



河海大学

# 《多元统计分析与 R 建模》 课程作业报告

学号姓名 \_\_\_\_\_ 李宗霖

专业年级 \_\_\_\_\_ 20 级大数据管理与应用

指导老师 \_\_\_\_\_ 韦庆明

报告主题 \_\_\_\_\_ 基于电影数据的电影个性化推荐

2022 年 12 月

# 基于电影数据的电影个性化推荐

## 目录

一、 选题背景.....	3
二、 选题目的.....	4
三、 指标确定.....	5
3.1 通过词云图分析 .....	5
3.2 通过电影数据分析 .....	6
3.2.1 数据介绍.....	6
3.2.2 数据预处理.....	7
3.2.3 数据分析.....	7
四、 个性化电影推荐.....	9
4.1 数据爬取与处理 .....	9
4.1.1 数据爬取.....	9
4.1.2 上映时间转为时间戳.....	9
4.1.3 上映国家和地区提取.....	9
4.1.4 电影类型与时长.....	10
4.1.5 评论数据.....	10
4.1.6 导演与演员数据.....	10
4.2 因子分析 .....	11
4.3 聚类分析 .....	14
4.4 多元线性回归分析 .....	16
五、 课程心得.....	17
六、 附录.....	18
6.1 Python 程序 .....	18
6.2 R 项目 .....	18
6.3 原始数据与编号 .....	18
6.4 情感分析结果 .....	18
6.5 因子分析结果 .....	18
6.6 聚类分析结果 .....	18
七、 参考文献.....	18

## 一、选题背景

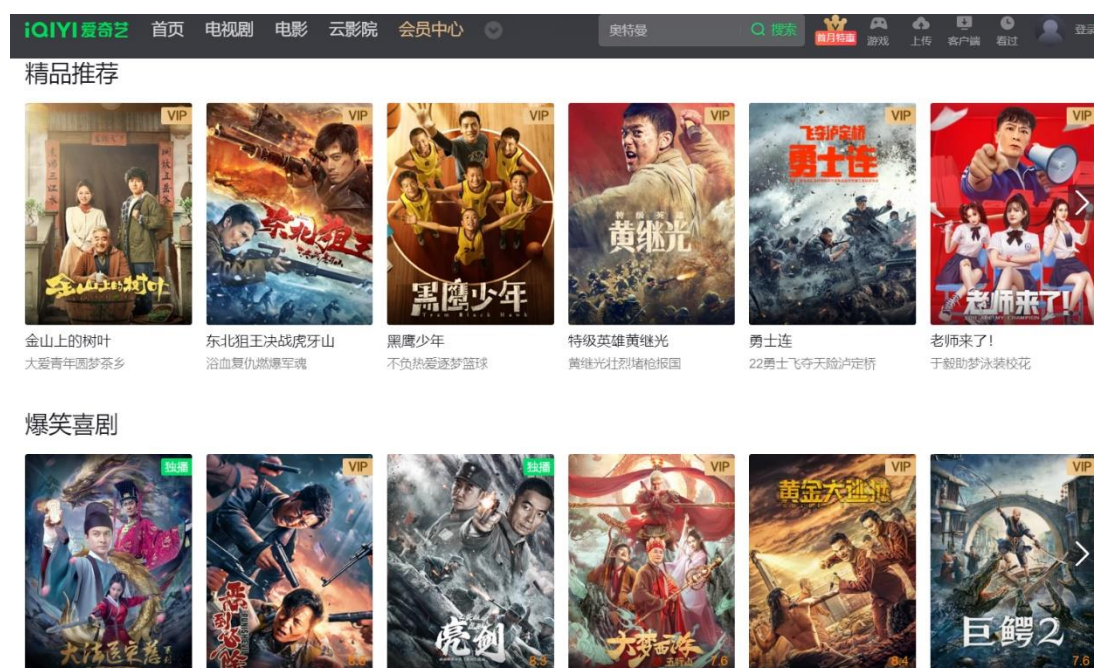
电影，也被称为运动画面或动态画面，即“映画”，是作品视觉艺术形式，通过使用移动图像来表达沟通思想，故事，认知，情感，价值观，或各类大气模拟体验。这些图像通常伴随着声音，很少有其他感官刺激。

