Data Analysis Tools and Practice(Using R)

2021.06.15

课程总结

北京大学 软件与微电子学院
School of Software and Microelectronics, Peking University

Huiping Sun(孙惠平)
sunhp@ss.pku.edu.cn

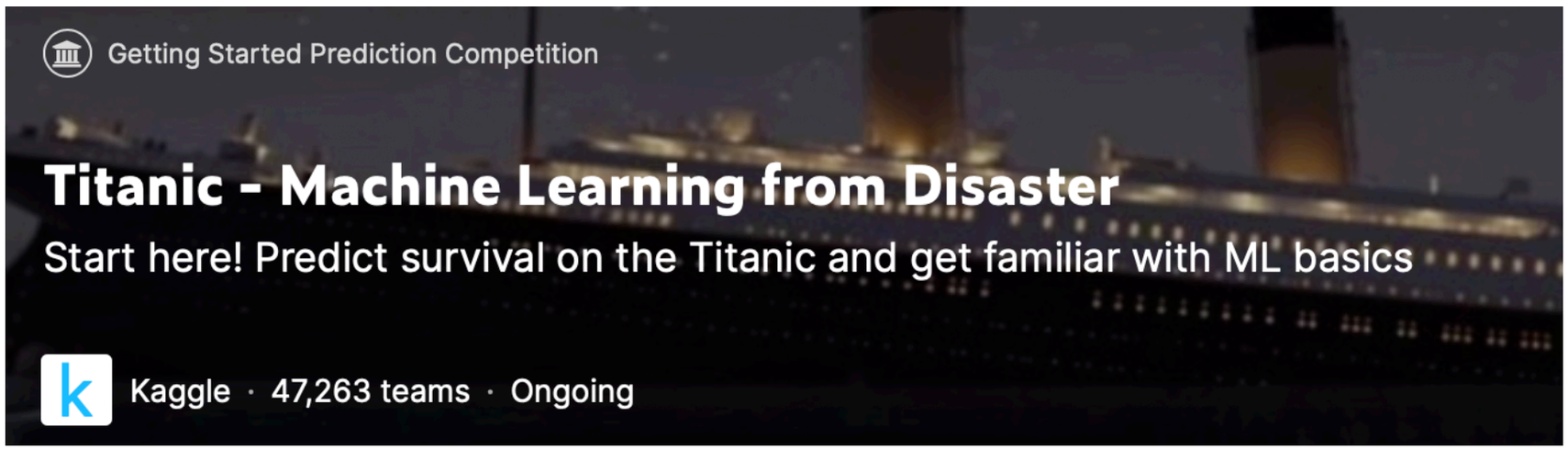
课堂测试时间

课堂测试14

先用电脑完成
40分钟 然后誊抄纸上

- 泰坦尼克号数据库，见**titanic.zip**

- 对数据进行预处理，包括查看统计特征、插补空值等操作
- 可视化分析得出初步结论
- 利用决策树进行建模，并且作出分析 caret
- 修正模型并利用ROC曲线检验模型的准确度如何 ROCR



