北京郵電大學
BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS

Data Analytics with R
语言数据分析

基础编程

艾新波 / 2018 • 北京

课程体系

- R语言数据分析
  - 上部：论道 
    - 第1章 气象万千、数以等观
    - 第2章 所谓学习、归类而已
    - 第3章 格言联璧话学习
    - 第4章 源于数学、归于工程
  - 中部：执具 
    - 第5章 工欲善其事必先利其器
    - 第6章 基础编程
    - 第7章 数据对象
  -
    - 第8章 人人都爱tidyverse
    - 第9章 最美不过数据框
  - 下部 博术 
    - 第10章 观数以形
    - 第11章 相随相伴、谓之关联
    - 第12章 既是世间法、自当有分别
    - 第13章 方以类聚、物以群分
    - 第14章 庐山烟雨浙江潮

一个完整的数据分析流程

http://en.wikipedia.org/wiki/CRoss_Industry_Standard_Process_for_Data_Mining

学生文理分科Mini案例

学生成绩表

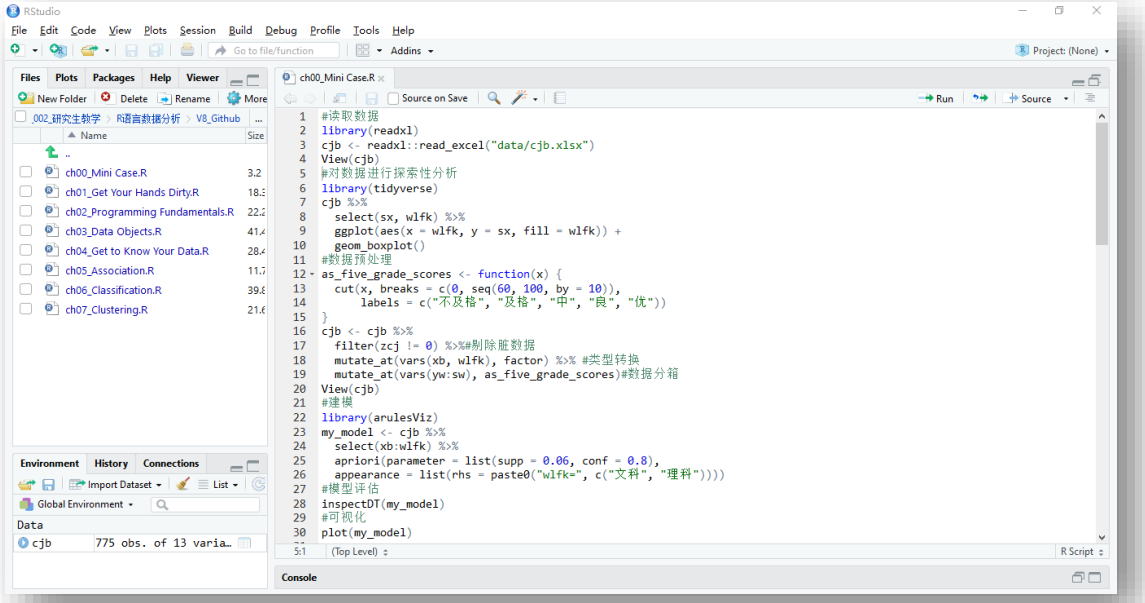
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	xm	bj	xb	yw	sx	wy	zz	ls	dl	wl	hx	sw	wlfk
2	周黎	1101	女	94	82	96	97	97	98	95	94	88	文科
3	汤海明	1101	男	87	94	89	95	94	94	90	90	89	文科
4	舒江辉	1101	男	92	79	86	98	95	96	89	94	87	文科
5	翁柯	1101	女	91	84	96	93	97	94	82	90	83	文科
6	祁强	1101	男	85	92	82	93	87	88	95	94	93	文科
7	湛容	1101	女	92	82	85	91	90	92	82	98	90	文科
8	穆伶俐	1101	女	88	72	86	94	87	88	89	98	94	文科
9	韦永杰	1101	男	81	89	87	97	94	96	81	88	83	文科
10	龚兰秀	1101	女	88	77	95	94	84	94	87	94	82	文科
11	舒亚	1101	女	94	81	88	91	85	98	81	88	88	文科
12	宰玲玲	1101	女	87	83	92	91	86	94	84	90	87	文科
13	邰友生	1101	男	88	82	91	89	81	98	89	98	75	文科
14	历阳	1101	男	79	84	91	87	91	87	85	96	90	文科
15	卜杰	1101	男	78	81	83	86	88	98	85	90	99	文科

