统计学原理(Statistic)

胡华平

西北农林科技大学

经济管理学院数量经济教研室

huhuaping01@hotmail.com

2022-03-26

第二章 数据收集、整理和清洗

2.3 资料整理和数据清洗

资料整理的流程

资料整理的记录

资料的检索

资料的安全

数据清洗的内容

数据清洗的记录

数据清洗的安全

数据清洗的操作

资料整理的流程

在数据收集过程中，重要的是条分缕析。

- 分类存储
  - 依据数据的载体类型、研究的时间需来进行分类，采用合适存放工具进行存放。
  - 纸版问卷，不能随便堆放，需要按照一定分类标准进行存放，便于后续工作。
- 建立目录
  - 存放的目的不仅仅只为了存储，更重要的是为了便于使用，建立目录就是便于利用的方式之一。
  - 目录是用于检索的，对调查获得数据建立目录，也是为了方便检索。
- 编制索引
  - 对于复杂数据，还需要在目录与存储之间建立关联，这就是索引

资料整理的记录

数据收集和整理中不仅需要核实，还需要记录。主要记录：

- 数据来源信息：
  - 如调查项目，调查人，采集人，采集时间，地点，对象。
- 数据载体类型信息：
  - 具体是什么载体？比如，纸张的、数字的。
- 数据描述信息：
  - 有多大规模，什么内容，关联什么主题，等等。
- 数据分类信息：
  - 无论是按照载体形态分类还是按照其他标准分类，一个大型项目需要对原始数据根据数据使用，建立基本分类。
- 数据存储信息：
  - 数据以什么样的载体，什么样的方式存储在什么位置？
  - 与数据安全相关的信息，如存储的版本、份数、时间变化关系等。

资料的安全

> “老师，能把上星期发给我的课件再发一遍吗？我忘记放到哪了？”
> “老师，非常的崩溃！电脑的硬盘坏了，写的东西都没有了！”

上述记录信息要尽可能的保存若干个版本。

- 纸和笔的传统版本。便于在需要的时候翻阅，尤其是使用范围相对较广的数据。
- 数字化可检索的版本。为什么要做数字化的可检索版本？
  - **目录树法**（相对简单的数据）
  - 建立专门的**数据库**（针对异常复杂或庞大的数据）

资料的安全

数字化数据有一些需要特别注意的问题：

- 数据存储。随时都有若干个备份！
  - 数字化的数据从最初的纸袋到今天的磁盘、硬盘，有各种介质。由于介质的可靠性不同，数据的安全性也不相同。
  - 美国的“911”事件。美国联储会的主席格林斯潘知道这个消息的第一时间，他担心的 不是“911”的伤亡情况，而是美国金融数据的安全。
- 数据安全。安全的风险，要么来自于使用者的误操作，要么来自于内部或者外部的有意攻击。
  - 离线保存的目的不仅仅是为了应对各种预想不到的不测，更重要的是为了防止数据泄露。
  - **斯诺登事件**：任何在线数据事实上都是不安全的，都有安全隐患。

资料的安全

文本数据的安全：

- 文本数据的安全威胁主要来自于不可抗力的一些因素，比如说自然灾害、风蚀等。
- 当然也来源于人为因素。比如说错误的识别，本来是很重要的数据，却被当作了废纸。

非数字化数据的安全：

- 图片数据的载体形态比较复杂，胶片、图片由于介质存储特征的差异，不可以混合放置而保管。如胶片就需要防潮，通常要使用防潮器皿。
- 实物数据的安全具有独特性，应根据实物实物特征进行科学整理和安全管理。比如说兵马俑，那就在兵马俑的原址上盖一个博物馆进行整理。

数据清洗的内容

数据整理主要是分类和梳理，**数据清洗**主要讨论的则是检错。通过这两部分的工作减少人为的误差，降低调查误差。数据清洗包括四个工作内容：

- **真实性评估**：确认数据是真实的，不是道听途说，不是张冠李戴，更不是杜撰臆想。
  - “假新闻”现象就是调查数据在真实性层面出现的问题。
  - 微信群里“令人发指”的各类长辈转发
- **完整性评估**：数据应与研究工作的目标要求相符，研究不需要的就不应该出现在数据中，研究需要的在数据中就不应该缺少。
  - 如果需要补值，就应继续补充收集数据。

