

PISA 数据集数据可视化-最终报告

组员：何果财、秦晓东

1 简介

数据可视化主要旨在借助于图形化手段，清晰有效地传达与沟通信息，是数据挖掘领域的重要研究方向，是一个极为活跃而又关键的方面。为了有效地传达思想概念，美学形式与功能需要齐头并进，通过直观地传达关键的方面与特征，从而实现对于相当稀疏而又复杂的数据集的深入洞察。数据可视化与信息图形、信息可视化、科学可视化以及统计图形密切相关。一直以来，数据可视化就是一个处于不断演变之中的概念，其边界在不断地扩大；因而，最好是对其加以宽泛的定义。数据可视化指的是利用图形、图像处理、计算机视觉以及用户界面，通过表达、建模以及对立体、表面、属性以及动画的显示，对数据加以可视化解释。

本项目主要想解决的问题：通过有效的视觉可视化手段对复杂结构数据中蕴含的各种潜在知识进行表达，数据可视化是一个数据分析和可视化迭代的过程。如图 1 所示，数据准备和数据分析是可视化的必备条件，而可视化又能反过来影响数据分析，改进数据分析的目标。

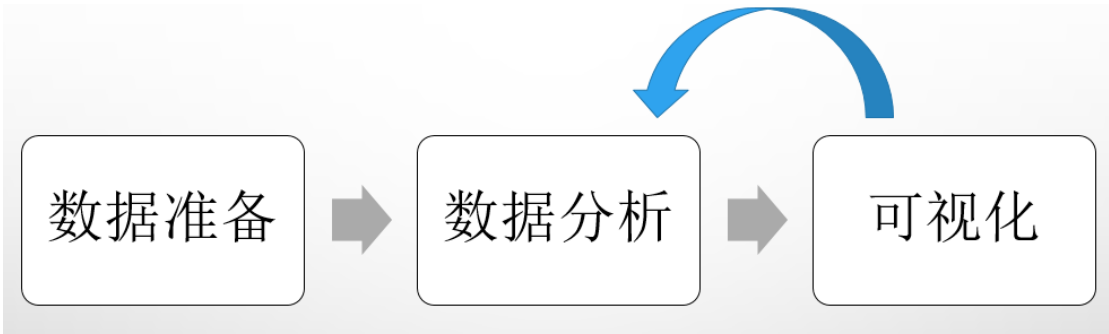


图 1：数据可视化流程和意义

2 问题描述

2.1 背景

PISA 是 OECD 经济合作与发展组织的项目，主要对接近完成基础教育的 15 岁学生进行评估，测试学生是否掌握参与社会所需的知识与技能。2015 年的 PISA，共有代表 72 个国家和地区 15 岁的五十万学生参加，此次项目在科学，数学，阅读，协作解决问题和金融知识进行了评估。PISA 项目有大量的数据产生，OECD 也为此举办了数据可视化竞赛。

本项目的目标：

- (1) 对数据集提出一些问题。
- (2) 使用数据分析方法，挖掘 PISA 数据中蕴藏的知识。
- (3) 进而使用数据可视化工具，将挖掘出的内容以视觉方式展现。

2.2 数据集

PISA 数据集是 OECD 在全球范围内举办的青少年知识评估项目产生的，它主要包含一下几个方面的数据：

- (1) 学生的问卷
- (2) 学校的调查问卷
- (3) 家庭的调查问卷

其中，数据的维度相当大，即字段非常地多，如学生的问卷数据就多达 636 个。好在从数据源获得的数据已经是结构化的，这简化了于我们的数据清洗工作。数据集共包含 485490 学生的考察数据，主要包括学生成绩、家庭和学校情况三方面的数据。如图 2 所示，数据量比较大，但是下载下来的数据已经结构化处理，方便使用 R 语言进行数据分析和建模。






















 computerItem2012.rda	2017/2/23 5:39
 computerItem2012dict.rda	2017/2/23 5:39
 computerParent2012.rda	2017/2/23 5:39
 computerParent2012dict.rda	2017/2/23 5:39
 computerSchool2012.rda	2017/2/23 5:39
 computerSchool2012dict.rda	2017/2/23 5:39
 computerScoredItem2012.rda	2017/2/23 5:39
 computerScoredItem2012dict.rda	2017/2/23 5:39
 computerStudent2012.rda	2017/2/23 5:39
 computerStudent2012dict.rda	2017/2/23 5:39
 item2012.rda	2017/2/23 5:39
 item2012dict.rda	2017/2/23 5:39
 parent2012.rda	2017/2/23 5:39
 parent2012dict.rda	2017/2/23 5:39
 school2012.rda	2017/2/23 5:39
 school2012dict.rda	2017/2/23 5:39
 scoredItem2012.rda	2017/2/23 5:39
 scoredItem2012dict.rda	2017/2/23 5:39
 student2012.rda	2017/2/23 5:39
 student2012dict.rda	2017/2/23 5:39
 student2012weights.rda	2017/2/23 5:39