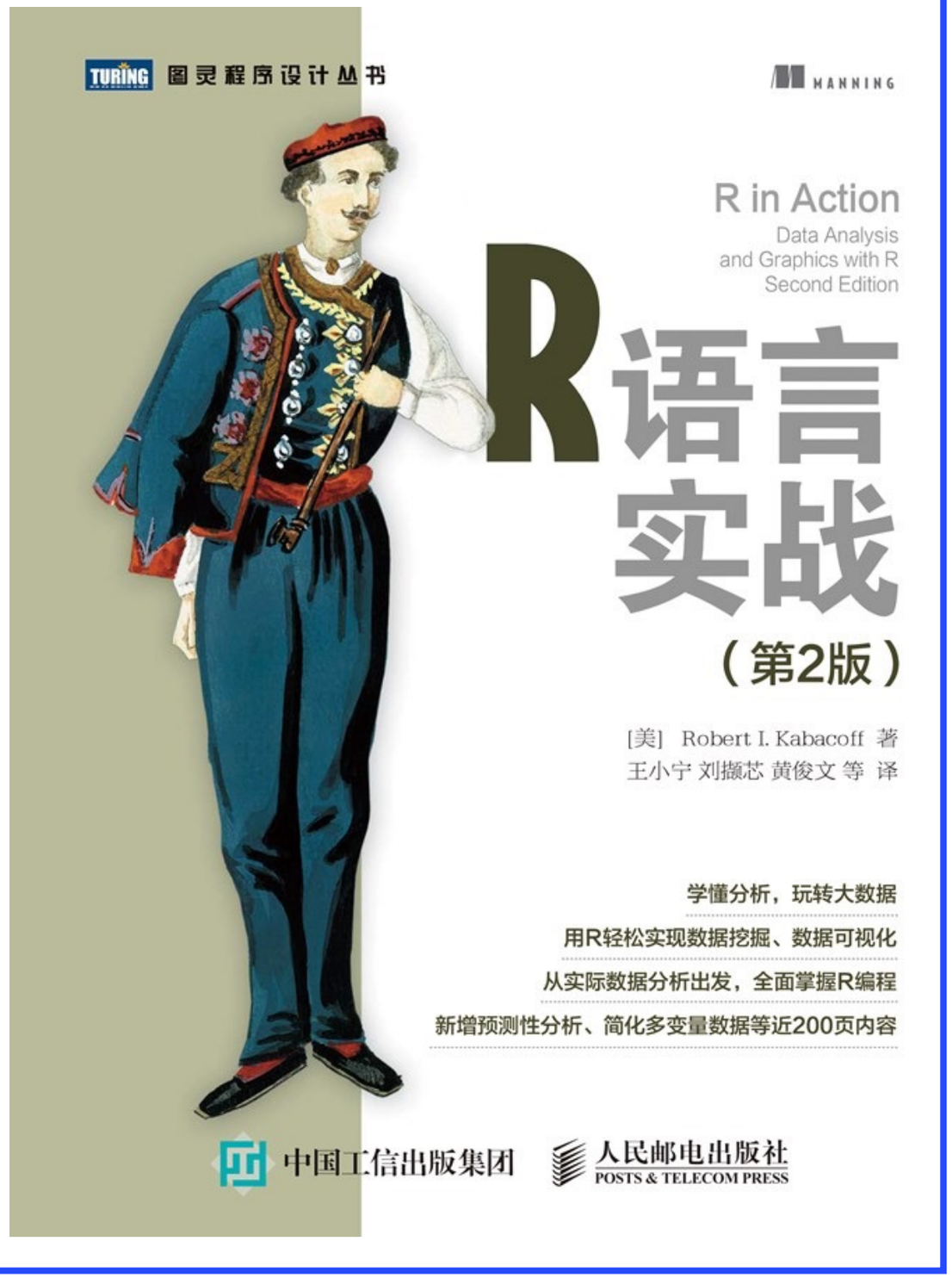
检索

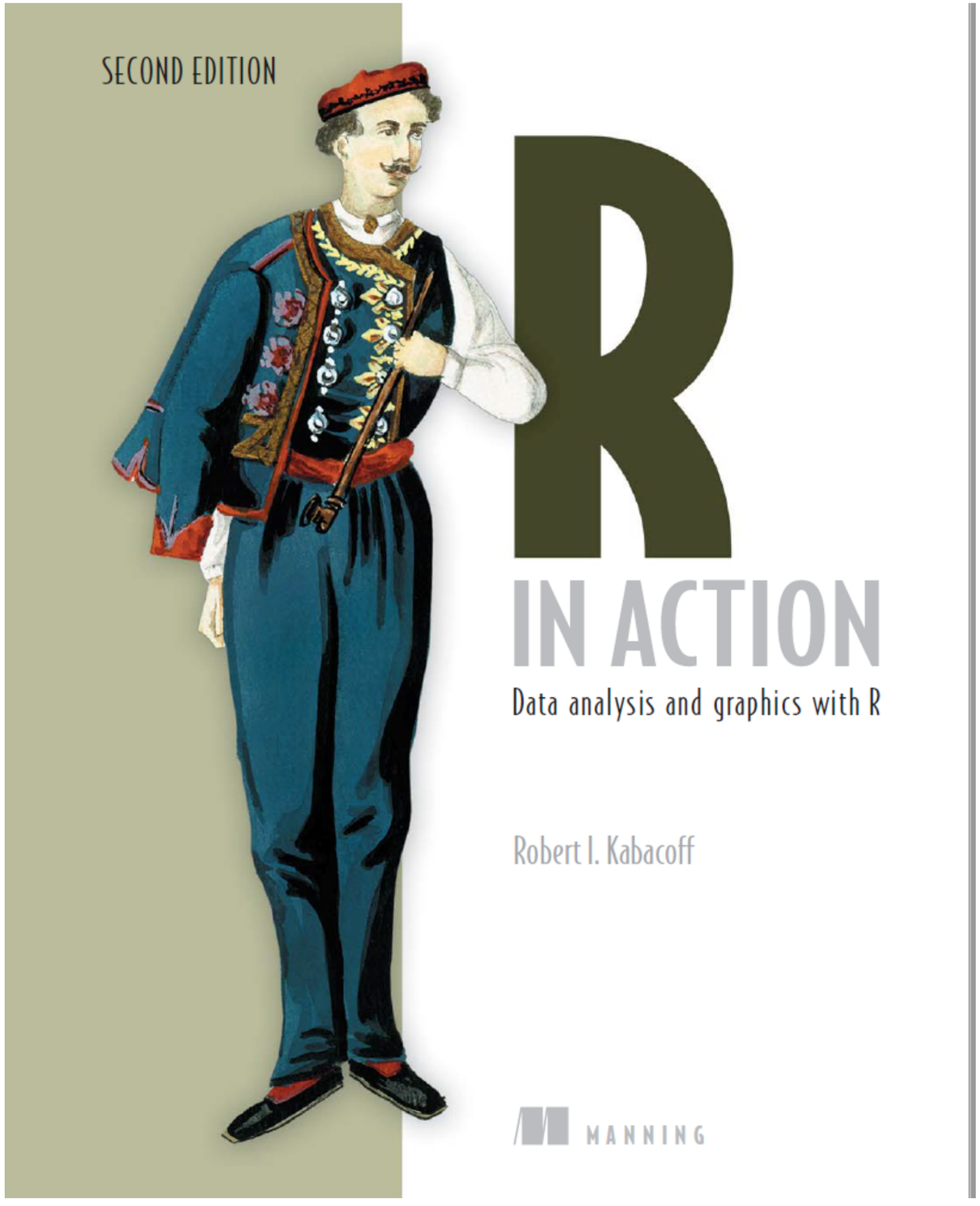
https://www.kaggle.com/c/titanic/

课程教材

Course Summary

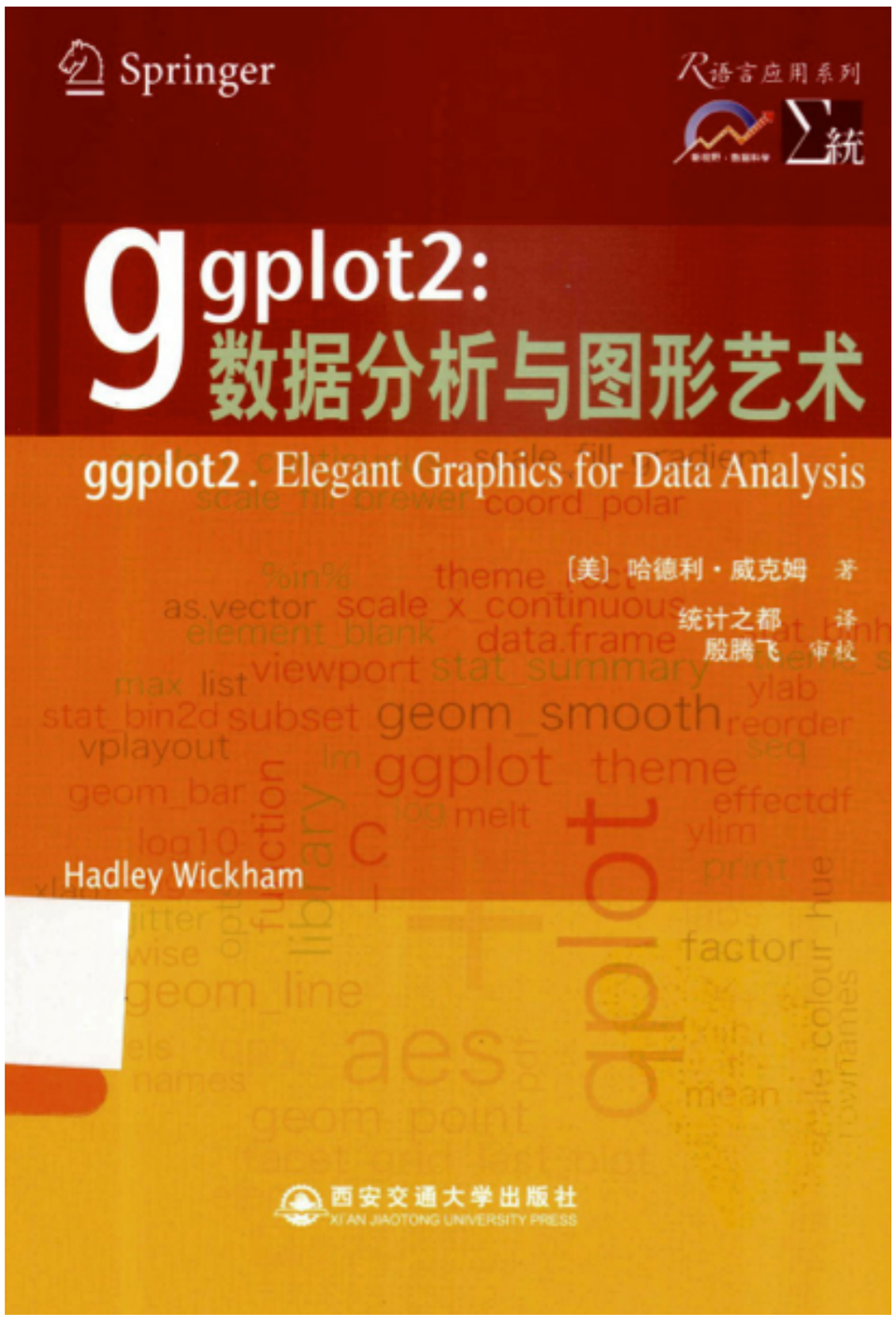
课程教材-主教材

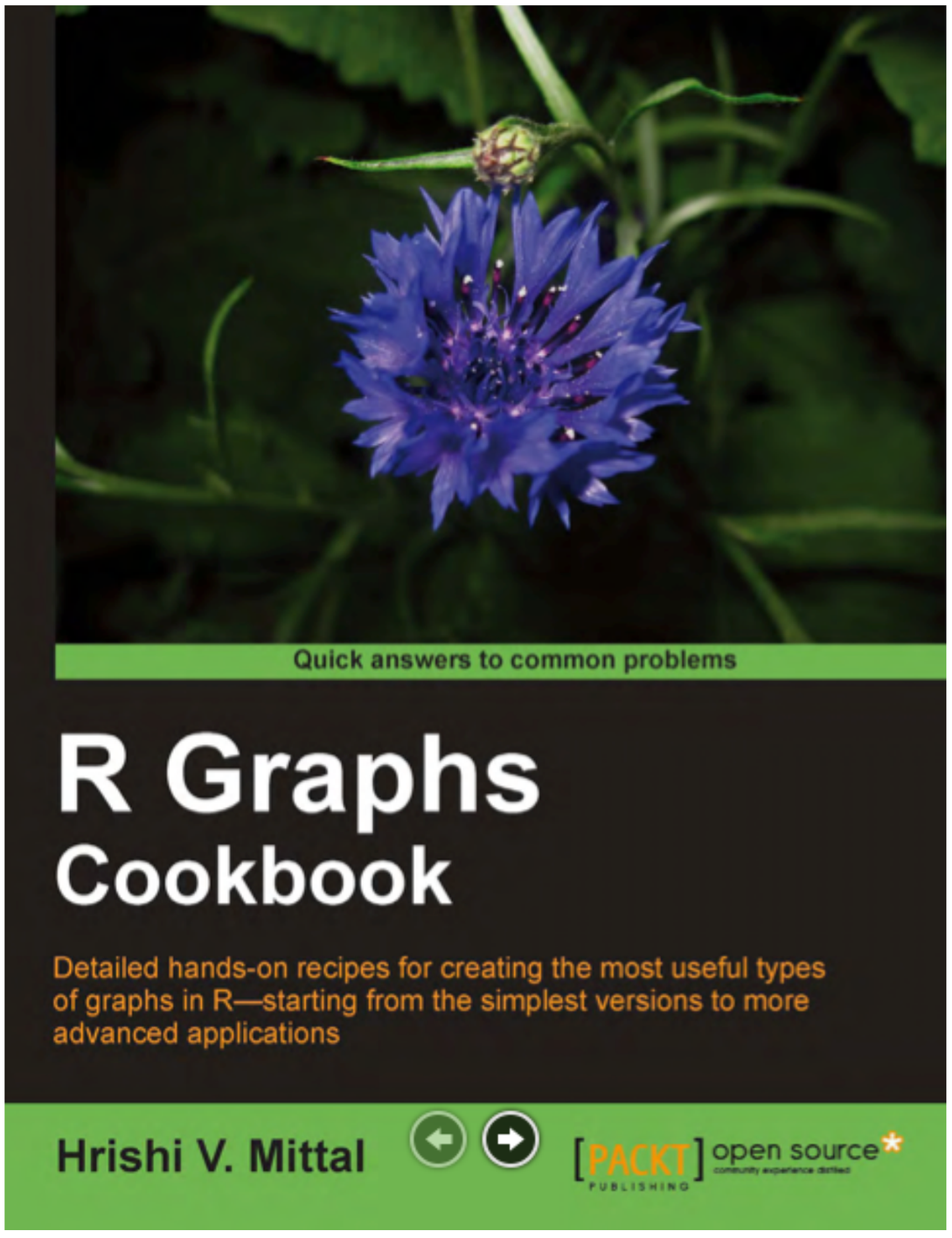




Course Summary

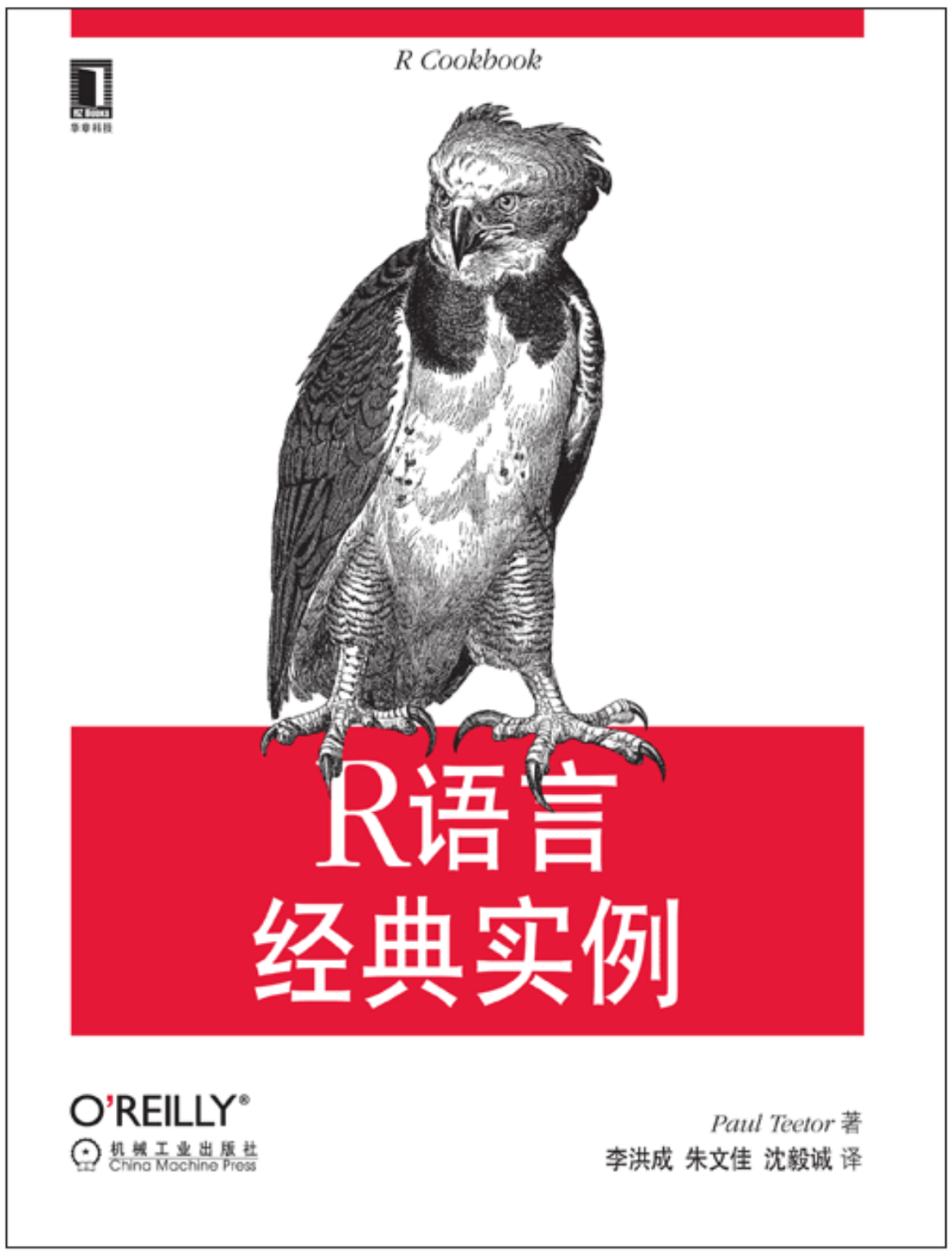
课程参考书-图形





Course Summary

课程教材-问题导向



课程教材-其余

O'REILLY®

TURING 图灵程序设计丛书

全彩印刷

R数据科学

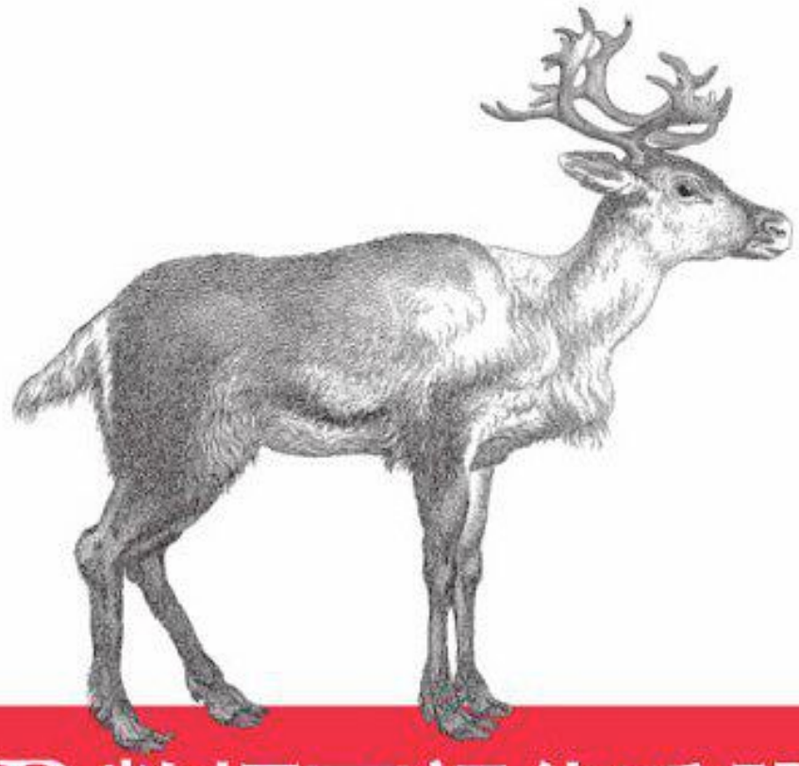
R for Data Science

摒弃其他R语言工具书从头到尾讲统计的陋习
从实用的R包出发，带你重新认识R和数据科学

[新西兰] 哈德利·威克姆 [美] 加勒特·格罗勒芒德 著
陈光欣 译

中国工信出版集团 人民邮电出版社 POSTS & TELECOM PRESS

R Graphics Cookbook



R数据可视化手册

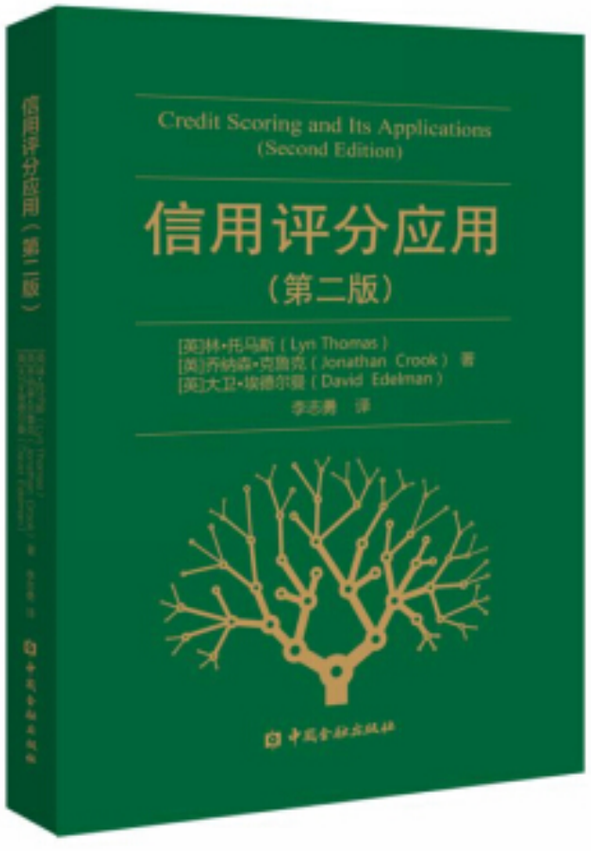
[美] Winston Chang 著
肖楠 邓一硕 魏太云 译
邱怡轩 审校

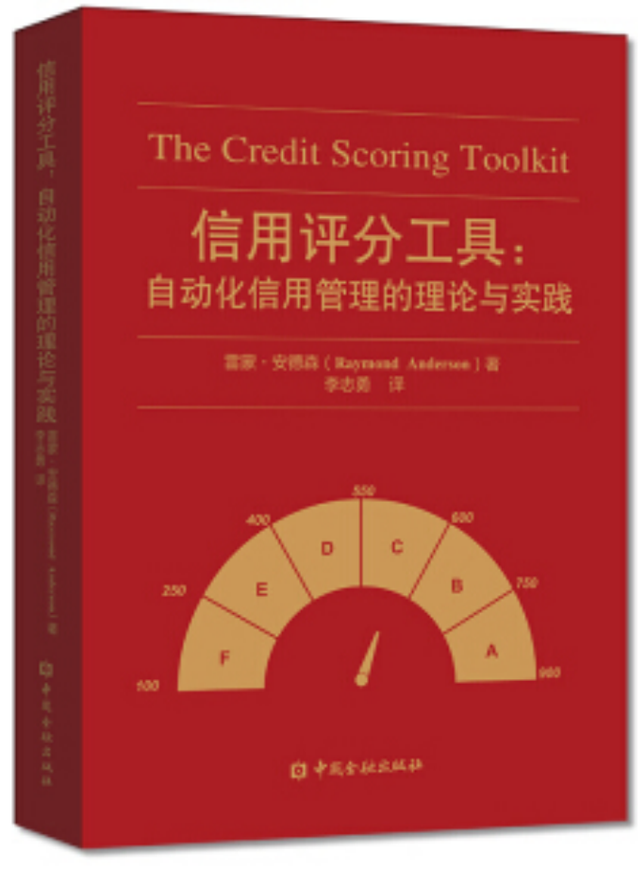
O'REILLY®

人民邮电出版社 POSTS & TELECOM PRESS

Course Summary

课程教材-信用评分





课程内容

Course Summary

课程内容

- 00: 课程简介
- 01: R的用户界面
- 02: R的数据对象1
- 03: R的数据对象2
- 04: R的语言语法1
- 05: R的语言语法2
- 06: R的基本图形1
- 07: R的基本图形2
- 08: 大作业中期汇报
- 09: R的统计分析1
- 10: R的统计分析2
- 11: R的统计分析3
- 12: ggplot2绘图1
- 13: ggplot2绘图2
- 14: R的一些扩展包1
- 15: R的一些扩展包2
- 16: R的一些扩展包3
- 17: 大作业最终汇报

Course Summary

简介

- 课程简介:
- 数据分析简介: 数据; 数据分析; 过程;目的; 数据展现; 数据分析报告;
- *R*软件简介: 历史; 优点; 安装; *R Gui*; *RStudio*;
- *R*基本使用: 解释性语言; >; <-; #, 表达式; 区分大小写;
- *demo*: *graphics; image; Hershey; persp;*
- 包函数: *install.package(); library(); search(); update.package();*
- 工作空间函数: *getwd(); setwd(); history(); savehistory(); loadhistory(); save.image();*
- 输入输出函数: *source(); sink(); dev.off(); pdf(); png(); jpeg(); bmp();*
- 帮助函数: *help.start(); help(); ?; example(); data();*
- *R*软件操作: *DA01.R*

c(); mean(); sd(); cor(); plot(); map(); rnorm(); density(); option(); summary(); hist(); runif();

Course Summary