xm:姓名　　bj:班级　　xb:性别　　yw:语文　　sx:数学　　wy:外语

zz:政治　　ls:历史 dl:地理　　hx:化学　　sw:生物　　wlfk:文理分科

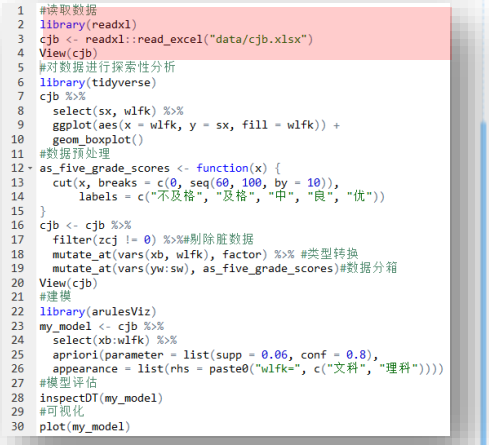
本数据集可从https://github.com/byaxb/RDataAnalytics下载

R语言数据分析Mini案例



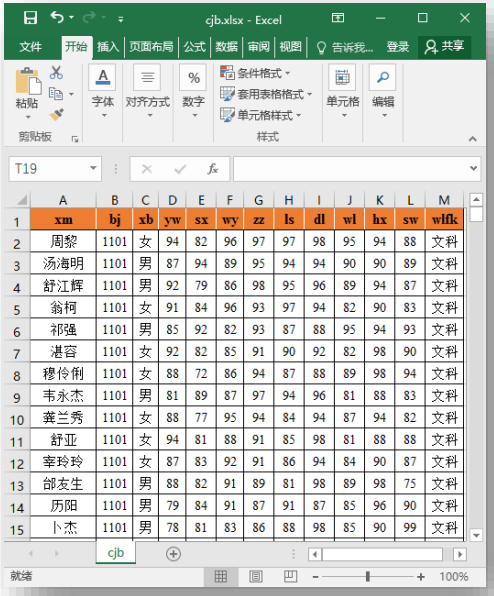
R语言数据分析Mini案例

```r
#读取数据
library(readxl)
cjb <- read_excel("data/cjb.xlsx")
View(cjb)
#对数据进行探索性分析
library(tidyverse)
cjb %>%
  select(sx, wlfk) %>%
  ggplot(aes(x = wlfk, y = sx, fill = wlfk)) +
  geom_boxplot(width = 0.5)
```

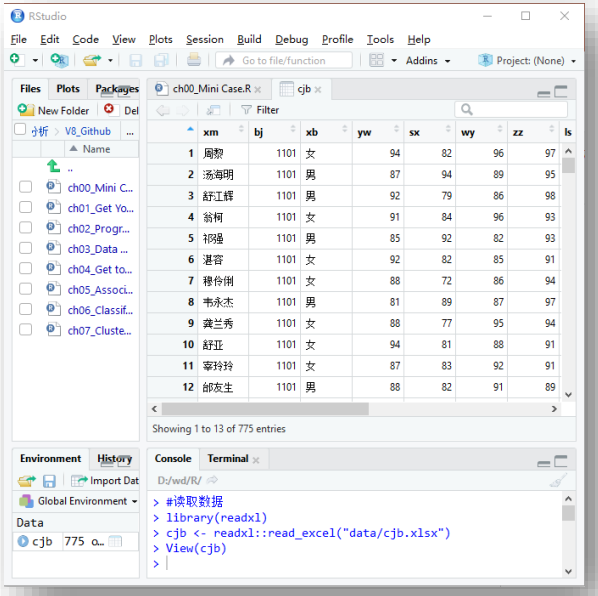


```r
#读取数据
library(readxl)
cjb <- readxl::read_excel("data/cjb.xlsx")
View(cjb)
#对数据进行探索性分析
library(tidyverse)
cjb %>%
  select(sx, wlfk) %>%
  ggplot(aes(x = wlfk, y = sx, fill = wlfk)) +
  geom_boxplot()
#数据预处理
as_five_grade_scores <- function(x) {
  cut(x, breaks = c(0, seq(60, 100, by = 10)),
      labels = c("不及格", "及格", "中", "良", "优"))
}
cjb <- cjb %>%
  filter(zcj != 0) %>%#剔除脏数据
  mutate_at(vars(xb, wlfk), factor) %>% #类型转换
  mutate_at(vars(yw:sw), as_five_grade_scores)#数据分箱
View(cjb)
#建模
library(arulesViz)
my_model <- cjb %>%
  select(xb:wlfk) %>%
  apriori(parameter = list(supp = 0.06, conf = 0.8),
  appearance = list(rhs = paste0("wlfk=", c("文科", "理科"))))
#模型评估
inspectDT(my_model)
#可视化
plot(my_model)
```

