**广州商学院**

**课 程 论 文**

**题目：基于Python的二手房数据爬取及可视化分析**

课 程 名 称 **数据分析与处理实验**

考 查 学 期 **2021/2022 学年 第二学期**

考 查 方 式 **考查（论文）**

姓 名 王 弦

学 号 202006120123

专 业 软件工程

成 绩

指 导 教 师 邹锋

**摘要**

网上的房源信息比较繁杂，为了帮助人们更全面地了解广东省广州市的二手房价变化趋势，获取精准的房源数据，本文利用爬虫技术爬取链家网站上的二手房信息，并经过清洗和可视化处理，将广州各地区的二手房信息以更直观的形式呈现出来并进行比较，从而为购房者提供更多的帮助。

**关键词：**Python；爬虫；可视化处理；二手房

1. **绪论**
2. **研究问题的提出**

1.用Python语言自行编写爬虫框架，爬取链家的二手房相关数据，按照标题、朝向、地区、价格等信息进行分类和显示。

2.在此基础上可以进一步获取用户反馈信息，评价不同二手房的优劣。

1. **研究的背景、目的和意义**
2. **研究背景**

网络爬虫是可以自动抓取网页上各种类型数据的程序。网络爬虫使用网页的链接地址查找网页的内容，并直接返回用户所需的数据，而无需手动操纵浏览器来获取它。爬虫是搜索引擎中的重要组成部分，可在互联网上为搜索引擎抓取数据。通用搜索引擎（例如谷歌和百度)等逐渐成为人们访问互联网的入口。但是，由于它们的通用性限制，在抓取网页时没有针对性，因而也无法对其抓取的结果进行针对特定领域的进一步分析，导致查询结果不够深入和专业化;另外，一般的搜索引擎通常会返回一些与用户所寻找的主题无关的结果，从而造成信息过载。

本文提出的爬虫程序通过模拟登录二手房网站，获取相关数据，并将这些数据保存到本地，为进一步的数据挖掘和分析提供了方便。最后，本文使用Python可视化方法对数据进行简单的数据分析。

1. **目的和意义**

利随着互联网的高速发展、数据库存储技术的成熟，高性能的存储设备和存储介质日益普及，人们在生活、工作等产生的数据量以指数形式爆炸式增长，大数据发展势不可挡，但如何利用大数据分析为人们的生活提供便利就成了人类的共同话题。同时由于人们的住房压力越来越大，新房价格一般来说更高，因此有些人会考虑购买二手房。所以有必要探讨二手房价格背后的影响因素。

大数据时代的到来，让人们可以在互联网上获得越来越多的信息，其中包括二手房信息。但网络数据来源广泛，且数据量庞大，为了把网页上的数据下载到本地，且节约时间，网络爬虫技术便应运而生。本文选择链家网广州二手房为数据来源，并利用python对数据进行爬取，爬取的字段包括发布的二于房标题、位置、楼层。修建年份.面积.朝向、户型、总价、单价等，一共九个字段。最后对获得的数据进行数据清洗、可视化。分析广州二于房楼层、面积等对房价的影响情况。

1. **系统结构**

**（一）应用技术介绍**

**1.网络爬虫**

网络爬虫(Web Crawler )是基于http协议中的请求与响应机制，按照一定的规则自动提取Web 网页的应用程序或者脚本，可以完整爬取整个网页的源代码3。开发者通过相关设备发送请求并传输url到目标服务器，目标服务器通过相关处理和优化分析后返回访问者所需的信息，访问者得到相关HTML信息后可利用Beautiful Soup、正则表达式、Xpath等技术获取想要的网页固定信息。