数据-01

- 数据结构定义: *c(); matrix(); array(); data.frame(); factor(); list();*
- 数据结构访问: 下标; 下标向量; 逻辑向量; 负下标;
- 向量: :; *seq(); rep();*
- 算术运算符: +; -; *; /; **; ^; %%; %/%;
- 逻辑运算: >; <; >=; <=; ==; !==; !; |; &; ***isTRUE()***; *identical(); any(); all();*
- 属性函数: *length(); dim(); class(); names(); head(); tail();*
- 排序函数: *order(); sort(); sort.list(); which(); which.max(); which.min();*
- 运算函数: *max(); min(); range(); sum(); prod(); sqrt(); abs();*
- 类型函数: *is.numeric(); is.integer(); is.logical(); is.character(); as.xxxx();*
- 其余函数: *attach(); detach(); with(); $; t(); diag(); solve(); eigen();*

Course Summary

数据-02

- 矩阵运算: *t(); det(); array(); crossprod(); tcrossprod(); diag(); solve(); eigen();*
- 缺失值: *NA; is.na(); na.rm = TRUE; na.omit();*
- 类型函数: *is.numeric(); is.integer(); is.logical(); is.character(); as.xxxx()*
- 字符处理： *nchar(); substr(); strsplit(); toupper(); tolower(); paste();*
- 日期和时间： *Sys.Date(); date(); difftime(); format(); as.Date(); %d, %a,%A, %m, %b, %B, %y,%Y;*
- 统计函数： *mean(); median(); sd(); var(); max(); min(); range(); sum(); quantile(); diff(); scale();*
- 数据集合合并: *rbind(); cbind();*
- 其余: *apply();*

编程

- 流程控制: ***if-else; ifelse;***
- 循环控制: ***repeat; for; while;***
- 数据输入输出函数: *read.table(); write.table(); read.csv(); write.csv();*
- 函数： ***function();***
- *apply*族函数： ***lapply(); sapply(); vapply(); tapply();***

Course Summary

画图-01

- 图形函数：
  - ✴ plot(); barplot(); pie(); hist(); boxplot();
- 图形参数：
  - ✴ col; font; pch; cex; lty; lwd; xlab; ylab; xlim; ylim; type; main; horiz; beside;
- 图例函数：
  - ✴ legend(location, title, legend, …);
- 图形组合：
  - ✴ par(); layout();
- 其余函数：
  - ✴ title(); abline(); line(); text(); mtext();