数据清洗的内容

- **可用性评估：**数据是不是可以用于数据库化了？如果不能，还需要做怎样的数据加工？
  - 比如对图片数据、 音频、视频数据，甚至文本数据是不是还要做数字化工作。
  - 对于**痕迹数据**，尤其是大数据，如果不是直接采用大数据分析，而是应用于单机分析或服务器分析，是不是还要根据数据量进行抽样。
  - **脱敏化**处理。对有可能泄露受访者隐私、泄露传感器使用者隐私的部分，还需要做匿名化工作。
- **错误性评估：**评估数据可能的错误来源、可能的错误大小，及其对数据质量的影响。

（示例）调查问卷的数据清洗内容

以**调查问卷数据**的清洗为例：

- 真实性的清洗：要确认数据来自于受访者。
- 完整性的清洗：主要看样本无应答，也就是一整份问卷没有应答。以及选项无应答，也就是应该应答的访题没有应答。
- 可用性的清洗：主要是看编码是否完成，权数是否可行，以及缺失值如何标记和处理。
- 错误性的清洗：主要是清洗调查环节的错误，比如样本错误、应答人错误、应答方式错误。

数据清洗的记录

清洗数据工作中的每一项活动都要有记录。记录信息包括：

- 清洗工作的信息记录：
  - 数据清洗每一个步骤的做法、参与人、时间、地点、过程信息。
- 与清洗内容相关联的信息记录：
  - 数据真实性信息。比如是否真实？是否存在编造、作弊嫌疑？哪些部分存在不真实？怎么样不真实？。
  - 数据完整性信息。比如是否完整？是否有缺失？如果有缺失，哪些部分缺失？缺失哪些数据？
  - 数据可用性信息。比如问卷数据是否加权？痕迹数据是否数据化了？大数据如何处理？是运用云计算策略，还是裁剪为单机计算容量？
  - 数据错误性信息。比如问卷数据中的缺失，文献数据中的差错等。

示例：一次对原始数据的清洗记录！

EDA-xian-market.Rmd

D:/github/wp-covxian/EDA-xian-market.Rmd

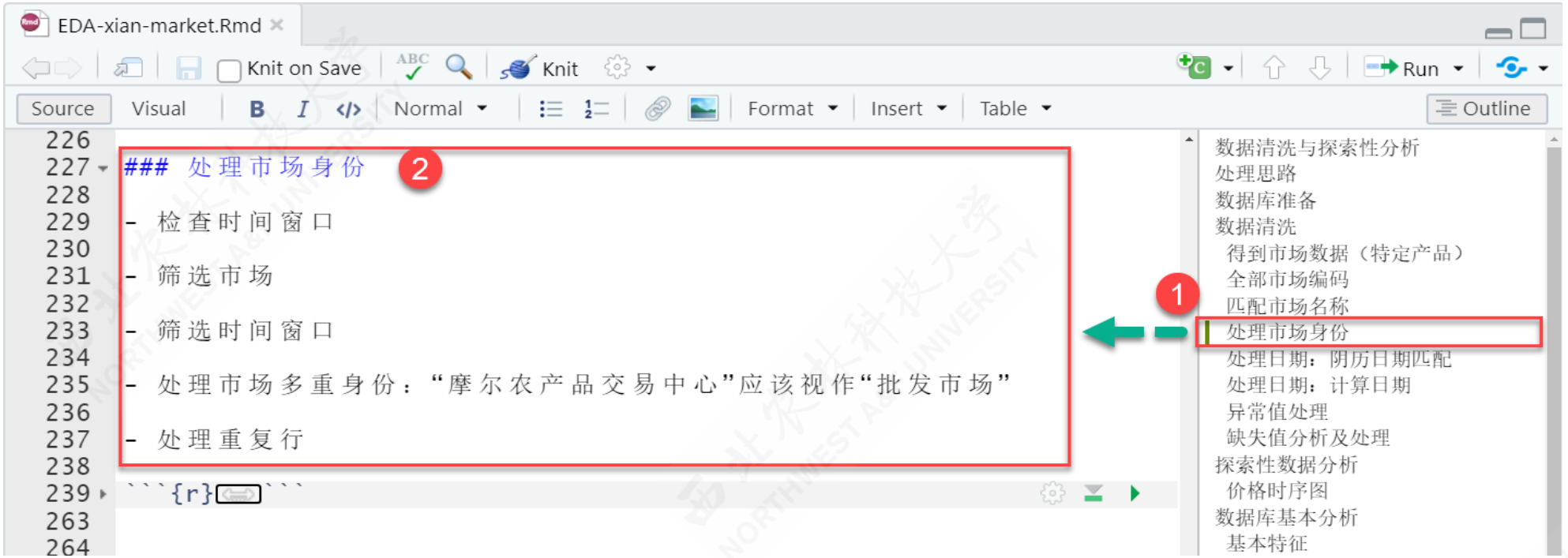
Knit | Run | Source | Visual | Normal | Format | Insert | Table | Outline

```
1  ---
2  title: "数据清洗与探索性分析"
3  subtitle: "西安市农产品价格监测数据"
4  author: "胡华平"
5  date: "`r Sys.Date()`"
6  params:
7    year_public: 2021
8    year_lead: 2020
9    year_report: 2019
10   year_base: 2018
11   year_public_char: "2021"
12   year_lead_char: "2020"
13   year_report_char: "2019"
14   year_base_char: "2018"
15   save_xlsx: false # options for saving ggplot to xlsx
16 knit: (function(inputFile, encoding) {
17   outFile <- sub(pattern = "(.*)\\..*$", replacement = "\\1",
     basename(inputFile));
18   out_dir <- 'public';
19   rmarkdown::render(inputFile,
20                     encoding=encoding,
21                      output_file=file.path(dirname(inputFile),
     out_dir, outFile)) })
```