R语言数据分析Mini案例



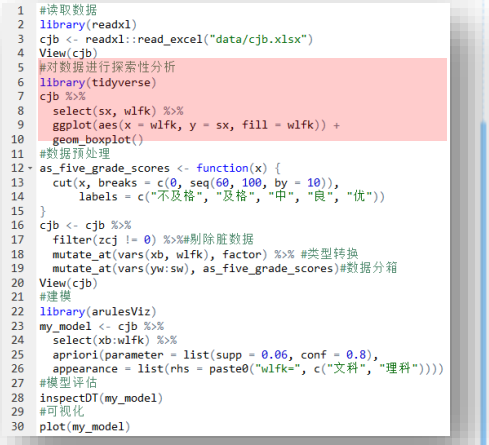
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	xm	bj	xb	yw	sx	wy	zz	ls	dl	wl	hx	sw	wlfk
2	周黎	1101	女	94	82	96	97	97	98	95	94	88	文科
3	汤海明	1101	男	87	94	89	95	94	94	90	90	89	文科
4	舒江辉	1101	男	92	79	86	98	95	96	89	94	87	文科
5	翁柯	1101	女	91	84	96	93	97	94	82	90	83	文科
6	祁强	1101	男	85	92	82	93	87	88	95	94	93	文科
7	谌容	1101	女	92	82	85	91	90	92	82	98	90	文科
8	穆伶俐	1101	女	88	72	86	94	87	88	89	98	94	文科
9	韦永杰	1101	男	81	89	87	97	94	96	81	88	83	文科
10	龚兰秀	1101	女	88	77	95	94	84	94	87	94	82	文科
11	舒亚	1101	女	94	81	88	91	85	98	81	88	88	文科
12	宰玲玲	1101	女	87	83	92	91	86	94	84	90	87	文科
13	邰友生	1101	男	88	82	91	89	81	98	89	98	75	文科
14	历阳	1101	男	79	84	91	87	91	87	85	96	90	文科
15	卜杰	1101	男	78	81	83	86	88	98	85	90	99	文科



	xm	bj	xb	yw	sx	wy	zz	ls
1	周黎	1101	女	94	82	96	97	
2	汤海明	1101	男	87	94	89	95	
3	舒江辉	1101	男	92	79	86	98	
4	翁柯	1101	女	91	84	96	93	
5	祁强	1101	男	85	92	82	93	
6	谌容	1101	女	92	82	85	91	
7	穆伶俐	1101	女	88	72	86	94	
8	韦永杰	1101	男	81	89	87	97	
9	龚兰秀	1101	女	88	77	95	94	
10	舒亚	1101	女	94	81	88	91	
11	宰玲玲	1101	女	87	83	92	91	
12	邰友生	1101	男	88	82	91	89	