画图-02

- 图例：
  - ✴ 坐标; 边界标注; 标注(mar);horiz=TRUE;
- 线图：
  - ✴ grid(); abline(); line(); lm(); arrows();
- 条形图：
  - ✴ 堆积(beside); horiz=TRUE;
  - ✴ 显示数字; 宽度、颜色和边界; 显示标注; 增加误差线
- 散点图：
  - ✴ point();type="n"; corplot(); 增加抖动;
- 其余：
  - ✴ par(); axis(); mtext(); jitter();

Course Summary

ggplot-01

- ggplot2
- qplot() :
  - ✴ data; log; colour; shape; alpha;
- geom :
  - ✴ point; smooth; jitter; boxplot;path; line; histogram; freqpoly; density; bar;
  - ✴ binwidth; fill; weight; scale_y_continous(); smooth;
- facets :
- ggplot():
  - ✴ +; %+%; layer(); geom_xxx(); stat_xxx(); aes(); group;

Course Summary

统计

- 基本统计
  - ✴ mean; median; quantile; weighhted.mean; length; min; max; var; sd;
- 概率函数:
  - ✴ dnorm 密度; pnorm 分布; qnorm 分位数; rnorm 随机数;
  - ✴ norm 正态; binom 二项; unit 均匀; beta; exp 指数; …; sample; set.seed;
- 总结:
  - ✴ summary; sapply; HMisc::describe; pastecs::stat.desc; psych::describe;
  - ✴ stat_bin; stat_bin2d; stat_binhex; stat_density2d; stat_summary;
- 回归分析:
  - ✴ lm; format; residuals; anova; predict;
  - ✴ 简单线性回归; 多项式回归; 多元线性回归; 有交互项的多元线性回归;
- 因子分析:
  - ✴ cor; cor.test; factanal 因子; princomp 主成分; screeplot; biplot;

Course Summary

信用评分

- 信用评分概述
  - ✳ 定义、需求、目的、历史、例子
- 信用评分计算
  - ✳ 概率、案例、贝叶斯、贝叶斯评分卡、评价
- 评分卡建模
  - ✳ dnorm 密度; pnorm 分布; qnorm 分位数; rnorm 随机数;
  - ✳ 好坏样本; 数据来源; 开发过程; 特征分析;
  - ✳ 线性回归; 逻辑回归; 分类树; 神经网络; SVM; KNN; … ;
- 其余
  - ✳ 中小微企业信用评分

Course Summary

R数据科学

- dplyr
  - ✴ filter; arrange; select; mutate; summarize;
- 数据处理：
  - ✴ tibble; readr; dplyr;
  - ✴ line; vline; hline; abline; rect; text; arrow;
- 字符：
  - ✴ strsplit 拆分; grep/grepl 查询; regexpr/gregexpr/regexec 查询-位置;
  - ✴ 正则; stringr; lubridate; rmarkdown;
- 编程：
  - ✴ %>%; purr; colour;modelr; tidyverse;

RiA其余

- 方差分析
  - ✴ aov; TukeyHSD 多重比较; glht; qqplot;
  - ✴ interaction.plot; plotmeans; interaction2wt;
  - ✴ 单因素; 协方差; 双因素; 多元;
- 缺失值处理
  - ✴ is.na; is,nan; is.infinite;
  - ✴ mice::md.pattern; aggr; matrixplot;
  - ✴ 完整行删除; 多重插补; 成对删除; 简单插补;
- RCurl

Datacamp

Course Summary

练习 - 0007

INTERACTIVE COURSE

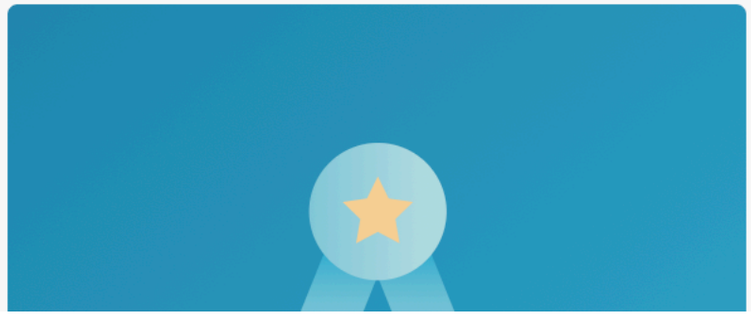
Introduction to R

Practice Now | Replay Course | Bookmarked

4 hours | 0 Videos | 62 Exercises | 1,470,780 Participants | 6,200 XP | ALSO AVAILABLE ON MOBILE

Course Description

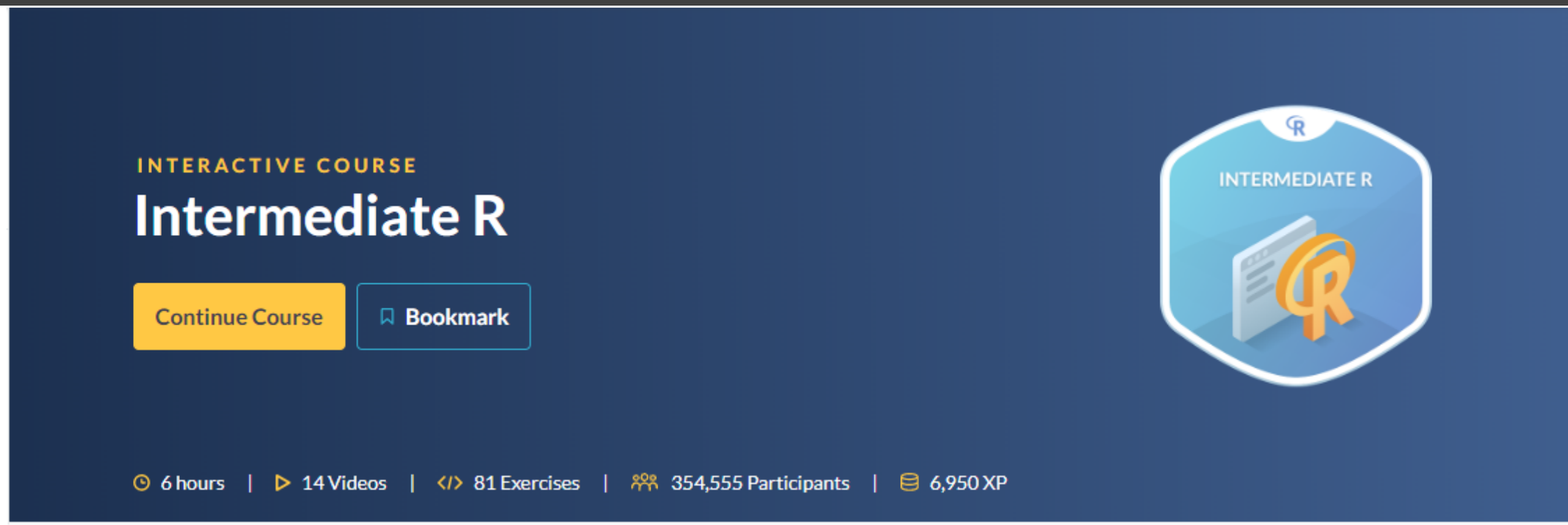
In Introduction to R, you will master the basics of this widely used open source language, including factors, lists, and data frames. With the knowledge gained in this course, you will be ready to undertake your first very own data analysis. Oracle



