

中國地質大學



《商务智能上机报告》

报告题目： 大学生体测数据挖掘与分析

指导老师： 李四福

小组成员： 贲雅雯、郭思琪、袁晴、徐嘉艺、

袁应安、普叶、马宸晨

学生专业： 信息管理与信息系统

学生班级： 086191、086192、086193

报告日期： 2022年4月15日

目录

1 概述.....	4
2 数据预处理.....	4
2.1 数据总体概述.....	4
2.2 数据预处理需求.....	5
2.2.1 缺失数据处理.....	5
2.2.2 数据维度统一化.....	6
2.2.3 数据格式转换.....	6
2.2.4 新数据维度生成.....	7
2.3 数据预处理结果.....	8
3 描述性统计分析.....	9
3.1 地理热力图.....	9
3.2 2014-2017 样本分布.....	11
3.2.1 2014-2017 体测性别对比.....	11
3.2.2 2014-2017 体测年级对比.....	14
3.2.3 2014-2017 学院-等级桑基图.....	17
3.2.4 2014-2017 生源地地图.....	19
3.2.5 2014-2017 年总成绩分布密度图.....	25
3.2.6 2014-2017 等级分布条形图.....	28
3.2.7 2014-2017 年单项成绩分布山脊图.....	30
3.2.8 2014-2017 单项成绩平均分.....	32
3.3 2017 男女数据对比.....	34
3.3.1 总成绩分布密度图.....	34
3.3.2 单项成绩分布山脊图.....	35
3.3.3 单项成绩平均分雷达图.....	37
3.3.4 等级分布弦图.....	38
3.4 2017 生源地数据对比.....	40
3.4.1 总成绩平均分地图.....	40
3.4.2 等级分布弦图.....	42
3.5 2017 民族数据对比.....	43
3.5.1 总成绩小提琴图.....	43
3.5.2 单项成绩小提琴图.....	44
3.5.3 等级分布弦图.....	45
3.6 2017 学院数据对比.....	47
3.6.1 总成绩小提琴图.....	47
3.6.2 单项成绩小提琴图.....	48
3.6.3 等级比率条形图.....	49
3.7 经管学院各专业数据对比.....	50
3.7.1 各专业总成绩小提琴图.....	50
3.7.2 2017 年经济管理学院单项成绩分布山脊图.....	51

3.7.3 2017 年经管各专业单项成绩平均分雷达图.....	51
3.7.4 经管学院各专业等级分布弦图.....	54
3.8 2017 年各年级数据对比.....	55
3.8.1 总成绩分布密度图.....	55
3.8.2 2017 年各年级单项成绩分布山脊图.....	57
3.8.3 2014-2017 单项成绩平均分雷达图.....	59
3.8.4 2014-2017 等级分布弦图.....	62
3.8.5 2014-2017 等级比率.....	63
3.9 2014 级学生四年成绩变化.....	64
3.9.1 四年的总成绩小提琴图/箱线图.....	64
3.9.2 2014 级学生大学期间单项成绩分布山脊图.....	65
3.9.3 2014 级学生单项成绩平均分变化折线图.....	67
4 推断性统计分析.....	70
4.1 主成分分析.....	70
4.1.1 模型建立.....	70
4.1.2 数据准备.....	71
4.1.3 分析结果.....	71
4.2 影响体测成绩因素决策树分析.....	72
4.2.1 模型的建立.....	73
4.2.2 模型的求解与分析.....	74
4.3 速度与耐力项目聚类分析.....	78
4.3.1 模型的建立.....	78
4.3.2 模型的求解与分析.....	79
5 结论与建议.....	83
6 参考文献.....	89
附录一：小组分工.....	90
附录二：贲雅雯课程学习报告.....	91
附录三：郭思琪课程学习报告.....	106
附录四：袁晴课程学习报告.....	117
附录五：徐嘉艺课程学习报告.....	123
附录六：袁应安课程学习报告.....	128
附录七：普叶课程学习报告.....	136
附录八：马宸晨课程学习报告.....	142

1 概述

如今，随着我国经济和社会的高速发展，大学生的体质健康状况已成为越来越多的高校、社会人员甚至科研人员关注的重点。为了贯彻落实健康第一的指导思想，切实加强学校体育工作，促进学生积极参加体育锻炼，养成良好的锻炼习惯，提高体质健康水平制定，我国制定了国家学生体质锻炼标准。2019年6月，党中央、国务院发布《关于实施健康中国行动的意见》，提出了为加快推动以治病为中心转变为以人民健康为中心，动员全社会落实预防为主方针，实施健康中国行动^[1]。《意见》提出把高校学生体质健康状况纳入高校的考核评价，高校学生体质健康测试是促进学生体质健康提升，保证全民健康的重要支持。高校大学生体质健康测试工作已经成为各类高校考核评价的重要标准，经历不断的改革与创新，学生体质测试工作已经成为学校的一项政策性事宜^[2-3]。2014年6月教育部印发《学生体质健康监测评价办法》和《高等学校体育工作标准》等多个文件，并制定了《国家学生体质健康标准(2014年修订)》^[4]，以下简称《标准》。我国青少年体质健康状况呈现20年以上连续下降趋势，^[5]究其原因发现，影响学生体质健康状况下降的因素有很多，学生体质健康水平下降是各种因素综合作用的结果。大学生作为未来祖国建设的后备人才，健康的身体和强壮的体魄无疑是重要的，因此，研究他们的身体素质具有重要意义。于是本次课设将对体测数据进行研究，动态显示近几年该校学生体测工作的变化规律及发展趋势。

R是一套完整的数据处理、计算和制图软件系统。其功能包括：数据存储和处理系统；数组运算工具（其向量、矩阵运算方面功能尤其强大）；完整连贯的统计分析工具；优秀的统计制图功能；简便而强大的编程语言：可操纵数据的输入和输出，可实现分支、循环，用户可自定义功能。本文以R语言为工具，对所拥有的体测数据数据进行基本的描述性统计分析，发现数据的分布规律，总结一般性的描述性结论，并进行数据的可视化。对数据进行进一步的深层次挖掘，对体测项目间，项目单项成绩与总成绩之间的的相关关系进行分析，以便对体测信息进行更深一层的分析。

2 数据预处理

2.1 数据总体概述

我们所应用的原始数据为中国地质大学2014年至2017年总共四年间的学生体测真实数据，原始数据主要包含两个维度，分别是学生个人基础信息（年级编号、班级编号、班级名称、学籍号、民族代码、姓名、性别、出生日期、学生来源、家庭住址）与学生体测成绩数据（身高、体重、肺活量、50米跑、立定跳远、坐位体前屈、800米跑、1000米跑）。其中，由于每年录入指标略有不同，导致数据略微存在差异。例如，在2017年的原始数据中，缺少了“家庭住址”的原始信息）。对于该情况我们将在后续操作中进行统一化处理。原始数据如图2-1所示。

级编	班级编号	班级名称	学籍号	民族代码	姓名	性别	出生日期	学生来源	家庭住址	身高	体重	肺活量	50米跑	立定跳远	位体前屈	100米跑	分钟仰卧起坐	引体向上
41	163143	音乐学(音)	20141001547	汉族	范吉昊	男	1996/7/22	42000000	湖北省	166	42.6	3468	8	205	11.4	3' 38		1
41	86142	信息管理与	20141003941	汉族	樊木	男	1996/3/7	42000000	湖北省	166.8	42.8	3895	8.4	205	14.9	4' 12		2
41	91142	英语	20141000990	汉族	何翔	男	1995/10/29	43000000	湖南省	161.3	43.5	3335	8.1	200	4.6	4' 16		4
41	86142	信息管理与	20141002410	汉族	杨甜野	男	1995/11/9	13000000	河北省	166.4	43.5	652	7.9	243	6.3	3' 54		2
41	173141	公共事业管	20141001110	汉族	吴荣波	男	1996/4/3	33000000	浙江省	169.4	45	3654	7.9	228	11.4	3' 51		3
41	83141	国际经济与	20141000721	汉族	玉作乾	男	1996/4/7	45000000	广西壮族自	169.5	45.7	4217	7.8	237	13.2	4' 09		7
41	71143	电子信息工	20141000073	汉族	熊起	男	1997/1/14	61000000	陕西省	164.9	45.8	3286	8.1	215	18.2	4' 0		0
41	122141	物理学(光)	20141000536	汉族	陈子明	男	1996/6/19	15000000	内蒙古自治	172.6	46.5	3457	8.8	189	0.8	6' 0		1
41	122142	物理学(光)	20141003055	汉族	冯昭阳	男	1996/11/15	41000000	河南省	159.7	46.6	2176	7.6	245	20.1	3' 54		0
41	122141	物理学(光)	20141001915	汉族	李星乾	男	1995/2/4	51000000	四川省	159.4	46.7	3060	8.4	212	10.4	4' 19		7
41	122141	物理学(光)	20141000298	汉族	郭少俊	男	1996/9/10	44000000	广东省	164.1	47	3872	8.4	202	11.1	4' 07		2
41	121141	数学与应用	20141002450	汉族	刘琦	男	1996/4/10	13000000	河北省	167.9	47.1	3441	9.3	218	16.3	3' 52		0
41	173141	公共事业管	20141001794	汉族	叶锟	男	1996/8/31	36000000	江西省	154.4	47.2	2903	7.7	197	10.8	4' 02		10
41	231141	自动化	20141003838	汉族	周颖	男	1995/10/10	42000000	湖北省	154.7	47.7	2988	7.6	220	21.1	4' 04		13
41	88141	统计学	20141003897	汉族	阮明仔	男	1994/8/18	42000000	湖北省	160	47.7	4540	8.8	189	15.7	4' 22		3
41	172142	行政管理	20141000770	汉族	雷都甫	男	1995/1/13	45000000	广西壮族自	154.4	47.9	1128	8.3	197	8.7	4' 03		3
41	173141	公共事业管	20141004010	汉族	唐文泽	男	1996/10/1	42000000	湖北省	170.4	47.0	4076	7.4	222	11.4	4' 02		2

图 2-1 学生体测原始数据 (2017)

2.2 数据预处理需求

为了更好地方便后续描述性统计分析与推断性统计分析的开展，我们需要对原始数据进行预处理，其主要操作包含如图 2-2 所示。

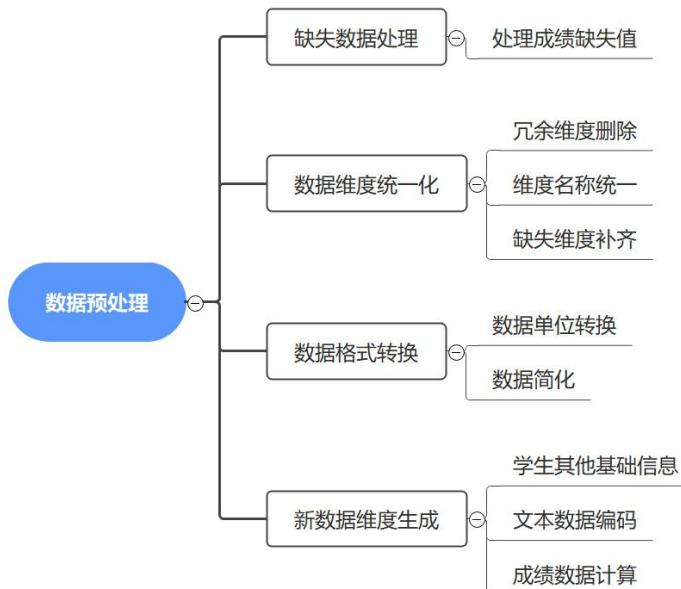


图 2-2 数据预处理步骤

2.2.1 缺失数据处理

通过观察数据，我们发现在这四年间的每一年体测数据中，主要缺失值存在于部分体测成绩的缺失。由于成绩数据的特殊性，我们考虑到成绩缺失的部分由人为丢失的情况较小，多为学生主动或因客观无法避免的因素缺考，所以我们将缺失的体测数据一律赋值为 0，代表该学生未曾参加测试，不拥有成绩。

2.2.2 数据维度统一化

在这一步，我们观察了四年间不同数据维度的细微差异（其差异如表 2-1 所示），我们发现了如下的数据处理需求并做了相关的处理：

- 2014 年和 2015 年拥有“一分钟跳绳”和“ 50×8 往返跑”的维度，而 2016 年和 2017 年没有。所以为了数据统一化，我们删去了这两个维度；
- 2015 年数据的“身份证号”维度对应的是其他三年数据中的“学生来源”维度，且其数据内容是相同的。因此，我们将 2015 年数据的“身份证号”维度概为“学生来源”维度；
- 在 2017 年数据中，缺少了家庭住址这一维度，因此，我们通过提取前三年之间的“学生来源——家庭住址”的映射关系，将 2017 年数据中的家庭住址补齐。

表 2-1 2014 年至 2017 年数据维度对比

年份	2014	2015	2016	2017
数据维度	年级编号	年级编号	年级编号	年级编号
	班级编号	班级编号	班级编号	班级编号
	班级名称	班级名称	班级名称	班级名称
	学籍号	学籍号	学籍号	学籍号
	民族代码	民族代码	民族代码	民族代码
	姓名	姓名	姓名	姓名
	性别	性别	性别	性别
	出生日期	出生日期	出生日期	出生日期
	学生来源	身份证号	学生来源	学生来源
	家庭住址	家庭住址	家庭住址	-
	身高	身高	身高	身高
	体重	体重	体重	体重
	肺活量	肺活量	肺活量	肺活量
	50 米跑	50 米跑	50 米跑	50 米跑
	立定跳远	立定跳远	立定跳远	立定跳远
	100 米跑	100 米跑	100 米跑	100 米跑
	一分钟仰卧起坐	一分钟仰卧起坐	一分钟仰卧起坐	一分钟仰卧起坐
	引体向上	引体向上	引体向上	引体向上
	一分钟跳绳	一分钟跳绳	-	-
	50×8 往返跑	50×8 往返跑	-	-

2.2.3 数据格式转换

为了方便后续进行相关的数据分析，我们有必要对其中的某些数据值进行相应的转换，其操作主要如下：

- 由于 BMI 的计算公式中，身高是以米为单位，而原数据中的身高数据以厘米为单位。

因此，将“身高”数据转换为以“米”为单位；

- 将 800 米跑与 1000 米跑的“分秒”计时格式转换为以秒计时；
- 我们观察到，“学生来源”维度中的内容为实际的个人身份证件编号，由于隐私保护，原数据只保留了前两位的数字，因此，我们只提取出前两位数字，用于作为省份编码。

2.2.4 新数据维度生成

由于原始数据存在一定的限制性，我们需要在原始数据中生成一些新的数据维度，其主要包含三方面的数据维度类别：学生其他基础信息生成、文本数据编码和成绩数据计算。

(1) 学生其他基础信息生成操作的目的是提取更详细的学生信息，例如学院信息等。其主要处理内容及处理方式如下：

- **学院信息：**取班级编号的前两位；
- **专业信息：**取班级编号的前三位。

(2) 文本数据编码操作目的是方便后续的推断性统计分析，其处理的内容及处理方式如下：

- **班级名称：**取班级编号的前三位；
- **民族代码：**统计所有出现的民族，并依次附上编码；
- **性别：**女生为 1，男生为 2。

(3) 成绩数据计算的目的是为了利用原始体测数据统计出学生个人成绩，包括单项成绩与总成绩。单项成绩与总成绩的计算方式（学年成绩=BMI * 15%+肺活量 * 15%+50m 跑 * 20%+坐位体前屈 * 10%+立定跳远 * 10%+引体向上/仰卧起坐 * 10%+1000m/800m * 20%）参考全国统一标准，其详情如下图：

等级	单项得分	肺活量 (ml)		50米		体前屈		立定跳远		仰卧起坐		800米	
		大一	大三	大一	大三	大一	大三	大一	大三	大一	大三	大一	大三
优秀	100	3400	3450	7.5	7.4	25.8	26.3	207	208	56	57	3'18"	3'16"
	95	3350	3400	7.6	7.5	24	24.4	201	202	54	55	3'24"	3'22"
	90	3300	3350	7.7	7.6	22.2	22.4	195	196	52	53	3'30"	3'28"
	85	3150	3200	8	7.9	20.6	21	188	189	49	50	3'37"	3'35"
良好	80	3000	3050	8.3	8.2	19	19.5	181	182	46	47	3'44"	3'42"
	78	2900	2950	8.5	8.4	17.7	18.2	178	179	44	45	3'49"	3'47"
	76	2800	2850	8.7	8.6	16.4	16.9	175	176	42	43	3'54"	3'52"
	74	2700	2750	8.9	8.8	15.1	15.6	172	173	40	41	3'59"	3'57"
及格	72	2600	2650	9.1	9	13.8	14.3	169	170	38	39	4'04"	4'02"
	70	2500	2550	9.3	9.2	12.5	13	166	167	36	37	4'09"	4'07"
	68	2400	2450	9.5	9.4	11.2	11.7	163	164	34	35	4'14"	4'12"
	66	2300	2350	9.7	9.6	9.9	10.4	160	161	32	33	4'19"	4'17"
不及格	64	2200	2250	9.9	9.8	8.6	9.1	157	158	30	31	4'24"	4'22"
	62	2100	2150	10.1	10	7.3	7.8	154	155	28	29	4'29"	4'27"
	60	2000	2050	10.3	10.2	6	6.5	151	152	26	27	4'34"	4'32"
	50	1960	2010	10.5	10.4	5.2	5.7	146	147	24	25	4'44"	4'42"
	40	1920	1970	10.7	10.6	4.4	4.9	141	142	22	23	4'54"	4'52"
	30	1880	1930	10.9	10.8	3.6	4.1	136	137	20	21	5'04"	5'02"
	20	1840	1890	11.1	11	2.8	3.3	131	132	18	19	5'14"	5'12"
	10	1800	1850	11.3	11.2	2	2.5	126	127	16	17	5'24"	5'22"

图 2-3 大学生体测成绩表 (女)

3 描述性统计分析

3.1 地理热力图

地理热力图是以特殊高亮的形式显示用户的地理位置，借助热力图，可以直观地观察到用户的总体情况和不同地域的指标分布情况。由于研究群体的共性指标为身高体重（譬如男生体测长跑项目为1000米而女生为800米，男生需要测引体向上而女生不需要等等），基于此我们绘制了不同地域学生群体的平均身高-平均体重热力分布图，从描述性统计的角度进行两方面研究：

（1）身高与体重间是否存在相关关系；（2）地域的不同是否会对身高产生显著影响。

值得注意的是，由于身高和体重数据平均值并无显著差异，倘若直接绘制热力图将难以得到直观信息，因此此部分对各省份学生群体的身高、体重数据先进行标准化然后扩大100倍以放大差异，使得热力图更加直观。

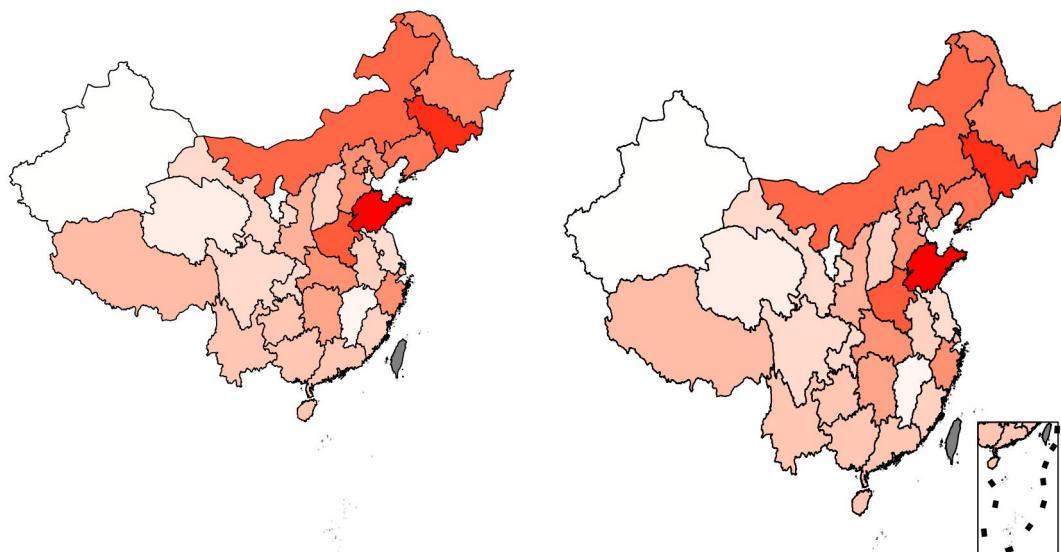


图 1 身高-体重分布情况地理热力图

从上述身高-体重分布情况地理热力图可以看出，对于身高情况（A部分），北方城市、东北部分的学生群体要显著高于其他地方的学生。A、B（体重情况）部分的横向比较也得得到身高与体重基本上是呈现正相关的，这是从相关数据的描述性统计得到的信息，严谨的判断还需要借助假设检验等推断性统计。

主要程序：

1. library(rgdal)
2. library(ggplot2)
3. library(cowplot)
4. library(dplyr)
5. china <- readOGR("E:\\Rplotexercise\\bou2_4p.shp")
6. china2 <- fortify(china)

```

7. mymap <- china@data
8. mymap$id <- 0:924
9.
10. china.map2 <- plyr::join(mymap, china2)
11.
12. mydata <- read.csv('E:\\Rplotexercise\\code87data.csv')
13.
14. datamap <- full_join(china.map2, mydata)
15. p1 <- ggplot() + geom_polygon(data=datamap,
16.                                     aes(x=long, y=lat, group = group, fill=HBOR),
17.                                     colour="black", size = .2) +
18.                                     scale_fill_gradient2(low = "green", mid = 'white', high = "red",
19.                                     midpoint = 1) +
20.                                     theme_void() + labs(x="", y="") +
21.                                     guides(fill = guide_colorbar(title='Height')) +
22.                                     theme(legend.position = c(0.15, 0.2))
23. df_China <- datamap
24. Width<-9
25. Height<-9
26. long_Start<-124
27. lat_Start<-16
28. df_Nanhai<-df_China[df_China$long>106.55 & df_China$long<123.58,]
29. df_Nanhai<-df_Nanhai[df_Nanhai$lat>4.61 & df_Nanhai$lat<25.45,]
30. min_long<-min(df_Nanhai$long, na.rm = TRUE)
31. min_lat<-min(df_Nanhai$lat, na.rm = TRUE)
32. max_long<-max(df_Nanhai$long, na.rm = TRUE)
33. max_lat<-max(df_Nanhai$lat, na.rm = TRUE)
34. df_Nanhai$long<-(df_Nanhai$long-min_long)/(max_long-min_long)*Width+long_Start
35. df_Nanhai$lat<-(df_Nanhai$lat-min_lat)/(max_lat-min_lat)*Height+lat_Start
36. df_Nanhai$class<-rep("NanHai", nrow(df_Nanhai))
37. df_China$class<-rep("Mainland", nrow(df_China))
38. df_China<-rbind(df_China, df_Nanhai)
39. df_NanHaiLine <- read.csv("E:\\Rplotexercise\\中国南海九段线.csv")
40. colnames(df_NanHaiLine)<-c("long", "lat", "ID")
41. df_NanHaiLine$long<-(df_NanHaiLine$long-min_long)/(max_long-min_long)*Width+long_Start
42. df_NanHaiLine$lat<-(df_NanHaiLine$lat-min_lat)/(max_lat-min_lat)*Height+lat_Start
43. p2 <- ggplot() +
44.   geom_polygon(data=df_China, aes(x=long, y=lat, group=interaction(class, group),
45.                                     fill=HBOR), colour="black", size=0.25) +

```

```

46. geom_rect(aes(xmin=long_Start, xmax=long_Start+Width+0.8,
47.             ymin=lat_Start-0.6, ymax=lat_Start+Height),fill=NA,
48.             colour="black",size=0.25)+  

49. geom_line(data=df_NanHaiLine, aes(x=long, y=lat, group=ID),
50.             colour="black", size=1)+  

51. scale_fill_gradient2(low = "green",mid = 'white', high = "red",
52.                       midpoint = 1)+  

53. coord_cartesian()+
54. ylim(15,55)+  

55. theme_void()+labs(x="", y="")+
56. theme(  

57. legend.position=c(0.15,0.2),
58. legend.background = element_blank())
59. plot_grid(p1,p2,ncol=2,align = 'h',labels = c('A','B'))  

60.  

61.

```

3.2 2014-2017 样本分布

3.2.1 2014-2017 体测性别对比

一、2014 年体测性别对比

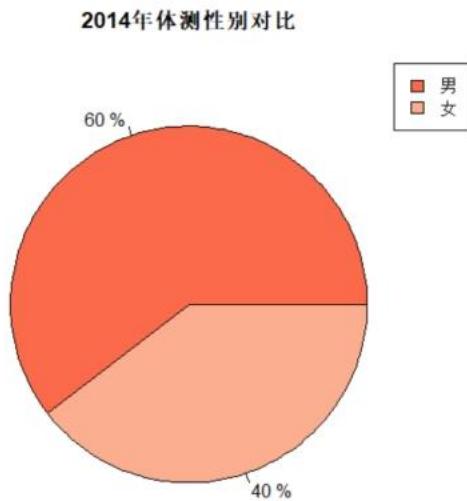


图 1 2014 年体测性别对比

描述：2014 年参加体测的同学中，男生占 60%，女生占 40%。

代码：

```

info = c(2835, 1857)  

names = c("男", "女")# 命名  

cols = c("#FB6A4A", "#FCAE91")# 涂色

```

```
piepercent = paste(round(100*info/sum(info)), "%")# 计算百分比  
pie(info, labels=piepercent, main = "2014 年体测性别对比", col=cols, family='GB1') # 绘图  
legend("topright", names, cex=0.8, fill=cols) # 添加颜色样本标注
```

二、2015 年体测性别对比

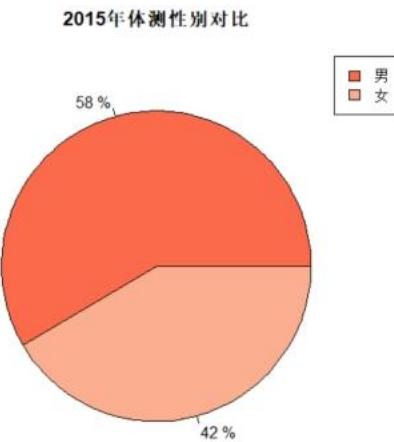


图 2 2015 年体测性别对比

描述：2015 年参加体测的同学中，男生占 58%，女生占 42%。

代码：

```
info = c(2897,2056) # 数据准备  
names = c("男", "女")# 命名  
cols = c("#FB6A4A","#FCAE91")# 涂色  
piepercent = paste(round(100*info/sum(info)), "%")# 计算百分比  
pie(info, labels=piepercent, main = "2015 年体测性别对比", col=cols, family='GB1') # 绘图  
legend("topright", names, cex=0.8, fill=cols) # 添加颜色样本标注
```

三、2016 年体测性别对比

2016年体测性别对比

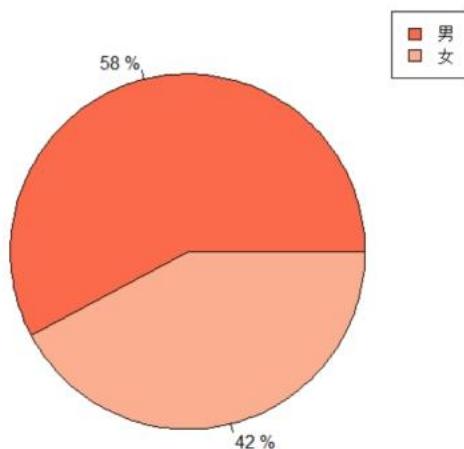


图 3 2016 年体测性别对比

描述：2016 年参加体测的同学中，男生占 58%，女生占 42%。

代码：

```
info = c(2925,2133) # 数据准备  
names = c("男", "女")# 命名  
cols = c("#FB6A4A","#FCAE91")# 涂色  
piepercent = paste(round(100*info/sum(info)), "%")# 计算百分比  
pie(info, labels=piepercent, main = "2016 年体测性别对比", col=cols, family='GB1') # 绘图  
legend("topright", names, cex=0.8, fill=cols) # 添加颜色样本标注
```

四、2017 年体测性别对比

2017年体测性别对比

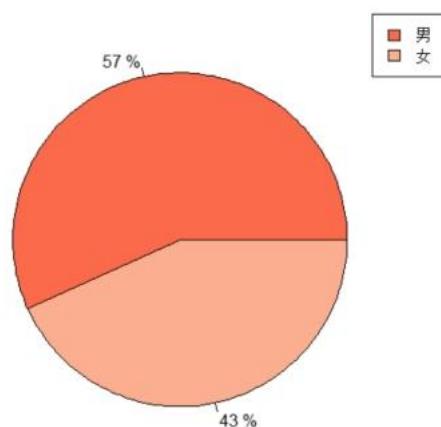


图 4 2017 年体测性别对比

描述：2017 年参加体测的同学中，男生占 57%，女生占 43%。

代码：

```

info = c(2882,2189) # 数据准备
names = c("男", "女")# 命名
cols = c("#FB6A4A", "#FCAE91")# 涂色
piepercent = paste(round(100*info/sum(info)), "%")# 计算百分比
pie(info, labels=piepercent, main = "2017 年体测性别对比", col=cols, family='GB1')# 绘图
legend("topright", names, cex=0.8, fill=cols) # 添加颜色样本标注

```

3.2.2 2014-2017 体测年级对比

一、2014 年体测年级对比

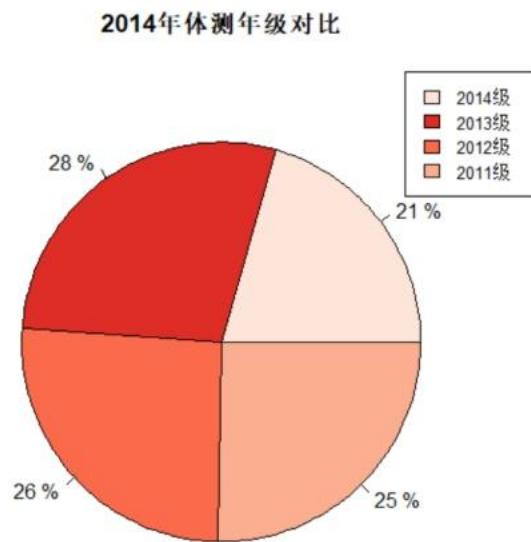


图 5 2014 年体测年级对比

描述：2014 年参加体测的同学中，2014 级同学占 21%，2013 级同学占 28%，2012 级同学占 26%，2011 级同学占 25%。

代码：

```

info = c(964,1330,1212,1186) # 数据准备
names = c("2014 级", "2013 级", "2012 级", "2011 级")# 命名
cols = c("#FEE5D9","#DE2D26","#FB6A4A","#FCAE91")# 涂色
piepercent = paste(round(100*info/sum(info)), "%")# 计算百分比
pie(info, labels=piepercent, main = "2014 年体测年级对比", col=cols, family='GB1')# 绘图
legend("topright", names, cex=0.8, fill=cols) # 添加颜色样本标注

```

二、2015 年体测年级对比

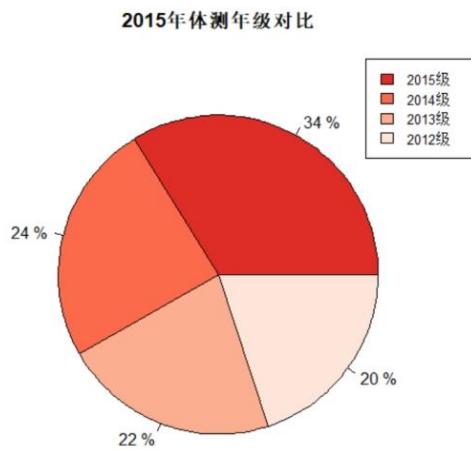


图 6 2015 年体测年级对比

描述: 2015 年参加体测的同学中, 2015 级同学占 34%, 2014 级同学占 24%, 2013 级同学占 22%, 2012 级同学占 20%。

代码:

```
info = c(1674,1213,1078,988) # 数据准备
names = c("2015 级", "2014 级", "2013 级", "2012 级")# 命名
cols = c("#DE2D26", "#FB6A4A", "#FCAE91", "#FEE5D9")# 涂色
piepercent = paste(round(100*info/sum(info)), "%")# 计算百分比
pie(info, labels=piepercent, main = "2015 年体测年级对比", col=cols, family='GB1') # 绘图
legend("topright", names, cex=0.8, fill=cols) # 添加颜色样本标注
```

三、2016 年体测年级对比

2016年体测年级对比

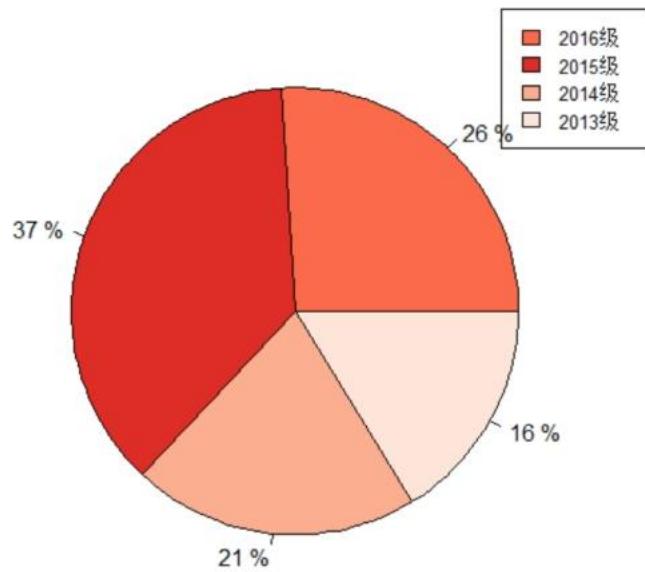


图 7 2016 年体测年级对比

描述: 2016 年参加体测的同学中, 2016 级同学占 26%, 2015 级同学占 37%, 2014 级同学占 21%, 2013 级同学占 16%。

代码:

```
info = c(1316,1871,1048,823) # 数据准备  
names = c("2016 级", "2015 级","2014 级", "2013 级")# 命名  
cols = c("#FB6A4A","#DE2D26" , "#FCAE91", "#FEE5D9")# 涂色  
piepercent = paste(round(100*info/sum(info)), "%")# 计算百分比  
pie(info, labels=piepercent, main = "2016 年体测年级对比", col=cols, family='GB1') # 绘图  
legend("topright", names, cex=0.8, fill=cols) # 添加颜色样本标注
```

四、2017 年体测年级对比

2017年体测年级对比

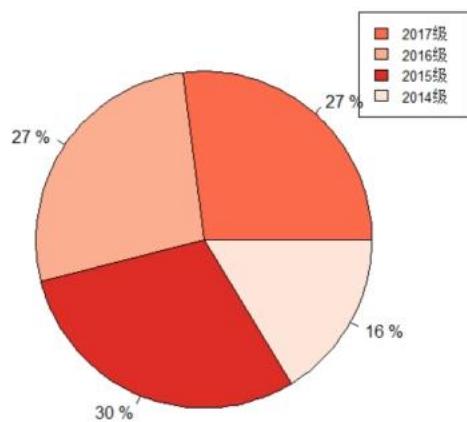


图 8 2017 年体测年级对比

描述: 2017 年参加体测的同学中, 2017 级同学占 27%, 2016 级同学占 27%, 2015 级同学占 30%, 2014 级同学占 16%。

代码

```
info = c(1371,1365,1508,827) # 数据准备
names = c("2017 级", "2016 级","2015 级", "2014 级")# 命名
cols = c("#FB6A4A","#FCAE91","#DE2D26","#FEE5D9")# 涂色
piepercent = paste(round(100*info/sum(info)), "%")# 计算百分比
pie(info, labels=piepercent, main = "2017 年体测年级对比", col=cols, family='GB1') # 绘图
legend("topright", names, cex=0.8, fill=cols) # 添加颜色样本标
```

3.2.3 2014-2017 学院-等级桑基图

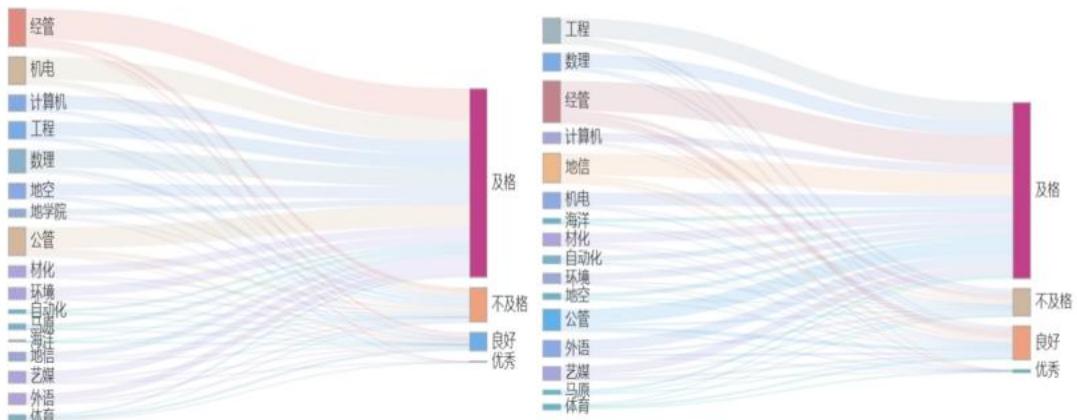


图 9 2014 年学院-等级桑基图

图 10 2015 年学院—等级桑基图

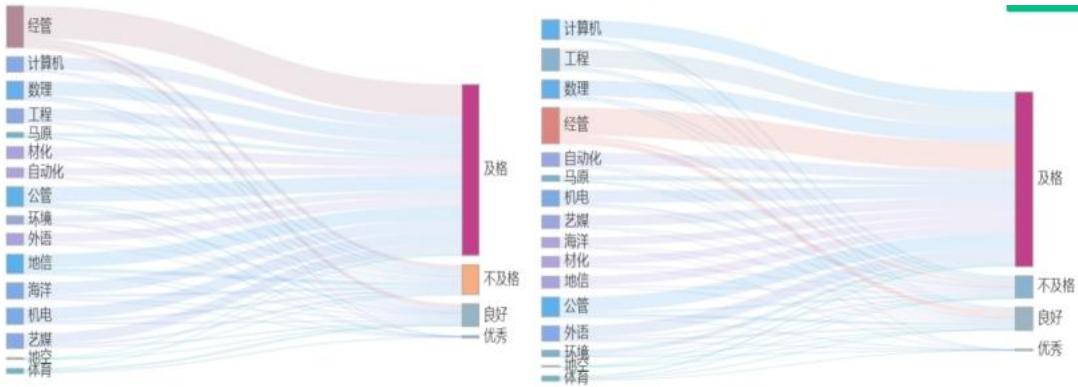


图 11 2016 年学院-等级桑基图

图 12 2017 年学院—等级桑基图

描述：在桑基图内部，不同的线条代表了不同的流量分流情况，线条的宽度代表此分支所代表的数据大小。综合四年情况分析，不同年份各学院的等级分布无明显差异。自 2014 年至 2017 年，各学院的绝大部分学生等级为及格，极少学生为优秀。其中，等级不及格、良好、优秀中各学院的人数均匀分布。

代码：

```

library(dplyr)
library(echarts4r)
library(highcharter)
library(htmlwidgets)

data<-read.csv(file="C:/Users/23842/Desktop/2014.csv",header=T,encoding="GBK")
Counts1->table(data$学院)
Counts2->table(data$等级)
t1 <- sample(x = c("材化","环境","工程","地空","机电","经管","外语","地信","数理","体育","艺媒
","公管","马原","计算机","自动化","海洋"),Counts1 , size = 100, replace=TRUE)
t2 <- sample(x = c("不及格", "及格","良好","优秀") ,Counts2, size = 100, replace=TRUE)
d <- bind_cols(t1,t2)
d%>%data_to_sankey()%>%hchart('sankey')

data<-read.csv(file="C:/Users/23842/Desktop/2015.csv",header=T,encoding="GBK")
Counts1->table(data$学院)
Counts2->table(data$等级)
t1 <- sample(x = c("材化","环境","工程","地空","机电","经管","外语","地信","数理","体育","艺媒
","公管","马原","计算机","自动化","海洋"),Counts1 , size = 100, replace=TRUE)
t2 <- sample(x = c("不及格", "及格","良好","优秀") ,Counts2, size = 100, replace=TRUE)
d <- bind_cols(t1,t2)
d%>%data_to_sankey()%>%hchart('sankey')

```

```

data<-read.csv(file="C:/Users/23842/Desktop/2016.csv",header=T,encoding="GBK")
Counts1->table(data$学院)
Counts2->table(data$等级)
t1 <- sample(x = c("材化","环境","工程","地空","机电","经管","外语","地信","数理","体育","艺媒
","公管","马原","计算机","自动化","海洋"),Counts1 , size = 100, replace=TRUE)
t2 <- sample(x = c("不及格", "及格","良好","优秀") ,Counts2, size = 100, replace=TRUE)
d <- bind_cols(t1,t2)
d%>%data_to_sankey()%>%hchart('sankey')
data<-read.csv(file="C:/Users/23842/Desktop/2017.csv",header=T,encoding="GBK")
Counts1->table(data$学院)
Counts2->table(data$等级)
t1 <- sample(x = c("材化","环境","工程","地空","机电","经管","外语","地信","数理","体育","艺媒
","公管","马原","计算机","自动化","海洋"),Counts1 , size = 100, replace=TRUE)
t2 <- sample(x = c("不及格", "及格","良好","优秀") ,Counts2, size = 100, replace=TRUE)
d <- bind_cols(t1,t2)
d%>%data_to_sankey()%>%hchart('sankey')

```

3.2.4 2014-2017 生源地地图

一、2014 生源地地图

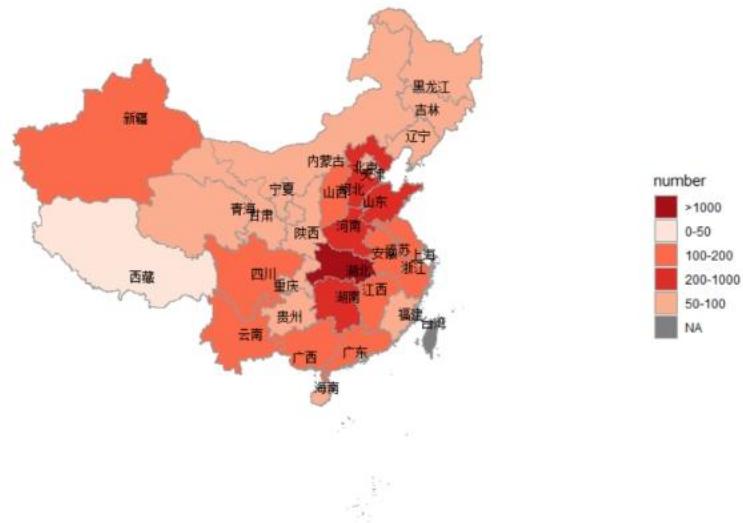


图 13 2014 生源地地图

描述：2014 年参加体测的地大学生生源地遍及中国大陆各个省及直辖市；其中，只有来自西藏自治区和上海市的学生数量小于 50，属于生源数量很少的地区；来自黑龙江省、内蒙古自治区、吉林省、辽宁省、甘肃省、北京市、天津市、陕西省、宁夏回族自治区、青海省、重庆市、贵州省、

福建省、海南省等 14 个省或直辖市的学生数量在 50 和 100 之间，属于生源数量较少的地区；来自新疆维吾尔自治区、山西省、安徽省、江苏省、四川省、浙江省、江西省、云南省、广西壮族自治区、广东省等 10 个省或直辖市的学生数量在 100 到 200 之间，属于生源数量中等的地区；来自河北省、山东省、河南省、湖南省等 4 个省的学生数量在 200 道 1000 之间，属于生源数量较多的地区，同时这些地区确实是高考大省；来自湖北省的学生数量超过 1000，属于生源数量很多的地区。

代码：

```
library(mapdata)
library(maptools)
library(ggplot2)
library(plyr)#引用包
china_map=readShapePoly("C:/Rfiles/map/bou2_4p.shp")#读取地图数据
ggplot(china_map,aes(x=long,y=lat,group=group))      +geom_polygon(fill="white",colour="grey")
+coord_map("polyconic")#绘制并投影得到可用地图
x <- china_map@data#读取行政信息
xs <- data.frame(x,id=seq(0:924)-1)#含岛屿共 925 个形状
china_map1 <- fortify(china_map) #转化为数据框
china_map_data <- join(china_map1, xs, type = "full") #合并两个数据框
mydata<-read.csv("C:\\Rfiles\\2014 附总成绩.csv")#读取业务数据
china_data <- join(china_map_data, mydata, type="full")#合并两个数据框
province_city <- read.csv("C:/Rfiles/china-cities.csv") #读取省市经纬度
memory.limit(10000000)# 设置约为 1G内存，否则会显示无法分配
ggplot(china_data,aes(long,lat))+ 
  geom_polygon(aes(group=group,fill=number),colour="grey60")+
  scale_fill_manual(values=c("#A50F15","#FEE5D9","#FB6A4A","#DE2D26","#FCAE91"))+
  coord_map("polyconic") +
  geom_text(aes(x = lon,y = lat,label = province), data =province_city, size=3 )+
  theme(
    panel.grid = element_blank(),
    panel.background = element_blank(),
    axis.text = element_blank(),
    axis.ticks = element_blank(),
    axis.title = element_blank()
  )#确定颜色并作图
```

二、2015 生源地地图

2015年生源地地图

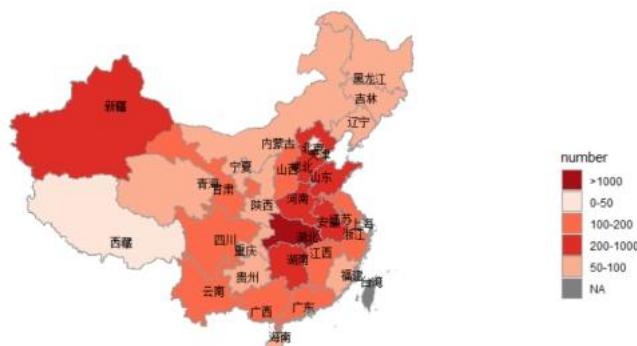


图 14 2015 生源地地图

描述：2015 年参加体测的地大学生生源地遍及中国大陆各个省及直辖市。其中，来自西藏自治区、上海市和北京市等 3 个省或直辖市的学生数量小于 50，属于生源数量很少的地区；来自黑龙江省、内蒙古自治区、吉林省、辽宁省、天津市、陕西省、宁夏回族自治区、青海省、重庆市、贵州省、福建省、海南省等 12 个省或直辖市的学生数量在 50 和 100 之间，属于生源数量较少的地区；来自甘肃省、山西省、江苏省、四川省、浙江省、江西省、云南省、广西壮族自治区、广东省等 9 个省或直辖市的学生数量在 100 到 200 之间，属于生源数量中等的地区；来自新疆维吾尔自治区、河北省、山东省、河南省、安徽省、湖南省等 6 个省的学生数量在 200 到 1000 之间，属于生源数量较多的地区；来自湖北省的学生数量超过 1000，属于生源数量很多的地区。