**2. Pandas**

Pandas是用于Python的数据分析包。它包含许多高级数据结构和操作工具，可以让人们使用Python进行数据分析更加轻松快捷。

**3.Numpy**

Numpy是使用Python进行科学计算的基础包。它包括一个功能强大的N维数组对象Array。它具有成熟的函数库，例如矩阵数据类型和矢量处理等。

**4.Matplotlib**

数据可视化是数据分析的结果，可以更直观，更优雅地显示结果。Matplotlib库是Python数据可视化的重要类库之一。它包含大量的数据可视化资源，包括地图，3D等。可以将数据以图表形式呈现出来，更为直观地展示给用户，进而将数据的特征显示出来。

Matplotlib有四个部分:

1. Matplotlib的基本figure类型;
2. 调整figure和样式以及颜色;
3. 在图形上添加注释(包括坐标轴范围，长宽比或坐标轴等);

(4) 其他复杂图形。

1. **具体实现**
2. **数据来源分析**

确定自己采集的数据是什么，并且这些数据可以从哪里获取到。

在众多的房地产网站中，本文选取较为权威的“链家”房地产网站爬取广州市的新售楼盘信息，主要爬取一些影响楼盘价格的结构化数据，包含楼盘名称、楼盘位置、楼盘均价格和楼盘面积。

1. **数据爬取过程**

首先通过链家网址进入目标网站，通过源码查看特定信息标签，如图1所示，其次对目标网站的结构进行分析，根据标签信息获取目标数据，最后根据目标网址 requests.get(url)获取数据。



图1 目标网站结构分析

通过以下代码把获取的HTML字符串转化成selector对象：

title = selector\_1.css('.main::text').get() # 标题

selling\_point = selector\_1.css('.sub::text').get() # 卖点

price = selector\_1.css('.price .total::text').get() # 总价

unitPrice = selector\_1.css('.unitPrice .unitPriceValue::text').get() # 单价

house\_type = selector\_1.css('.room .mainInfo::text').get() # 户型

subInfo = selector\_1.css('.room .subInfo::text').get().split('/') # 楼层

floor = subInfo[0] # 楼层

num = re.findall('\d+', subInfo[1])[0] # 共有楼层数

furnish = selector\_1.css('.type .subInfo::text').get().split('/')[-1] # 装修

face = selector\_1.css('.type .mainInfo::text').get() # 朝向

date = re.findall('\d+', selector\_1.css('.area .subInfo::text').get()) # 建造时间

selector\_1.css('.area .mainInfo::text').get().replace('平米', '') # 面积

community = selector\_1.css('.communityName .info::text').get() # 小区

areaName\_info = selector\_1.css('.areaName .info a::text').getall() # 区域

areaName = areaName\_info[0] # 所属区

region = areaName\_info[1] # 区域

scale = selector\_1.css('div.content ul li:nth-child(10)::text').get() # 梯户比例

elevator = selector\_1.css('div.content ul li:nth-child(11)::text').get() # 是否有电梯

houseProperty = selector\_1.css('div.content li:nth-child(2) span:nth-child(2)::text').get() # 房屋属性代码实现

