 中存放的是考生的个人信息,包含个人基本信息以及考试成绩等，而在 T_TDD.dbf 中存放的是考生的投档信息，包含批次科类代码等,以及最重要的录取专业信息等。新高考省份由于引入了自由选课的机制，所以对于成绩项有了新的管理方式，即在非新高考省份表格中提到的 T_KSHKCJ.dbf。

另外和非新高考省份还有一个比较重要的不同在于专业的处理，新高考省份将招生专业记录在了 td_zydh 和 t_jhk 两个文件：t_jhk 中记录的是该省份该批次招生计划，它并没有记录专业的名称等具体信息，而 td_zydh 记录的是专业的具体信息包括了专业名称等。具体文件结构见表二

表二：新高考省份 DBF 文件结构

T_BMK.dbf	考生的个人基本信息，包含姓名、政治面貌、高考成绩等
T_TDD.dbf	考生投档信息记录表，包含考生的投档有关的信息，如科类代码、投档成绩以及录取专业
td_zydh.dbf	专业代号，该表为新高考省份专业目录，其中包含了专业代号和名称及其一些相关信息，T_TDD 中的 LQZY 的代号对应的即为这张表
t_jhk.dbf	记录该省份该批次的招生计划，包含专业代码，招生计划数，新高考省份的该表为表明专业的名称，需要到 td_zydh 中根据代码匹配
T_QBJHK.dbf	记录该省份所有批次的招生计划，包含专业代码和招生计划数等信息，专业名称与 t_hjk 同理
T_TJXX.dbf	考生的体检信息
t_zykmx.dbf	志愿信息记录表，记录每个考生的所有志愿信息，如第一志愿到第六志愿，包含序号与志愿专业序号
T_KSJL.dbf	每一个考生的经历，包含其高中时期的学校和担任职务及其起始日期和终止日期，考生在年级上升时，会导致证明人不同，所以一个考生会有

	多条记录
T_KSXKCJ.dbf	考生学考成绩表，记录每个考生的所有学考科目
t_tddw.dbf	投档单位代码库，将 T_TDD 中投档单位代码（TDDWDM）诠释为名称，其中也包含了当前录取结果的批次以及科类，不同身份有不用的表述，如杭州电子科技大学或 xxx 专业
t_tddxx.xls	桌面投档单，今年将考虑弃用
td_xkcjdj.dbf	学考成绩登记表，为空
TD_YXCJXDM.dbf	意义不明
td_zydhdmdz.dbf	专业代号与专业代码对应表，意义不明
TD_BMDWDM.dbf	报名单位代码库，将 T_BMK 的报名单位代码（BMDW）诠释为名称
TD_BYLBDM.dbf	毕业类别库，将 T_BMK 中的毕业类别代码（BYLBDM）诠释为名称
TD_CCDM.dbf	层次代码库，如本科、专科等，目的不明
TD_CJXDM.dbf	成绩项代码，该表对 T_BMK 中的只有标号的成绩项进行解释，比如 GKCJX01 对应语文，该表将在下面进行详细说明
TD_CZLBDM.dbf	残障类别代码库，将 T_BMK 中的残障类别代码（CZLBDM）诠释为名称
TD_DQDM.dbf	地区代码字典库，将 T_BMK 中的地区代码（DQDM）诠释为名称
TD_HJLBDM.dbf	户籍类别代码库，将 T_BMK 中的户籍类别代码（HJLBDM）诠释为名称
TD_JHLBDM.dbf	计划类别代码库，将 T_TDD 中的计划类别代码（JHLBDM）诠释为名称
TD_JHXZDM.dbf	计划性质代码库，将 T_TDD 中计划性质代码（JHXZDM）诠释为名称
TD_KLDM.dbf	科类代码库，将 T_TDD 中科类代码（KLDM）诠释为名称
TD_KMDM.dbf	科目代码库，对应 T_BMK 中选考科目的编号以及招生计划中某专业需要报考的科目列表
TD_KSLBDM.dbf	考生类别代码库，将 T_BMK 中考生类别代码（KSLBDM）诠释为名称
TD_KSLXDM.dbf	考生类型代码库，将 T_BMK 中考生类型代码（KSLXDM）诠释为名称
TD_KSTZDM.dbf	考生特征代码库，将 T_BMK 中考生特征代码（KSTZDM）诠释为名称，并在这有分数的加成（ZGFS）
TD_KSZGDM.dbf	意义不明
TD_LQLXDM.dbf	录取类型代码库，将 T_TDD 中录取类型代码（LQLXDM）诠释为名称
TD_MZDM.dbf	名族代码库，将 T_BMK 中名族代码（MZDM）诠释为名称
TD_PCDM.dbf	批次代码库，将 T_TDD 中批次代码（PCDM）诠释为名称
TD_TDLXDM.dbf	投档类型代码库，将 T_TDD 中投档类型代码（TDLXDM）诠释为名称
TD_TDYYDM.dbf	退档原因代码库，将 T_TDD 中退档原因代码（TDYYDM）诠释为名称，对于录取的考生来说，该字段为空
TD_TJJLDM.dbf	体检记录代码库，将 T_TJXX 中体检记录代码（TJJLDM）诠释为名称
TD_WYYZDM.dbf	外语语种代码库，将 T_BMK 中外语语种代码（WYYZDM）诠释为名称
TD_XBDM.dbf	性别代码库，将 T_BMK 中性别代码（XBDM）诠释为名称
TD_XZDM.dbf	学制代码库，将 td_zydh 中学制代码（XZDM）诠释为名称
TD_ZJLXDM.dbf	证件类型代码库，将 T_BMK 中证件类型代码（ZJLXDM）诠释为名称
TD_ZKLXDM.dbf	招考类型
TD_ZYLBDM.dbf	专业类别代码库，将 td_zydh 中专业类别代码（ZYDMDM）诠释为名称，如普通类、师范类等
TD_ZYTJLXDM.dbf	志愿调剂类型代码库，将 T_TDD 中志愿调剂类型代码（ZYTJLXDM）诠释为名称
TD_ZYTZDM.dbf	志愿特征代码库，将 T_TDD 中志愿特征代码（ZYTZ）诠释为名称，可以理解为是否调剂
TD_ZZMMDM.dbf	政治面貌代码库，将 T_BMK 中政治面貌代码（ZZMMDM）诠释为名称