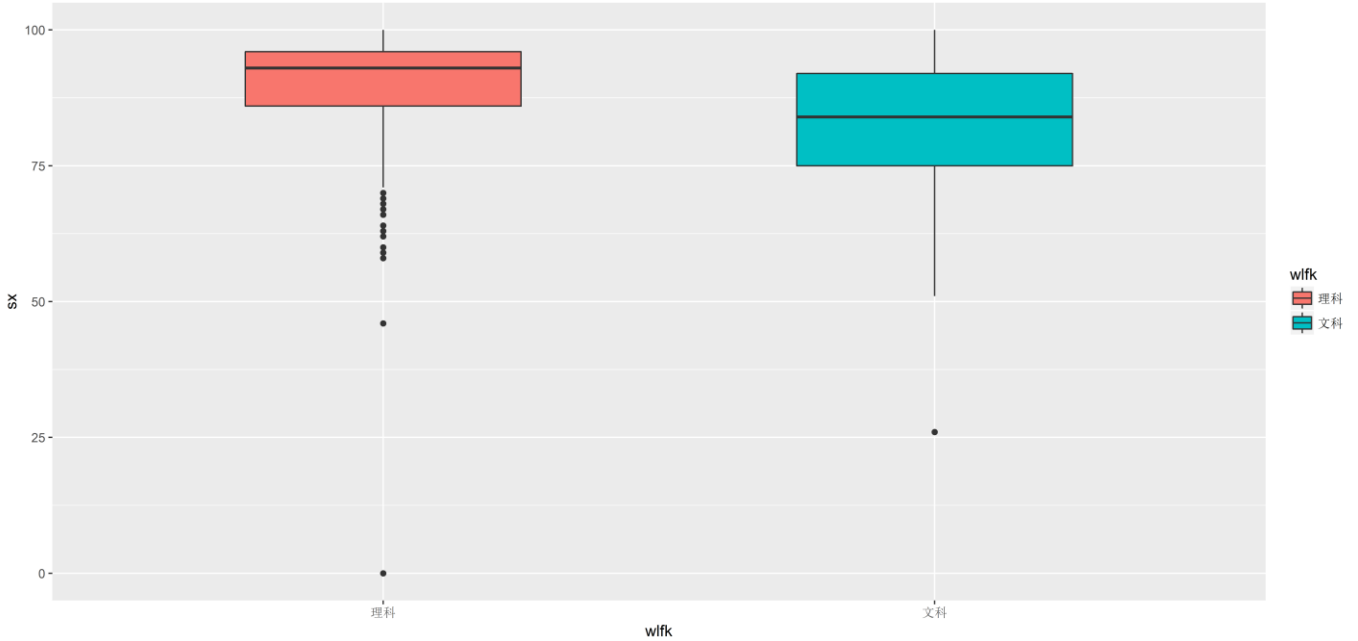
R语言数据分析Mini案例

```r
#读取数据
library(readxl)
cjb <- read_excel("data/cjb.xlsx")
View(cjb)
#对数据进行探索性分析
library(tidyverse)
cjb %>%
  select(sx, wlfk) %>%
  ggplot(aes(x = wlfk, y = sx, fill = wlfk)) +
  geom_boxplot(width = 0.5)
```



```r
#读取数据
library(readxl)
cjb <- readxl::read_excel("data/cjb.xlsx")
View(cjb)
#对数据进行探索性分析
library(tidyverse)
cjb %>%
  select(sx, wlfk) %>%
  ggplot(aes(x = wlfk, y = sx, fill = wlfk)) +
  geom_boxplot()
#数据预处理
as_five_grade_scores <- function(x) {
  cut(x, breaks = c(0, seq(60, 100, by = 10)),
      labels = c("不及格", "及格", "中", "良", "优"))
}
cjb <- cjb %>%
  filter(zcj != 0) %>%#剔除脏数据
  mutate_at(vars(xb, wlfk), factor) %>% #类型转换
  mutate_at(vars(yw:sw), as_five_grade_scores)#数据分箱
View(cjb)
#建模
library(arulesViz)
my_model <- cjb %>%
  select(xb:wlfk) %>%
  apriori(parameter = list(supp = 0.06, conf = 0.8),
  appearance = list(rhs = paste0("wlfk=", c("文科", "理科"))))
#模型评估
inspectDT(my_model)
#可视化
plot(my_model)
```

