



数据分析及实践

Analysis and Practice of the Data

实验课 (三)

刘 淇

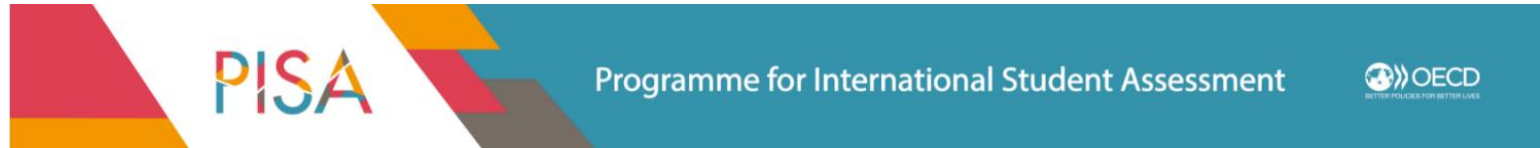
Email: qiliuql@ustc.edu.cn

课程主页:

<http://staff.ustc.edu.cn/~qiliuql/AD2022.html>

PISA介绍

2



□ 国际学生评估项目(PISA)

- 对世界各地 **15 岁** 学生进行的三年一次调查，评估他们获得了充分参与社会和经济生活所必需的**关键知识和技能**的程度。
- **覆盖范围广**，被称为全球教育的“奥林匹克盛会”
- 问卷评估+认知项目测验

600 000 students

representing about **32 million** 15-year-olds
in the schools of the **79 participating
countries and economies** sat the **2-hour**
PISA test in 2018



PISA介绍



oecd.org/pisa/



FOLLOW US



Programme for International Student Assessment



Home

About

PISA
Test

Innovation

Data

Publications

Webinars

Join
PISA

FAQ

[Français](#) [Deutsch](#)

What is PISA?

PISA is the OECD's Programme for International Student Assessment. PISA measures 15-year-olds' ability to use their reading, mathematics and science knowledge and skills to meet real-life challenges.