https://www.datacamp.com/courses

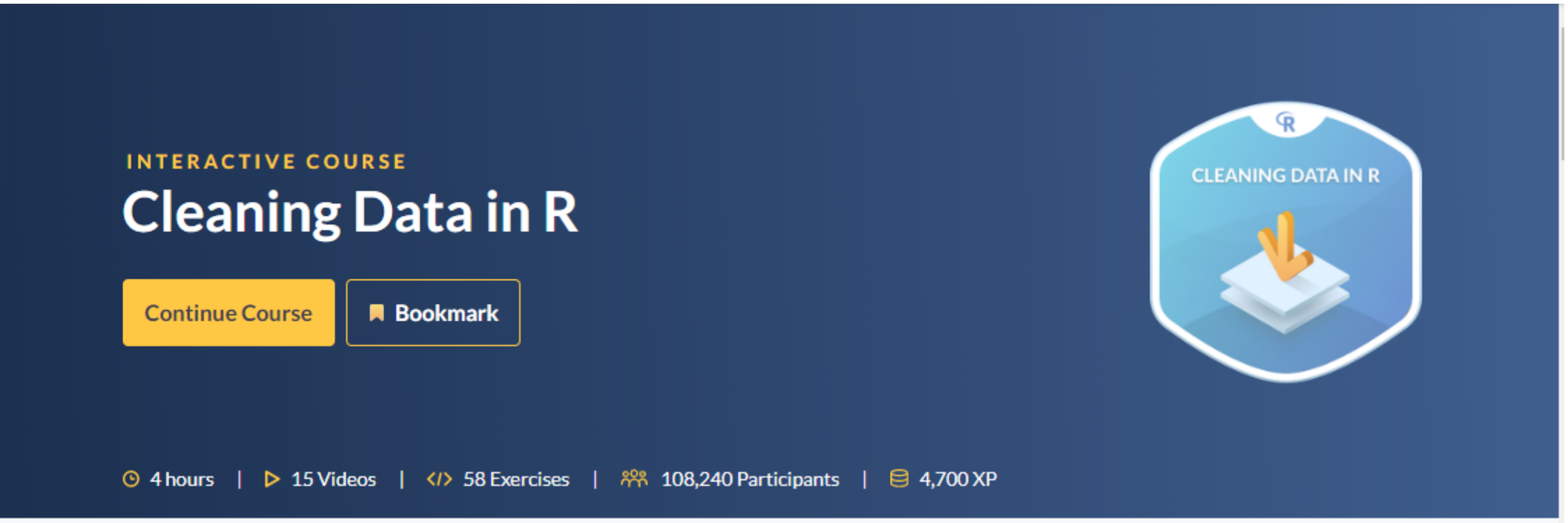
Course Summary

练习 - 0009



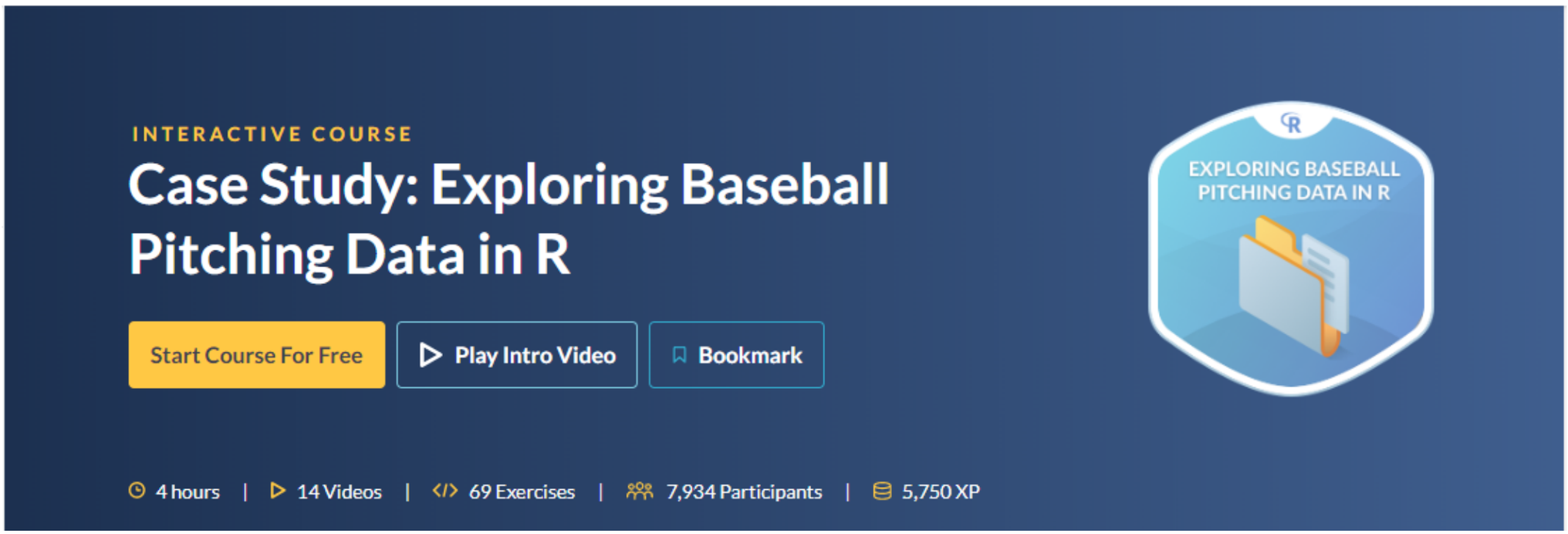
提交方式和上节课一样！

https://www.datacamp.com/courses



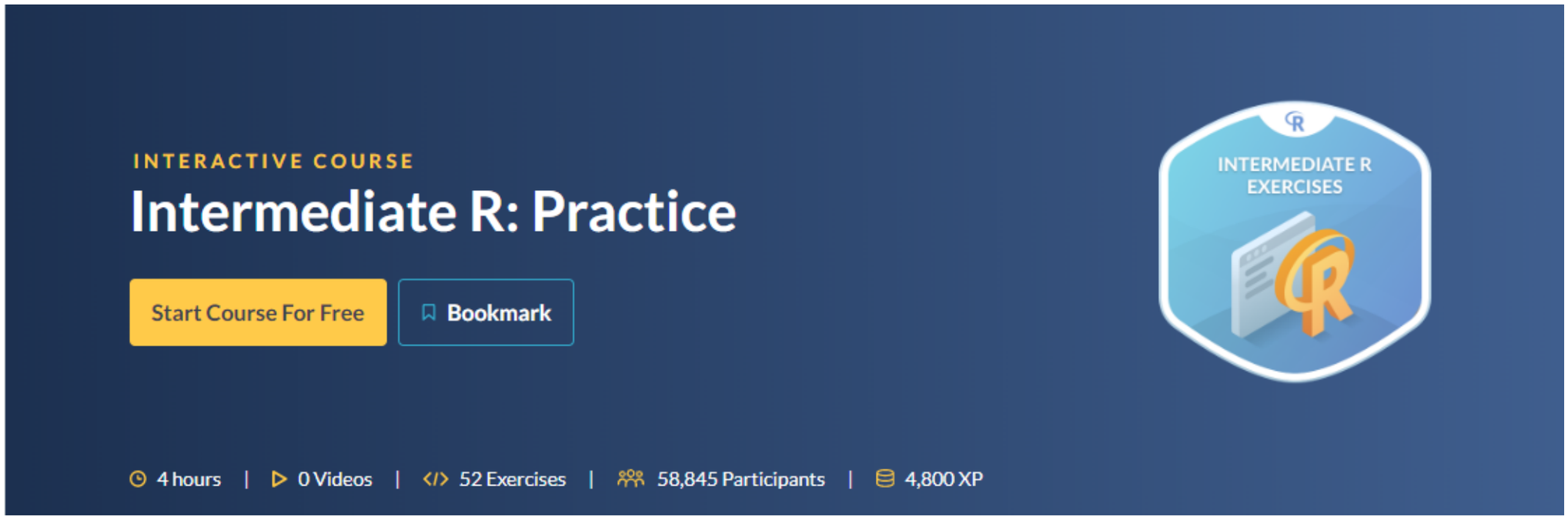
Course Summary

练习 – 0013



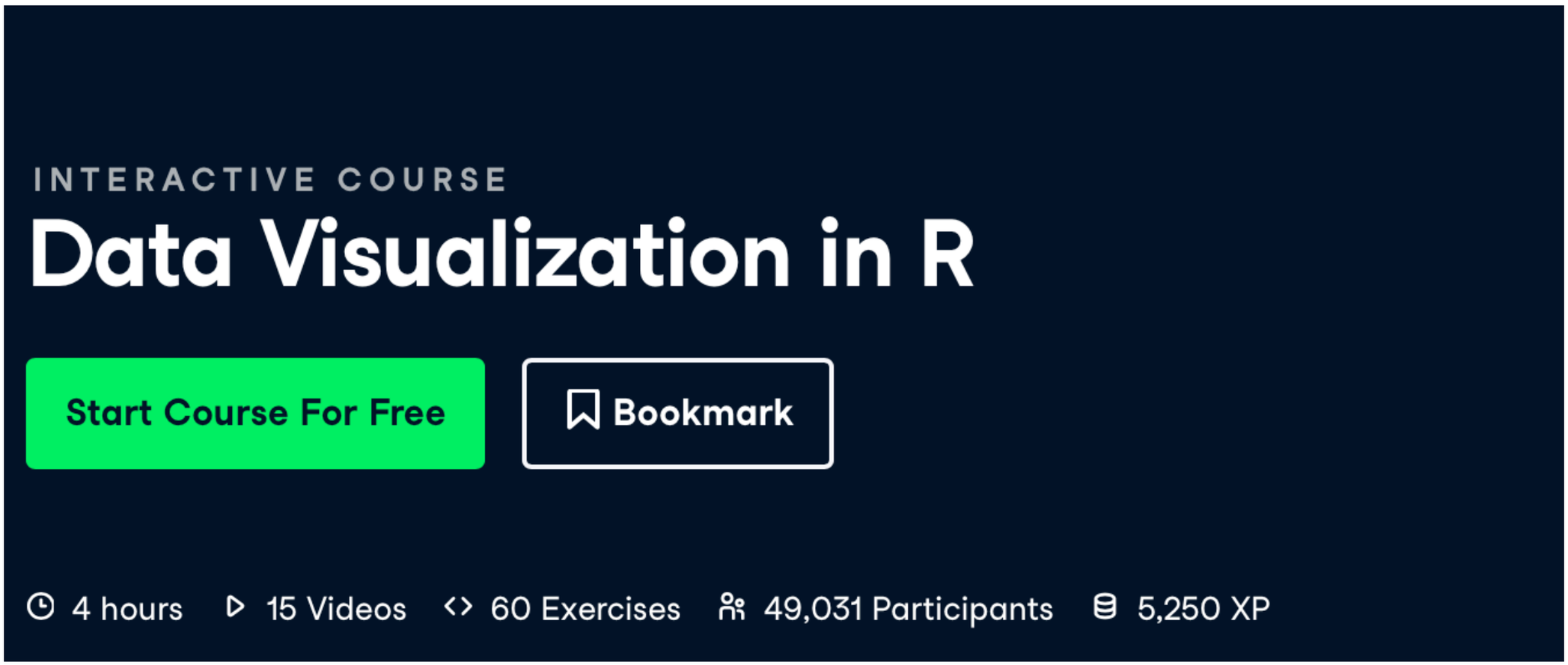
提交方式和上节课一样！

https://www.datacamp.com/courses



Course Summary

练习 – 0020

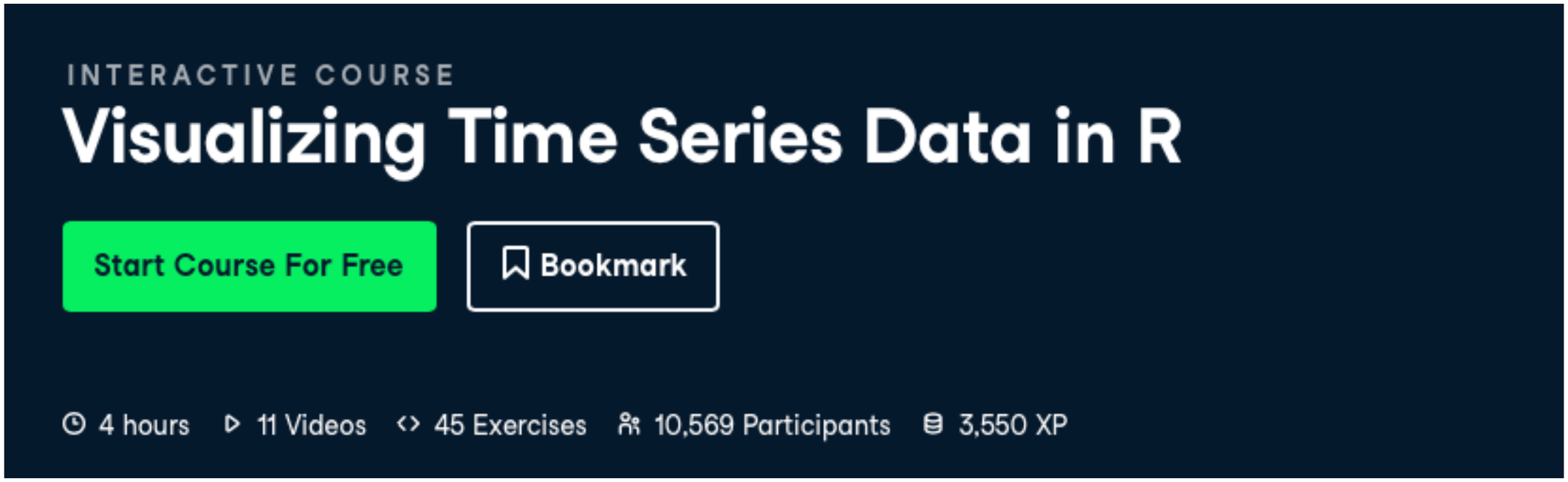


提交方式和上节课一样！

https://www.datacamp.com/courses

Course Summary

练习 – 0023

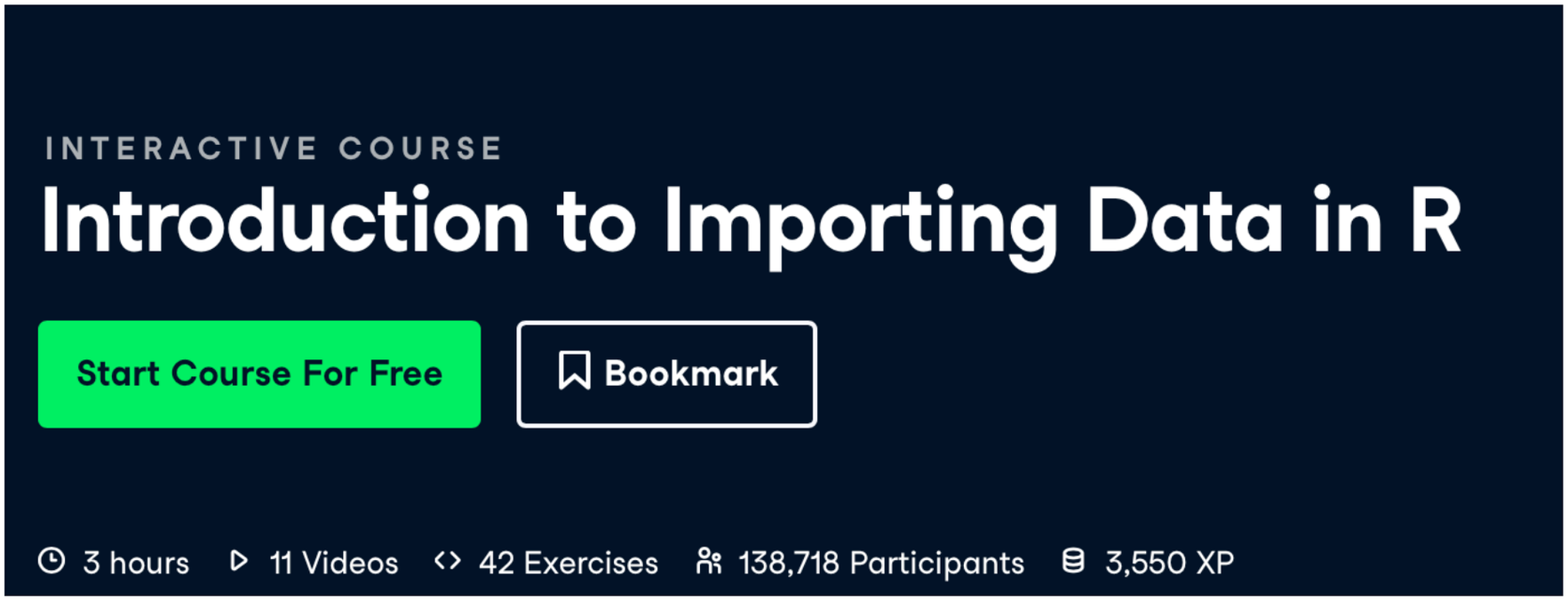


提交方式和上节课一样！

https://www.datacamp.com/courses

Course Summary

练习 – 0024

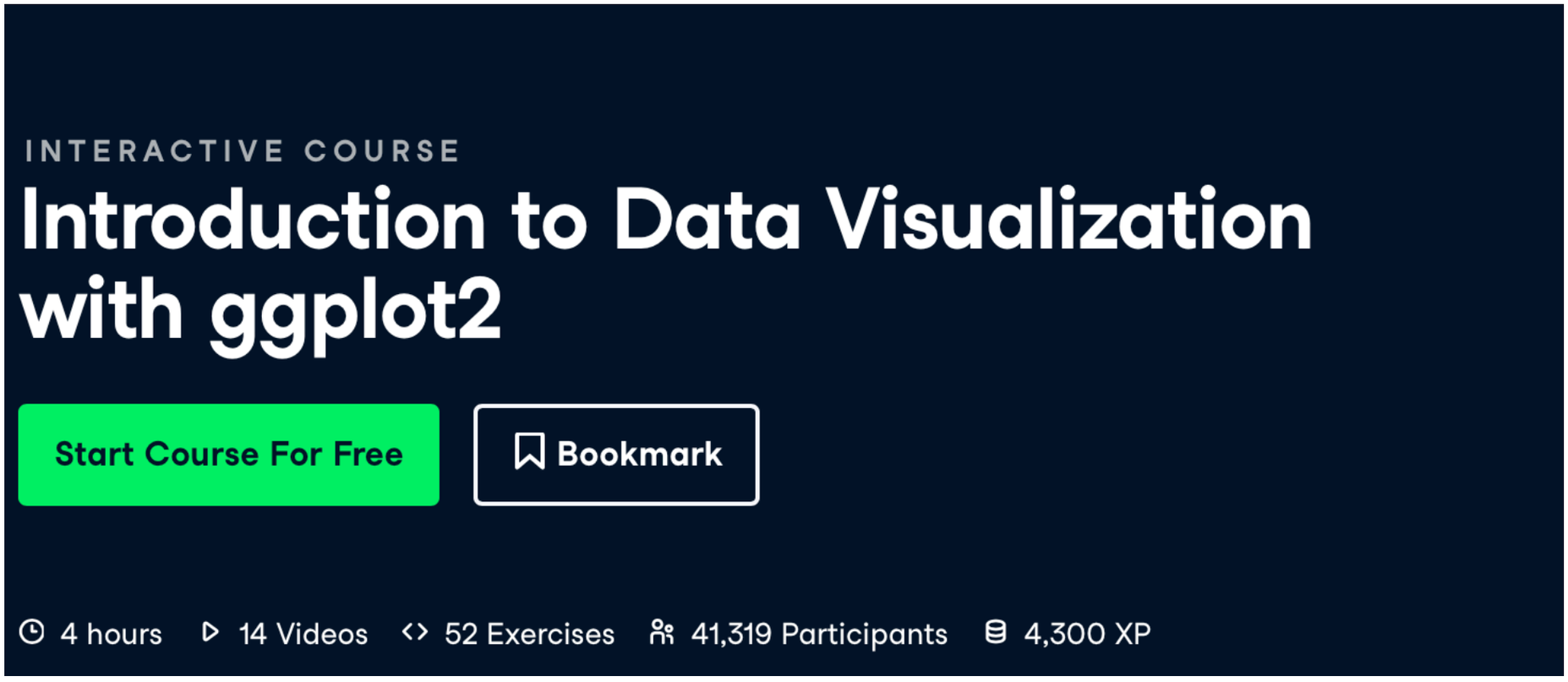


提交方式和上节课一样！

https://www.datacamp.com/courses

Course Summary

练习 – 0026

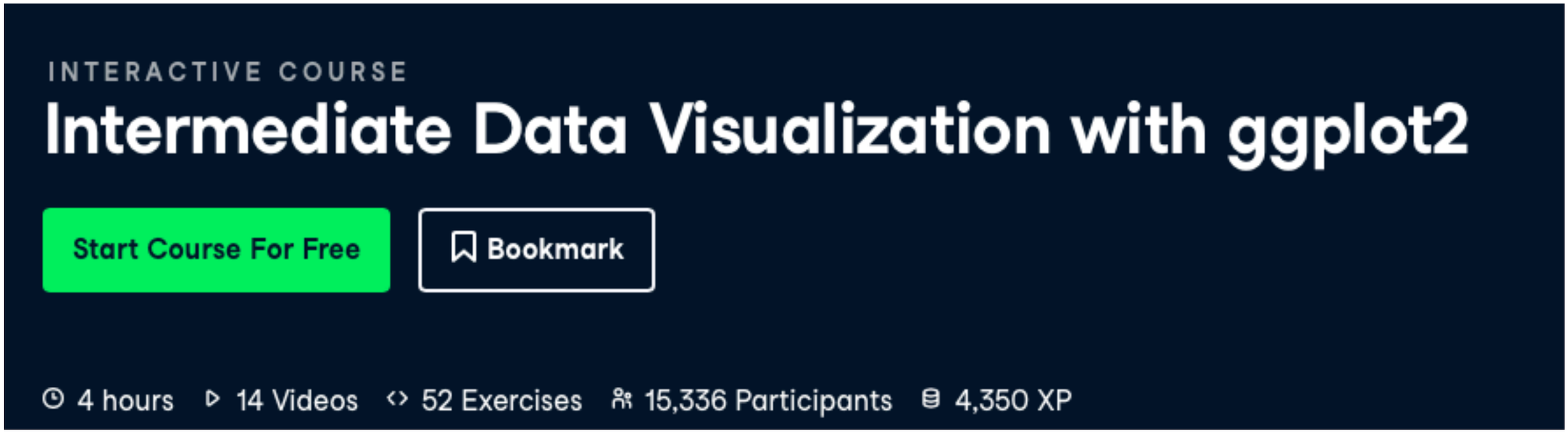


提交方式和上节课一样！

https://www.datacamp.com/courses

Course Summary

练习 – 0028

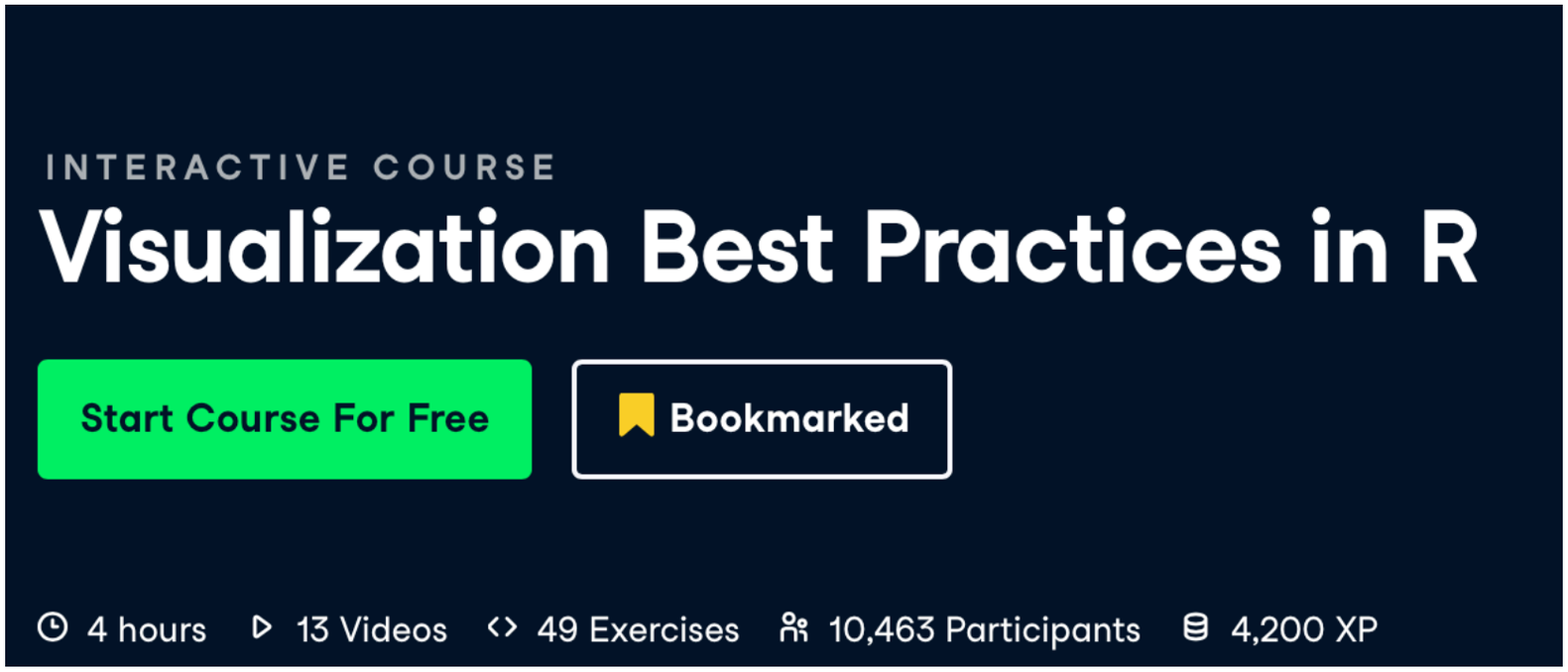


提交方式和上节课一样！

https://www.datacamp.com/courses

Course Summary