图 2：整个 PISA 项目的数据集，dict 是字段的解释文件

2.3 预期成果

我们将从成绩与学校因素的关系和成绩与其它因素两个方面进行讨论。其中，与学校相关的因素有：教师水平、计算机多媒体设备、图书馆、学生入学时间、授课方式等等。其它因素主要包括性别、国家地区、学科、书籍等等。

3 技术方案

首先我们要对数据集提出一些问题，问题应该在数据集中有所体现。针对提出的问题，我们对数据进行分析，然后绘制可视化的图表，视觉化地展现对问题的解答。

4 实现与实验结果

下面，我们将按照提出问题，分析数据，可视化展示来进行实验，并进行实验结果的展示。

4.1 学校相关因素对成绩影响

1. 学校和家庭的支持对学生成绩的影响？

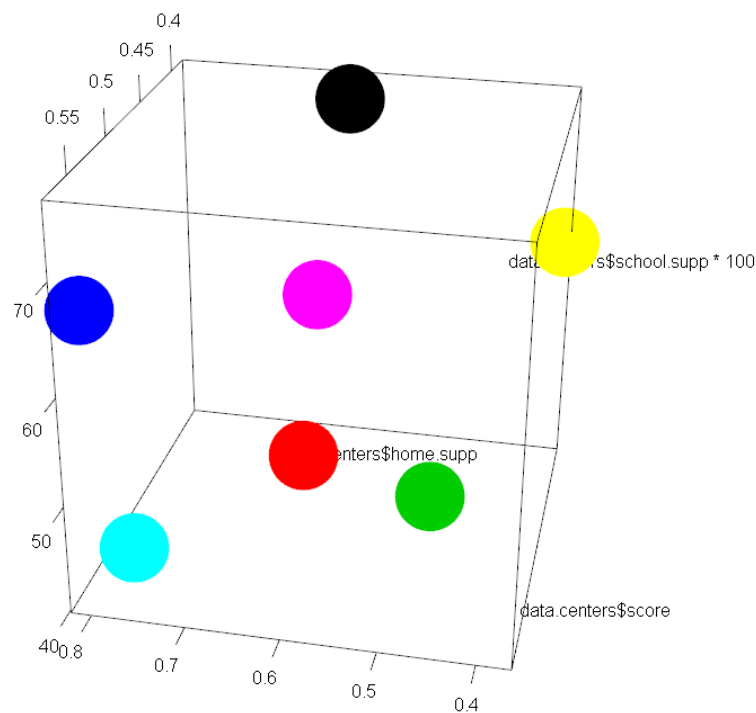


图 3：学生成绩与家庭和学校支持的关系

2. 逃学率与数学成绩的关系

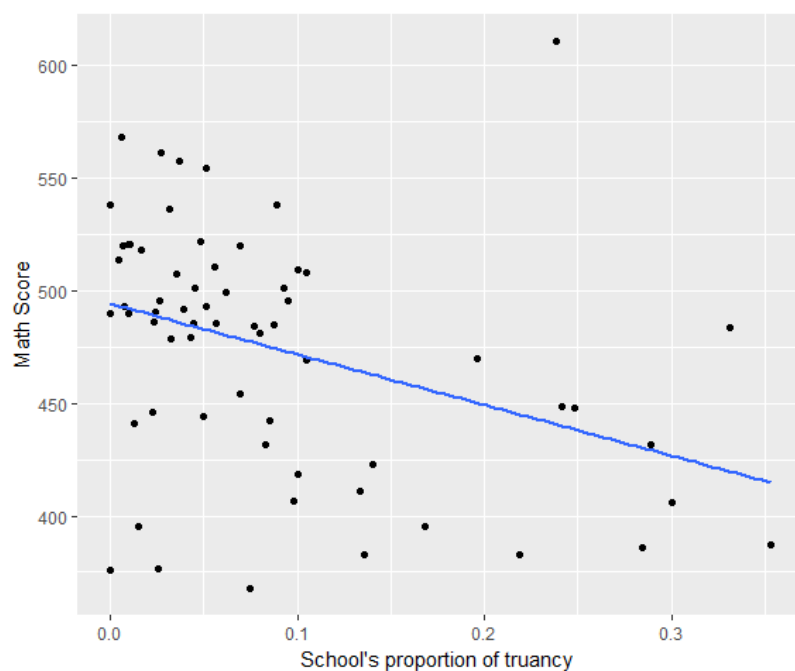


图 4：学校逃学率与平均数学成绩的关系