<https://www.oecd.org/pisa/>

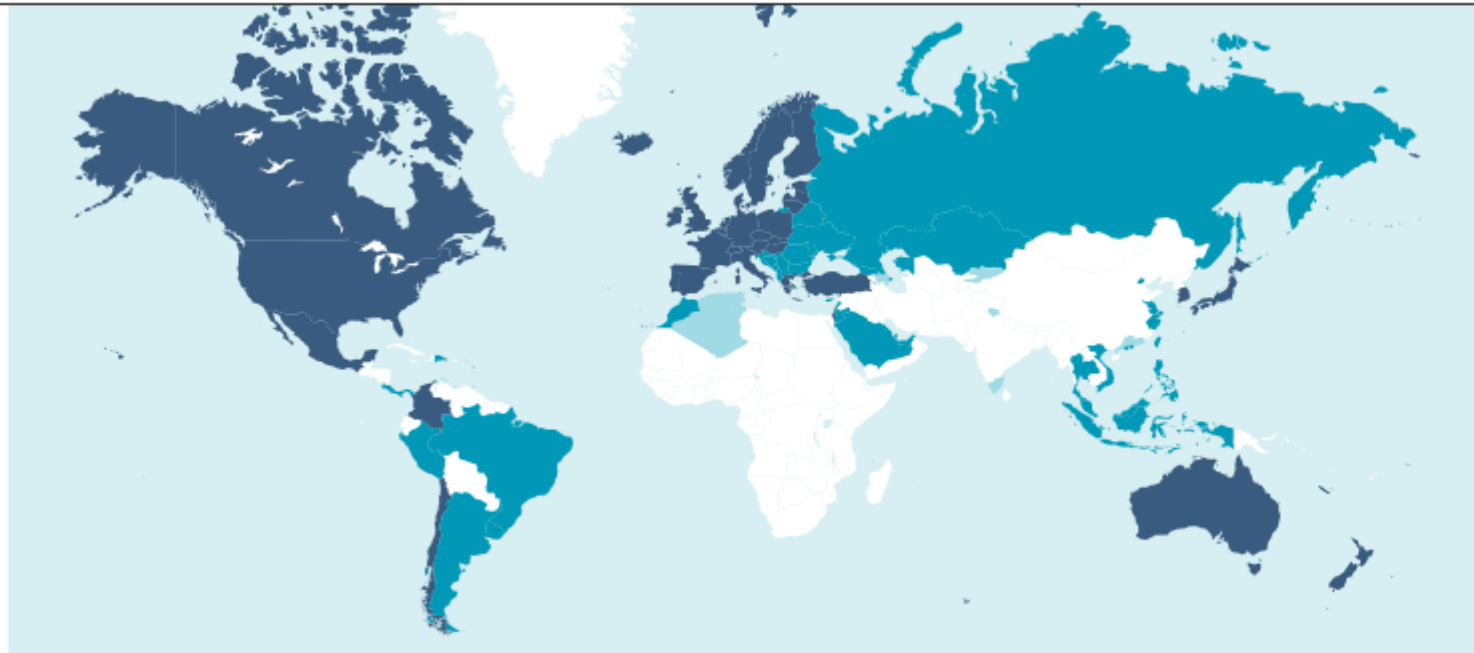
PISA介绍

4

□ 具体覆盖范围

■ 37 世界经合（OECD）组织国家+ 42 伙伴国家与经济体

Map of PISA countries and economies



OECD member countries

Partner countries and economies in PISA 2018

Partner countries and economies in previous cycles

PISA介绍

□ OECD 学生能力国际评价项目

- 在国家层面上对学生总体做出推断，而不是评估学生个体的知识掌握情况
- 认知项目测验考察领域：科学、阅读、数学和金融素养
- 调查问卷：学生（含家长）+ 教师 + 学校



例如，考察阅读

以下哪个观点归纳了芝诺的学说？

- A. 我们应该体谅他人，如此我们能过快乐的生活。
- B. 我们应该在意我们的外貌。
- C. 我们不该让欲望控制我们。
- D. 我们不该试图改变过去。

PISA介绍

□ OECD 学生能力国际评价项目

- 在国家层面上对学生总体做出推断，而不是评估学生个体的知识掌握情况
- 认知项目测验考察领域：科学、阅读、数学和金融素养
- 调查问卷：学生（含家长）+ 教师 + 学校

PISA 2015 单元名称：坛子

任务1

你需要设计出一个最好的能存储食物的坛子。

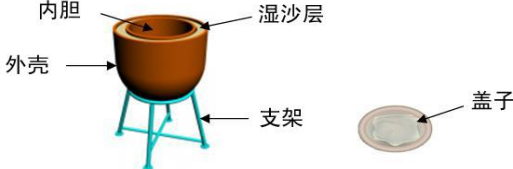
食物最好保存在一个4℃且无空气无菌的环境下。

使用模拟设备，通过改变粘土的厚度和沙层的水分情况，设计出能存储最大量食物且食物保持新鲜（温度为4℃）的坛子

你能够做仿真实验并重复或移动任何数据。

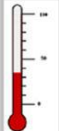
4℃环境下食物最大存储量


千克



沙层厚度 (cm)	食物数量 (kg)	沙子湿度 (潮湿/干燥)	温 度

常数变量





 空气温度 30℃


 湿度 20%

沙层厚度 (cm) 1 2 3 4 5

食物数量 (kg) 0 4 8 12 16 20

沙子湿度 潮湿 干燥

保存数据
清除数据

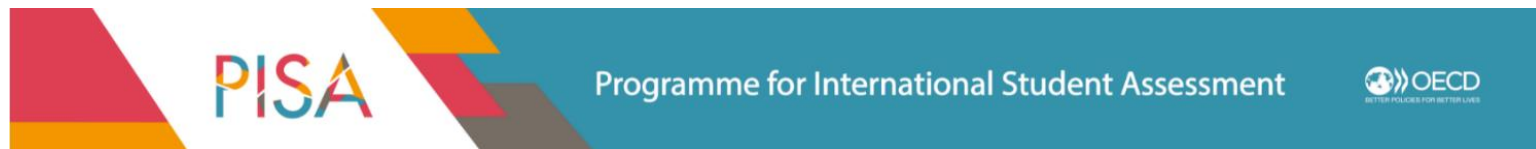
例如，考察科学

□ 教育专家设计的问卷评估框架

	STQ	SCQ	TGQ	TTQ	PAQ	ICQ	ECQ
Policy area: Reading education							
Teacher qualifications and professional development		X	X	X			
Teaching practices for reading	X		X	X		X	
Out-of-school reading experiences	X					X	
Policy area: Equity							
Student SES, family and home background	X				X	X	
Migration and culture	X	X					
Educational pathways in early childhood	X				X		X
Policy area: Broader Educational outcomes beyond achievement							
Dispositional & school-focused variables	X					X	X
Dispositions for global competence	X	X	X	X	X		X
Policy area: Supportive school context							
Learning time and curriculum	X	X	X	X			X
School context and resources		X	X	X		X	