练习 - 0030



提交方式和上节课一样！

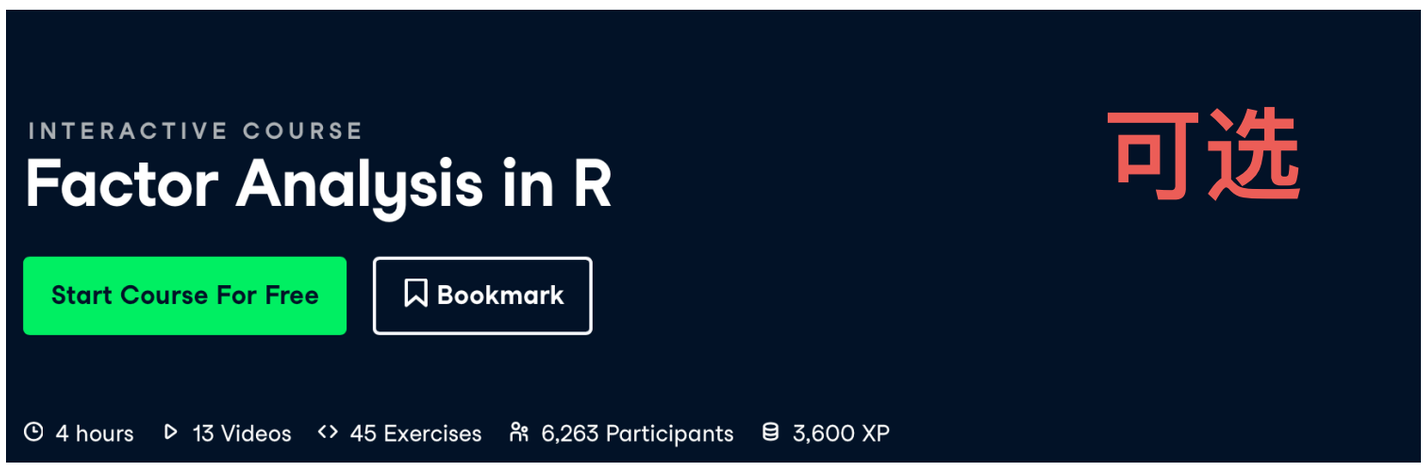
https://www.datacamp.com/courses

Course Summary

练习 - 0031

提交方式和上节课一样！

https://www.datacamp.com/courses



可选

R包

R包

- dplyr & tidyr
- jirba & wordcloud
- tidymodels
- tidyquant
- svm & xgboost
- stringr & lubridate

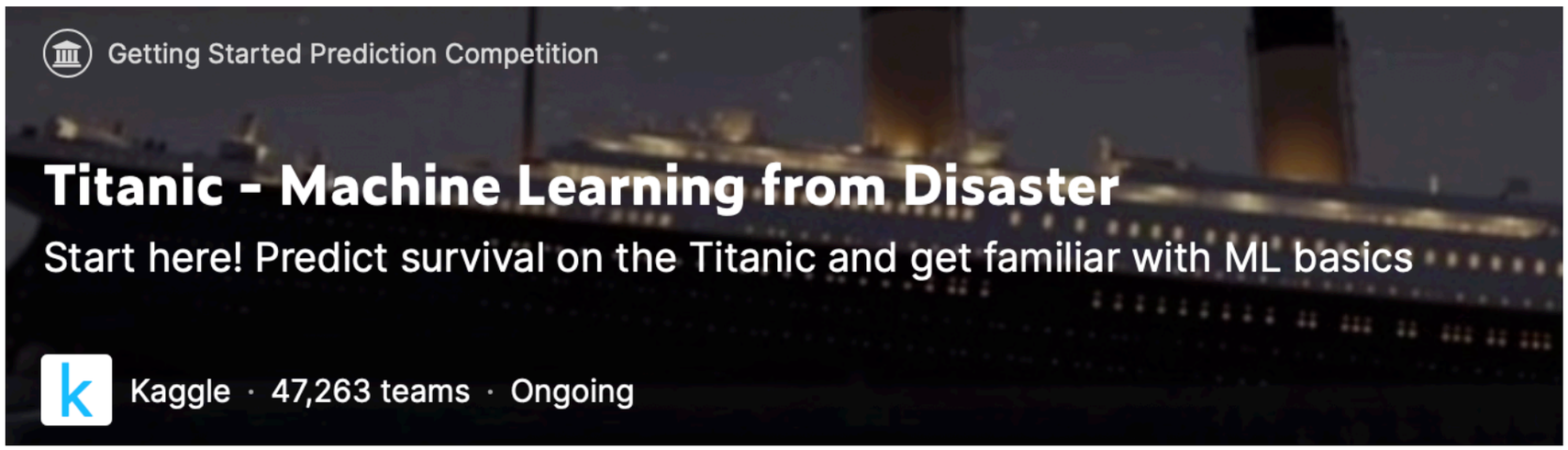
课后和课堂练习

课堂测试14

先用电脑完成
40分钟 然后誊抄纸上

- 泰坦尼克号数据库，见**titanic.zip**

- 对数据进行预处理，包括查看统计特征、插补空值等操作
- 可视化分析得出初步结论
- 利用决策树进行建模，并且作出分析 caret
- 修正模型并利用ROC曲线检验模型的准确度如何 ROCR



检索

https://www.kaggle.com/c/titanic/

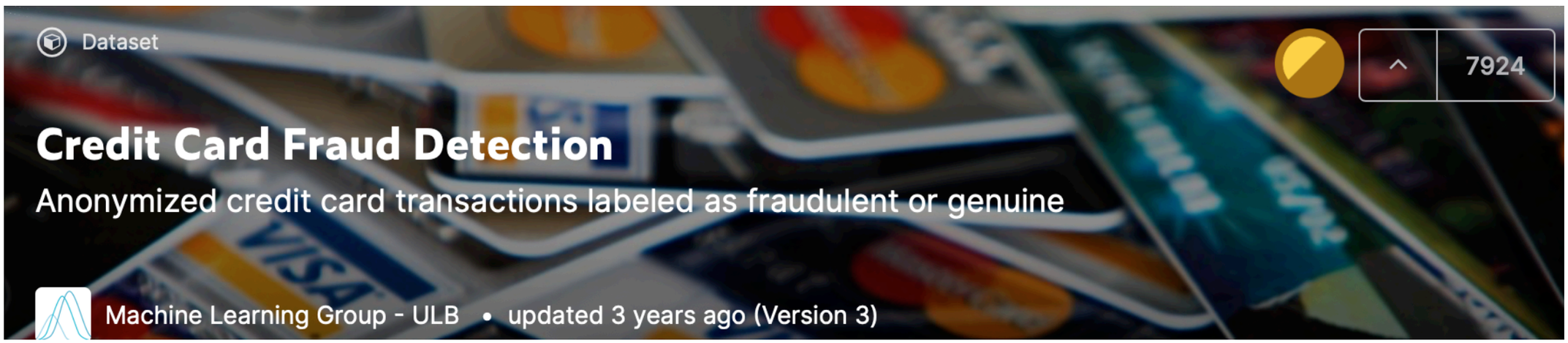
练习 - 0035

下周上课前提交方式和以前一样

- 信用卡欺诈数据库，见**creditcard.csv**

	U	V	W	X	Y	Z	AA	AB	AC	AD	AE
	V20	V21	V22	V23	V24	V25	V26	V27	V28	Amount	Class
6	0.2514121	-0.0183068	0.27783758	-0.1104739	0.06692807	0.12853936	-0.1891148	0.13355838	-0.0210531	149.62	0
1	-0.0690831	-0.2257752	-0.638672	0.10128802	-0.3398465	0.1671704	0.12589453	-0.0089831	0.01472417	2.69	0
