中国人民大學
RENMIN UNIVERSITY OF CHINA

数据的可视化分析

李扬

中国人民大学统计学院

2020 年 6 月 16 日

中国人民大學
RENMIN UNIVERSITY OF CHINA

目录

研究设计

总体和抽样框

- **目标总体：**
  - 全国范围内（不包括港澳台）的 55.3 万行政村、2.6 亿农户、7.9 亿农村居民。(数据来源：各省 2017 年统计年鉴整理)
- **抽样框：**

数据	发布方	数据源
县级行政单位目录	民政部	http://www.mca.gov.cn
行政村目录	国家统计局	http://www.stats.gov.cn
县级农业人口	公安部治安管理局	《中华人民共和国全国分县市人口统计资料（2012 年)》
贫困县目录	国务院扶贫开发领导小组办公室	http://www.cpad.gov.cn
分层目录		《中国自然地理》《中国地理》教材
县级第一产业增加值	各县级行政单位	《2016 年国民经济和社会发展统计公报》

总体和抽样框

- **三级抽样框**

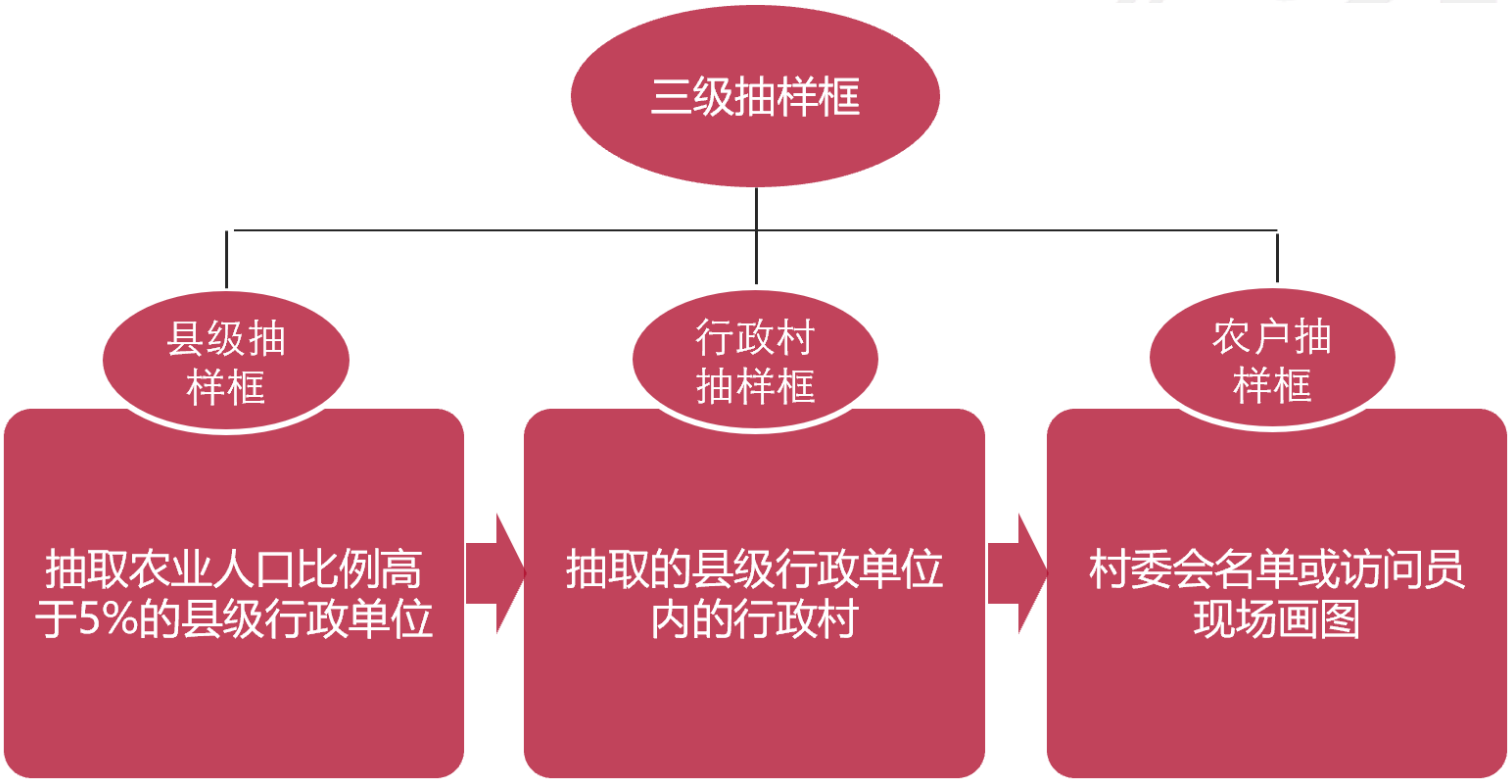


图 1: 三级抽样框

抽样方案概述

- **三阶段抽样**
  1. **第一阶段：**
     - 县级行政单位
     - 分层 PPS
     - 分层依据为地理区域、是否国家级贫困县；PPS 抽样以各县级行政单位农业人口比例为辅助变量
  2. **第二阶段：**
     - 行政村
     - 简单随机抽样
     - 对每个抽取的县级行政单位内符合条件的行政村，采取简单随机抽样抽取特定个数的行政村
  3. **第三阶段：**
     - 农户
     - 等距抽样、地图抽样
     - 访员自行整理行政村农户信息，等距抽样或地图抽样

抽样方案概述

样本量设计

- 在 95% 置信度和全国相关指标相对误差 20% 内的要求下，由公式 $\frac{z_{\alpha/2}\sqrt{P(1-P)}}{P\sqrt{n_{SRS}}} \leq r$ 可计算得项目样本量全国为 9,600 农户，根据第一阶段各层农业人口比例，可得各层待抽取数量。

第一阶段抽样

图 2: 第一阶段抽样效果

- 其中：红色点代表贫困县层; 棕色点代表华中层

第二阶段抽样

图 3: 第二阶段抽样效果

第三阶段抽样

图 4: 农户[1]抽样示例

- 其中：黑色线为港田村边界; 白色线为最佳调研路线；红色数字为建筑编号农户抽样框未列出

[1]江西省南昌市新建县乐化镇港田村

县级样本轮换

县级样本单元采用样本轮换方法

- **轮换周期**
  - 每两年轮换
- **轮换方法**
  - 每年在各层内，将已调查县级行政单位按照人均第一产业从高到低排序，采用等距抽样法抽取 20% 县级行政单位替换为新的县级行政单位
- **轮换抽样**
  - 新的县级行政单位和行政村的抽取与前述第一阶段抽样方法一致

计划调查情况

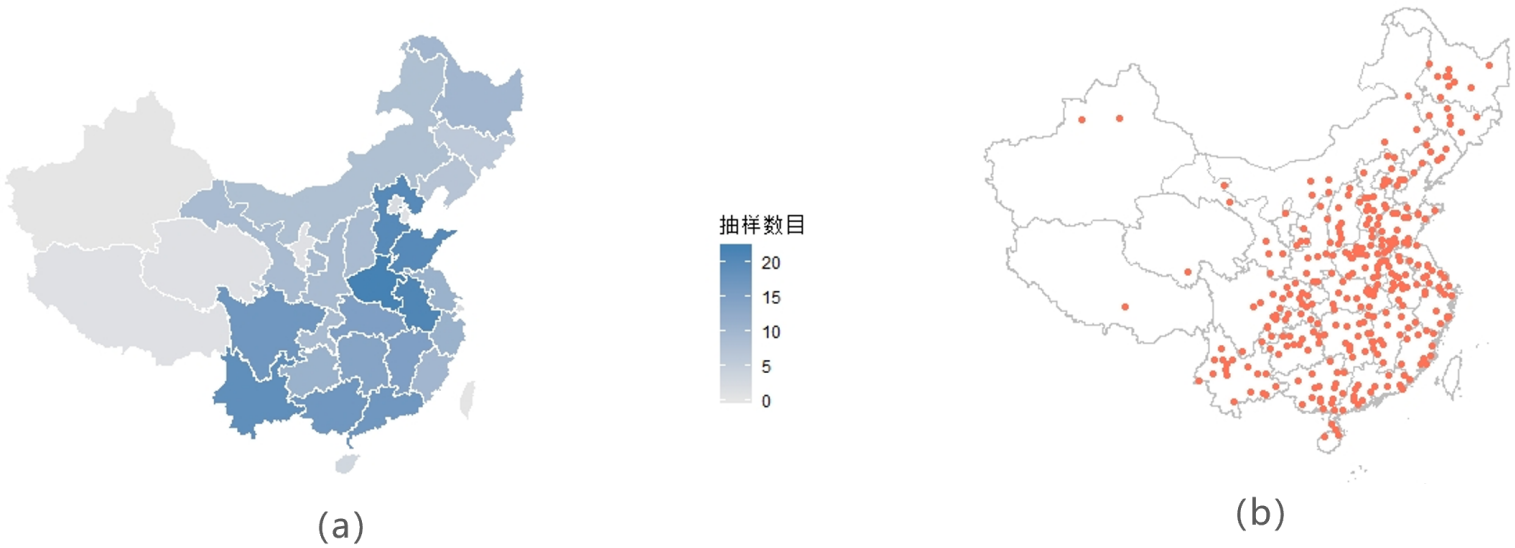
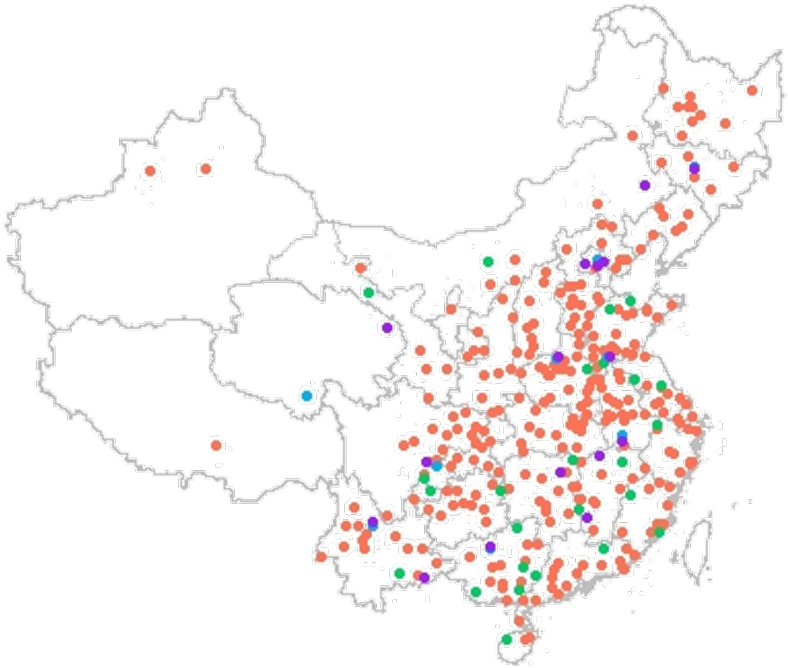


图 5: 抽取的行政村的地理分布情况

- 图（a）为各省级行政单位计划调查的行政村样本密度图，颜色越深代表该省级行政单位中计划调查的行政村数目越多
- 图（b）为计划调查的行政村样本在全国的分布，各坐标点表示计划调查的行政村位置

2018 实际调研情况对比

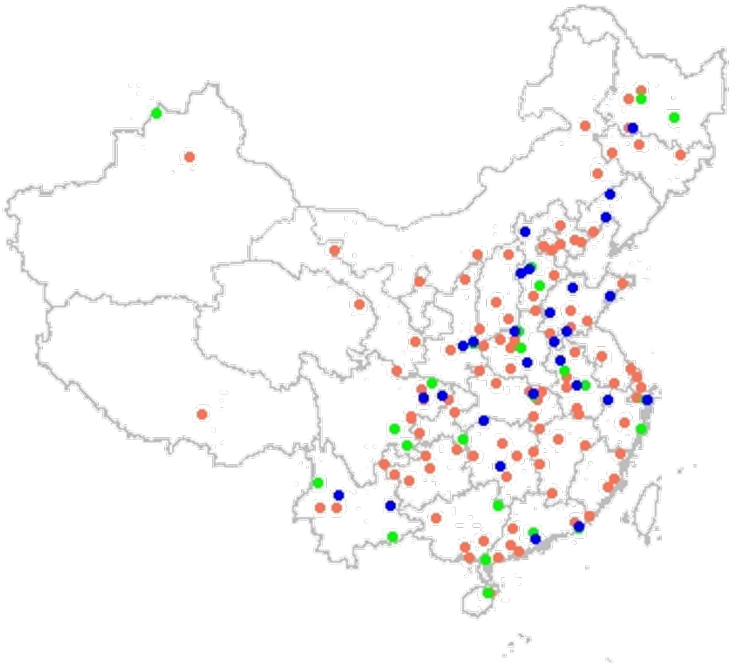


红色坐标点：
表示计划调查并实际调查了的行政村样本，共279个。
绿色坐标点：
表示计划调查但实际未调查，也未被替换的行政村样本，共25个。
紫色坐标点：
表示不在计划调查的样本中，但实际调查了的行政村样本，共16个。
蓝色坐标点：
表示计划调查但实际未调查，并对其替代样本进行了调查的行政村样本，共16个。

