PISA2018 DATA ANALYSIS

作者姓名：陈鸿绪　学科专业：数据科学与大数据技术 导师姓名：刘淇

（中国科学技术大学，安徽合肥）

摘　要：本文主要通过数据预处理、相关性分析、特征组合与提取等一系列数据分析手段，分析提取并组合 PISA2018 数据集的各种数据特征，在分析相关性与图像直观性的基础上，试图提取出与目标列特征“REPEAT”具有一定相关性的特征，并进行分析。

关键词：数据分析；特征工程；相关性分析；PISA2018;

1. 数据统计与分析

a) 数据整体显示

CSV 文件导入后，非完整数据展示：

	Unnamed: 0	index	CNTRYID	CNT	NatCen	STRATUM	SUBNATIO	OECD	ADMINMODE	LANGTEST_COG	...	EMOSUPP	PQSCHOOL	PASCHPOL
0	0	11956	152.0	CHL	15200	CHL0206	1520000	1.0	2.0	156.0	...	0.739	0.7457	1.6215
1	1	11958	152.0	CHL	15200	CHL0414	1520000	1.0	2.0	156.0	...	0.739	2.0484	1.3268
2	2	11960	152.0	CHL	15200	CHL0308	1520000	1.0	2.0	156.0	...	0.739	-0.7951	-0.6368
3	3	11961	152.0	CHL	15200	CHL0414	1520000	1.0	2.0	156.0	...	0.739	-1.4552	-1.1177
4	4	11965	152.0	CHL	15200	CHL0308	1520000	1.0	2.0	156.0	...	0.739	0.3731	0.5620

5 rows × 487 columns

b) 数据缺失值情况

```
Part of missing values for every column
Unnamed: 0     0.000000
index          0.000000
CNTRYID        0.000000
CNT            0.000000
NatCen         0.000000
                 ...
ATTIMMP        0.643233
INTCULTP       0.644679
GCAWAREP       0.636926
BODYIMA        0.256829
SOCONPA        0.229040
Length: 487, dtype: float64
```