数据采集的过程如图2所示

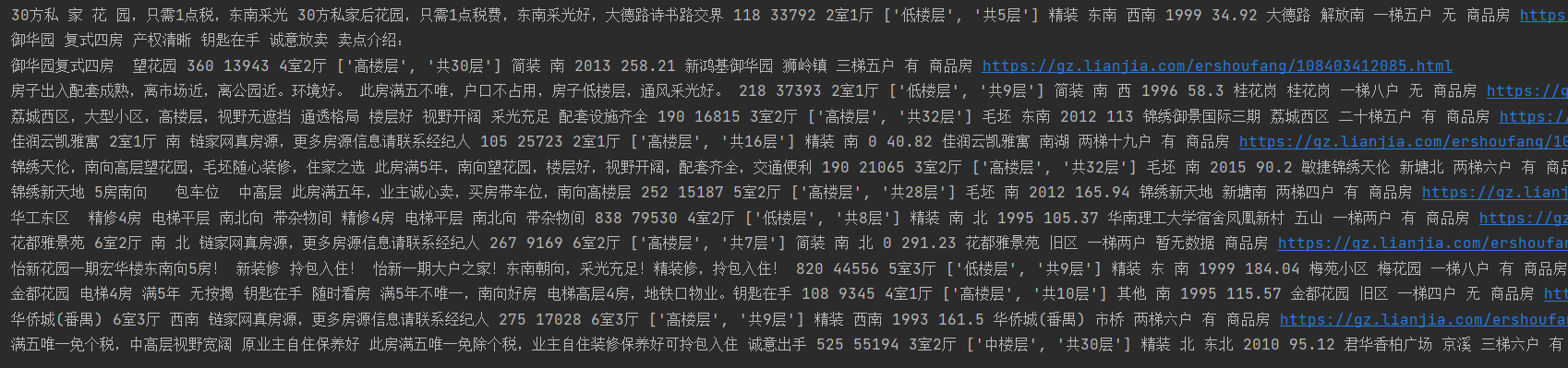


图2数据采集过程

这样就得到当前页的楼盘数据，采用for循环进行翻页爬取操作，提取出目标字段并保存中间网页的数据结果，将获取的数据执行语句f = open('二手房多页gz11.csv', mode='a', encoding='utf-8', newline='')写入到csv文件。

1. **数据清洗**

数据清洗是对所爬取的数据进行规范化操作，将数据中多余的符号及字符做删除或者替换处理，并且通过使用正则表达式等工具对数据进行一致性处理的过程。由于爬取出来的数据很乱，有各种换行符，制表符以及空格隐藏在数据中，或者是字段出现错位，存在缺失值异常值等情况，因此需要先将数据进行清洗补充缺失数据等数据处理再进行探索分析。

在数据分析前需要对数据进行检查，查看是否有重复数据和空缺值数据。使用第三方库panda读取数据df = pd.read\_csv('二手房多页gz.csv', encoding='gbk')，对数据进行读取操作后发现房子是否有电梯存在空缺值Nan，采用语句df['是否有电梯'].fillna('暂无数据',inplace=True)清除未能正常显示的数据，如图3所示，从而得到一份完整的数据。