图 6: 抽样行政村的调研情况图

- 图中坐标点表示行政村的位置，计划调查行政村样本 320 个，实际调查行政村样本 295 个。

2019 实际调研情况对比

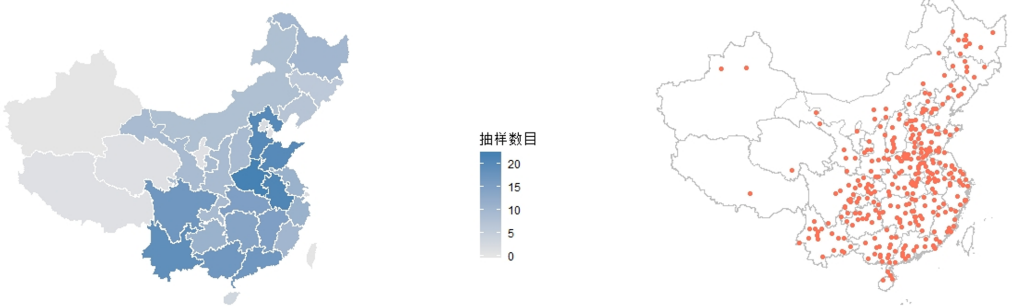


红色坐标点：
表示计划调查并实际调查了的行政村样本，共100个。
绿色坐标点：
表示计划调查但实际未调查，也未被替换的行政村样本，共41个。
蓝色坐标点：
表示不在计划调查的样本中，但实际调查了的行政村样本，共28个。

图 7: 抽样行政村的调研情况图

- 图中坐标点表示行政村的位置，计划调查行政村样本 141 个，实际调查行政村样本 128 个。

可视化内容



(a)

(b)

图 8: 可视化内容

可视化原因

什么是可视化

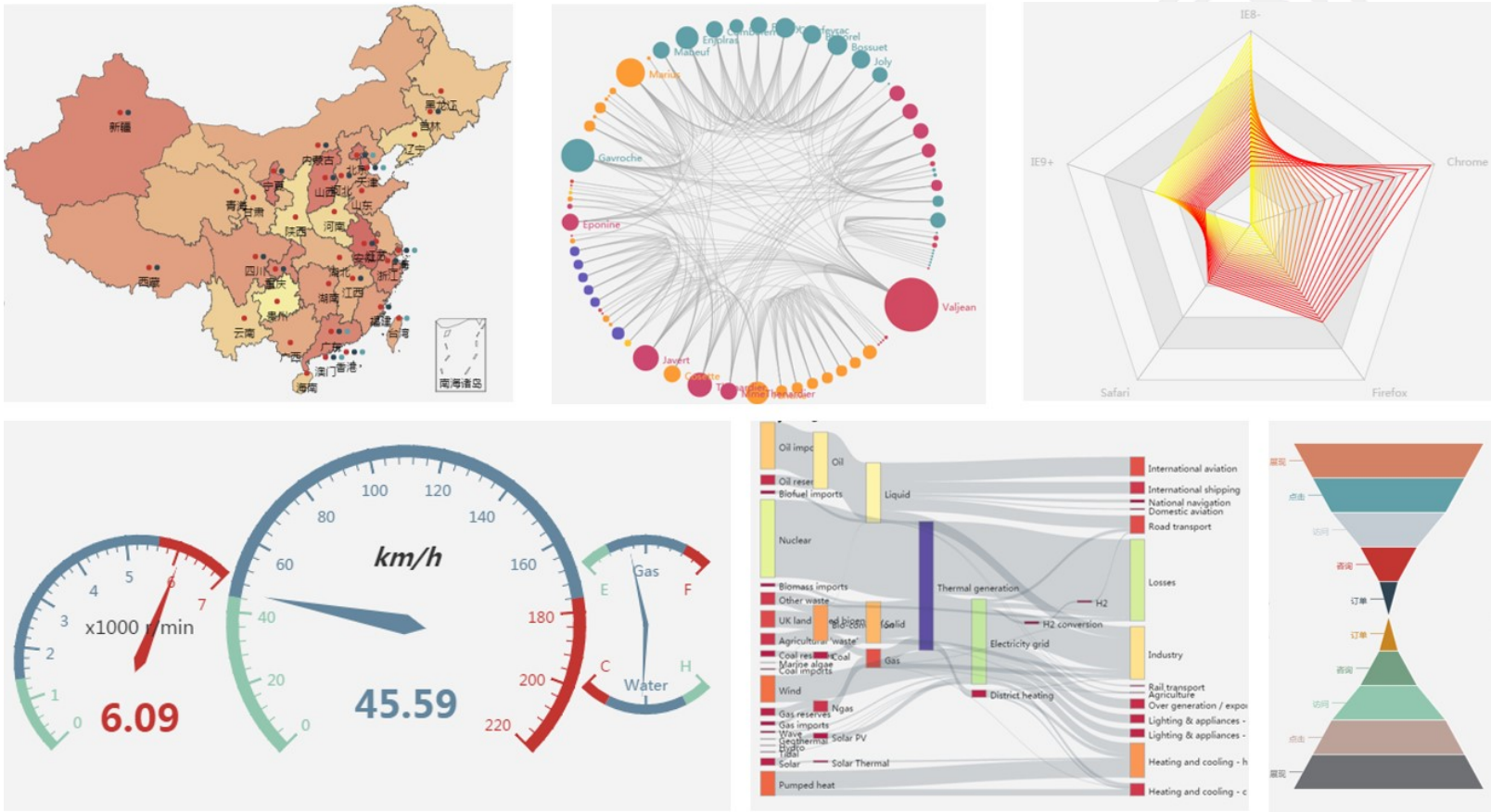


图 9: 可视化示例

为什么要可视化

表 1: Anscombe's Quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	10.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	13.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	9.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	11.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	14.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	6.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	4.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	12.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	7.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	5.0	6.89

为什么要进行可视化

这四组数据中：

- x 值的平均数都是 9.0，y 值的平均数都是 7.5
- x 值的方差都是 10.0，y 值的方差都是 3.75
- 它们的相关度都是 0.816
- 线性回归线都是 $y = 3 + 0.5x$