可以发现，对于不同列的数据，缺失情况是由非常大的区别的。

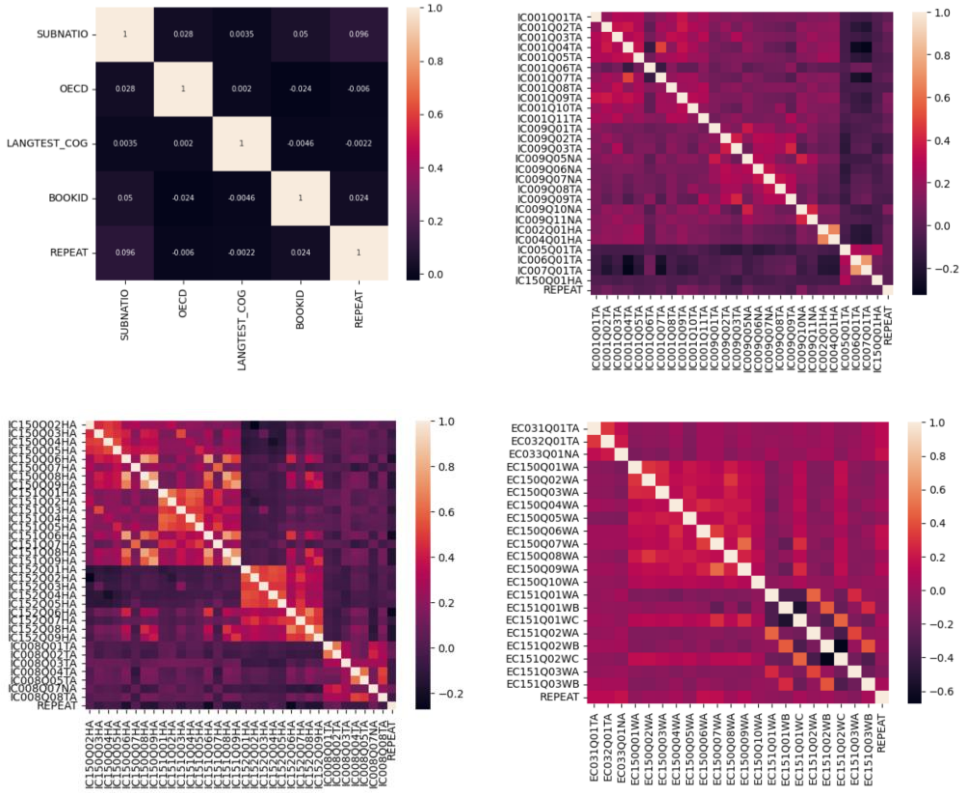
c) 数据基本统计量描述

	Unnamed: 0	index	CNTRYID	NatCen	SUBNATIO	OECD	ADMINMODE	LANGTEST_COG	LANGTEST_PAQ	BOOKID	...
count	42176.000000	42176.000000	42176.000000	42176.000000	4.217600e+04	42176.000000	42176.0	42176.000000	15612.000000	42176.000000	...
mean	21087.500000	59311.667536	598.035352	59803.535186	5.980354e+06	0.888230	2.0	156.007943	156.010056	22.389321	...
std	12175.306813	22074.043971	196.749989	19674.998860	1.967500e+06	0.315087	0.0	1.437761	1.256523	17.028105	...
min	0.000000	11956.000000	152.000000	15200.000000	1.520000e+06	0.000000	2.0	156.000000	156.000000	1.000000	...
25%	10543.750000	45627.750000	484.000000	48400.000000	4.840000e+06	1.000000	2.0	156.000000	156.000000	9.000000	...
50%	21087.500000	64732.500000	724.000000	72400.000000	7.240000e+06	1.000000	2.0	156.000000	156.000000	18.000000	...
75%	31631.250000	77064.250000	724.000000	72400.000000	7.240000e+06	1.000000	2.0	156.000000	156.000000	37.000000	...
max	42175.000000	89406.000000	724.000000	72400.000000	7.240000e+06	1.000000	2.0	451.000000	313.000000	72.000000	...

8 rows × 485 columns

由此可以知道数据行数为 42176 行，也知道了特征数据的基本分布情况。

d) 对几乎所有特征进行了与“REPEAT”特征的相关性分析后，得到如下热图(HOTMAP)
(注：并没有完整展示所有图片)



这些图可以看出各种特征与“REPEAT”目标列的相关系数大小，从而得到选取单个特征值的初步判断。

2. 特征提取与变换

a) 特征提取出的单个列属性名与其相应描述

只考虑相关系数大小，不考虑特征含义的情况下进行特征提取得到的特征

Field	description
ST127Q01TA	Have you ever repeated a ? At <ISCED 1>
ST127Q02TA	Have you ever repeated a ? At <ISCED 2>
ST127Q03TA	Have you ever repeated a ? At <ISCED 3>
ST001D01T	Student International Grade (Derived)
EC031Q01TA	Did you change schools when you were attending <ISCED 1>?
EC032Q01TA	Did you change schools when you were attending <ISCED 2>?
EC033Q01NA	Have you ever changed your study programme?
EC150Q05WA	Find out about future study or types of work: I spoke to a outside of my school.
EC150Q06WA	Find out about future study or types of work: I completed a questionnaire to find out about my interests and abilities.
EC150Q07WA	Find out about future study or types of work: I researched the Internet for information about careers.
EC150Q08WA	Find out about future study or types of work: I went to an organised tour in an <ISCED 3-5> institution.
OCOD1	ISCO-08 Occupation code - Mother
OCOD2	ISCO-08 Occupation code - Father
OCOD3	ISCO-08 Occupation code - Self
GRADE	Grade compared to modal grade in country

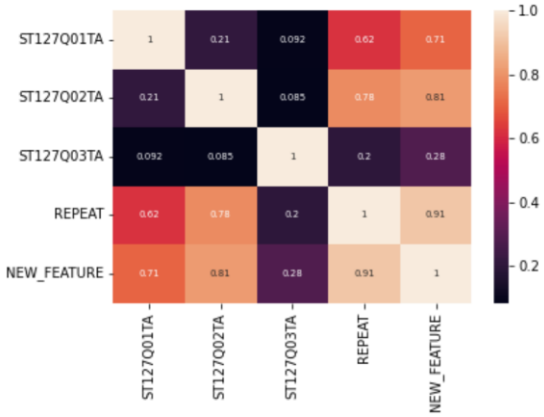
PROGN	Unique national study programme code
COBN_S	Country of Birth National Categories- Self
COBN_M	Country of Birth National Categories- Mother
COBN_F	Country of Birth National Categories- Father