5	0.52497973	0.24799815	0.7716794	0.90941226	-0.689281	-0.3276418	-0.1390966	-0.0553528	-0.0597518	378.66	0
7	-0.2080378	-0.1083005	0.0052736	-0.1903205	-1.1755753	0.64737603	-0.2219288	0.06272285	0.06145763	123.5	0
5	0.40854236	-0.0094307	0.79827849	-0.1374581	0.14126698	-0.2060096	0.50229222	0.21942223	0.21515315	69.99	0
8	0.08496767	-0.2082535	-0.5598248	-0.0263977	-0.3714266	-0.2327938	0.10591478	0.25384422	0.08108026	3.67	0
5	-0.2196326	-0.1677163	-0.2707097	-0.1541038	-0.7800554	0.75013694	-0.2572368	0.03450743	0.00516777	4.99	0
1	-0.1567419	1.94346534	-1.0154547	0.05750353	-0.649709	-0.4152666	-0.0516343	-1.2069211	-1.0853392	40.8	0
7	0.05273567	-0.0734251	-0.2680916	-0.2042327	1.0115918	0.37320468	-0.3841573	0.01174736	0.14240433	93.2	0

- 建模分析信用卡欺诈
- 分析模型有效性
- 使用全部数据
- 使用更复杂模型和工具



https://www.kaggle.com/mlg-ulb/creditcardfraud

练习 - 0034

- **天眼查**每个企业均有详细信息和企业评分
- 写一个爬虫，下载1000个企业信息
- 根据企业信息项，分析预测天眼查现有企业评分的计算方法，用下载的数据进行回归分析，再爬一些企业作为测试数据，分析方法的有效性
- 自己设计一个企业评分算法，说明设计原理，并用数据检验