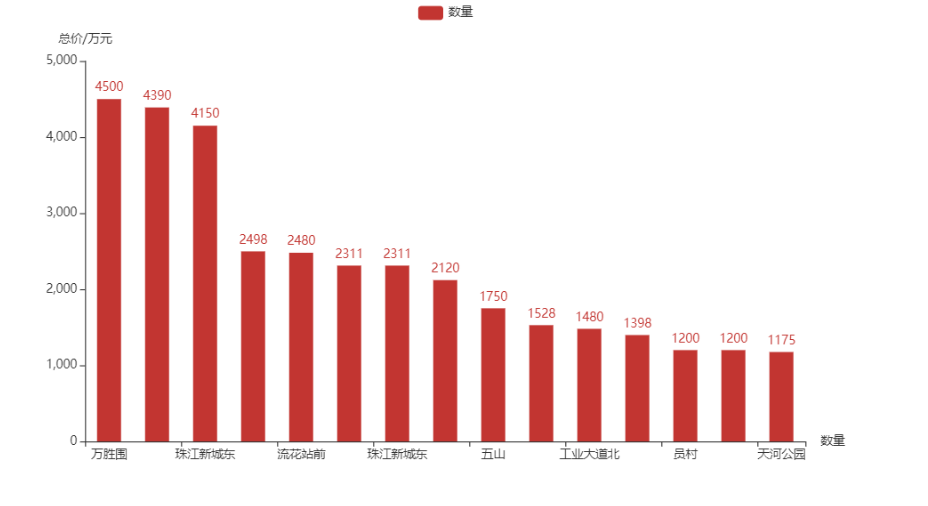
**图3 清楚空缺值nan**

1. 数据可视化分析

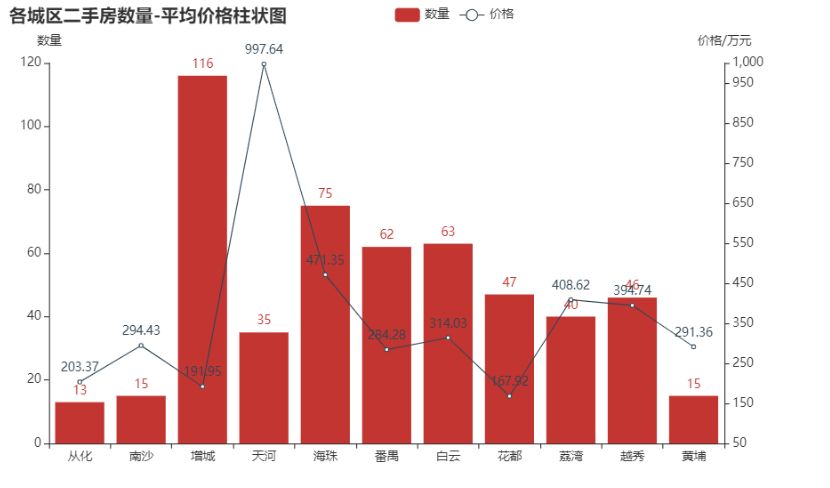
数据可视化技术将数据转换成图形图表，为决策提供依据。数据可视化技术的研究已得到了快速发展并取得相应的成就

为了更好地呈现广州市的房价变化趋势，本文采用Python的第三方库 Matplotlib把数据中蕴含的信息以图表的形式展示出来，

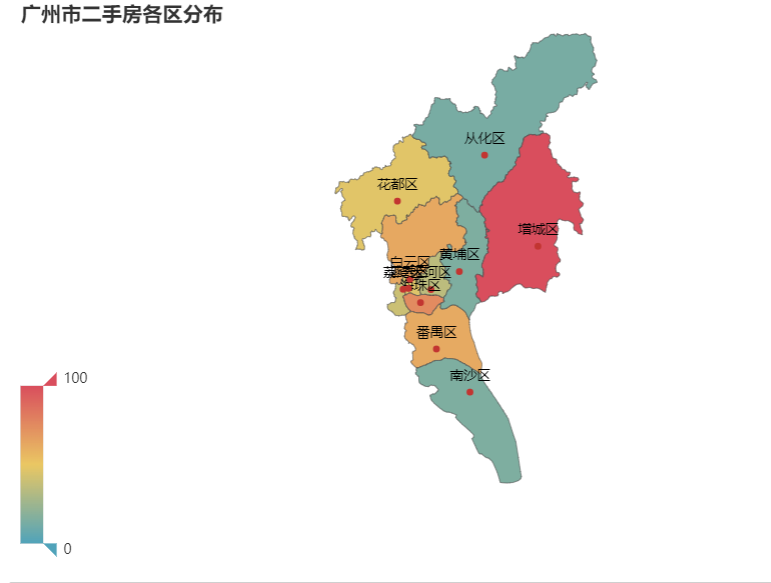
通过Bar和Map函数对广州市每个区域二手房数量进行分析并且分析每个区域二手房总价均值,可视化结果分别如图4图5图6。