不考虑相关系数大小，只考虑特征含义的情况下进行特征提取得到的特征 CNTRYID ,ST005Q01TA,ST007Q01TA。

b) 特征组合与特征构造（注：在以下所有操作之前都进行过数据预处理）

（1）ST127Q01TA,ST127Q02TA,ST127Q03TA 分析

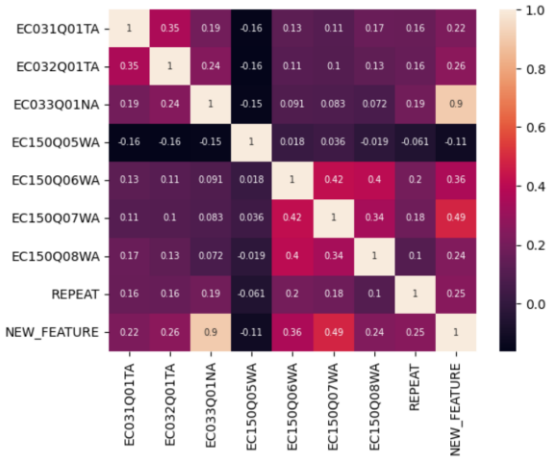
由该特征的描述可知，这三个特征与“REPEAT”目标列必然存在绝对联系。特征构造如下：NEW_FEATURE 取值为三个目标属性值之和。将该构造的特征值加入热图中，显示得到：



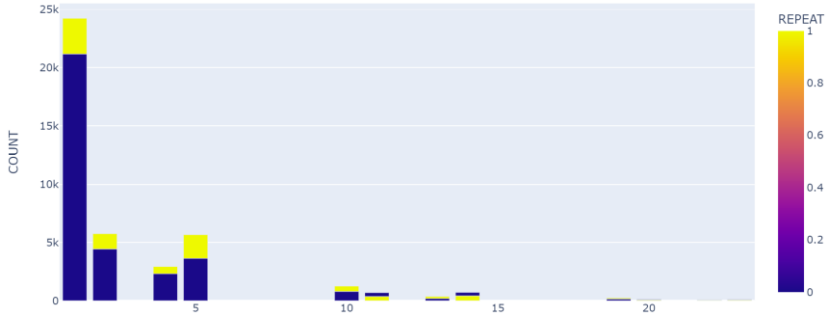
注意到，NEW_FEATURE 与 REPEAT 目标列的相关系数为 0.91，也就是说，构造的特征值在原有的数据集中几乎可以大概率预测 REPEAT 的状态。说明 ST127Q01TA，ST127Q02TA，ST127Q03TA 这三个与 REPEAT 强相关的特征可以构造出与 REPEAT 更加强相关的特征。

（2） EC031Q01TA, EC032Q01TA,EC033Q01NA,EC150Q05WA,EC150Q06WA, EC150Q07WA,EC150Q08WA 分析

特征构造如下：将其中与 REPEAT 目标列相关系数较大的三个特征提取出来，具体是 EC033Q01NA，EC150Q07WA，EC150Q06WA 这三列特征，将它们作笛卡尔积并一一映射到 1—27 整数。该特征值为 NEW_FEATURE。将该构造的特征值加入热图中，显示得到：



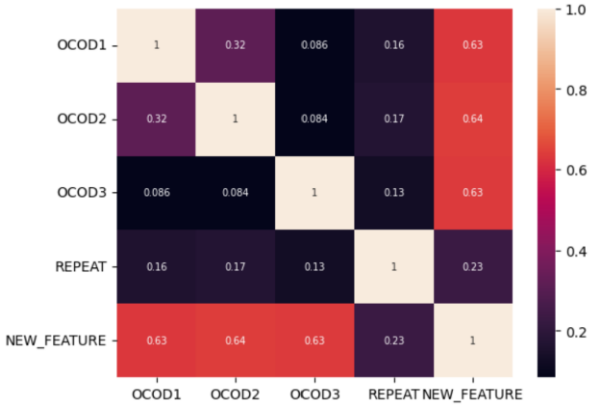
可得到新的特征值与 REPEAT 的相关系数为 0.25，此为弱相关。为了让结果直观，可以将相同特征值对应的 REPEAT 大致占比显示出来：