为什么要进行可视化

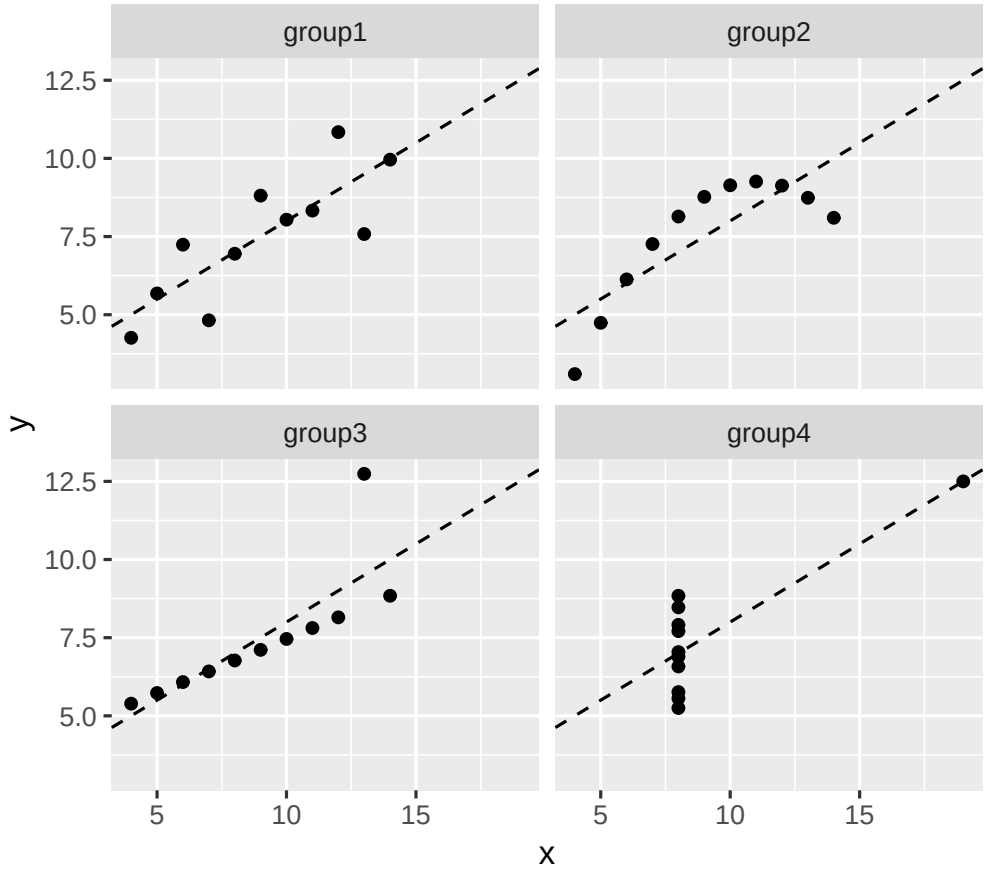


图 10: 散点图示例

为什么要进行可视化

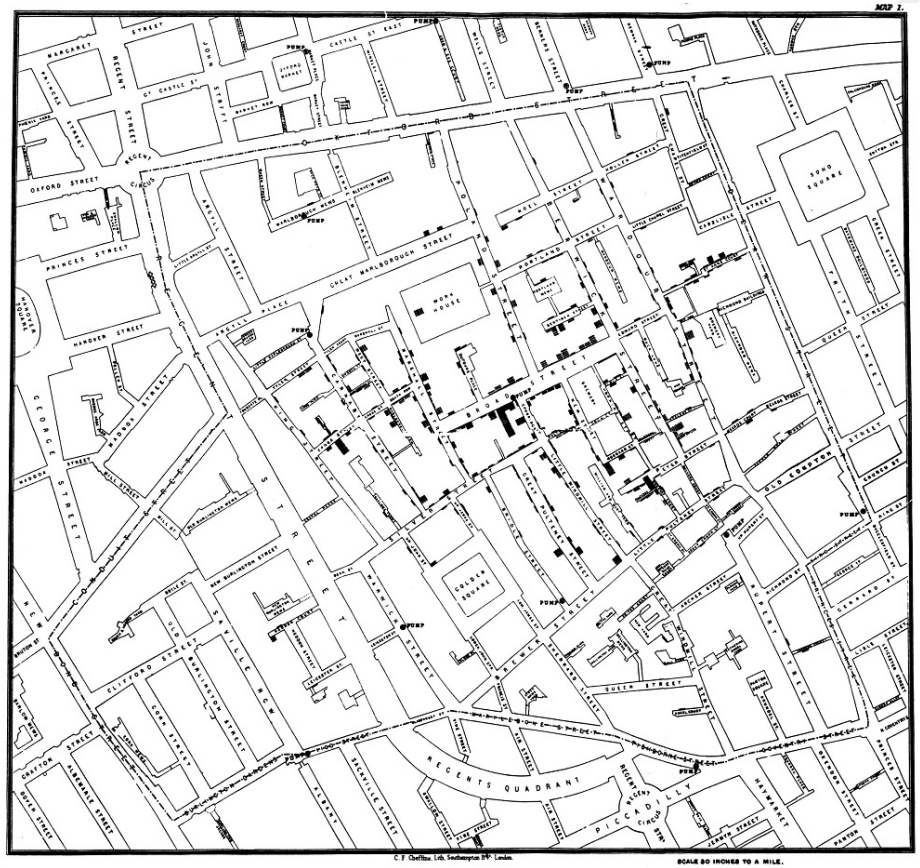


图 11: 伦敦霍乱示意图

可视化分析

可视化分析（Visualized Data Analytics）

1. 检查数据问题
2. 探索关系与趋势
3. 建模准备

检查数据问题

- 缺失

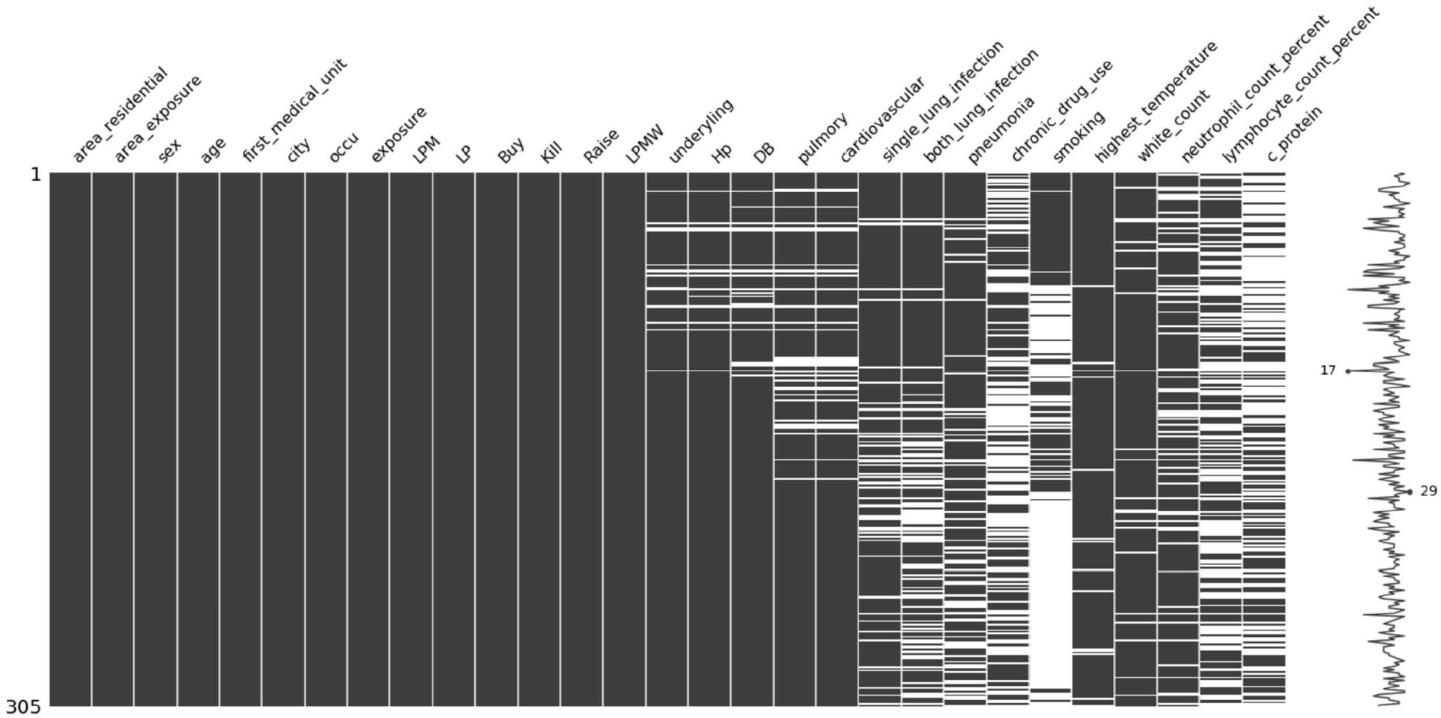


图 12: 数据缺失情况可视化

检查数据问题

- 异常

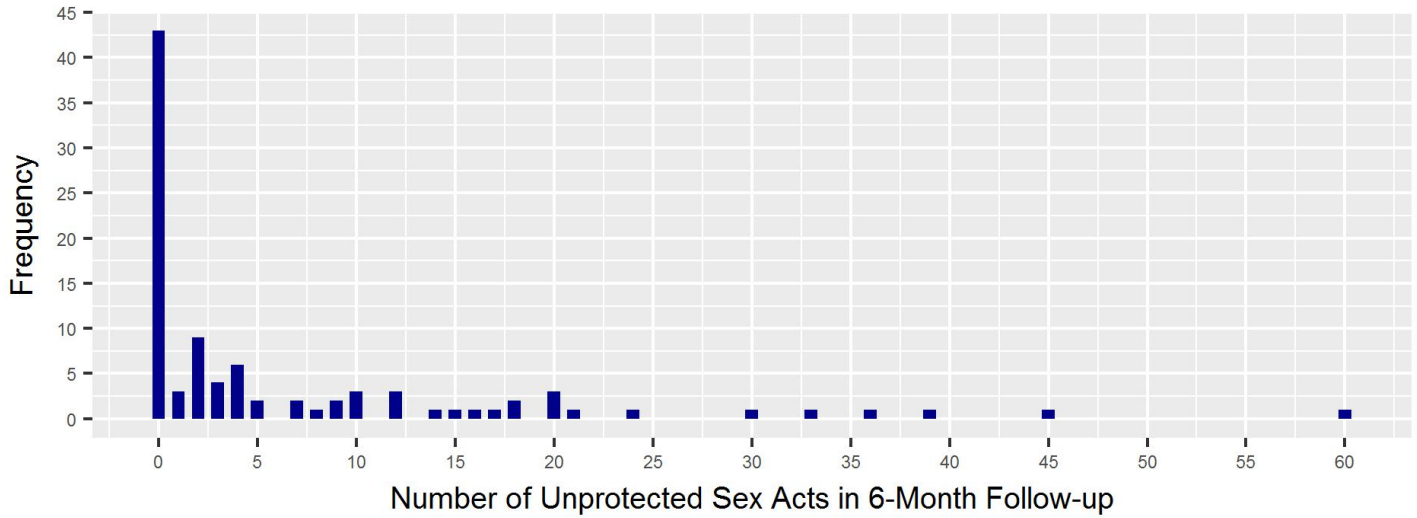


图 13: 数据异常情况可视化

检查数据问题

- 异常

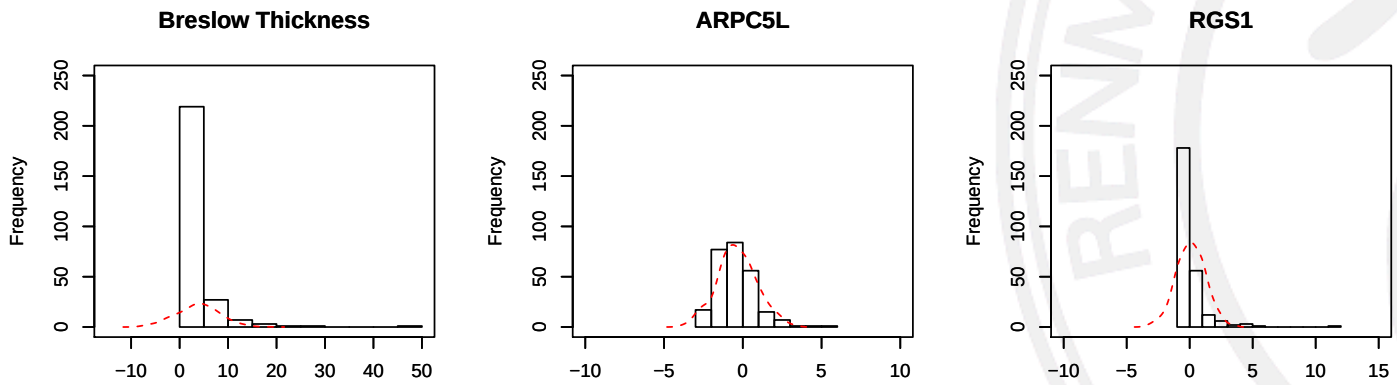
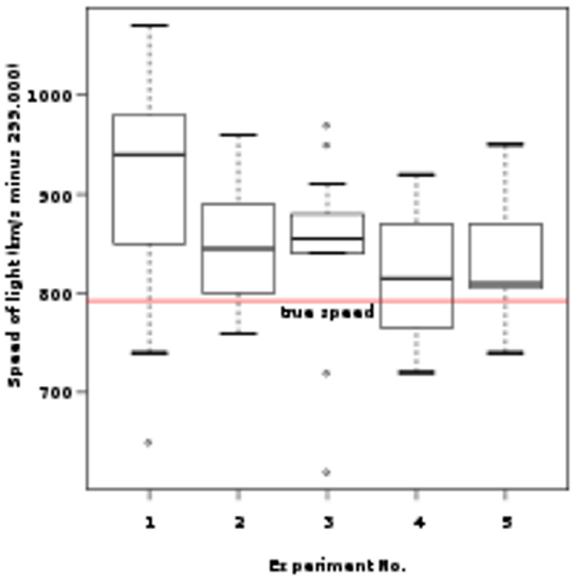


图 14: 数据异常情况可视化

检查数据问题

- 异常



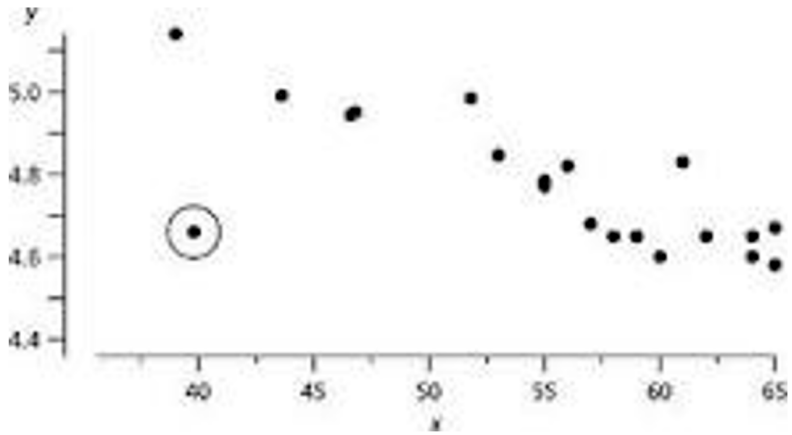
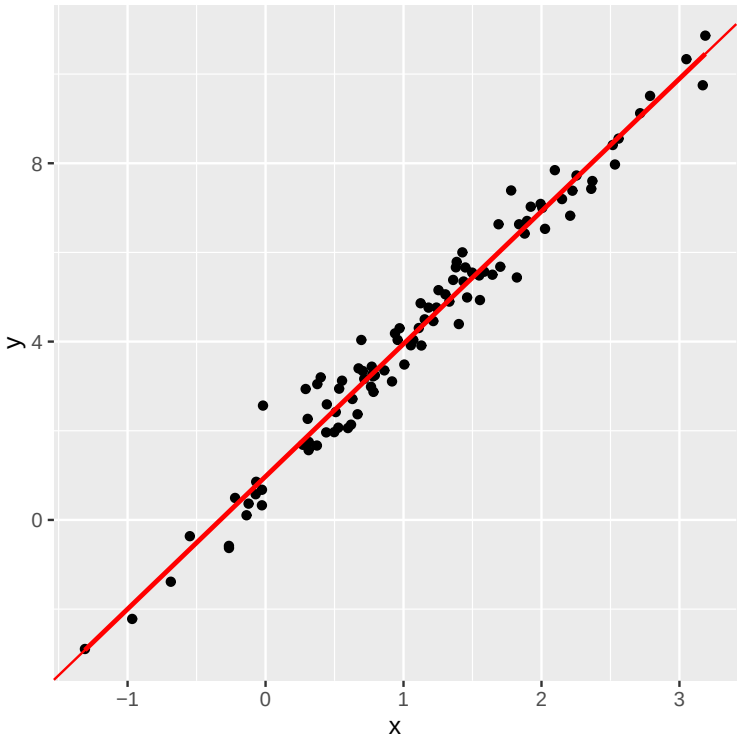
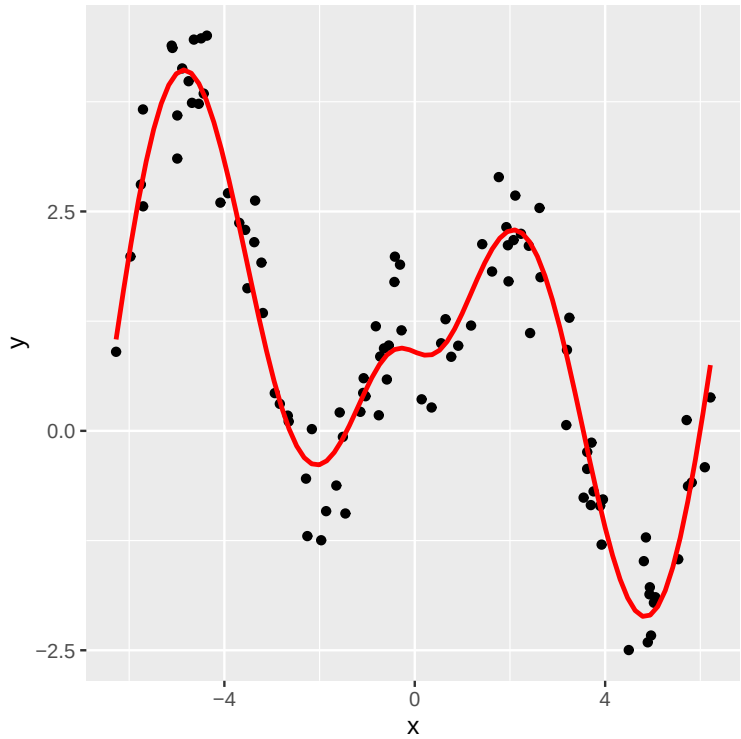