**图4 广州各区域二手房平均总价**



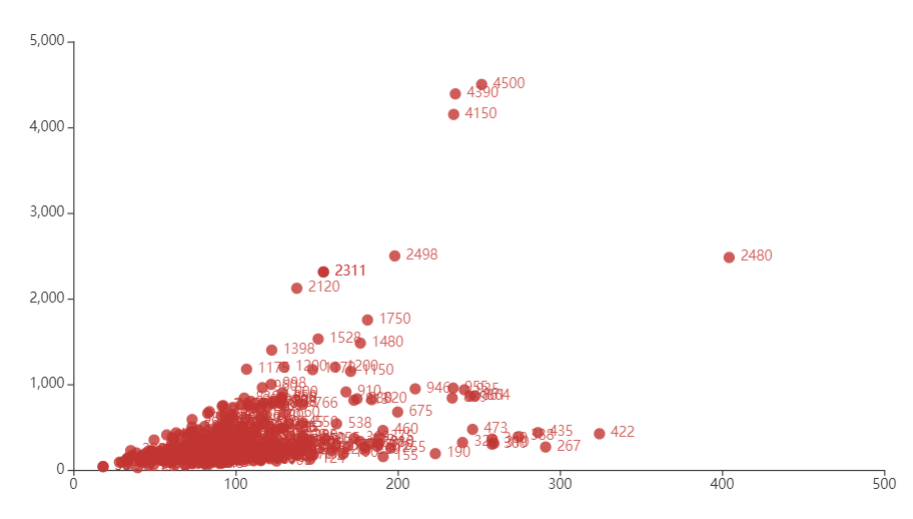
**图5 广州各所属区二手房数量及平均总价**



**图6 广州各所属区二手房数量分布情况**

从图5图6可以看出增城区二手房数量最多，远远高于其他区域，但是房屋每平米价格却不是最高的，天河区二手房每平米单价均值最高，可能是因为天河属于中心位置，同时占据交通和经济两大优势。另外,通过图4的柱状图结果可以得出，武汉各区房屋总价都在5000万以下,而且房屋总价差距比较大，最高的是万胜围，房屋总价达到了4500万,最低的不到1200万。

