**《数据世界探秘》教学大纲**

(2022.9-2022.12)

**课程：** 数据世界探秘

**授课教师：** 吴纯杰 教授

答疑时间： 周二下午15:30-16:30或者微信

办公室：统计1313，65901084

E-mail: [wumaths@mail.shufe.edu.cn](mailto:wumaths@mail.shufe.edu.cn)

**课程类别：** 通识限定选修课

**课程对象：** 各年级本科生

**课程安排说明：** 2022年9月5日—2022年12月19日

教室： 科研实验大楼B103

班级： 合班

时间： 周一18:00-19:40

课程调整：根据学校相关节假日通知。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **学分** | **总学时** | **理论教学学时** | **实践教学学时** | **实验教学学时** |
| **2** | **32** | **22** | **6** | **4** |

**课件网址：** 见学校BB教学资源平台

**教材和参考书目：**

指定教材：

王汉生，《数据思维：从数据分析到商业价值》，中国人民大学出版社，2018年

参考书目：

（1）格鲁斯，《数据科学入门》，人民邮电出版社，2016年。

（2）朝乐门，《数据科学理论与实践》，清华大学出版社，2017年。

（3）杰弗里˙班尼特，威廉˙L.˙布里格斯，马里奥˙F.˙崔奥拉，《妙趣横生的统计学》，人民邮电出版社，2016年。

（4）吴喜之，《统计学：从概念到数据分析（第二版）》，高等教育出版社，2016年。

（5）Robert I. Kabacoff，《R语言实战（第2版）》，人民邮电出版社，2016年。

（6）Norman Matloff，《R语言编程艺术》，机械工业出版社，2013年。

（7）吴喜之，《复杂数据统计方法——基于R的应用（第三版）》，中国人民大学出版社，2017年。

（8）Megan Squire，《干净的数据：数据清洗入门与实践》，人民邮电出版社，2016年。

（9）Winston Chang，《R数据可视化手册》，人民邮电出版社，2014年。

（10）Nathan Yau，《鲜活的数据：数据可视化指南》，人民邮电出版社，2012年。

（11）戴维·萨尔斯伯格，《女士品茶》，江西人民出版社，2016年。

（12）周英、卓金武、卞月青，《大数据挖掘：系统方法与实例分析》，机械工业出版社，2016年。

（13）刘小平，《统计学——理论、案例、实训》，电子工业出版社，2017年。

（14）城市数据团，《数据不说谎：大数据之下的世界》，清华大学出版社，2017年。

（15）达莱尔·哈夫，《统计数字会撒谎》，中国城市出版社，2009年。

（16）戴维·穆尔、威廉·诺，《统计学的世界》，中信出版社，2017年

**先修课程：**

无

**预备知识**

无

**课程达成目标：**

课程授课对象以低年级学生为主，通过介绍数据科学及大数据背景下，数据形态、数据采集、统计调查方法的最新动态和变革，介绍数据分析工具以及可视化技术，帮助学生增强统计认知。课程通过介绍国内统计最新发展动态，尤其是突出大数据背景下，在大数据领域卓有成效的企业。通过数据解读中国发展，增强学生爱国敬业的情感，树立建设富强民主社会主义强国的梦想。课程为学生参与大学生创新创业项目、暑期各类社会实践活动等打好基础，帮助学生提升认知社会和科学研究的能力。

**课程设置知识要求：**

课程具有较强的实践性和应用性。积极组织和鼓励学生参与如“设计调查问卷——借助网络调查平台开展调查——有效避免有偏样本形成——借助数据分析工具完成初步建模和可视化分析”的完整过程。课程加大通识经典读物的阅读量，对于部分课外材料要求同学们至少阅读1-2本，分成小组撰写通识论文报告。

**课程设置能力要求**：

课程建设有助于培养学生数据思维的能力，提升学生动手实践的本领，提高学生数据分析的水平；有助于为后期参与创新创业项目、暑期各类社会实践活动等打好基础、提供帮助，并有助于学生提升认知社会和科学研究的能力。

**考核形式**

期末考试采用课程论文方式，学生的最后的总分计算方法如下：

考勤 10%

团队作业 40%

课程论文 50%

**试卷结构**

课程论文，重在考查学生运用统计软件采用本课程学习的各种统计分析方法进行数据分析的能力，及研究报告的撰写能力。课程论文认定采用的几种方式：

1. 数据分析报告；

2. 项目设计和方法报告；

3. 通识阅读报告或本学期内参加通识阅读大赛获奖根据等级认定分数；

4. 本学期内各类二课相关比赛（数据分析相关）根据获奖级别给予不同分数的认定。

**学术诚实**

涉及学生的学术不诚实问题主要包括考试作弊；抄袭；伪造或不当使用在校学习成绩；未经老师允许获取、利用考试材料。对于学术不诚实的最低惩罚是最终考核成绩给予0分。其它的惩罚包括报告学校相关部门并按照有关规定进行处理。

《数据世界探秘》教学要点

**教学大纲**

**第一章 欢迎来到数据科学时代**

§1 引言

§2 数据科学的内涵和发展

§3 数据科学的学科地位

§4 数据科学的成熟度曲线

§5 数据科学的理论体系

**本章教学要求：**通过讲解近几年几个比较著名的大数据应用案例，引出数据科学时代的到来，同时提升大数据时代对数据应用的重视。接着，对数据科学的相关概念进行简要介绍。包括数据科学的内涵和发展；学科地位；成熟度曲线；理论体系等概念，以期学生对数据科学时代有初步的了解。

**第二章 数据及数据来源**

§1 数据定义和类型

§2 传统的数据来源

§3 大数据数据来源

**本章教学要求：**掌握对数据的正确理解；掌握4种角度下数据的不同类型；了解传统的数据来源方式（统计调查、实验、观察研究）；理解大数据数据的主要来源。

**第三章 数据分析的道与术**

§1 数据分析定义与价值

§2 统计建模工具

2.1 SPSS

2.2 SAS

2.3 R语言

2.4 Python

§3 探索性数据分析框架

3.1 数值描述分析

3.2 数据可视化

3.3 数据清洗和转换

3.4 实例分析

§4 统计建模技术

4.1 抽样分布

4.2 假设检验

4.3 预测分析

4.4 聚类分析

**本章教学要求：**掌握数据分析的基本框架。在理解数据分析定义的基础上，介绍几种最常用的数据分析工具，掌握数据的探索性分析和统计分析技术的基本内容。提供相关学习材料，辅助学生对数据分析框架有初步了解。

**第四章 数据陷阱**

§1 有偏的样本

§2 幸存者偏差

§3 精心挑选的平均数

§4 相关关系的误解

§5 辛普森悖论

§6 统计显著与经济显著

**本章教学要求：**理解在数据分析过程中，经常容易陷入数据陷阱。从而在以后进行数据分析或者翻阅数据分析报告时，警惕可能碰到的数据陷阱。以案例的形式介绍了一些常见的数据陷阱，并阐述了这些数据陷阱背后的相关统计知识和理论。

**第五章 数据世界专题课**

§1 经济管理大数据案例及其应用

§2 以国家战略为主导向，探统计分析之新思维

§3 工欲善其事，必先利其器--解读大创项目和建模比赛

§4 社会生活中的统计思维

**本章教学要求：**通过讲解我国社会经济领域发展的应用案例和大学生身边的创新训练项目、建模比赛以及数据诊所等解读数据世界，有助于学生了解国内外最新态势，通过介绍各行业的大数据发展和应用，融入思政育人元素，弘扬社会主义核心价值观。