图 15: 数据异常情况可视化

探索关系与趋势

- 线性非线性





探索关系与趋势

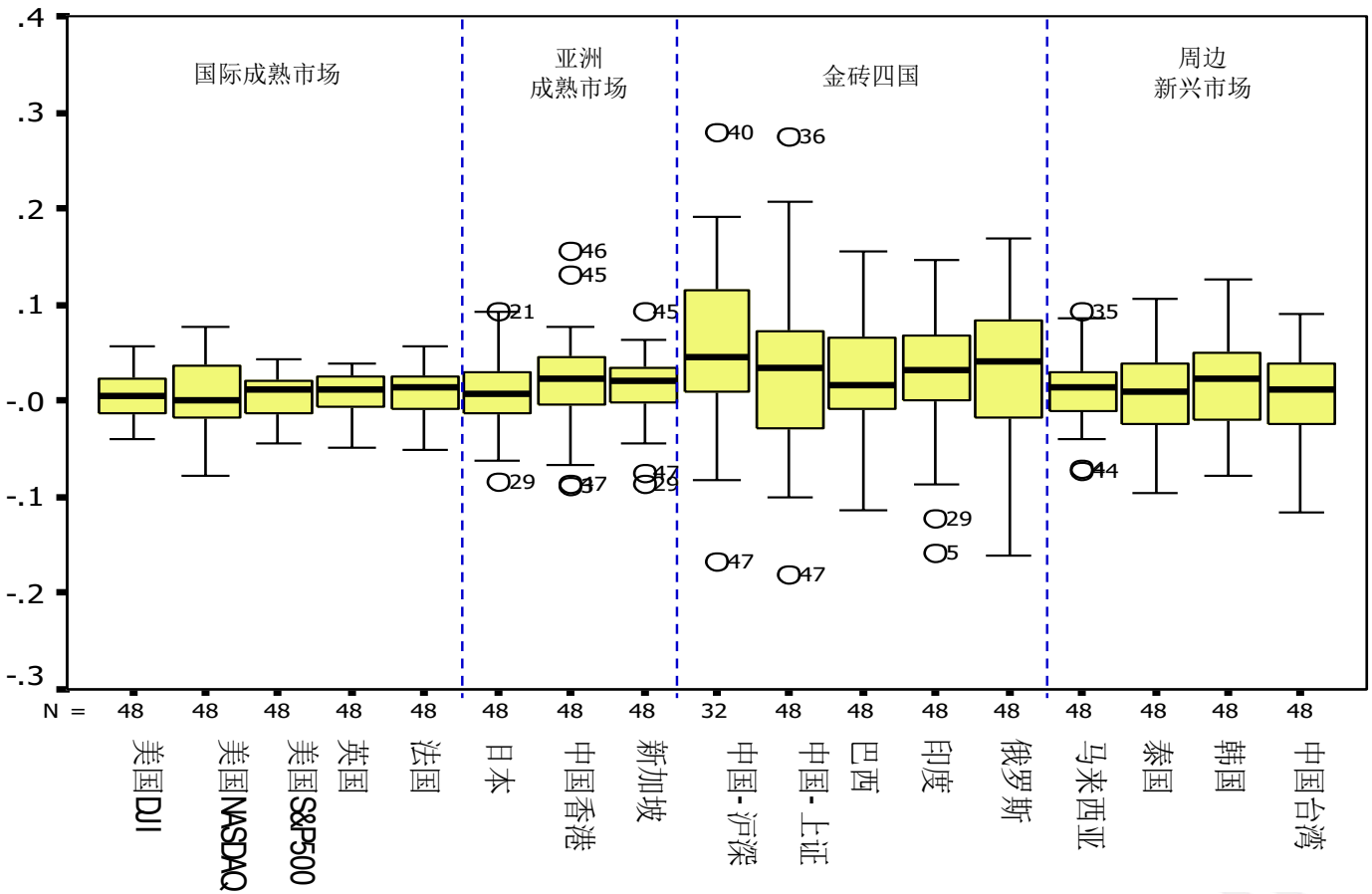


图 16: 多组 box-plot

探索关系与趋势

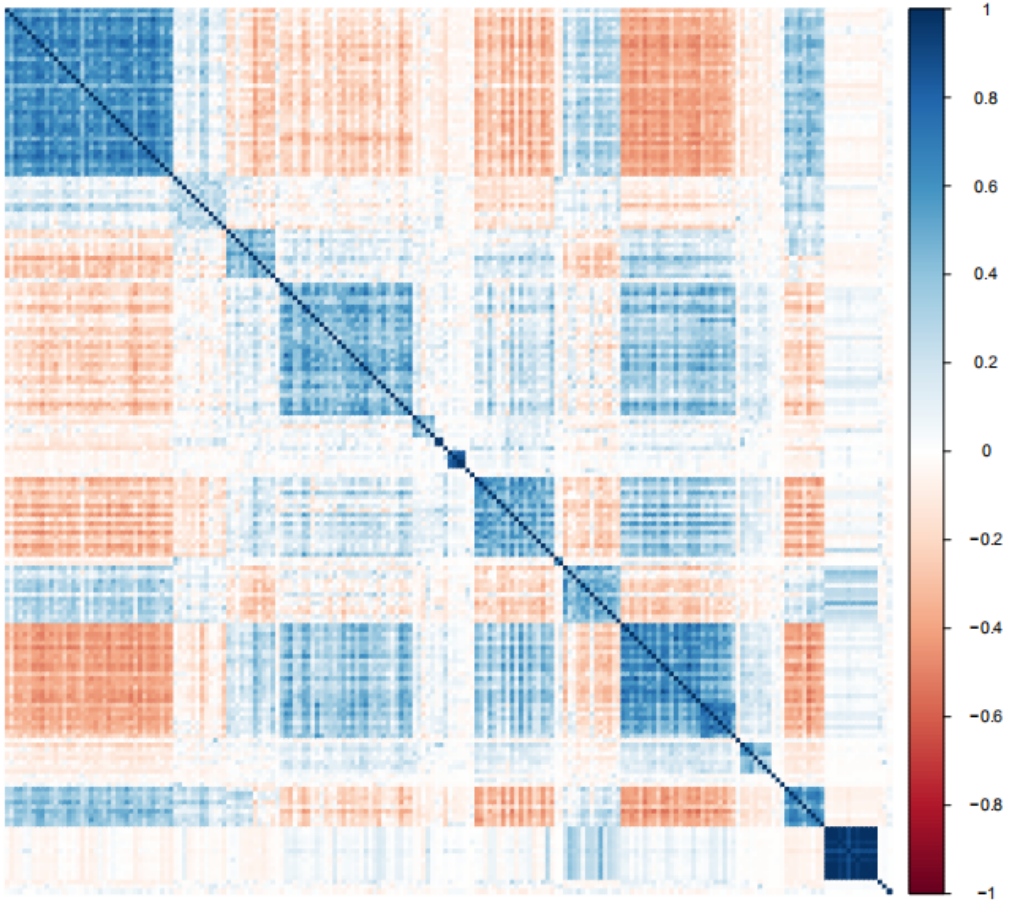


图 17: 多变量相关结构

探索关系与趋势

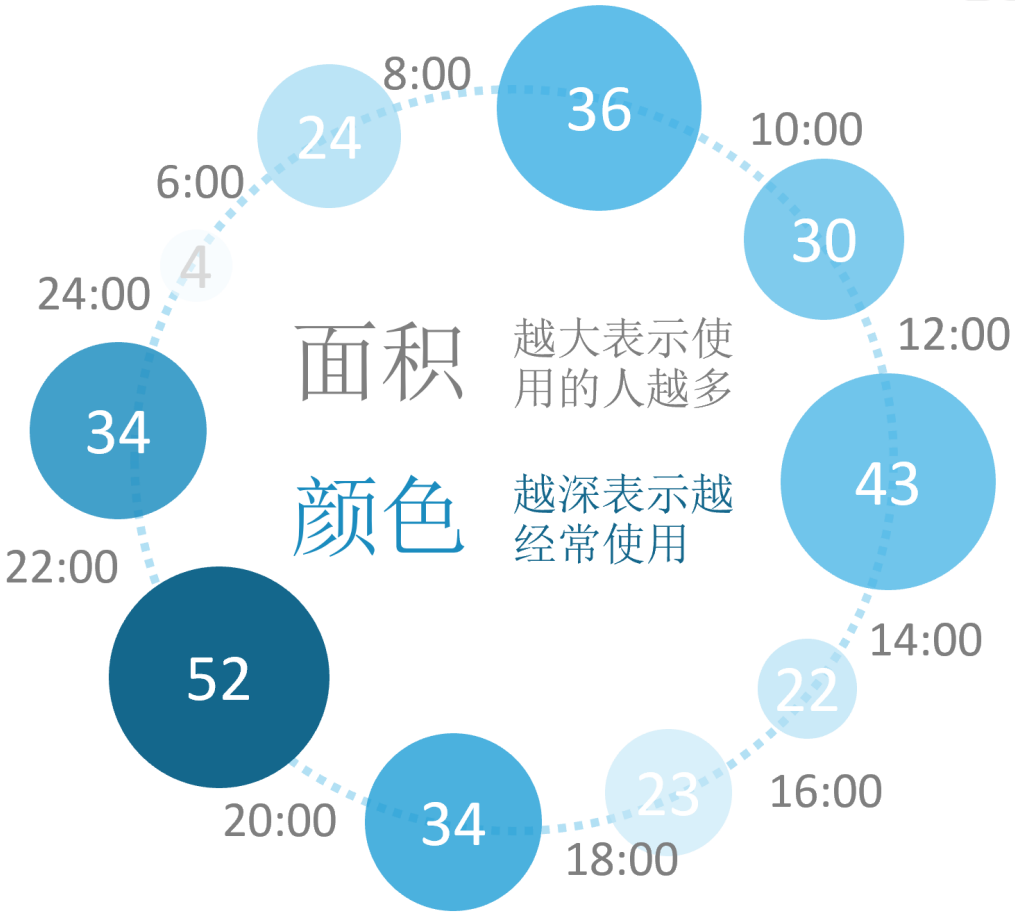


图 18: 多变量相关结构

建模准备

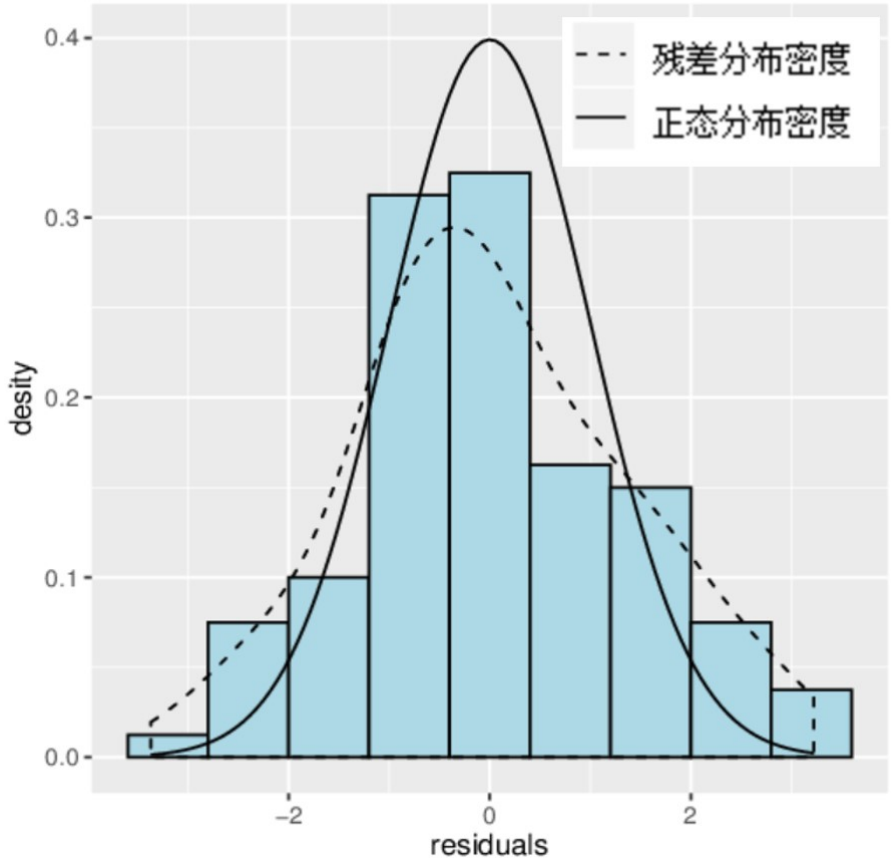


图 19: 分布密度图

建模准备

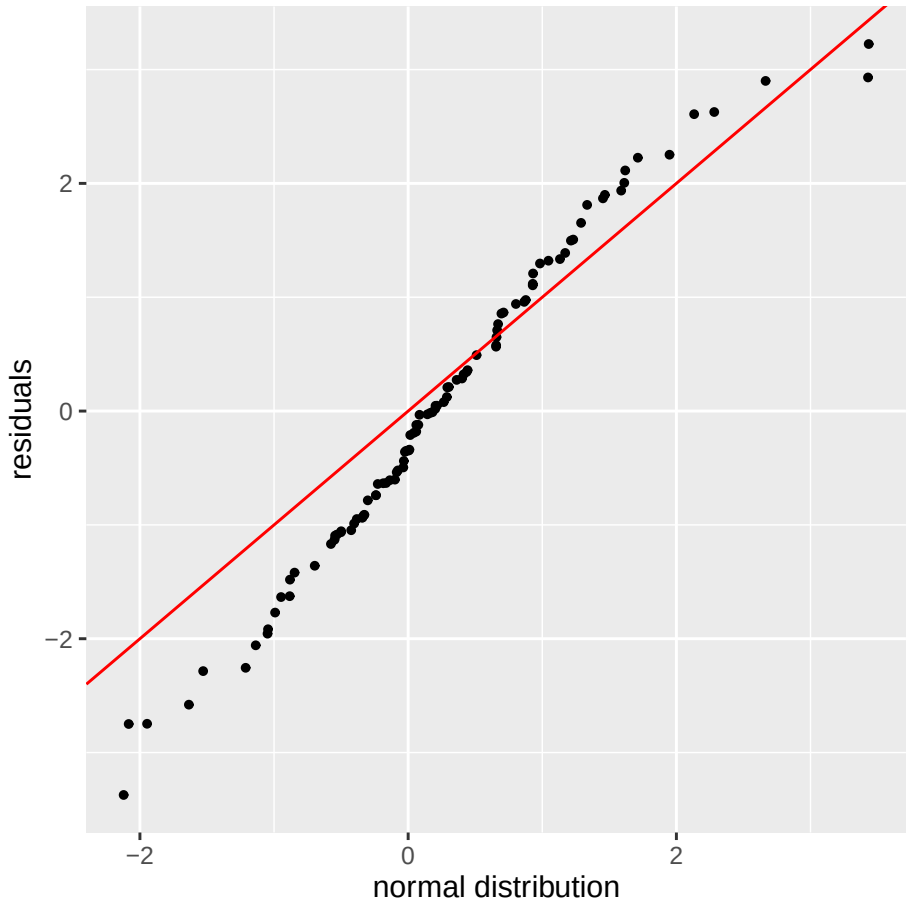


图 20: qq 图

建模准备

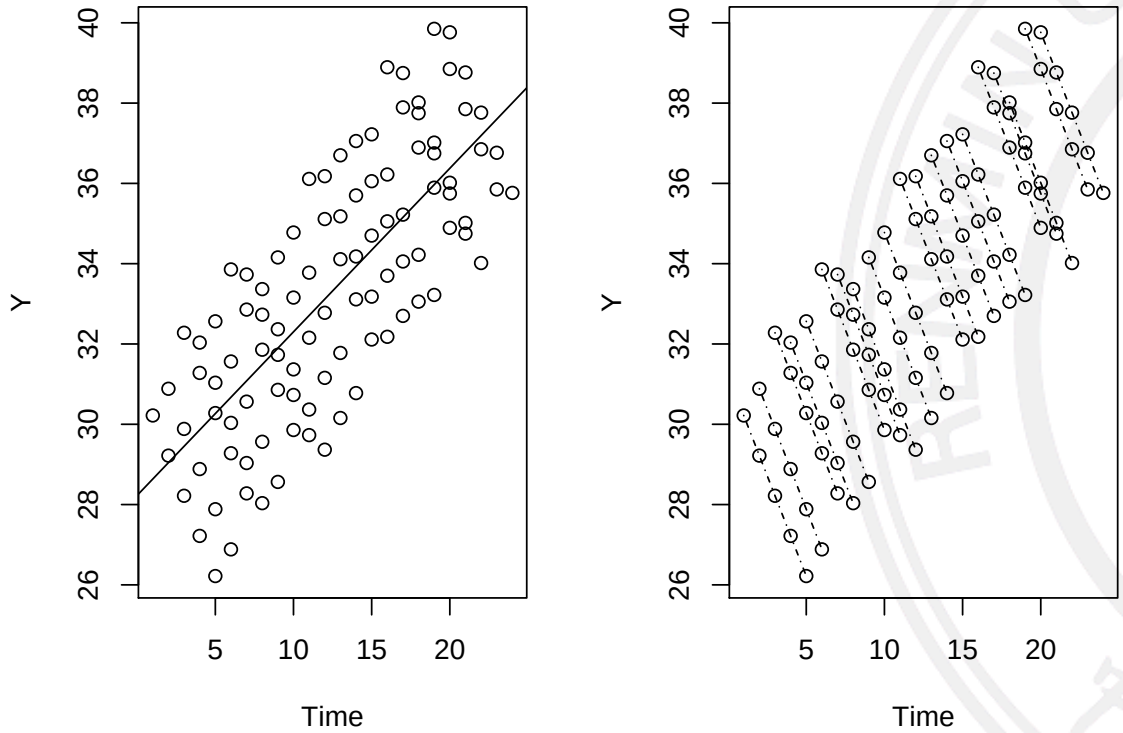


图 21: 模型选择