二、 数据库结构制作

经过以上对数据库的数据源做了分析之后，可以大概清楚每个省份的数据有哪几部分组成以及需要从哪边寻找总数据库的各个字段，下面分析新高考省份和非新高考省份的数据源和总表的对应关系。对于数据库结构中的字段，有一部分是可以直接在数据源中查找字段名称获取数据，但还是有部分无法做到一一对应，需要特殊处理，以下分析根据这两个情况进行阐述。另外成绩是数据库中最复杂的一个部分，该文档将单独进行分析。

注意：因为新高考省份的数据来源于两张表（T_TDD 和 T_BMK），所以首先需要将这两张表根据 KSH 进行合并，从这张合并表进行查找，以下将这样合并表也叫做 T_TDD。

1. 简单字段

简单情况是指字段名可以直接从 T_TDD 中获取到该字段，从人工手段来讲其实就是可以把 TDD 中的字段直接复制过来。

表三：简单字段与数据源的对应情况

字段名	字段名称	数据来源
KSH	考生号	T_TDD 表中的 KSH
ZKZH	准考证号	T_TDD 表中的 ZKZH
XM	姓名	T_TDD 表中的 XM
XBDM	性别	T_TDD 表中的 XBDM,如果需要将其改为中文时,需要从 TD_XBDM.dbf 获取中文
CSNY	出身年月	特殊情况
ZZMMDM	政治面貌	T_TDD 表中的 ZZMMDM，如果需要将其改为中文时，需要从 TD_ZZMMDM.dbf 获取中文
MZDM	民族	T_TDD 表中的 MZDM,如果需要将其改为中文时,需要从 TD_MZDM.dbf 获取中文
KSLBDM	考生类别	T_TDD 表中的 KSLBDM，如果需要将其改为中文时，需要从 TD_KSLBDM.dbf 获取中文
WYYZDM	外语语种代码	T_TDD 表中的 WYYZDM，如果需要将其改为中文时，需要从 TD_WYYZDM.dbf 获取中文
HKDJ	会考等级	T_TDD 表中的 HKDJ（一般为空）
BMDW	报名单位	T_TDD 表中的 BMDW（非新高考省份并没有找到中文对应表）
KSTZ	考生特征	T_TDD 表中的 KSTZ
XTDW	考生系统单位	T_TDD 表中的 XTDW，如果需要将其改为中文时，需要从 TD_XTDWDM.dbf 获取中文（新高考省份没有这张表）
DQDM	地区代码	T_TDD 表中的 DQDM,如果需要将其改为中文时,需要从 TD_DQDM.dbf 获取中文
YZBM	邮政编码	T_TDD 表中的 YZBM
HKKH	会考考号	T_TDD 表中的 HKKH
KSTC	考生特长	T_TDD 表中的 KSTC
KSJLHCF	考生奖励或处分	T_TDD 表中的 KSJLHCF
WYKS	外语口试	T_TDD 表中的 WYKS
ZSYJ	政审意见	T_TDD 表中的 ZSYJ
KSLXDM	考试类型代码	T_TDD 表中的 KSLXDM，如果需要将其改为中文，需要从