R语言数据分析Mini案例



R语言数据分析Mini案例

```r
#数据预处理
as_five_grade_scores <- function(x) {
  cut(x, breaks = c(0, seq(60, 100, by = 10)),
      labels = c("不及格", "及格", "中", "良", "优"))
}
cjb <- cjb %>%
  mutate(zcj = rowSums(.[4:12])) %>%
  filter(zcj != 0) %>%#剔除脏数据
  mutate_at(vars(xb, wlfk), factor) %>% #类型转换
  mutate_at(vars(yw:sw), as_five_grade_scores)#数据分箱
View(cjb)
```

```r
#读取数据
library(readxl)
cjb <- readxl::read_excel("data/cjb.xlsx")
View(cjb)
#对数据进行探索性分析
library(tidyverse)
cjb %>%
  select(sx, wlfk) %>%
  ggplot(aes(x = wlfk, y = sx, fill = wlfk)) +
  geom_boxplot()
#数据预处理
as_five_grade_scores <- function(x) {
  cut(x, breaks = c(0, seq(60, 100, by = 10)),
      labels = c("不及格", "及格", "中", "良", "优"))
}
cjb <- cjb %>%
  filter(zcj != 0) %>%#剔除脏数据
  mutate_at(vars(xb, wlfk), factor) %>% #类型转换
  mutate_at(vars(yw:sw), as_five_grade_scores)#数据分箱
View(cjb)
#建模
library(arulesViz)
my_model <- cjb %>%
  select(xb:wlfk) %>%
  apriori(parameter = list(supp = 0.06, conf = 0.8),
  appearance = list(rhs = paste0("wlfk=", c("文科", "理科"))))
#模型评估
inspectDT(my_model)
#可视化
plot(my_model)
```

R语言数据分析Mini案例

	xm	bj	xb	yw	sx	wy	zz	ls	dl	wl	hx	sw	wlfk
1	周黎	1101	女	94	82	96	97	97	98	95	94	88	文科
2	汤海明	1101	男	87	94	89	95	94	94	90	90	89	文科
3	舒江辉	1101	男	92	79	86	98	95	96	89	94	87	文科
4	翁柯	1101	女	91	84	96	93	97	94	82	90	83	文科
5	祁强	1101	男	85	92	82	93	87	88	95	94	93	文科
6	湛容	1101	女	92	82	85	91	90	92	82	98	90	文科

	xm	bj	xb	yw	sx	wy	zz	ls	dl	wl	hx	sw	wlfk
1	周黎	1101	女	优	良	优	优	优	优	优	优	良	文科
2	汤海明	1101	男	良	优	良	优	优	优	良	良	良	文科
3	舒江辉	1101	男	优	中	良	优	优	优	良	优	良	文科
4	翁柯	1101	女	优	良	优	优	优	优	良	良	良	文科
5	祁强	1101	男	良	优	良	优	良	良	优	优	优	文科
6	湛容	1101	女	优	良	良	优	良	优	良	优	良	文科

R语言数据分析Mini案例

```r
#建模
library(arulesViz)
my_model <- cjb %>%
  select(xb:wlfk) %>%
  apriori(parameter = list(supp = 0.06, conf = 0.8),
  appearance = list(rhs = paste0("wlfk=", c("文科", "理科"))))
#模型评估
inspectDT(my_model)
#可视化
plot(my_model)
```