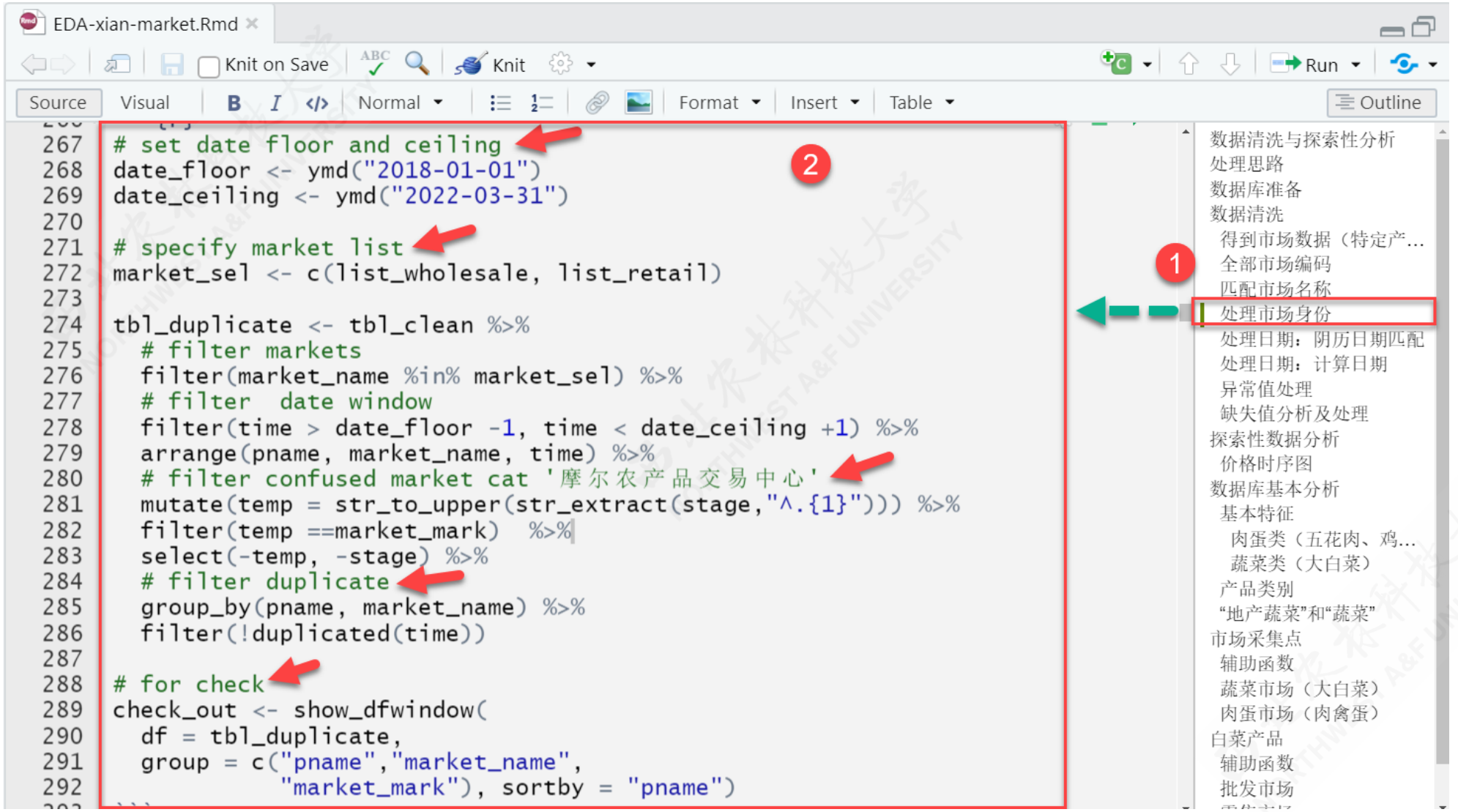
1

2

示例：一次对原始数据的清洗记录2



示例：一次对原始数据的清洗记录3



数据清洗的安全：笔记清洗

数据清洗的记录信息应尽可能地保留若干不同的版本。一般包括纸笔版本和数字化版本。纸笔版本便于随时翻阅，数字化版本，便于交流，也便于检索。

- **笔记的清洗**。不管是哪一类的笔记，所有的笔记都有私用和公用之别，通常人们做笔记都是做给自己看的（**私用笔记**）。
  - 你把自己的笔记给别人看，别人能看懂吗？
  - 在正式使用之前，需要把笔记数据通过清洗，变成任何使用者都可读的笔记（**公用笔记**），
  - 这就是格式化问题，就是把你个人的笔记清洗为数据笔记。

数据清洗的安全：音视频清洗

- **对音频要抄录：**
  - 把语音文档，不管是磁带录音，还是数字录音，抄录为文字，表述为文字或者文字加图片这样的格式。
  - 数字录音还有一个**格式清洗**问题，不同数字设备的录音，可能会 采用不同的格式。
  - 比如olympus的早期设备，采用的就是它自己的格式，DSS格式；如果不是采用它自己的软件就读不出来，最好呢，是转化为通用的格式，比如mp3格式。
- **对视频清洗编码：**
  - 如果是非数字录像，最好先转化为数字格式
  - 如果已经是数字录像，对视频清洗编码需要给出**时间记录码**。

数据清洗的安全：几点忠告

> 哦，已经数字化了，可以扔了，那个没用了，可以扔了。

- 不要轻易地丢弃任何一段看起来没有用处的信息，信息载体。
- 清洗不是仍东西，是清洗数据，让数据清晰化。
- 清洗的目的就是将特异性的数据，转化为公共性的数据、分析研究者都可以读的数据。