可视化原则

可视化原则

1. 完整性和易读性
2. 忠实度和美观性
3. 可视化空间内平衡
4. Proportional ink
5. 编号与标题
6. ……

完整性和易读性

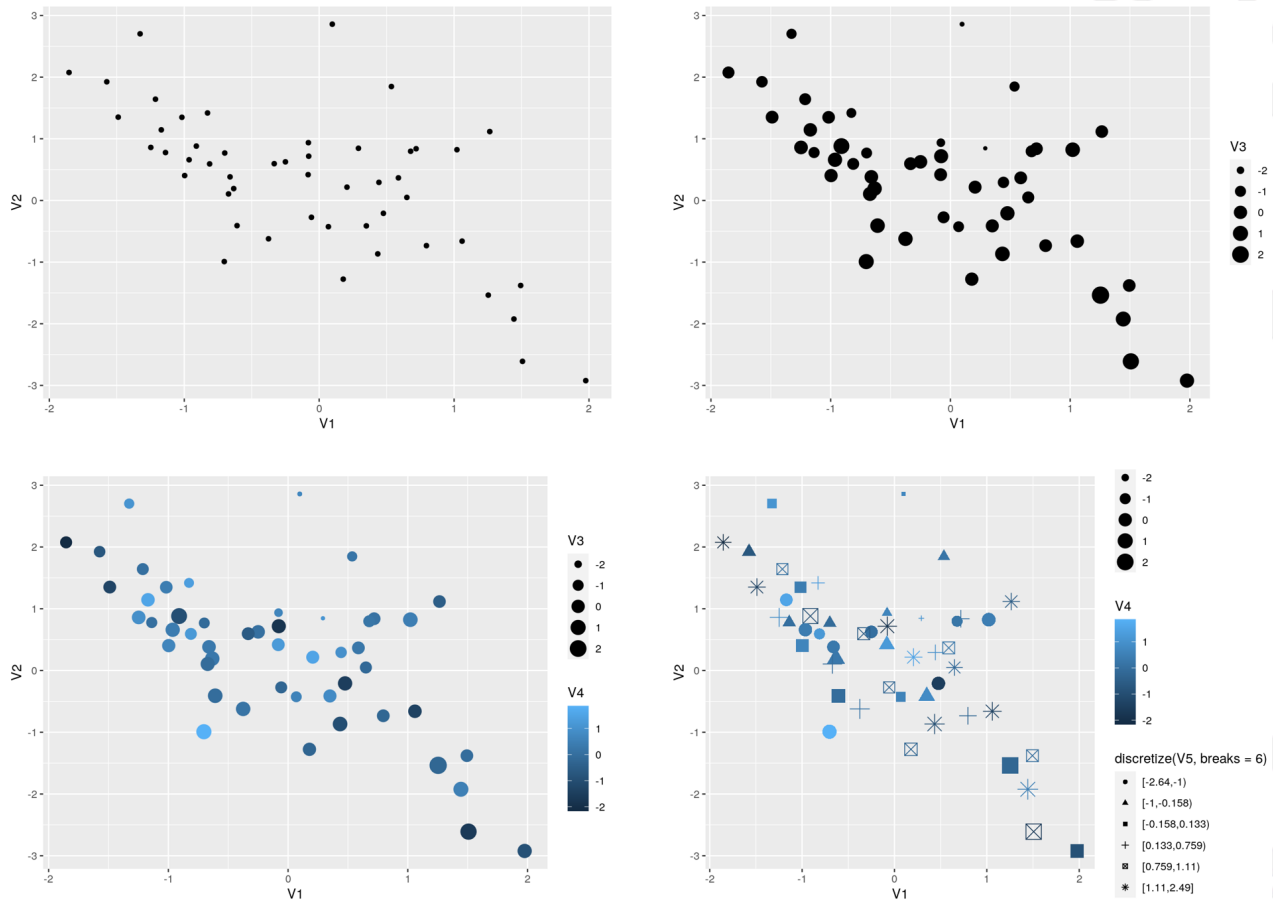
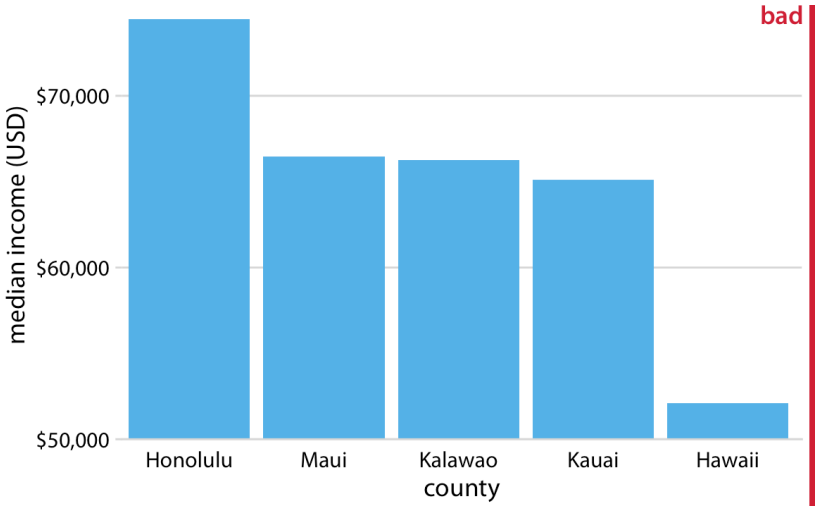
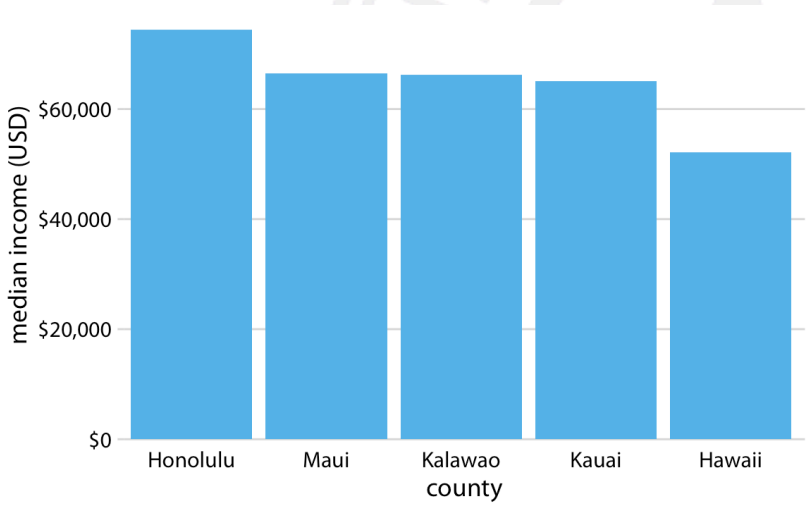


图 22: 完整性和易读性

忠实度和美观性


bad
median income (USD)
$70,000
$60,000
$50,000
Honolulu
Maui
Kalawao
Kauai
Hawaii
county


median income (USD)
$60,000
$40,000
$20,000
$0
Honolulu
Maui
Kalawao
Kauai
Hawaii
county