相较于 2014 年，生源数量区间发生变化的地区有北京市、新疆维吾尔自治区、甘肃省、安徽省。

代码：

```
library(mapdata)
library(maptools)
library(ggplot2)
library(plyr)#引用包
china_map=readShapePoly("C:/Rfiles/map/bou2_4p.shp")#读取地图数据
ggplot(china_map,aes(x=long,y=lat,group=group))      +geom_polygon(fill="white",colour="grey")
+coord_map("polyconic")#绘制并投影得到可用地图
x<-china_map@data#读取行政信息
xs<-data.frame(x,id=seq(0:924)-1)#含岛屿共 925 个形状
china_map1<-fortify(china_map) #转化为数据框
```

```

china_map_data <- join(china_map1, xs, type = "full") #合并两个数据框

mydata<-read.csv("C:\\Rfiles\\2015 附总成绩.csv")#读取业务数据
china_data <- join(china_map_data, mydata, type="full")#合并两个数据框
province_city <- read.csv("C:/Rfiles/china-cities.csv") #读取省市经纬度
memory.limit(10000000)# 设置约为 1G内存，否则会显示无法分配
ggplot(china_data,aes(long,lat))+

  geom_polygon(aes(group=group,fill=number),colour="grey60")+
  scale_fill_manual(values=c("#A50F15","#FEE5D9","#FB6A4A","#DE2D26","#FCAE91"))+
  coord_map("polyconic") +
  geom_text(aes(x = lon,y = lat,label = province), data = province_city, size=3 )+
  ggtitle("2015 年生源地地图")+
  theme(
    panel.grid = element_blank(),
    panel.background = element_blank(),
    axis.text = element_blank(),
    axis.ticks = element_blank(),
    axis.title = element_blank()
  )#确定颜色并作图

```

三、2016 生源地地图

2016年生源地地图

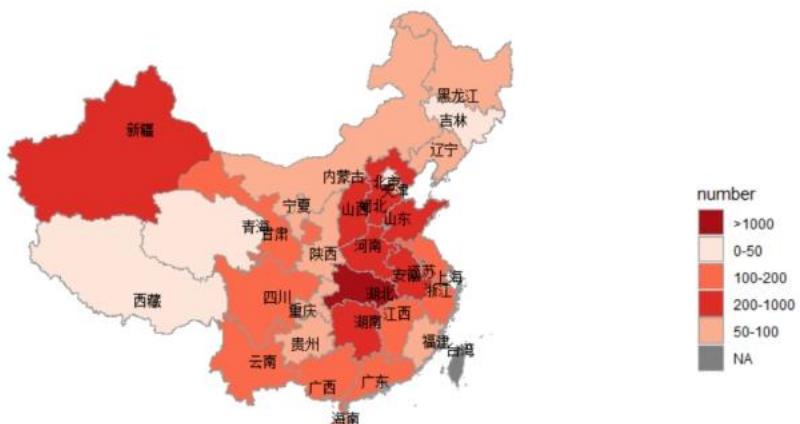


图 15 2016 生源地地图

描述：2016 年参加体测的地大学生生源地遍及中国大陆各个省及直辖市。其中，来自吉林省、青海省、西藏自治区、上海市和北京市等 5 个省或直辖市的学生数量小于 50，属于生源数量很少的地区；来自黑龙江省、内蒙古自治区、辽宁省、天津市、陕西省、宁夏回族自治区、重庆市、贵州省、福建省、海南省等 10 个省或直辖市的学生数量在 50 和 100 之间，属于生源数量较少的地区；来自甘肃省、江苏省、四川省、浙江省、江西省、云南省、广西壮族自治区、广东省等 8 个省或直辖市的学生数量在 100 到 200 之间，属于生源数量中等的地区；来自新疆维吾尔自治区、河北省、山西省、山东省、河南省、安徽省、湖南省等 7 个省的学生数量在 200 到 1000 之间，属于生源数量较多的地区；来自湖北省的学生数量超过 1000，属于生源数量很多的地区。

相较于 2015 年，生源数量区间发生变化的地区有吉林省、青海省、山西省。

代码

```
library(mapdata)
library(maptools)
library(ggplot2)
library(plyr)#引用包
china_map=readShapePoly("C:/Rfiles/map/bou2_4p.shp")#读取地图数据
ggplot(china_map,aes(x=long,y=lat,group=group))      +geom_polygon(fill="white",colour="grey")
+coord_map("polyconic")#绘制并投影得到可用地图
x <- china_map@data#读取行政信息
xs <- data.frame(x,id=seq(0:924)-1)#含岛屿共 925 个形状
china_map1 <- fortify(china_map) #转化为数据框
china_map_data <- join(china_map1, xs, type = "full") #合并两个数据框
mydata<-read.csv("C:\\Rfiles\\2016 附总成绩.csv")#读取业务数据
china_data <- join(china_map_data, mydata, type="full")#合并两个数据框
province_city <- read.csv("C:/Rfiles/china-cities.csv") #读取省市经纬度
memory.limit(10000000)# 设置约为 1G 内存，否则会显示无法分配
ggplot(china_data,aes(long,lat))+ 
  geom_polygon(aes(group=group,fill=number),colour="grey60")+
  scale_fill_manual(values=c("#A50F15","#FEE5D9","#FB6A4A","#DE2D26","#FCAE91"))+
  coord_map("polyconic") +
  geom_text(aes(x = lon,y = lat,label = province), data = province_city, size=3 )+
  ggtitle("2016 年生源地地图")+
  theme(
```

```

panel.grid = element_blank(),
panel.background = element_blank(),
axis.text = element_blank(),
axis.ticks = element_blank(),
axis.title = element_blank()
)#确定颜色并作图

```

四、2017 生源地地图

2017年生源地地图

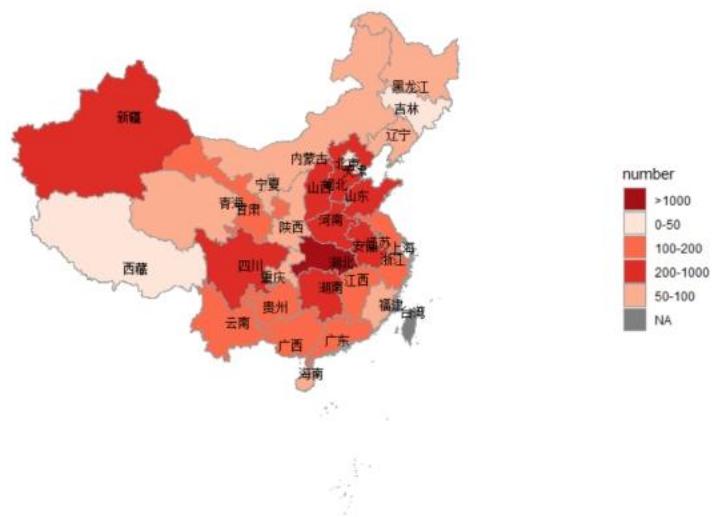


图 16 2017 生源地地图

描述：2017 年参加体测的地大学生生源地遍及中国大陆各个省及直辖市。其中，来自吉林省、天津市、西藏自治区、上海市和北京市等 5 个省或直辖市的学生数量小于 50，属于生源数量很少的地区；来自黑龙江省、内蒙古自治区、辽宁省、陕西省、宁夏回族自治区、重庆市、青海省、福建省、海南省等 9 个省或直辖市的学生数量在 50 和 100 之间，属于生源数量较少的地区；来自甘肃省、江苏省、浙江省、江西省、云南省、贵州省、广西壮族自治区、广东省等 8 个省或直辖市的学生数量在 100 到 200 之间，属于生源数量中等的地区；来自新疆维吾尔自治区、河北省、山西省、山东省、河南省、安徽省、四川省、湖南省等 8 个省的学生数量在 200 到 1000 之间，属于生源数量较多的地区；来自湖北省的学生数量超过 1000，属于生源数量很多的地区。

相较于 2016 年，生源数量区间发生变化的地区有天津市、青海省、四川省和贵州省。

代码：

```

library(mapdata)
library(maptools)
library(ggplot2)
library(plyr)#引用包

```

```

china_map=readShapePoly("C:/Rfiles/map/bou2_4p.shp")#读取地图数据
ggplot(china_map,aes(x=long,y=lat,group=group))      +geom_polygon(fill="white",colour="grey")
+coord_map("polyconic")#绘制并投影得到可用地图

x <- china_map@data#读取行政信息
xs <- data.frame(x,id=seq(0:924)-1)#含岛屿共 925 个形状
china_map1 <- fortify(china_map) #转化为数据框
china_map_data <- join(china_map1, xs, type = "full") #合并两个数据框
mydata<-read.csv("C:\\Rfiles\\2017 附总成绩.csv")#读取业务数据
china_data <- join(china_map_data, mydata, type="full")#合并两个数据框
province_city <- read.csv("C:/Rfiles/china-cities.csv") #读取省市经纬度
memory.limit(10000000)# 设置约为 1G内存，否则会显示无法分配
ggplot(china_data,aes(long,lat))+

  geom_polygon(aes(group=group,fill=number),colour="grey60")+
  scale_fill_manual(values=c("#A50F15","#FEE5D9","#FB6A4A","#DE2D26","#FCAE91"))+
  coord_map("polyconic") +
  geom_text(aes(x = lon,y = lat,label = province), data = province_city, size=3 )+
  ggtitle("2017 年生源地地图")+
  theme(
    panel.grid = element_blank(),
    panel.background = element_blank(),
    axis.text = element_blank(),
    axis.ticks = element_blank(),
    axis.title = element_blank()
  )#确定颜色并作图

```

3.2.5 2014-2017 年总成绩分布密度图

一、2014 总成绩分布密度图

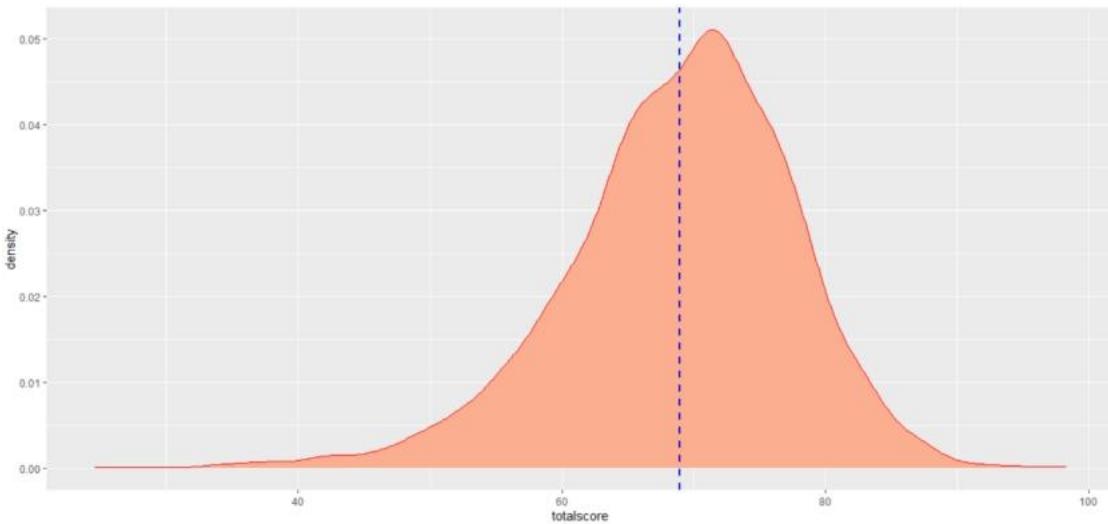


图 17 2014 总成绩分布密度图

描述：2014 年参加体测的全体同学的总成绩基本集中在 40-90 之间，平均值约为 69 分，且在 72 分左右的同学数量最多。

代码

```
library(ggplot2)
data<-read.csv("C:\\Rfiles\\2014 总成绩.csv", header=T)
p <- ggplot(data, aes(x=totalscore))+geom_density(color="red", fill="#FCAE91")
p+ geom_vline(aes(xintercept=mean(totalscore)),
              color="blue", linetype="dashed", size=1)
```

二、2015 年总成绩分布密度图

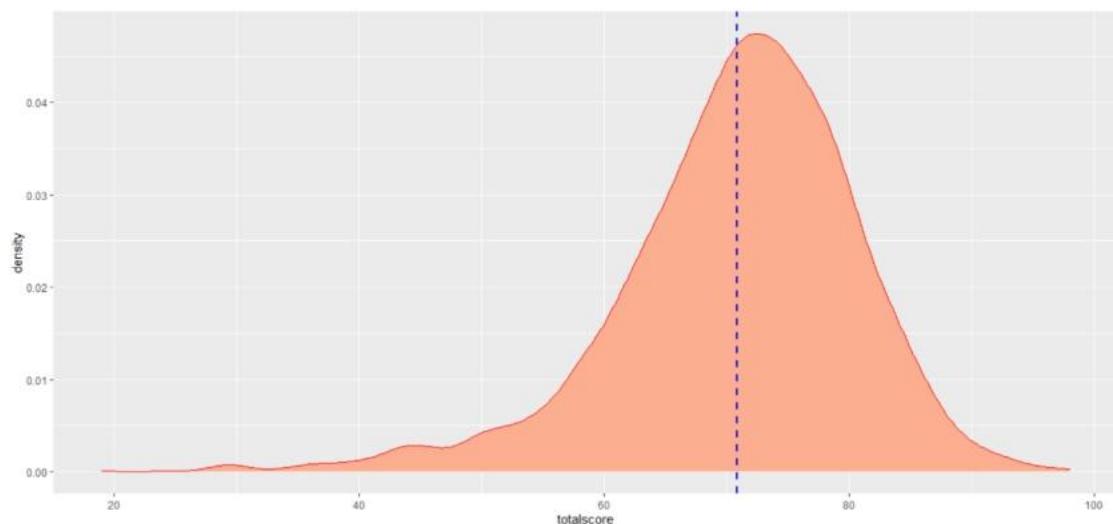


图 18 2015 总成绩分布密度图

描述：2015 年参加体测的全体同学的总成绩主要集中在 40-90 之间，平均值约为 71 分，且在 73 分左右的同学数量最多。

代码

```

library(ggplot2)
data<-read.csv("C:\\Rfiles\\2015 总成绩.csv", header=T)
p <- ggplot(data, aes(x=totalscore))+geom_density(color="red", fill="#FCAE91")
p+ geom_vline(aes(xintercept=mean(totalscore)),
              color="blue", linetype="dashed", size=1)

```

三、2016年总成绩分布密度图

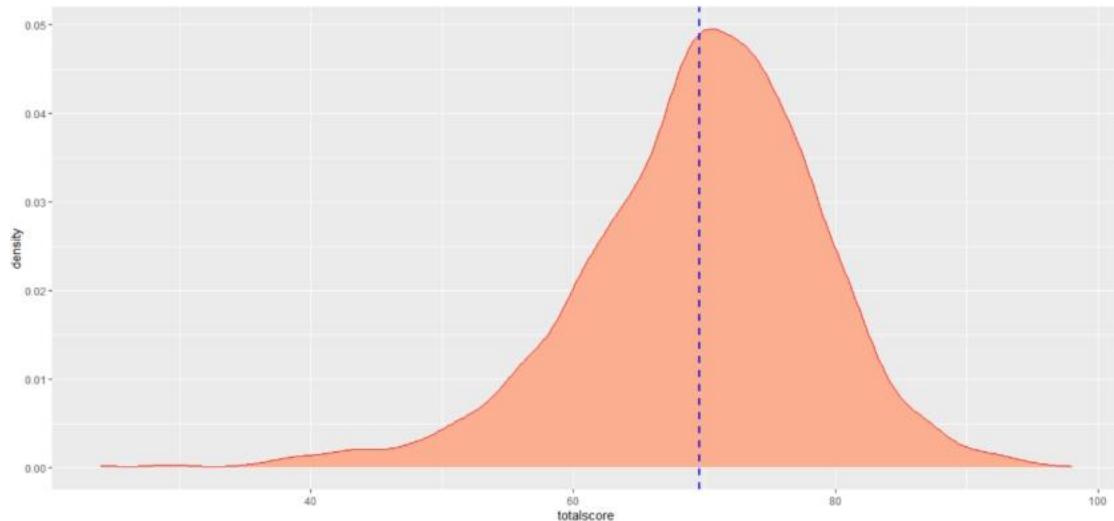


图 19 2016 总成绩分布密度图

描述：2016 年参加体测的全体同学的总成绩主要集中在 40-90 之间，平均值约为 69 分，且在 71 分左右的同学数量最多。

代码

```

library(ggplot2)
data<-read.csv("C:\\Rfiles\\2016 总成绩.csv", header=T)
p <- ggplot(data, aes(x=totalscore))+geom_density(color="red", fill="#FCAE91")
p+ geom_vline(aes(xintercept=mean(totalscore)),
              color="blue", linetype="dashed", size=1)

```

四、2017年总成绩分布密度图

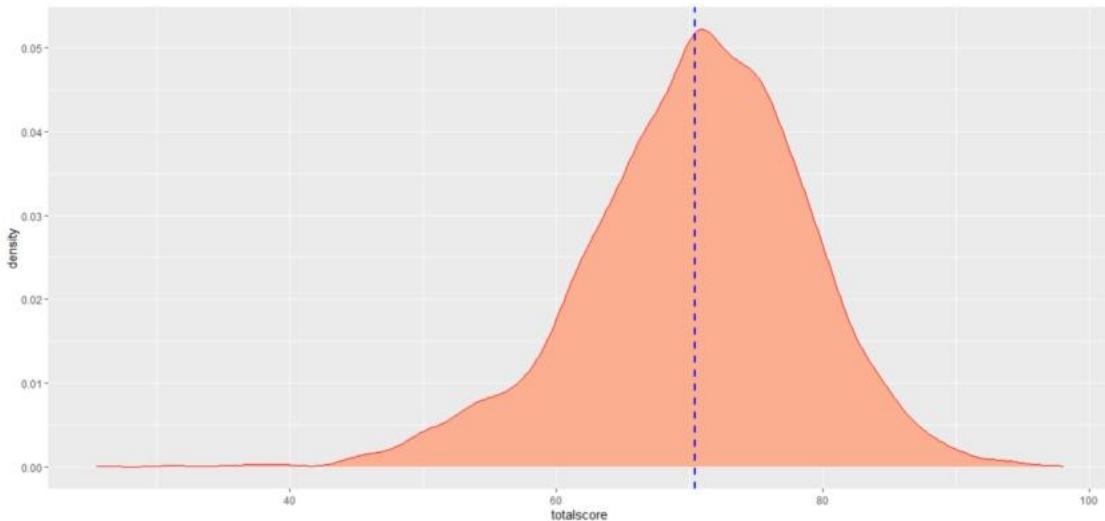


图 20 2017 总成绩分布密度图

描述：2017 年参加体测的全体同学的总成绩主要集中在 45-90 之间，平均值约为 70 分，且在 71 分左右的同学数量最多。

代码

```
library(ggplot2)
data<-read.csv("C:\\Rfiles\\2017 总成绩.csv", header=T)
p <- ggplot(data, aes(x=totalscore))+geom_density(color="red", fill="#FCAE91")
p+ geom_vline(aes(xintercept=mean(totalscore)),
              color="blue", linetype="dashed", size=1)
```

3.2.6 2014-2017 等级分布条形图

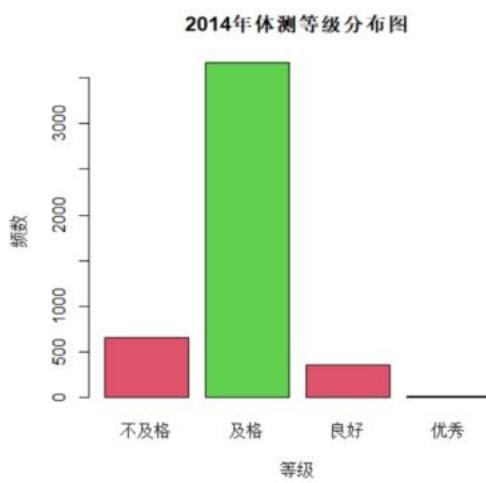


图 21 2014 体测等级分布图

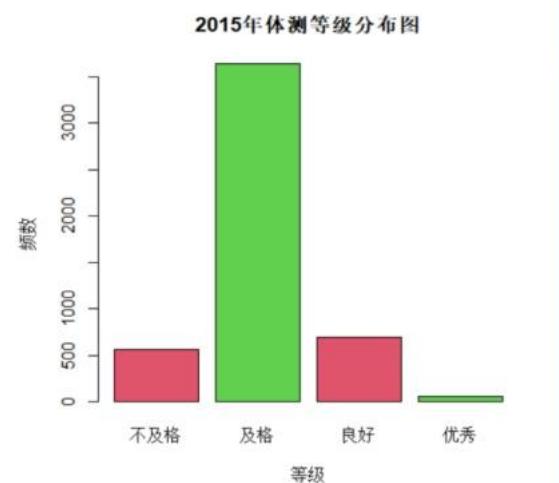


图 22 2015 体测等级分布图



图 23 2016 体测等级分布图

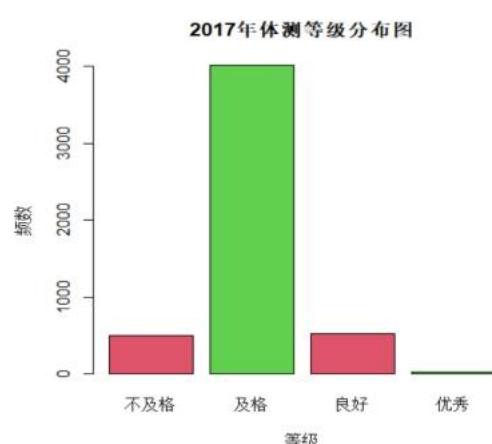


图 24 2017 体测等级分布图

描述：上图为 2014-2017 年每年全部学生的等级分布情况，由图可以看出，不同年份学生等级分布无明显差异。其中，等级为及格的人数最多，浮动在 4000 人左右；等级为优秀的人数最少，每年不过百人。

代码：

```

data<-read.csv(file="C:/Users/23842/Desktop/2014.csv",header=T,encoding="GBK"
count<-table(data$等级)
Barplot(count,xlab="等级",ylab="频数", main="2014 年体测等级分布图",col=1:1)

data<-read.csv(file="C:/Users/23842/Desktop/2015.csv",header=T,encoding="GBK"
count<-table(data$等级)
Barplot(count,xlab="等级",ylab="频数", main="2015 年体测等级分布图",col=1:1)

data<-read.csv(file="C:/Users/23842/Desktop/2016.csv",header=T,encoding="GBK"
count<-table(data$等级)
Barplot(count,xlab="等级",ylab="频数", main="2016 年体测等级分布图",col=1:1)

data<-read.csv(file="C:/Users/23842/Desktop/2017.csv",header=T,encoding="GBK"
count<-table(data$等级)
Barplot(count,xlab="等级",ylab="频数", main="2017 年体测等级分布图",col=1:1)

```

3.2.7 2014-2017年单项成绩分布山脊图

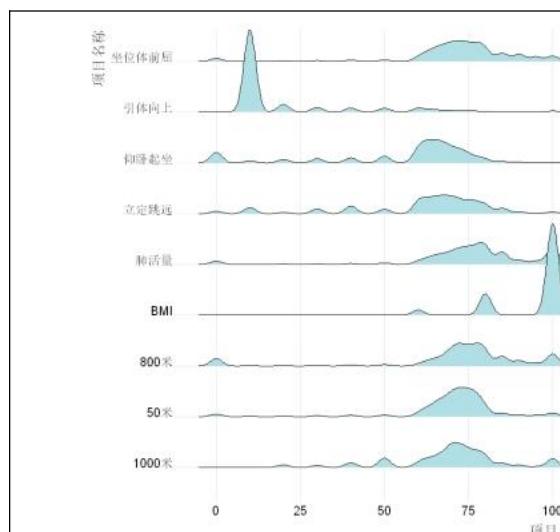


图 25 2014 年单项成绩分布图

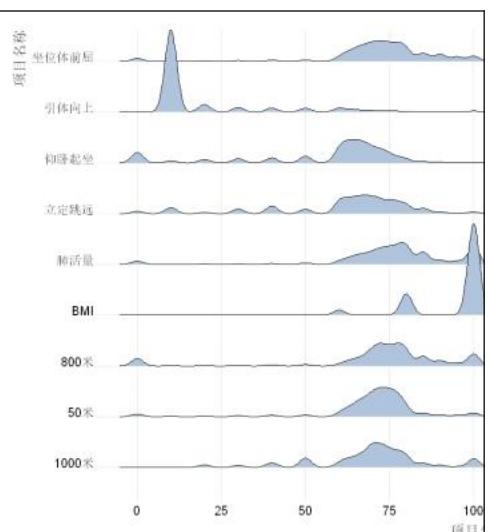


图 26 2015 年单项成绩分布图

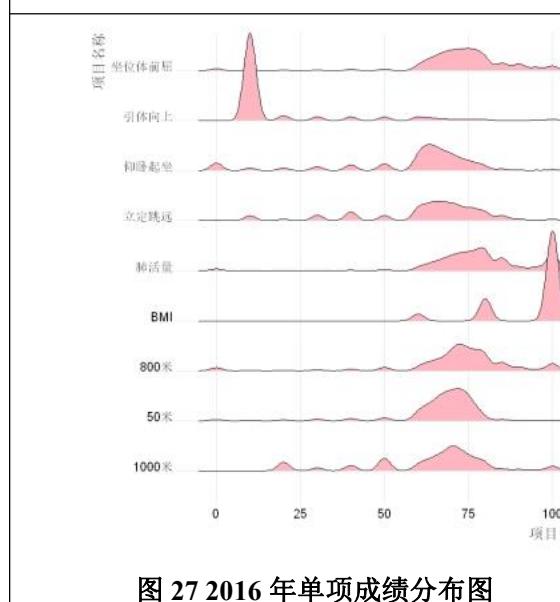


图 27 2016 年单项成绩分布图

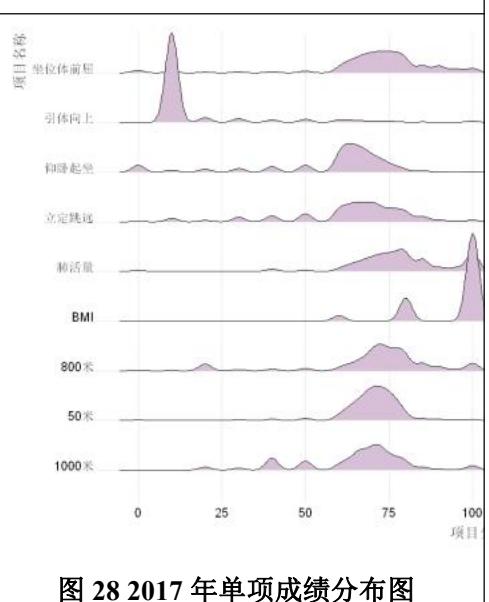


图 28 2017 年单项成绩分布图

描述：综合四年情况分析，不同年份单项成绩分布不呈现明显差异，坐位体前屈、仰卧起坐、立定跳远、肺活量、800m、50m成绩分布大多集中在中上成绩段，引体向上多集中于低分段，近两年 1000m 中低分段人数较 2014、2015 年人数略有增长，应加强对低分段男生的引体向上与长跑训练强度。

预处理代码

```
import numpy as np
import pandas as pd
from tqdm import trange
df = pd.read_csv(r'D:\Desktop\bi\data\2014.csv')
row = df.shape[0]
print(row)
```

```

data=[]
for i in range(row):
    ID = str(df['学号'].values[i])
    data.append([ID[:4],df['性别'].values[i],df['专业'].values[i],"BMI",df['BMI分数'].values[i]])
    data.append([ID[:4],df['性别'].values[i],df['专业'].values[i],"肺活量",df['肺活量分数'].values[i]])
    data.append([ID[:4],df['性别'].values[i],df['专业'].values[i],"50 米",df['50 米分数'].values[i]])
    data.append([ID[:4],df['性别'].values[i],df['专业'].values[i],"立定跳远",df['立定跳远分数'].values[i]])
    data.append([ID[:4],df['性别'].values[i],df['专业'].values[i],"坐位体前屈",df['坐位体前屈分数']
                ].values[i]])
    data.append([ID[:4],df['性别'].values[i],df['专业'].values[i],"800 米",df['800 米分数'].values[i]])
    data.append([ID[:4],df['性别'].values[i],df['专业'].values[i],"1000 米",df['1000 米分数'].values[i]])
    data.append([ID[:4],df['性别'].values[i],df['专业'].values[i],"仰卧起坐",df['仰卧起坐分数'].values[i]])
    data.append([ID[:4],df['性别'].values[i],df['专业'].values[i],"引体向上",df['引体向上分数'].values[i]])

score = pd.DataFrame(data,columns=['年级','性别','专业','项目名称', '项目分数'])
score = score.drop(score[pd.isnull(score['项目分数'])].index)
score.to_csv('score14.csv',index=False)

```

r语言代码

```

data1 <- read.csv(file="score17.csv")
ggplot(data1, aes(x = 项目分数 ,y = 项目名称, fill = "cut")) +
  geom_density_ridges() +
  theme_ridges() +
  theme(legend.position = "none")+
  scale_fill_manual(values=c("thistle"))
data2 <- read.csv(file="score16.csv")
ggplot(data2, aes(x = 项目分数 ,y = 项目名称, fill = "cut")) +
  geom_density_ridges() +
  theme_ridges() +
  theme(legend.position = "none")+
  scale_fill_manual(values=c("lightpink"))
data3 <- read.csv(file="score15.csv")
ggplot(data3, aes(x = 项目分数 ,y = 项目名称, fill = "cut")) +
  geom_density_ridges() +
  theme_ridges()

```

```

theme(legend.position = "none")+
scale_fill_manual(values=c("lightsteelblue"))

data4<- read.csv(file="score14.csv")

ggplot(data3, aes(x = 项目分数 ,y = 项目名称, fill = "cut")) +
geom_density_ridges() +
theme_ridges() +
theme(legend.position = "none")+
scale_fill_manual(values=c("powderblue"))

```

3.2.8 2014-2017 单项成绩平均分

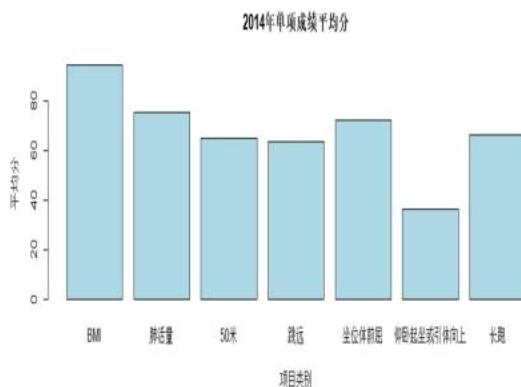


图 29 2014 年单项成绩平均分

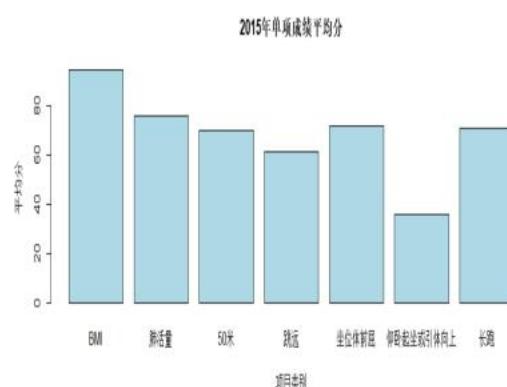


图 30 2015 年单项成绩平均分

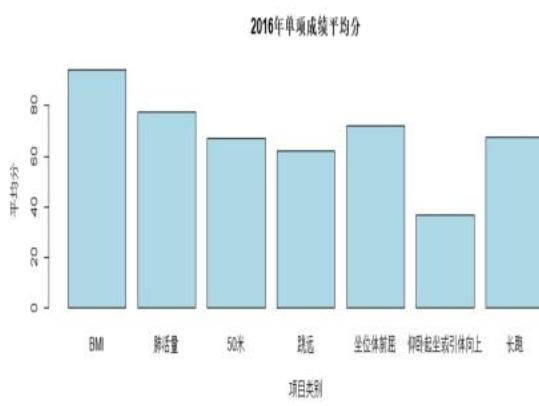


图 31 2016 年单项成绩平均分

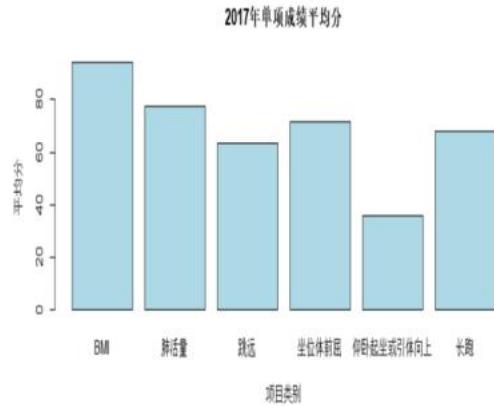


图 32 2017 年单项成绩平均分

描述：上图为 2014-2017 年每年全部学生的单项平均分分布情况，由图可以看出，不同年份学生等级分布无明显差异。综合每年成绩，学生 BMI 成绩最高，接近一百分，说明绝大部分学生身材匀称。肺活量、50 米、跳远、坐位体前屈、长跑成绩都位于 60-80 之间，而仰卧起坐或引体向上成绩最低，不到 40 分，说明学生亟待加强该方面体育锻炼。

代码：

```

> data<-read.csv(file="C:/Users/23842/Desktop/2017.csv",header=T,encoding="GBK")
> bmi=mean(data$BMI分数)
> vc=mean(data$肺活量分数)
> run=mean(data$50 米分数)
> jump=mean(data$立定跳远分数)
> sit=mean(data$坐位体前屈分数)
> up=mean(data$仰卧引体分数)
> run=mean(data$长跑分数)
> solo=c(bmi,vc,jump,sit,up,run)
> label_solo=c("BMI","肺活量","跳远","坐位体前屈","仰卧起坐或引体向上","长跑")
> barplot(solo,names.arg=label_solo,xlab="项目类别",ylab="平均分",col="lightblue",main="2017 年
单项成绩平均分")

> data<-read.csv(file="C:/Users/23842/Desktop/2014.csv",header=T,encoding="GBK")
> bmi=mean(data$BMI分数)
> vc=mean(data$肺活量分数)
> run=mean(data$五十米分数)
> jump=mean(data$立定跳远分数)
> sit=mean(data$坐位体前屈分数)
> up=mean(data$仰引分数)
> runn=mean(data$长跑分数)
> solo=c(bmi,vc,run,jump,sit,up,runn)
> label_solo=c("BMI","肺活量","50 米","跳远","坐位体前屈","仰卧起坐或引体向上","长跑")
> barplot(solo,names.arg=label_solo,xlab="项目类别",ylab="平均分",col="lightblue",main="2014 年
单项成绩平均分")

> data<-read.csv(file="C:/Users/23842/Desktop/2015.csv",header=T,encoding="GBK")
> bmi=mean(data$BMI分数)
> vc=mean(data$肺活量分数)
> run=mean(data$五十米分数)
> jump=mean(data$立定跳远分数)
> sit=mean(data$坐位体前屈分数)
> up=mean(data$仰卧引体分数)
> runn=mean(data$长跑分数)
> solo=c(bmi,vc,run,jump,sit,up,runn)

```

```

> label_solo=c("BMI","肺活量","50 米","跳远","坐位体前屈","仰卧起坐或引体向上","长跑")
> barplot(solo,names.arg=label_solo,xlab="项目类别",ylab="平均分",col="lightblue",main="2015 年
单项成绩平均分")

> data<-read.csv(file="C:/Users/23842/Desktop/2016.csv",header=T,encoding="GBK")

> bmi=mean(data$BMI分数)
> vc=mean(data$肺活量分数)
> run=mean(data$五十米分数)
> jump=mean(data$立定跳远分数)
> sit=mean(data$坐位体前屈分数)
> up=mean(data$仰卧引体分数)
> runn=mean(data$长跑分数)
> solo=c(bmi,vc,run,jump,sit,up,runn)

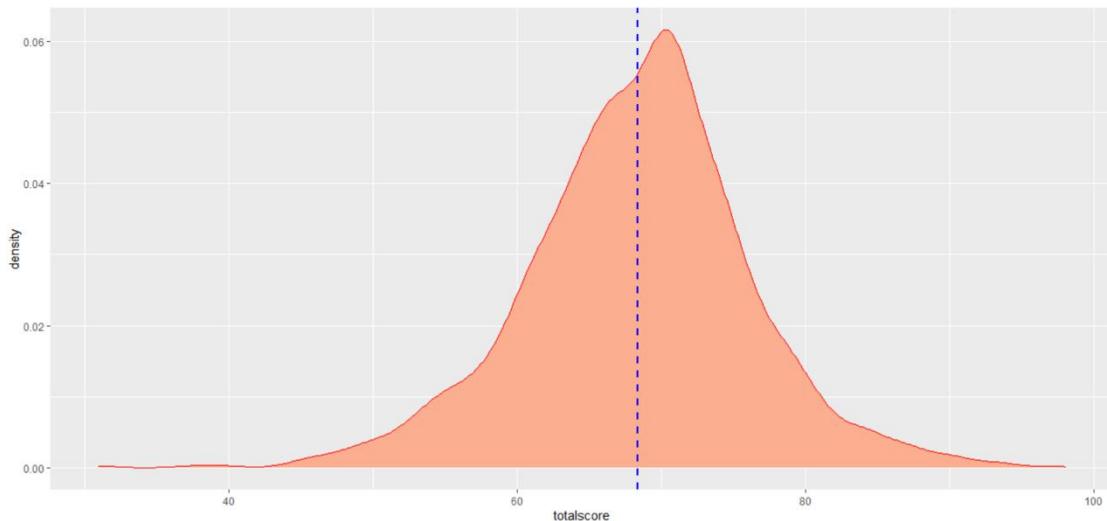
> label_solo=c("BMI","肺活量","50 米","跳远","坐位体前屈","仰卧起坐或引体向上","长跑")
> barplot(solo,names.arg=label_solo,xlab="项目类别",ylab="平均分",col="lightblue",main="2016 年
单项成绩平均分")

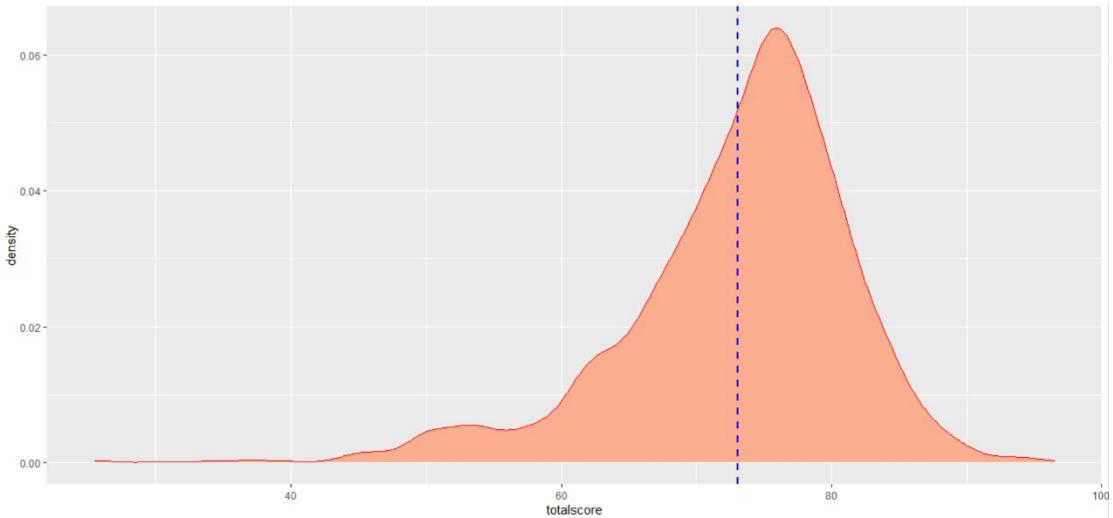
```

3.3 2017 男女数据对比

3.3.1 总成绩分布密度图

(1) 图形





(2) 描述

2017 年参加体测的全体男性同学的总成绩主要集中在 45-95 之间，平均值约为 68 分，且在 70 分左右的同学数量最多；2017 年参加体测的全体女性同学的总成绩主要集中在 45-95 之间，平均值约为 73 分，且在 77 分左右的同学数量最多。

(3) R 语言代码

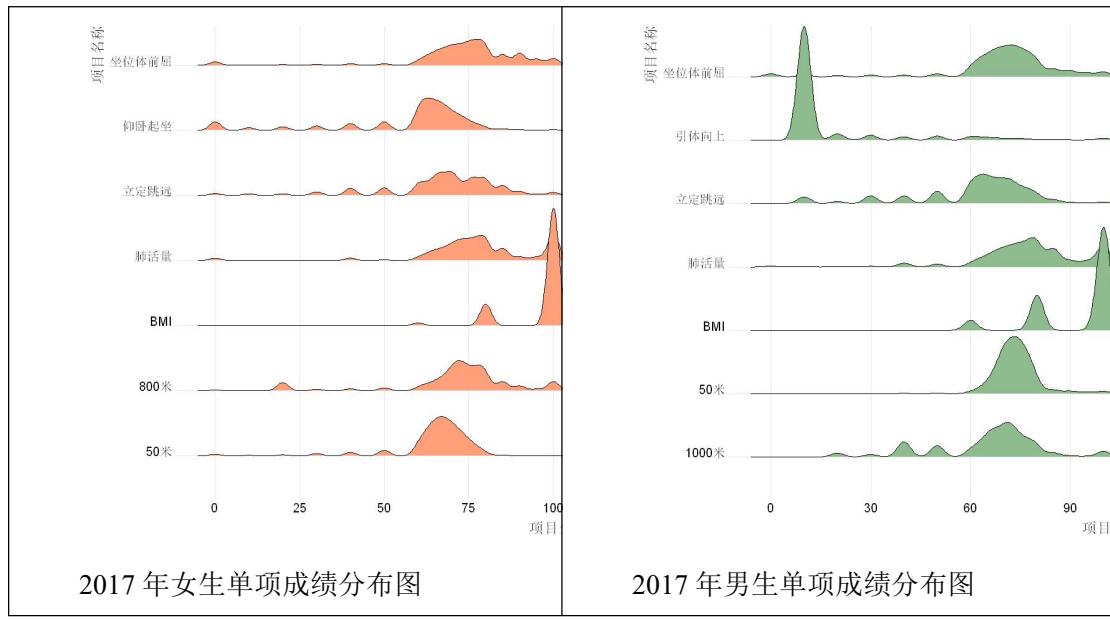
```
library(ggplot2)

data<-read.csv("C:\\Rfiles\\2017 男生总成绩.csv", header=T)
p <- ggplot(data, aes(x=totalscore))+geom_density(color="red", fill="#FCAE91")
p+ geom_vline(aes(xintercept=mean(totalscore)),
              color="blue", linetype="dashed", size=1)

data<-read.csv("C:\\Rfiles\\2017 女生总成绩.csv", header=T)
p <- ggplot(data, aes(x=totalscore))+geom_density(color="red", fill="#FCAE91")
p+ geom_vline(aes(xintercept=mean(totalscore)),
              color="blue", linetype="dashed", size=1)
```

3.3.2 单项成绩分布山脊图

(1) 图形



(2) 描述

在 BMI、50 米跑、肺活量这三项上，男生成绩分布情况优于女生。而在坐位体前屈这一项中，男生虽然在中低分段人数多于女生，但在高分段还是女生更多，说明女生柔韧性要优于男生。在立定跳远这一项上也是如此，也能表明女生跳跃水平高于男生。但在引体或仰卧起坐这一方面，可以看出男生本方面得分大不如女生。0 分段以及低分段男生极多，女生得分情况亦较低。所以男女生普遍缺乏上肢以及腰腹力量锻炼，需要加强。

(3) R 语言代码

```

data17fm <- read.csv(file="score17fm.csv")
ggplot(data17fm, aes(x = 项目分数 ,y = 项目名称, fill = "cut")) +
  geom_density_ridges() +
  theme_ridges() +
  theme(legend.position = "none")+
  scale_fill_manual(values=c("lightsalmon"))
data17m <- read.csv(file="score17m.csv")
ggplot(data17m, aes(x = 项目分数 ,y = 项目名称, fill = "cut")) +
  geom_density_ridges() +
  theme_ridges() +
  theme(legend.position = "none")+
  scale_fill_manual(values=c("darkseagreen"))
  
```

3.3.3 单项成绩平均分雷达图

(1) 图形



(2) 描述

除了男女生分别特有的项目外，2017年参加体测男女生平均分数相差不大。其中，男生在50米项目平均分数高于女生，女生在BIM、肺活量、立定跳远、坐位体前屈项目平均分数高于男生。

(3) R 语言代码

```
install.packages("fmsb")
library(fmsb)

# 构建测试数据集
data <- data.frame(row.names = c('男','女'),
                    "BIM" = c(92.56,96.17),
                    "肺活量" = c(77.60,77.82),
                    "五十米"= c(73.49,64.77),
                    "立定跳远"= c(61.20,66.78),
                    "坐位体前屈" = c(69.71,73.93),
                    "仰卧起坐"= c(0,57.51),
                    "引体向上" = c(19.48,0),
                    "八百米"= c(0,70.95),
                    "一千米"=c(65.56,0))

#定义每个变量的范围
max_min <- data.frame(row.names = c("Max", "Min"),
```

```

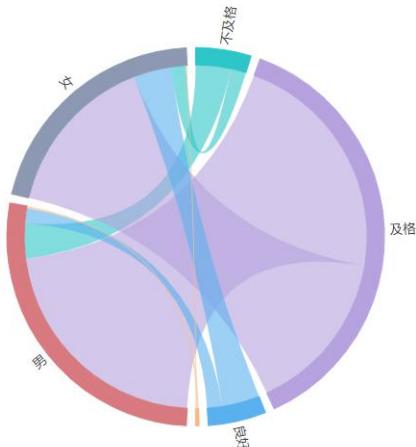
    "BIM" = c(100,0),
    "肺活量" = c(100,0),
    "五十米"= c(100,0),
    "立定跳远"= c(100,0),
    "坐位体前屈"= c(100,0),
    "仰卧起坐"= c(100,0),
    "引体向上" = c(100,0),
    "八百米"= c(100,0),
    "一千米"= c(100,0))

# 合并数据
data_pro <- rbind(max_min,data)
color <- c("#0E36FF", "#FF022C")
plot02 <- radarchart(data_pro,title = c("2017 年男女单项成绩平均分雷达图"),
                      caxislabels = c(0, 20, 40, 60, 80,100),
                      pcol = color,
                      #pfcoll = scales::alpha(color, 0.5),
                      plwd = 2, plty = 1,
                      cglcol = "grey", cglty = 1, cglwd = 0.8,
                      axislabcol = "grey",
                      vlabels = colnames(data),vlcex = 1,
)
# 添加图例
legend(
  x=1.3,y=1.2, legend = c('男','女'),
  bty = "n", pch = 20 , col = color,
  text.col = "black", cex = 1, pt.cex = 3.
)

```

3.3.4 等级分布弦图

(1) 图形



(2) 描述

由图可知，男女生中大部分都位于及格等级，极少数位于优秀等级。但在男生人数多于女生的情况下，男生良好等级人数少于女生，不及格人数远远大于女生，说明 2017 年度女生体育成绩较好。