R语言数据分析Mini案例

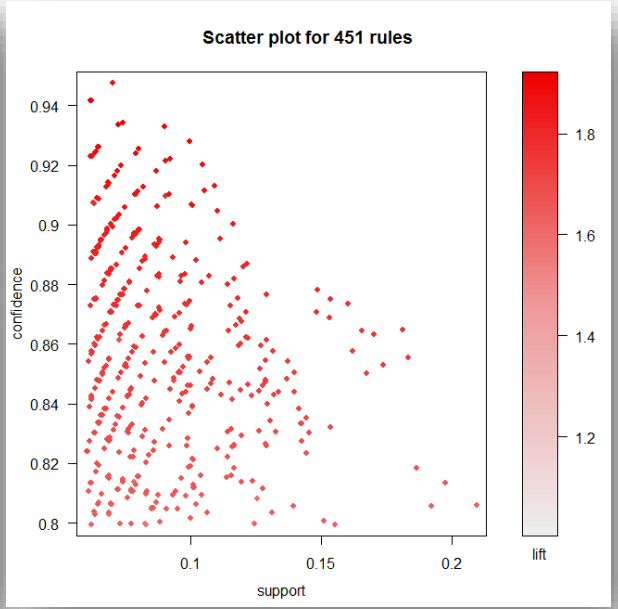
Viewer Zoom

Show 10 entries Search:

	LHS	RHS	support	confidence	lift	count
	All	All	All	All	All	All
[24]	{xb=女,zz=优,sw=中}	{wlfk=文科}	0.070	0.947	1.861	54.000
[347]	{xb=男,sx=优,wy=优,wl=优,hx=优,sw=优}	{wlfk=理科}	0.062	0.941	1.917	48.000
[429]	{xb=男,sx=优,wy=优,dl=优,wl=优,hx=优,sw=优}	{wlfk=理科}	0.062	0.941	1.917	48.000
[363]	{xb=男,sx=优,ls=优,wl=优,hx=优,sw=优}	{wlfk=理科}	0.074	0.934	1.903	57.000
[366]	{xb=男,sx=优,dl=优,wl=优,hx=优,sw=	{wlfk=理科}	0.090	0.933	1.901	70.000

Showing 1 to 10 of 451 entries

Previous 1 2 3 4 5 … 46 Next



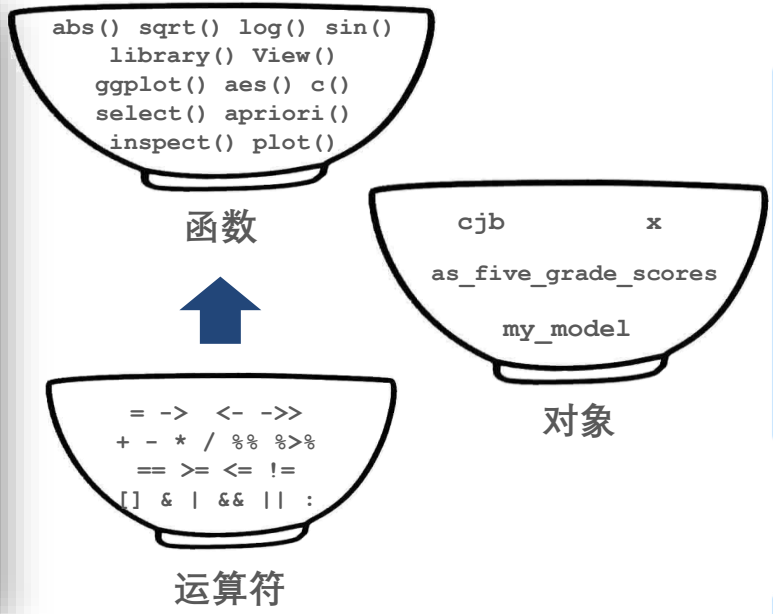
R语言数据分析Mini案例

```r
#读取数据
library(readxl)
cjb <- read_excel("data/cjb.xlsx")
View(cjb)
#对数据进行探索性分析
library(tidyverse)
cjb %>%
  select(sx, wlfk) %>%
  ggplot(aes(x = wlfk, y = sx, fill = wlfk)) +
  geom_boxplot()
#数据预处理
as_five_grade_scores <- function(x) {
  cut(x, breaks = c(0, seq(60, 100, by = 10)),
      labels = c("不及格", "及格", "中", "良", "优"))
}
cjb <- cjb %>%
  filter(zcj != 0) %>%#剔除脏数据
  mutate_at(vars(xb, wlfk), factor) %>% #类型转换
  mutate_at(vars(yw:sw), as_five_grade_scores)#数据分箱
View(cjb)
#建模
library(arulesViz)
my_model <- cjb %>%
  select(xb:wlfk) %>%
  apriori(parameter = list(supp = 0.06, conf = 0.8),
  appearance = list(rhs = paste0("wlfk=", c("文科", "理科"))))
#模型评估
inspectDT(my_model)
#可视化
plot(my_model)
```