可视化空间内平衡

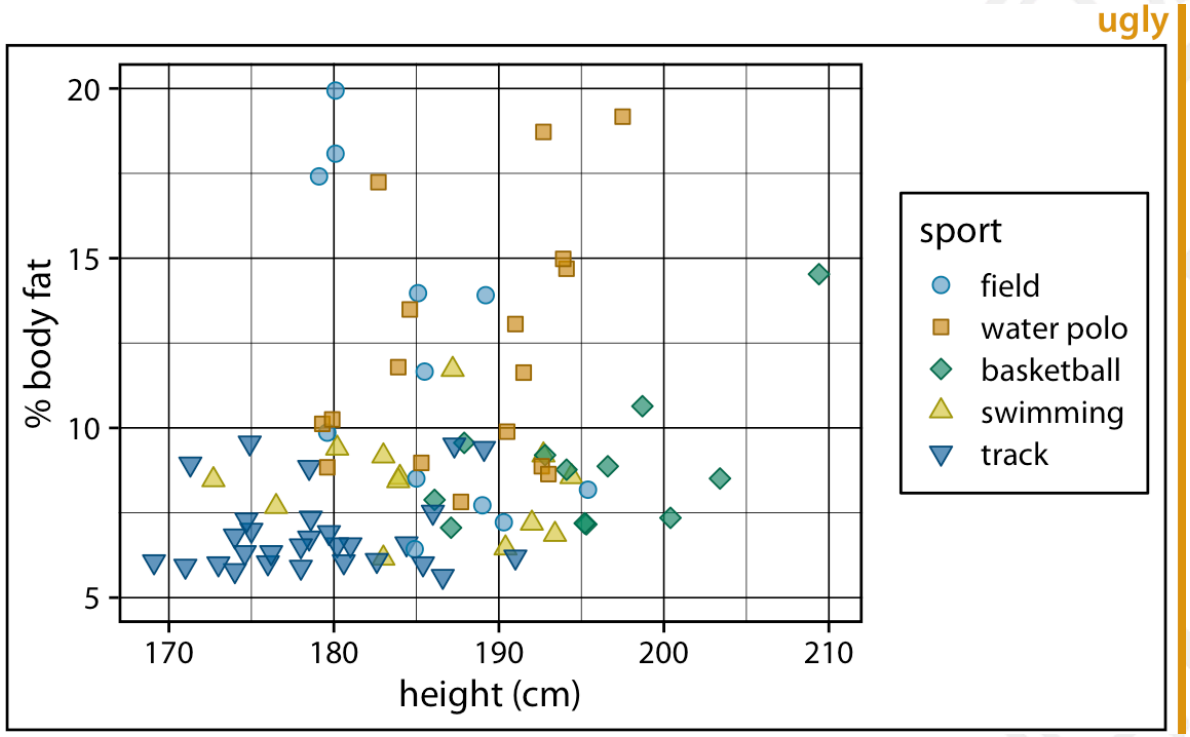


图 23: 可视化空间内平衡

可视化空间内平衡

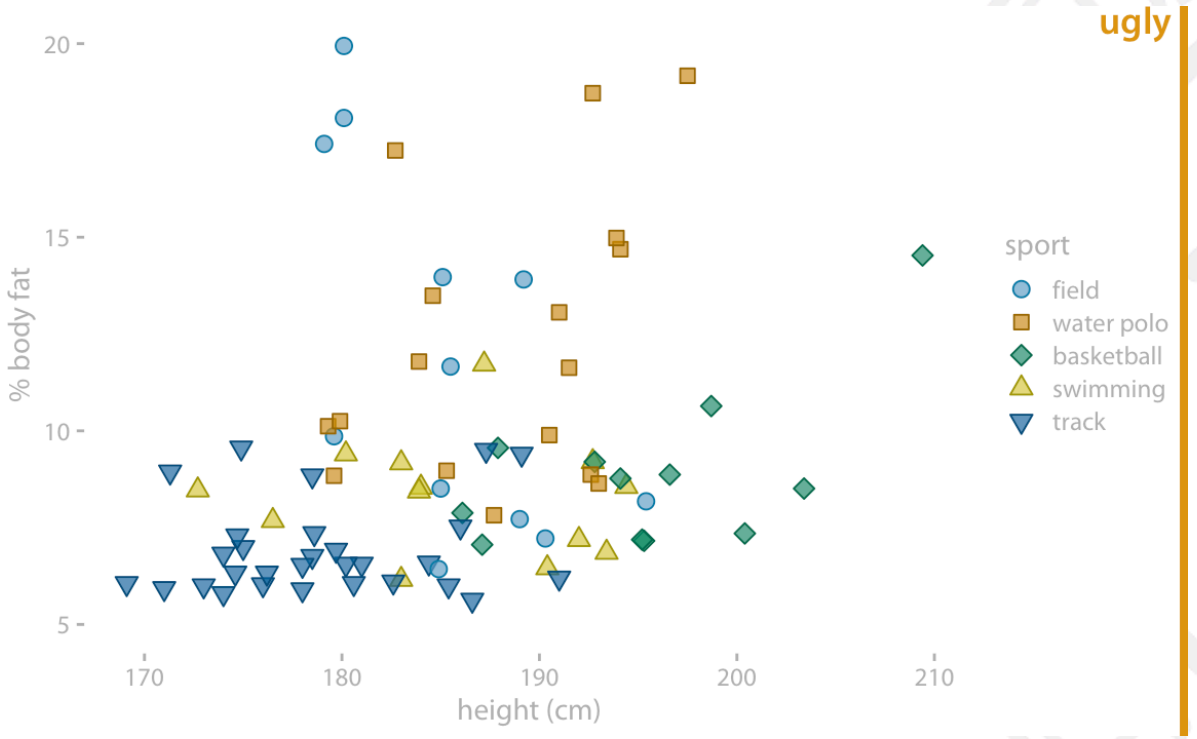


图 24: 可视化空间内平衡

可视化空间内平衡

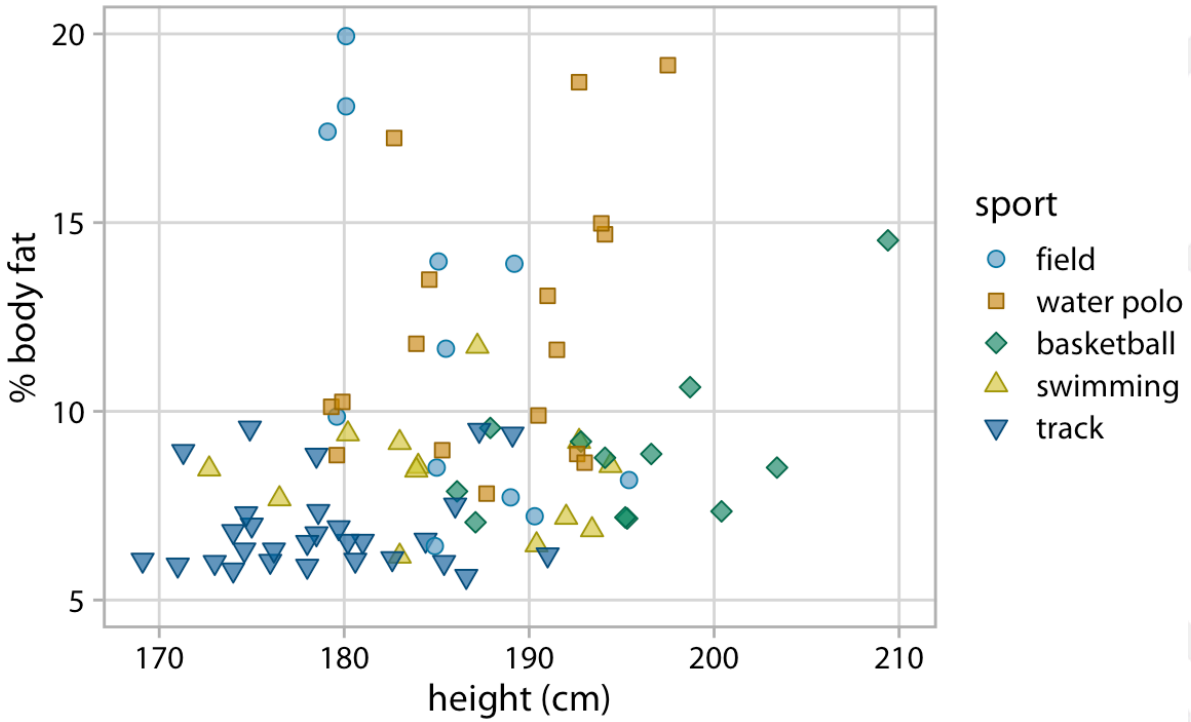
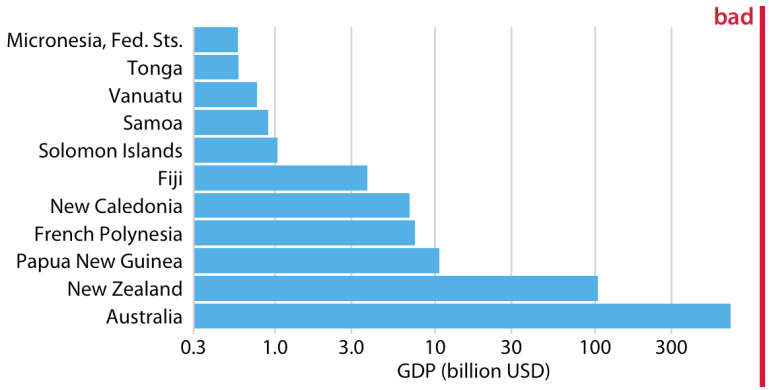
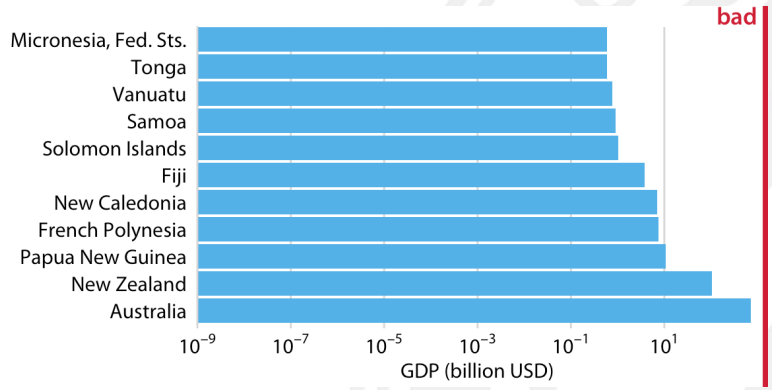
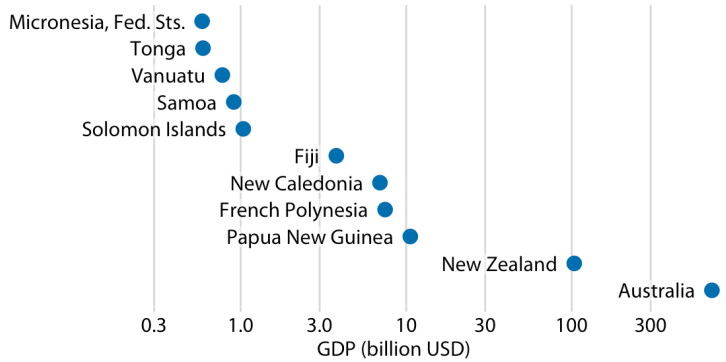


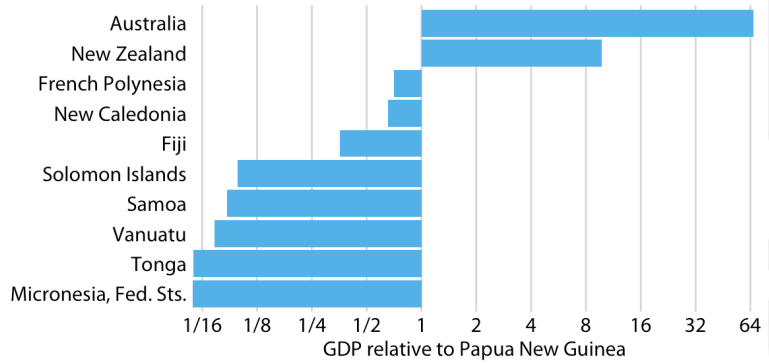
图 25: 可视化空间内平衡

Proportional ink

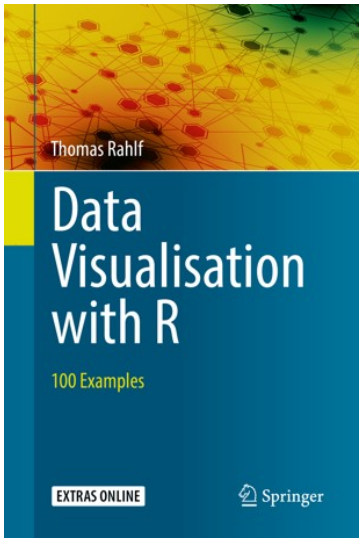


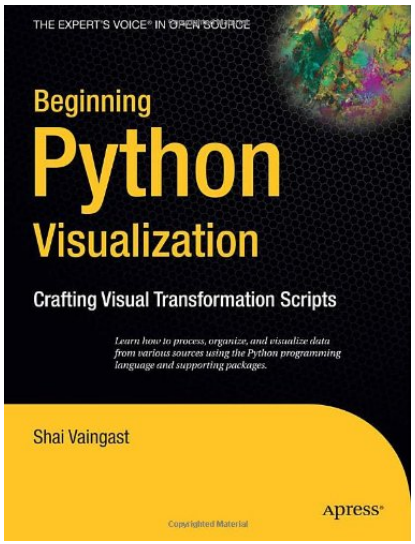


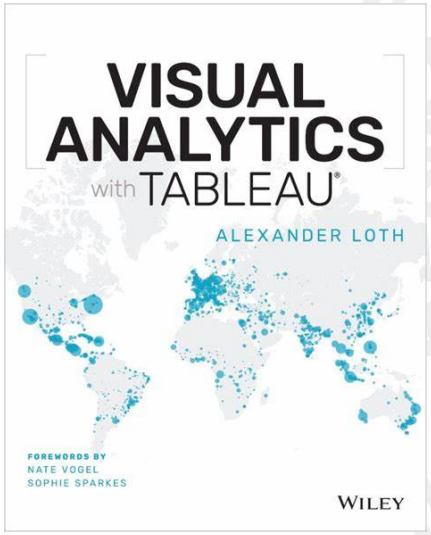


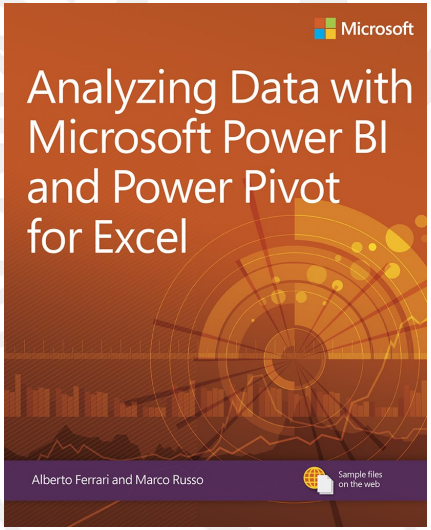


可视化分析的工具


Thomas Rahlf
Data Visualisation with R
100 Examples
EXTRAS ONLINE
Springer


THE EXPERT'S VOICE® IN OPEN SOURCE
Beginning
Python
Visualization
Crafting Visual Transformation Scripts
Learn how to process, organize, and visualize data from various sources using the Python programming language and supporting packages.
Shai Vaingast
Apress®
Copyrighted Material