这张图可以直观的表现出新的特征值与 REPEAT 之间的弱关系。

（3）OCOD1,OCOD2,OCOD3 的分析

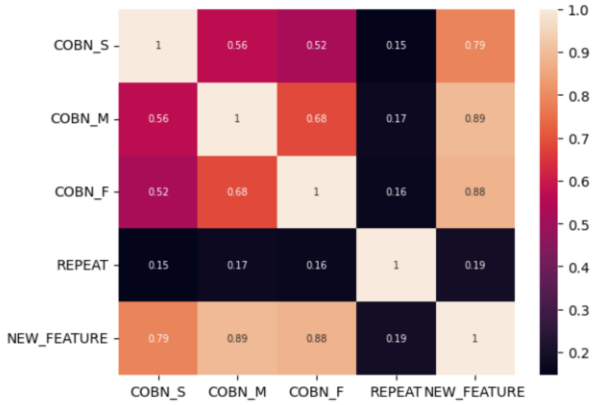
对该三个特征值组合而成的新特征值构造如下：NEW_FEATURE 的值等于该三个特征值乘积后取三分之一次方，即取三个数的几何平均。将该构造的特征值加入热图中，显示得到：



从这张热图可以观察出原本 OCOD1,OCOD2,OCOD3 与目标列的相关系数分别为 0.16，0.17，0.13，但在做了该特征变换后得到的新特征值与目标列的相关系数变成 0.23，具有明显提升，该构造的特征值可以认为与“REPEAT”成弱相关。

（4）COBN_S,COBN_M,COBN_F 的分析

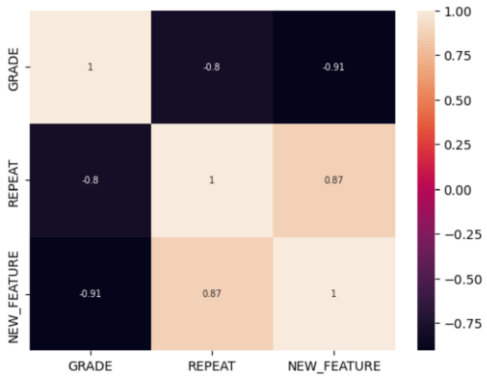
对该三个特征值组合而成的新特征值构造如下：NEW_FEATURE 等于该三个特征值之和。将该构造的特征值加入热图中，显示得到：



在这张热图中原本 COBN_S,COBN_M,COBN_F 与目标列的相关系数分别为 0.15，0.17，0.16 然而在组合后的特征值与“REPEAT”的相关系数为 0.19，具有一定提升，但该相关性仍然非常弱。

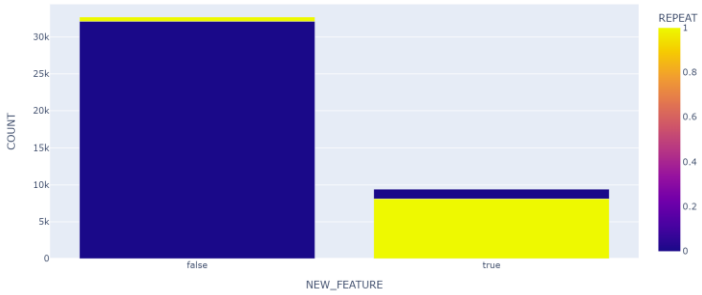
（5）GRADE 的分析

GRADE 与 REPEAT 原热图中即可以看出两者具有很强的相关性，具体可以得到两者相关性为 0.8，对 GEADE 做如下特征变换使得两者相关性更加强烈：将所有 GRADE 不为 0 的数全部映射到 1，0 全部映射到 0，变换得到一个新的特征值 NEW_FEATURE。将该构造的特征值加入热图中，显示得到：



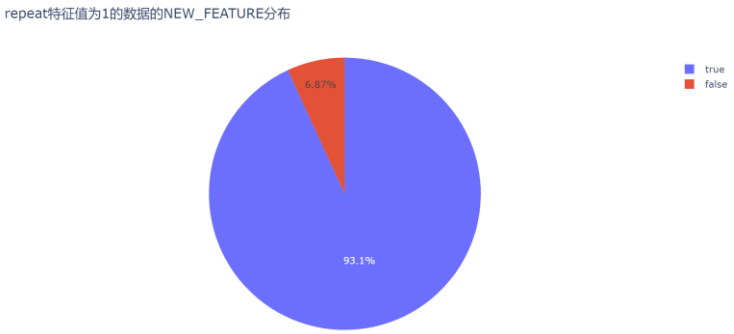
观察得到新特征值与目标列的相关系数为 0.87，大于原特征值 0.8 的相关系数，由“GRADE”变换得到的特征值与原目标列具有强相关性。