(3) R 语言代码

```

data<-read.csv(file="C:/Users/23842/Desktop/l.csv",header=T,encoding="GBK")
library(circlize)
test<-d0[d0$year0==1960,-1]
chordDiagram(x = test,
              directional = 1,
              order = d1$region,
              grid.col = d1$col1,
              annotationTrack = "grid",
              direction.type = c("diffHeight","arrows"),
              )
circos.track(track.index = 1, bg.border = NA,
            panel.fun = function(x, y) {
              xlim = get.cell.meta.data("xlim")
              sector.index = get.cell.meta.data("sector.index")
              reg1 = d1 %>% filter(region == sector.index) %>% pull(reg1)
              reg2 = d1 %>% filter(region == sector.index) %>% pull(reg2)
              circos.text(x = mean(xlim), y = ifelse(is.na(reg2), 3, 4),labels = reg1, facing =
              "bending", cex =0.8)
            }
          )
  
```

```

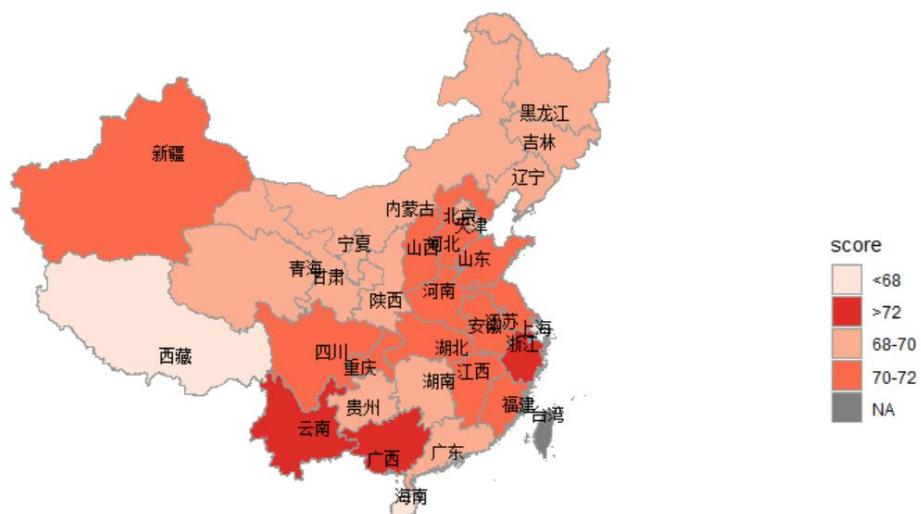
circos.text(x = mean(xlim), y = 2.75, labels = reg2, facing = "bending", cex = 0.8)
circos.axis(h="top", labels.cex = 0.6,labels.niceFacing = FALSE, labels.pos.adjust
= FALSE)
})

```

3.4 2017 生源地数据对比

3.4.1 总成绩平均分地图

(1) 图形



(2) 描述

2017年总平均成绩为70.4分。其中，平均成绩小于68分的地区有3个，分别是西藏自治区、上海市和海南省，属于低于平均水平的地区；平均成绩介于68和70的地区有13个，分别是黑龙江省、内蒙古自治区、吉林省、辽宁省、甘肃省、北京市、天津市、陕西省、宁夏回族自治区、青海省、湖南省、贵州省和广东省，属于略低于平均水平的地区；平均成绩介于70和72之间的地区有12个，分别是新疆维吾尔自治区、河北省、山西省、山东省、河南省、安徽省、江苏省、四川省、湖北省、重庆市、江西省和福建省，属于基本平均水平的地区；平均成绩大于72分的地区有3个，分别是浙江省、云南省和广西壮族自治区，属于高于平均水平的地区。

(3) R 语言代码

```
library(mapdata)
```

```

library(maptools)
library(ggplot2)
library(plyr)#引用包

china_map=readShapePoly("C:/Rfiles/map/bou2_4p.shp")#读取地图数据

ggplot(china_map,aes(x=long,y=lat,group=group))      +geom_polygon(fill="white",colour="grey")
+coord_map("polyconic")#绘制并投影得到可用地图

x <- china_map@data#读取行政信息
xs <- data.frame(x,id=seq(0:924)-1)#含岛屿共 925 个形状
china_map1 <- fortify(china_map) #转化为数据框
china_map_data <- join(china_map1, xs, type = "full") #合并两个数据框

mydata<-read.csv("C:\\Rfiles\\2017 附总成绩.csv")#读取业务数据
china_data <- join(china_map_data, mydata, type="full")#合并两个数据框

province_city <- read.csv("C:/Rfiles/china-cities.csv") #读取省市经纬度

memory.limit(10000000)# 设置约为 1G 内存，否则会显示无法分配

ggplot(china_data,aes(long,lat))+

  geom_polygon(aes(group=group,fill=score),colour="grey60")+
  scale_fill_manual(values=c("#FEE5D9","#DE2D26","#FCAE91","#FB6A4A"))+
  coord_map("polyconic") +
  geom_text(aes(x = lon,y = lat,label = province), data =province_city, size=3)+

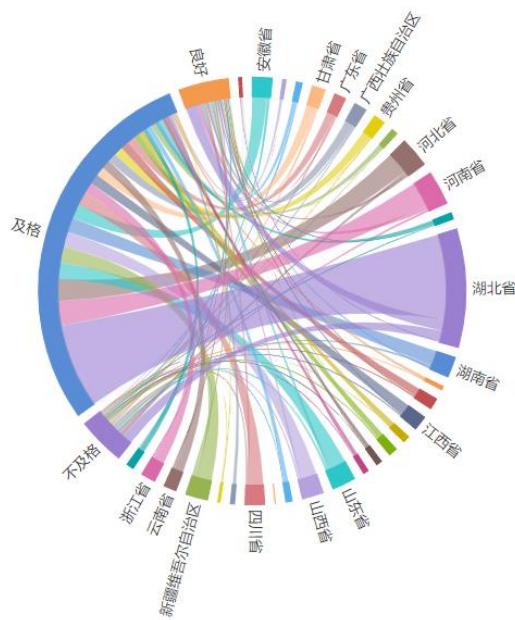
  ggtitle("2017 年生源地平均成绩图")+
  theme(
    panel.grid = element_blank(),
    panel.background = element_blank(),
    axis.text = element_blank(),
    axis.ticks = element_blank(),

```

```
axis.title = element_blank()  
)#确定颜色并作图
```

3.4.2 等级分布弦图

(1) 图形



(2) 描述

由图可知，各生源地中大部分都位于及格等级，极少数位于优秀等级。各生源地等级分布较均匀。

(3) R 语言代码

```
data<-read.csv(file="C:/Users/23842/Desktop/2.csv",header=T,encoding="GBK")  
library(circlize)  
test<-d0[d0$year==1960,-1]  
chordDiagram(x = test,  
             directional = 1,  
             order = d1$region,  
             grid.col = d1$col1,  
             annotationTrack = "grid",  
             direction.type = c("diffHeight","arrows"),  
             )  
circos.track(track.index = 1, bg.border = NA,  
            panel.fun = function(x, y) {
```

```

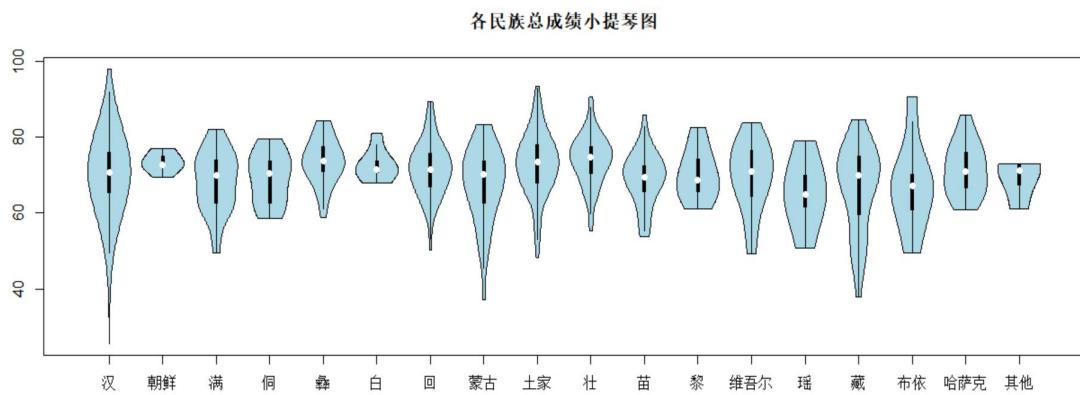
xlim = get.cell.meta.data("xlim")
sector.index = get.cell.meta.data("sector.index")
reg1 = d1 %>% filter(region == sector.index) %>% pull(reg1)
reg2 = d1 %>% filter(region == sector.index) %>% pull(reg2)
circos.text(x = mean(xlim), y = ifelse(is.na(reg2), 3, 4), labels = reg1, facing =
"bending", cex = 0.8)
circos.text(x = mean(xlim), y = 2.75, labels = reg2, facing = "bending", cex = 0.8)
circos.axis(h = "top", labels.cex = 0.6, labels.niceFacing = FALSE, labels.pos.adjust
= FALSE)
})

```

3.5 2017 民族数据对比

3.5.1 总成绩小提琴图

(1) 图形



(2) 描述

各民族总成绩从中位数来看相差不大，都位于 60-80 之间，其中瑶族最低，壮族最高。从核密度来看，各民族相差不多，密度最大部分都位于 70 上下。从上下限来看，汉族上限最高、下限最低，成绩差距大，最不稳定；朝鲜族上限与下限差距最小，成绩最稳定。

(3) R 语言代码

```

> a<-data[data$民族编号==1,"总分数"]
> b<-data[data$民族编号==2,"总分数"]
> c<-data[data$民族编号==3,"总分数"]
> d<-data[data$民族编号==4,"总分数"]
> e<-data[data$民族编号==5,"总分数"]

```

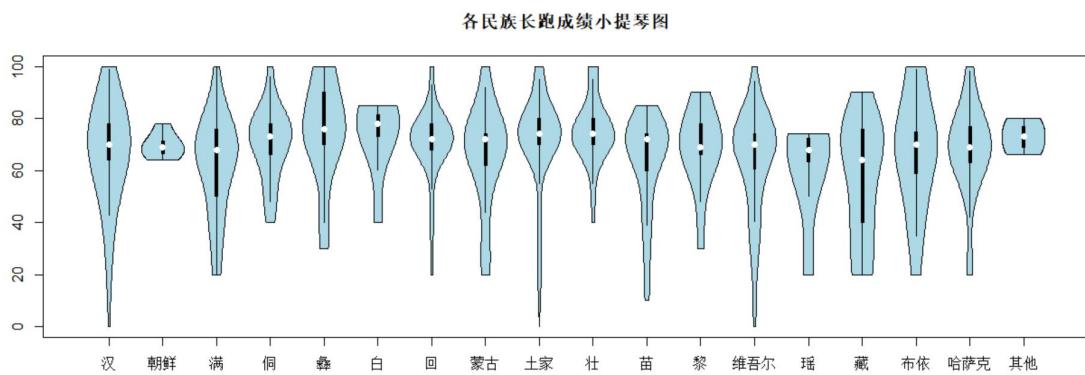
```

> f<-data[data$民族编号==1,"总分数"]
> g<-data[data$民族编号==6,"总分数"]
> h<-data[data$民族编号==7,"总分数"]
> i<-data[data$民族编号==8,"总分数"]
> j<-data[data$民族编号==9,"总分数"]
> k<-data[data$民族编号==10,"总分数"]
> l<-data[data$民族编号==11,"总分数"]
> m<-data[data$民族编号==12,"总分数"]
> n<-data[data$民族编号==13,"总分数"]
> o<-data[data$民族编号==16,"总分数"]
> p<-data[data$民族编号==17,"总分数"]
> q<-data[data$民族编号==18,"总分数"]
> r<-data[data$民族编号==19,"总分数"]
> vioplot(a,b,c,d,e,f,g,h,i,j,k,l,m,n,o,p,q,r,col="lightblue",names=c("汉","朝鲜","满","侗","彝","白",
回","蒙古","土家","壮","苗","黎","维吾尔","瑶","藏","布依","哈萨克","其他"),main="各民族总成绩小
提琴图" )

```

3.5.2 单项成绩小提琴图

(1) 图形



(2) 描述

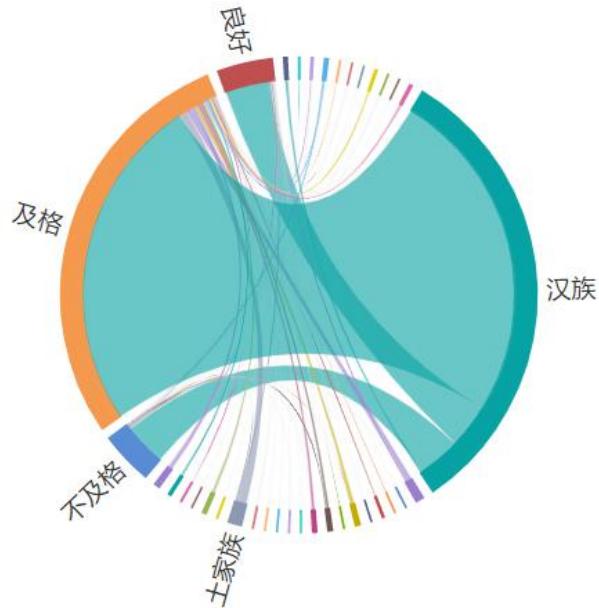
选取差异最大的长跑成绩进行可视化，各民族长跑成绩从中位数来看相差不大，都位于 60-80 之间，其中藏族最低，白族最高。从核密度来看，各民族相差不多，密度最大部分都位于 70 上下。从上下限来看，汉族、土家族、维吾尔族上限最高、下限最低，长跑成绩差距大，最不稳定；朝鲜族上限与下限差距最小，成绩最稳定；瑶族上限低但下限也很低，说明瑶族长跑成绩不好。

(3) R 语言代码

```
> a<-data[data$民族编号==1,"长跑分数"]  
> b<-data[data$民族编号==2,"长跑分数"]  
> c<-data[data$民族编号==3,"长跑分数"]  
> d<-data[data$民族编号==4,"长跑分数"]  
> e<-data[data$民族编号==5,"长跑分数"]  
> f<-data[data$民族编号==6,"长跑分数"]  
> g<-data[data$民族编号==7,"长跑分数"]  
> h<-data[data$民族编号==8,"长跑分数"]  
> i<-data[data$民族编号==9,"长跑分数"]  
> j<-data[data$民族编号==10,"长跑分数"]  
> k<-data[data$民族编号==11,"长跑分数"]  
> l<-data[data$民族编号==12,"长跑分数"]  
> m<-data[data$民族编号==13,"长跑分数"]  
> n<-data[data$民族编号==16,"长跑分数"]  
> o<-data[data$民族编号==17,"长跑分数"]  
> p<-data[data$民族编号==18,"长跑分数"]  
> q<-data[data$民族编号==19,"长跑分数"]  
> r<-data[data$民族编号==20,"长跑分数"]  
> vioplot(a,b,c,d,e,f,g,h,i,j,k,l,m,n,o,p,q,r,col="lightblue",names=c("汉","朝鲜","满","侗","彝","白","回","蒙古","土家","壮","苗","黎","维吾尔","瑶","藏","布依","哈萨克","其他"),main="各民族长跑成绩  
小提琴图")
```

3.5.3 等级分布弦图

(1) 图形



(2) 描述

由图可知，各民族中大部分都位于及格等级，极少数位于优秀等级，各民族等级分布较均匀。以主体汉族来看，等级为良好和不及格的学生人数大致相同。

(3) R 语言代码

```

data<-read.csv(file="C:/Users/23842/Desktop/3.csv",header=T,encoding="GBK")
library(circlize)
test<-d0[d0$year0==1960,-1]
chordDiagram(x = test,
             directional = 1,
             order = d1$region,
             grid.col = d1$col1,
             annotationTrack = "grid",
             direction.type = c("diffHeight","arrows"),
             )
circos.track(track.index = 1, bg.border = NA,
            panel.fun = function(x, y) {
              xlim = get.cell.meta.data("xlim")
              sector.index = get.cell.meta.data("sector.index")
              reg1 = d1 %>% filter(region == sector.index) %>% pull(reg1)
              reg2 = d1 %>% filter(region == sector.index) %>% pull(reg2)
            }
)

```

```

circos.text(x = mean(xlim), y = ifelse(is.na(reg2), 3, 4), labels = reg1, facing =
"bending", cex = 0.8)

circos.text(x = mean(xlim), y = 2.75, labels = reg2, facing = "bending", cex = 0.8)

circos.axis(h = "top", labels.cex = 0.6, labels.niceFacing = FALSE, labels.pos.adjust
= FALSE)

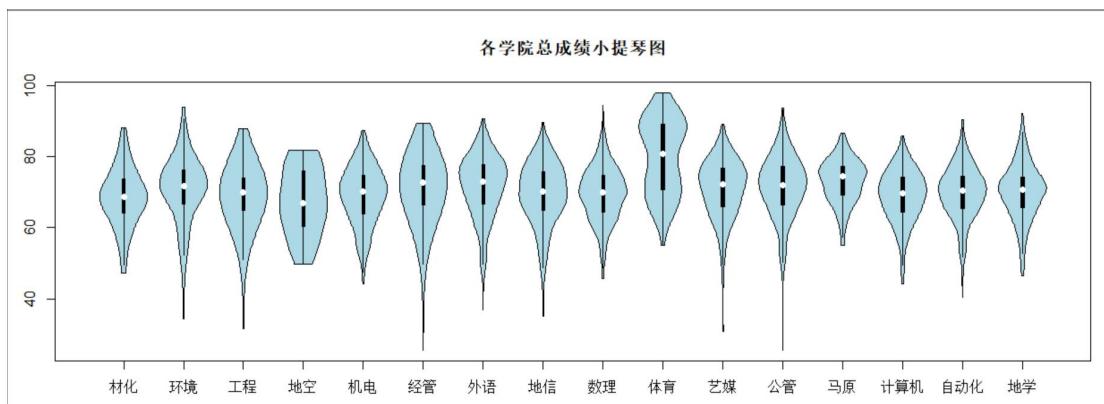
})

```

3.6 2017 学院数据对比

3.6.1 总成绩小提琴图

(1) 图形



(2) 描述

各学院除体育学院总成绩从中位数来看相差不大，都位于 60-80 之间，其中地空最低，马原最高，而体育学院遥遥领先，突破 80 分，说明体育学院学生体质远远强于其他学院学生。从核密度来看，各学院相差不多，密度大部分都位于 70 上下，而体育学院成绩大多在 90 分左右。从上下限来看，公管、经管上限最高、下限最低，总成绩差距大，最不稳定；马原、地空上限与下限差距最小，成绩最稳定；体育学院上限高但下限也很高，说明体育学院每个人的体育成绩都偏好。

(3) R 语言代码

```

> data<-read.csv(file="C:/Users/23842/Desktop/2017.csv",header=T,encoding="GBK")

> a<-data[data$学院==3,"总分数"]

> b<-data[data$学院==4,"总分数"]

> c<-data[data$学院==5,"总分数"]

> d<-data[data$学院==6,"总分数"]

> e<-data[data$学院==7,"总分数"]

> f<-data[data$学院==8,"总分数"]

> g<-data[data$学院==9,"总分数"]

> h<-data[data$学院==11,"总分数"]

```

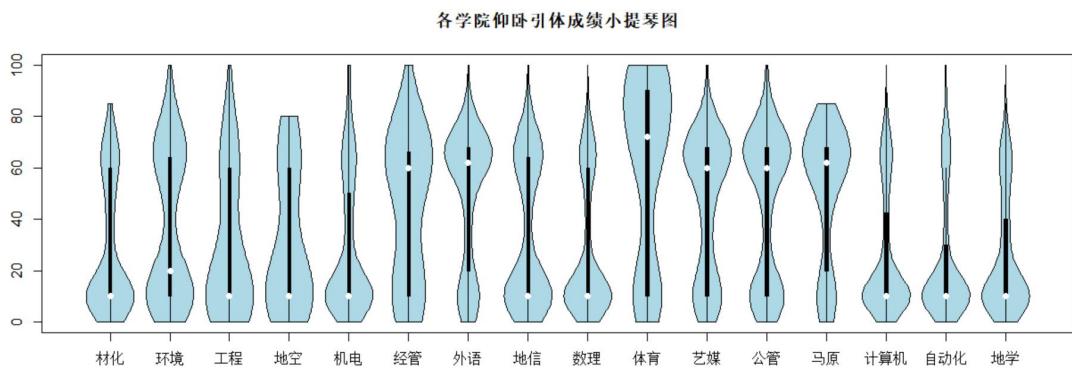
```

> i<-data[data$学院==12,"总分数"]
> j<-data[data$学院==13,"总分数"]
> k<-data[data$学院==16,"总分数"]
> l<-data[data$学院==17,"总分数"]
> m<-data[data$学院==18,"总分数"]
> n<-data[data$学院==19,"总分数"]
> o<-data[data$学院==23,"总分数"]
> p<-data[data$学院==91,"总分数"]
> vioplot(a,b,c,d,e,f,g,h,i,j,k,l,m,n,o,p)
> vioplot(a,b,c,d,e,f,g,h,i,j,k,l,m,n,o,p,col="lightblue",names=c("材化","环境","工程","地空","机电",
  "经管","外语","地信","数理","体育","艺媒","公管","马原","计算机","自动化","地学"),main="各学院
  总成绩小提琴图")

```

3.6.2 单项成绩小提琴图

(1) 图形



(2) 描述

选取差异最大的仰卧起做或引体向上成绩可视化，从中位数来看相差很大，经管、外语、体育、艺媒、公管、马原中位分数在 60-80 之间，而材化、环境、地空、工程、机电、数理等专业中位分數极低，位于 0-20 之间。从核密度来看，各学院相差也很大，经管、外语、艺媒、公管、马原分數集中于 70 左右，材化、环境、地空、工程、机电、数理集中在 10 左右。从上下限来看，各学院相差不大，上限都很高，下限都很低，成绩不稳定。由此可以看出，除体育最好的体育专业学生外，文科类学生体质强于理科类学生。

(3) R 语言代码

```

> a<-data[data$学院==3,"仰卧引体分数"]
> b<-data[data$学院==4,"仰卧引体分数"]

```

```

> c<-data[data$学院==5,"仰卧引体分数"]
> d<-data[data$学院==6,"仰卧引体分数"]
> e<-data[data$学院==7,"仰卧引体分数"]
> f<-data[data$学院==8,"仰卧引体分数"]
> g<-data[data$学院==9,"仰卧引体分数"]
> h<-data[data$学院==11,"仰卧引体分数"]
> i<-data[data$学院==12,"仰卧引体分数"]
> j<-data[data$学院==13,"仰卧引体分数"]
> k<-data[data$学院==16,"仰卧引体分数"]
> l<-data[data$学院==17,"仰卧引体分数"]
> m<-data[data$学院==18,"仰卧引体分数"]
> n<-data[data$学院==19,"仰卧引体分数"]
> o<-data[data$学院==23,"仰卧引体分数"]
> p<-data[data$学院==91,"仰卧引体分数"]
> vioplot(a,b,c,d,e,f,g,h,i,j,k,l,m,n,o,p,col="lightblue",names=c("材化","环境","工程","地空","机电",
,"经管","外语","地信","数理","体育","艺媒","公管","马原","计算机","自动化","地学"),main="各学院
仰卧引体成绩小提琴图")

```

3.6.3 等级比率条形图

(1) 图形



(2) 描述

由图可知，各学院及格率最高，优秀率最低，其中体育学院优秀率最高，不及格率最低，说明体育学院体质远远强于其他学院。且地信学院、数理学院以及海洋学院不及格率偏高，说明这三个

学院的学生需加强体育锻炼。

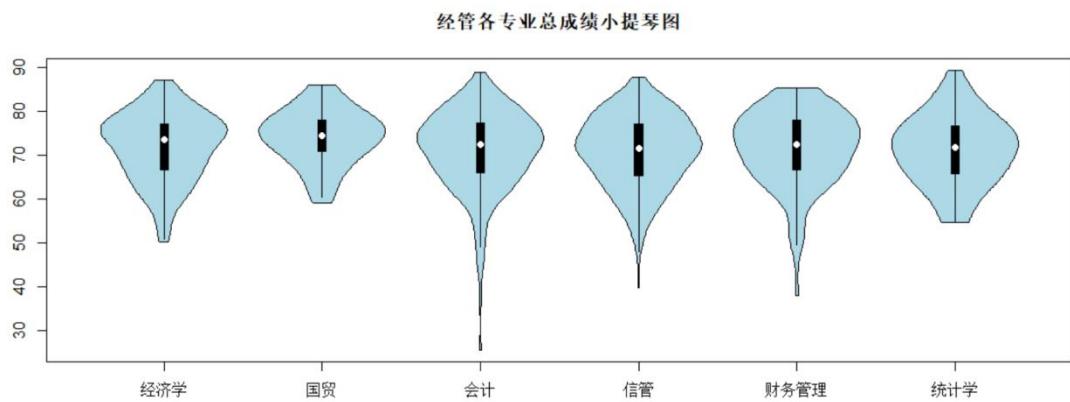
(3) R 语言代码

```
> data<-read.csv(file="C:/Users/23842/Desktop/2017.csv",header=T,encoding="GBK")
> counts<-table(data$等级,data$学院)
> lable=c("材化","环境","工程","地空","机电","经管","外语","地信","数理","体育","艺媒","公管",
,"马原","计算机","自动化","海洋")
> barplot(counts,main="2017 学院 等级 对比 ",xlab="学院 ",ylab="频数",
",names.arg=lable,col=c("lightgreen","lightblue","lightyellow","red"),legend=rownames(counts))
```

3.7 经管学院各专业数据对比

3.7.1 各专业总成绩小提琴图

(1) 图形:



(2) 描述: 各专业从中位数来看相差不大, 都位于 70-80 之间, 其中国贸专业最高。从核密度来看, 各专业相差不多, 密度最大部分都位于 75 上下。从上下限来看, 会计专业上限最高、下限最低, 总成绩差距大, 最不稳定; 国贸上限与下限差距最小, 成绩最稳定。

(3) 代码:

```
library(zoo)
> library(sm)
> library(vioplot)
> data<-read.csv(file="C:/Users/23842/Desktop/2017.csv",header=T,encoding="GBK")
> x1<-data$经济学
> x2<-data$国际经济与贸易
> x3<-data$会计
> x4<-data$信管
> x5<-data$财务管理
```

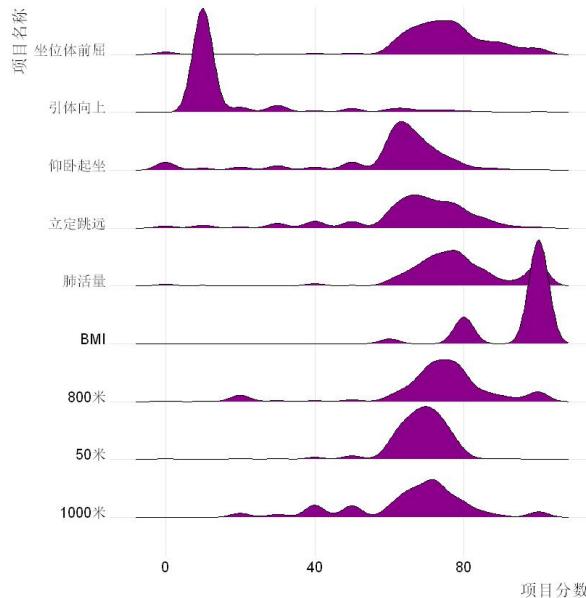
```

> x6<-data$统计学
> vioplot(x1,x2,x3,x4,x5,x6,col="lightblue",names=c("经济学","国贸","会计","信管","财务管理",
统计学"),main="经管各专业总成绩小提琴图")

```

3.7.2 2017 年经济管理学院单项成绩分布山脊图

(1) 图形



2017 年经济管理学院单项成绩分布图

(2) 描述

2017 年经济管理学院单项成绩分布和全校单项成绩分布情况基本一致，坐位体前屈、仰卧起坐、立定跳远、肺活量、800m、50m 成绩分布大多集中在中上成绩段，引体向上多集中于低分段，应对该项目进行重点加强。

(3) r 语言代码

```

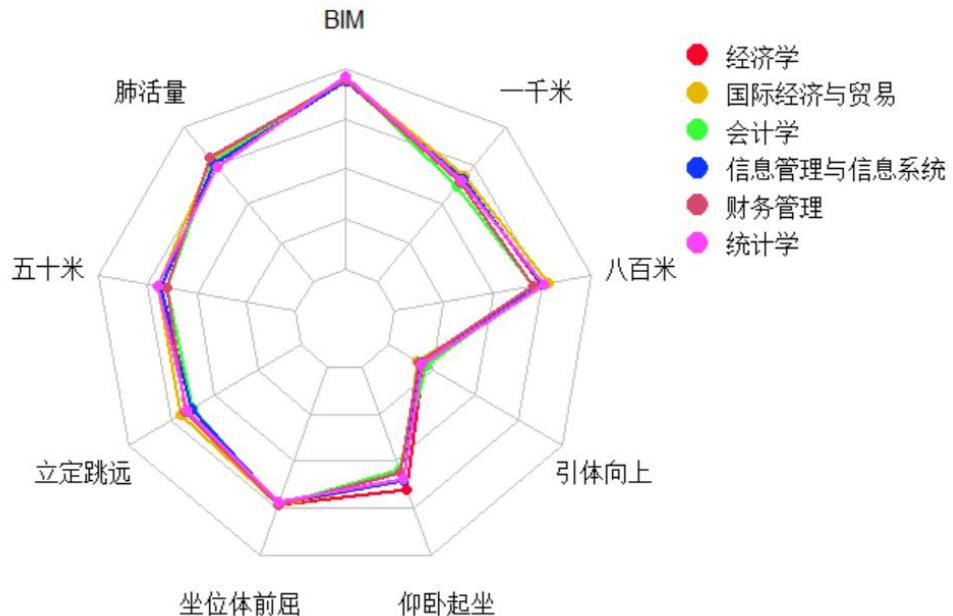
jg <- read.csv(file="jg.csv")
ggplot(jg, aes(x = 项目分数 ,y = 项目名称, fill = "cut")) +
  geom_density_ridges() +
  theme_ridges() +
  theme(legend.position = "none")+
  scale_fill_manual(values=c("darkmagenta"))

```

3.7.3 2017 年经管各专业单项成绩平均分雷达图

(1) 图形

2017年经管各专业单项成绩平均分雷达图



(2) 描述

2017年参加体测的经管各专业学生在各个项目的平均分均相差不大。其中，经济学专业在仰卧起坐项目平均分数最高，国际经济与贸易专业在立定跳远、800米和1000米项目中平均分数最高，会计学专业在引体向上项目平均分数最高，财务管理专业在肺活量和坐位体前屈项目平均分数最高，统计学在BIM和50米项目中平均分数最高。

(3) r 语言代码

```
library(fmsb)

# 构建测试数据集
data <- data.frame(row.names = c('经济学','国际经济与贸易','会计学','信息管理与信息系统','财务管理','统计学'),
"BIM" = c(94.55,95.00,95.10,94.01,94.31,95.67),
"肺活量" = c(76.04,79.23,77.79,76.52,79.91,74.68),
"一千米"= c(68.36,69.83,65.38,68.59,65.11,69.67),
"八百米"= c(66.04,70.08,62.89,63.65,66.67,65.78),
"引体向上" = c(72.76,72.83,72.85,72.75,73.29,71.27),
"仰卧起坐"= c(65.11,60.51,54.71,60.42,56.52,59.31),
"坐位体前屈" = c(19.77,16.58,21.47,18.00,17.56,18.90),
"五十米"= c(66.04,70.08,62.89,63.65,66.67,65.78),
"立定跳远"= c(66.04,70.08,62.89,63.65,66.67,65.78))
```

```

    "八百米"= c(74.56,78.44,70.26,74.93,70.69,75.59),
    "一千米"=c(66.67,67.95,61.76,66.14,64.07,65.48))

#定义每个变量的范围
max_min <- data.frame(row.names = c("Max", "Min"),
                        "BIM" = c(100,0),
                        "肺活量" = c(100,0),
                        "五十米"= c(100,0),
                        "立定跳远"= c(100,0),
                        "坐位体前屈" = c(100,0),
                        "仰卧起坐"= c(100,0),
                        "引体向上" = c(100,0),
                        "八百米"= c(100,0),
                        "一千米"= c(100,0))

# 合并数据
data_pro <- rbind(max_min,data)
color<- c("#FF022C", "#E7B800","#3AFF38","#0E36FF","#D9486D","#FF44FF")
plot02 <- radarchart(data_pro,title = c("2017 年经管各专业单项成绩平均分雷达图"),
                      caxislabels = c(0, 20, 40, 60, 80,100),
                      pcol = color,
                      #pfcol = scales::alpha(color, 0.5),
                      plwd = 2, plty = 1,
                      cglcol = "grey", cglty = 1, cglwd = 0.8,
                      axislabcol = "grey",
                      vlabels = colnames(data),vlcex = 1,
)
# 添加图例
legend(
  x=1.3,y=1.2, legend = c('经济学','国际经济与贸易','会计学', '信息管理与信息系统', '财务管理',
  '统计学'),
```

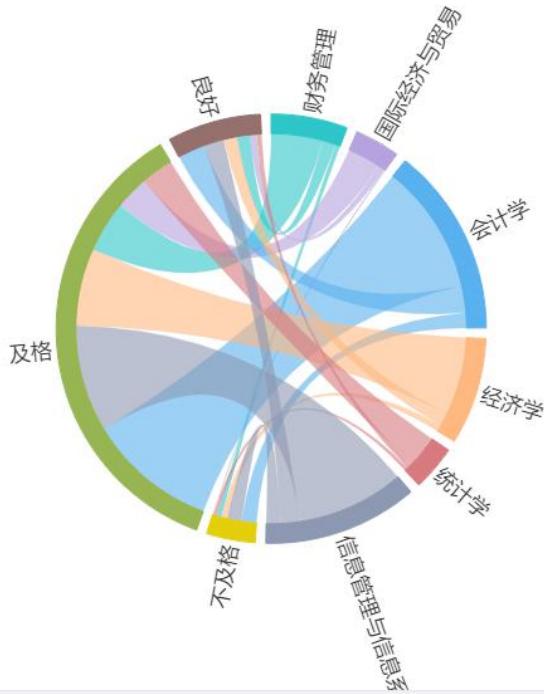
```

bty = "n", pch = 20 , col = color,
text.col = "black", cex = 1, pt.cex = 3.
)

```

3.7.4 经管学院各专业等级分布弦图

(1) 图形



(2) 描述由图可知，各专业中大部分都位于及格等级，少数位于不及格等级，没有学生位于优秀等级。各专业等级按人数比例分布较均匀。

代码：

```

data<-read.csv(file="C:/Users/23842/Desktop/4.csv",header=T,encoding="GBK")
library(circlize)
test<-d0[d0$year0==1960,-1]
chordDiagram(x = test,
              directional = 1,
              order = d1$region,
              grid.col = d1$col1,
              annotationTrack = "grid",
              direction.type = c("diffHeight","arrows"),
)
circos.track(track.index = 1, bg.border = NA,
            panel.fun = function(x, y) {

```

```

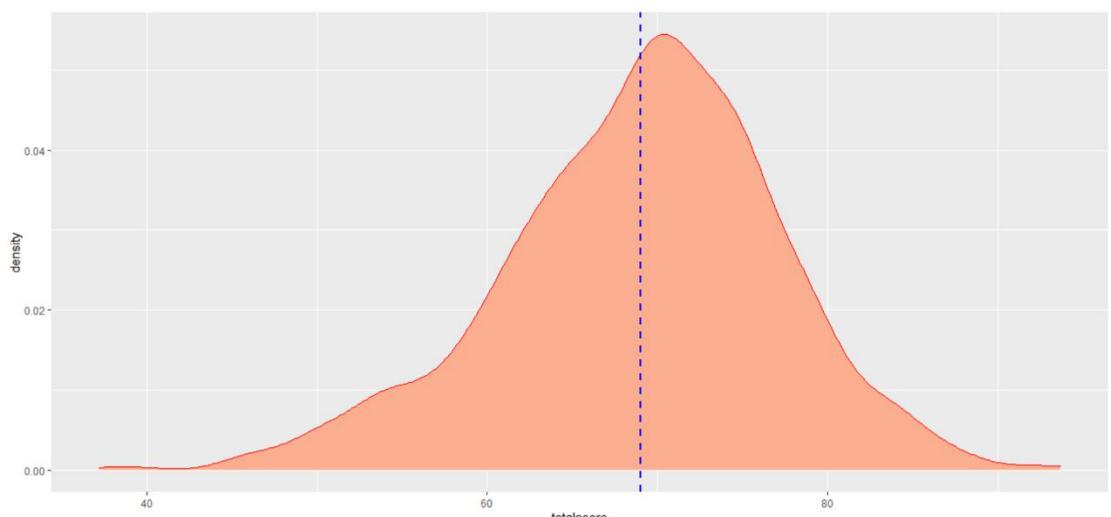
        xlim = get.cell.meta.data("xlim")
        sector.index = get.cell.meta.data("sector.index")
        reg1 = d1 %>% filter(region == sector.index) %>% pull(reg1)
        reg2 = d1 %>% filter(region == sector.index) %>% pull(reg2)
        circos.text(x = mean(xlim), y = ifelse(is.na(reg2), 3, 4), labels = reg1, facing =
        "bending", cex = 0.8)
        circos.text(x = mean(xlim), y = 2.75, labels = reg2, facing = "bending", cex = 0.8)
        circos.axis(h = "top", labels.cex = 0.6, labels.niceFacing = FALSE, labels.pos.adjust
        = FALSE)
    })
}

```

3.8 2017 年各年级数据对比

3.8.1 总成绩分布密度图

一、2017 年 2017 级总成绩分布密度图



2017 年参加体测的全体 2017 级同学的总成绩主要集中在 45-95 之间，平均值约为 69 分，且在 70 分左右的同学数量最多。

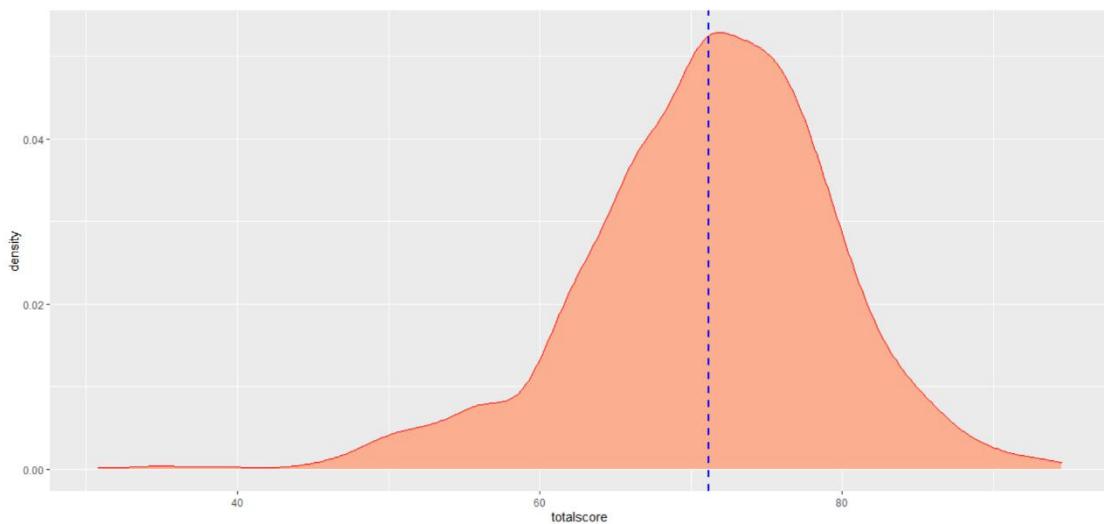
r 语言代码：

```

library(ggplot2)
data<-read.csv("C:\\Rfiles\\2017 年 2017 级总成绩.csv", header=T)
p <- ggplot(data, aes(x=totalscore))+geom_density(color="red", fill="#FCAE91")
p+ geom_vline(aes(xintercept=mean(totalscore)),
              color="blue", linetype="dashed", size=1)

```

二、2017 年 2016 级总成绩分布密度图

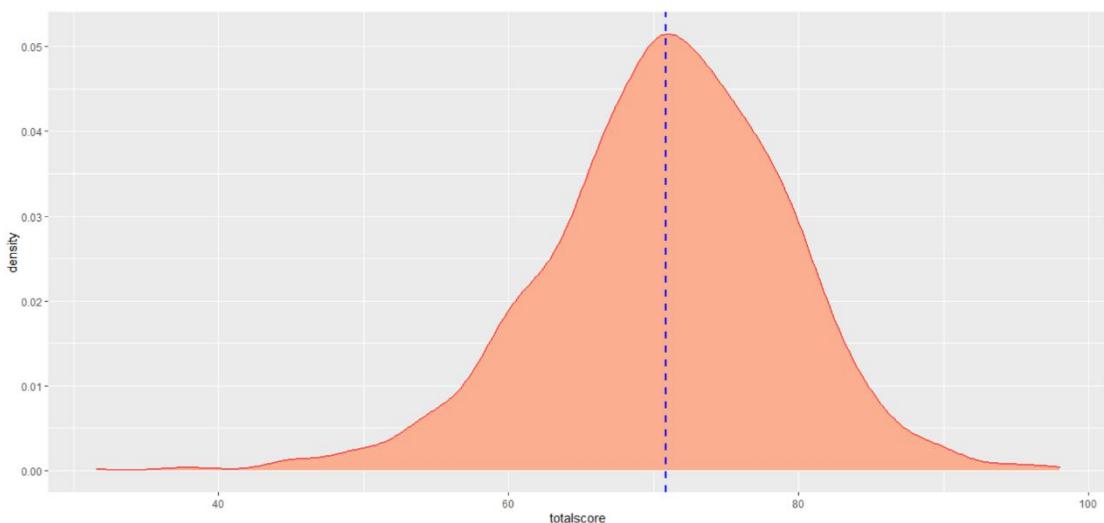


2017 年参加体测的全体 2016 级同学的总成绩主要集中在 45-95 之间，平均值约为 71 分，且在 71 分左右的同学数量最多。

r 语言代码：

```
library(ggplot2)  
data<-read.csv("C:\\Rfiles\\2017 年 2016 级总成绩.csv", header=T)  
p <- ggplot(data, aes(x=totalscore))+geom_density(color="red", fill="#FCAE91")  
p+ geom_vline(aes(xintercept=mean(totalscore)),  
              color="blue", linetype="dashed", size=1)
```

三、2017 年 2015 级总成绩分布密度图



2017 年参加体测的全体 2015 级同学的总成绩主要集中在 45-98 之间，平均值约为 71 分，且在 71 分左右的同学数量最多。

r 语言代码：

```

library(ggplot2)

data<-read.csv("C:\\Rfiles\\2017 年 2015 级总成绩.csv", header=T)

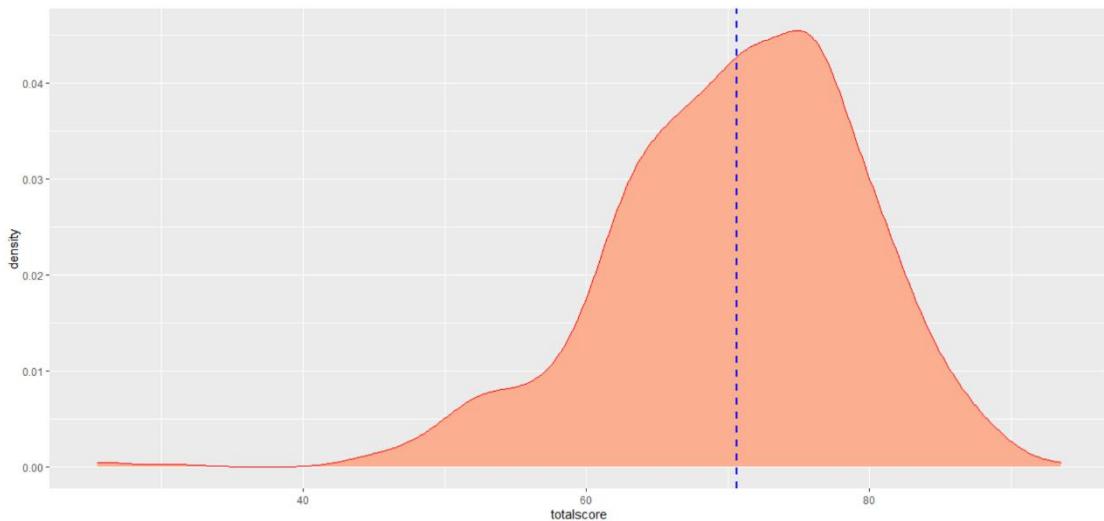
p <- ggplot(data, aes(x=totalscore))+geom_density(color="red", fill="#FCAE91")

p+ geom_vline(aes(xintercept=mean(totalscore)),  

              color="blue", linetype="dashed", size=1)

```

四、2017 年 2014 级总成绩分布密度图



2017 年参加体测的全体 2014 级同学的总成绩主要集中在 40-95 之间，平均值约为 70 分，且在 75 分左右的同学数量最多。

r 语言代码：

```

library(ggplot2)

data<-read.csv("C:\\Rfiles\\2017 年 2014 级总成绩.csv", header=T)

p <- ggplot(data, aes(x=totalscore))+geom_density(color="red", fill="#FCAE91")

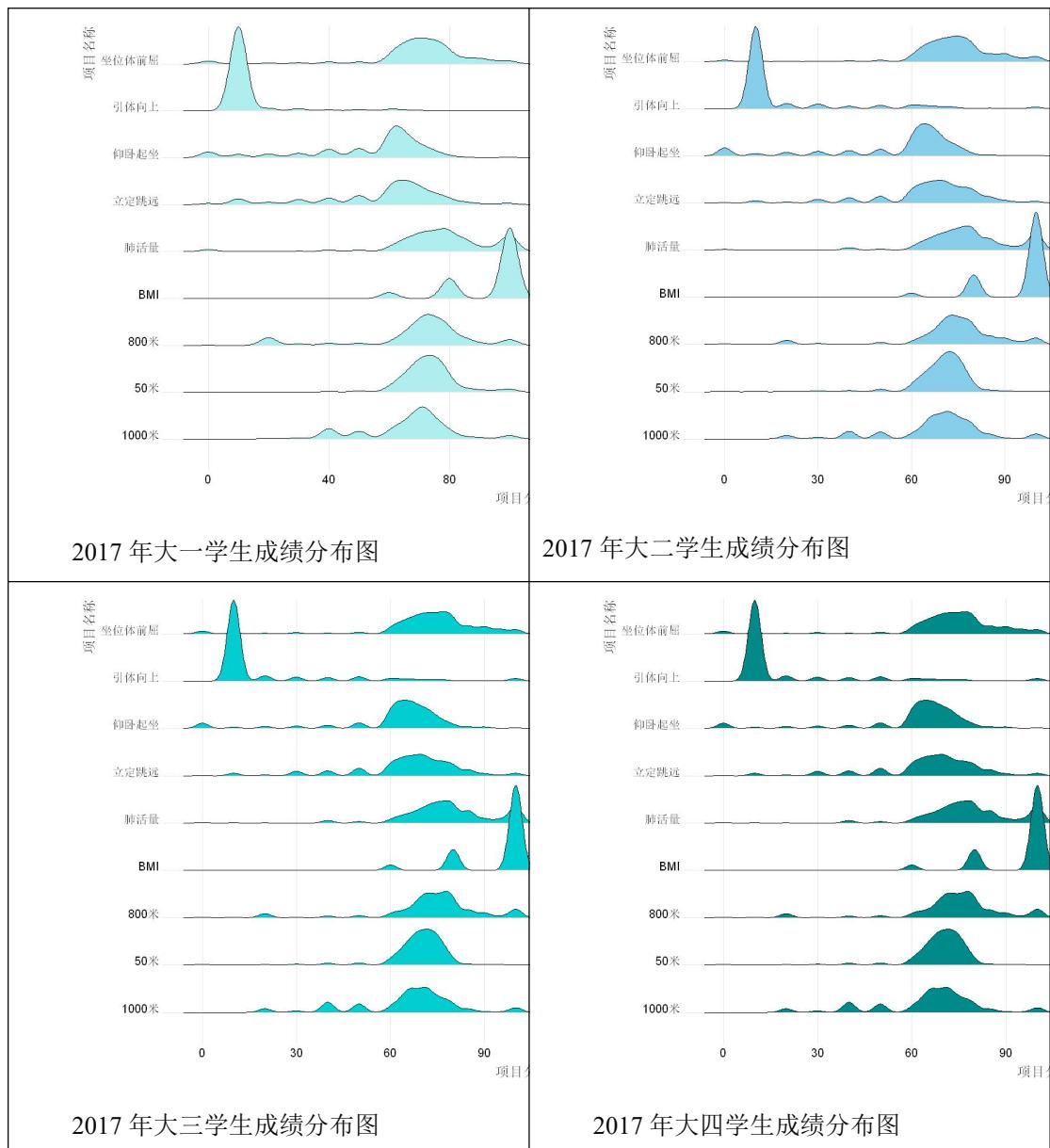
p+ geom_vline(aes(xintercept=mean(totalscore)),  

              color="blue", linetype="dashed", size=1)