VISUAL ANALYTICS
with TABLEAU
ALEXANDER LOTH
FOREWORDS BY
NATE VOGEL
SOPHIE SPARKES
WILEY


Microsoft
Analyzing Data with Microsoft Power BI and Power Pivot for Excel
Alberto Ferrari and Marco Russo
Sample files on the web

可视化示例

示例 1 花费构成

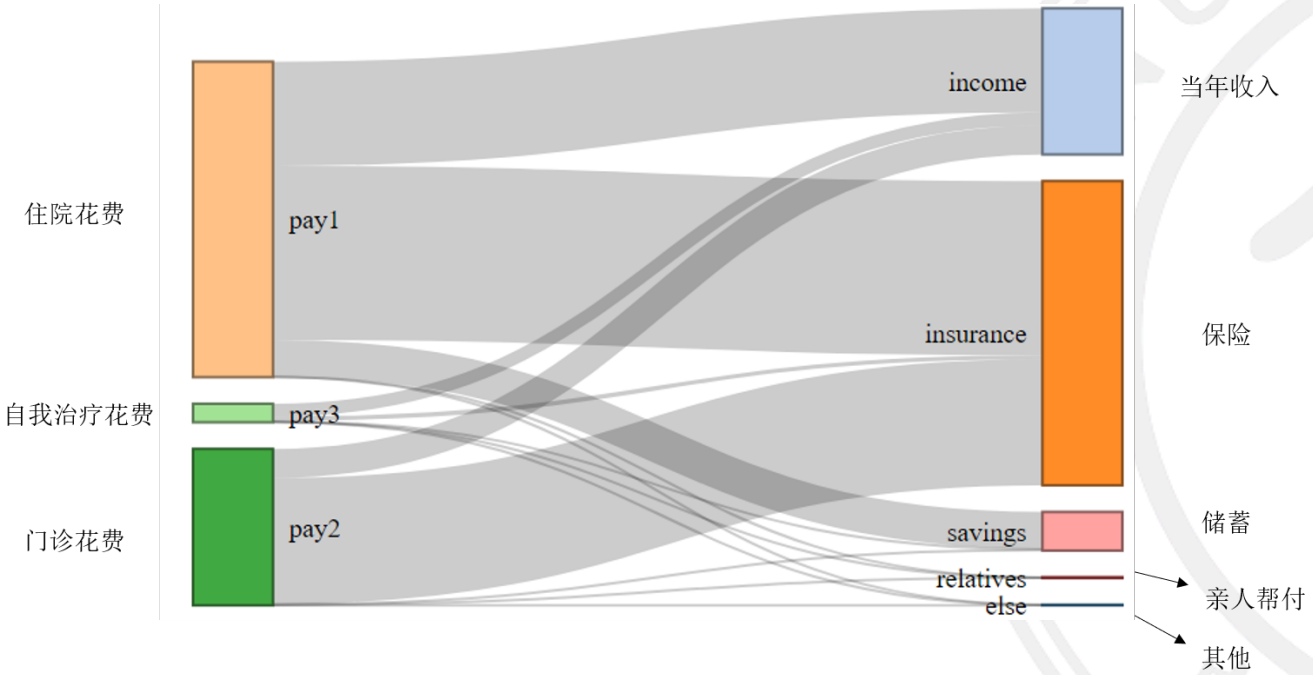
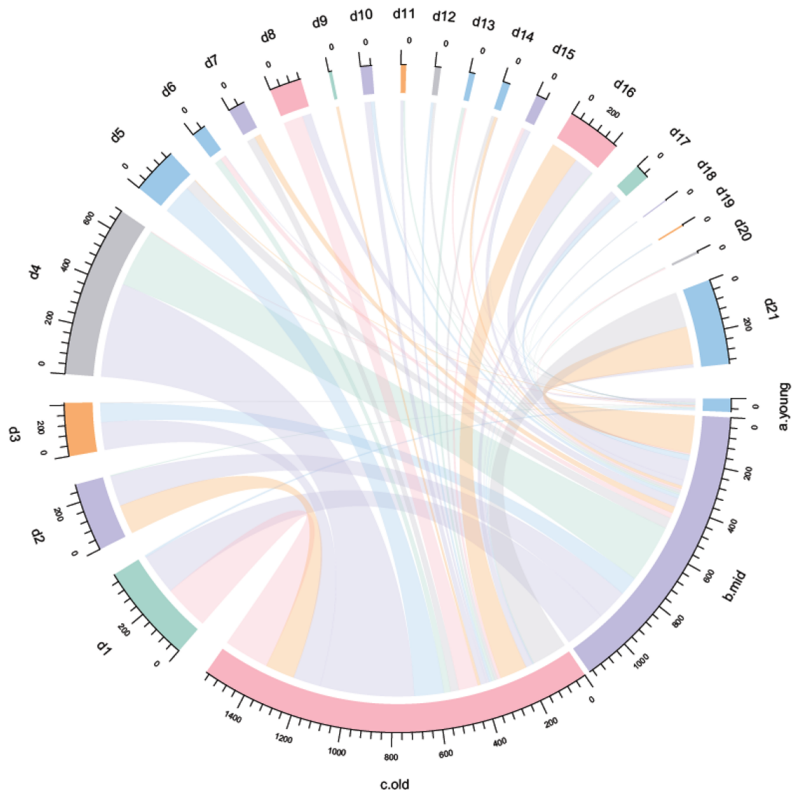


图 26: 桑基图

示例 2 网络结构



符号	含义	符号	含义
d1	关节炎	d13	慢性阻塞性肺部疾患
d2	风湿病	d14	哮喘
d3	糖尿病（血糖高）	d15	慢性肝病（肝硬化）
d4	高血压	d16	胃部或消化系统疾病
d5	脑血管病	d17	肾脏疾病
d6	风湿性心脏病	d18	老年痴呆
d7	缺血性心脏病	d19	帕金森综合症
d8	冠心病	d20	精神类疾病
d9	中风	d21	其他
d10	肺心病	a.young	年龄小于40岁
d11	脑卒中	b.mid	年龄在40-60间
d12	肿瘤	c.old	年龄大于60岁