根据图 4 可以看到，逃学率与数学成绩基本呈负相关，当然也有部分离群点（中国），逃学率高反而数学成绩很好。

3. 数学老师的缺乏对数学成绩的影响？

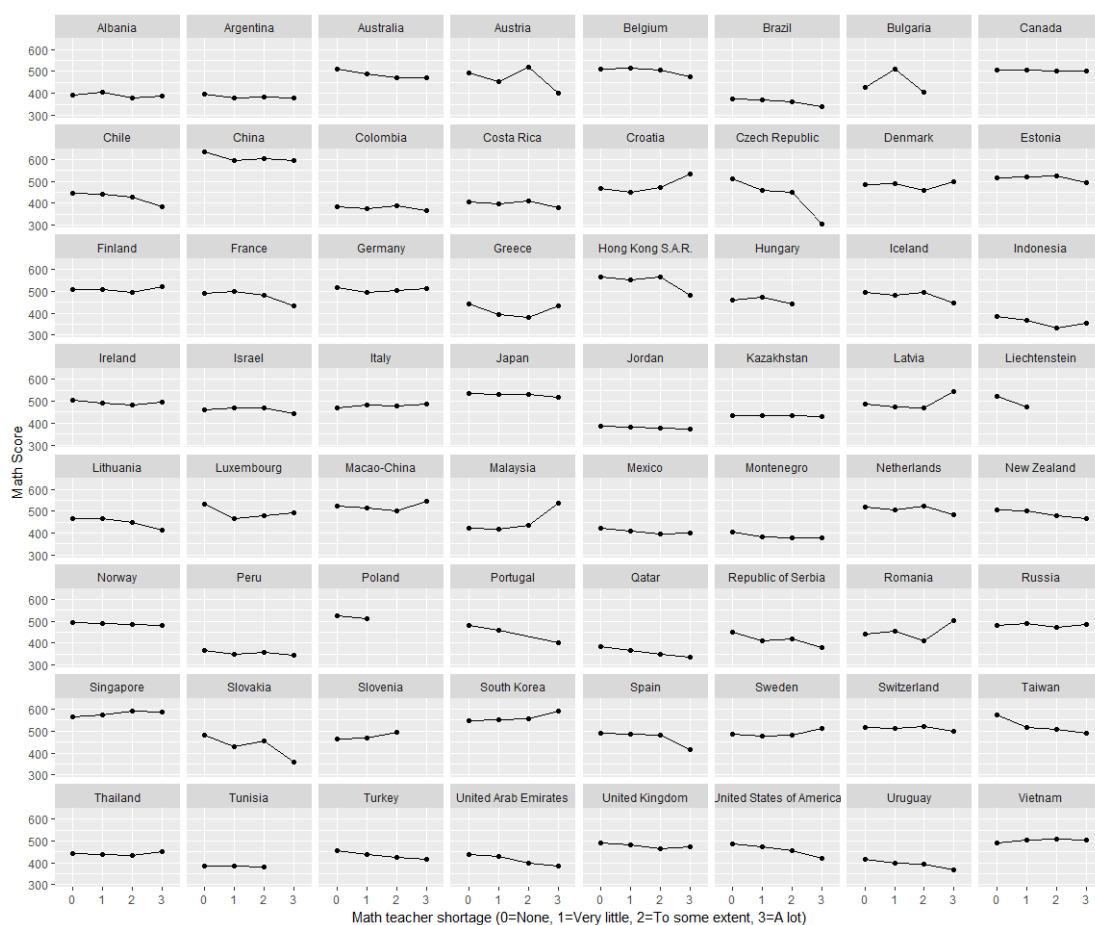


图 5：缺乏数学老师对学生成绩的影响

根据图 5 可以看到，对于大多数国家，数学老师的缺乏会降低学生的数学成绩，符合我们的常识。

4. 学生入学时间对数学、阅读、科技学科的成绩影响？

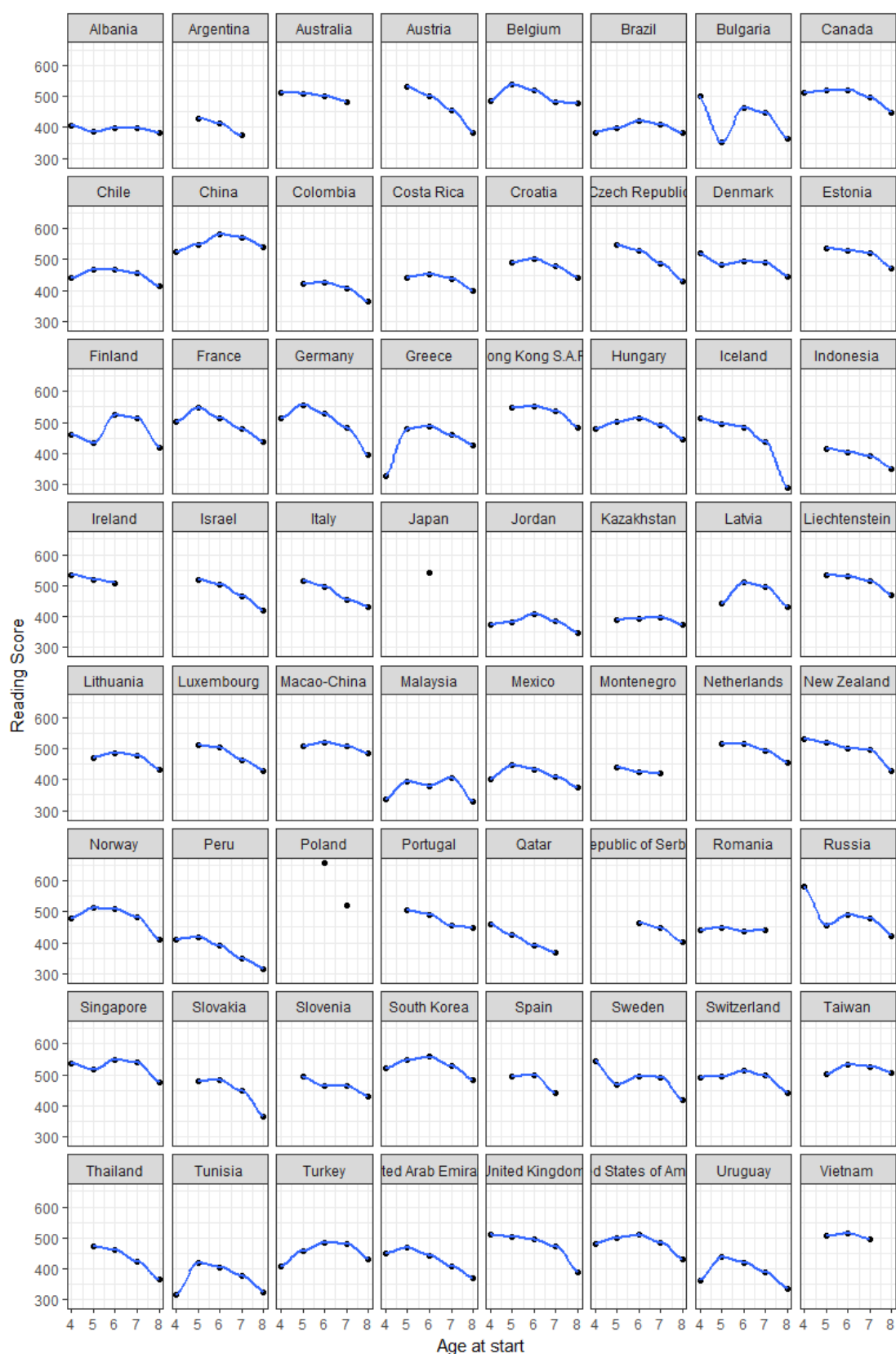