```

3.8.2 2017 年各年级单项成绩分布山脊图

(1) 图形



(2) 描述

2017年不同年级学生单项成绩分布大体一致，大三、大四年级学生坐位体前屈、立定跳远、800m、引体向上略优于低年级学生，1000m成绩分布略差于低年级学生。

(3) 7.2.3 r 语言代码

```

nj1 <- read.csv(file="1711.csv")
ggplot(nj1, aes(x = 项目分数 ,y = 项目名称, fill = "cut")) +
  geom_density_ridges() +
  theme_ridges() +
  theme(legend.position = "none")+
  scale_fill_manual(values=c("paleturquoise"))
nj2 <- read.csv(file="1722.csv")

```

```

ggplot(nj2, aes(x = 项目分数 ,y= 项目名称, fill = "cut")) +
  geom_density_ridges() +
  theme_ridges() +
  theme(legend.position = "none")+
  scale_fill_manual(values=c("skyblue"))

nj3 <- read.csv(file="1733.csv")

ggplot(nj3, aes(x = 项目分数 ,y= 项目名称, fill = "cut")) +
  geom_density_ridges() +
  theme_ridges() +
  theme(legend.position = "none")+
  scale_fill_manual(values=c("darkturquoise"))

nj4 <- read.csv(file="1744.csv")

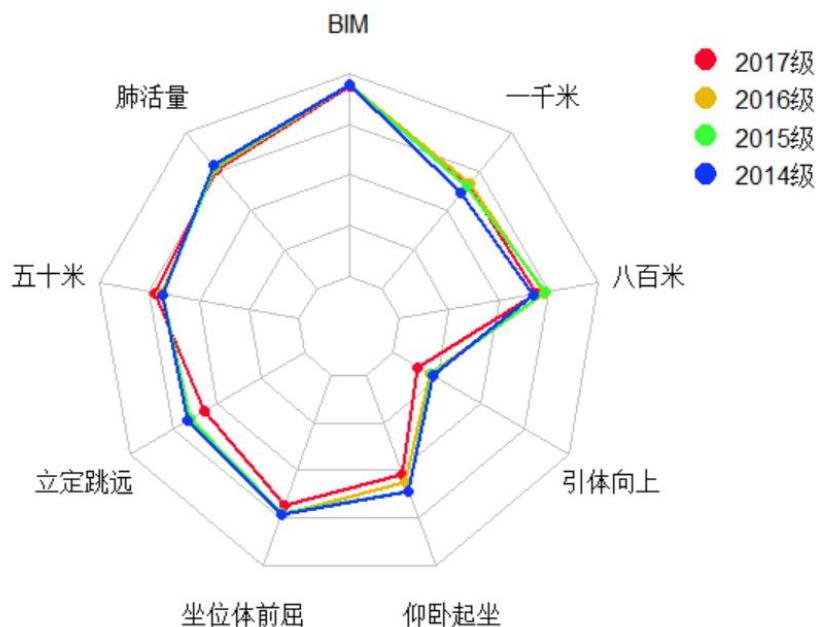
ggplot(nj3, aes(x = 项目分数 ,y= 项目名称, fill = "cut")) +
  geom_density_ridges() +
  theme_ridges() +
  theme(legend.position = "none")+
  scale_fill_manual(values=c("darkcyan"))

```

3.8.3 2014-2017 单项成绩平均分雷达图

(1) 图形

2017年各年级单项成绩平均分雷达图



(2) 描述

2017年参加体测的各年级学生在各个项目的平均分均相差不大。其中，2017级在50米项目中平均分数最高，2016级在坐位体前屈和1000米项目中平均分数最高，2015级在800米项目中平均分数最高，2014级在BIM、肺活量、立定跳远、仰卧起坐和引体向上项目中平均分数最高。

(3) r 语言代码

```
library(fmsb)

# 构建测试数据集
data <- data.frame(row.names = c('2017 级','2016 级', '2015 级', '2014 级'),
                    "BIM" = c(93.45,94.39,94.27,94.51),
                    "肺活量" = c(76.53,77.66,77.93,79.25),
                    "五百米"= c(72.45,69.25,68.61,68.03),
                    "立定跳远"= c(57.55,65.67,65.38,67.00),
                    "坐位体前屈" = c(68.43,73.18,72.00,73.11),
                    "仰卧起坐"= c(52.28,56.21,60.91,61.27),
                    "引体向上" = c(14.26,20.81,21.54,22.95),
                    "八百米"= c(69.20,72.61,73.36,67.00),
                    "一千米"=c(66.09,67.41,65.65,61.12))
```

```

#定义每个变量的范围

max_min <- data.frame(row.names = c("Max", "Min"),
                       "BIM" = c(100,0),
                       "肺活量" = c(100,0),
                       "五十米"= c(100,0),
                       "立定跳远"= c(100,0),
                       "坐位体前屈" = c(100,0),
                       "仰卧起坐"= c(100,0),
                       "引体向上" = c(100,0),
                       "八百米"= c(100,0),
                       "一千米"= c(100,0))

# 合并数据

data_pro <- rbind(max_min,data)

color <- c("#FF022C", "#E7B800","#3AFF38","#0E36FF")

plot02 <- radarchart(data_pro,title = c("2017 年各年级单项成绩平均分雷达图"),
                      caxislabels = c(0, 20, 40, 60, 80,100),
                      pcol = color,
                      #pfcol = scales::alpha(color, 0.5),
                      plwd = 2, plty = 1,
                      cglcol = "grey", cglty = 1, cglwd = 0.8,
                      axislabcol = "grey",
                      vlabels = colnames(data),vlcex = 1,
)

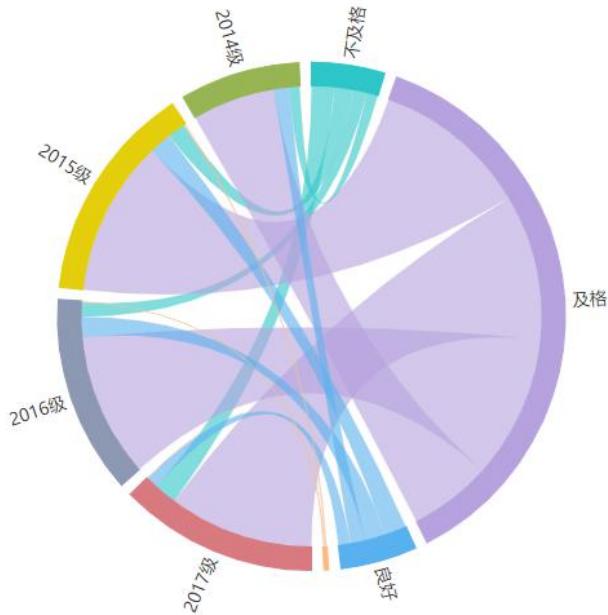
# 添加图例

legend(
  x=1.3,y=1.2, legend = c('2017 级','2016 级', '2015 级','2014 级'),
  bty = "n", pch = 20 , col = color,
  text.col = "black", cex = 1, pt.cex = 3.
)

```

3.8.4 2014-2017 等级分布弦图

(1) 图形



(2) 描述

由图可知，各年级中大部分都位于及格等级，极少数位于优秀等级。2017 级相比于 2015、2016 级，人数大致相同，但良好人数偏少、不及格人数偏多，且 2014 级无优秀等级，说明从年级来看，大二、大三阶段体育成绩好于大一、大四阶段。

(3) 代码

```
data<-read.csv(file="C:/Users/23842/Desktop/5.csv",header=T,encoding="GBK")
library(circlize)
test<-d0[d0$year0==1960,-1]
chordDiagram(x = test,
              directional = 1,
              order = d1$region,
              grid.col = d1$col1,
              annotationTrack = "grid",
              direction.type = c("diffHeight","arrows"),
              )
circos.track(track.index = 1, bg.border = NA,
            panel.fun = function(x, y) {
```

```

xlim = get.cell.meta.data("xlim")
sector.index = get.cell.meta.data("sector.index")
reg1 = d1 %>% filter(region == sector.index) %>% pull(reg1)
reg2 = d1 %>% filter(region == sector.index) %>% pull(reg2)
circos.text(x = mean(xlim), y = ifelse(is.na(reg2), 3, 4), labels = reg1, facing =
"bending", cex = 0.8)
circos.text(x = mean(xlim), y = 2.75, labels = reg2, facing = "bending", cex = 0.8)
circos.axis(h = "top", labels.cex = 0.6, labels.niceFacing = FALSE, labels.pos.adjust
= FALSE)
})

```

3.8.5 2014-2017 等级比率

(1) 图形



(2) 描述

由图可知，各年级及格率最高，优秀率最低，其中 2015 级优秀率最高。而不及格率 2017 级最高，说明大一新生要加强体育锻炼。

(3) 代码

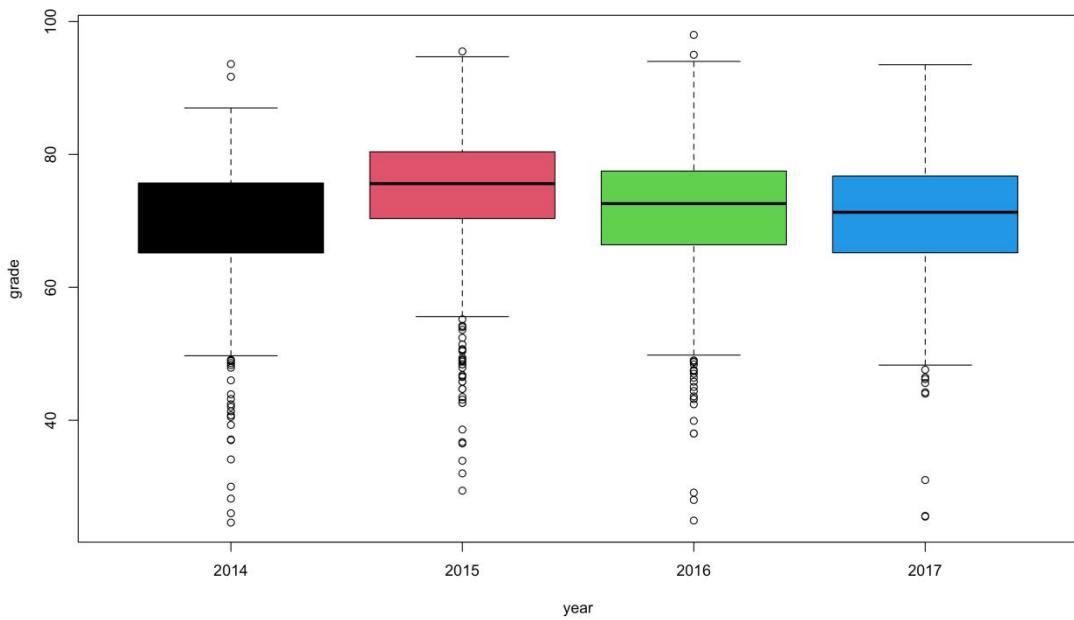
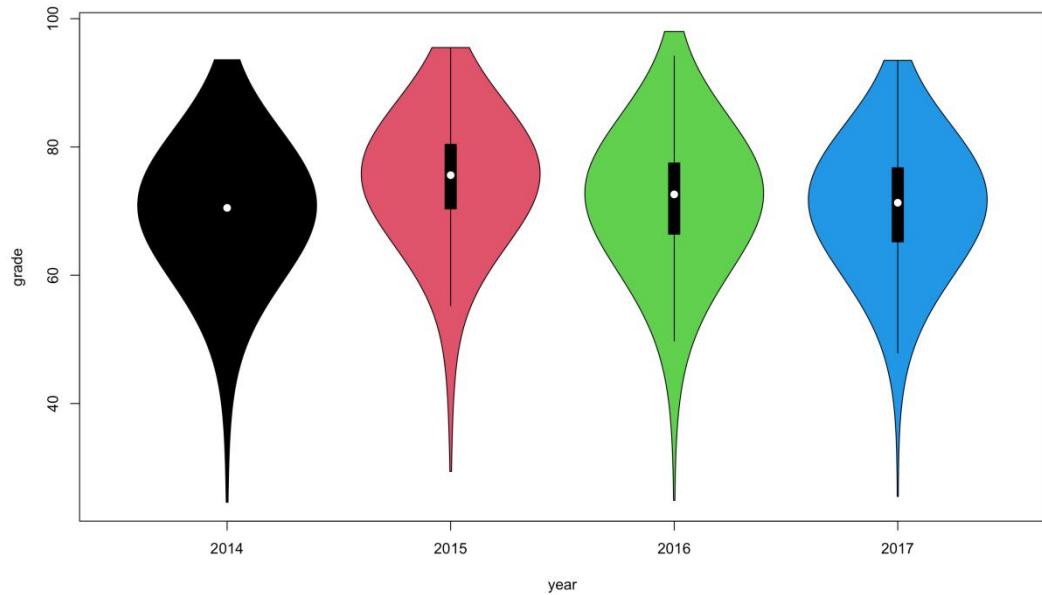
```

> data<-read.csv(file="C:/Users/23842/Desktop/2017.csv",header=T,encoding="GBK")
> counts<-table(data$等级,data$年级编号)
> lable=c("2017 级","2016 级","2015 级","2014 级")
> barplot(counts,main="2017 年 级 等 级 对 比 ",xlab=" 年 级 ",ylab=" 频 数
",names.arg=lable,col=c("lightgreen","lightblue","lightyellow","red"),legend=rownames(counts))

```

3.9 2014 级学生四年成绩变化

3.9.1 四年的总成绩小提琴图/箱线图



(2) 描述

2014 级学生从大一到大四总成绩从中位数来看相差不大，都位于 60-80 之间，其中大二时最高，大一时最低。从核密度来看，各阶段相差不多，密度最大部分都位于 70-80，其中大二时位于 80 左右的人数最多。说明在大学阶段，大部分学生在大二时体质最好，大一时体质最差。

(3) 代码

```

df_2014 <- read.csv("2014 附总成绩.csv")
grade_2014 <- df_2014[df_2014$年级编号 == 41, "总分数"]
df_2015 <- read.csv("2015 附总成绩.csv")
grade_2015 <- df_2015[df_2015$年级编号 == 41, "总分数"]
df_2016 <- read.csv("2016 附总成绩.csv")
grade_2016 <- df_2016[df_2016$年级编号 == 41, "总分数"]
df_2017 <- read.csv("2017 附总成绩.csv")
grade_2017 <- df_2017[df_2017$年级编号 == 41, "总分数"]
grade <- data.frame(year = c(rep(2014, length(grade_2014)),
                             rep(2015, length(grade_2015)),
                             rep(2016, length(grade_2016)),
                             rep(2017, length(grade_2017))),
                      grade = c(grade_2014, grade_2015, grade_2016, grade_2017))

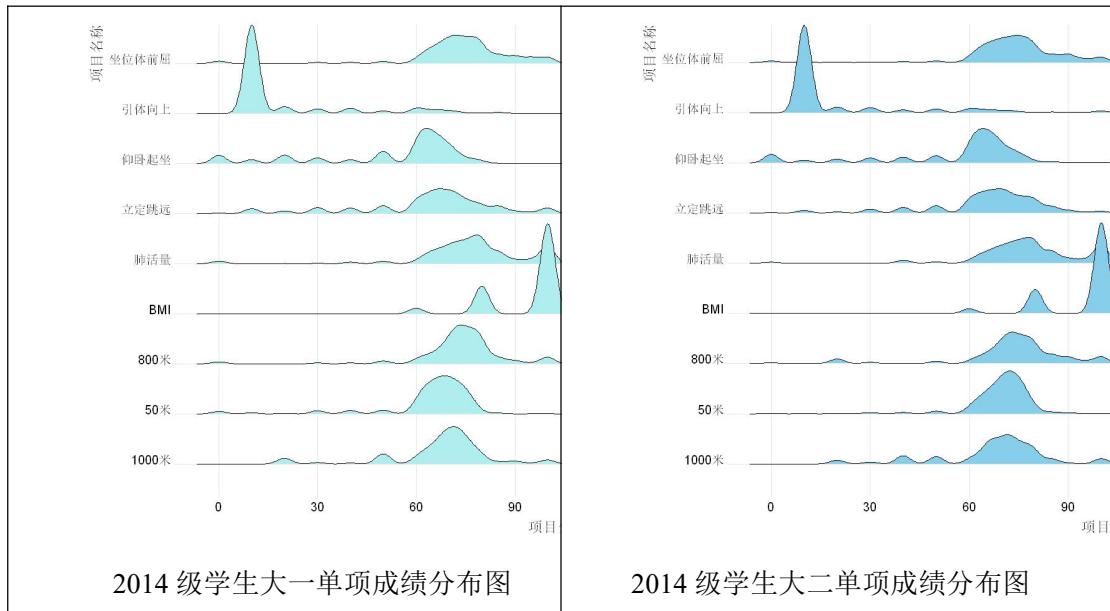
library(vioplot)

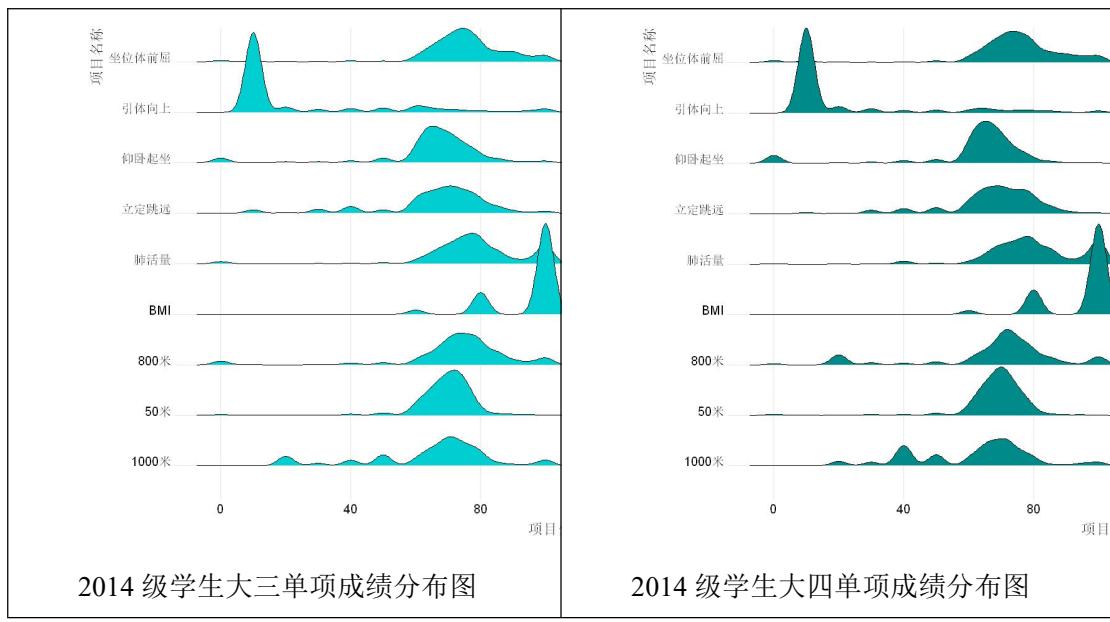
vioplot(grade ~ year, data = grade, col = 1:4)
boxplot(grade ~ year, data = grade, col = 1:4)

```

3.9.2 2014 级学生大学期间单项成绩分布山脊图

(1) 图形





(2) 描述

2014 级学生大一到大四期间单项成绩分布无较大变化，坐位体前屈、仰卧起坐、立定跳远中高分段人数有所上升，800m、50m、1000m 低分段人数有所上升，表明 2014 级学生大学期间柔韧度、弹跳能力有所提升，但随着年级升高，存在体测“摆烂”现象。

(3) r 语言代码

```

bh1 <- read.csv(file="bh1.csv")
ggplot(bh1, aes(x = 项目分数 ,y = 项目名称, fill = "cut")) +
  geom_density_ridges() +
  theme_ridges() +
  theme(legend.position = "none")+
  scale_fill_manual(values=c("paleturquoise"))
bh2 <- read.csv(file="bh2.csv")
ggplot(nj2, aes(x = 项目分数 ,y = 项目名称, fill = "cut")) +
  geom_density_ridges() +
  theme_ridges() +
  theme(legend.position = "none")+
  scale_fill_manual(values=c("skyblue"))
bh3 <- read.csv(file="bh3.csv")
ggplot(bh3, aes(x = 项目分数 ,y = 项目名称, fill = "cut")) +
  geom_density_ridges() +
  theme_ridges()

```

```

theme(legend.position = "none")+
scale_fill_manual(values=c("darkturquoise"))

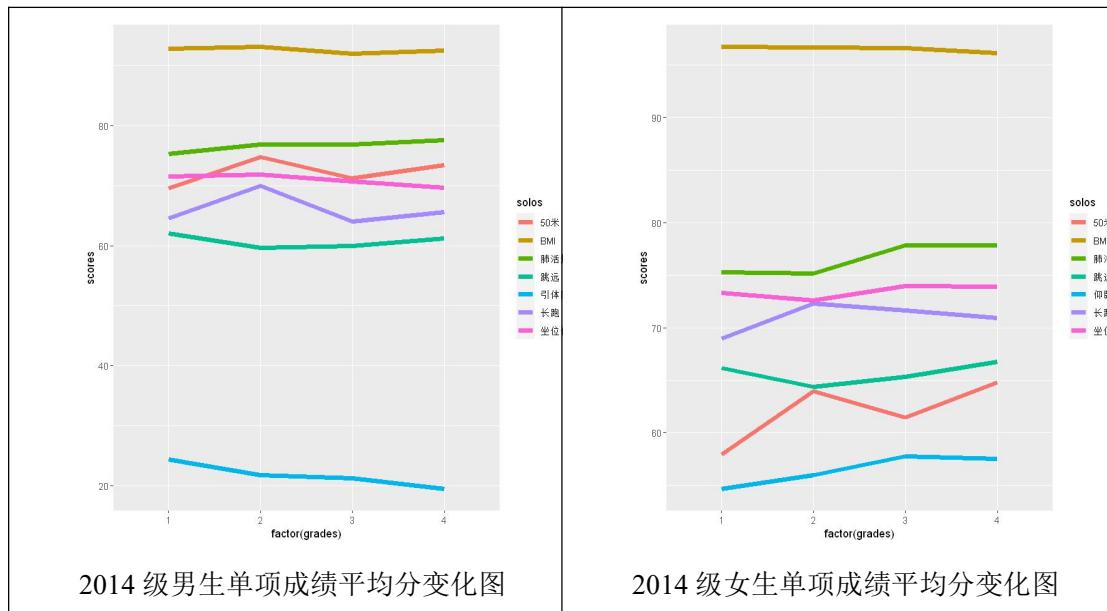
bh4<- read.csv(file="bh4.csv")

ggplot(bh4, aes(x = 项目分数 ,y= 项目名称, fill = "cut")) +
geom_density_ridges() +
theme_ridges() +
theme(legend.position = "none")+
scale_fill_manual(values=c("darkcyan"))

```

3.9.3 2014 级学生单项成绩平均分变化折线图

(1) 图形



(2) 描述

通过对 2017 年男生大一到大四单项成绩变化图进行分析，可以得出，男生引体向上成绩在大二时到达巅峰，大四又有所下降，与之趋势相同的是 50 米跑成绩和跳远成绩。BMI 得分比较平稳。坐位体前屈成绩四年差距不大。长跑成绩大二处于峰值，但在大三大四又有所下降，肺活量成绩大二最好，大三有所下降，大四有所上升。

通过 2017 年女生大一到大四单项成绩变化图进行分析，可以得出，女生 BMI 分数、肺活量成绩相对平稳。50 米成绩自大一到大三逐年下降，到大四有所提升。仰卧起坐成绩自大一到大三逐年提升，大三达最大值，大四有所下降。长跑成绩自大一到大二有所上升，大二到大三基本不变，大三到大四有所下降。坐位体前屈成绩大二达到巅峰，大三有所下降，大四又有所提升，跳远成绩大二达到巅峰，大二到大四无明显变化。

(3) r 语言代码

```
data1 <- read.csv(file="2014.csv")
data2 <- read.csv(file="2015.csv")
data3 <- read.csv(file="2016.csv")
data4 <- read.csv(file="2017.csv")
data1m=data1[data1$性别=="男",]
data2m=data2[data2$性别=="男",]
data3m=data3[data3$性别=="男",]
data4m=data4[data4$性别=="男",]
bmi1m=mean(data1m$BMI 分数)
bmi2m=mean(data2m$BMI 分数)
bmi3m=mean(data3m$BMI 分数)
bmi4m=mean(data4m$BMI 分数)
vc1m=mean(data1m$肺活量分数)
vc2m=mean(data2m$肺活量分数)
vc3m=mean(data3m$肺活量分数)
vc4m=mean(data4m$肺活量分数)
run_501m=mean(data1m$短跑分数)
run_502m=mean(data2m$短跑分数)
run_503m=mean(data3m$短跑分数)
run_504m=mean(data4m$短跑分数)
jump1m=mean(data1m$立定跳远分数)
jump2m=mean(data2m$立定跳远分数)
jump3m=mean(data3m$立定跳远分数)
jump4m=mean(data4m$立定跳远分数)
sit_bf1m=mean(data1m$坐位体前屈分数)
sit_bf2m=mean(data2m$坐位体前屈分数)
sit_bf3m=mean(data3m$坐位体前屈分数)
sit_bf4m=mean(data4m$坐位体前屈分数)
run1m=mean(data1m$长跑分数)
run2m=mean(data2m$长跑分数)
run3m=mean(data3m$长跑分数)
run4m=mean(data4m$长跑分数)
```

```

up1m=mean(data1m$引体向上分数)
up2m=mean(data2m$引体向上分数)
up3m=mean(data3m$引体向上分数)
up4m=mean(data4m$引体向上分数)

solos=c("BMI","BMI","BMI","BMI","肺活量","肺活量","肺活量","肺活量","50 米","50 米"
,"50 米","跳远","跳远","跳远","跳远","坐位体前屈","坐位体前屈","坐位体前屈","坐位体前屈","长跑"
,"长跑","长跑","长跑","引体向上","引体向上","引体向上")

grades=c("1","2","3","4")

scores=c(bmi1m,bmi2m,bmi3m,bmi4m,vc1m,vc2m,vc3m,vc4m,run_501m,run_502m,run_503m,run_
504m,jump1m,jump2m,jump3m,jump4m,sit_bf1m,sit_bf2m,sit_bf3m,sit_bf4m,run1m,run2m,run3m,run4
m,up1m,up2m,up3m,up4m)

tgg=data.frame(solos,grades,scores)

ggplot(tgg, aes(x=factor(grades),y=scores,colour=solos,group=solos)) + geom_line(size=2)

data1w=data1[data1$性别=="女",]
data2w=data2[data2$性别=="女",]
data3w=data3[data3$性别=="女",]
data4w=data4[data4$性别=="女",]

bmi1w=mean(data1w$BMI 分数)
bmi2w=mean(data2w$BMI 分数)
bmi3w=mean(data3w$BMI 分数)
bmi4w=mean(data4w$BMI 分数)

vc1w=mean(data1w$肺活量分数)
vc2w=mean(data2w$肺活量分数)
vc3w=mean(data3w$肺活量分数)
vc4w=mean(data4w$肺活量分数)

run_501w=mean(data1w$短跑分数)
run_502w=mean(data2w$短跑分数)
run_503w=mean(data3w$短跑分数)
run_504w=mean(data4w$短跑分数)

jump1w=mean(data1w$立定跳远分数)
jump2w=mean(data2w$立定跳远分数)
jump3w=mean(data3w$立定跳远分数)
jump4w=mean(data4w$立定跳远分数)

```

```

sit_bf1w=mean(data1w$坐位体前屈分数)
sit_bf2w=mean(data2w$坐位体前屈分数)
sit_bf3w=mean(data3w$坐位体前屈分数)
sit_bf4w=mean(data4w$坐位体前屈分数)

run1w=mean(data1w$长跑分数)
run2w=mean(data2w$长跑分数)
run3w=mean(data3w$长跑分数)
run4w=mean(data4w$长跑分数)

up1w=mean(data1w$仰卧起坐分数)
up2w=mean(data2w$仰卧起坐分数)
up3w=mean(data3w$仰卧起坐分数)
up4w=mean(data4w$仰卧起坐分数)

solos=c("BMI","BMI","BMI","BMI","肺活量","肺活量","肺活量","肺活量","50米","50米",
"50米","跳远","跳远","跳远","跳远","坐位体前屈","坐位体前屈","坐位体前屈","坐位体前屈","长跑",
"长跑","长跑","长跑","仰卧起坐","仰卧起坐","仰卧起坐","仰卧起坐")

grades=c("1","2","3","4")

scores=c(bmi1w,bmi2w,bmi3w,bmi4w,vc1w,vc2w,vc3w,vc4w,run_501w,run_502w,run_503w,run_50
4w,jump1w,jump2w,jump3w,jump4w,sit_bf1w,sit_bf2w,sit_bf3w,sit_bf4w,run1w,run2w,run3w,run4w,up1
w,up2w,up3w,up4w)

tgg=data.frame(solos,grades,scores)

ggplot(tgg, aes(x=factor(grades),y=scores,colour=solos,group=solos)) + geom_line(size=2)

```

4 推断性统计分析

4.1 主成分分析

4.1.1 模型建立

主成分分析（Principal Component Analysis，PCA），是一种统计方法。通过正交变换将一组可能存在相关性的变量转换为一组线性不相关的变量，转换后的这组变量叫主成分。

在用统计分析方法研究多变量的课题时，变量个数太多就会增加课题的复杂性。人们自然希望变量个数较少而得到的信息较多。在很多情形，变量之间是有一定的相关关系的，当两个变量之间有一定相关关系时，可以解释为这两个变量反映此课题的信息有一定的重叠。主成分分析是对于原先提出的所有变量，将重复的变量（关系紧密的变量）删去多余，建立尽可能少的新变量，使得这些新变量是两两不相关的，而且这些新变量在反映课题的信息方面尽可能保持原有的信息。

设法将原来变量重新组合成一组新的互相无关的几个综合变量，同时根据实际需要从中可以取出几个较少的综合变量尽可能多地反映原来变量的信息的统计方法叫做主成分分析或称主分量分析，也是数学上用来降维的一种方法。

4.1.2 数据准备

取出原始数据可用数据即各项分数。计算数据间的相关系数。得到矩阵如下：

	BMI分数	肺活量分数	X50米分数	立定跳远分数	坐位体前屈分数	仰卧引体分数	长跑分数
BMI分数	1.00000000	-0.05823382	0.09493142	0.2288135	0.05689562	0.20385473	0.19489354
肺活量分数	-0.05823382	1.00000000	0.07752583	0.1323592	0.19062126	0.05445748	0.06674533
X50米分数	0.09493142	0.07752583	1.00000000	0.2508628	-0.01648901	-0.14429805	0.21206799
立定跳远分数	0.22881349	0.13235924	0.25086282	1.0000000	0.20670799	0.29843513	0.23572248
坐位体前屈分数	0.05689562	0.19062126	-0.01648901	0.2067080	1.00000000	0.19994547	0.06627704
仰卧引体分数	0.20385473	0.05445748	-0.14429805	0.2984351	0.19994547	1.00000000	0.22991691
长跑分数	0.19489354	0.06674533	0.21206799	0.2357225	0.06627704	0.22991691	1.00000000

图 2 变量相关系数结果图

4.1.3 分析结果

作主成分分析，分析结果如下所示：

Importance of components:	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5
Standard deviation	1.3701046	1.0979816	1.0667481	0.9047205	0.8619768
Proportion of Variance	0.2681695	0.1722234	0.1625645	0.1169313	0.1061434
Cumulative Proportion	0.2681695	0.4403929	0.6029574	0.7198887	0.8260321
	Comp. 6	Comp. 7			
Standard deviation	0.8436473	0.71136084			
Proportion of Variance	0.1016773	0.07229061			
Cumulative Proportion	0.9277094	1.00000000			

图 3 主成分分析结果图

Standard deviation 行表示的是主成分的标准差，也就是特征值 $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6, \lambda_7$ 的开方，Proportion of Variance 行表示的是方差的贡献率，Cumulative Proportion 行表示的是方差的累计贡献率。前三个特征值均大于 1，第四个特征值接近于 1，表明其所对应的主成分变量包含的信息较多，五个主成分的贡献率分别为 26.81%、17.22%、16.27%、11.69%、10.61%，累计方差解释率 82.6%。因此，确定主成分的个数为 5 比较合理。

loadings（载荷）的内容如下所示：

Loadings:	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5	Comp. 6	Comp. 7
BMI分数	0.369	0.135	0.463	0.394	0.632	0.276	
肺活量分数	0.207	-0.174	-0.692	-0.267	0.612		
X50米分数	0.240	0.698	-0.303	0.176	-0.162		-0.553
立定跳远分数	0.535			0.226	-0.170	-0.590	0.524
坐位体前屈分数	0.321	-0.448	-0.320	0.449	-0.360	0.511	
仰卧引体分数	0.429	-0.442	0.311	-0.270		-0.312	-0.590
长跑分数	0.436	0.247	0.111	-0.646	-0.189	0.462	0.266

图 4 载荷结果图

它实际上是主成分对于原始变量各项分数的系数。也是特征值对应的特征向量，它们是线性无关的单位向量。第 1 列表示第 1 主成分 Comp1 的得分系数，依次类推。

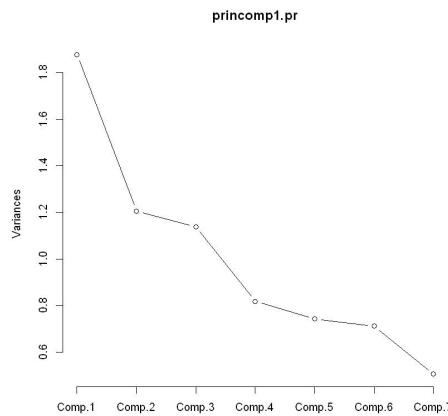


图 5 第一主成分得分系数结果图

由碎石图可以看出，第四个主成分之后，图线变化趋于平稳，因此可以选择前四个主成分做分析。

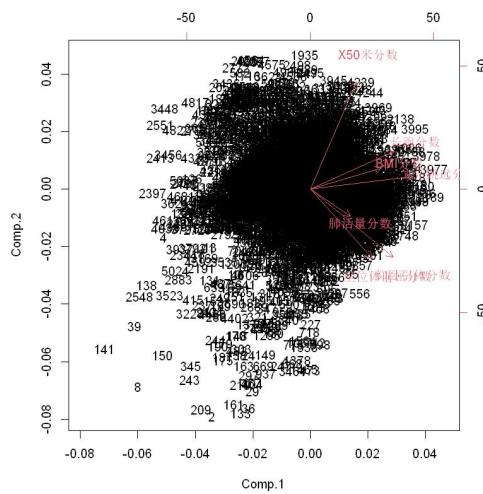


图 6 碎石图

在各主成分的表达式中，各标准化指标 X_i 前面的系数与该主成分所对应的特征值之平方根的乘积是该主成分与该指标之间的相关系数。系数的绝对值越大，说明该主成分受该指标的影响也越大。因此，决定第 1 主成分 Comp1 大小的主要为立定跳远分数；决定第 2 主成分 Comp2 大小的主要为 50 米分数；决定第 3 主成分 Comp3 大小的主要为肺活量分数；决定第 4 主成分 Comp 大小的主要为长跑分数。

4.2 影响体测成绩因素决策树分析

该部分主要是利用 2017 年中国地质大学（武汉）体质测试数据进行决策树分析，以此来探究

影响体测成绩的主要因素。男女生体测项目存在差异，因此在进行分析时该部分将男生、女生数据分离来分别进行决策树分析。

4.2.1 模型的建立

决策树的计算就是对数据进行挖掘分类的过程，利用决策树这一数据分类器，可以将一些无序的数据分类进行分析和推导。其中根节点、内部节点、叶节点是决策树用于分类计算的显著特征。决策树方法的每个根节点到叶节点都有相应的路径将数据按照一定规则进行分类。利用决策树对大学生体测成绩进行具体的研究，可以更加直观的显示出相关因素对体测成绩的影响。

一、划分选择

决策树的思想类似于我们做出选择，大部分的步骤都比较简单，但属性的选择则会直接影响我们决策树的预测准确度。一般而言决策树会通过不断划分节点，来使每一个分支都尽可能的归属到同一类。在决策树的划分中，主要存在三种准则，即信息增益准则、信息增益率准则、基尼指数准则。

1、信息增益准则

用信息增益表示分裂前后跟的数据复杂度和分裂节点数据复杂度的变化值，计算公式表示为：

$$Info_Gain = Gain - \sum_{i=1}^n Gain_i$$

其中 Gain 表示节点的复杂度，Gain 越高，说明复杂度越高。信息增益说白了就是分裂前的数据复杂度减去孩子节点的数据复杂度的和，信息增益越大，分裂后的复杂度减小得越多，分类的效果越明显。

2、信息增益率准则

使用信息增益作为选择分裂的条件有一个不可避免的缺点：倾向选择分支比较多的属性进行分裂。为了解决这个问题，引入了信息增益率这个概念。信息增益率是在信息增益的基础上除以分裂节点数据量的信息增益（听起来很拗口），其计算公式如下：

$$Info_Ratio = \frac{Info_Gain}{Intrinsic\ Info}$$

其中 Info_Ratio 表示信息增益， Intrinsic_Info 表示分裂子节点数据量的信息增益，其计算公式为：

$$Intrinsic\ Info = -\sum_{i=1}^m \frac{n_i}{N} * \log(\frac{n_i}{N})$$

其中 m 表示子节点的数量， n_i 表示第 i 个子节点的数据量，N 表示父节点数据量，说白了，其

实 IntrinsicInfo 是分裂节点的熵，如果节点的数据链越接近，IntrinsicInfo 越大，如果子节点越大，IntrinsicInfo 越大，而 Info_Ratio 就会越小，能够降低节点分裂时选择子节点多的分裂属性的倾向性。信息增益率越高，说明分裂的效果越好。

3、基尼指数准则

基尼值计算公式如下：

$$Gini = 1 - \sum_{i=1}^n p_i^2$$

其中 P_i 表示类 i 的数量占比。其同样以上述熵的二分类例子为例，当两类数量相等时，基尼值等于 0.5；当节点数据属于同一类时，基尼值等于 0。基尼值越大，数据越不纯。

二、停止分裂的条件

- 1、当节点的数据量小于一个指定的数量时，不继续分裂。
- 2、熵和基尼值的大小表示数据的复杂程度，当熵或者基尼值过小时，表示数据的纯度比较大，如果熵或者基尼值小于一定程度数，节点停止分裂。
- 3、决策树的深度是所有叶子节点的最大深度，当深度到达指定的上限大小时，停止分裂。
- 4、当已经没有可分的属性时，直接将当前节点设置为叶子节点。

三、剪枝

决策树是充分考虑了所有的数据点而生成的复杂树，它在学习的过程中为了尽可能的正确的分类训练样本，不停地对结点进行划分，因此这会导致整棵树的分支过多，造成决策树很庞大。决策树过于庞大，有可能出现过拟合的情况，决策树越复杂，过拟合的程度会越高。所以，为了避免过拟合，需要对决策树进行剪枝。

R 语言的 rpart 包提供了一种剪枝方法--复杂度损失修剪的修剪方法。并且 printcp 这个函数会告诉你分裂到的每一层，对应的某个点的复杂度是多少，平均相对误差是多少。然后我们就可以使用具有最小交叉验证误差的某个点的复杂度的方式进行剪枝。

4.2.2 模型的求解与分析

一、代码实现

1、工作目录和数据集的准备

```
setwd("D:/K")#设定当前的工作目录  
audit2<-read.csv("D:/K/ceshi.csv",header=T,fileEncoding = "GBK")  
str(audit2) #转成字符串类型的
```

2、做训练集和测试集

```
set.seed(1)  
sub<-sample(1:nrow(audit2),round(nrow(audit2)*2/3))  
length(sub)  
data_train<-audit2[sub,]#取 2/3 的数据做训练集
```

```

data_test<-audit2[-sub,]#取 1/3 的数据做测试集
dim(data_train)#训练集行数和列数
dim(data_test) #测试集的行数和列数
table(data_train$是否转化) #看该列分布的
table(data_test$是否转化)
3、做决策树模型
## rpart.control 对树进行一些设置
## xval 是 10 折交叉验证
## minsplit 是最小分支节点数，这里指大于等于 20，那么该节点会继续分划下去，否则停止
## minbucket: 叶子节点最小样本数,这里设置 100
## maxdepth: 树的深度
## cp 全称为 complexity parameter，指某个点的复杂度，对每一步拆分,模型的拟合优度必须提高的程度
library(rpart)
ct<-rpart.control(xval=10,minsplit=20,minbucket=150,cp=0.00017)
library(rpart)
tree.both<-rpart(as.factor(立定跳远分
数)~ .,data=data_train,method='class',minsplit=20,minbucket=150,cp=0.00017)
summary(tree.both)
tree.both$variable.importance
printcp(tree.both)
plotcp(tree.both,lwd=2)

```

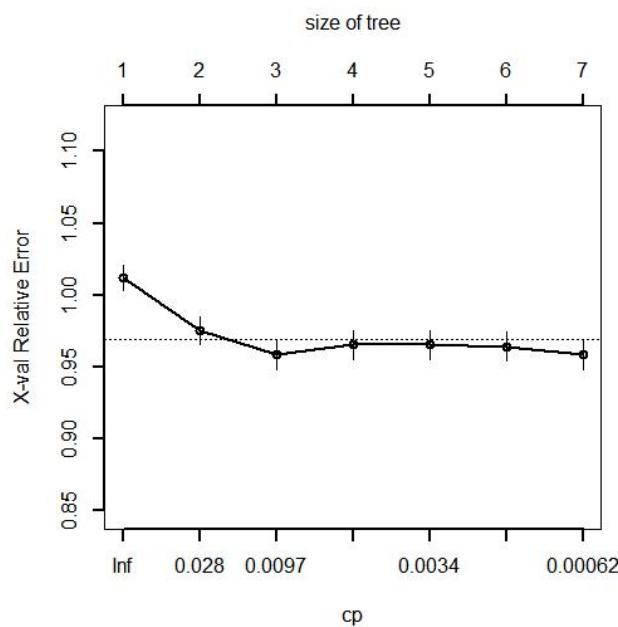


图 7 初始决策树规模可视化图

4、可视化决策树

```
library(rpart.plot)
```

```
rpart.plot(tree.both,branch=1,shadow.col="gray",box.col="green",border.col="blue",split.col="red",sp  
lit.cex=1.2,main="决策树")
```

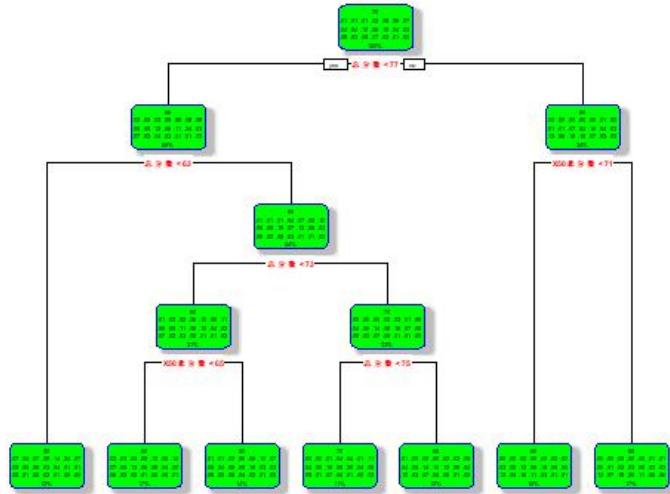


图 8 初始决策树可视化图

5、剪枝

```
printcp(tree.both)  
cp=tree.both$cptable[which.min(tree.both$cptable[, "xerror"]),"CP"]  
cp #cp=0.00049
```

6、剪枝之后再画图

```
tree.both2<-prune(tree.both,cp=tree.both$cptable[which.min(tree.both$cptable[, "xerror"]),"CP"])  
summary(tree.both2)  
tree.both2$variable.importance  
printcp(tree.both2)  
plotcp(tree.both2,lwd=2)
```

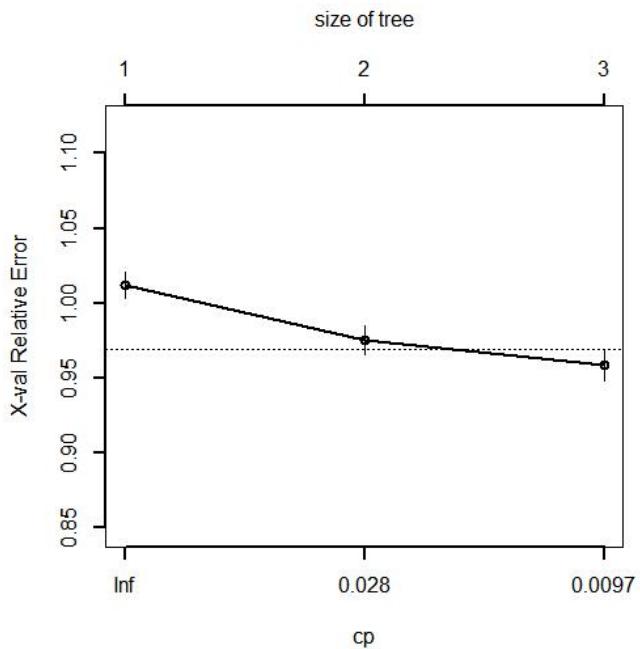


图 9 剪枝后决策树规模图

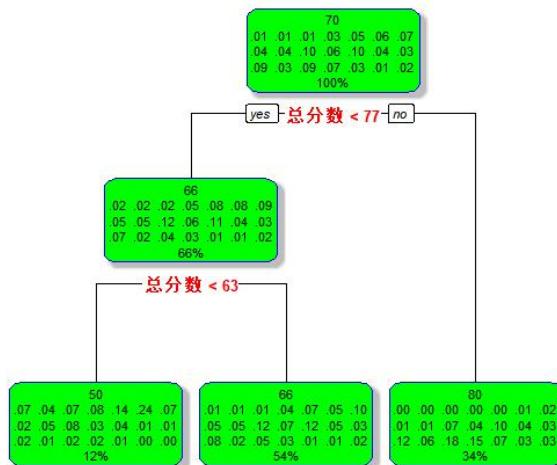


图 10 剪枝后决策树可视化

7、在测试集上做预测

```
library(pROC)
pred.tree.both<-predict(tree.both,newdata=data_test)
```