图 27: 不同年龄与各种慢病关系

示例 3 比例对比

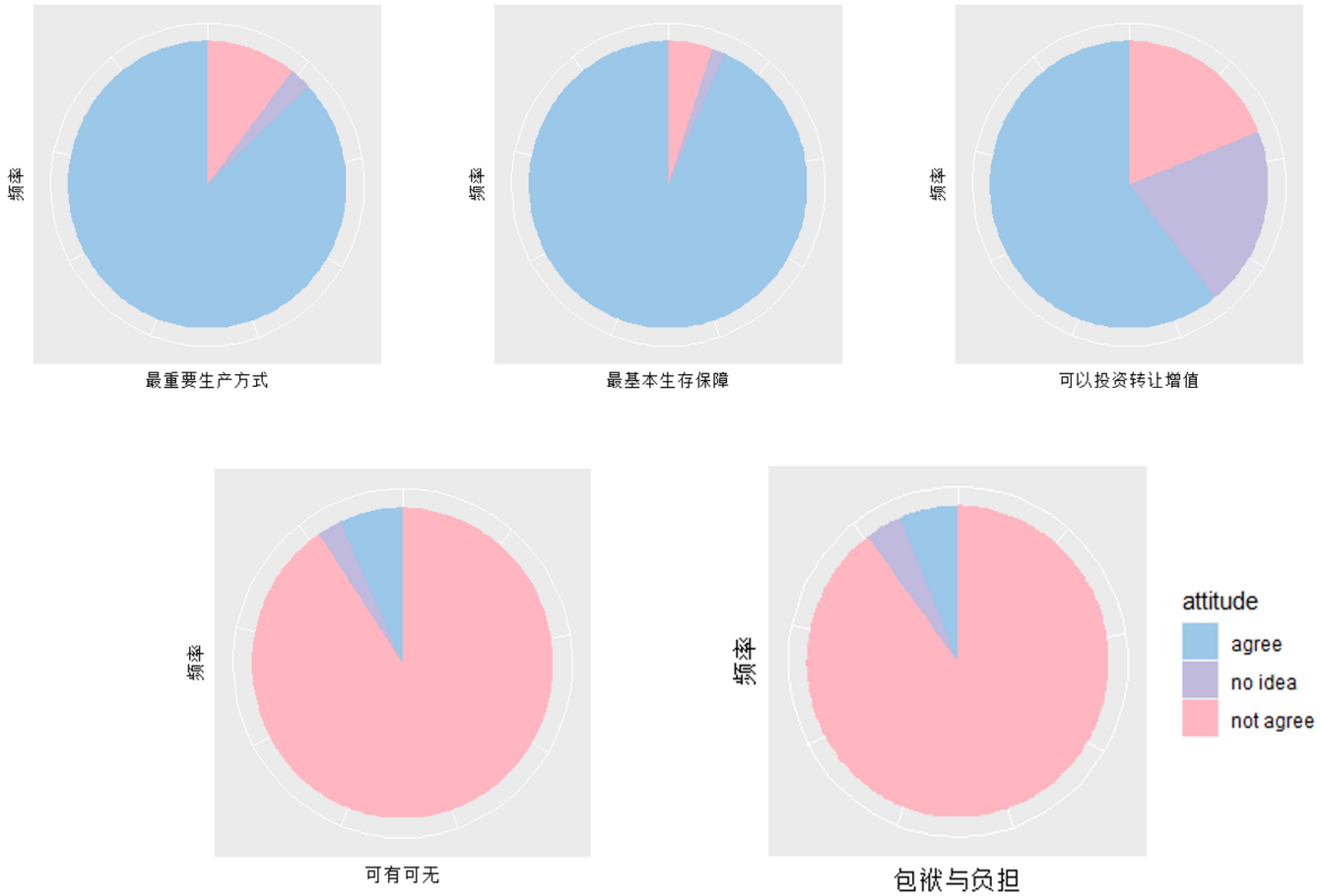


图 28: 比例对比示意图

示例 3 比例对比

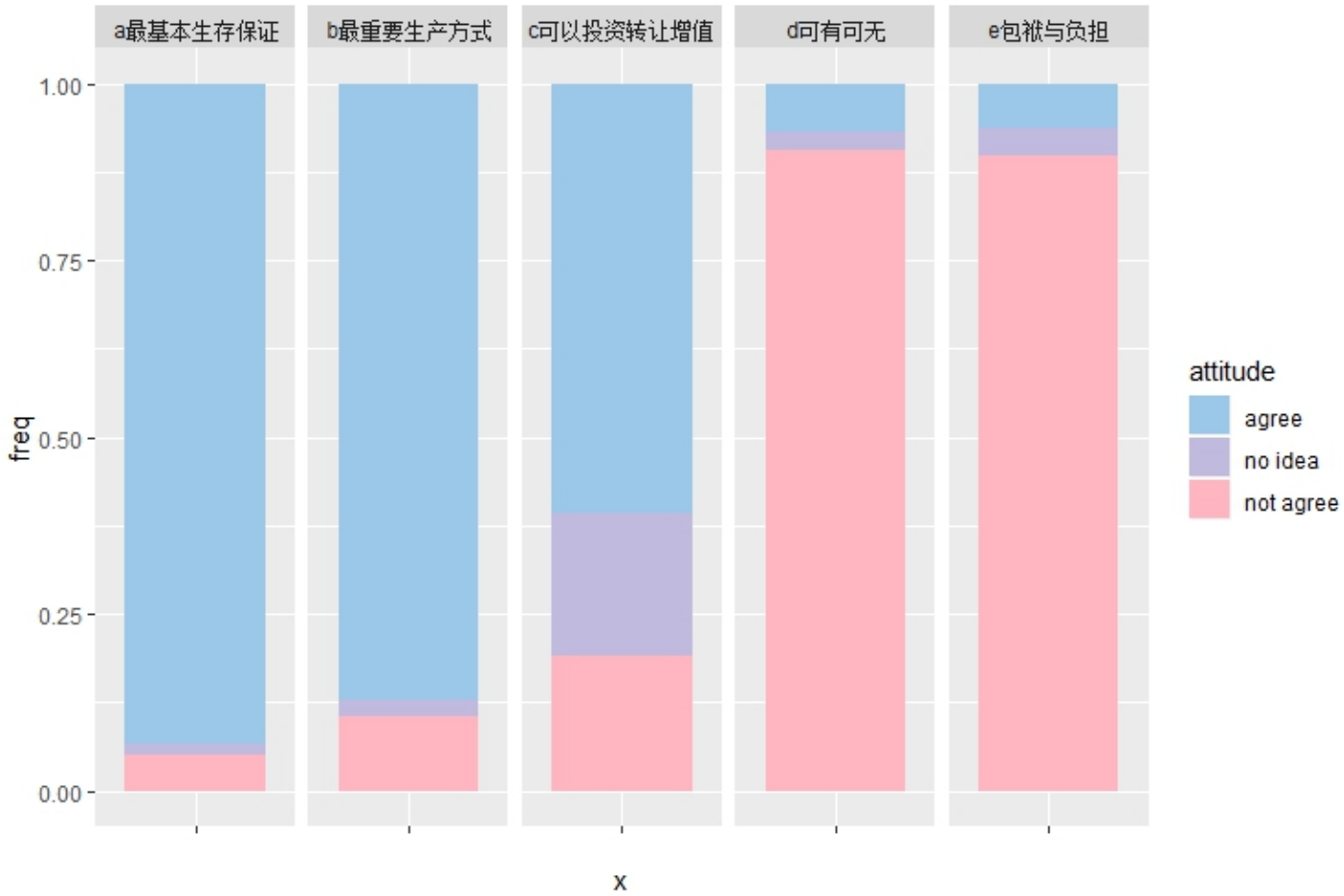


图 29: 比例对比示意图

示例 4 多组趋势比较

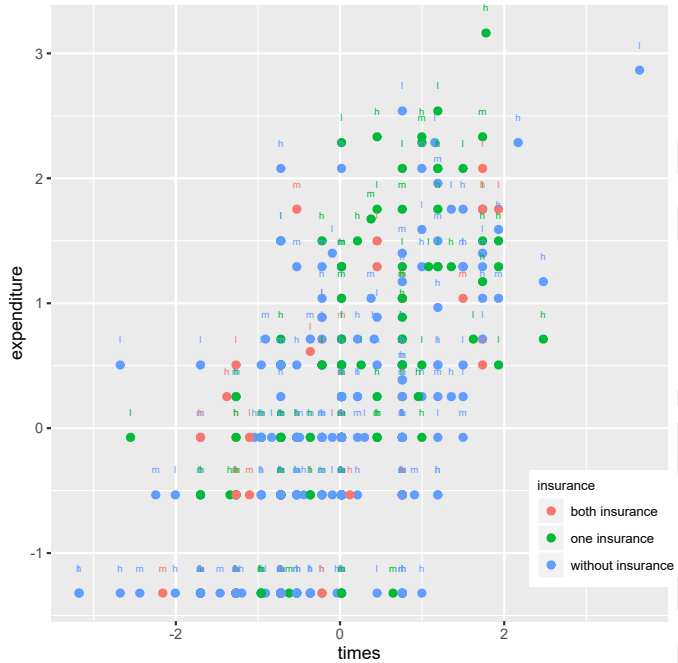


图 30: 二维多标签散点图

- x 轴表示门诊次数，y 轴表示门诊消费；标签 h/m/l 表示家庭收入分组；不同颜色点表示门诊与自我自疗保险使用情况

示例 5 地理分布

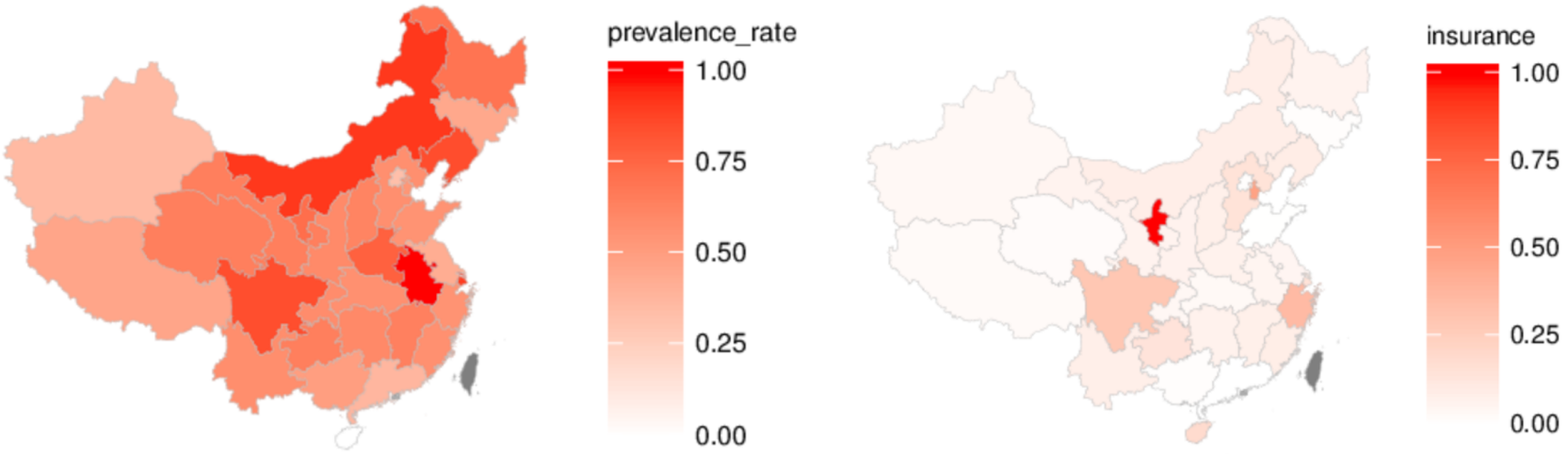


图 31: 左图为慢性疾病的分布，右图为平均保险花费的分布

示例 6 文本可视化

图 32: 文本词云图

示例 6 文本可视化

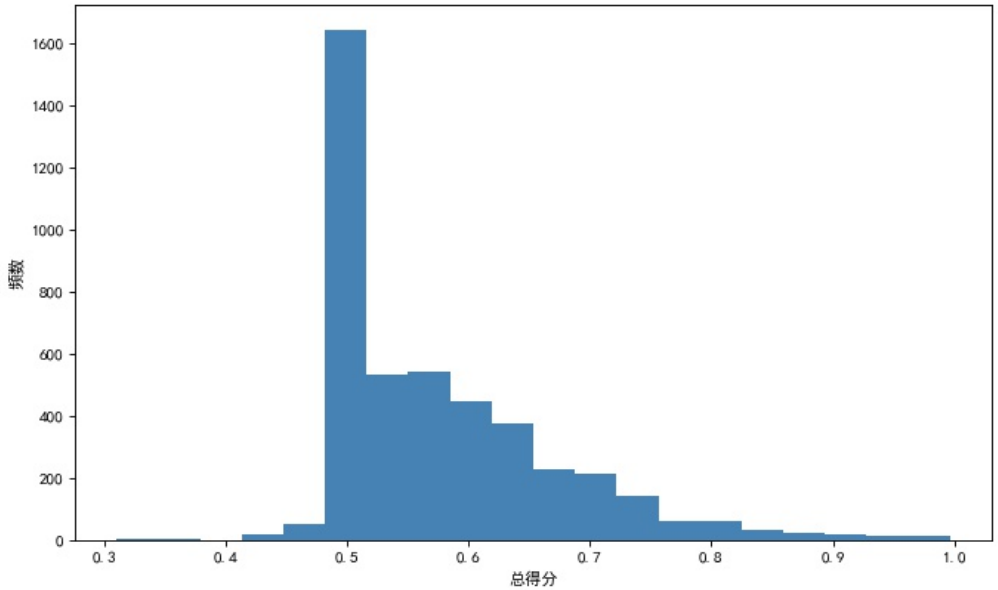


图 33: 总得分直方图

示例 6 文本可视化

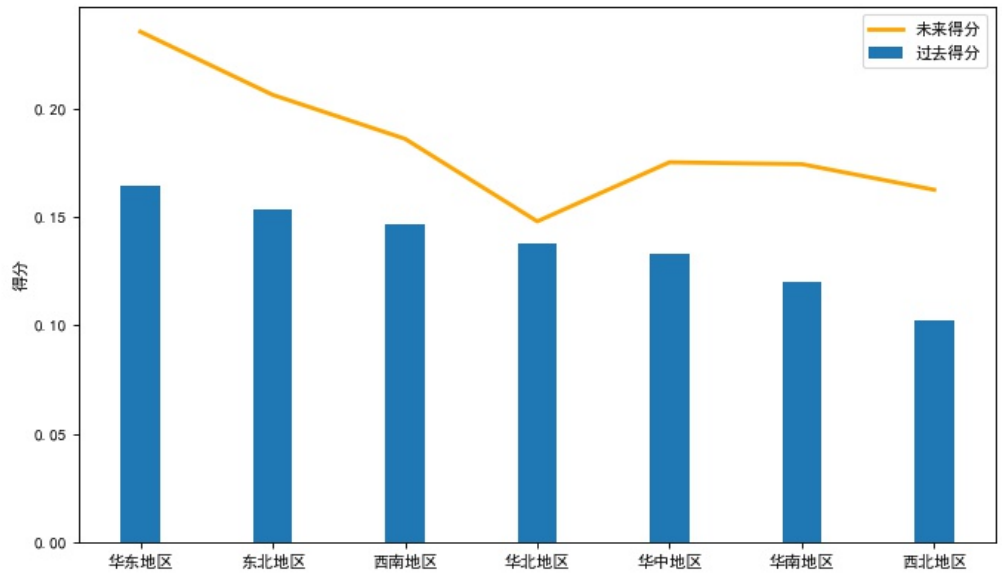


图 34: 情感得分的地区比较

示例 6 文本可视化

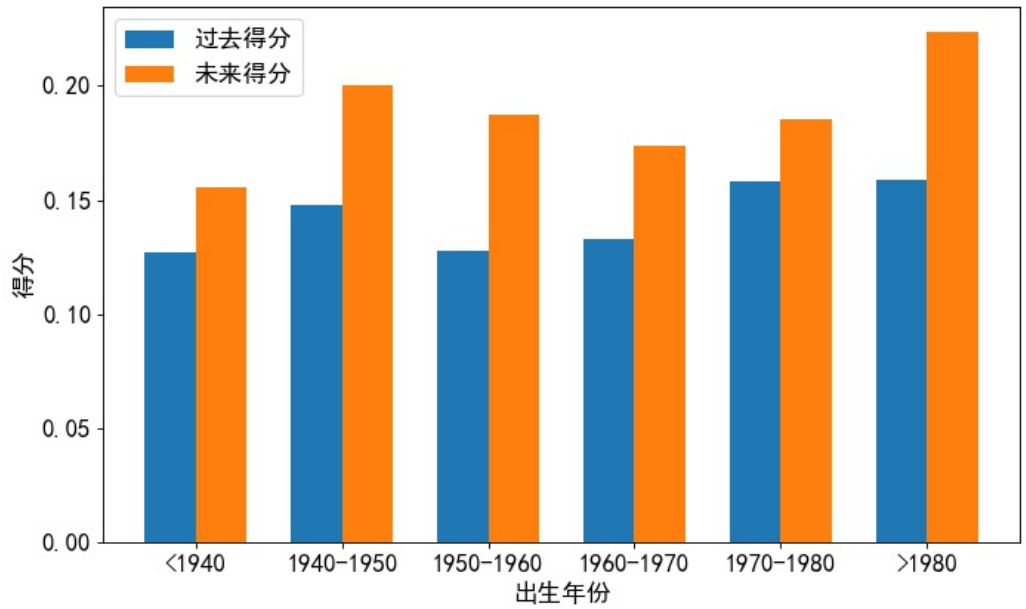
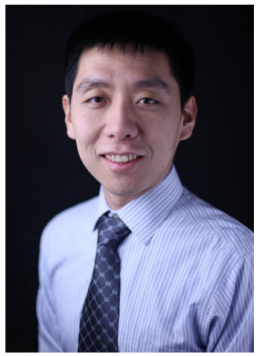


图 35: 情感得分的年龄比较

致谢


秦祎辰


王菲菲


孟珠峰


祁乐


杨昊宇


马伊莎


王彧


麻世钰

图 36: 感谢为本课件做出贡献的老师和同学们

中国人民大學
RENMIN UNIVERSITY OF CHINA

参考

- 《可视化分析》，中国人民大学出版社
- https://github.com/rucliyang/Visualized-Analysis

谢谢!