为了显示出这种强相关性，可以画出对于不同的 NEW_FEATURE 值，REPEAT 所占比例的大小情况。如下图所示：

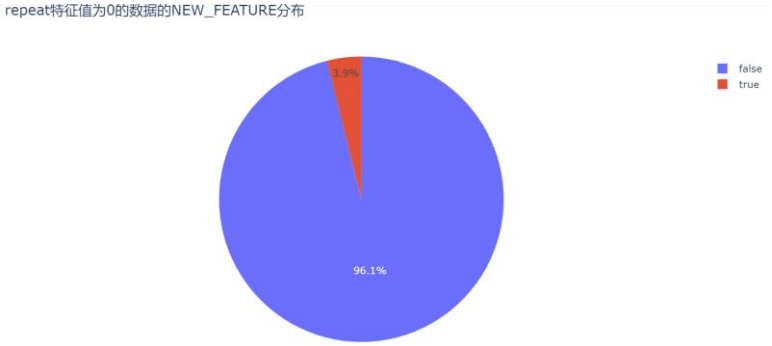


这张图直观显示出对于 NEW_FEATURE=0 时，REPEAT 几乎全部取值为 0，对于 NEW_FEATURE=1 时，REPEAT 几乎全部取值为 1。这显然体现了两者强相关性。

再考察 REPEAT=1 时，NEW_FEATURE 的分布比例：



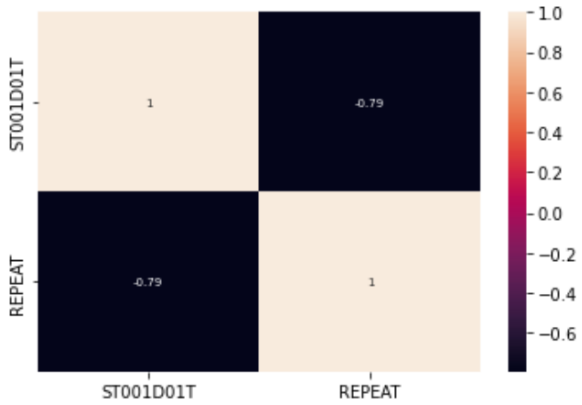
再考察 REPEAT=0 时，NEW_FEATURE 的分布比例：



对于同一种情况，NEW_FEATURE 取值相同比例已经达到 90%。两张比例图亦表现了两者强相关性。

（6）ST001D01T 的分析

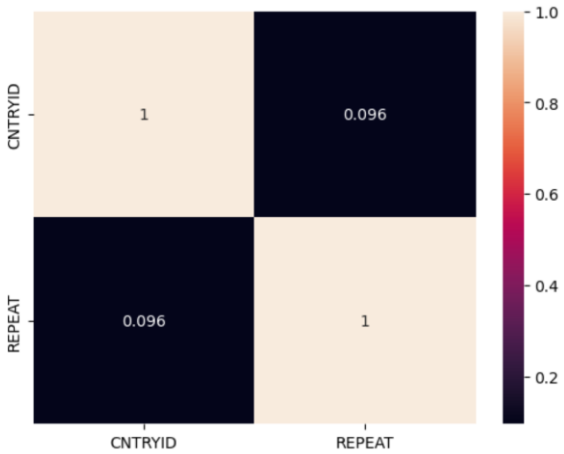
数据预处理采用平均值填充的方法，在原热图中，ST001D01T 与 REPEART 相关性即为强相关，具体值为-0.79。这里对该属性就不采用属性变换，得到与 REPAEAT 相关性热图如下：



c) 由列属性的描述，可以初步判断一些属性与目标列的相关性

(1) CNTRYID 的分析

热图：



由两者相关性大小可以发现两者几乎不存在相关性

（2）ST005Q01TA,ST007Q01TA 的分析

由于父母亲的教育程度会一定程度上影响到孩子教育，所以做出假设，这两个特征与 REPEAT 具有一定相关性，对两个特征平方平均得到构造的新特征加入热图有：

两者具有 0.21 的相关性，可以认为两者弱相关。

3. 总结：

a). 可以发现 ST127Q01TA,ST127Q02TA,ST127Q03TA,ST001D01T,GRADE 是与目标列具有强相关性。完全可以作为预测 REPEAT 列的依据。

b). 对 EC031Q01TA, EC032Q01TA,EC033Q01NA,EC150Q05WA,EC150Q06WA, EC150Q07WA,EC150Q08WA 等若干属性原本与 REPEAT 相关性，也只有百分之十几，通过函数构造出新的属性，提高了相关性若干百分点，达到了构造的新属性与 REPEAT 弱相关的程度。