- R仅有的命令形式是返回结果的函数和表达式
- 赋值是一种常见的操作：对象的读取、转换、模型的建立等
- 赋值给新的对象，往往也就意味着数据的流转：读取、转换、探索、建模、评估等操作

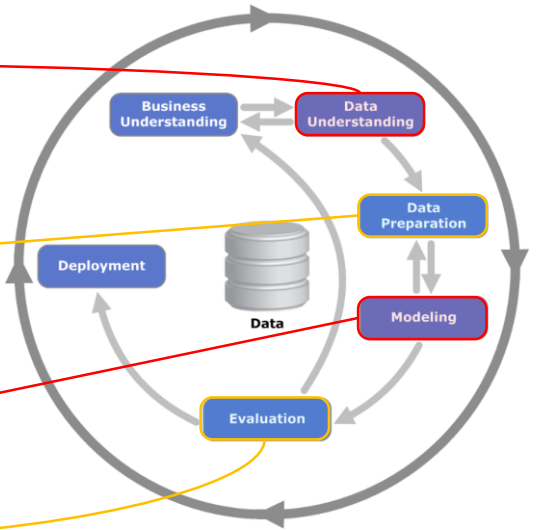
R语言数据分析Mini案例

```r
#读取数据
library(readxl)
cjb <- read_excel("data/cjb.xlsx")
View(cjb)
#对数据进行探索性分析
library(tidyverse)
cjb %>%
  select(sx, wlfk) %>%
  ggplot(aes(x = wlfk, y = sx, fill = wlfk)) +
  geom_boxplot()
#数据预处理
as_five_grade_scores <- function(x) {
  cut(x, breaks = c(0, seq(60, 100, by = 10)),
      labels = c("不及格", "及格", "中", "良", "优"))
}
cjb <- cjb %>%
  filter(zcj != 0) %>%#剔除脏数据
  mutate_at(vars(xb, wlfk), factor) %>% #类型转换
  mutate_at(vars(yw:sw), as_five_grade_scores)#数据分箱
View(cjb)
#建模
library(arulesViz)
my_model <- cjb %>%
  select(xb:wlfk) %>%
  apriori(parameter = list(supp = 0.06, conf = 0.8),
  appearance = list(rhs = paste0("wlfk=", c("文科", "理科"))))
#模型评估
inspectDT(my_model)
#可视化
plot(my_model)
```



R语言数据分析Mini案例

```r
#读取数据
library(readxl)
cjb <- read_excel("data/cjb.xlsx")
View(cjb)
#对数据进行探索性分析
library(tidyverse)
cjb %>%
  select(sx, wlfk) %>%
  ggplot(aes(x = wlfk, y = sx, fill = wlfk)) +
  geom_boxplot()
#数据预处理
as_five_grade_scores <- function(x) {
  cut(x, breaks = c(0, seq(60, 100, by = 10)),
      labels = c("不及格", "及格", "中", "良", "优"))
}
cjb <- cjb %>%
  filter(zcj != 0) %>%#剔除脏数据
  mutate_at(vars(xb, wlfk), factor) %>% #类型转换
  mutate_at(vars(yw:sw), as_five_grade_scores)#数据分箱
View(cjb)
#建模
library(arulesViz)
my_model <- cjb %>%
  select(xb:wlfk) %>%
  apriori(parameter = list(supp = 0.06, conf = 0.8),
  appearance = list(rhs = paste0("wlfk=", c("文科", "理科"))))
#模型评估
inspectDT(my_model)
#可视化
plot(my_model)
```



谢谢聆听

Thank you

教师个人联系方式

艾新波

手机：13641159546

QQ： 23127789

微信：13641159546

E-mail: 13641159546@126.com

axb@bupt.edu.cn

地址：北京邮电大学科研楼917室

课程网址：https://github.com/byaxb/RDataAnalytics