PISA的优势

8



□ 政策导向

- 分析学生表现与背景、学习态度、校内外关键因素
- 确定表现良好的教育系统的特征

□ 创新概念——“素养”

- 学生在关键领域应用知识和技能的能力
- 识别、解释和解决问题时进行有效的分析、推理和沟通的能力

□ 终身学习相关

- 询问、记录学生学习动机、自身信念、学习策略

□ 规律性

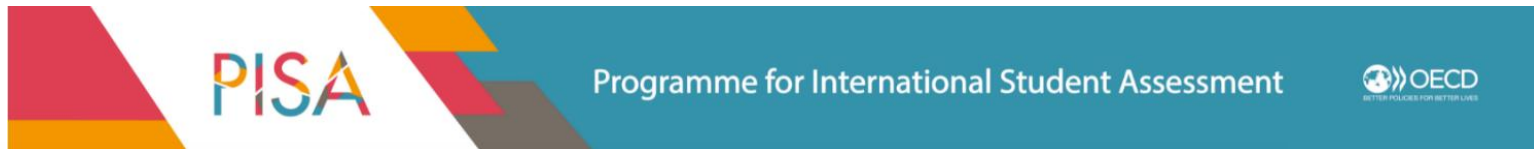
- 三年一次：帮助各国检测实现关键学习目标的进展

□ 覆盖范围

- 37 世界经合（OECD）组织国家+ 42 伙伴国家与经济体
- 11 伙伴国家与经济体

PISA的优势

9



PISA2018测试结果正式发布

2019-12-04 来源：教育部

近日，经济合作与发展组织（OECD）公布了2018年国际学生评估项目（PISA 2018）测试结果。在参与测试的85个国家和地区中，我国四省市（北京、上海、江苏、浙江）在阅读、数学、科学三大基础学科中均位居第一。

最新PISA结果出炉！中国包揽三项第一



观察者网

发布时间：2019-12-04 08:13 | 上海观察者信息技术有限公司官方帐号

关注

文汇报12月3日消息，巴黎时间当天上午9点，世界经合组织（OECD）发表了第七轮国际学生评估结果（PISA 2018）：北京、上海、江苏省和浙江省组成的中国部分地区联合体分别在阅读、数学和科学素养三项测试均位居第一。

在共有来自79个国家和地区约60万学生参与的测试中，由北京、上海、江苏、浙江组成的中国部分地区联合体在阅读（555）、数学（591）和科学（590）三项测试中遥遥领先，新加坡紧随其后。OECD国家在上述三项上的平均成绩分别为487分、489分和489分。

从2009年、2012年连续两年排名第一，到2015年的排名第十，再到如今的回归第一，上海师范大学国际与比较教育研究院院长张民选认为，阅读、数学、科学三大基础学科都取得了第一，值得庆贺。但，三个基础学科不是基础教育的全部。

PISA2015测试结果：新加坡第一 中国排名惨烈

2016年12月07日 16:03 新浪教育 微博

欧洲中部时间12月6日10点45分，经合组织（OECD）公布了2015PISA测试的结果——在最新出炉的PISA报告中，新加坡学生力压群雄，以数学564、阅读535、科学556的成绩获得第一；发挥较好的OECD国家还包括日本、爱沙尼亚、芬兰和加拿大；而北京、上海、江苏、广东组成的中国部分地区联合体（B-S-J-G, China）在此次测试中仅位居总分第十。



实验要求

□ 实验要求

- 给定一个数据集和预测目标，需要分析数据、统计以及抽取特征
- 数据分析、统计，如：
 - 单个特征的分布
 - 统计缺失值
 - 特征间的相关性
 - 推测特征的含义
 - 异常样本
 - 数据抽样...
- 特征抽取，如：
 - 特征的变换，如：str 转 int, 取log
 - 尝试组合特征
 - 特征子集选择
 - ...