8、检测测试效果

```
predictScore<-data.frame(pred.tree.both)
rownames(predictScore) #看这个矩阵行的名字
colnames(predictScore)#看这个矩阵列的名字
```

二、结果分析

利用上述方法，对男生女生数据集进行决策树分析后，我们可以发现男生数据的决策树根节点是引体向上，女生数据决策树的根节点是立定跳远。根据决策树结果我们可以得出影响男生体测成

绩的因素的前三位排序依次是引体向上、1000米、立定跳远，影响女生体测成绩的因素的前三位排序依次是立定跳远、50米和800米。

从男生的前三位影响因素我们可以看出如果一个男生要想尽快提高体测成绩，那么他应该首先控制体重，锻炼上肢的肌肉力量和耐力，同时如果能辅以爆发力的训练，成绩提升将会更加显著。对于女生而言如果想要尽快提高体测成绩，那么她应该加强腿部力量的训练，如果能辅以耐力、持久力的训练那么成绩提高将更加显著。

4.3 速度与耐力项目聚类分析

4.3.1 模型的建立

(一) 体测项目训练体系的构建

通过体测工作获得的学生体质数据，可使大学生了解自己的发育程度、机能水平、身体素质和运动能力以及各个时期体质的发展与变化，从而引导大学生关注自己的体质状况，激发与培养大学生科学锻炼身体的自觉性与积极性。

本文建立了体测项目训练体系，将大学生体测项目划分为速度测试和耐力测试。由于男女生体质差异，个别测试项目会有所调整。

表 1 体测项目训练体系表

测试类别	性别	项目
速度测试	男、女	50 米短跑
		立定跳远
耐力测试	男	1000 米长跑
	女	引体向上 800 米长跑 仰卧起坐

根据前人研究，大学生体测速度与耐力素质测试内容为：通过50米短跑和立定跳远测试学生速度测试，通过男生1000米跑、引体向上以及女生800米跑、仰卧起坐测试学生的耐力测试。

(二) k-means 模型的构建

通过体测工作获得的学生速度与耐力测试数据，本文采用k-means聚类分析研究可使学校有关部门了解大学生针对速度与耐力体质变化的客观规律，并可检测与衡量学校体育教学对增强大学生特定体质的效果。下面将介绍K-Means算法实现大学生体测速度与耐力项目聚类分析。

K-Means算法的思想很简单，对于给定的样本集，按照样本之间的距离大小，将样本集划分为K个簇。让簇内的点尽量紧密的连在一起，而让簇间的距离尽量的大。

如果用数据表达式表示，假设簇划分为(c_1, c_2, \dots, c_k)，则我们的目标是最小化平方误差E：

$$E = \sum_{i=1}^k \sum_{x \in C_i} \|x - \mu_i\|_2^2 \quad (1)$$

其中 μ_i 是簇 C_i 的均值向量，有时也称为质心，表达式为：

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \quad (2)$$

我们随机选择了 n 个 k 类所对应的类别质心，然后分别求样本中所有点到 n 个质心的距离，并标记每个样本的类别为和该样本距离最小的质心的类别。经过计算样本和 n 个质心的距离，我们得到了所有样本点的第一轮迭代后的类别。此时，我们对我们当前标记为新的质心。重复上述过程，即：将所有点的类别标记为距离最近的质心的类别并求新的质心。最终我们得到的 n 个类别。

我们在每轮迭代时，要计算所有的样本点到所有的质心的距离，这样会比较的耗时。那么对于距离的计算，本文采用 elkan K-Means 算法加以改进。它的目标是减少不必要的距离的计算。

elkan K-Means 利用了两边之和大于等于第三边，以及两边之差小于第三边的三角形性质，来减少距离的计算。

第一种规律是对于一个样本点 x 和两个质心 j_1, j_2 。如果我们预先计算出了这两个质心之间的距离 $D(j_1, j_2)$ 。当发现满足式 (3) 条件时，我们得到式 (4)。此时我们不需要再计算 $D(x, j_2)$ ，也就是说省了一步距离计算。

$$2D(x, j_1) \leq D(j_1, j_2) \quad (3)$$

$$D(x, j_1) \leq D(x, j_2) \quad (4)$$

第二种规律是对于一个样本点 x 和两个质心 j_1, j_2 。我们可以根据如下式子节省距离计算。

$$D(x, j_2) \geq \max\{0, D(x, j_1) - D(j_1, j_2)\}$$

利用上边的两个规律，elkan K-Means 比起传统的 K-Means 迭代速度有很大的提高。但是如果我们的样本的特征是稀疏的，有缺失值的话，这个方法就不使用了，此时某些距离无法计算，则不能使用该算法。

4.3.2 模型的求解与分析

本次研究从 2017 年大学生体质测试数据，对体测中速度与耐力项目的成绩进行了聚类分析，通过体测数据了解当前学生速度与耐力素质现状，为高校今后的体育教学及学生的自我科学锻炼提供参考依据。

(一) 女大学生体质聚类分析

首先，通过 R 语言 factoextra 包中的 fviz_nbclust 函数，确定最佳聚类数目，如图所示。

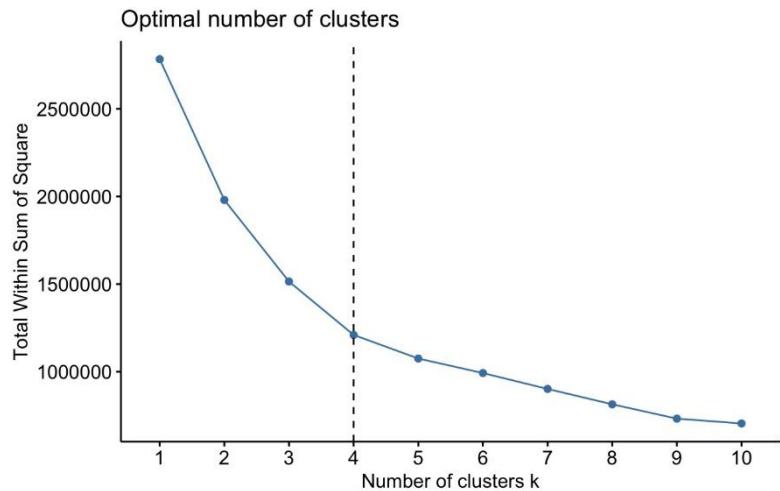


图 11 最佳聚类数量图

从指标的坡度变化来看，我们可以发现聚为四类最合适。因此，本文取 $k=4$ 。利用 k-means 模型进行聚类，得到以下结果。

```
> table(km$cluster)
```

	1	2	3	4
270	306	1436	177	

图 12 k-means 聚类结果图

本文将聚类结果进行可视化展示，如图所示。

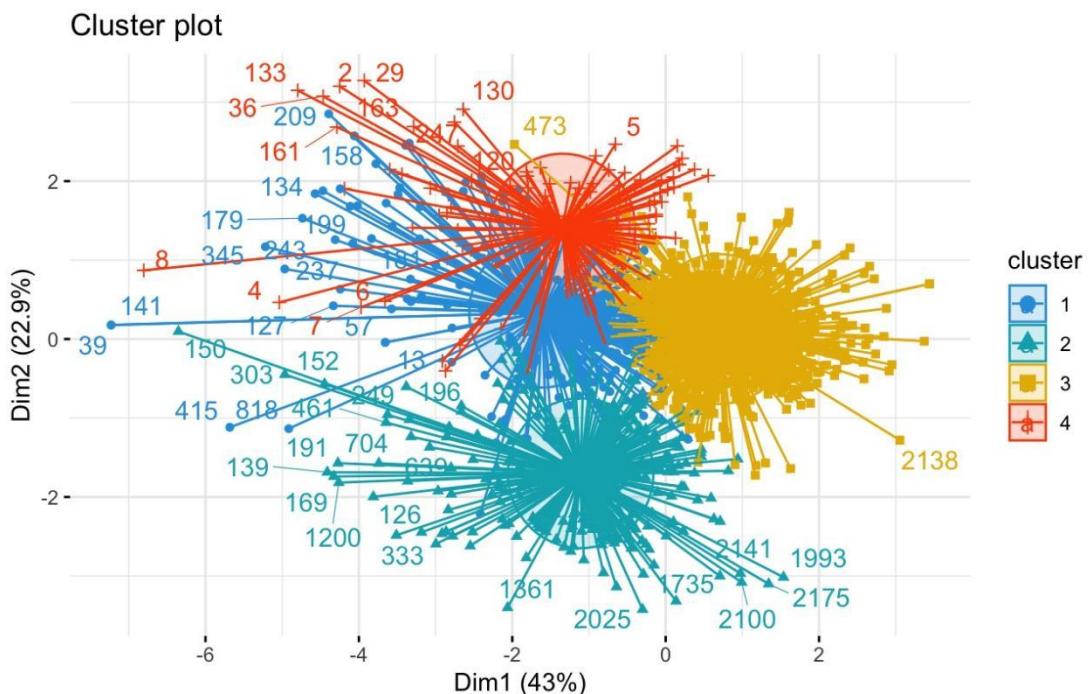


图 13 聚类结果可视图

我们可以发现，本文调查的女大学生素质主要集中在第三类人群。该类人群的速度素质较高，爆发力较强，尤其是立定跳远项目。第三类人群体测成绩主要分布在及格和良好。身高集中在160-165，体重分布跨度大。

(二) 耐力素质聚类分析

首先，通过 R 语言 factoextra 包中的 fviz_nbclust 函数，确定最佳聚类数目，如图所示。

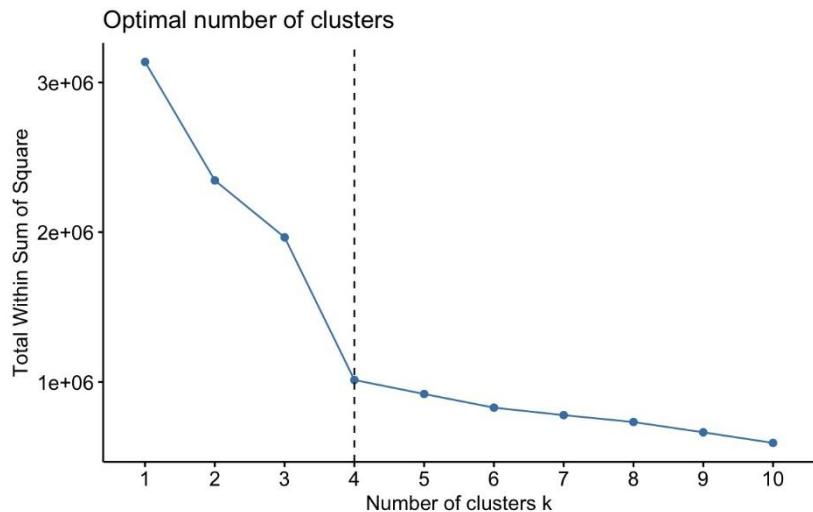


图 14 最佳聚类数量图

从指标的坡度变化来看，我们可以发现聚为四类最合适。因此，本文取 $k=4$ 。利用 k-means 模型进行聚类，得到以下结果。

```
> table(km$cluster)
```

	1	2	3	4
	437	1593	403	449

图 15 k-means 聚类结果图

本文将聚类结果进行可视化展示，如图所示。

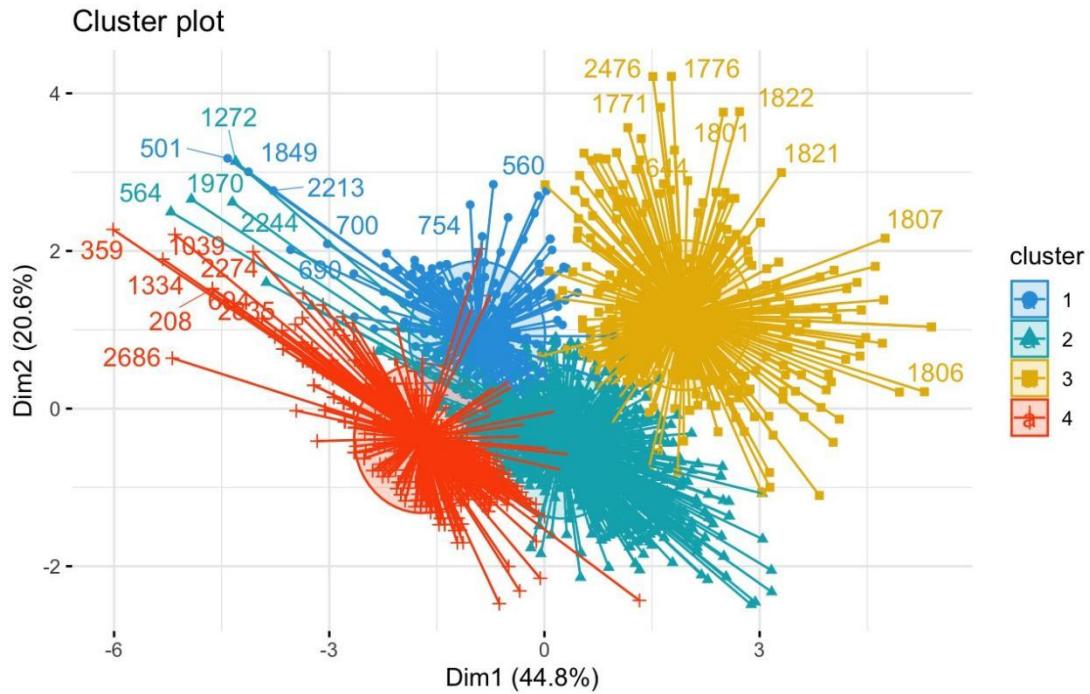


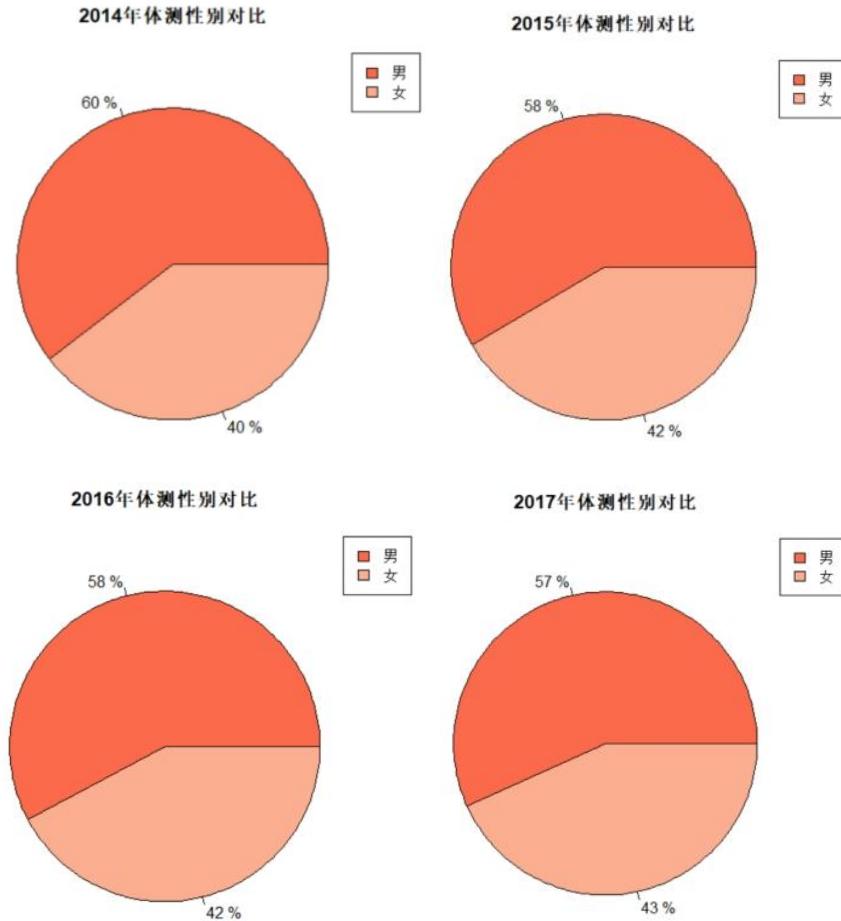
图 16 聚类结果可视图

我们可以发现，本文调查的男大学生素质主要集中在第二类人群。该类人群的速度与耐力素质表现跨度较大，第三类人群体测成绩主要分布在良好。对于立定跳远项目和引体向上项目整体分数相对较低。身高集中在 170-175cm，体重主要集中在 60-70kg，体型偏瘦。

5 结论与建议

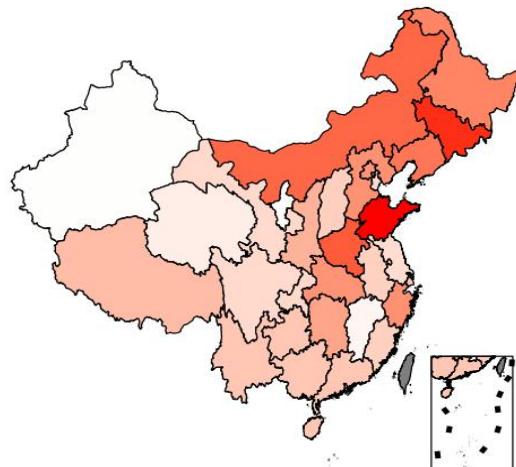
(1) 参加体测学生的性别比

对比 2014 年至 2017 年的参测同学性别比，可以发现男女比例基本维持在 6: 4。



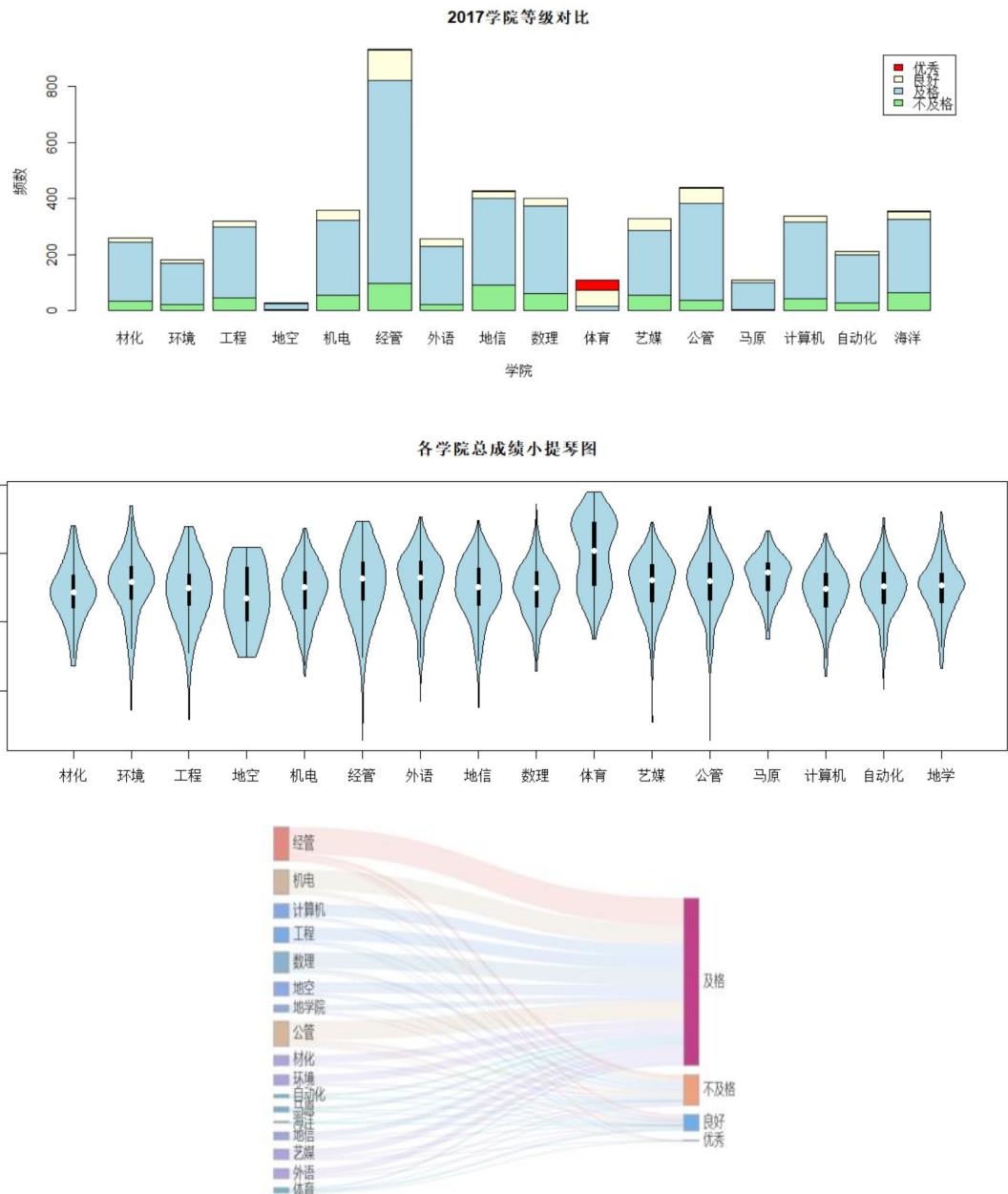
(2) 学生的身高-体重与地理位置分布情况

经统计可知，北方城市、东北部分的学生群体要显著高于其他地方的学生。



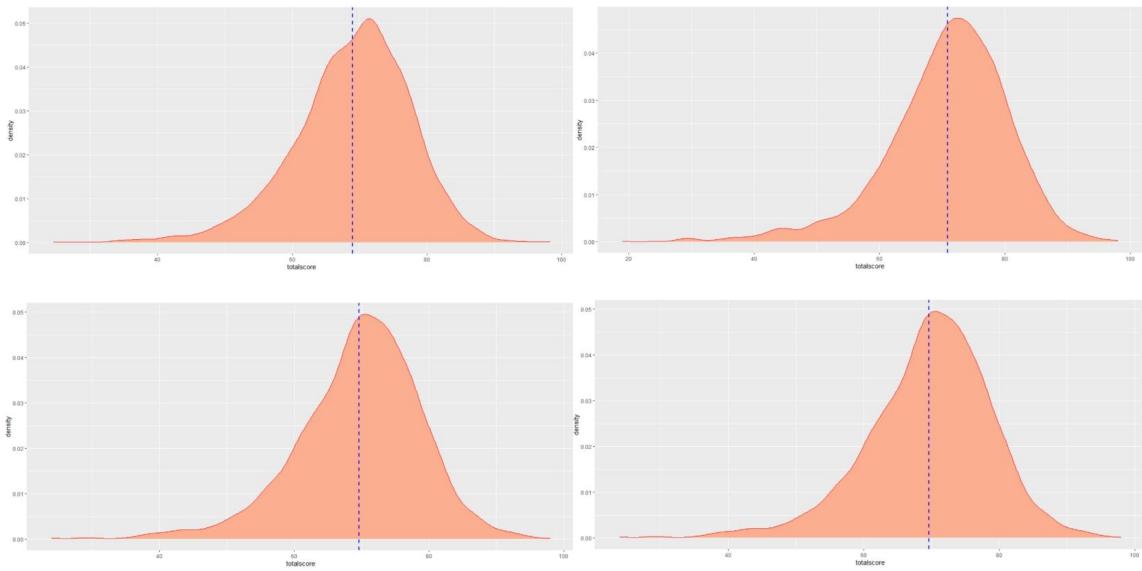
(3) 体测结果与学院分布

各学院除体育学院总成绩从中位数来看相差不大，都位于 60-80 之间，其中地空最低，马原最高，而体育学院遥遥领先，突破 80 分，说明体育学院学生体质远远强于其他学院学生。从核密度来看，各学院相差不多，密度最大部分都位于 70 上下，而体育学院成绩大多在 90 分左右。从上下限来看，公管、经管上限最高、下限最低，总成绩差距大，最不稳定；马原、地空上限与下限差距最小，成绩最稳定；体育学院上限高但下限也很高，说明体育学院每个人的体育成绩都偏好。



(4) 体测成绩总体情况

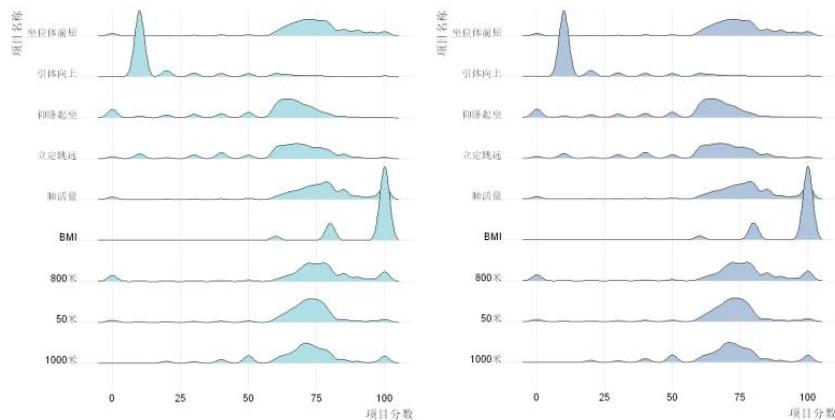
不同年份学生等级分布无明显差异。其中，等级为及格的人数最多，浮动在 4000 人左右；等级为优秀的人数最少，每年不过百人。

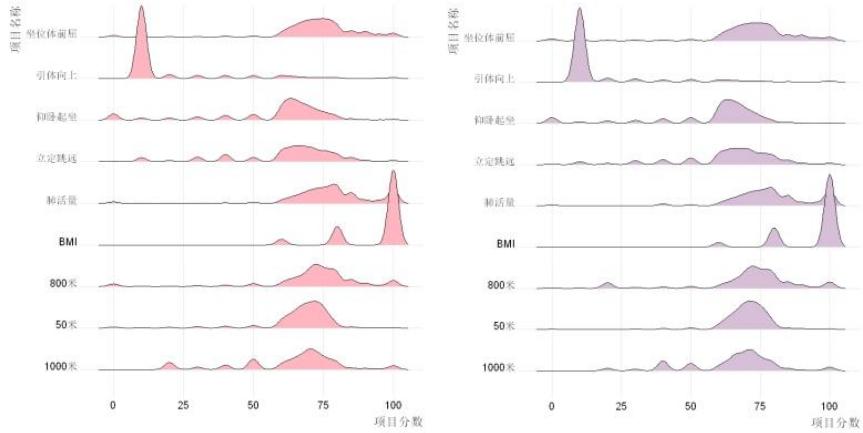


(5) 体测单项成绩情况

综合四年情况分析，不同年份单项成绩分布不呈现明显差异，坐位体前屈、仰卧起坐、立定跳远、肺活量、800m、50m 成绩分布大多集中在中上成绩段，引体向上多集中于低分段，近两年 1000m 中低分段人数较 2014、2015 年人数略有增长，**应加强对低分段男生的引体向上与长跑训练强度。**

不同年份学生等级分布无明显差异。综合每年成绩，学生 BMI 成绩最高，接近一百分，说明绝大部分学生身材匀称。肺活量、50 米、跳远、坐位体前屈、长跑成绩都位于 60-80 之间，而**仰卧起坐或引体向上成绩最低，不到 40 分，说明学生亟待加强该方面体育锻炼。**

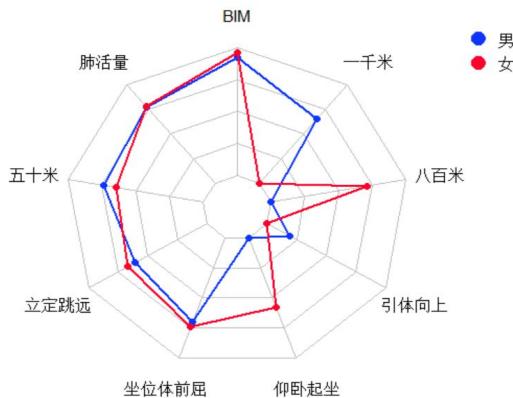




(6) 男女体测成绩对比

在 BMI、50 米跑、肺活量这三项上，男生成绩分布情况优于女生。而在坐位体前屈这一项中，男生虽然在中低分段人数多于女生，但在高分段还是女生更多，说明女生柔韧性要优于男生。在立定跳远这一项上也是如此，也能表明女生跳跃水平高于男生。但在引体或仰卧起坐这一方面，可以看出男生本方面得分大不如女生。0 分段以及低分段男生极多，女生得分情况亦较低。所以男女生普遍缺乏上肢以及腰腹力量锻炼，需要加强。

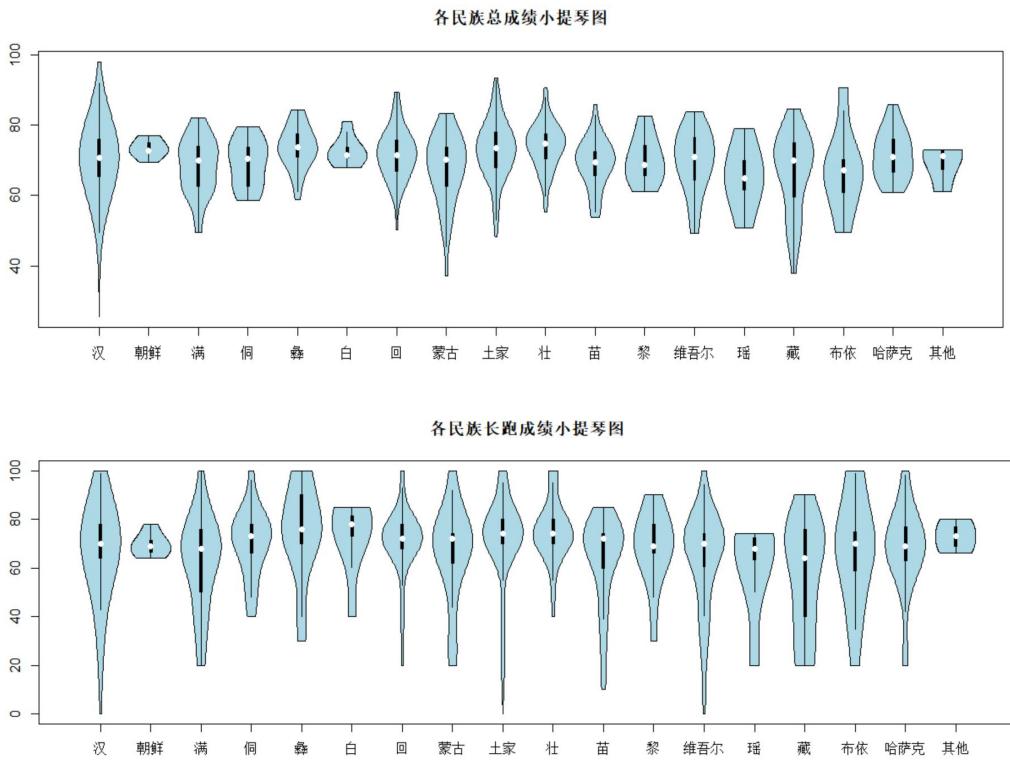
2017年男女单项成绩平均分雷达图



除了男女生分别特有的项目外，2017 年参加体测男女生平均分数相差不大。其中，男生在 50 米项目平均分数高于女生，女生在 BIM、肺活量、立定跳远、坐位体前屈项目平均分数高于男生。由图可知，男女生中大部分都位于及格等级，极少数位于优秀等级。但在男生人数多于女生的情况下，男生良好等级人数少于女生，不及格人数远远大于女生，说明 2017 年度女生体育成绩较好。

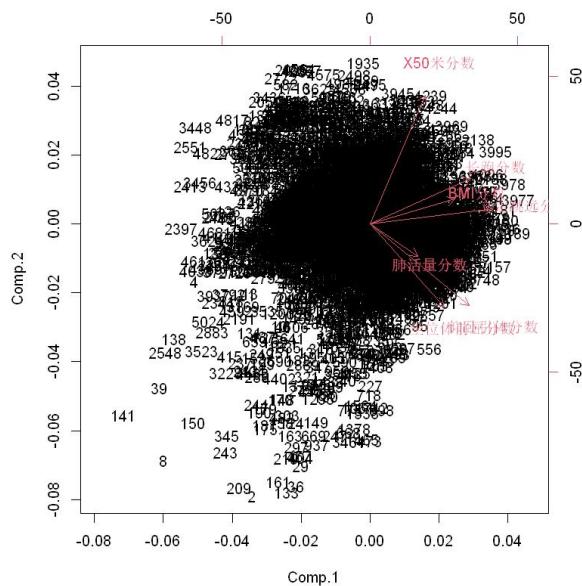
(7) 体测成绩与民族差异

各民族总成绩从中位数来看相差不大，都位于 60-80 之间，其中瑶族最低，壮族最高。从核密度来看，各民族相差不多，密度最大部分都位于 70 上下。从上下限来看，汉族上限最高、下限最低，成绩差距大，最不稳定；朝鲜族上限与下限差距最小，成绩最稳定。



(8) 体测各个指标的主成分分析

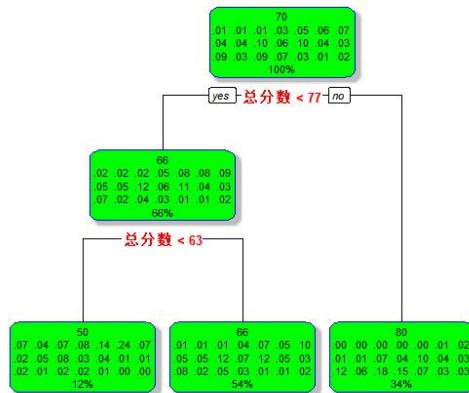
在各主成分的表达式中，各标准化指标 X_i 前面的系数与该主成分所对应的特征值之平方根的乘积是该主成分与该指标之间的相关系数。系数的绝对值越大，说明该主成分受该指标的影响也越大。因此，决定第 1 主成分 Comp1 大小的主要为立定跳远分数；决定第 2 主成分 Comp2 大小的主要为 50 米分数；决定第 3 主成分 Comp3 大小的主要为肺活量分数；决定第 4 主成分 Comp 大小的主要为长跑分数。



(9) 影响体测成绩的因素

利用上述方法，对男生女生数据集进行决策树分析后，我们可以发现男生数据的决策树根节点是引体向上，女生数据决策树的根节点是立定跳远。根据决策树结果我们可以得出影响男生体测成绩的因素的前三位排序依次是引体向上、1000米、立定跳远，影响女生体测成绩的因素的前三位排序依次是立定跳远、50米和800米。

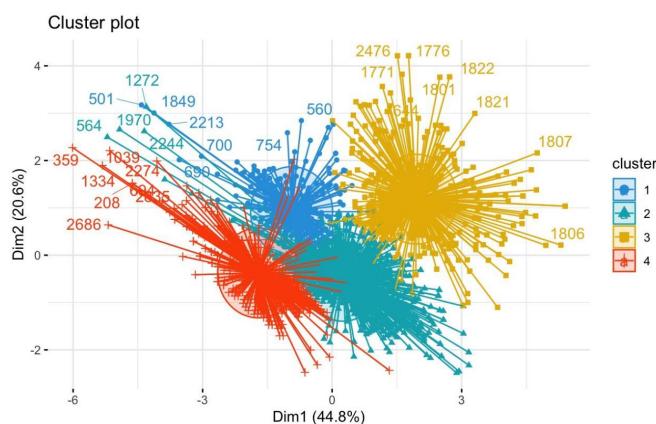
从男生的前三位影响因素我们可以看出如果一个男生要想尽快提高体测成绩，那么他应该首先控制体重，锻炼上肢的肌肉力量和耐力，同时如果能辅以爆发力的训练，成绩提升将会更加显著。对于女生而言如果想要尽快提高体测成绩，那么她应该加强腿部力量的训练，如果能辅以耐力、持久力的训练那么成绩提高将更加显著。



(9) 速度与耐力的分析

女生：我们可以发现，本文调查的女大学生素质主要集中在第三类人群。该类人群的速度素质较高，爆发力较强，尤其是立定跳远项目。第三类人群体测成绩主要分布在及格和良好。身高集中在160-165cm，体重分布跨度大。

男生：我们可以发现，本文调查的男大学生素质主要集中在第二类人群。该类人群的速度与耐力素质表现跨度较大，第三类人群体测成绩主要分布在良好。对于立定跳远项目和引体向上项目整体分数相对较低。身高集中在170-175cm，体重主要集中在60-70kg，体型偏瘦。



6 参考文献

- [1] 国务院关于实施健康中国行动的意见 [OL]. <http://tyfw.jschina.com.cn/zcfw/.shtml>,2016.
- [2] 郭建军, 杨桦. 中国青少年体育发展报告(2015) [M]. 北京: 社会科学文献出版社, 2015: 23.
- [3] 钟秉枢. 精准施策行胜于言 [J]. 中国学校体育, 2016(6): 2-3.
- [4] 教育部关于印发《国家学生体质健康标准(2014年修订)》的通知 [R]. 2014.
- [5] 刘悦. 菏泽医学专科学校大学生体质健康测试与评价研究 [D]. 济宁: 曲阜师范大学, 2017.

附录一：小组分工

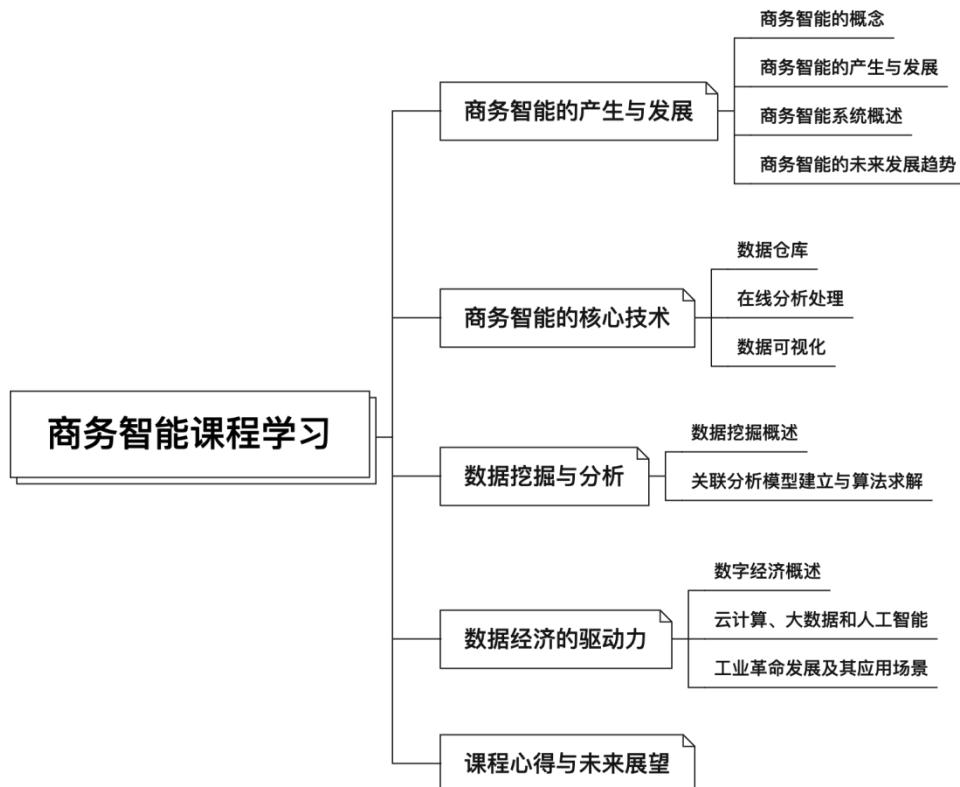
内容	具体分工
1 概述	贲雅雯
2 数据预处理	徐嘉艺
3 描述性统计分析	郭思琪： 袁 晴： 马宸晨： 袁应安：3.1 地理分布热力图
4 推断性统计分析	袁应安：4.1 主成分分析 普 叶：4.2 影响体测成绩因素决策树分析 贲雅雯：4.3 速度与耐力项目聚类分析
5 结论与建议	徐嘉艺

附录二：贾雅雯课程学习报告

一、课程报告概述

现今，商务智能越来越受到学术界和产业界的青睐，逐渐成为目前国内外企业界和软件开发界备受关注的一个研究热点。作为一项新兴的技术，在过去的十多年间，围绕商务智能的理论、方法、技术等的研究和应用已经取得了许多令人瞩目的成就。商务智能已发展成不仅仅只是软件产品和工具，而是一种整体应用的解决方案，甚至升华为一种管理思想，体现的是一种理性的经营管理决策的能力，即全面、准确、及时、深入分析和处理数据与信息的能力。

通过这学期对商务智能这门课程的学习，我了解到了很多关于商务智能相关的理论知识以及应用案例。本文将重点阐述在课程中对所学习内容的理解，主要包括商务智能的产生与发展、商务智能的核心技术、数据挖掘与分析、商务智能的应用。最后，我将分享这门课程的个人学习心得。



二、商务智能的产生与发展

1、商务智能的概念

在课堂上，老师为我们详细介绍了商务智能的概念。通过老师的讲解以及教材的理解，我梳理了有关商务智能定义的发展，如表 1 所示。商务智能作为一门学科由几个相关活动组成、包括数据挖掘、在线分析处理(OLAP)、查询和报告。有学者提出商务智能是组织中的大规模决策支持系统(DSS)的总称。现在，商务智能依然是组织中最大的 IT 投资领域，也被评为全球最优先的技术领域。

表 2-1 商务智能定义发展整理表

年份	学者	定义内容
1958 年	Hans Peter Luhn	一种用于生意处理上的信息系统。
1989 年	Howard Dressner	一种描述一系列概念和方法并通过应用基于事实的支持系统来辅助商业决策的制定。
2013 年	一些学者	一种用于分析组织原始数据的各种软件应用程序的总称。

我认为，商务智能是融合了先进信息技术与创新管理理念的结合体，从企业的数据仓库中敏捷提取能够创造商业价值的信息并呈现可操作的信息，以指导决策者做出更好决策的信息技术手段。商务智能可以服务企业战略，还可以提升企业绩效。他能够对企业的内外部数据进行分析，支持企业战略管理。商务智能更多地是用来解决管理问题。通过商务智能能从企业多年运营的数据中，挖掘有效的模式辅助管理决策。

2、商务智能的产生与发展

商务智能是随着 Internet 的高速发展和企业信息化的不断深入而产生的。其发展也是一个渐进的、复杂的演变过程，而且目前仍然处于发展之中。它经历了事物处理系统(Transaction Processing System, TPS)、高级管理人员信息系统(Executive Information System, EIS)、管理信息系统(Management Information System, MIS) 和决策支持系统(Decision Support System, DSS)等几个不同阶段，最终演变成今天的企业商务智能系统(BIS)。它是一个可包含企业所有知识的系统，服务于管理决策层或部门执行经理，帮助其进行分析和决策。我整

理了这四个阶段下商务智能发展的特征，如图 1 所示。

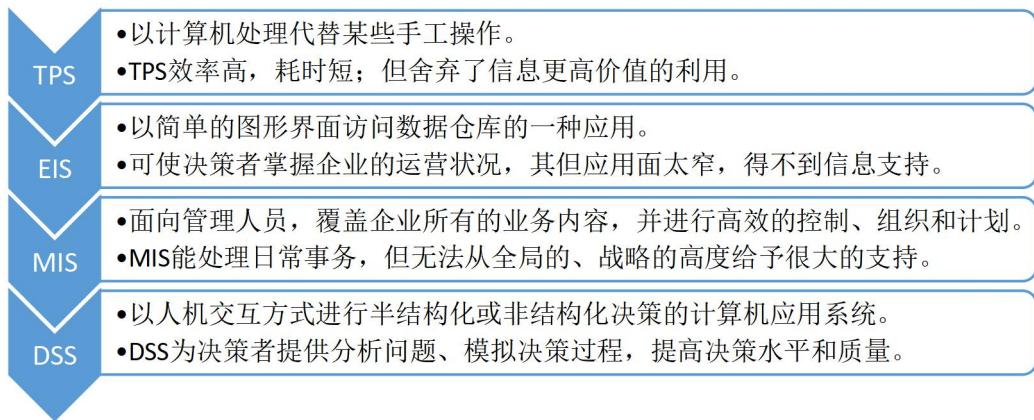


图 2-1 商务智能演变阶段

3、商务智能系统概述

由于企业的不断发展，数据的累计和海量信息的增长，使得激烈的市场竞争下企业越来越依赖通过信息共享平台对商务数据进行多维度分析，以满足信息资源的集中式和精确化管理，进而及时准确地满足决策的需要，显然这种为企业提供全面服务的信息系统是一种商务智能系统。

商务智能系统是通过数据仓库、在线分析和数据挖掘技术来处理和分析商业数据，并提供针对不同行业特点或特定应用领域的解决方案来协助用户解决在商务活动中所遇到的复杂问题，从而帮助企业决策者面对商业环境的快速变化做出敏捷的反应和更好、更合理的商业决策的系统。商务智能系统是一种整合系统。它运用数据仓库、联机分析和数据挖掘技术来处理和分析商业数据。它能从不同的数据源搜集的数据中提取有用的数据，并对这些数据进行清洗与整理，以确保数据的正确性。然后对数据进行转换、重构等操作，并将其存入数据仓库或数据集市中。同时运用合适的查询、分析、数据挖掘、OLAP 等管理分析工具对信息进行处理，使信息变为辅助决策的知识，并将知识以适当的方式展示在决策者面前，供决策者使用。商务智能系统有助于提高企业工作效率，建立有利的客户关系，增加产品的销售，帮助企业从现有的“知本”中提炼更多的价值。

通过文献查阅及老师的讲解，我绘制了商务智能系统框架图，如图 3-1 所示。商务智能涉及一个很宽的领域，集收集、合并、分析、提供信息存取功能于一体，包括抽取、转换、装载软件工具、数据仓库、数据查询和报告、联机数据分析、数据挖掘和可视化等工具，能够在线分析和挖掘知识，为决策者提供特定的决策解决方案。从商务智能系统内数据流程可

以看出，商务智能系统框架通常由数据产生层（数据源）、数据交换层、数据整合层、数据服务层、数据应用层和用户访问层（信息展示）组成。

数据产生层，也称作数据源层，是整个数据仓库的基础，也是商业智能的基础，包括企业内部数据和外部数据。内部数据主要来自经营过程中产生的各种业务数据，如 ERP、SCM 中产生的信息，这里有结构化数据和非结构化数据。外部信息主要指企业收集的来自网络、行业期刊等有关市场、竞争对手情况的信息。这些数据可以是结构化的，也可以是非结构化的。