图 6：学生阅读成绩与上学年纪的关系

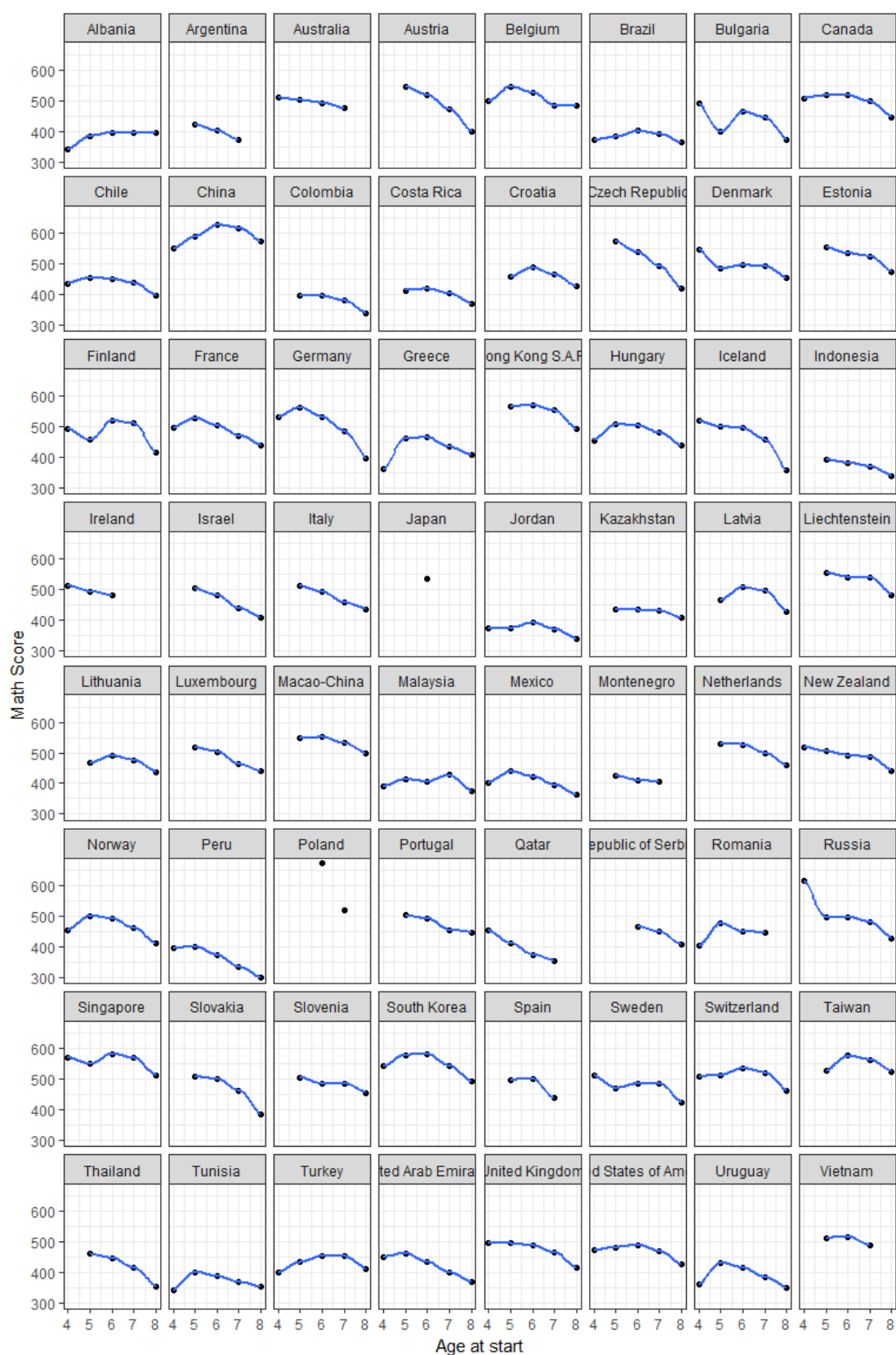


图 7：学生数学成绩与上学年纪的关系

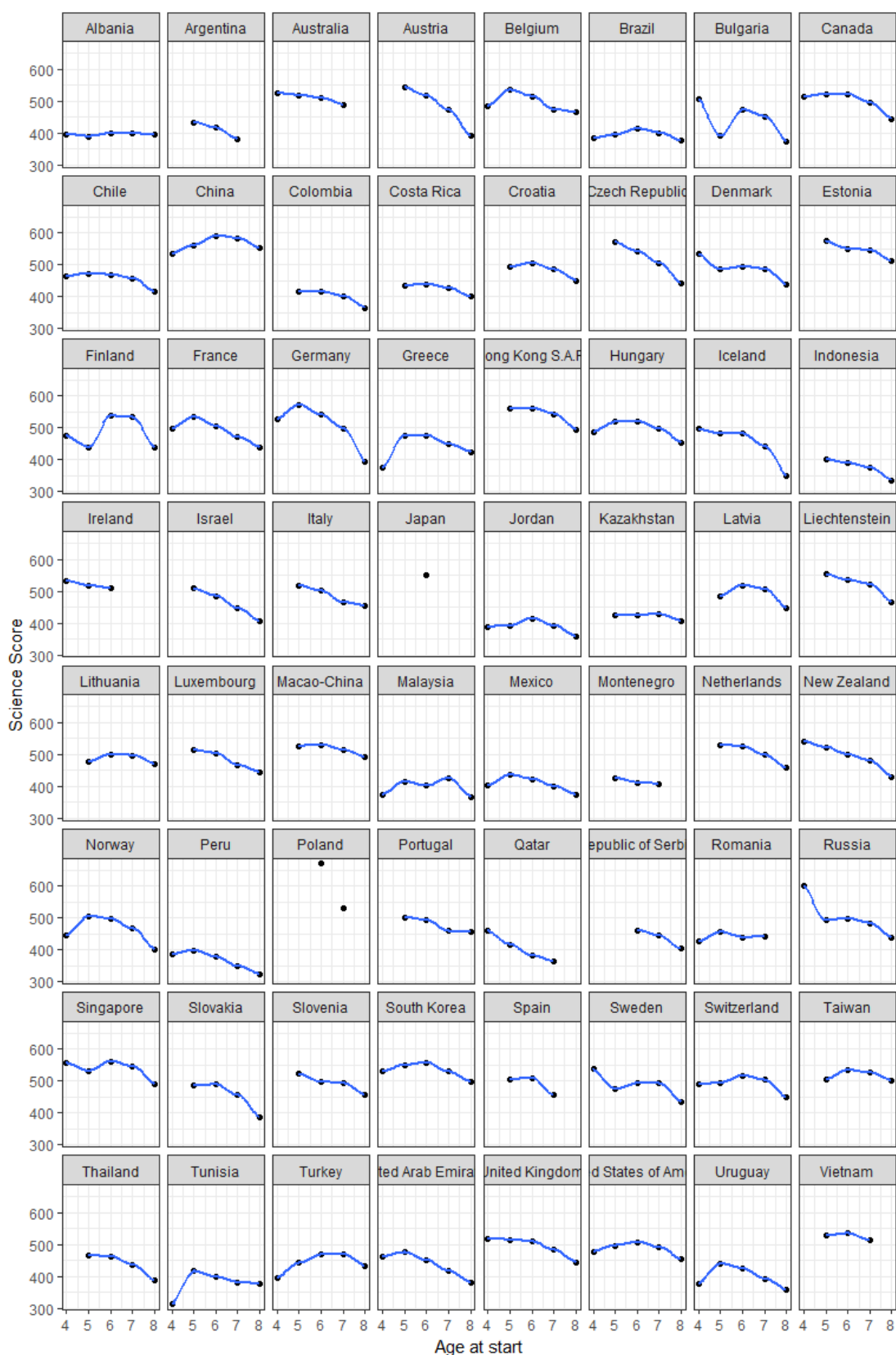


图 8：学生科学题成绩与上学年纪的关系

根据图 6、7、8，可以发现学生的成绩和入学年级基本呈负相关的关系，5 岁上学的学生的成绩比 8 岁上学的学生成绩要好。所以我们分析，对于学生而言，政府不应该限制入学年龄。