企业最好针对一个行业

法定代表人	季 季昊 任职 8 家企业，分布如下 北京（共8家） 北京国际大数据交易… 等	经营状态	存续	天眼评分	评分 68
		成立日期	2020-04-29		
		注册资本	625万人民币		
		实缴资本	-	工商注册号	110108028700442
统一社会信用代码	91110108MA01R2FT36	纳税人识别号	91110108MA01R2FT36	组织机构代码	MA01R2FT3
营业期限	2020-04-29 至 2070-04-28	纳税人资质	-	核准日期	2021-01-06
公司类型	其他有限责任公司	行业	软件和信息技术服务业	人员规模	-
参保人数	-	登记机关	北京市海淀区市场监督管理局		
曾用名	-	英文名称	-		
注册地址	北京市海淀区丹棱街1号院1号楼26层2601室 附近公司				

下周上课前提交方式和以前一样

Course Summary

练习 - 0033

- 12306泄漏数据库，见@12306.txt

```
274667266@qq.com----6837605----郑一峰----332522198705040011----z6837605----15068860664----274667266@qq.com
zaistar@163.com----tianxia512----池善卿----350426197909063O1X----chitang520----18105013289----zaistar@163.com
weizhongjie55@163.com----wzj27713----卫忠杰----210602198711260513----wzj871126----18707734000----weizhongjie55@163.com
xujsh2004@yahoo.com.cn----19830307----许家圣----340103198303072554----xujsh2012----18225513108----xujsh2004@yahoo.com.cn
793925564@qq.com----793925564----李靖男----410183199307210015----lijingnan741----18024105681----793925564@qq.com
chenkan588@163.com----chengkang----陈侃----362326198306270039----chenkan588----18258288023----chenkan588@163.com
kangjie109@163.com----159648sl----康焕卉----430503198706130038----kangjie109----13716008430----kangjie109@163.com
a2135336@163.com----a2135336----池鹏----331081198601210014----cp165147----18888731462----a2135336@163.com
daqi1003@163.com----liudaqi----刘大奇----230103198509121352----daqi1003----15810596619----daqi1003@163.com
```

- 统计口令数量：仅有数字、仅有字母、包含特殊字符、字母＋数字，画图展示
- 统计最常用的100个口令，并用词云画出来
- 口令长度统计分析，画图展示
- 口令是日期形式的统计分析，画图展示
- 检索强口令的定义，找出有哪些强口令
- 通过数据分析你能发现现在的账户和口令存在哪些问题不，简答总结并用数据说明

下周上课前提交方式和以前一样

Course Summary

练习 - 0032

Give Me Some Credit 数据

https://www.kaggle.com/c/GiveMeSomeCredit

数据描述

缺失值处理

异常值处理

好坏样本选择

特征选择

特征工程

模型构建

逻辑回归模型

模型评测

Lending Club数据

提交代码和报告

其余课堂测试

课堂测试08

先用电脑完成
40分钟 然后誊抄纸上

- 数据集aqi_combine.csv描述：AQI指数（空气质量指数）AQI的指数的取值范围为0~500，其中0～50、51～100、101～200、201～300和大于300，分别对应国家空气质量标准中日均值的 I级、II级、III级、IV级和V级标准的污染物浓度限定数值。
  - ➡ I级：空气质量评估为优，对人体健康无影响；
  - ➡ II级：空气质量评估为良，对人体健康无显著影响；
  - ➡ III级：为轻度污染，健康人群出现刺激症状；
  - ➡ IV级：中度污染，健康人群普遍出现刺激症状；
  - ➡ V级：严重污染，健康人群出现严重刺激症状。
- 主要污染物
  - ➡ 六项污染物质的浓度：其中PM2.5（粒径小于等于2.5μm的颗粒物，也称细颗粒物），PM10（粒径小于等于10μm的颗粒物，也称可吸入颗粒物），SO2（二氧化硫），NO2（二氧化氮）以及CO（一氧化碳）的浓度全部为24小时平均值，O3浓度值为8小时的滑动平均值。
- 时间跨度：2015年1月1日至2017年6月30日，共有912条记录。

课堂测试08

先用电脑完成
40分钟 然后誊抄纸上

- 使用ggplot2里的画图函数完成以下的练习：
  - ➡ 载入绘图相关数据包并加载数据集文件aqi_combine.csv，打印数据概况
  - ➡ 污染等级的频率和频数表
  - ➡ AQI指数的频数直方图
  - ➡ 数据集的第一列日期，请提取其中的年份，并转换成因子类型，画出分年份AQI密度曲线,设置主题为theme_bw
  - ➡ 主要污染物的频数统计，以及分污染等级对主要污染物进行频数统计
  - ➡ AQI指数与各类污染物的矩阵散点图
  - ➡ 是否下雨条件下分组AQI密度曲线,设置主题为theme_bw
  - ➡ 日均温度-AQI散点图和拟合曲线，分是否下雨情况下的日均温度箱线图
  - ➡ 温度-各类污染物散点图，设置布局为两行三列，是否下雨-各污染物浓度分组箱线图，布局为一行六列

课堂测试09

先用电脑完成
40分钟 然后誊抄纸上

- 1、下表是一个一个村庄儿童年龄和平均身高的统计数据
  - ✴ （1）画出平均身高height和年龄age关系的散点图
  - ✴ （2）建立回归模型并提取结果输出，在(1)中的图中表示生成的模型

平均身高与年龄			
年龄（月）	平均身高（厘米）	年龄（月）	平均身高（厘米）
18	76.1	24	79.9
19	77	25	81.1
20	78.1	26	81.2
21	78.2	27	81.8
22	78.8	28	82.8
23	79.7	29	83.5

- 2、 revenue.txt中记录了财政收入(y)和第一产业GDP X_1、第二产业GDP X_2、第三产业GDPx_3、人口数X_4、社会消费品零售总额X_5 、受灾面积X_6 、等情况的统计数据。要求:写出多元线性回归模型。

课堂测试09

先用电脑完成
40分钟 然后誊抄纸上

- 3、某公司想要了解消费者购买牙膏时更追求什么样的目标,于是通过商场拦访对30个人进行访谈，用7级里克特量表询问他们对以下陈述的认同程度(即1表示非常不同意，7表示非常同意,V1:购买预防蛀牙的牙膏是重要的;V2:我喜欢使牙齿亮泽的牙膏；v3:牙膏应当保护牙龈；V4:我喜欢使口气清新的牙膏；V5:预防坏牙不是牙膏提供的一项重要功效；V6:购买牙膏时最重要的考虑是富有魅力的牙齿:
  - ✴ 将调查样本存储于文本文档 yagao.txt。请使用R函数factanal对数据进行因子分析，根据载荷系数矩阵，写出因子和原变量之间的线性关系式。
- 4、某地区农业生态经济系统的各区域单元相关指标数据在文本文件agriculture.txt中，使用R中的主成分分析的函数princomp选取更少的指标来描述该地区的农业生态经济系统。写出主成分和原变量之间的线性关系式。

课堂测试10

先用电脑完成
40分钟 然后誊抄纸上

- 1、针对Give Me Some Credit 数据
  - ✳ （1）通过可视化开分析缺失值和异常值
  - ✳ （2）处理缺失值和异常值
  - ✳ （3）分析变量的相关性
  - ✳ （4）通过分箱、WOE和IV来检查各变量预测能力
  - ✳ （5）变量和特征选择
  - ✳ （6）用逻辑回归建立一个模型
  - ✳ （7）检验模型有效性（F1、ROC）

可以使用
ScoreCard包

课堂测试11

先用电脑完成
40分钟 然后誊抄纸上

- 12306泄漏数据库，见@12306.txt

```
274667266@qq.com----6837605----郑一峰----332522198705040011----z6837605----15068860664----274667266@qq.com
zaistar@163.com----tianxia512----池善卿----350426197909063O1X----chitang520----18105013289----zaistar@163.com
weizhongjie55@163.com----wzj27713----卫忠杰----210602198711260513----wzj871126----18707734000----weizhongjie55@163.com
xujsh2004@yahoo.com.cn----19830307----许家圣----340103198303072554----xujsh2012----18225513108----xujsh2004@yahoo.com.cn
793925564@qq.com----793925564----李靖男----410183199307210015----lijingnan741----18024105681----793925564@qq.com
chenkan588@163.com----chengkang----陈侃----362326198306270039----chenkan588----18258288023----chenkan588@163.com
kangjie109@163.com----159648sl----康焕卉----430503198706130038----kangjie109----13716008430----kangjie109@163.com
a2135336@163.com----a2135336----池鹏----331081198601210014----cp165147----18888731462----a2135336@163.com
daqi1003@163.com----liudaqi----刘大奇----230103198509121352----daqi1003----15810596619----daqi1003@163.com
```

- 统计口令数量：仅有数字、仅有字母、包含特殊字符、字母＋数字
- 统计最常用的100个口令
- 分离出账号和口令有关系的信息，分离身份证和口令有关的信息
- 统计泄漏的年龄分布

注意：数据量较大，可以选择一个子集上完成（10万）

课堂测试12

先用电脑完成
40分钟 然后誊抄纸上

- US baby names数据集：
- 美国1880年到2008年Top1000的男婴和女婴的名字
- 258000条记录
- year
- name
- sex
- percent
- 见： bnames.csv

year	name	percent	sex
1880	John	0.081541	boy
1880	William	0.080511	boy
1880	James	0.050057	boy
1880	Charles	0.045167	boy
1880	George	0.043292	boy
1880	Frank	0.02738	boy
1880	Joseph	0.022229	boy
1880	Thomas	0.021401	boy

```
> head(bnames,15)
   year     name  percent sex
1  1880     John 0.081541 boy
2  1880  William 0.080511 boy
3  1880    James 0.050057 boy
4  1880  Charles 0.045167 boy
5  1880   George 0.043292 boy
6  1880    Frank 0.027380 boy
7  1880   Joseph 0.022229 boy
8  1880   Thomas 0.021401 boy
9  1880    Henry 0.020641 boy
10 1880   Robert 0.020404 boy
11 1880   Edward 0.019965 boy
12 1880    Harry 0.018175 boy
13 1880   Walter 0.014822 boy
14 1880   Arthur 0.013504 boy
15 1880     Fred 0.013251 boy
```

课堂测试12

先用电脑完成
40分钟 然后誊抄纸上

- 该数据集中每年有多少记录
- 数据集中男孩和女孩各自排名
- 男孩名和女孩名的Top100

dplyr

- Top100中男孩名和女孩名的所占比例
- 画图显示每一年男孩名和女孩名在Top100的比例
- 哪些名字仅仅在一年中使用，哪些名字每一年都使用
- 显示每个名字的平均百分比
- 那个名字被使用的时间最长

谢谢！

孙惠平

sunhp@ss.pku.edu.cn