这里再具体说明一下数据应用层，也就是数据分析层，该层是数据存储和前端分析工具的桥梁，能按照用户的要求设计、生成具有多维分析功能的分析主题，予以组织，以便进行多角度、多层次的分析，并发现趋势。它们响应前端用户的分析请求，将多维数据传送给前端的分析工具显示。主要的技术包括数据挖掘和联机分析处理技术。

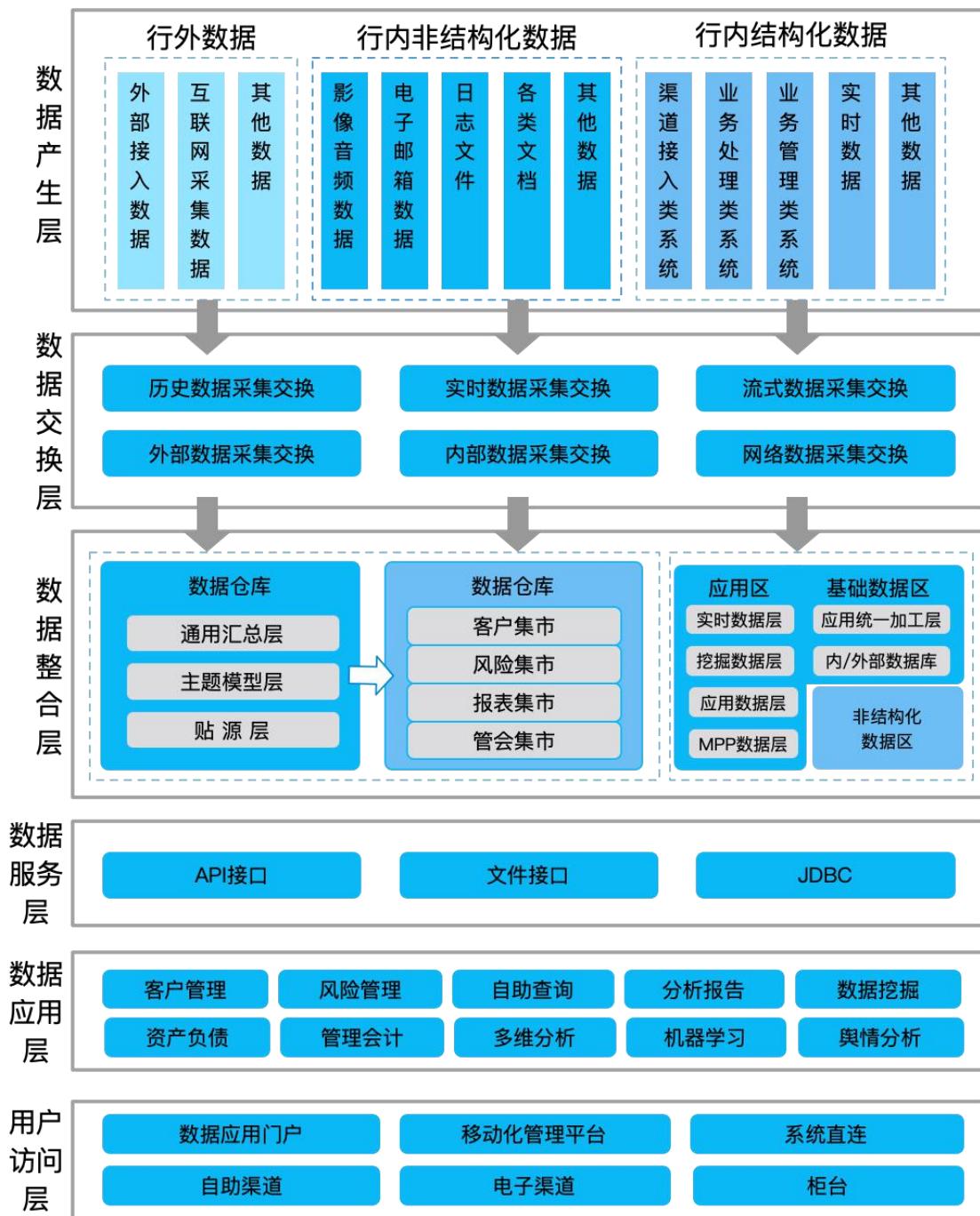


图 3-1 商务智能系统框架图

在课后，我查询了有关商务智能系统的应用案例。高爽等人（2020 年）提出了企业销售与分销商务智能系统的设计与实现。该研究以 SAP 商务智能为主要的研究对象，利用商务信息仓库、高性能分析器等核心技术，以某公司销售与分销的项目实施为背景，进行了模型设计、模型建立、数据抽取，以及报表的设计与展示，使决策者从宏观到微观全面了解企业状况，获取更加直观有效的企业经营决策信息。马小民（2017 年）设计了甘肃联通商务智能系统的集成优化方案，针对系统集成优化的总体架构、系统数据仓库以及系统 OLAP

应用进行分析设计。在原有架构的优化基础上，集成优化后的商务智能系统具有创新价值和应用价值的数据分析功能。



图 2-2 甘肃联通商务智能系统集成优化内容图

4、商务智能的未来发展趋势

我认为，商务智能的发展趋势可以归纳为以下几点。

(1)具有可配置性、灵活性、可变化的功能。

商务智能系统的范围从为部门的特定用户服务扩展到为整个企业所有用户服务。同时，由于企业用户在职权、需求上的差异，商务智能系统提供广泛的、具有针对性的功能。从简单的数据获取，到利用 Web 和局域网、广域网进行丰富的交互、决策信息和知识的分析和使用。

(2)解决方案更开放、提供客户化的界面。

针对不同企业的独特需求，商务智能系统在提供核心技术的同时，使系统又具个性化，即在原有方案基础上加入自己的代码和解决方案，增强客户化的接口和扩展特性；可为企业提供基于商务智能平台的定制工具，使系统具有更大的灵活性和使用范围。

(3)从单独的商务智能向嵌入式商务智能发展。

这是目前商务智能应用的一大趋势，即在企业现有的应用系统中，如财务、人力、销售等系统中嵌入商务智能组件，使普遍意义上的事物处理系统具有商务智能的特性。考虑商务智能系统的某个组件而不是整个商务智能系统并非一件简单的事，如将联机分析处理技术应用到某一个应用系统，一个相对完整的商务智能开发过程，如企业问题分析、方案设计、原型系统开发、系统应用等过程是不可缺少的。

(4)从传统功能向增强型功能转变。

增强型的商务智能功能是相对于早期用 SQL 工具实现查询的商务智能功能。目前应用中的商务智能系统除实现传统的商务智能系统功能之外，大多数已实现了数据分析层的功能。而数据挖掘、企业建模是商务智能系统应该加强的应用，以更好地提高系统性能。

此外，“商务智能”关键词在学术研究上，萧文龙等人（2020年）利用 Cite Space 软件，选取国内外商业智能和大数据分析相关文献，从文献结构特征、研究热点与研究趋势等方面进行可视化分析，讨论商业智能和大数据分析领域的未来发展。研究结果显示，商业智能研究领域越来越成熟，国际上发文量呈稳步上升的趋势，我国相关研究虽起步较晚，仍处在探索和发展阶段，但是发文量呈逐年上升趋势。从关键词共现来看，国际上研究热点包括系统、管理、绩效、模型、技术和知识；国内研究热点包括数据挖掘、数据仓库、决策支持、企业管理；国外研究与管理和决策有关，而国内研究更侧重于技术和应用。从关键词突现图来看，国外研究前沿由技术转向商业分析；国内研究前沿由数据挖掘技术转向人工智能。

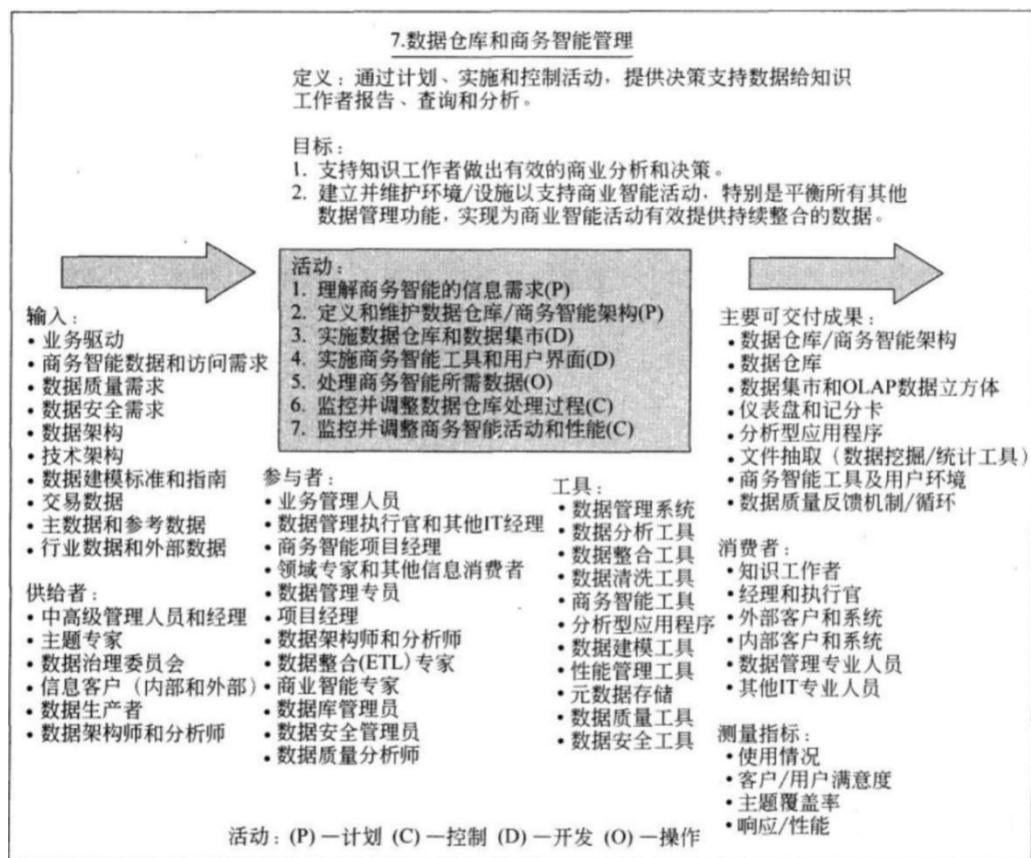
三、商务智能的核心技术

1、数据仓库

在课堂上，李老师详细地为我们讲解了数据仓库是什么，它的产生与发展以及它的概念与特征表现。通过学习，我了解到数据仓库是指一个面向主题的、集成的、相对稳定的、反映历史变化的数据集合，用于支持管理决策。建立数据仓库的目的是建立一种体系化的数据存储环境，将决策分析所需的大量数据从传统的操作环境中分离出来，使分散、不一致的操作数据转换成集成、统一的信息，为用户提供查询和决策分析的依据。

其中，老师就“数据在操作性系统中的存储与在数据仓库中的存储的差异性”为我们对“数据仓库面向主题”特征做了详细的解答。在操作型系统中，我们使用独立的应用程序来存储数据例如，关于汽车保险政策的索赔在自动保险应用中处理，汽车保险的索赔数据在这个应用中组织。同样，工人赔偿保险的索赔数据也在工人赔偿保险应用中组织。在数据仓库中，数据是为主题而不是为应用而存储的例如，索赔对于一家保险公司来说就是非常重要的主题。在保险公司的数据仓库中，索赔数据按照索赔的主题进行组织，而不是根据像汽车保险或是工人赔偿保险这样的单独应用来进行组织。其次，数据仓库还有集成、相对稳定、反应历史变化等特征。

其次，数据仓库与商务智能存在着一定的关系。下图通过 DAMA 知识体系解读数据仓库和商务智能管理之间的联系。数据仓库包括两部分：数据库和软件程序。两者结合可以支持数据统计、分析和商务智能的各种需求。从广义上来说，能够为商务智能提供数据支持的任何数据抽取或者数据存储都可以称之为数据仓库（DW）。



2、在线分析处理

OLAP 是使分析人员、管理人员或执行人员能够从多角度对信息进行快速、一致、交互地存取，从而获得对数据的更深入了解的一类软件技术。OLAP 的目标是满足决策支持或者满足在多维环境下特定的查询和报表需求，通过把一个实体的多项重要的属性定义为多个维，用户能对不同维上的数据进行比较。因此 OLAP 也可以说是多维数据分析工具的集合。

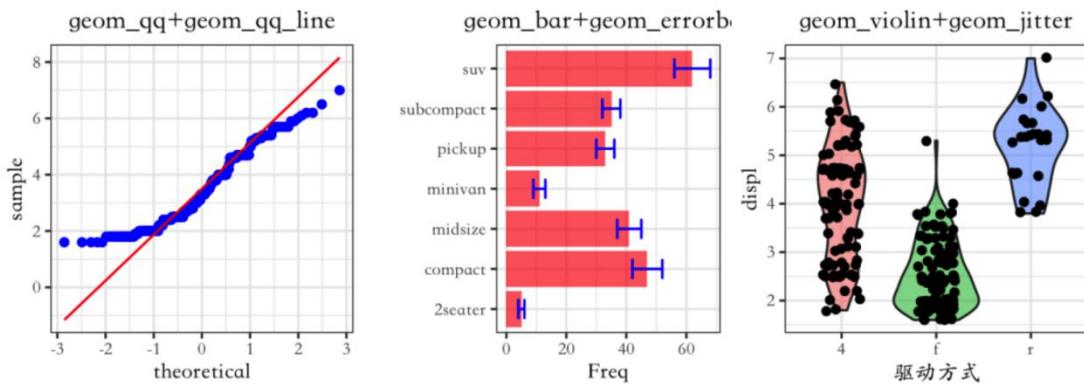
OLAP 是基于多维模型定义了一些常见的面向分析的操作类型，使这些操作显得更加直观。OLAP 的多维分析操作包括钻取、上卷、切片、切块以及旋转。

3、数据可视化

数据可视化可以让数据分析更加便捷，对于“信息管理与信息系统”专业的同学来讲，熟练掌握可视化工具可以在对数据进行处理的过程中，更加方便、快捷与精准。信管专业其中一个就业方向就是“IT 咨询”。通过老师课题上讲述的案例，让我知道具备这样的能力可以让我们在人群中脱颖而出。这样的数据分析不仅能更加贴近人们的生活，还能满足人们的实际生活需要。此外，数据可视化还具有很好的交互性，不仅设计功能良好，且使用过程中更加有意义，更加容易被人们理解和接受。

对于我们小组完成的“大学生体测数据分析”作用，可以让我们更好地掌握利用 R 语言来实现数据可视化。在进行实验过程中，我发现 R 语言绘制的图形更美观，而且绘图非常简

便。其中，在基于 R 语言的数据可视化方面，`ggplot2` 包已经发展成为最受欢迎的 R 包，并且在 `ggplot2` 包的基础上，还衍生出了各种各样的 R 包用来丰富 `ggplot2` 的绘图功能，将这些包和 `ggplot2` 包结合使用，能够获得更加精美的图像。



四、数据挖掘与分析

1、数据挖掘概述

数据挖掘使数据处理技术进入了一个更高级的阶段。它不仅能对过去的数据进行查询，并且能够找出与过去数据之间的潜在联系，进行更高层次的分析，以便更好地做出理想的决策、预测未来的发展趋势等。通过数据挖掘，有价值的知识、规则或高层次的信息就能从数据库的相关数据集合中抽取出来，从而使大型数据库作为一个丰富、可靠的资源为知识，的提取服务。

2、关联分析建模建立与算法求解

在老师推荐的教材中，详细地介绍了分类分析、关联分析、聚类分析、深度学习、Web 挖掘技术等数据挖掘分析技术。这里，由于课堂上老师具体给我们介绍了关联分析的应用。因此，作为课堂回归与深入学习，我将分享有关关联分析方法的模型建立与 Apriori 算法求解。

首先，结合老师课堂讲述的内容与文献查阅，我整理出了关联分析的相关概念，如图所示。



① 确定最小支持度和最小置信度

最小支持度和最小置信度都是描述事件发生的概率，所以取值范围在 0 和 1 之间。假如最小支持度设定过高，就会导致一些重要但不频繁的项集被过滤掉；设定过低，一方面，会影响计算性能，另一方面，一些无实际意义的数据也会被留下来，最小置信度也是同理。但判断它们是否“合适”的感觉很微妙，没有特定的标准答案，可以根据过往经验、试错法、事务出现的最小频率等去思考，

② 找出频繁项集和强关联规则

这一步，Apriori 的算法主要依赖两个性质。一个项集如若是频繁的，那它的非空子集也一定是频繁的。一个项集如若是非频繁的，那包含该项集的项集也一定是非频繁的。算法一旦找到某个不满足条件的“非频繁项集”，包含该集合的其他项集不需要计算，更不用对比，通通绕开。凭借新生成的频繁项集，就可以开始“制造”关联规则了。

因为频繁 1 项集只有一个项，无法构成具有指导意义的关联关系 (≥ 2 项)，可直接忽略。Apr 算法会先产生一系列后件项数为 1 的关联规则，与最小置信度进行比较，得到一部分强关联规则。

然后，频繁项集继续生成后件的项数为 2 的关联规则，再对它们的置信度进行比较，又收获一批强关联规则。当无法从剩下的频繁项集中生成新的关联规则时，该过程就结束了。若是这些强关联规则正好是你想要的，那就进一步计算它们的提升度。相反，若是你对筛选出来的强关联规则不满意，那我们就得重新调整最小支持度和最小置信度，再计算一次。

③ Python 调用 apriori 函数

尽管 Apr 算法已经对原始的关联分析做了优化，但手动计算依然繁琐，特别是当我们想要调整最小支持度或者最小置信度的时候。如果能把 Apr 算法的计算流程抽象成函数，

将最小支持度、最小置信度和最小提升度设置成参数，通过调整参数来查看关联规则，想怎么调就怎么调，那就最好了。实际上，Python 已经实现了这一切，那就是 apriori 函数 apyori 模块属于 Python 的第三方模块，在本地使用它，需要先安装一下。

apriori 函数常用的参数有 4 个：transactions（事务的集合），min_support（最小支持度），min_confidence（最小置信度），min_lift（最小提升度）。

apriori(transactions, min_support, min_confidence, min_lift)

功能：执行apriori算法，并返回包含关联规则的数据

参数	说明	示例
transactions	事务的集合，值可以是嵌套列表或者 Series 对象	apriori([[['薯条', '可乐'], ['可乐']]])
min_support	最小支持度，默认值为0.1	apriori([[['薯条', '可乐'], ['可乐']]], min_support=0.3)
min_confidence	最小置信度，默认值为0.0	apriori([[['薯条', '可乐'], ['可乐']]], min_confidence=0.5)
min_lift	最小提升度，默认值为0.0	apriori([[['薯条', '可乐'], ['可乐']]], min_lift=1)

3、数据挖掘在商务智能的应用

随着科学技术的发展和商业物流行业信息系统的不断完善，信息资源的开发越来越受到重视。要对海量信息进行科学的分析处理，为决策者提供决策支持，从而适应激烈的市场竞争，就必须提高数据组织和分析的技术水平。数据挖掘就是从海量的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。将数据挖掘技术应用在商业物流领域是本文研究的重点。

欧阳升乙（2009 年）以某商业物流总公司的需求为背景，在对现代数据挖掘和数据仓库技术分析与研究的基础上，开展面向数据仓库系统相关模块的研究与设计。根据商品供应物流行业数据仓库结构特点，从数据库的安全性角度考虑，进行数据仓库的逻辑建模。在深入学习 ETL 技术基础上，确定商品分析系统方案，实现商品分析系统的分群分类。进一步的开展系统模型的研究与设计，根据系统总体功能结构要求，确定建立系统模型的步骤，实现系统关键技术，根据物化视图选取的策略和方法，研究物化视图选取的代价模型和处理方法，完成了数据立方体的物化及选取。

杨斌等人（2012 年）就此学习研究了数据挖掘和商务智能的相关知识，分析了目前针对电子商务网站数据挖掘的主要研究方向 Web 日志挖掘，针对基于 Web 日志处理来获取访客行为数据所存在的不足，对数据挖掘技术在智能商务中的应用进行了分析，将实时在线挖掘与定期人工挖掘相补充的商务客户行为分析技术进行了研究。

五、数字经济的驱动力

1、数字经济的概述

数字经济是指以使用数字化的知识和信息作为关键生产要素、以现代信息网络作为重要载体、以信息通信技术的有效使用作为效率提升和经济结构优化的重要推动力的一系列经济活动。对于数字经济的理解，李老师有着独特的见解。数字经济也可以拆成两部分，一个是互联网，另一个是“互联网+”，学名叫产业数字化（德国工业4.0）或数字产业化（美国工业互联网）。

对于“产业数字化”和“数字产业化”的区分与理解，老师通过生动的案例给我们详细地讲述。产业数字化的体现其中就有智慧城市建设。智慧城市是涵盖范围十分广泛的系统，涉及智慧交通、智慧社区、智慧工地等诸多产业与应用场景。产业数字化本质就是分析产业行为数据，根据行为习惯提供相应服务，省去人工环节，使产业运营、发展、管理更加精细化、智能化、数字化。而数字产业化，就是让数据为产业发展提供服务。人的每一个行为、产业流程的每一个环节，都会产生大量数据，最终构建起以数据为关键要素的数字经济。让数据为现代化经济发展服务、让数据成为产业，同时发挥数据的基础资源作用和创新引擎作用，从而形成以技术创新和数字化为主要引领和支撑的数字经济。

2、云计算、大数据和人工智能

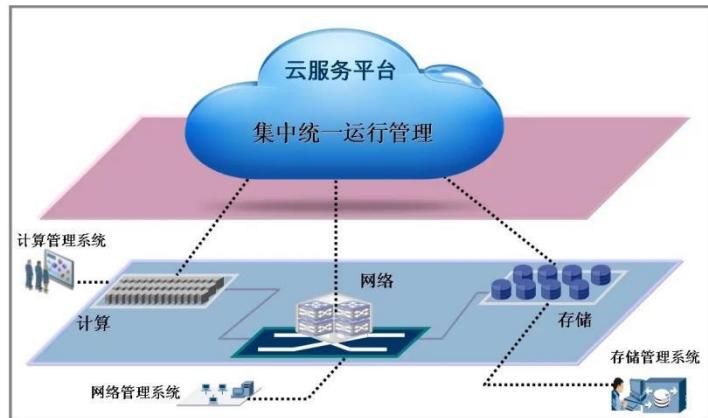
大数据产业正蓬勃发展，借助大数据的风口，云计算和人工智能也同时走进我们的视野，他们三者之间有着不可分割、相互影响的关联。

对于云计算的理解，我认为是基于互联网的相关服务的增加、使用和交付模式，这种模式提供可用的、便捷的、按需的网络访问，进入可配置的计算资源共享池（资源包括网络，服务器，存储，应用软件，服务），这些资源能够被快速提供，只需投入很少的管理工作，或与服务供应商进行很少的交互。通常涉及通过互联网来提供动态易扩展且经常是虚拟化的资源。过去在图中往往用云来表示电信网，后来也用来表示互联网和底层基础设施的抽象。因此，云计算强大的计算能力可以模拟核爆炸、预测气候变化和市场发展趋势。我们在日常生活中通过电脑、笔记本、手机等方式接入数据中心，也可以按自己的需求进行运算。

大数据是需要新处理模式才能具有更强的决策力、洞察力和流程优化能力的海量、高增长率和多样化的信息资产。简言之，从各种各样类型的数据中，快速获得有价值信息的能力，就是大数据技术。我认为，大数据时代已经来临，它将在众多领域掀起变革的巨浪。但大数据的核心在于为客户挖掘数据中蕴藏的价值，而不是软硬件的堆砌。因此，针对不同领域的大数据应用模式、商业模式研究将是大数据产业健康发展的关键。我相信，在国家的统筹规划与支持下，通过各地方政府因地制宜制定大数据产业发展策略，通过国内外企业以及众多创新企业的积极参与，大数据产业未来发展前景十分广阔。进充分利用大数据的价值。

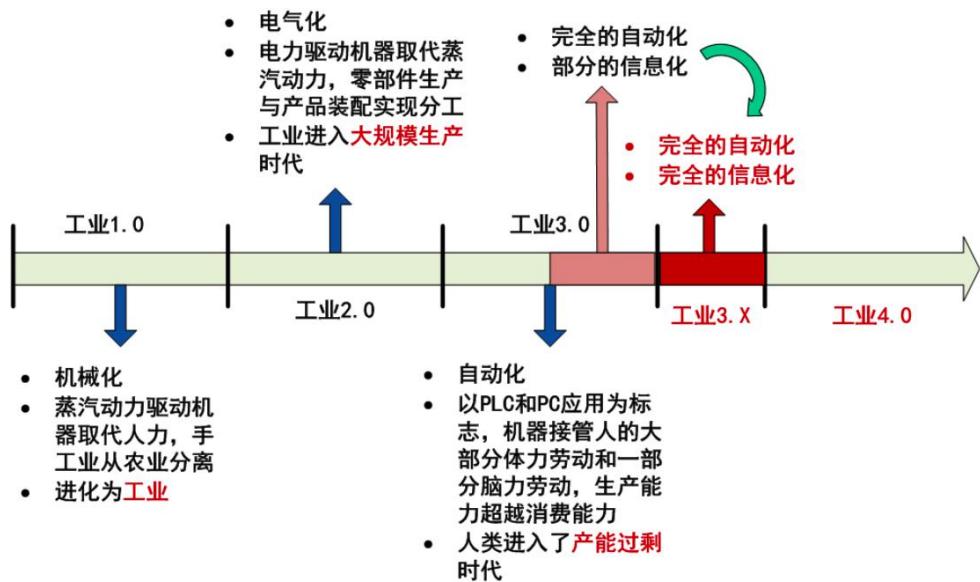
人工智能是研究使计算机来模拟人的某些思维过程和智能行为（如学习、推理、思考、

规划等)的学科，主要包括计算机实现智能的原理、制造类似于人脑智能的计算机，使计算机能实现更高层次的应用。人工智能将涉及到计算机科学、心理学、哲学和语言学等学科。可以说几乎是自然科学和社会科学的所有学科，其范围已远远超出了计算机科学的范畴，人工智能与思维科学的关系是实践和理论的关系，人工智能是处于思维科学的技术应用层次，是它的一个应用分支。



3、工业革命发展及其应用场景

在“工业革命发展及其应用场景”这个部分的学习中，我认为老师的见解是非常独特的，并提出了“工业 3.X”的概念，认为一些企业并不是从工业 3.0 直接过渡到工业 4.0。其中，还有经过“企业完全自动化、信息化”阶段，才能进入工业 4.0。



在工业 3.0 时期，工业企业中的主要有两大核心部门，即生产部门和业务部门；以及两大系统，即 MES(制造执行系统，生产部门)和 ERP(管理信息系统，业务部门)。其中，ERP 更倾向于财务信息的管理，而 MES 更倾向于生产过程的控制，主要负责监控和管理生产这

些瓶子的每一个步骤和工序如何实现。当 ERP 给 MES 下达生产计划指令后，MES 在生产过程中发生了与计划偏差的事项，MES 会根据车间的实际情况进行调整。但是 ERP 是不知道的，它会继续按照原本的计划执行订单，时间久了，财务系统和工厂的实际情况就会出现非常大的偏差。

这是因为 ERP 和 MES 的开发公司通常是两拨人，搞财务的和搞生产的合作，不但互相不懂对方的职业术语，而且互相看不上对方。业务部门和生产部门在公司里通常是分开运营，各自的领导有各自喜欢的供应商。此时，工厂车间通常会定期把 MES 的调整项做成一个表，交给业务部门，然后由业务部门手动在 ERP 中调整过来。这个问题只是工厂内系统断层的一个问题缩影，事实上工厂里还有非常多的其他系统，如设计、采购、办公等。这些系统都是一一个个的信息孤岛，互相不知道对方在干嘛，干到哪一步了。只有等到问题出现了，才能退回来，所有系统再一个个改。

随着人们越来越喜欢个性化的东西，这使得传统工业必须要快速、小批量、定制化的生产。要实现这样的生产方式，先得做点准备工作，就是把 ERP 和 MES 等等信息系统彻底打通，生产的原材料和生产设备连接起来(RFID 射频识别技术)，将工业 3.0 进化为工业 3.X，即完全的自动化和完全的信息化，就能冲击工业 4.0 了。

六、课程心得与未来展望

在这门课程中，我主要学到了商务智能的产生与发展、商务智能的核心技术、数据挖掘与分析、商务智能的应用、数字经济的驱动力、R 语言实践应用。对于课程中提到的“数据挖掘、云计算、大数据、人工智能、物联网”等概念在课堂之前接触过，直到大三学期上了李四福老师讲授的商务智能课程后，对此才有了更加深刻的理解，让我将这些思想与方法与案例实现融会起来。并且，在某些课程内容的讲解时，老师提出了自己的研究成果与独特的见解，这对于我们来说是十分宝贵的。

在完成小组作业的时候，作为组长最开始的想法是分成数据预处理、描述性统计分析、推断性分析这三个部分分工完成。在开始工作前，老师为我们提供了一些开展思路，例如：在数据分析前，小组可以先根据大学生体质核算公式，计算出各项分数及总分数，便于后续的研究分析；最好针对不同人群特征进行分析，这里指的是不同身份的对象想要掌握的信息是不同的，这个最好可以利用上学期学到的“面向对象分析设计”的思想来解决这个问题。在小组完成作业工程中，我们进行了多次学习讨论、互相合作，包括软件操作、r 语言实现、数据可视化。在代码实现过程中，我个人也遇到了一些问题，例如：中文不兼容、数据读取乱码、包的安装与使用等，通过同学之间的帮助学习及 csdn 网站的资料查阅，逐一排查问题，找出问题根源，最终完成了我负责的工作任务。

商务智能课程的学习，不仅仅提高的是理论知识，还有未来就业的帮助。通过这门课程的学习，掌握了一些实用的工作技能，对未来就业数据分析、IT 咨询等岗位十分有帮助。

由于我个人对 IT 咨询岗位非常感兴趣，因此在完成小组作业时也十分努力，同时在课后也额外查阅资料，并与学长学姐交流有关就业发展相关的问题。

总而言之，非常感谢李四福老师和助教学姐的指导，在每次课后问老师问题或者思路的时候，老师都非常耐心地回答我们，并且还讲很多关于专业以及未来读研或就业等方面的问题，真正地把我们当作自己的孩子一样。助教学姐每次也都非常认真，在课下就和朋友一样与我们相处，在小组作业的完成过程中也帮助了我们很多，再次表示感谢。

七、参考文献

- [1] 高爽, 李艳玲, 宫毅. 企业销售与分销商务智能系统的设计与实现[J]. 软件工程, 2020, 23(12): 54-56.DOI: 10.19644/j.cnki.issn2096-1472.2020.12.016;
- [2] 马小民. 中国联通甘肃分公司商务智能系统集成优化研究[D].兰州大学, 2017;
- [3] 萧文龙, 王镇豪, 陈豪, 徐瑀婧.国内外商务智能及大数据分析研究动态和发展趋势分析[J].科技与经济, 2020, 33(06):66-70.DOI:10.14059/j.cnki.cn32-1276n.2020.06.014.
- [4] 李博.APRIORI 数据挖掘算法在商务智能中的应用[J].电脑迷, 2018(07):155-156.
- [5] 杨斌, 纪东升.商务智能系统中客户行为数据挖掘研究综述[J].甘肃科技, 2012, 28(18):23-26.

附录三：郭思琪课程学习报告

0 前言

BI 所以合变也，以成表里之数，杂取成商之实，望其谋而事其治，教其谋而谋之，进其力而谋之，涉其谋而治思也。欲业要术，使物反利，智能之用，在于胜事。昔 IBM 之用，公会数，析本；今智能无涉不通之域，官治之和，教有司，俱为智能所被。

商务智能者，非空中楼阁，犹有后业之助也。摄于智能之学，练于工用之功，析于未就之数，揄咨之职，十有补焉。斯则用数言者之世，亦因数竞之世，感世之五十强企业，十有八九，皆立数为司。彼国愈众俱信数成智利，数虑之裁，方成日倚之术。数交俱动，便掇取新。较前统数条，互联网之数分内自入临非数匮，乃立过馀。是故互联网之数分，师欲不绝于立术，造罢之。

为数分者，一须有孚。是故必得其业，便其业，公于其流，贵其独见。如离其业，便为脱缕风筝，无大者；二须通判，立数裁架，披数析义，计度而应于萧、管，不闲于理，则难施检架，后数亦难施矣；三须通论，指之以数，析之以数，随之以事，因之以展数，凡策之道云云，如察法分类、察数分构、察法漏辨、论数分方，大说之法亦为其用，曰披法、曰归法、曰类、曰类、曰法、曰主成分、曰时；四须解器，领分掌其事具也。夫数者，道之常也。法者，法之用也。必以渐大，不可以术割，必因强分，并用成数；五须知设，知设为者，以表效数而师析之，则目肆矣，图者门大学问，如图、牒、色配者，所以得其道也。

所长数分者，必得九能。统计为基石，可以助吾以科学，渐近于真知，率而布之、置信区畔、假令参验、相关与归分，归之陈谈。如欲更甚者，请得其本法，如决策树、神经网络之属。再深入则可习深习、图象等法。可视化主由编程、非编程二器通行，近岁佳具，如 TB、QLK 皆极言可观。精技在库，数分之学必贵之，主以 Mysql 通制也。数知其事至重，则数为之，而数有集，全府库可定，高从其数，内成其数，而外能全其性也。立掘多项目计课心应比，须执谴傅之、欣乐就迁。智能与掘算法于数处亦要。此外，百务所须，不相语殊。以昏涩之计繁数，R 愈优。若跨 gpu 行 NLP 或神经网络之术，则 Python 为甚善之选；如欲就固之功、面成之数，又贵操作、资刀斧，则 Java 或 Scala 之固好择。

幸从四福老师顾视知化工信程。文明也，因地而造生，度其所近。文者，谓殊方之士，独习于长生。文以殊方异众之别，而文明以述性，别人物之祖。长河开历，文明有新。先时文明，民务采猎；及机杼之兴，大较之为朴也；至当涂之时，人于网络取上直。夫文明之变，总出西土三工三潮，缘人之擢物也。三共于蒸，演于电气，穷于信息，大益生产之术，尽变商务之文，而防遏沟渎，表里之致。

物之所通，天财之所生也。以数电为效，言科技彭勃之年。信息科技，深变万端。观夫往时之息行，已历两潮，观于台、匠、用而见其章，我方居三浪之始，夫美国险起贸易战，

压华才之实背，5G 通讯、智能终端等科技风口之擅断，是弃诸国市律及次第、霸权主义、强权之体现也。什有八会，公曰：纲施强国之谋，共目之令，得民之道。夫总此者，所以为法也。行此者，所以为存也。视余者，所以为速也。视货财而为新也，行术者，所以尽数而用之，竭智而用之。

故能定开阖，以成形势。是故抑本者之产，经者之地，经历天下之利，以劝开新。资新者之态以培新产，合网甚高，两策之会也。须谋“跨界、融合、创新”，此乃“互联网+”之至键。联网越交，为生之要也。罔集之音技，靡而用之，使其信不徒他业也。目有所积，意有所归，望有所归，毕有所归，目有所舍，形有所归，望有所归。望此者，毕也，吾将示之以息也。十者之本，思形之用。凡三信立浪，国之要络，可谓遭其时矣。

夫素信与工融，党中央、务院之大要也；时为强国、构产、长，深合势、新攻。当今之世，百年无有，海内深变。新也者，速于使名。由总而观之，是两化深融，当由名要之会，方步深化，宜速创意，表急变之速。开塞深融，出新息术，于未形、全规、毕功、损益众用。新动数产，总为强国。深开党中央国务院，深化新代之数，与造业融博之规，按工业及信息化部十四五，规其本末，为之《规划》。故规为新古之位，接于未和之基，合于数解之统，合于两化深融之机，为举错足之计，计合于今，焦合于重，而离于统，合方制之利，明于五年之举，而修于未举。其于为数驱节、为义、为台支、为事增、为智能为主，进为强国进数之要。

构台为新，构深之物也。今海内通行者盖承云以奉其事，及图厥成，不副厥深。《规划》其言，庶作其欢。合网逐台，周而成务。开其数，引其深，参其成。疾极台之数，化柔功之复，养得行，极得台之高，甚得台之式，进得台之心。凡为工业，连衡建数，引企计，审其方，治其功，建典型，以成用。凡百工联衡，治财用金，利窦连功，疾措意造数，攒化台化，为本于工台之产。

昔 ERP 系统，略失效率；传统 DSS，稍逊风骚。譬之他统，则智能优势显著。夫以交易之统，通变之绪，靡所与定，各由其序，与其初则，变而不改。而商务智能尽于商者也。知能化以治新，则不足以塞其求。以术言之，则法、法、类、元、告、柔，皆为变者也。智能之致战也，在于虑统。条贯所随者，新交也，无穷之数也。万物之所终始，靡然多故，而少选之所终始。故吾将以务智能于工台之铸。高能条贯，能为干纪。均之阴阳，备之使人。然亦何事于繁末之中而治末之迹矣。因果关通，梳理型思，对宰艺师。上有观，下有谋，可以同业。务司之通方及领力甚高，所请如此，唯与知终始，统纪治会，计士两修并力，才堪产业、家国之升擢。

壬寅之春，幸得恩师循循善诱，发蒙启蔽，苦心孤诣，鱼渔双授。修身求知，师以身作则，行端表正，不言之教，桃下之蹊。何以述师教育之功？艟艨巨舰，非桨舵导引之助不能乘风破浪；北溟鲲鹏，非长风托举之力不能奋翼九天。恩长笔短，伏惟珍摄，节劳为盼！

1 课程认知方面

商务智能是融合了先进信息技术与创新管理理念的结合体，集成了企业内外的数据，进行加工并从中提取能够创造商业价值的信息，面向企业战略并服务于管理层、业务层，指导企业经营决策，提升企业竞争力，涉及企业战略、管理思想、业务整合和技术体系等层面，促进信息到知识再到利润的改变，从而实现更好的绩效。事实上，商务智能应用的核心不在其功能，而在于对业务的优化，IBM 公司更强调数据集成和数据分析基础上的业务分析和优化。目前，商务智能的应用已延伸到了非商业领域，政府和教育部门等也成为商务智能的应用领域。

商务智能这门学问不是虚无缥缈的理论，通过这门课程的学习，可以掌握许多实用的工作技能，对未来就业十分有帮助。这是一个用数据说话的时代，也是一个依靠数据竞争的时代。世界 500 强企业中，有 90%以上都建立了数据分析部门。IBM、微软、Google 等知名公司都积极投资数据业务，建立数据部门，培养数据分析团队。各国政府和越来越多的企业意识到数据和信息已经成为企业的智力资产和资源，数据的分析和处理能力正在成为日益倚重的技术手段。互联网本身具有数字化和互动性的特征，这种属性特征给数据搜集、整理、研究带来了革命性的突破。与传统的数据分析师相比，互联网时代的数据分析师面临的不是数据匮乏，而是数据过剩。因此，互联网时代的数据分析师必须学会借助技术手段进行高效的数据处理。更为重要的是，互联网时代的数据分析师要不断在数据研究的方法上进行创新和突破。

2 职业发展方面

担任数据分析师，一要懂业务。从事数据分析工作的前提就会需要懂业务，即熟悉行业知识、公司业务及流程，最好有自己独到的见解，若脱离行业认知和公司业务背景，分析的结果只会是脱了线的风筝，没有太大的使用价值。二要懂管理。一方面是搭建数据分析框架的要求，比如确定分析思路就需要用到营销、管理等理论知识来指导，如果不熟悉管理理论，就很难搭建数据分析的框架，后续的数据分析也很难进行。另一方面的作用是针对数据分析结论提出有指导意义的分析建议。三要懂分析。指掌握数据分析基本原理与一些有效的数据分析方法，并能灵活运用到实践工作中，以便有效的开展数据分析。基本的分析方法有：对比分析法、分组分析法、交叉分析法、结构分析法、漏斗图分析法、综合评价分析法、因素分析法、矩阵关联分析法等。高级的分析方法有：相关分析法、回归分析法、聚类分析法、判别分析法、主成分分析法、因子分析法、对应分析法、时间序列等。四要懂工具。指掌握数据分析相关的常用工具。数据分析方法是理论，而数据分析工具就是实现数据分析方法理论的工具，面对越来越庞大的数据，我们不能依靠计算器进行分析，必须依靠强大的数据分析工具帮我们完成数据分析工作。五要懂设计。懂设计是指运用图表有效表达数据分析师的分析观点，使分析结果一目了然。图表的设计是门大学问，如图形的选择、版式的设计、颜

色的搭配等等，都需要掌握一定的设计原则。

一个合格的、高级的大数据分析师必须要掌握以下 9 种技能。统计学是数据分析的基石，可以帮助我们以更科学的角度看待数据，逐步接近这个数据背后的“真相”，包括概率分布、置信区间、假设检验、相关性与回归分析等基本理论。如果想要更进一步，请掌握一些主流算法的原理，比如决策树、神经网络等。再深入一点，还可以掌握文本分析、深度学习、图像识别等相关的算法。数据可视化主要通过编程和非编程两类工具实现，近几年冒出来的优秀工具，如 TB、qlk 都强调可视化。近几年冒出来的 BI 之秀，如 TB、qlk 都强调可视化。sql 在数据库里是核心技术，在数据分析学习时一定要重视这些内容，主要以 MySQL 为主，MySQL 就是互联网行业的通用标准。数据分析中的工作最重要的就是数据处理工作，而数据仓库具有集成、稳定、高质量等特点，基于数据仓库为数据分析提供数据，往往能够更加保证数据质量和数据完整性。数据挖掘是大多数项目计划的核心应用程序，需要掌握 matlab、python 与 R。人工智能与挖掘算法对于数据分析来说也很重要。此外，对于各种需求的实现，不同编程语言具有不同的优势。如果对晦涩的统计运算进行繁重的数据分析工作，R 更有优势。如果跨 GPU 进行 NLP 或密集的神经网络处理，那么 Python 是很好的选择。如果想要一种加固的、面向生产环境的数据流解决方案，又拥有所有重要的操作工具，Java 或 Scala 绝对是出色的选择。

3 宏观认知方面

有幸跟随李四福老师回顾领悟工业化与信息化历程、深刻认识我国战略机遇。在人类历史的长河里，文明阶段不断更迭，在早期文明阶段，人类的生产方式主要是采集和狩猎；随着大机器生产的兴起，现代化的工业基础建设与原始资本的积累初步完成。而到了当今时代，人们通过网络来获取高级直接经济。文明的演变主要源于西方的三次工业革命与信息化浪潮，其演变机理在于人类发展过程中对自然万物认知的提升和积累，。三次工业革命将人类由蒸汽时代送到电气时代再过渡到信息时代，极大地提高了社会生产力，彻底改变了商务活动的形式与内容，更潜移默化地影响着人类日常生活更防范的领域，对世界格局的演变产生了广泛深远的影响。

继物质与能源之后，信息成为人类社会生存和发展的第三大战略资源。以数字电子计算机发明为标志，信息科技蓬勃发展了 70 年。信息科技及其在社会经济生活方方面面的应用广泛并深刻地影响和改变了人类社会。回顾过去的信息化进程，已经经历了两次大的浪潮，从技术平台、管理资源和应用模式等方面看，信息化在不断演化并呈现出明显的阶段性特征，当前，我们正处于信息化建设第三次浪潮的起始期，面对美国恶意挑起贸易战，打压华为的现实，我国坚决反对这种无视国际贸易规则与秩序、霸权主义、强权政治的行为。党的十八届五中全会公报指出：“实施网络强国战略，实施‘互联网+’行动计划，发展分享经济，实施国家大数据战略。”“实施‘互联网+’行动计划”，侧重于在各行各业中充分利用互

联网技术，全面推进社会和经济的发展、转型；“发展分享经济”强调促进因信息快速便捷的流通，有效分配物理世界的各类资源，从而产生的新经济形态；“实施国家大数据战略”则充分利用数据，萃取知识，产生效益，带动一个庞大的数据产业，提升经济活力、社会生产力和国家治理能力。

我国确定了实施创新驱动发展战略，实现转型发展的战略目标，调结构、稳增长成为一项重要的任务。一方面需要抑制传统低端产业，通过实施“一带一路”战略实现国际产能合作、共赢发展，同时也要创造新的就业机会，鼓励大众创业、万众创新；另一方面，需要发展新经济、新业态，“互联网+”正是这两方面战略的结合。

实施“互联网+”战略，“跨界、融合、创新”是三个重要的关键词。互联网技术及思想与传统产业的跨界融合，是培育产业发展新生态的重要途径，使得信息技术不仅是其他行业/产业的催化剂和倍增器，而且在越来越多的行业/产业出现颠覆式创新，成为其颠覆者。就这个意义而言，“互联网+”将是我国信息化3.0建设的基础设施、思维模式和实施指南。面对大数据驱动的第三次信息化建设浪潮，我国的“网络强国”战略可以说是“恰逢其时”了。

持续深化信息化与工业化融合发展，是党中央、国务院做出的重大战略部署，是新发展阶段制造业数字化、网络化、智能化发展的必由之路，是数字时代建设制造强国、网络强国和数字中国的扣合点。而工业互联网平台是新一代信息技术与制造业深度融合的产物。目前，国内不少通用型工业互联网平台提供的是接入上云服务，在解决企业生产经营实际问题时，不能满足其深层次需求为此，《规划》提出，全面激发企业融合发展活力。发展专业化工业互联网平台模式，围绕具体行业打造若干细分领域的工业互联网平台，构建产业生态。为实现数字化目标，组织开展活动，引导各界深度参与到平台发展中。加快发展基于平台的数字化软件工具和工业APP，培育和推广“平台+产品”“平台+模式”“平台+行业/区域”等解决方案，提升平台服务水平。同时，编制发布工业互联网平台发展指数和数据地图，引导企业通过评价结果明确平台应用水平提升方向，提炼工业互联网平台应用优秀成果，树立典型标杆，持续引导平台深化应用。工业互联网平台建设及推广是一项系统工程，要加强产学研用金合作，开展关键技术产品联合攻关，加快研发设计、生产制造、经营管理以及物流等社会资源的数字化改造、在线平台化共享，打造基于工业互联网平台的产业，发展新生态。

4 BI 应用与对比

因此，我们要重视商务智能在工业信息平台建设中的作用。高性能的商务智能系统能够为企业管理会计师提高分析的全面性、及时性和准确性提供必备条件，然而，如何在复杂的商业环境中把握绩效指标的因果联系，梳理出分析模型和思路，对管理会计师对业务的深入洞察，高级别的逻辑思考能力，同业务部门之间的沟通技巧以及领导力等方面均有很高的要求。只有在商务智能与分析系统建设和管理会计人才培养两方面同时做出努力，才能有效提

升企业的数字化分析能力。

4.1 BI 优势

同其他系统相比，商务智能具有显著优势。以交易系统为例，交易系统把交易强加于业务之上，不管谁来进行一项业务，都得遵循同样的程序和规则，而且一旦一个交易系统设计出来以后，轻易不会改变。而商务智能则能适用商务，因为商务智能是一个学习分析型系统，能适应商务的不断变化。若商务智能不能变化以解决新的问题，就不能满足商务的需求。从技术的角度讲，商务智能中变化的是数据、数据类型、元数据、报告和应用软件。商务智能的真正挑战就在于设计和管理一个总在变化的系统。两者所管理的数据类型不同。交易系统跟踪的是最近的交易情况，保留极其有限的历史数据。而商务智能系统维持来自多个交易系统的、多年的交易情况，且数据量很大。