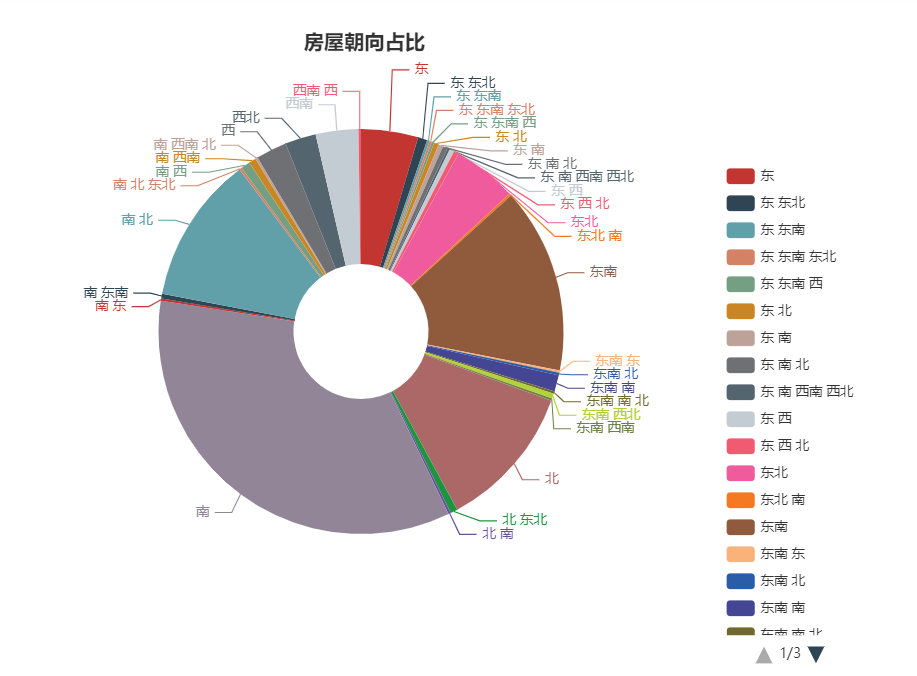
使用Scatter函数绘制散点图对房子的总面积和总价进行分析如图7所示：



**图7 房子面积以及总价分布情况**

通过图7可以看出大部分房子的面积都处于100平米左右内，少部分房子超过200平米。且大部分房子的价格都低于1000万。面积和价格整体上成线性关系。

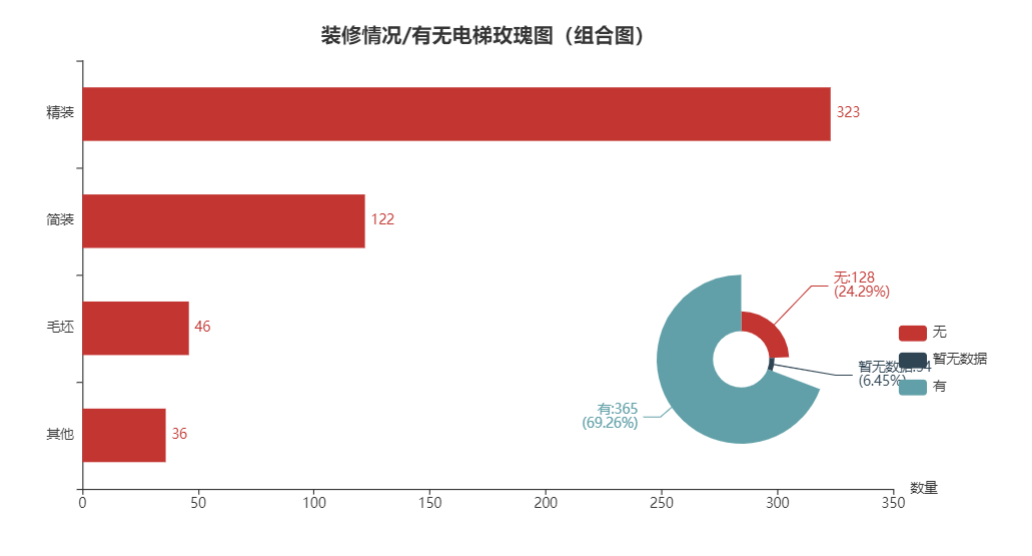
使用Pie函数绘制房屋朝向比饼状图如图8所示：



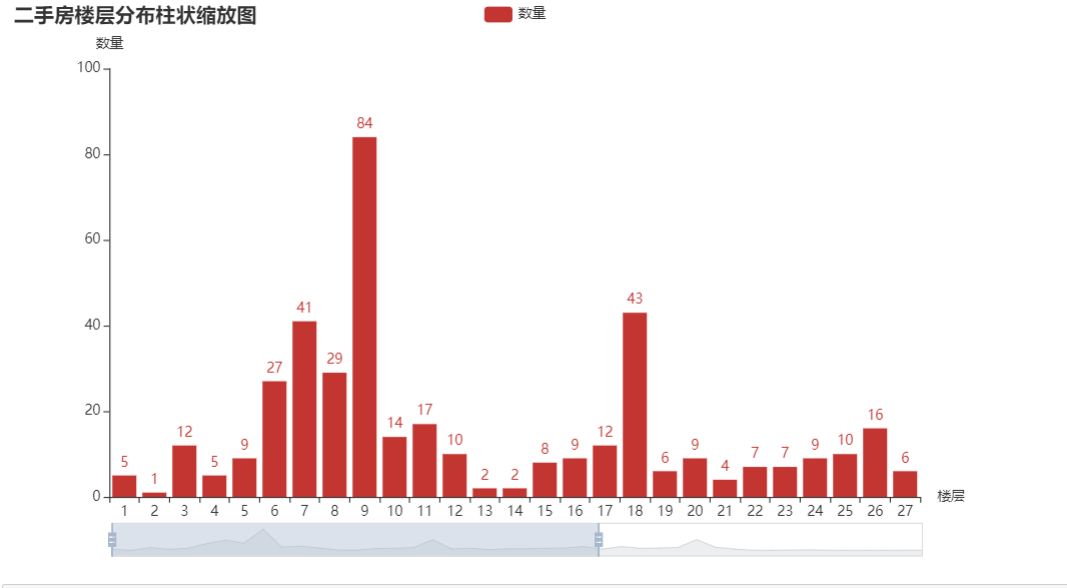
**图8 房屋朝向占比**

可以看出广州大部分二手房朝向都是南

最后用Bar函数和Pie函数绘制房屋装修情况/有无电梯玫瑰图和二手房楼层分布柱状图。如图9和图10所示：



**图9 装修情况/有无电梯玫瑰图**



**图10 二手房楼层分布情况**

通过观察图9可知，广州市精装二手房数量最多，简装次之，毛坯较少，有电梯的二手房远多于没有电梯的。

通过观察图10可知，大多数楼房都是9楼。9楼兼顾了高层住宅和高层住宅优点，既能够有不错的自然通风光照标准，又有较为安稳，整体生活的质量比较高。

为了探索二手房标题后的规律,使用jieba分词,对标题进行切分,再用可视化的方法，画出标题的词云图。由图11可知二手房为了更快的卖出房屋，标题通常包含户型方正，近地铁，看房方便等关键词。



**图11 标题词云图**

1. **总结**

本文基于Python语言采用网络爬虫技术获取链家网站的楼盘数据，对广州市的楼盘数据进行爬取、存取、清洗和可视化，并把影响房价的要索以一种合理的形式呈现出来。为人们发布房源，购买二手房提供了信息。与传统分析方法相比，大数据分析技术在数据采集、数据挖掘和可视化应用方面的优势明显。

参考文献

[1]潘晓英,陈柳,余慧敏,等.主题爬虫技术研究综述[J.计算机应用研究,2020,37(4):961-965.

[2]闫志国,宛楠,严迪，等.基于Scrapy爬取电商药品数据及数据可视化分析处理[J.轻工科技,2021,37(3):98-100.

[3]文鹏,袁小艳.基于Python的招聘信息爬取和分析[J].信息与电脑,2021,33(9):65-67.