- 在清洗的过程中，千万要保留原始观察记录。
  - 一般而言，原始问卷至少要保留十年以上，访谈记录和观察笔记一般要求永久保留。

数据清洗操作：观测性数据

以**观察性研究**中数据的清洗为例：

- 观察性数据有一个特点就是差异性，对同一个场景、同一个事件，不同的人去观察，看到的并非完全一致。
- 每个人的观察记录，都有自己的习惯，有的习惯于采用速写和密写，比如说有些人为了防止别人看他的笔记，长采用密写的方式。即使是结构式的观察，不同的观察者也会有特异性。
- 观察性数据的清洗就需要把各类个性化的个人观察数据转变为标准化的观察记录。

数据清洗操作：文献数据

以**文献数据**的清洗为例：

- 笔记的清洗。比如说：研究用的素材如文献的阅读、标注与笔记、摘录，如果希望未来继续使用，那就需要**格式化清洗**，把素材转化为数据。如果有必要，还可以为下一步的数据库化做准备，比如编码。
- 文献的清洗。对阅读过的文献，如果已经获得了数字版本，就需要与数字版本关联的编目信息、阅读信息关联起来整理，结合后边讨论的数据库化工作，把它们转化为个人档案馆。如果没有数字化的版本，则需要将文献信息与阅读笔记信息关联，结合后边讨论的数据库化工作，把它们变成个人的档案阅读目录数据馆。

数据清洗操作：痕迹数据

对痕迹数据的“四性”评估和清洗，一般是直接依据数据的来源来确认的。比如，来自于**网络爬取**的数据，和来自于**数据拥有者机构**提供的数据，其它的**平行数据**等等。

一般而言，如果数据来源的渠道没有问题，数据的四性就不会有太大的问题。
清洗痕迹数据最重要的一项工作，就是把**非格式化数据** 清洗为**格式化数据**
（Why？至少目前的分析工具还不支持直接分析非格式化的数据）