4.2 BI 同传统报表对比

传统的报表系统和商务智能存在着本质的区别。传统的业务报表系统一般被设计成扁平系统，主要是针对分离的事物处理，但对结构化的分析和统计却无能为力。一个独立的商务智能系统，能够从多种异构的应用系统中获取各类业务数据，并通过数学模型建立多层次的分析系统，最终将其转化为具有一定商业意义的信息。商务智能的应用需求往往复杂多变，而且它的实施过程的复杂性也要远远超过传统的报表系统。所以在进行商务智能系统的实施过程中，绝不能受传统事物处理系统思维模式的影响和制约。商务智能和传统的报表系统在应用对象及目的上也是有区别的。一般而言，商务智能更加关注企业长期的战略决策，甚至更侧重于商业趋势和业务单元的联系；而传统的报表系统则注重企业的短期运作支持，更加强调的是具体的数据和精确度。

4.3 BI 同 ERP 对比

商务智能与 ERP 的共性就是使企业运行效率更高、响应更及时及易于整合。从基础架构的角度上看，二者有以下几点相似之处：都是采用分布式结构存储海量数据，都能为大范围终端用户提供深度访问的能力，都具有高度的分布性和应用程序的可扩展性，都是利用直接或者间接数据作为预测工作的信息参考。尽管二者之间存在许多共同之处，但绝不是同一个事物或是同一个事物体的两个方面，而是互补的系统。因而，两者之间也存在以下区别。都是基于现代信息技术进行商业判断，只是其功能特点各有不同，分别侧重于商务智能与业绩跟踪。通过整合，ERP 系统涉及的所有业务流程得到了充分的协调，从而打破了原有的部分分割局面。不仅企业内部所有环节的信息获知能力都得到了提升，打破了企业内外的业务处理瓶颈，其响应速度也得到了极大的改善。商务智能使得用户在一些关键领域的信息获取能力和掌控精度得到了极大的提高，主要表现在以下几个方面：首先，极大程度地改良了报告的格式，通过整合用户数据使报告进行得更快、更及时、更精确；其次，信息传输也更加实时化，极大地缩短了信息在企业内部各部门之间周转的时间；最后，能够及时发现业务处理流程中可能出现的问题及错漏，能准确迅速地实施纠错。通过商务智能，原先分散、孤立的

企业数据按历史记录顺序彼此相关了，而且能按高效、易于提取的结构进行存储。

4.4 BI 同 DSS 对比

作为一种新型的决策支持系统，与传统的决策支持系统相比，商务智能在很多方面都存在显著的优点。在使用对象上。传统决策支持系统仅局限于企业的高层决策者、分析人员，而商务智能的使用对象扩展到企业组织内外的各类人员，为他们提供决策支持服务，既包括企业的领导、企业内部各部门的职能人员，也有客户、供应商、合作伙伴等企业外部用户。在具有的功能上。与传统的决策支持系统相比，商务智能具有传统的决策支持系统所不具备的功能强大的数据管理、数据分析与知识发现能力。在知识库状态方面。在建成的决策支持系统系统中预先设置好知识库是传统的决策支持系统系统的特点，而且知识库中的知识一般很少发生变化，即便是发生变化，也只是采用定期人为更新的方法。但商务智能系统中的知识库是动态变化的，其数据大多是从企业各应用系统中抽取的，且可以对已有的数据仓库或数据集市进行数据挖掘等操作，从而发现新知识，并随时对知识库中的内容进行补充和修正。就实施的目标而言，二者都是为了提高企业决策的效率和准确性，然而，商务智能在一些方面也存在不足之处。利用数据分析、知识发现等工具，商务智能为企业提供了有价值的辅助决策的信息和知识，然后用户再将这些信息和知识与企业的知识和经验相结合进行判断，最后做出明智的决定，其智能决策的能力非常有限，且不具备群体决策能力。

5 核心技术

在课上，老师为我们悉心讲解了数据仓库、在线分析处理、数据挖掘、数据可视化等核心技术以及 R 语言等重要工具，对相关内容总结如下：

5.1 数据仓库与数据库比较

(1) 面向主题。操作型数据库的数据组织面向事务处理任务，各个业务系统之间各自分离，而数据仓库中的数据是按照一定的主题域进行组织。

(2) 集成的。面向事务处理的操作型数据库通常与某些特定的应用相关，数据库之间相互独立，并且往往是异构的。而数据仓库中的数据是在对原有分散的数据库数据抽取、清理的基础上经过系统加工、汇总和整理得到的，必须消除源数据中的不一致性，以保证数据仓库内的信息是关于整个企业的一致的全局信息。

(3) 相对稳定的。操作型数据库中的数据通常实时更新，数据根据需要及时发生变化。数据仓库的数据主要供企业决策分析之用，所涉及的数据操作主要是数据查询，一旦某个数据进入数据仓库以后，一般情况下将被长期保留，也就是数据仓库中一般有大量的查询操作，但修改和删除操作很少，通常只需要定期的加载、刷新。

(4) 反映历史变化。操作型数据库主要关心当前某一个时间段内的数据，而数据仓库中的数据通常包含历史信息，系统记录了企业从过去某一时间点(如开始应用数据仓库的时点)到目前的各个阶段的信息，通过这些信息，可以对企业的发展历程和未来趋势做出定量分析。

和预测

5.2 OLAP 技术

5.2.1 OLAP 的定义

在线分析处理是使管理人员能够从多种角度对从原始数据中转化出来的、能够真正为用户所理解的并真实反映业务维特性的信息进行快速、一致和交互的存取，从而获得对数据更深入的理解。

5.2.2 OLAP 与 OLTP 的区别

OLAP 与 OLTP 有较大的区别。OLAP 是数据仓库系统的主要应用，支持复杂的分析操作，侧重决策支持，并且提供直观易懂的查询结果；OLTP 是传统的关系型数据库的主要应用，主要是基本的、日常的事务处理，例如银行交易。OLAP 是决策人员和高层管理人员对数据仓库进行信息分析处理，而 OLTP 是操作人员和低层管理人员利用计算机网络对数据库中的数据进行查询、增加、删除和修改等操作，以完成事务处理工作。OLTP 和 OLAP 的不同，主要通过以下 5 点区分开来。

(1) 用户和系统的面向性：OLTP 是面向顾客的，用于事务和查询处理；OLAP 是面向市场的，用于数据分析。

(2) 数据内容：OLTP 系统管理当前数据；OLAP 系统管理大量历史数据，提供汇总和聚集机制。

(3) 数据库设计：OLTP 采用实体-联系（E-R）模型和面向应用的数据库设计；OLAP 采用星状或雪片模式和面向主题的数据库设计

(4) 视图：OLTP 主要关注一个企业或部门内部的当前数据，不涉及历史数据或不同组织的数据；OLAP 则相反。

(5) 访问模式：OLTP 系统的访问主要由短的原子事务组成，这种系统需要并行和恢复机制；OLAP 系统的访问大部分是只读操作。

二者及其数据对比如下：

	OLTP	OLAP
用户	操作人员、低层管理人员	决策人员、高层管理人员
功能	日常操作处理	分析决策
DB 设计	面向应用	面向主题
数据	当前的，最新的，细节的，二维的，分立的	历史的，聚集的，多维的，集成的，统一的
存取	读/写数十条记录	读上百万条记录

工作单 位	简单的事务	复杂的查询
	上千个	上百万个
DB 大 小	100MB-GB	100GB-TB

OLTP 数据	OLAP 数据
原始数据	导出数据
细节性数据	综合性数据
当前值数据	历史数据
可更新	周期性刷新
单次处理量小	单次处理量大
面向应用，数据驱动	面向分析，分析驱动
支持日常操作	支持管理需求

5.2.3 OLAP 多维数据分析操作类型

1. 钻取
2. 上卷：钻取的逆操作，即从细粒度数据向高层的聚合
3. 切片：选择维中特定的值进行分析
4. 切块：选择维中特定区间的数据或者某批特定值进行分析
5. 旋转：即维的位置的互换，就像是二维表的行列转换

5.3 可视化工具

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。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。这意味着面对一大堆杂乱的数据，你无法嗅觉其中的关系，但通过可视化的数据呈现，你能很清晰地发觉其中的价值。目前，已经有很多数据可视化工具可以满足各种可视化需求。主要包括用于日常办公的 Excel，信息图表工具 GoogleChartAPI、D3.js、Folt、Echarts、Raphael，地图工具 ModestMaps、Leaflet、PolyMaps、OpenLayers、Kartograph、Google Fusion Tables、QuatumGIS，时间线工具 Timetoast、Xtimeline、Timeslide、Dipity，以及高级分析工具 Precessing、NodeBox、R、Weka 和 Gephi 等。

在以往课程学习中，我最经常使用 python、origin 进行绘图，在商务智能的课程学习中，我学会了使用 R 语言进行关联规则挖掘、主成分分析、聚类分析、熵权法分析及相应图

表绘制，并使用 R 语言实战、Rgallery 网站等资源自学各类描述统计图表的绘制。通过对 R、Python 进行对比，我发现二者有如下不同：

1. 适用场景

R 适用于数据分析任务需要独立计算或单个服务器的应用场景。Python 作为一种粘合剂语言，在数据分析任务中需要与 Web 应用程序集成或者当一条统计代码需要插入到生产数据库中时，使用 Python 更好。

2. 任务

在进行探索性统计分析时，R 胜出。它非常适合初学者，统计模型仅需几行代码即可实现。Python 作为一个完整而强大的编程语言，是部署用于生产使用的算法的有力工具。

3. 数据处理能力

有了大量针对专业程序员以及非专业程序员的软件包和库的支持，不管是执行统计测试还是创建机器学习模型，R 语言都得心应手。Python 最初在数据分析方面不是特别擅长，但随着 NumPy、Pandas 以及其他扩展库的推出，它已经逐渐在数据分析领域获得了广泛的应用。

4. 开发环境

对于 R 语言，需要使用 RStudio 或者给 jupyter 安装 R 核心。对于 Python，有很多 PythonIDE 可供选择，其中 Spyder 和 IPython Notebook 是最受欢迎。

6 案例感悟

除理论、工具讲解外，李四福老师为我们讲解了煤炭矿井决策系统、Target 零售、体测动态权重、体测面板研究等多个案例。其中，令我感触最深的就是老师讲解过的煤炭矿井重点工程网络计划决策系统，这个案例解决了我对未来职业前景的一些困惑：信管人毕业之后能够在政企部门担任什么角色？发挥什么作用？如何在自己不熟悉的领域，参与多学科交叉项目？我们应如何应用所学的系统设计知识、数学模型知识以及 ERP、供应链知识为政企组织谋发展？

我们知道只由少数几项工作组成的任务其安排是否合理，凭经验或进行简单分析是可以解决的。但对于大型的工程项目，其生产活动错综复杂、工序繁多，如何最合理地组织好生产，使生产中各个环节互相密切配合，协调一致，使任务完成得既好又快且省，这就不是单凭经验或稍加分析所能解决的。必须要有科学的组织和严密的计划，对生产上出现的不平衡情况，要及时通过信息进行周密预测、调整和处理，才能保证生产的连续进行和充分有效地利用现有人力、物力、财力，以取得良好的经济效果。

网络计划技术作为一种科学的决策方法，适用于大型工程项目的计划、组织、监控过程，而且越是复杂的、多头绪的、时间紧迫的任务运用网络分析技术就越能取得较大的经济效益。但是，应用网络计划技术进行工程项目管理时，现有的项目管理软件还没能有效解决一些技术应用问题，面临诸多技术“瓶颈”，导致这一优秀的管理方法无法得到有效应用。

在听完老师在课堂上的介绍后，我到学校图书馆借阅了《煤炭矿井工程网络计划决策系统》这本书籍，详细了解了这篇案例，并结合书中的一些模型与设计思路，对我另一门课程——决策支持系统的系统开发作业进行了打磨。我们的选题是企业碳资产管理系统，具体优化细节如下：仿照煤矿系统工程信息和工序信息的信息化管理以及工程进度的实时查询与控制功能，提高了系统开发作业的数据共享水平；仿照煤矿系统的网络计划图的自动绘制功能，完善系统开发作业的可视化功能；仿照煤矿系统的网络计划优化功能，完善系统开发作业的动态控制功能。

7 实践感悟

纸上得来终觉浅，绝知此事要躬行。我对课程所学知识的实践包括如下几方面：

课堂作业方面，采用 Apriori 关联规则方法挖掘超市购物车数据并绘制相关图形；

小组作业方面，制定体测成绩描述统计思路并使用 R 语言实战、Rgallery 网站等资源自学山脊图、弦图等描述统计图表的绘制，对描述统计思路进行实现，同时参与数据预处理、推断统计部分工作。

课外实践方面，在市场调研大赛中使用 R 语言进行 ridit 检验，在决策支持系统课程学习中采用商务智能思想与方法进行设计与开发。

在上述实践过程中，我深深感受到 R 语言的优美，其作为脚本语言凭借其良好的互动性和丰富的扩展包资源可以方便地解决大部分数据处理、变换、统计分析、可视化的问题，并可以重现所有的细节。此外，除了技术上的硬性能力，数据敏感力、逻辑思维能力、归纳能力、批判性思维能力、交流沟通能力、责任心这些软性的技能也是优秀分析师必须具备的素质。另外，如果分析师能站在更高的角度思考问题，有管理者的思维，则能在众多分析师中能脱颖而出。成为优秀的数据分析师需要具备过硬的业务素养和技术能力，这绝非一朝一夕之功，需要在实践中不断成长和升华。道阻且长，以此自勉。

附录四：袁晴课程学习报告

商务智能学习感悟

(袁晴 20191001494)

一、课前引导

在学期初，由于武汉疫情，我们持续上了好几周网课。我清晰得记得李四福老师给我们上的第一节课，李老师给我们讲从古代到近代和现代的中国工业史，从中我加深了对三次科技革命的了解，也更加明白科技和智能对我们现在生活的重要意义。

我了解到 18 世纪由于英国资产阶级统治一方面积极发展海外贸易，进行殖民统治，积累了丰富的资本，扩展了广阔的海外市场和最廉价的原料产地，另一方面，进一步推行"圈地运动"，获得了大量的廉价劳动力，蓬勃发展的工场手工业，积累了丰富的生产技术知识，增加了产量，但还是无法满足不断扩大市场需要，因此，人类历史上一场生产手段的革命呼之欲出。18 世纪 60 年代，在英国的资本主义生产中，大机器生产开始取代工厂手工业，最终生产力得到突飞猛进的发展，人类历史上，把这一过程称为"工业革命"。第一次工业革命是技术发展史上的一次巨大革命，它密切加强了世界各地之间的联系，改变了世界的面貌，最终确立了资产阶级对世界的统治地位，率先完成了工业革命的英国，很快成为世界霸主。它也开创了以机器代替手工劳动的时代，这不仅是一次技术改革，更是一场深刻的社会变革，推动了经济领域、政治领域、思想领域、世界市场等诸多方面的变革。

19 世纪中期，欧洲国家和美国、日本的资产阶级革命或改革的完成，促进了经济的发展。19 世纪 60 年代后期，开始第二次工业革命，人类由此进入了"电气时代"。第二次工业革命尤其以电器的广泛应用最为显著：比如 19 世纪六七十年代开始，出现了一系列的重大发明。1866 年，德国西门子制成了发电机；到 70 年代，实际可用的发电机问世。由此电器开始用于代替机器，成为补充和取代以蒸汽机为动力的新能源。随后，电灯、电车、电影放映机相继问世。第二次工业革命极大地推动了生产力的发展要求，对人类社会的经济、政治、文化、军事，科技、和生产力产生了深远的影响，使社会面貌发生翻天覆地的变化，形成西方先进、东方落后的局面，资本主义逐步建立起对世界的统治，世界逐渐成为一个整体。第二次工业革命进一步增强了人们的生产能力，交通更加便利快捷，改变了人们的生活方式，扩大了人们的活动范围，加强了人与人之间的交流。

而"第三次科技革命"，是从 20 世纪四五十年代，开始的新科学技术革命，以原子能技术、航天技术、电子计算机技术的应用为代表，还包括人工合成材料、分子生物学和遗传工程等高新技术。这次科技革命被称为"第三次科技革命"。值得一提的是，第三次科技革命的出现，既是由于科学理论出现重大突破，一定的物质、技术基础的形成，也是由于社会发展的需要，特别是第二次世界大战期间和第二次世界大战后，各国对高科技迫切需要的结果。

第三次科技革命是人类文明史上继蒸汽技术革命和电力技术革命之后科技领域里的又一次重大飞跃。第三次科技革命以原子能、电子计算机、空间技术和生物工程的发明和应用为主要标志，涉及信息技术、新能源技术、新材料技术、生物技术、空间技术和海洋技术等诸多领域的一场信息控制技术革命。第三次科技革命不仅极大地推动了人类社会经济、政治、文化领域的变革，而且也影响了人类生活方式和思维方式，随着科技的不断进步，人类的衣、食、住、行、用等日常生活的各个方面也在发生了重大的变革。

而当今，在第四次工业革命的驱动下，快速发展的社会越来越需要智能化，以人工智能为基础的创新实现了技术创新机理的重大转变，具有重要的基础创新价值。人工智能的基本理念形成于 20 世纪中叶，在经历过前两轮重要的发展阶段之后，在过去 10 年内进一步接近于发现推动人类走向智能时代的“密码”。人工智能通过对人类大脑的解构与模拟，更加精确化的深度学习方法的研究与迭代，尤其是相关技术与互联网和大数据的进一步融合，使得当前人工智能已经在部分领域实现了对人类自身的模仿甚至超越。第四次工业革命的技术创新与扩散的速度非常快，新兴技术和各领域创新成果传播的速度和广度要远远超过前几次革命。具体到智能领域，相关技术创新活动已显示出极强的“颠覆性”特征。人工智能带来技术创新的同时，对技术创新的核心主体——人才也提出了更高要求。人工智能的发展“绝不意味着超级计算机会彻底取代人脑”，但却呼唤人类知识结构的迅速变化尤其是创新人才培养质量的提高。中国在错失了前两次人工智能热潮之后，随着改革开放以来尤其是近年来综合国力和在人工智能领域的突飞猛进，正逐步成为全球人工智能研究和应用的有力竞争者。而面对中国人工智能人才需求旺盛、供给不足、培养缓慢的基本现状，李四福老师鼓励我们要努力培养创新意识，为中国未来科技智能发展做出我们自己的贡献。

二、课程学习

后面的课程中，李老师主要讲解了《商务智能》这门课的主要知识，从中我了解到商务智能的概念和特点。随着世界经济全球化的迅猛发展,生产国际化的趋势不断加强，企业必须能够在瞬息万变的环境下及时做出反应。为了迎接市场的挑战,企业需要对市场有准确的把握，分析顾客的消费趋势，找出企业经营中出现的问题,加强与供应链合作伙伴的关系,挖掘新的商业机会,并能够对未来进行预测。如何充分利用数据资产，挖掘出决策者需要的信息，做出高质量的决策是企业管理者需要考虑的问题。近年来，数据集成、数据分析、大容量数据存储与并行处理等技术不断成熟，成本不断下降，企业各种应用软件积累了大量的数据。这些因素促进了商务智能的发展。商务智能(Business Intelligence, BI)可以将各种数据及时地转换为支持决策的信息和知识，帮助企业管理者了解顾客的需求与消费习惯，预测市场的变化趋势以及行业的整体发展方向，进行有效的决策，从而在竞争中占据有利地位。

商务智能专家王苗在总结了商务智能的众多版本之后给商务智能下的定义：“商务智能是企业利用现代信息技术收集、管理和分析结构化和非结构化的商务数据和信息,创造和积

累商务知识和见解,改善商务决策水平,采取有效的商务行动,完善各种商务流程,提升各方面商务绩效,增强综合竞争力的智慧和能力。利用现代信息技术是这一定义中的关键之一。现代信息技术的发展催生了信息经济和信息社会,在这一新型的经济和社会形态中,信息的爆炸式增长又产生了对能够处理和控制信息的技术的强烈需求,商务智能正是新的信息技术在商务分析中的有效应用。总结上述观点,商务智能是融合了先进信息技术与创新管理理念的结合体,它集成了企业内外的数据,进行加工并从中提取能够创造商业价值的信息,面向企业战略并服务于管理层、业务层,指导企业经营决策,提升企业竞争力,涉及企业战略、管理思想、业务整合和技术体系等层面,促进信息到知识再到利润的转化,从而实现更好的绩效。事实上,商务智能应用的核心不在其功能,而在于对业务的优化,IBM公司更强调数据集成和数据分析基础上的业务分析和优化(Business Analytics and Optimization, BA0)。目前,商务智能的应用已延伸到了非商业领域,政府和教育部门等也成为了商务智能的应用领域。

商务智能是在计算机软硬件、网络决策分析等多种技术成熟的基础上出现的,是通过对数据整理与分析为决策提供依据的一项技术,商务智能技术是运用了数据仓库、OLAP和数据挖掘等技术来处理和分析数据的技术,能够帮助企业进行经营分析、战略支持和绩效管理。数据仓库技术、OLAP、数据挖掘技术是商务智能系统的三大支撑技术,其中数据仓库是商务智能的基础,OLAP与数据挖掘是商务智能系统中的数据分析工具。数据仓库的作用是为系统中的分析工具提供数据基础,OLAP和数据挖掘的作用是要把数据仓库中的数据变成知识,把潜在的知识变成可以为工作所用的知识,帮助我们在业务管理和发展上及时做出正确的判断,为决策者提供问题解决方案以及决策依据。学习了这些理论知识,我对商务智能系统有了更详细和深刻的理解,我也成功将其运用到这学期的计算机三级考试中,并且也会运用到今后的学习和生活中。

在课程学习中,我还有一个印象深刻的章节——关联分析。关联分析用于发现隐藏在大型数据集中令人感兴趣的关联关系,描述数据之间的密切度。其中,支持度和置信度是描述关联规则的两个重要概念。支持度用于衡量关联规则在整个数据集中的统计重要性,简单地说,支持度度量的是在所有行为中规则A, B同时出现的概率。置信度用于衡量关联规则的可信程度,即置信度度量的是出现A的情况下,B出现的概率。

以一道题目为例:根据7位顾客购买的物品回答下列三个问题:(1)计算(苹果,香蕉) \rightarrow 榴莲的支持度和置信度。(2)如果要求最低支持度为0.4,请问(香蕉) \rightarrow 榴莲满足该条件吗?(3)如果要求最低置信度为0.8,请问(苹果,樱桃) \rightarrow 榴莲满足该条件吗?

序号	购买物品
1	苹果, 香蕉, 樱桃, 榴莲
2	苹果, 榴莲
3	香蕉, 榴莲

4	榴莲, 香蕉, 樱桃
5	香蕉, 榴莲
6	苹果, 香蕉
7	苹果, 樱桃, 榴莲

(1) ①Support ((苹果, 香蕉) → 榴莲) = 1/7 = 14.29% (7位顾客中有三位顾客都购买了三件商品) ;

②Confidence ((苹果, 香蕉) → 榴莲) = 1/2 = 50% (同时购买了苹果和香蕉的有两位顾客, 其中一位顾客也购买了榴莲, 所以置信度是 50%) ;

(2) support ((香蕉) → 榴莲) = 4/7 = 57.14% > 0.4, 故满足条件;

(3) confidence ((苹果, 樱桃) → 榴莲) = 2/2 = 100% > 0.8, 故满足条件。

通过李老师的耐心讲解和课堂例题的训练, 我掌握了支持度和置信度的计算和分析, 也将其运用到了计算机三级考试中。

三、课后作业

除了课堂上的学习, 我在小组作业中也学到了很多实用的知识, 更锻炼了自己信息检索能力、r 语言编程能力和 debug 能力。我负责的小组作业模块是描述性统计中的地图、密度图、饼图和雷达图的绘制与描述。

首先在处理数据方面, 由于进行数据预处理的 csv 数据在我的电脑上打开完全乱码, 我就在百度搜索解决方法。我找到了两个方案, 其中方案一是: 创建一个新的 Excel 文件—切换至“数据”菜单—选择数据来源为“自文本”选择 CSV 文件—出现文本导入向导—选择“分隔符号”—文件原始格式选择“65001: Unicode(UTF-8)”—下一步—勾选“逗号”—去掉“Tab 键”—下一步—完成—在“导入数据”对话框里, 直接点确定。经尝试解决乱码失败。方案二是: 使用记事本打开 CSV 文件——点击菜单: 文件-另存为, 编码方式选择 ANSI——保存完毕后, 再用 EXCEL 打开这个文件。方案二更简便快捷, 经尝试解决乱码基本成功, 只是“ID”列仍是乱码, 但接下来的读取数据用不到“ID”列, 所以预处理数据乱码问题解决。

由于以前的课程中未用过 r 语言进行图形的编程, 所以在开始绘制地图的时候, 我要熟悉这个软件并且安装绘图需要的包。在安装包 (install.packages (“包名”)) 和引用包 (library(“包名”)) 后才开始了正式的绘制图形工作。

首先是绘制地图。由于需要将各地区生源数量划分为一定区间, 经规划, 我觉得分为 [0.50], [50,100], [100,200], [200,1000] 和 [1000, +∞], 并从预处理数据中统计出生源数量相关数据, 提取为 r 可读取的 csv 文件。由于 ggplot 的地图绘制程序包绘制的中国地图不完整, 我决定在相关官网下载完整的中国地图素材, 得到 bou2_4p.pdf、bou2_4p.shp、bou2_4pshx 等文件, 读取 bou2_4p.pdf 文件, 并利用代 ggplot(china_map,aes(x=long,y=lat,group=group))

+geom_polygon(fill="white", colour="grey") +coord_map("polyconic")绘制并投影得到可用地图，随后“x <- china_map@data”读取行政信息，“xs <- data.frame(x,id=seq(0:924)-1)”确定地图含岛屿共 925 个形状，“china_map1 <- fortify(china_map)”转化为数据框，“china_map_data <- join(china_map1, xs, type = "full")”合并两个数据框，此时再进行 ggplot 绘图并规定颜色即可得出分区域上色的中国地图。但是此时我发现，地图上没有省份的名称，这非常不利于看图，通过思考如何能在省份区域内出现省份名称和上网查询相关方法，我最终确定了利用地区经纬度确定文字位置的方式，于是我下载并读取盛世经纬度数据 china-cities.csv，再进行 ggplot 画图。本以为可以成功，但此时突然出现“无法分配 82.6M”，我猜想可能内存不足，通过搜索顺利找到解决方案：利用“memory.limit(1000000)”设置约为 1G 内存，然后顺利画出地图。但是我觉得地图区域为随机颜色不美观，于是从颜色表中选出“#A50F15”, "#FEE5D9", "#FB6A4A", "#DE2D26", "#FCAE91”等五种深浅不一的红色来标明数据量的大小。同理做出其他年份的生源地地图。至此，地图已绘图完毕。

其次是绘制密度图。绘制密度图需要 ggplot2 包，进行安装并 library 引用，但在第一步读取的时候就开始报错，经调试发现是由于数据默认将第一行读为具体数据，添加“header=T”后，读取 csv 文件就将第一行用于列名称，具体数据从第二行开始。随后利用“p <- ggplot(data, aes(x=totalscore))+geom_density(color="red", fill="#FCAE91)”画出曲线为“red”、填充为“#FCAE91”的总成绩密度图。密度图有了，但是我想到将平均成绩也体现在密度图上会更好，于是上网查询并添加代码“p+ geom_vline(aes(xintercept=mean(totalscore)), color="blue", linetype="dashed", size=1)”，在已生成的密度图上添加一条数值为平均成绩的蓝色虚线，以更好的做描述和对比。同理做出其他年份、分性别和分年级的密度图。至此，密度图已绘图完毕。

然后是绘制饼图。将需要绘制饼图的各组数据计算出来备用，例如参加 2014 年体测的男女生数量分别是 2835 和 1857，“info = c(2835, 1857)”进行数据赋值，“names = c("男", "女")”将组别进行命名，“cols = c("#FB6A4A", "#FCAE91)”确定填充颜色，之后“pie(info, labels=pielpercent, main = "2014 年体测性别对比", col=cols, family='GB1')”即可汇出饼图。但是没有标出颜色对应的区域，利用“legend("topright", names, cex=0.8, fill=cols)”添加颜色样本标注。此时我又在想，以及显示了大概区域比例，如何显示具体所占百分率呢，我又查询并确定添加“pielpercent = paste(round(100*info/sum(info)), "%)”语句计算百分比。同理做出其他年份和年级的饼图。至此，饼图部分已绘制完毕。

最后是雷达图。绘制雷达图需要“fmsb”包，安装（install.packages("fmsb")）并引用（library(fmsb））。在多次尝试读取文件失败后，我尝试将所需数据提取并计算出来，例如绘制 2017 年男女单项成绩平均分雷达图时，运用代码“data <- data.frame(row.names = c('男', '女'), "BIM" = c(92.56, 96.17), "肺活量" = c(77.60, 77.82), "五十米" = c(73.49, 64.77), "立定跳远" = c(61.20, 66.78), "坐位体前屈" = c(69.71, 73.93), "仰卧起坐" = c(0, 57.51), "引体向上" =

`c(19.48,0), "八百米"= c(0,70.95), "一千米"=c(65.56,0))` 构建数据集，`“max_min <- data.frame(row.names = c("Max", "Min"), "BIM" = c(100,0), "肺活量" = c(100,0), "五十米"= c(100,0), "立定跳远"= c(100,0), "坐位体前屈" = c(100,0), "仰卧起坐"= c(100,0), "引体向上" = c(100,0), "八百米"= c(100,0), "一千米"= c(100,0))` 定义每个变量的范围，随后合并数据集并确定颜色、线段参数，并添加图例得到雷达图。同理绘制 2017 年经管各专业和整体各年级的相关雷达图。至此，雷达图已绘制完毕。

四、小结

经过李四福老师的课前引导、课堂学习和课后作业，我对《商务智能》这门课有了非常深刻的印象和理解，我深知科技智能的重要性，并在今后的学习中努力培养相关意识；我也了解了商务智能是企业利用现代信息技术收集、管理和分析结构化和非结构化的商务数据和信息，创造和积累商务知识和见解，改善商务决策水平，采取有效的商务行动，完善各种商务流程，提升各方面商务绩效，增强综合竞争力的智慧和能力，并且运用了例如 OLAP 等各种技术；同时在完成小组作业的过程中我重新捡起对 r 语言的记忆，并且 r 语言代码编写能力和 debug 能力大大提高。

感谢李四福老师的悉心教学和耐心指导，也感谢助教学姐的帮助。我会将在这门课上学到的知识和能力运用到今后的学习和科研中，也相信会给今后的学习和科研带来帮助，对今后的人生带来积极的影响。

附录五：徐嘉艺课程学习报告

商务智能课程学习心得

徐嘉艺 20191000960

在大三下学期，我修读了李四福老师的商务智能课程，这也是继上学期信息系统分析与设计课程后，我修读李四福老师的第二门课。信息系统分析与设计这门课侧重点在于一个系统如何从用户需求开始，一步一步分析、设计直到系统开发完成并落实。而本学期的商务智能课程，则聚焦于使用现代技术，例如数据仓库、数据分析处理、数据挖掘等，去实现商业价值。在这篇心得里，我将从三个内容阐述本课程令我印象深刻的体悟，其分别为求学之道、数据分析与系统建设。

一 大学之道，求学之路

首先，在学会做学问之前，必先学会做人。而大学之道的内容，也被老师从信息系统分析与设计的课程一直提及到了商务智能课程——这也是使我非常难忘的内容。我很赞同，无论学习任何形式的知识、技能，都必须先贯彻落实如何为人处世。儒学经典传授了许多深刻的人生道理，而“大学之道”则是作为《大学》的开篇第一句，深刻体现了其重要性。大学之道：指穷理、正心、修身、治人的根本原则。“大学”一词在古代有两种含义：一是“博学”的意思；二是相对于小学而言的“大人之学”。大学的宗旨在于弘扬光明正大的品德，学习和应用于生活，使人达到最完善的境界。知道应达到的境界才能够志向坚定；志向坚定才能够镇静不躁；镇静不躁才能够心安理得；心安理得才能够思虑周详；思虑周详才能够有所收获。每样东西都有根本有枝末，每件事情都有开始有终结。明白了这本末始终的道理，就接近事物发展的规律了。

古代那些要想在天下弘扬光明正大品德的人，先要治理好自己的国家；要想治理好自己的国家，先要管理好自己的家庭和家族；要想管理好自己的家庭和家族，先要修养自身的品性；要想修养自身的品性，先要端正自己的心思；要想端正自己的心思，先要使自己的意念真诚；要想使自己的意念真诚，先要探究事物原理。

通过探究事物原理才能获得智慧。获得智慧意念才能真诚；意念真诚后心思才能端正；心思端正后才能修养品性；品性修养后才能管理好家庭和家族；管理好家庭和家族后才能治理好国家；治理好国家后天下才能太平。上自国家元首，下至平民百姓，人人都要以修养品性为根本。若这个根本被扰乱了，家庭、家族、国家、天下要治理好是不可能的。不分轻重缓急，本末倒置却想做好事情，这也同样是不可能的。

因此，在做任何事情之前，必先内心谙熟“大学之道”，悟其根本，察其内在，方能作为个人的求学准则，在茫茫学海中竖起一盏指导自我前进的明灯。

二 数据分析——塔吉特案例

塔吉特是商务智能课程中提到的一个关于数据分析的案例，在此我希望结合塔吉特案例来讲述一些我对数据分析的感悟。塔吉特公司是美国的第二大零售商，一直是零售业的先锋代表。其之所以能做到如此零售业的巨头，与其细致入微的数据分析是离不开的。塔吉特充分收集了来自顾客的有效数据，并通过合理地数据分析，为客户提供更周到合适地服务。我们不妨举一个例子来解释，这个例子就是著名的《比父亲更早知道女儿怀孕》。

曾经有一位男性顾客到一家塔吉特店中投诉，商店竟然给他还在读书的女儿寄婴儿用品的优惠券。这家全美第二大零售商，会搞出如此大的乌龙？但经过这位父亲与女儿进一步沟通，才发现自己女儿真的已经怀孕了。

一家零售商是如何比一位女孩的亲生父亲更早得知其怀孕消息的呢？

每位顾客初次到塔吉特刷卡消费时，都会获得一组顾客识别编号，内含顾客姓名、信用卡卡号及电子邮件等个人资料。日后凡是顾客在塔吉特消费，计算机系统就会自动记录消费内容、时间等信息。再加上从其他管道取得的统计资料，塔吉特便能形成一个庞大数据库，运用于分析顾客喜好与需求。塔吉特的统计师们通过对孕妇的消费习惯进行一次次的测试和数据分析，得出了一些非常有用结论：孕妇在怀孕头三个月过后会购买大量无味的润肤露；有时在头 20 周，孕妇会补充如钙、镁、锌等营养素；许多顾客都会购买肥皂和棉球，但当有女性除了购买洗手液和毛巾以外，还突然开始大量采购无味肥皂和特大包装的棉球时，说明她们的预产期要来了。在塔吉特的数据库资料里，统计师们根据顾客内在需求数据，精准地选出其中的 25 种商品，对这 25 种商品进行同步分析，基本上可以判断出哪些顾客是孕妇，甚至还可以进一步估算出她们的预产期，在最恰当的时候给她们寄去最符合她们需要的优惠券，满足她们最实际的需求。依靠分析消费者数据，塔吉特的年营收从 2002 年的 440 亿美元扩大到 2010 年的 670 亿美元。这家成立于 1961 年的零售商能有今天的成功，数据分析功不可没。

塔吉特和其他公司在技术变革上遇到的问题告诉我们：敏捷开发非常强大，但那远远不够。想要拥有高度有效的数字化组织，公司需要给敏捷开发之路装上“减速带”，软件开发不能一味求快。敏捷开发常伴随以下三个阻碍：刚性架构、人才管理落后、缺乏产品思维。

塔吉特的明显优势在于两个点：

一、打造现代化系统，方便用户体验

在 IT 行业，多年来一直在膨胀的数据库和升级补丁已经让许多公司的技术架构失去了灵活性。在大多数公司，提供给用户的应用软件都是在更科学的设计架构出现前开发完成的，饱受刚性架构之苦，太多的功能都被耦合在一起，如果你想要对其中某一段代码进行变更，产生的影响可能如同雪崩。塔吉特针对这些问题做出了决策，他们将打造现代化系统置于优先地位，公开常用的关键数据，例如物品价格和供应量等，对于运行良好的遗留事物系统则

没有改动。这使得团队能够集中精力优化用户体验，让用户能够更便捷地搜索、兑换物品，而不是把时间浪费在从几个定价系统中选择哪个最准确。

二、重视优秀人才，注重核心技术集中

人才是数字化运营模式中的核心，高管也应该清楚他们需要招募的是能够负担得起的最优秀的人才。然而，过去几年的经验告诉我们，想要招到对的人来提升 IT 企业的敏捷性太难了。塔吉特发现他们采用的传统招聘模式并没有招到适合的技术人才，把公司引领到目标的位置。企业对于技术人才外包的依赖，限制了它对于打造软件工程师社区的远见。说回塔吉特，他们在经历失败之后，选择转向开源技术，对外宣布了公司进行转型并发展数字科技的雄心。塔吉特此前坚持使用自己的专有代码，导致难以聚集足够多优秀的开发人才，因为许多优秀的开发人才都更偏爱在开源数据上工作。塔吉特的这一举措有助于吸引并留住优秀的稀缺技术人才，减少对第三方的依赖性。

三 系统建设——煤炭矿井重点工程

在系统建设方面，我们在课上学习到了关于煤炭矿井重点工程的案例，关于这个案例，我想讲述一些自己对于系统建设方面的心得。

其案例背景如下：PMW 矿建于 1956 年，1958 年底投产，由于 PMW 矿所产煤为气肥煤，而瓦斯含量高，所以煤尘爆炸指数高，长期严重亏损。2002 年底以王矿长为首的新一届领导来到 PMW 矿，开始了实施新的“四驱驱动”经营战略和“四位一体”经营管理体系，PMW 盈利逐渐上升，被评为省级“五优”矿井。为了适应行使其发展的要求，PMW 矿于 2006 年进行了技术改造，为了满足国民经济发展的需要，需新建、扩建大批矿井，从而举办了集团公司网络计划工作会议，会议主要讨论网络计划技术的使用，虽然网络计划技术有利于项目主管对偏离计划轨道的行为及时进行整改，也有利于各部门围绕一个明确的目标紧密配合，客服主观随意性、忙乱、窝工等现象，但是由于技术力量和网络计划技术的掌握程度有限，要在短期绘制出一个规范的网络计划图比较困难，PMW 矿面临许多挑战。后来，由于校企的合作，有了计算机技术和网络计划技术的引进，项目方向日渐明确，网络计划图绘制完成，在研发小组成员卓越成效的努力下，不到半年就完成了整个系统的开发工作。

在此，我主要想论述关于该案例思考的两点：

一、系统建设过程中需要注意的事项

1) 系统要和公司/用户相匹配

很多人认为，直销商系统和公司存在着不可逆转的博弈现象，更有人将系统和公司列到对立的层面。而笔者认为，直销商系统一定要和公司相匹配。直销商系统和公司之间是鱼和水的关系，公司搭建了平台才有直销商发挥的空间。虽然这里面有经销商的文化，但说到底经销商的文化也是从公司的母文化中延伸出来的，切不可与公司的母文化相背离。当下有很

多新起盘的公司，一些领导人加盟新的公司后，自然也要建立团队和系统的文化，这时切不可把以前的文化全盘照搬过去，要知道，每个公司的制度、产品、文化都不同，你以前所在的系统一定是和那个公司相匹配产生的。简单的照搬很可能会出现水土不服的情况。比如级差制的公司延伸出的系统文化和双轨制衍生出来的文化就一定会有不一样的地方。

2) 掌握系统的本质规律

关于系统的认识是众说纷纭。国内外学者给系统所下的定义不下几十个，“仁者见仁，智者见智”，对直销商运营系统的认识也是如此。尽管系统的具体说法有这样那样的差异。但我们必须知道系统的三项普遍的、本质的东西：首先系统具有整体性。就好像汽车，你不能指着某一个部分或零件就说这是车，只有当所有的部分组合起来时才能称上是汽车。脱开整体的观念去谈系统是不完全的。所以在一次培训会结束后，一位朋友问我对系统的看法，我说：每个人都是系统，每个人又都不是系统。所有的工具、会议、人员、培训有关的一切合起来才能称上系统。第二是系统由着相互作用和相互依存的要素所组成，既然是有不同的元素的融合，那么系统就要有极强的包容性。系统最重要的组成部分就是人，而每一个人来系统之前有不同的背景和经历，所以他们有不同的思维，也有不同的智慧，这就意味着需要不断的教育、培养、沟通和磨合方能达到价值观的相同或相近，思维模式和行为模式的统一，这需要一个足够长的时间。第三是系统受环境影响和干扰，和环境相互发生作用。系统是环境的产物，所以在环境改变后，系统也要做出相应的改变。当直销发展到今天，我们不能只一味的以过去的经验来指导今天的市场。事物是发展的，系统也应该随着时代的不同而不断的发展。今天系统的运作要有行业的全局观、发展观。当然，发展不意味着完全的摒弃，而是在继承的基础上发展。

3) 关于简单和复杂的问题

本专题一开始就提到过简单——复杂——简单的过程，注意是复杂到简单，而不是简单到简单。所有的领导者都是将复杂的东西简单化，这个道理大家都懂，但做起来就不是那么容易了。要将复杂的东西简单化，这对领导人有非常高的要求，一个真正的系统建造者，不只是懂得这个生意怎么去建立，还要精通营销、管理、运营、策划、心理、哲学、人性等一系列领域的知识，同时还要有极强的个人魅力、超出常人的坚定、崇高的品格。这就不是一般的直销人士可以做到的。这两年，我们接触过很多所谓的系统，恕我直言，很多“系统”的领导人远没达到这样的水平。很多的一些系统复制的东西其实还停留在简单到简单的阶段，严格意义上讲，只能称之为团队，一些系统开始进入了简单到复杂的过程，但最难的就是复杂到简单。

4) 做执行者还是拥有者

讲到了简单和复杂，我们就要进一步思考这个问题了。现在想自己创建系统的人很多，但问题是，你有多大的能力可以自己创建一个完善的系统呢？在本专题里，我们反复提到的一个观念，当我们看系统时，要从广意上去理解，我们也可以把一个运作成功的公司说成是

一个系统，那么我们来看这样一个统计报告：100 家公司同时成立，五年后只有十家存活下来，这五年考核的是公司经营者的眼光，而十年后只有一家成功经营，成为商业的典范，这五年考核的是企业综合能力：人、财、物、进、销、存、产，任何一个方面出了问题，都会让企业走向失败。那么，建立一个系统也是这样。这两年，我们不断的听到一些系统，但也看到一些所谓优秀的系统很快消失了，留下一批又一批失望的经销商。所以要考虑建立自己系统的朋友要衡量一下自己的实力。做系统的设计者和拥有者固然痛快，但其要承受的压力和所需要的能力也非常高。对于一般的经销商而言，做一个系统的使用者，执行者不失为一个更好的选择，先学习，跟随你的系统，当对系统的理解和掌握到一定程度时，再思考创立系统的事情。

那么，为什么 PMW 矿采用校企合作的开发模式？

因为 PMW 矿缺乏相应的专业人员，而高校正是专业人员的聚集地。选择校企合作的开发模式，不仅能最大限度的节约时间和成本，而且能将网络计划技术应用的更好。还可以选择使用当下商业中正在使用的相关软件还有可供选择的方案——企企合作。其实从长远的发展来看，该矿还是应该采取企企合作。目前的校企合作只是权宜之计，学校老师的本职工作还是教书育人，不可能一直留在那里指导他们，而且一般学校老师虽然有过硬的理论基础，但是他们的实践能力还是比不上企业方，可能他们在考虑成本问题时没有结合市场因素；再者，企业与企业之间是追求利益的共同体，他们之间沟通更加顺畅，而且有共同的目标，由于是业务范围内的事，更加可以全心全意的去实施网络计划技术，这样对 PMW 矿的长远发展来说，更加有利。

四 总结

最后，在商务智能课程中我学到了许多广博的东西，提升了自己的眼界和能力。最后，仍十分感谢李四福老师的教导与助教学姐的帮助！

附录六：袁应安课程学习报告

商务智能学习心得

(袁应安 20191000223)

商务智能(Business Intelligence)是结合商业分析、数据挖掘、数据可视化、数据工具和基础结构，以及最佳实践，可帮助组织更多地利用数据进行决策的一门学科^[1]。通过全面了解数据，利用数据推动变革，消除效率低下的环节并快速适应市场或供应变化，借助 IT 实现服务解决方案是商务智能所关注的问题领域。

本学期在李四福老师的带领下，我们对时空视角下的商务智能发展史、商务智能的理论、方法和技术、现代视角下商务智能的应用场景等课程内容进行了学习，此外老师还拓展了诸如 Target 的巧妙营销、煤炭矿井重点 PERT 决策支持系统等案例，抛出了大数据时代下数据的体量“大”重要还是“数据”的质量、维度等属性更加重要等极具启发意义、对应试思维产生巨大冲击的问题，让我们对于现代商业社会下对数据分析人才的需求与期望有了新的认识。本次学习心得将结合课程所学、案例所感、问题所思，从以下五个方面详细展开。

1 课程学习

1.1 时空视角下的商务智能发展史

在《商务智能》课程中老师带领我们回顾了时空视角下的工业化与信息化历程，人类文明经过了“农业文明——工业文明——信息文明”的演进历程。农业文明时代的核心资源是土地，直接经济的获取方式和来源是日出而作日落而息的辛苦劳作，到了以物产为核心资源的工业文明时代，大机器生产的迂回生产消费经济初步奠定了现代化的工业基础并进行原始资本的积累。而到了以信息为核心资源的信息文明时代，人们通过网络来获取高级直接经济，网络外部性也打破了微观经济学中“边际效用递减”的定律，随着互联网用户的递增而获得指数组级增长的效益。

谈现代商务智能的发展和来源不可避免地提及西方的三次工业革命及信息化的三次浪潮。西方三次工业革命将人类由蒸汽时代送到电气时代再过渡到信息时代，极大地提高了社会生产力，彻底改变了商务活动的形式与内容，更潜移默化地影响着人类日常生活更防范的领域，对经济增长和社会变革以及世界格局的演变产生了广泛深远的影响。回顾信息化的三次浪潮，第一次浪潮的背景是数据孤岛亟需解决，而个人计算机和信息处理技术的发展实现了数据资源的获取和积累，第二次浪潮的时代特点是“人机”二元融合，互联网的普及促进数据资源的流通和汇聚，而第三次浪潮下，云计算、大数据、物联网技术的发展促进了“人机物”的三元融合，在信息爆炸的时代背景下通过多源数据的融合分析呈现信息应用的类人智能，帮助人类更好认知事物和解决问题，庆幸的是中国及时抓住了第三次浪潮，在世纪之交

实现弯道超车赶超诸多曾经领先的欧美发达国家。

站在当今繁荣发达的商业社会的时间节点回溯，我认为农业文明、工业文明与信息文明的本质特征在于人类发展过程中对自然万物认知的提升和积累，从而提升了对各种资源和整合、加工和利用。在智能革命以前，旧的生产关系，是雇佣关系，是一方把另一方当生产资料的关系。而智能革命以后，新的生产关系，将是契约关系，是双方平等、共同合作的关系。