5. 学生入学年纪在各个国家地区的分布

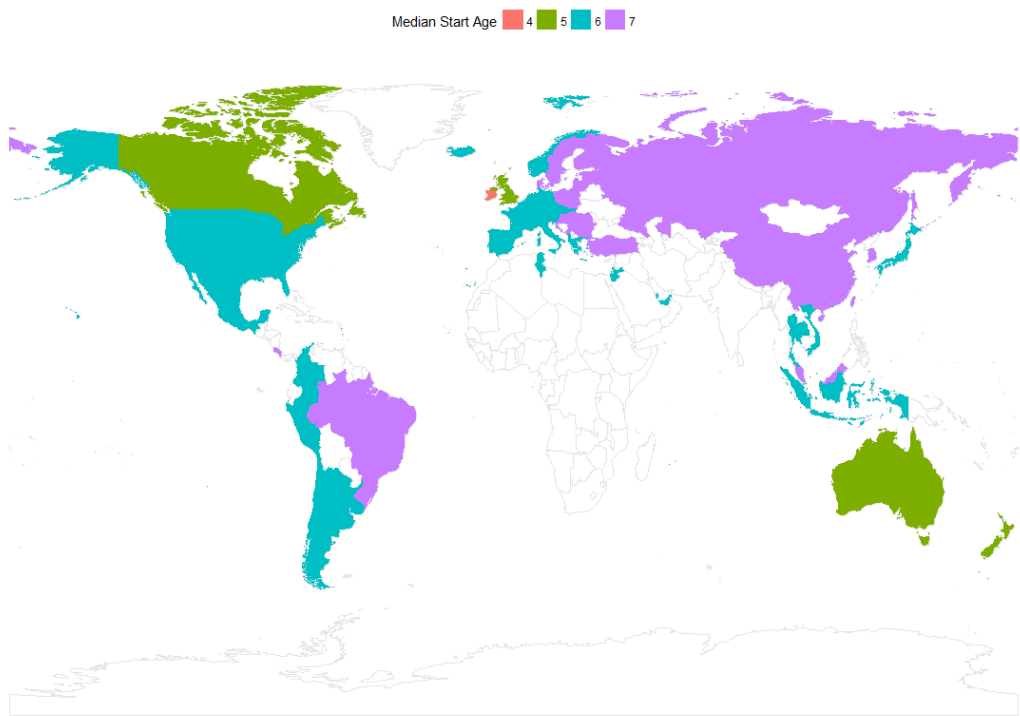


图 9：学生入学年纪分布地图

4.2 其它因素

1. 学生成绩与家里书籍的数量存在关系吗？

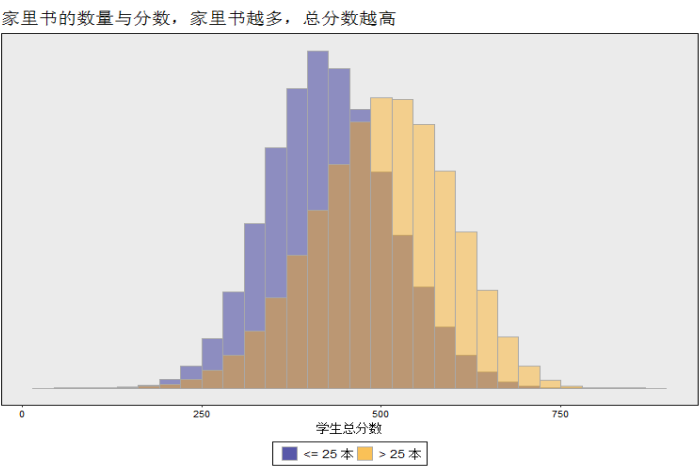