数据格式化：把混杂在一堆数据中的各类数据清洗出来，分门别类。比如说日志数据中的用户行为数据，以淘宝数据为例，订单数据、发单数据、物流数据等等，分门别类整理出来。

数据结构化：把各类数据和变量进行多维度关联。比如把以上日志数据中的各个子集关联到用户之下，形成类似于问卷调查数据的每个**样本数据**。

数据清洗操作：大数据

如果**痕迹数据**是**大数据**，情况就有些不同了。

在清洗数据之前，需要把**清洗策略**测试一遍，然后就可以直接采用大数据的清洗模式了。

- 从大数据中抽取数据，或者是从网页上爬取数据，在处理中尽管不一定会用到云计算，在处理逻辑上还是一致的。
- 大数据的清洗，目前运用比较普遍的是Hadoop框架下的Map Reduce。

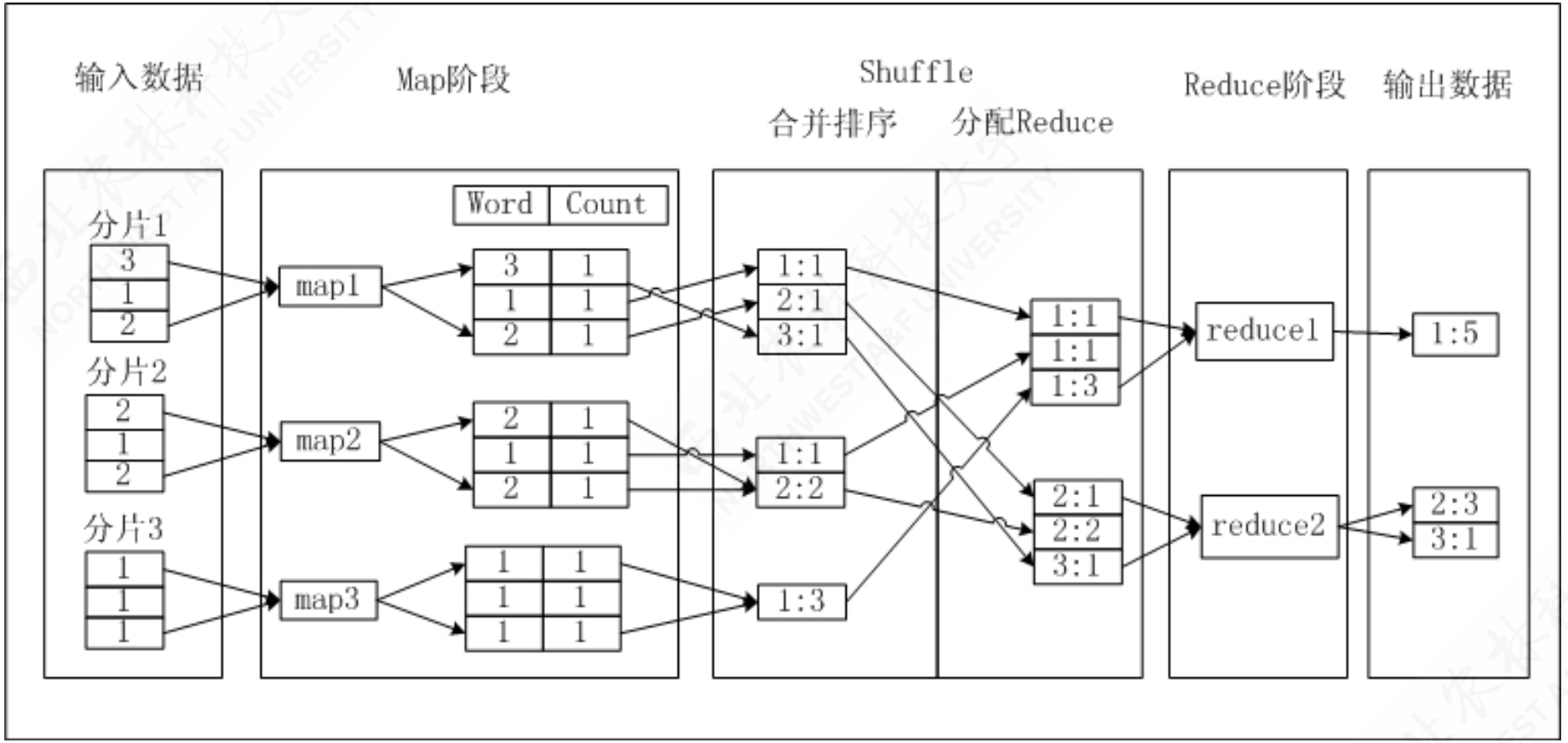
（示例）阿里巴巴的大数据清洗1/2

阿里巴巴有淘宝、天猫、一淘等等业务，这些业务每时每刻都在产生数据，这些数据涉及到信用、金融、物流、管理等等业务操作。 所有这些操作的数据都会汇集到数据交换平台，由此构成了阿里巴巴的数据动态。

2014年的双十一期间，6个小时之内的处理量就已经达到了100个PB。在产生的这些数据中，既有结构化的数据，也有非结构化的数据，进出数据平台的数据不是个，不是匹，而是流。这些数据流，通过数据处理就变成了中间层的数据，可以运用和应用于服务，中间服务，既可以对内，又可以对外。

（示例）阿里巴巴的大数据清洗2/2

问题是，这些数据是怎么处理的呢？数据清洗关心的正是这个问题。



（示例）CGSS数据的清洗：介绍

```
Warning in library(package, lib.loc = lib.loc, character.only = TRUE,
logical.return = TRUE, : 不存在叫'sjmisc'这个名字的程辑包
```

中国综合社会调查（China General Society Survey，**CGSS**）：

- CGSS2013, **CGSS2015**
- 属于**混合截面数据**：也即不同年份的观测单位不是固定的。
- **CGSS2015**一共有样本数10968，变量总数有1398

（示例）CGSS数据的清洗：数据视图

随机抽取300份样本的前20个变量。CGSS2015数据（样本数=10968）数据如下：

id	s1	s41	s42	s43	s44	s45	token	a0	a11011	a11012
178	1	4	7	20	34	61	11058189315255	2015-10-17 17:07:00	妻子	41
192	1	4	7	20	35	63	11062104543903	2015-10-17 15:51:00	妻子	61