□ 数据集-PISA2015（筛减版）

□ 本次试验针对PICA2015中的学生调查问卷数据集

□ 助教已经做了筛选，现包含西语地区的32130个学生、429个特征

	CNTSCHID	Region	STRATUM	SUBNATIO	OECD	ADMINMODE	Option_CPS	Option_FL	Option_ICTQ	Option_ECQ	...	ST078Q04NA	ST078Q05NA	ST078Q06NA	ST078Q07NA	ST078Q08NA	ST078Q09NA	ST07
0	97100001	72409	ESP9017	7240900	2	2	1	1	1	1	...	2	1	2	1	1	1	1
1	97100001	72409	ESP9017	7240900	2	2	1	1	1	1	...	1	1	1	1	1	1	1
2	97100001	72409	ESP9017	7240900	2	2	1	1	1	1	...	2	1	1	2	2	2	2
3	97100001	72409	ESP9017	7240900	2	2	1	1	1	1	...	2	1	2	1	1	1	2
4	97100001	72409	ESP9017	7240900	2	2	1	1	1	1	...	2	1	2	1	1	1	1
...
32125	97100980	72416	ESP1634	7241600	2	2	1	1	1	1	...	2	1	1	1	1	1	2
32126	97100980	72416	ESP1634	7241600	2	2	1	1	1	1	...	2	2	2	2	1	1	1
32127	97100980	72416	ESP1634	7241600	2	2	1	1	1	1	...	2	1	2	1	1	1	1
32128	97100980	72416	ESP1634	7241600	2	2	1	1	1	1	...	1	1	1	1	1	1	1
32129	97100980	72416	ESP1634	7241600	2	2	1	1	1	1	...	2	1	1	1	1	1	1

□ CNTSCHID代表学校id， Region代表学生所在地区

□ ST开头的是学生问卷回答特征

□ 每个特征的具体含义可以参考codebook

□ Codebook是数据集每个特征的详细说明

□ 预测任务

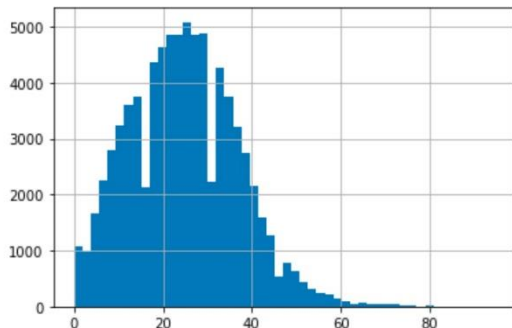
□ 预测学生是否复读，即**REPEAT**列。

实验要求

注意事项

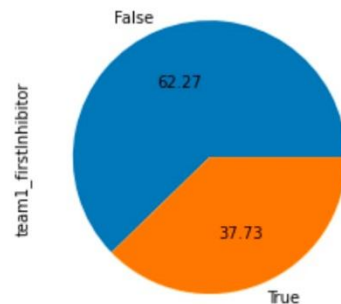
- 该实验需要掌握：Jupyter、numpy、pandas库、matplotlib库
- 每位同学将代码和图表保存在 Jupyter notebook 中

```
df['team1_kills']=df['player1_kills']+df['player2_kills']+df['player3_kills']  
df['team2_kills']=df['player6_kills']+df['player7_kills']+df['player8_kills']  
df['team1_kills'].hist(bins=50)  
plt.show()  
df['team2_kills'].hist(bins=50)  
plt.show()
```



```
t1fi=df['team1_firstInhibitor'].value_counts('TRUE')  
t1fi.plot.pie(autopct='%0.2f')
```

<AxesSubplot:ylabel='team1_firstInhibitor'>





实验要求

注意事项

- 该实验需要掌握：Jupyter、numpy、pandas库、matplotlib库
- 每位同学将代码和图表保存在 Jupyter notebook 中
- 实验报告中记录数据分析结论和提取的特征，注意不要将代码放在实验报告中
- 数据集会发布于QQ群中
- 注意：该实验不需要把所有的特征、数据条目都分析，可以结合自己的判断和精力进行特征子集（若干特征）和数据子集（抽样）的分析