立足当下展望未来，面对美国恶意挑起贸易战，打压华为的现实的背后，是对 5G 通讯领域、智能终端等未来科技风口的垄断，是无视国际贸易规则与秩序、霸权主义、强权政治的体现，这也侧面反映出我国在信息技术、互联网科技领域的卓越成就。

1.2 商务智能的工具、理论、方法与技术

商务智能是通过应用基于现实数据的支持系统来辅助商业决策的制定，其技术包括数据仓库、数据挖掘、数据集成和存储管理、数据分析和建模、联机分析处理等。下面将从商务智能的工具、理论、方法和技术四个角度回顾并总结课程所学。

1.2.1 工具

在老师的介绍以及自主探索后可用于商务智能的工具可总结为 R 语言、Python 语言、Julia 语言、Tableau 以及 SPSS 等。

R 语言是用于统计分析、绘图的语言，是一个用于统计计算和统计制图的优秀工具，基于基础绘图系统，Lattice 绘图系统，ggplot2 绘图系统可以绘制出可解释性强、十分美观的图表，为商业决策提供有力支持。

Python 语言在数据科学、商务智能领域也广受应用，其语法相对简单、完全面向对象，有高可拓展性的绘图库 Matplotlib 以及可用于科学计算的 Numpy、Pandas、Scipy 等库也使其在数据分析领域备受青睐。此外，基于 Python 语言的 TensorFlow、Pytorch|、Keras 等主流深度学习框架也使其在深度学习领域大放异彩，赋能商务智能的实现。

Julia 语言是一种即时编译语言，而非像 Python、R 等类似的脚本解释语言，这种特性适得其反无需浪费时间走一趟解释器就能被编译为可在 CPU 上直接执行的机器代码，因此此计算速度、计算效率有优于 Python、R 甚至是 C 语言。在数据科学、人工智能领域，Julia 因其高效、优雅的语法而大放异彩，因此在大规模、海量商务数据分析时发挥着重要作用。

Tableau 是用于可视分析数据的商业智能工具。用户可以创建和分发交互式和可共享的仪表板，以图形和图表的形式描绘数据的趋势，变化和密度。Tableau 可以连接到文件，关系数据源和大数据源来获取和处理数据。该软件允许数据混合和实时协作，十分适于视觉数据分析。同时其操作更需对商业逻辑、业务流程的理解，这使得商务数据分析人员可将精力集中于理解业务本身无非繁琐的编程的细节。

SPSS 是由 IBM 推出的一系列用于统计学分析计算、数据挖掘、预测分析和决策支持的软件产品及相关服务的总称，它集数据录入、整理、分析功能于一身，包括数据管理、统计

分析、图表分析、输出管理等，操作简单，可为商务分析人员提供很大的帮助。

1.2.2 理论

数据仓库是一种面向商务智能活动（尤其是分析）的数据管理系统，其仅适用于查询和分析，通常涉及大量的历史数据。数据仓库具有以下特点：

- 1、**面向主题：**数据仓库可以高效分析关于特定主题或职能领域（例如销售）的数据。
- 2、**集成：**数据仓库可在不同来源的不同数据类型之间建立一致性。
- 3、**相对稳定：**进入数据仓库后，数据将保持稳定，不会发生改变。
- 4、**反映历史变化：**数据仓库分析着眼于反映历史变化。

一个精心设计的数据仓库支持高速查询、高数据吞吐量，能够凭借出色的灵活性帮助用户细分数据或降低数据量，进而执行更加细致的数据检查，满足高层级和精细化数据管理等各种需求。同时，它还能为中间件 BI 环境（为最终用户提供报告、仪表盘和更多其他界面）提供一个坚实的功能性基础。

当今的数据处理可分为 OLAP (OnLine Analytical Processing, 联机分析处理) 和 OLTP (OnLine Transaction Processing, 联机事务处理) 两大类。OLTP 是传统的关系型数据库的主要应用，主要是基本的、日常的事务处理，例如银行交易。OLAP 是数据仓库系统的主要应用，支持复杂的分析操作，侧重决策支持，并且提供直观易懂的查询结果。OLTP 系统强调数据库内存效率，强调内存各种指标的命令率，强调绑定变量，强调并发操作；OLAP 系统则强调数据分析，强调 SQL 执行市场，强调磁盘 I/O，强调分区等。其具体功能区别如下表所示：

表 1 OLTP 与 OLAP 的分析与比较

	OLTP	OLAP
用户	操作人员、低层管理人员	决策人员、高层管理人员
功能	日常操作处理	分析决策
DB 设计	面向应用	面向主题
数据	当前的、最新细节的、二维的、分立的	历史的、聚集的、多维的、集成的
存取	读/写数十条记录	读上百万条记录
工作单位	简单的事务	复杂的查询
用户数	上千个	上百万个

DB 大小	100MB-GB	100GB-TB
时间要求	具有实时性	对时间的要求不严格
主要应用	数据库	数据仓库

1.2.3 方法

商务智能的主要应用方法有线性回归、决策树、关联分析、聚类等机器学习与数据挖掘方法。课堂上老师着重向我们介绍了 Apriori 算法。

Apriori 是一种常用的数据关联规则挖掘方法，它可以用来找出数据集中频繁出现的数据集合。找出这样的一些频繁集合有利于决策，例如通过找出超市购物车数据的频繁项集，可以更好地设计货架的摆放。它是一种逐层迭代的方法，先找出频繁 1 项集 L1，再利用 L1 找出频繁 2 项集，然后以此类推。

1.2.4 技术

“好图胜千言”，数据可视化技术在商务智能领域也发挥着举足轻重的作用，上文提到的 R、Python 等语言，包括 Tableau 工具都是可视化分析与展示的利器。常用的数据可视化方法包括并行可视化、原位可视化、时序数据可视化等。

1.3 现代视角下商务智能的应用场景

老师在课堂上还向我们拓展了现代视角下商务智能的诸多应用场景，主要包括“数字经济”、“云计算”、“大数据”、“人工智能”、“物联网及其关键技术”、“工业 3.X 与 5G”、“工业 4.0 及其应用场景”、“区块链”、“元宇宙”等当下较为火热的学术概念。其中我想对“区块链”谈谈自己的认识和见解。

区块链的核心关键词是“非对称加密”、“去中心化”和“数字货币”，其技术实现层面为高安全性的哈希函数，通过计算哈希地址来获得工作量证明。区块链作为点对点网络、密码学、共识机制、智能合约等多种技术的集成系统，提供了一种在不可信网络中进行信息与价值传递交换的可信通道，凭借其独有的信任建立机制，与云计算、大数据、人工智能等新技术、新应用交叉创新，融合演进成为新一代网络基础设施，重构数字经济产业生态。但是由于资本的过度干预导致区块链的发展进入了一种“乱象”，我认为要想让区块链技术得以最终落地并广泛应用必须打击资金的过度干预现象，譬如炒币、炒作概念股等，这样会使投资者看不到区块链市场的发展前景进而丧失投资信心。

2 Target 案例启示

Target(塔吉特)是全美第二大零售商，其通过精确的数据挖掘与数据分析比一位女孩的

亲生父亲更早得到其怀孕的消息。Target 的分析师通过对孕妇的消费习惯进行了一次次的测试和数据分析，根据顾客内在需求数据精准地选出其中 25 种商品，并对这 25 种商品进行同步分析后基本就可判断哪些顾客是孕妇，甚至还可以进一步估算出预产期，在最恰当的时候给她们寄去最符合她们最需要的优惠券。实现此种精准营销的方式是每次顾客初到 Target 刷卡消费时都会获得一组包括顾客姓名、信用卡卡号、电子邮件等的顾客识别编号，日后凡是顾客在 Target 消费，计算机系统就会自动记录消费内容、时间等信息，再加上从其他渠道取得的统计资料，久而久之 Target 便能形成一个庞大的数据库并运用于分析顾客的喜好与需求。

此案例给我带来了两点启示：商务智能并非是“商务”和“智能”的简单加总以及本专业的价值所在。首先第一点，计算机技术的发展与应用为现代商业社会带来了极大的便利，正是数据库、数据挖掘、关联分析等技术帮助 Target 精准定位并预测孕妇人群。但是信息技术仅是实现商业逻辑的途径而非全部，商务智能是整体性问题而非仅是技术问题，市面上数不胜数的商业组件可以解决如何对数据进行存储、查询、分析等问题，但是解决这些组件如何自洽的实现商务逻辑却是一个很复杂的问题，需要决策者联系业务实际多加思索。以 Target 精确的数据挖掘与数据分析为例，如果单纯的想要寻找孕妇的共性，可以借助的数理、统计学分析理论、模型、方法有很多，但是更重要的是根据人群特点、购买习惯、购买特点进行实际性分析。

第二点是通过 Target 精准营销案例中也认识到了本专业的价值所在。如果在 BOSS 直聘、前程无忧等平台输入“数据分析”或“数据挖掘”等工作岗位的需求时，得到的技能需求基本是较高的数理思维逻辑、统计分析方法与技术和计算机、统计学等相关专业背景或工作经历，或是熟悉 Python、Java、Hadoop、云计算等主流大数据处理框架，这些对于信管专业学生来讲还是有相当高的入职门槛的。但是从 Target 精准营销案例可以看出，挖掘商业价值时对商业逻辑、商业流程的理解也同样重要，这需要既懂业务又懂技术更懂如何将二者逻辑自洽地、相辅相成地结合并用于解决实际问题的人才，而这正是信管专业的学生所擅长的。这也启示我们既要技术过硬，更要加强加深对实际场景下业务的应用与理解。

3 “大”VS“数据”之我见

课堂上老师抛出了一个很有价值的问题，即大数据时代下数据的体量“大”更重要还是“数据”的质量、维度等属性更加重要。基于上学期翁老师所授《机器学习与数据挖掘》的学科基础，我们知道对于机器学习尤其是深度学习而言，如果数据集的规模较小则会使得模型过拟合，使其难以充分学习全体数据集的相关特征；而如果数据集的“噪声”过多，即数据的质量较低、维度较少则会使得模型过拟合，使其鲁棒性较差，难以应对客观世界的诸多现实情境。

按照传统应试思维，此选择题将选择 C 选项，即“同等重要”，因为这是符合传统教育

体系下应试思维的逻辑产物，但是老师最终的结果却大大激发了我们的思维。“同等重要”是应试思维下不加思考、忽略现实场景的中庸之举，老师教导我们要联系现实，实际商业应用中“大”或“数据”在不同的场景下往往具有不同的重要性。下面将结合我最近学习到的自动驾驶领域的一个案例谈谈对老师这种思想的理解。

对道路近况、行驶信息等信息进行实时收集对于自动驾驶的实时分析、道路提取、异常监测等十分重要，而这依赖于车载传感器、高清摄像头的实时路况、车辆行驶信息采集的“高清地图”。自动驾驶领域的高清地图体现在两个方面：一方面是绝对坐标精度高，另一方面是信息更加丰富。传统导航地图通常只提供路网结构信息和粗略的几何点位置。而高精地图除了这些信息外，还会包含车道信息（车道线位置、类型，车道方向、车道交通限制信息等）、交通标志信息以及红绿灯、立交桥、高架桥等的路况位置信息。

自动驾驶领域对高清地图的采集与研究主要分为两个流派：一是认为“大”更重要，即牺牲一小部分清晰度和重要信息，尽可能地采集多的实时路况信息与车辆动态行驶状况，该流派主要以百度、蔚来等国内自动驾驶领域翘楚为主。该流派主要得益于国内百度云、阿里云、腾讯云所提供的效能卓越的实时在线计算、云计算以及云存储等能力，基于这些云计算平台，自动驾驶车企可以采集海量的路况进行实时流式分析并形成巨量的模型库、知识库与方法库为道路信息提供实时决策。从机器学习、深度学习的角度来看，海量的地理信息数据与计算机视觉数据可以使得模型得到充分的训练，得到较高的自动驾驶稳健性和鲁棒性，但是由于牺牲了部分清晰度，其对于道路突发状况以及特殊路况的处理能力还有巨大的提升空间。

另一个流派认为“数据”更重要，即不过度追求才采集道路信息的体量，而是尽可能多地采集和反映道路状况的细节信息，该流派主要以车企巨头特斯拉为主。该流派投入大量的测绘车在道路上行驶并采集原始的图像与激光数据以及部分静态控制点信息，利用感知系统感知定位元素，再和高精地图中的定位元素做匹配，以此来实现亚米级或是分米级的定位。该流派主要得益于测绘学、遥感学、地理信息科学、地图学、计算机科学等多学科多领域全生态的产业链及科研人才的投入与参与。从机器学习、深度学习的角度来看，更好的数据胜过更好的模型、算法，当数据集的噪声比例很小时，相对简单的模型也能在较短时间内得到精度较高的决策结果，但是由于数据集的体量较小，会导致模型欠拟合而稳健性、鲁棒性较差，可能出现正常驾驶情景下的突发应急反应。较为典型的案例便是特斯拉刹车失灵、高速骤然失去动力、变速箱控制器突发故障等安全事故。

基于上课时抛出的该问题以及此自动驾驶领域高清地图“路线之争”的问题，可以看出现代商业社会下，到底是数据体量“大”更重要还是“数据”的质量高更重要应结合实际商业背景、商业环境等进行分析。以上文分析的自动驾驶领域为例，不管是信息更加丰富所代表的“大”，还是“绝对坐标精度高”所代表的“数据”，本质上多有其在特定社会环境、商业氛围下得以广泛推广的自治逻辑。同时追求二者的最优在很多情况下是办不到的或者是没有意义的，例如既要求采集道路状况和车流实时信息的体量和质量最优不仅对其数据存储、数据传输等是巨

大的考验，而且在突发状况时选取哪些信息、如何选取信息以及选取后如何计算等都是莫大的难题，因此应在不同的场景下侧重于不同的维度，譬如大多数行驶情景下应采集海量的数据进行学习训练，而在极少数的突发情况下应该采集更多细节的数据为决策提供参考。

以上便是我对大数据时代下“大”VS“数据”问题及其相应案例的所感所悟，给我的启发是遇到新的问题要结合实际场景、实际案例多加分析思考，不能局限于应试思维下的非黑即白、非对即错式的思考方式，而是应该更多角度、更全方位、更多来源、更深层次地旁征博引、联系实际并多加思考。

4 煤炭矿井重点工程 PERT 决策支持系统学习心得

课堂上老师分享了煤炭矿井重点工程 PERT 决策支持系统的开发与设计，开发背景是企业面临着工程进度计划制定与控制方面的难题，具体表现为网络计划图不会画、工程计划不会优化，工程超期、成本超标等工程建设现状亟需良好解决，不可预见因素多、风险性高的地下作业环境，通风、提升、排水、洒水等复杂的工序关系，协作单位多导致的沟通协调困难，工程时间长，耗资大却见效慢等诸多因素更是加剧了这一困难性和复杂性。

在此背景下，煤炭矿井重点工程 PERT 决策支持系统的设计与开发便应运而生，该系统基于联合应用程序开发（JAD）和快速应用程序开发（RAD）两种方法相结合的协作开发方式，确定了网络图绘制、网络参数计算、优化模型的设计与求解、工程进度实时监控与决策分析等网络计划应用的关键环节和系统开发所要解决的关键技术问题。

系统主要借助于 Office Visio 2007 和 UML (Unified Modeling Language) 统一建模工具进行设计，Visio 可以绘制业务流程图、软件界面、工作流表图、数据库模型和软件图表等便于开发设计人员记录、设计和完全了解业务流程和系统的状态，深入了解复杂信息并利用这些知识做出更好的决策。UML 是用来对软件系统进行可视化建模的一种语言，UML 为面向对象开发系统的产品进行说明、可视化和编制文档的一种标准语言。

在数据库技术和集成开发技术的支撑下，矿井重点工程 PERT 决策支持系统有工程基础数据管理、网络图自动绘制、网络优化控制、工程进度查询与决策分析以及系统辅助管理五个关键模块。

通过学习煤炭矿井重点工程 PERT 决策支持系统的案例，不仅让我重温并加深了对 Java 面向对象程序设计、信息系统的分析与设计、决策支持系统、商务智能、运筹学、数据结构等相关课程进行回顾并整合，更对本专业有了更进一步的认知，认识到既要懂系统设计与开发的基本技能与技术，更要深入实际问题、实际案例、实际背景去分析业务流程与业务逻辑，不断学习。

5 大学生体质健康指标空间效应研究学习心得

助教学姐在最后两堂课对大学生体质健康指标空间效应研究进行了分析展示，并且此次

实训的内容也是对本校 2014-2017 年的体测信息进行分析。让我有了三方面的心得与认识。

- 1 开展一项研究前要做广泛、全面而充足的文献前沿分析与文献综述，挖掘研究价值与创新点：**学姐在开始正式分析展示前进行了广泛全年且充足的文献综述和可视化分析，向我们展示了当前关于大学生体质健康指标研究的现状以及存在的问题，援引国家相关评判指标并通过严密的分析指出其存在可改进之处，然后通过实证分析验证了所提出模型的有效性和创新性。这种研究范式对我今后的学术研究有了很大的启发，科研学术工作从来不是闭门造车，而是要广泛吸纳借鉴前人研究成果，在前人的肩膀上为本学科、本领域、本方向探索出一条新的前进方向，改进现有研究的不足或缺陷，这是我从学姐的分析中所感悟到的。
- 2 学习到了空间效应、空间面板计量这一计量经济学模型：**与以往我所接触到的应用统计分析不同的是，空间计量是将传统的统计分析加入了空间效应后做的系列回归。“橘生淮南则为橘，生于淮北则为枳，叶徒相似，其实味不同。所以然者何？水土异也”，即使是同一物种，不同的生长环境也会导致不同的生长结果，例如各个城市出台的限购政策不仅会影响当地城市的房价，还会通过诸如“人口流动”等影响另一个城市的房价。这为我们研究来自不同区域、不同地理的学生的相关体能素质情况提供了理论依据。
- 3 认识到了 R 语言绘图可视化功能的强大：**之前对 R 语言绘图功能的强大仅停留在老师的讲解层面，但是通过学姐在课堂上展示的 PPT 中各种精美的图表，以及实训时自己动手绘制的基于 ggplot 程序包的各种图表，让我对 R 语言绘图强大功能有了进一步直观的认识，其美观程度要远高于 Matlab、Python 等绘制的图表。

6 致谢

经过两个学期的共同学习与陪伴，李四福老师和学姐带领我们学习了《信息系统分析与设计》和《商务智能》这两门课，《信息系统分析与设计》让我们掌握了开发信息系统的基本流程、技能与方法，而《商务智能》这门课教会了我们如何使用信息系统从海量的商业数据中挖掘出有价值的、可供决策的信息。在课程学习外老师还分享了众多案例，譬如 Target 精准营销案例、煤炭矿井重点工程 PERT 决策支持系统等，这些案例都不同程度、不同层次地加深了我对本专业的认识，同时也钦佩于专业的专业知识和技能。

我认为更重要的是老师一再强调的“大学之道”的精神，特别是到了现在大三下学期面临诸多人生、职业选择的阶段，大学里面所形成的一些美好品质、习惯能加强自己的职业竞争力和综合素质，譬如团队协作能力、上进心、抗挫能力、准时自律等。“大学之道，在明明德，在亲民，在止于至善”，大学的作用更多的是教会我们明辨事理，有所为而有所不为，在于传授新的思想，大学的宗旨在于弘扬光明正大的品德，学习和应用于生活，使人达到或趋近于至臻至美的境界。

附录七：普叶课程学习报告

商务智能学习心得

(普叶 20191001713)

一、R 语言决策树

此次在完成体测数据分析时，我被分配到的任务是利用机器学习的相关知识，对体测数据进行分类，分类是一种基于一个或多个自变量确定因变量所属类别的技术。我说考虑过用于分类的算法主要有 K-近邻算法（K-NN）、支持向量机（SVM）、决策树分类。

K-近邻算法（K-NN）是一种最简单的分类算法，通过识别被分成若干类的数据点，以预测新样本点的分类。K-NN 是一种非参数的算法，是“懒惰学习”的著名代表，它根据相似性（如，距离函数）对新数据进行分类。简单来说，KNN 可理解为一种死记硬背式的分类器，记住所有的训练数据，对于新的数据则直接和训练数据匹配，如果存在相同属性的训练数据，则直接用它的分类来作为新数据的分类。K-NN 能很好地处理少量输入变量（ p ）的情况，但当输入量非常大时就会出现问题。

支持向量机（SVM）是基于定义决策边界的决策平面。决策平面可将一组属于不同类的对象分离开。在支持向量的帮助下，SVM 通过寻找超平面进行分类，并使两个类之间的边界距离最大化。SVM 中超平面的学习是通过将问题转化为使用一些某种线性代数转换问题来完成的。但使用 SVM 对高维数据分类，就要使用核函数，建模复杂，不便很好的对体测数据进行分类。

决策树的计算就是对数据进行挖掘分类的过程，利用决策树这一数据分类器，可以将一些无序的数据分类进行分析和推导。其中根节点、内部节点、叶节点是决策树用于分类计算的显著特征。决策树方法的每个根节点到叶节点都有相应的路径将数据按照一定规则进行分类。利用决策树对大学生体测成绩进行具体的研究，可以更加直观的显示出相关因素对体测成绩的影响。

最后对比分类效果，以及建立模型的难易程度我选择了决策树分类算法。利用最后输出的决策树，我们可以找出其根节点和内部节点的关系，对大学生体测影响因素进行排序。到此对分析方法和模型有了大概的想法，接下来就是通过 R 语言来实现相关模型。

我本以为选择模型，思考分类方法和预计达到的结果能较为简单的完成分析，但没想到在 R 语言代码的实现方面我也遇到了很多麻烦，甚至起了换个模型和方法的想法，但最后在尝试 KNN 分类算法无果后，还是转回决策树继续探究，最后经过多次试错才得出了想要的结果。以下是我对决策树实现时，遇到的几个主要问题。

首先是数据的读取，R 语言在使用数据地址时都要使用绝对地址，但我的的数据文件却

一直在报错，数据无法读出。刚开始我以为是路径中存在中文，然后重新改换了一个英文路径，但还是读取失败。然后在尝试了几种百度方法无果后，在小组同学那里找到了答案，要在读取代码后面加上 `fileEncoding = "GBK"`，这样就能顺利的读出 csv 文件了。

然后我将数据的 1/3 划分为测试数据 2/3 划分为训练数据，来检验模型的预测效果。同时通过设置交叉验证次数、最小分支节点数、叶子节点最小样本数、树的深度、某个点的复杂程度等相关参数作出决策树模型。

但是在建模代码运行时，我却发现 R 语言一直在转圈无响应，既没有报错，也不出结果。刚开始我以为是自己建模参数有问题，但是多次调参后都是没有响应。然后我随意选择几个数据运行了一下，发现没有问题，所以问题就是现有数据特征太多，导致决策树建模困难。最好的方法就是将一些无用的特征删除，考虑到本次探究的是各体测项目对体测成绩的影响，因此最终我将数据按照性别区分开，将除了体测项目分数、总分数、等级以外的其他数据删去，最终顺利画出了男生数据和女生数据的两个决策树。

观察初步得到的决策树，我发现决策树的枝叶还是太多，不能很好的得出准确的结果。所以我利用 R 语言的 `rpart` 包提供的一种剪枝方法--复杂度损失修剪的修剪方法。并且使用 `printcp` 这个函数显示分裂到的每一层，对应的某个点的复杂度是多少，平均相对误差是多少。然后我就可以使用具有最小交叉验证误差的某个点的复杂度的方式进行剪枝。

最后经过剪枝，我得到了男生体测影响因素和女生体测影响因素决策树，由此我推断出影响男生体测成绩的因素的前三位排序依次是引体向上、1000 米、立定跳远，影响女生体测成绩的因素的前三位排序依次是立定跳远、50 米和 800 米。并基于此给出了自己关于提升体测成绩的相关建议。

在上学期的机器学习课程中，我曾经使用 python 实现决策树分类，但那次调用的是自带的癌症数据包，实现起来会更加方便，实现效果也更加好看。当使用 R 处理外部数据时，因为想要达到想要的结果，数据处理时要考虑的更多。以我自己的使用感受来说 R 安包更加容易，可选的包更加丰富，更加简单一些，而 python 相对要复杂一点，但界面相对美观。

二、深入浅出数据分析读书笔记

1、分解数据

(1) 知识点

数据分解的分析思路主要是确定、分解、评估、决策。

确定问题：为了确定问题，通常需要进一步的客户信息。但是，有时客户不了解问题或目标，因此分析师需要过滤客户提供的实际情况不匹配的信息。换句话说，你不能听客户片面的话，而必须依靠数据来说话。

分解问题：其目的是将大问题分解成小问题，将大数据分解成更小的块。分解的一个重要因素是比较对象的发现。然后，对数据进行分解和汇总。

评估问题：根据已有的信息，得出各种结论。数据分析的核心是有效的比较。

决策问题：重新组合相关推论，作出正确的决策，以此来明确自己的假设和结论。

(2) 实际应用——化妆品公司提高销量

化妆品公司想要增加自己的销量，首先要确定自己想要解决的问题。然后从客户那里获取信息，分析假设条件。同时，拆分给定的数据和对应的问题。接着综合对问题和数据的观察作出评估，撰写分析报告。得出分析报告后，要继续寻找增加销量的方法，审视客户的假设是否有误，并提出自己的假设。最后，通过收集数据、数据挖掘，得出最终的决策结论

2、实验

(1) 知识点

统计和分析的最基本原则之一是比较法。数据有意义的唯一方法是比较它们。比较越多，分析结果就越准确。尤其是观察研究。

混杂因素：研究对象的个人差异，他们不是试图比较的因素，最终会导致分析结果的敏感度变差。观察分析法充满混杂因素。从对象池中随机选择对象是避免混杂因素的好办法。

好的实验总是有一个控制组（对照组）。控制组：也称作对照组。一组体现现状的处理对象，未经过任何新的处理。

数据分析的重点在于分析的结论有意义。要学会拆分数据块，管理混杂隐私。并且拆分的数据块要具有同质性。

观察研究法：被研究人自行决定自己属于哪个群体的一种研究方法。使用观察研究法时，应当假定其他因素会混杂你的结论。观察数据本身无法预示未来。

(2) 实际应用——寻找星巴克销量下降原因

星巴克由于销量下滑需要相应的解决方案来恢复销量。首先通过统计用户调查表我们发现质量感波动幅度较大，由此猜测出消费者可能觉得品牌不再物超所值或者猜测经济环境让人们对于价格更加敏感。这是 soho 分区经历提出了质疑，经过分析后发现店址是混杂因素。然后设计实验，一次来证明哪种猜测是根本原因。首先将大的地理区域分成小的地理区域，排除混杂因素。然后控制组维持现状，实验组 1 号降价，实验组 2 号游说。最终得出了结论，游说人们星巴克咖啡有价值是比降价和维持现状更有效的提高销量的方法。

3、最优化

(1) 知识点

约束条件就是无法控制的因素，如单位时间生产量。决策变量是可控因素，目标就是在不超出约束条件的情况下，对决策因素做一个组合，实现最大利润。

将决策变量，约束条件和希望最大化的目标合并成一个目标函数。任何最优化问题都有一些约束条件和目标函数。

一种产品越多就意味着另一种产品减少。不要假定两种变量是不相关的，创建模型时，

要规定假设中的各种变量的相互关系。同时，可以用电子表格实现最优化，比如 Excel 里的 Solver 求解器。根据实际情况的变化我们也要调整自己的模型，以此来提供更优结果。

(2) 实际应用——橡皮鸭、橡皮鱼生产问题

客户想要尽量提高利润，并保证橡皮鸭和橡皮鱼的产量都合适，找出理想产品组合。首先要了解产品的盈利情况和相关的约束条件，基于此构建出目标函数，并用图形绘制出所有的约束条件。绘制完图形后，发现模型有误，因此进行校正假设，重新规划约束，重新分析比较行业历史数据，从而发现新的约束条件，得出更加合理的模型。

4、数据图形化

(1) 知识点

可以利用散点图进行探索性数据分析，通过探究自变量和因变量的关系，发现相关的因果关系

要学会用数据思考，面对大量数据时，要记住目标，目光停留在和目标有关的数据上，无视其他。

要记住数据可视化对的初衷是为了作出正确的比较，好的数据图形化，通过有效的比较可以展示多个变量的关系。

可以利用 R 软件、Edward Tufte 来进行图形多元化处理。

(2) 实际应用——优化网站

进行优化网站时，首先要了解情况，然后可以利用 excel、illustrator、R 程序、Edward Tufte 等工具创建图形，进行参数均值和多因素的对比，从中找到最优设计方案。并且从方案中寻找因果关系，提出假设，并用已知数据证明。

5、假设检验

(1) 知识点

要学会观察数据变量，判断他们之间是正相关，还是负相关。线性关系大都存在于理想模型中，现实世界大都呈现网络关系。

假设检验的核心是证伪，证伪不是选出最合理的假设，而是剔除无法证实的假设。满意法是选出最可信的第一个假设。进行假设检验时，要使用证伪法，回避满意法。证伪法可以对各种假设保持敏锐，防止掉入

诊断性是证据所具有的一种功能，能够帮助评估所考虑的假设的相对似然性。如果证据具有诊断性，就能帮助对假设的排序。

(2) 实际应用——预测手机发布时间

要预测出公司下一个月手机壳厂商是否会推出新手机，首先要收集、汇总相关的信息，由此来确定出各变量间的正负相关性。然后利用工具画出关系网，设置合理的假设，并根据已知资料进行排除，对于一些不能排除的假设，可以根据多条证据来对假设评级，最终得出

最强假设。

6、贝叶斯检验

(1) 知识点

条件概率是以一件事的发生为前提的另一件事的发生概率。基础概率又叫事前概率。在根据试验结果分析之前，已经知道的概率。如果有基础概率，一定要考虑。在检验中可以将概率转变为整数，然后进行思考，这样可以有效的避免犯错。

贝叶斯规律可以反复使用，每当产生了一个新的概率，就可以把新的概率当成基础概率，不停的迭代验算。避免基础概率谬误的唯一方法就是对基础概率提高警惕，而且务必要将它整合到分析中去。

(2) 实际应用——阳性患病概率

可以利用贝叶斯规则，预测检测结果为阳性的患病概率。社会中患病人数 $1\% = \text{基础概率}$ ，患病检测结果阳性概率 90% 、未患病检测结果阳性 10% 、条件概率。基础概率和条件概率结合算出检测结果阳性和患病概率。

三、课堂感悟

对比上一学期的大象课程，这个学期第二次接触李四福老师的课，给了我很多不一样的体验，也有了更多的收获。如果说大象课程是基于面向对象的思想对实际问题进行建模解决，那么商务智能则更偏向于整合相关数据，为企业作出最明智的商业决策。一些重决策对于企业的发展起着决定性的作用，要想“运筹帷幄之中，决胜千里之外”，就必须“耳听六路，眼观八方”。因此商务智能融合了多方面的内容，利用先进的信息信息技术和管理理念加工处理数据，由此服务于企业不同层次的经营决策需求，提高企业的竞争力。

李四福老师在课上就时常强调要多学习新知识，关注最新的技术。商务智能中运用到的很多技术，例如数据仓库、在线分析处理、数据可视化、数据挖掘、数据的分类、聚类等方法，如果我们能够熟练的掌握，不但有利于我们后期完成自己的毕业设计。对于我们研究生阶段的研究或者找工作的面试，都大有裨益。信管毕业生要想提高自己的综合竞争力，以上技术确实值得我们好好钻研。同时在本学期的另一门课决策支持系统大作业上我也用到了例如数据挖掘、数据分类、数据可视化等方面的知识，也使得我的作业完成的更加顺利。

“读万卷书，不如行万里路”，我发现李四福老师更希望我们将自己学到的知识和实际结合起来去解决问题，老师和我们讲述的各类案例，和体测数据分析都是如此。我认为信管专业是一门急需要多动手的专业，代码自己手敲出来和只是在书本上看两眼，结果是完全不同的。虽然很多时候实践中会遇到各种问题，但是自己在一步步解决问题的过程中，也是在一步步的学习，譬如这次我使用决策树模型分析了体测数据，那么下次我在用决策树模型使用R语言进行分析时，相信就会比这次更加熟练。

很感谢李四福老师两门课的教导，我认为一个真正好的老师除了要教会学生理论知识，

更加重要的是如何教会学生能将其转化为现实。一些老师只注重理论，学生缺乏实践，不能很好的巩固所学的知识。而另一些老师在学生理论知识还不够丰富的时候，就要求完成一个巨大的工程，这无异于是揠苗助长，容易让学生丧失前进的动力。李四福老师则和前两种老师不同，他擅长引导学生分析实际问题，同时在实践方面也对学生耐心指导，布置的任务也比较巧妙的锻炼了学生。

再此我诚挚的向李四福老师和助教学姐表示感谢，感谢李四福老师兢兢业业，临退休还在培养新的信管专业年轻老师，还在用自己的人生智慧认真的教导我们。也感谢学姐及时为我们答疑解惑，为我们讲解体测数据论文。犹记得李四福老师上第一堂大象课时，让我们念的《大学之道》，曾今觉得梦回高中语文课，现在想起来却感慨万千。“师者，所以传道，授业，解惑也”感谢老师两门课程的教授！

附录八：马宸晨课程学习报告

商务智能学习心得

(马宸晨 20191000951)

大数据时代，随着云计算、5G、人工智能、物联网等新兴技术突飞猛进的发展，企业所处的环境更加的复杂多变，市场的竞争更加激烈，企业自身的组织结构越来越复杂，规模也越来越大，在这样的环境之下，数字化转型成为越来越多企业的共同目标。而商务智能（BI）是信息融合背景下企业精益化管理和科学决策能力的反映，体现了如何运用数据分析和数据挖掘等手段，将数据、信息转化为知识的过程。通过本学期商务智能课程的学习，我对其有了深入的理解，接下来我将结合课程所学，从商务智能发展历程、概述、核心技术，数字经济时代为何要发展商务智能，课上案例分析总结，R语言可视化心得以及个人感悟分别进行报告。

一、商务智能发展历程

20世纪以来，许多国家纷纷利用信息技术这一先进生产技术来推动本国的经济社会发展并将推动信息技术应用作为国家发展战略。从90年代初期开始，信息技术以它独有的渗透性、倍增性和创新性点燃了一场全球范围的信息革命，使整个世界发生着最为迅速、广泛、深刻的变化。这意味着：从商业公司的企业内部各种管理与运营数据，到个人移动终端与消费电子产品的社会化数据，再到互联网产生的海量信息数据等正在飞速增长，海量商务数据亟待智能处理。

正此时，商业智能概念逐渐深入，商业智能供应商、工具、技术逐渐成型，互联网的商业化开始形成，美国提供的信息服务器成为最流行的在线服务。21世纪初，云技术和基于互联网的软件成为实时系统，改进的可视化技术改变了数据的浏览方式，为商业智能带来了新的机会。到2010年底，35%的企业正在使用商业智能，而67%的“一流”公司都有某种形式的自助服务商业智能。如今，商业智能成为跨国企业到中小企业中所有人的标配工具，目前商业智能已经可以跨多个设备，并可以完成可交互式的分析推理。

目前，全球商业智能市场快速发展，根据查阅到的数据，2018年全球商业智能和分析软件解决方案市场规模达到216亿美元，同比增长11.7%，其中现代BI平台增长速度最快，增速达到23.3%。其中：中国商业智能软件行业规模约为16.6亿元，同比增长25.8%，发展迅速，且帆软以14.9%的市场份额排名国内市场第一，超越微软和SAP，说明未来中国BI行业的发展潜力巨大。这也印证了老师在课堂上讲的思政内容，新中国成立以来，我们党不仅领导人民进行大规模工业化建设，用几十年时间走完了发达国家几百年走过的工业化历程，创造了经济发展的“中国奇迹”，并且进入21世纪以来，我国坚持以信息化带动工业化，以工业化促进信息化，迅速缩短同发达国家的差距，不仅深刻改变了工业发展的面貌格局，也成就了世界网络大国的地位。

二、商务智能概述

BI 是一系列的概念、方法和过程的综合，通过这些概念、方法和过程来获取和分析数据，提取有用信息，帮助更好的决策，特别是战略决策。老师在课上反复强调，一个真正意义上的 BI 必须跟人员、组织和技术这三要素密切配合，监控企业的关键绩效指标，包括企业外部环境、顾客、供应商，竞争者等，及时给各层决策者提供智能支持，帮助企业构建更好的盈利模式。

BI 涉及到企业战略、组织、人员、技术、业务五个层面的整体解决方案，通过这五个层面，把企业整合成一个信息工厂。在 BI 的价值链中实现数据到信息、知识、智能、利润的价值增值，从而使企业取得竞争优势。同时在问题和决策之间，有信息的反馈，保证战略决策和执行对环境变化的适应性。

三、商务智能核心技术

商务智能的核心技术有三，分别为：数据仓库(DW)、联机分析处理(OLAP)和数据挖掘(DM)。

数据仓库是一种语义上一致的数据存储，是指从多个数据源收集的信息，以一种一致的存储方式保存所得到的数据集合。数据仓库的特点是面向主题的、集成的、与时间相关的、不可修改的数据集合。实施 BI 首先要从企业内部和企业外部不同的数据源，如 CRM、SCM、ERP 系统及其他应用系统等搜集有用的数据，进行转换和合并，因此需要数据仓库和数据集市技术的支持。

联机分析处理是一种软件技术，它使分析人员能够迅速、一致、交互地从各个方面观察信息，以达到深入理解数据的目的。OLAP 具备上钻、下钻、切片、切块和旋转 5 个基本功能。数据仓库与 OLAP 的关系是互补的，现代 OLAP 系统一般以数据仓库作为基础，即从数据仓库中抽取详细数据的一个子集并经过必要的聚集存储到 OLAP 存储器中供前端分析工具读取。作为商业智能 BI 软件的核心技术，OLAP 可以在使用多维数据模型的数据仓库或数据集市上进行，充分发挥 OLAP 的联机分析的功能和特性。

数据挖掘即数据库中的知识发现，是一个在数据中提取出有效的、新颖的、有潜在实用价值和易于理解知识模式的高级过程。数据挖掘技术以企业拥有的大量数据为对象，通过抽取、转换、装载等数据处理方法，发现数据的关联与趋势，探寻出其中的业务规律和模式，在关系数据库中存储多维数据集数据。在课堂上，我们着重学习了关联分析算法，它是数据挖掘领域的热点，关联规则反映一个对象与其他对象之间的相互依赖关系，如果多个对象之间存在一定的关联关系，那么一个对象可以通过其他对象进行预测。在课上，我按照老师讲的方式计算了频繁关联规则，并在课下用 R 实现了该题，这使我对关联分析算法有了更深刻的认识。

四、数字经济与商务智能

接下来，我将结合老师课上讲的数字经济的驱动，浅谈自己对数据经济时代为何需要商

务智能的理解。

国务院总理李克强在政府工作报告中介绍“十四五”时期主要目标任务时，指出“加快数字化发展，打造数字经济新优势，协同推进数字产业化和产业数字化转型，加快数字社会建设步伐，提高数字政府建设水平，营造良好数字生态，建设数字中国”。所谓“数字经济”，是指以使用数字化的知识和信息作为关键生产要素、以现代信息网络作为重要载体、以信息通信技术的有效使用作为效率提升和经济结构优化的重要推动力的一系列经济活动。数字经济时代下，数据作为关键生产要素，在经济活动中的作用越来越突出；与此同时，商务智能作为处理现有数据，并将其转换成知识、分析和结论，辅助业务或者决策者做出正确且明智决策的有力工具，其价值也越来越突显。

数字经济时代，数据量呈爆发式增长，且绝大部分为非结构化数据。由于非结构化数据的格式和标准不一，如何有效地利用这些数据资源为企业经营决策提供更多价值，就成为了关注的焦点。BI 融入大数据相关技术，通过数据采集、数据存储、数据分析和数据应用等环节对不同来源、不同类型的数据进行处理，有效地整合多渠道、多类型的数据，并用于分析和挖掘。

商务智能目前已经覆盖制造、零售、医药、教育、金融、航空与物流等各个行业，利用数据分析、挖掘与智能洞察，帮助众多知名企業解决了经营和发展中的各种瓶颈问题，实现了更大的发展。

可以预见，数字经济时代，随着越来越多的企业和组织开始意识到数据已成为关键生产力，数据资产管理已成为企业竞争力的重要来源，商务智能平台作为高效的数据分析与挖掘工具，将在帮助企业与组织分析和洞察数据背后价值方面，发挥着更加关键的作用。

五、案例分析

在课堂上，老师讲了许多与商务智能相关的案例，其中有三个给我留下了深刻的印象。

首先是美国零售巨头塔吉特的案例。孕妇对于零售商来说是含金量很高的群体，但她们一般会去专门的孕妇商店，忽视了 Target 有孕妇需要的商品。Target 为了将这部分细分顾客从孕妇商品专卖店的手中截留下来，收集了关于消费者行为的海量相关数据：以往的购物记录、信用卡记录、使用优惠券记录、调查问卷、邮寄退货单、访问网站记录等。Target 通过对顾客的消费数据进行数据挖掘发现，许多孕妇在固定的怀孕周期购买特定商品。Target 最终选出了 25 种典型商品的消费数据，构建了“怀孕预测指数”。通过这个指数，塔吉特能够在很小的误差范围内预测到顾客的怀孕情况，也就能早早地把孕妇优惠广告寄发给顾客，最终吸引了大量的顾客资源。我认为这个案例充分地显现了商务智能在帮助企业分析和洞察数据背后价值方面发挥的重要作用，它帮助 Target 收集、管理海量数据，并通过数据挖掘技术分析这些数据中的关联性，极大地提高了 Target 的市场竞争能力。

再是矿井决策支持系统案例，通过听老师上课细致的讲解、观看讲解视频，同时在课余时间搜集阅读李四福老师编写的《煤炭矿井重点工程网络计划决策系统建设》优秀案例，我

更加感叹商务智能的功能之强大。矿井的工程量大、消耗资金多，多项工程建设过程中不可预见因素多，施工方法、劳动组织和工程进度需要随时进行调整，并且工序制约关系复杂。矿井建设中人们习惯采用横道计划方法来编制施工计划和安排生产任务，但是这种图解并没有详细说明不同工序之间的相互制约关系，也没有表明生产成本对工期或工期对生产成本的影响，存在明显的局限性，而网络计划技术克服了传统横道计划方法难以表达工序之间逻辑关系，不便于进度计划的优化与调整等方面的缺陷，但矿井工作人员不会使用计算机绘制网络图，遇到了许多困难。尽管在我校科研团队的带领下完成了网络图的绘制，但也不能从根本上解决工作人员难以绘制计划图的难题，于是，网络计划决策系统应运而生。系统不仅根据工程总体目标以及工程任务的先后顺序、资源约束及其相互关系，制定各个工序的计划表和各个工序之间的逻辑关系，以绘制初始的网络计划图、计算网络参数并确定关键线路，还通过设计优化模型、模型求解，优化网络计划图，这为矿井工程的实施奠定了坚实的基础，带来了巨大的经济效益。该决策支持系统不要求管理人员的计算机专业性，便于理解操作，功能多样，有着重大意义。

最后是学姐最后两节课大学生体测论文的讲解。《大学生体质健康的动态权重评价模型研究》针对现有主观赋权评价法的不足，提出并构建了基于动态权重的大学生体质健康评价模型，动态权重评价模型根据最小相对信息熵原理，用拉格朗日乘子法将主客观赋权法相结合，构建动态权重模型，保留指标实际意义的同时，引入了大量标准数据为基础的客观权重，对学生体质进行客观科学的评价。《大学生体质健康空间效应研究》从学生体质健康现状在空间上相互作用表现的差异性入手，运用 ArcGIS 和 GeoDa 软件进行空间自相关检验，探索我国除港澳台外 31 个省市的学生体质健康指标与其生源地之间是否存在空间效应，以及不同年级之间空间效应的变化趋势。这两篇论文从不同角度对大学生体测数据进行了研究，但本质都是利用数据仓库管理数据、利用数据挖掘相关算法模型挖掘数据之间的联系，给我们的上机作业——体测数据分析拓展了分析思路。

六、R语言可视化心得

用 R 语言对我校大学生体测数据进行统计分析也是这门课上很重要的一部分内容。R 是一套完整的数据处理、计算和制图软件系统。其功能包括：据存储和处理，数组运算，完整连贯的统计分析工具，优秀的统计制图功能以及简便而强大编程语言。接触 R 语言以后，我很快就感受到了它的方便和强大。在描述性统计方面，R 语言中有非常多的函数和包，我们几乎不用自己去编一些复杂的算法，而往往只需要短短几行代码就能绘制出一幅幅精美的可视化图表。它又可操纵数据的输入输出，我们可以自定义功能，这意味着当找不到合适的函数或包来解决所遇的问题时，可以自己编程去实现各种具体功能，这也正是 R 语言的强大之处。在体测分析中，我们小组利用 R 绘制了小提琴图、桑基图、雷达图、山脊图等复杂的图形来描述统计不同类型的体测数据之间的差异，这些图形在 Python 中实现需要编写大段代码，而在 R 中只需短短几行函数调用就可实现，大大的降低了复杂程度。

同时我认识到，R语言只是我们解决问题的工具，而我们对问题的分析首先是要根据理论进行的，例如参数估计、假设检验以及线性回归、时间序列方面的知识，我们只有深刻理解这些理论背后的意义，才能用对R语言中的各个方法。描述性统计也是如此，首先必须要理清数据间的关联，确定分析的方向，用R绘图时才会有清晰的思路，才能全面地分析数据。

在实现上机作业时，我在编写代码时遇到了许多格式上的错误，才深知自己运用R语言并不熟练，我一定更加努力深入地学习R语言，在今后的学习实践中获得更多的知识，为即将到来的专业实习乃至工作打下坚实的基础。

七、个人感悟

在《商务智能》这门课上，我学到了许多BI的核心知识，了解了我国信息化的发展历程，惊叹于中国的大国力量；理解了BI的含义、架构以及核心技术；理清了BI与其他互联网技术的关系，以及它在如今发挥的巨大作用。并且，在上机阶段，我将专业理论、学姐课上讲的论文与上机实践相结合，更加加深了对BI的理解，深深明白了BI的巨大作用。

“物有本末，事有终始，知所先后，则近道矣。”世界上的很多事物都有它的根本和末梢，事情有开端和结尾，我们在对待处理的时候，应该知道孰先孰后，孰本孰末，加以区别对待，这也是我在课上除专业知识外牢记于心的道理。作为一个学生，我们应该始终把学习放在第一位，同时理清学习的思路，分清主次，才能提高学习的效率，做到学有所成。

“单丝不成线，独木难成林”，我还深刻体会到团队力量的强大。如果只靠自己一个人来全面地完成体测数据的预处理、描述性统计、推断性统计以及探索性统计是十分困难的。我们小组分工明确，每个人都能对自己的部分负责，同时积极帮助其他小组成员解决遇到的问题，群策群力，才能很好地完成体测数据的分析。

最后，感谢李四福老师和助教学姐的辛勤付出和悉心指导，润物无声俯亲躬，正是有了老师和学姐的亲力亲为，才能让我学到许多。在今后的学习生活中，我会做到学有所用，不让在课上学到的知识荒废。