232	1	4	7	17	30	55	11080923036905	2015-10-03 10:36:00	妻子	63

Showing 1 to 3 of 300 entries

Previous 1 2 3 4 5 … 100 Next

（示例）CGSS数据的清洗：变量视图(全景)

CGSS2015变量体系（变量数=1398）

题号	变量名	题项	变量类型	是否标签
1	id	问卷编号	double	false
2	s1	样本类型	double	true
3	s41	采访地点-省/自治区/直辖市编码	double	true
4	s42	采访地点-地级市编码	double	false
5	s43	采访地点-县/区编码	double	false

Showing 1 to 5 of 1,398 entries

Previous 1 2 3 4 5 … 280 Next

（示例）CGSS数据的清洗：变量视图(局部)

含有“收入”的变量（变量数=20）

题号	变量名	题项	标签
222	a8a	您个人去年全年的总收入	c(无法回答 = -8, 拒绝回答 = -3, 不知道 = -2, 不适用 = -1, 个人全年总收入高于百万位数 = 9999996)

（示例）CGSS数据的清洗：缺失值I

挑选出如下几个变量来观测：

题号	变量名	题项	标签
320	a5606	您在最近三个月内采取过以下哪些方式寻找工作-为自己经营做准备	c(无法回答 = -8, 拒绝回答 = -3, 不知道 = -2, 不适用 = -1, 否 = 0, 是 = 1)
457	b8b	您认为每月户平均收入高于多少元就属于富裕户了	c(拒绝回答 = -3, 不知道 = -2)

（示例）CGSS数据的清洗：缺失值2

挑选出如下几个变量来观测。CGSS2015回答情况一瞥(随机40个样本):

id	b8b	b1011	a5606
15924	10000	2	0
11179	4000	4	0
8053	-2	-8	0
1248	20000	-8	0
10092	10000	-8	
12562	3000	-8	
11713	10000	-8	
10653	5000	4	

Showing 1 to 8 of 40 entries

Previous 1 2 3 4 5 Next

（示例）CGSS数据的清洗：变量处理

变量重新命名前

id	a36	a10	a8a
63	4	1	20000
543	4	1	150000
675	5	4	43000
883	4	2	100000
1084	1	1	50000
1352	4	1	600000
2110	4	1	20000
2403	5	1	3000