实验三



□ 提交要求

- 将 jupyter 文件和实验报告打包成一个压缩文件，发送给助教：**18251859960@163.com**
- 邮件标题: 姓名_学号_exp3
压缩文件命名格式: 姓名_学号_exp3.zip (rar)
- 截止日期: **2022.4.20**

□ 评分标准:

- 格式是否规范
- 数据分析、特征提取是否详尽
- 提交是否及时

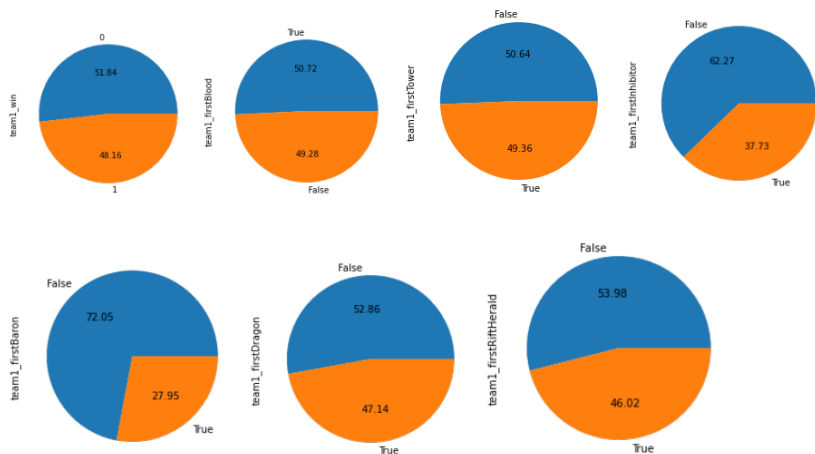
实验三---参考

Part I 数据分析

一、首先分析这个数据集单个特征的分布。

(1) 原有团队数据的分布情况

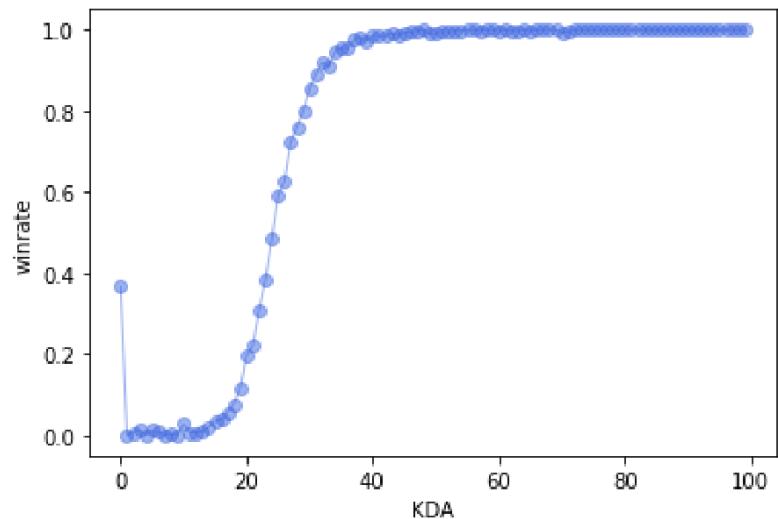
原有团队数据在 team1、team2 间的分布情况如饼图所示，包括 win, firstBlood, firstTower, firstInhibitor, firstBaron, firstDragon, firstRiftHerald。



可以看出数据基本分布均匀，但是 firstInhibitor 和 firstBaron 存在明显的分布不均匀问题，这是因为数据集中只记录了 team1 的获取情况，可能存在两个队伍没有击杀 Baron 的情况，无法在饼图中体现出来。

(5) 团队总 KDA 与 team1_win 间的相关性

将 KDA*10 离散化得到以下结果



得到 KDA 与 win1_win 呈正相关的结论，且当 KDA 高到一定程度时，胜率趋近于 1

实验三-参考资料

□ 参考资料：

- PISA官方网站：<https://www.oecd.org/pisa/>
- kaggle、天池等数据科学网站的初学者教程，如：
<https://www.kaggle.com/startupsci/titanic-data-science-solutions>
- 《利用Python进行数据分析-第2版》