在 1911 年意大利诗人和电影先驱者乔托·卡努杜发表的《第七艺术宣言》著论发表后，电影被世人普遍认可，并被列为位列文学、戏剧、绘画、音乐、舞蹈、雕塑之后的第七艺术。

伴随着摄影技术、后期制作技术的不断发展，电影行业进步神速，电影也被越来越多的人所接受，但与此同时，一些电影行业的问题也暴露出来。其中又以以下两个问题最为关键：

### （1）电影数量繁多、良莠不齐

随着人们生活水平的进步，电影不再是普通人消费不起的高端娱乐项目，电影行业变得热门。同时，电影学院培养出的大量导演、演员等人才的涌现，电影行业门槛的不断降低，使得许多小制作和网络电影出现在各大影视平台，这使得电影的质量参差不齐。



图一：爱奇艺网络电影

很多网络电影通过博人眼球的爽文情节、美女图片等内容，将不明所以的看客吸引点击，从而浪费他们生命的几个小时。即使是阅片无数的老饕，也难免被其迫害。

## （2）个性推荐仍显不足

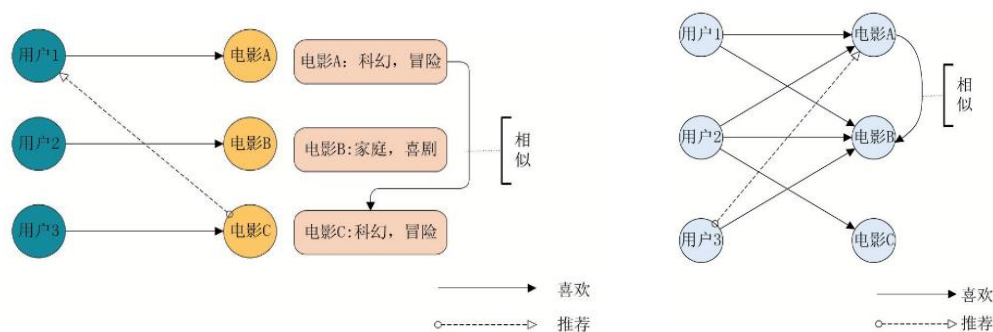
各大电影平台的热门和高分电影并不全是优质电影，榜单存在商业的行为。而根据用户喜好推荐的电影，相关性差，很难找到想看的影片。



图二：爱奇艺电影风云榜

继续以爱奇艺榜单为例，它主打的电影风云榜，并不会通过用户已观看过的电影进行个性化推荐，反而是将显眼的板块给商业化气息严重的网络电影和院线电影（正在电影院上映的电影）通过独播等垄断行为，来收取 VIP 观影费用或点播费用。

此外，现有的电影个性化推荐模型本身存在一定的不足。



图三：基于内容推荐（左一）和基于物品的协同过滤推荐（右一）

主流的两种个性化推荐模型为：基于内容推荐和基于物品的协同过滤推荐。总的说来，都是根据相同用户的行为的相似度来进行推荐，很少有直接根据用户喜好进行推荐。原因也较为容易理解，主要是用户的观影地点过于分散，每个平台的数据都不完整，同时数据量不够支持进行个性化推荐系统模型的构建。