Showing 1 to 8 of 100 entries

Previous 1 2 3 4 5 … 13 Next

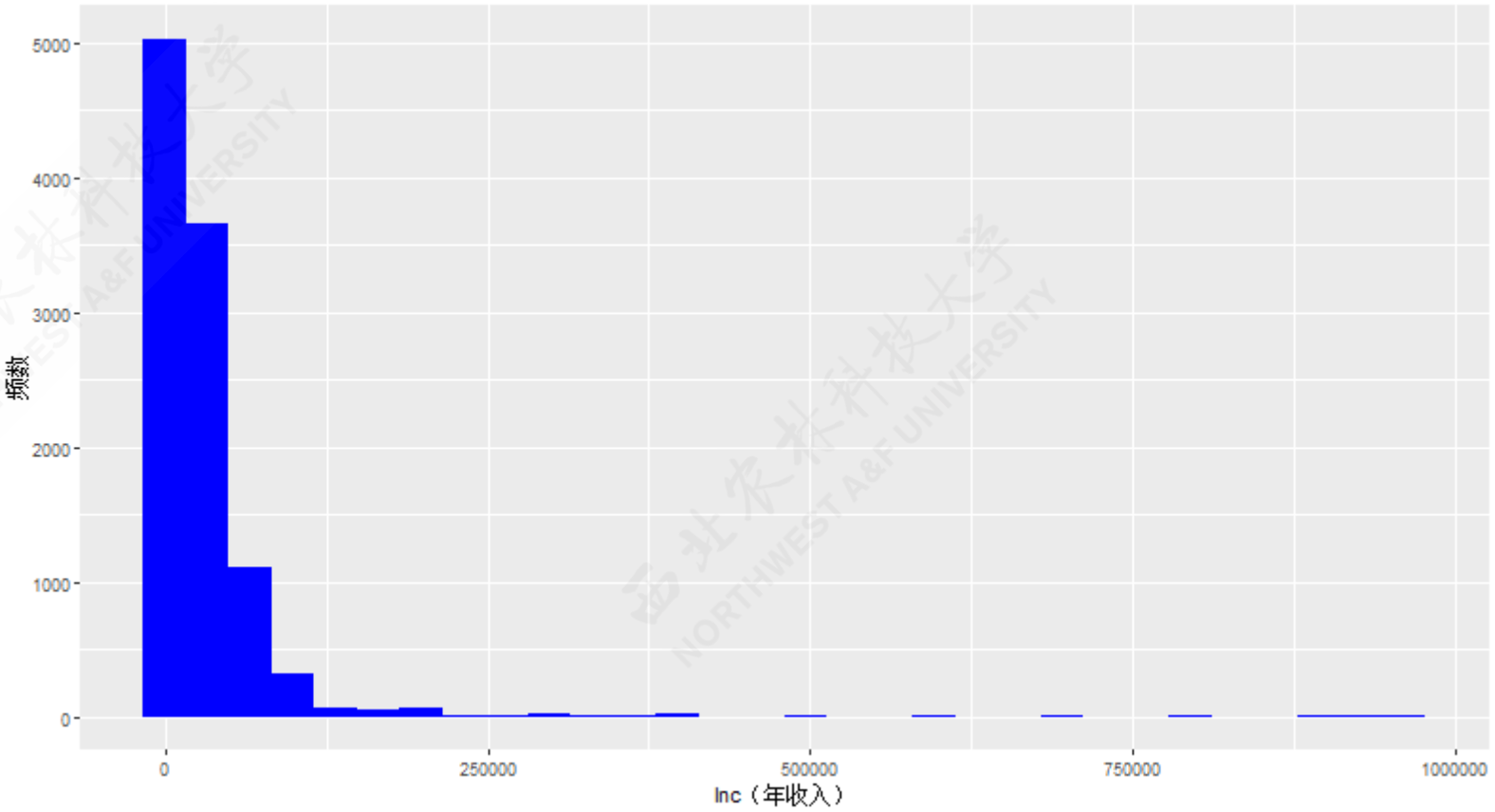
变量重新命名后

id	Happ	Pol	Inc	Lninc
63	4	1	20000	9.9
543	4	1	150000	11.92
675	5	4	43000	10.67
883	4	2	100000	11.51
1084	1	1	50000	10.82
1352	4	1	600000	13.3
2110	4	1	20000	9.9
2403	5	1	3000	8.01