图 10：学生成绩与家里存书的数量关系

2. 学生课外的学习与成绩的关系？

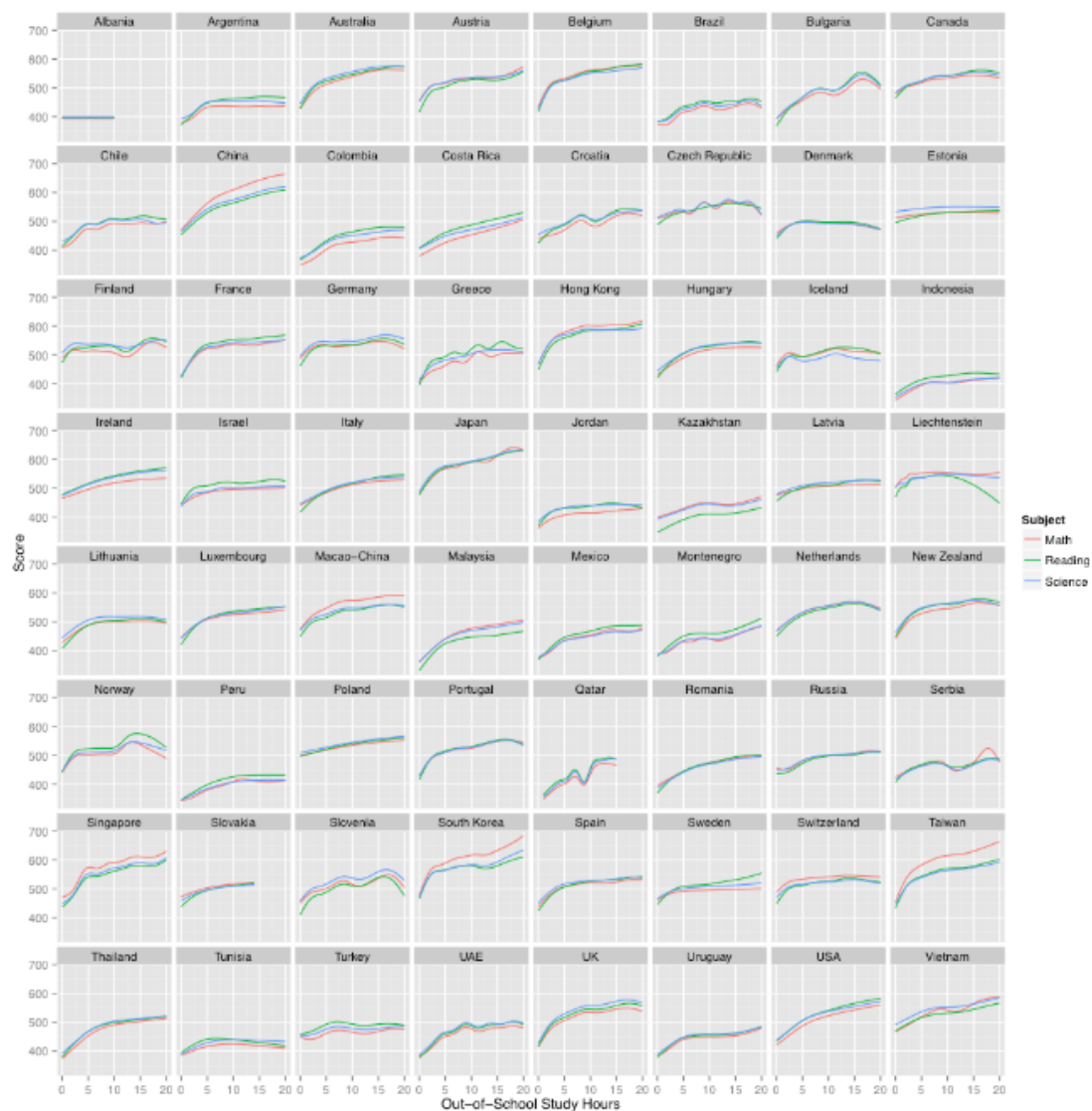


图 11: 课外花费的学习时间与成绩关系

3. 性别与数学成绩的关系?