		TD_KSLXDM.dbf 获取中文
SJR	收件人	T_TDD 表中的 KSJLHCF
YSJZDM	应试卷种代码	T_TDD 表中的 YSJZDM
WYTL	外语听力	T_TDD 表中的 WYTL
PCDM	批次代码	T_TDD 表中的 PCDM
KLDM	科类代码	T_TDD 表中的 KLDM
TDDW	投档单位	T_TDD 表中的 TDDWDM，如果需要将其改为中文，需要从 t_tddw.dbf 获取中文
JHXZ	计划性质	T_TDD 表中的 JHXZDM，如果需要将其改为中文，需要从 TD_JHXZDM.dbf 获取中文
CJ	成绩	T_TDD 表中的 CJ
TDCJ	投档成绩	T_TDD 表中的 TDCJ
TDZY	投档志愿	T_TDD 表中的 TDZY
BM	通知书编号	自制
XY	学院	根据录取专业从《学院专业对应表》中获取学院
SYD	生源地	即为省份
KL	科类	由 KLDM 从 TD_KLDM.dbf 获取中文，但后期需要调整（做统一）
PC	批次	由 PCDM 从 TD_PCDM.dbf 获取中文，但后期需要调整（做统一）

2. 特殊字段

特殊情况是指新高考省份和非新高考省份的有些字段的叫法是不一样的,根据 2020 年和 2021 的经验，如下几个字段的名称存在多种叫法。处理时可以先判断是否为新高考省份，再去 T 查找对应的字段。特殊字段如下表：

表四：新高考省份与非新高考省份字段不同叫法

字段名	字段名称	新高考省份	非新高考
ZXDM	中学代码	BYXXDM	ZXDM
ZXMC	中学名称	BYXXMC	ZXMC
SFZH	身份证号	ZJHM	SFZH
JTDZ	家庭地址	TXDZ	JTDZ
LXDH	联系电话	LSXJ、LSDH	LXDH

3. 录取专业获取

录取专业是学生最重要的信息，它的获取方式与该省是否为新高考省份有一定联系，但是区别不大，录取专业对应的都是 T_TDD 中的 LQZY 字段，其存放的时候该学生被录取的专业代码，但新高考和非新高考在诠释专业代码方面有一定的区别。

新高考省份：专业代码与专业名称对应表为 td_zydh.dbf，表中的 ZYDH 即为 T_TDD 中 LQZY，ZYMC 即为专业名称

非新高考省份：专业代码与专业名称对应为 t_jhk.dbf，即招生计划库，表中的 ZYDH 即为 T_TDD 中 LQZY，ZYMC 即为专业名称

录取专业这边没有很大的问题，但是有一种情况需要注意，即防止不同专业的同名情况，如 2021 年的“电子信息类（电子信息学院）”和“电子信息类（通信工程）”。在某些省份的专业目录中，这两个专业都被叫做“电子信息类”，并没有在 ZYMC 中做出区分。这种情况下，需要注意专业目录中的 ZKFX 等信息，以获取其具体对应的专业。

4. 成绩项分析

成绩对于所有省份来说都是一个难题，因为数据源中记录的科目顺序和我们数据库中的科目顺序是完全不一样的，所以需要进行一一映射，因此我们需要知道 T_TDD 中存放的 GKCJX??到底是什么科目。如图二为 T_TDD 中记录的成绩项数据，其没有指定科目。