Showing 1 to 8 of 100 entries

Previous 1 2 3 4 5 … 13 Next

（示例）CGSS数据的清洗：异常值处理前



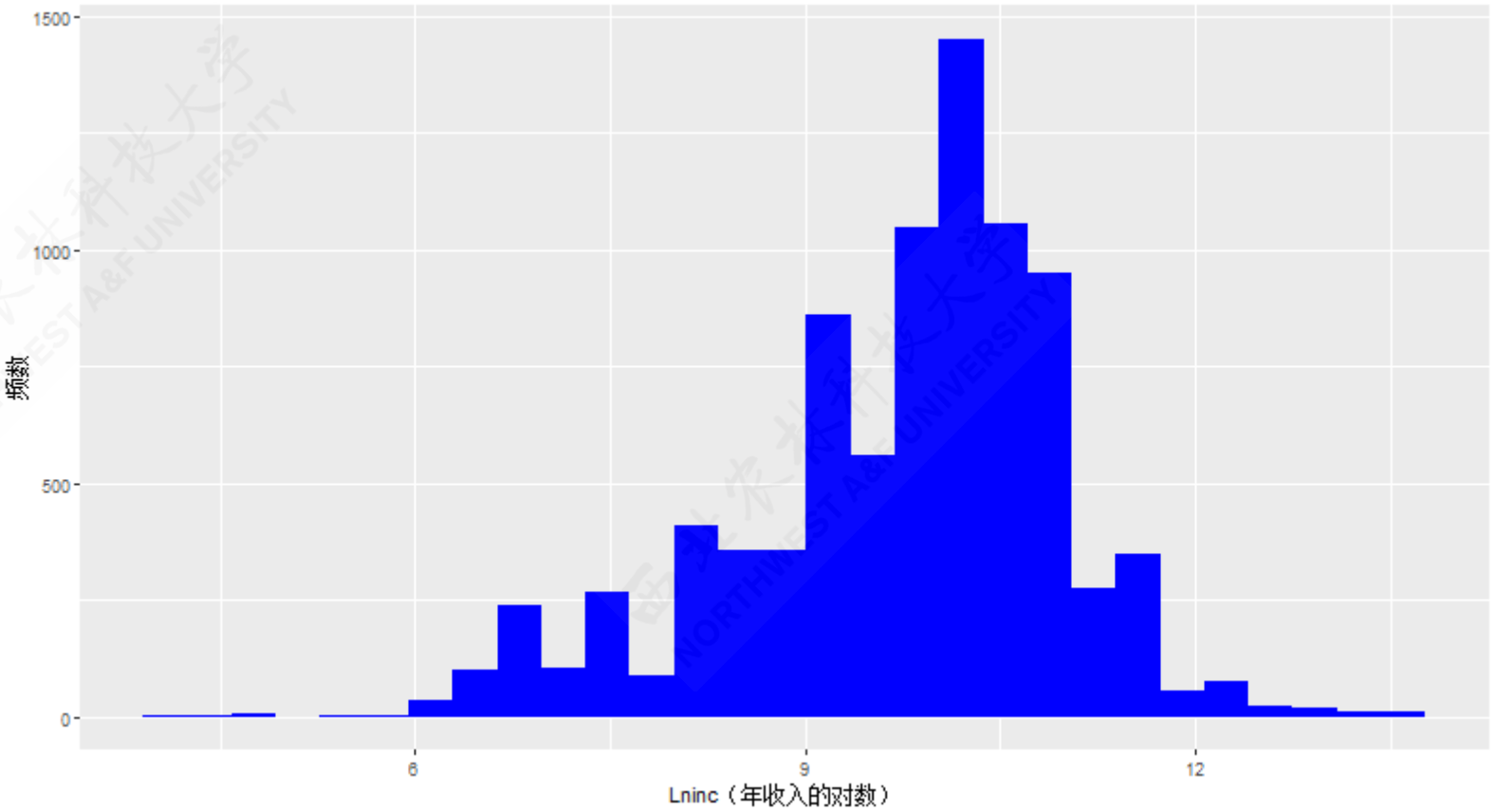
年收入的直方图

（示例）CGSS数据的清洗：异常值处理办法

异常值的处理办法：

- 截尾：比如截掉大于0.99分位数的观测值。
- 数据的转换：对于右偏分布较严重的变量，即右侧异常值较多，**自然对数**（ln()）可以使其更加对称。比如，年收入及其对数的分布。

（示例）CGSS数据的清洗：异常值对数化处理后



年收入对数的直方图

本节结束