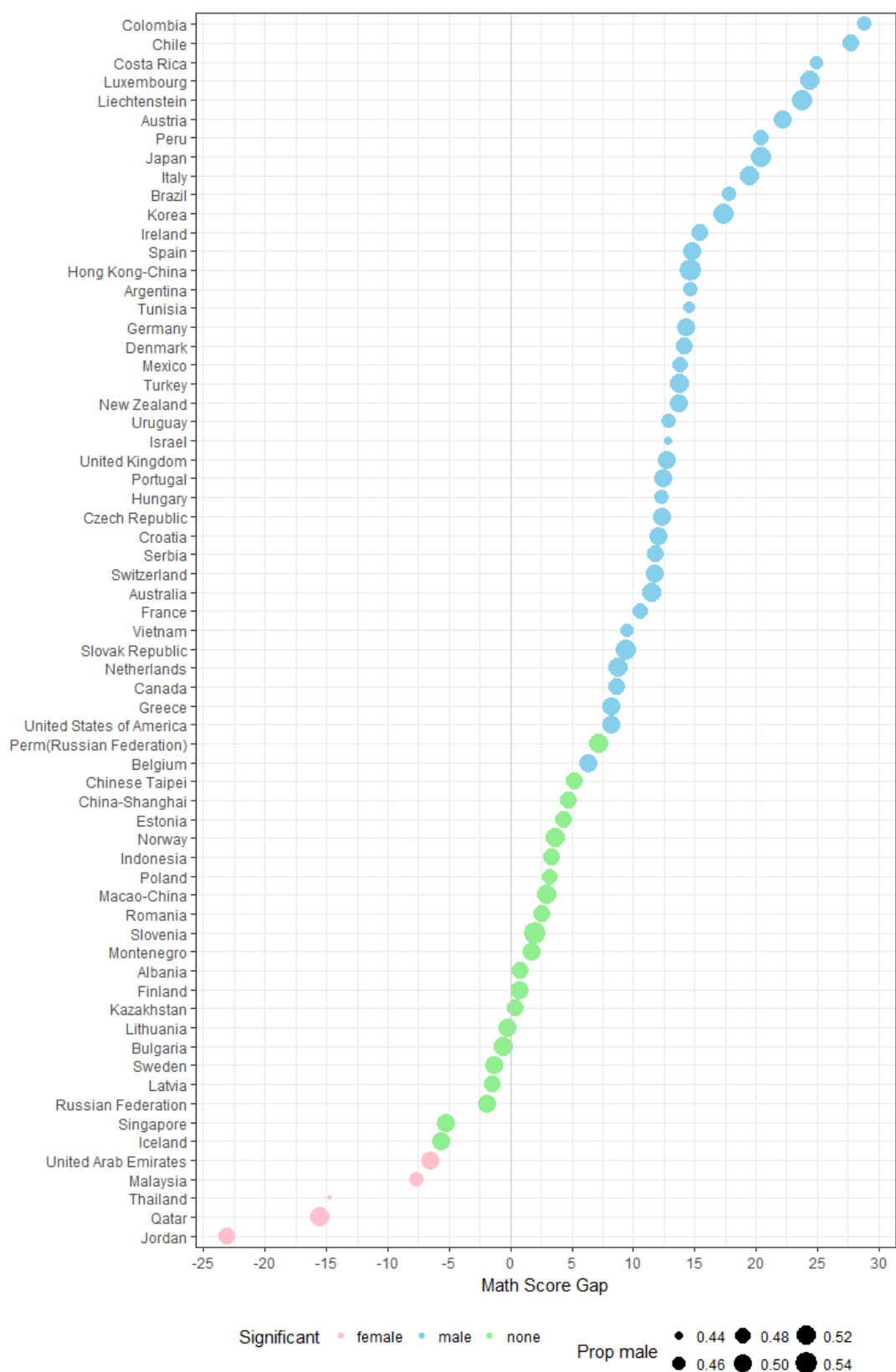


图 12: 性别在地区中的分布和性别与数学成绩的关系

5 结论

从实验结果看，我们的可视化实验展现了 PISA 数据集中蕴含的各种关系。主要的关注点是学生的成绩受各种因素的影响，挖掘出的知识不仅体现了我们对教育的常识认知，说明了方法的正确性。同时也挖掘了一些不常见的数据知识，证明了数据可视化是数据挖掘的一门重要的研究方向，可以帮助我们探索数据中蕴含的有用信息，以数据指导我们的工作。