GKCJX01	GKCJX02	GKCJX03	GKCJX04
102.000000000	120.000000000	107.000000000	92.0000000
95.000000000	118.000000000	123.000000000	84.0000000

图二：T_TDD 成绩项

	A	B
	CJX	CJXMC
	01	语文
	02	数学
	03	外语
	04	物理
	05	化学
	06	生物
	07	政治
	08	历史
	09	地理

图三：TD_CJXDM.dbf

关于科目的对应，可以寻找 TD_CJXDM.dbf 文件作为依据，如图三。该文件记录的是成绩项编号对应的科目，我们可以将数据库中的高考成绩项所对应的科目匹配 CJXDM 中的科目，再得到 CJX 的序号。比如我们的 GKCJX01 对应科目是“语文”，我们根据“语文”查找其在 T_TDD 中的序号，查得为“01”，那么在 T_TDD 中的 GKCJX01 就是“语文”。

	A	B	
	CJX	CJXMC	C
	01	语文分	0
	02	数学分	0
	03	外语分	0
	04	物理分	0
	05	化学分	0
	06	生物分	0
	07	政治分	0
	08	历史分	0

图四：成绩项特例

根据以上方法可以解决比较多省份，但是还是会有出现问题，那就是有些省份的“语文”不叫“语文”，如图四所示，该省的语文成绩叫做“语文分”。这个时候如果直接去精准匹配会导致很多省份匹配不上，这时候可以使用近似匹配，即根据字符串的相似度进行匹配。但在极端情况下近似匹配会出现科目匹配错误的情况，所以这里使用一个比较粗暴的方式，即经验主义：根据 2020 年的总表制作可以总结出每个科目可能出现的名称，在某一个省中寻找某一个科目时，遍历所有可能出现的情况，在 CJXDM 表中寻找，直到最后找到对应的序号，表五为科目可能会出现的名称。

举个例子，根据图四，我们人眼可以直接看出，语文分其实就是我们要找的语文成绩，但是如果让程序自动生成则不能这么智能，所以程序需要让“语文分” 在表五中查找，看它到底对应的是什么成绩，“语文分”出现在语文成绩这一行中，所以可以将“语文分”对应的成绩项作为语文成绩。

表五：科目可能字段表

字段名	字段名称	可能在 CJXDM 出现的名称
GKCJX01	语文成绩	"语文", "语文分", "语文成绩", "语文（含附加分）"
GKCJX02	数学成绩	"数学", "数学分", "理科数学", "数学理", "数学成绩", "数学（含附加分）"
GKCJX02X	数学成绩（文）	"文科数学", "数学文"
GKCJX03	外语成绩	"外语", "外语分", "英语", "英语成绩", "外语成绩"
GKCJX04	物理成绩	"物理", "物理分", "物理选测"
GKCJX05	化学成绩	"化学", "化学分", "化学选测"
GKCJX06	生物成绩	"生物", "生物分", "生物选测", "生命科学"
GKCJX07	政治成绩	"政治", "政治分", "政治必测", "思想政治"

GKCJX08	历史成绩	"历史","历史分","历史必测"
GKCJX09	地理成绩	"地理","地理分","地理必测"
GKCJX10	技术成绩	"技术"
GKCJX11	统考美术成绩	"统考美术成绩","美术专业分","美术总分","艺体专业成绩","美术统考总分","美术与设计学类","美术本科统考成绩","美术"
GKCJX12	综合成绩	"综合","综合成绩","理科综合","综合/对口专业","文综/理综"
GKCJX12X	综合成绩（文）	"文科综合"
GKCJX13	外语听力	"外语听力"

Ps：两个标黄需要特殊处理，由于有些省份一个批次出结果是包含了文理科的，另外为了节省时间，在程序制作总表数据库时，是直接把某一列直接拷贝到目标区域的，所以对于文理混合的省份来说，如果值开设一列存放数学成绩和综合成绩的话会导致部分数据确实，所以此处根据所有省份的数据特点另外开设了两个文科成绩通道，当然这需要在需要在最后进行手动删除处理。

以上即为录取总表数据库的制作总体逻辑，当知道了所有的映射关系之后，如果不嫌麻烦，可以直接复制粘贴，最后形成总表，但是这些其实是可以编写成一个脚本的。