[4]晋振杰,曹少中,项宏峰,等.基于python的电商书籍数据爬虫研究[J.北京印刷学院学报,2018,3(26): 39-42.

[5]曾悠．大数据时代背景下的数据可视化概念研究[D].杭州:浙江大学,2014.

**课程论文成绩评定表**

|  |  |  |  |
| --- | --- | --- | --- |
| 评价项目 | 评价内容 | 分值 | 得分 |
| 1.创新性、合理性 | 5分：表明设计具有创新性、合理性；  3-4分：表明设计具有新意、合理性；  1-2分：表明设计具有合理性。 | 5 |  |
| 2.选题难度、实现复杂性 | 9-10分： 难度系数高，设计实现功能复杂；  7-8分：难度系数低，设计实现功能较复杂；  6分及以下：难度系数很低，设计实现功能一般。 | 10 |  |
| 3.完成情况包括数据分析技术的应用、工作量 | 52-60分：有完整的符合标准的文档，有正确的数据分析技术设计方案，设计较全面，优化得当；  44-51分：有完整的符合标准的文档，有正确的数据分析技术方案，设置无明显错误；  36-43分：有完整的符合标准的文档，有基本的数据分析技术，设计方案正确，但有少数失误；  36分以下：有相对完整的符合标准的文档，有基本的数据分析技术，设计方案基本正确，但缺少部分必要功能，存在问题。 | 60 |  |
| 4.设计报告 | 23-25分：优秀。表明报告清楚详细地说明了设计的过程，格式规范、内容完整，阐述清晰，层次分明；  19-22分：良好。表明报告说明了设计的过程，格式较规范、内容比较完整，阐述清晰，有层次；  15-18分：中。表明报告说明了设计的过程说明不太清晰，格式不规范、内容比较简单；  15分以下：差。表明报告过于简单。 | 25 |  |
| 5.答辩 | 90-100分：优秀。思路清晰，表达清楚，设计正确、有创意；  80-89分：良好。思路清晰，有层次，设计正确；  70-79分：中。表达正确，内容比较简单，有少量错误；  60-69分：及格。表达不太清晰，有错误存在，能完成基本答辩；  60分以下：差。表达不清晰，设计没完成或错误。 | 100 |  |
| 实际得分合计（（评价项目1+评价项目2+评价项目3+评价项目4）\*0.5+评价项目5\*0.5） | | |  |