## 二、选题目的

本选题的目的主要有以下三点：

### （1）对电影数据进行分析

寻找电影评分和电影类型、电影时长、电影上映国家等因素的关系。确定会对观影者评分电影环节的关键指标，方便下文分析。

## （2）分析个人的观影数据

寻找个人喜欢的电影与其类型、电影时长、上映国家等其他因素的关系，通过聚类算法，将本人所喜欢的电影进行归类分析，查看这些电影的共性，方便之后电影的选择。

### (3) 实现电影的个性推荐

根据前两步过程，确定影响本人对电影喜好打分的评价指标，通过指标构建推荐模型，最终确定推荐哪些电影。

### 三、指标确定

### 3.1 通过词云图分析

下面是通过以电影为关键词进行检索，得到的相关结果进行的词云分析图。



图四：电影词云图

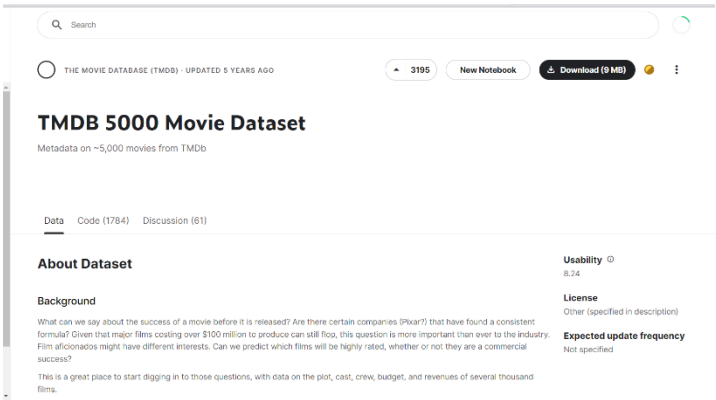
从词云中可以看到，电影艺术不分国界，中国、韩国、美国等多个国家的电影艺术，都是出现的高频词。同时，对于电影的上映时间，大家也同样关注，从图中的 2017、2018 等年份关键词中可以看出。此外，大家还关注了电影的类型，从图中的“爱情”、“经典”、“动画”等高频词可以看出。还有观影的方式，如“在线”、“下载”、“手机”、“电影院”等。这些指标都是在之后的分析中需要重点考虑的对象。

3.2 通过电影数据分析

为进一步确定上述指标和其他因素对电影评分、个人喜好的影响，本文通过收集电影数据，并对数据进行量化分析，以得出更为具体的上述指标的关系。

3.2.1 数据介绍

本部分的电影数据来自互联网电影资料库（Internet Movie Database，简称IMDb）隶属于亚马逊公司旗下网站。IMDb 是一个关于电影演员、电影、电视节目、电视明星和电影制作的在线数据库，包括了影片的众多信息、演员、片长、内容介绍、分级、评论等。对于电影的评分使用最多的就是 IMDb 评分。TMDB 是专门管理相关数据的平台，可提供 API 接口。



图五：数据集网站展示

选取了最为广泛使用的 5000 条热门电影的数据，原始数据文件包括 credits 与 movies 两个 CSV 文件。

字段名	意义
id	标识号
popularity	在 Movie Database 上的相对页面查看次数
budget	预算（美元）
revenue	收入（美元）
original_title	电影名称
cast	演员列表
director	导演列表，按
keywords	与电影相关的关键字
overview	剧情摘要
runtime	电影时长
genres	风格列表，按
production_companies	制作公司列表，按
release_date	首次上映日期
vote_count	评论次数
vote_average	平均评分
release_year	发行年份

图六：数据指标展示



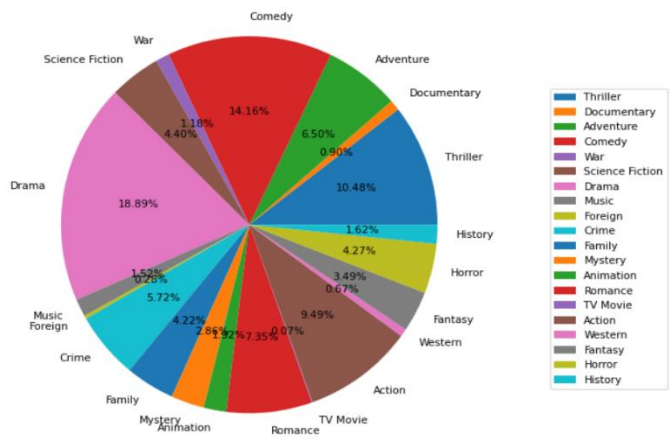
3.2.2 数据预处理

使用 python 对数据进行处理，对两个表格中的数据进行数据源处理、数据预处理（表格合并、去重、去除多余数据列、缺失值处理）、数据格式处理。处理合并后数据指标如图六所示。

3.2.3 数据分析

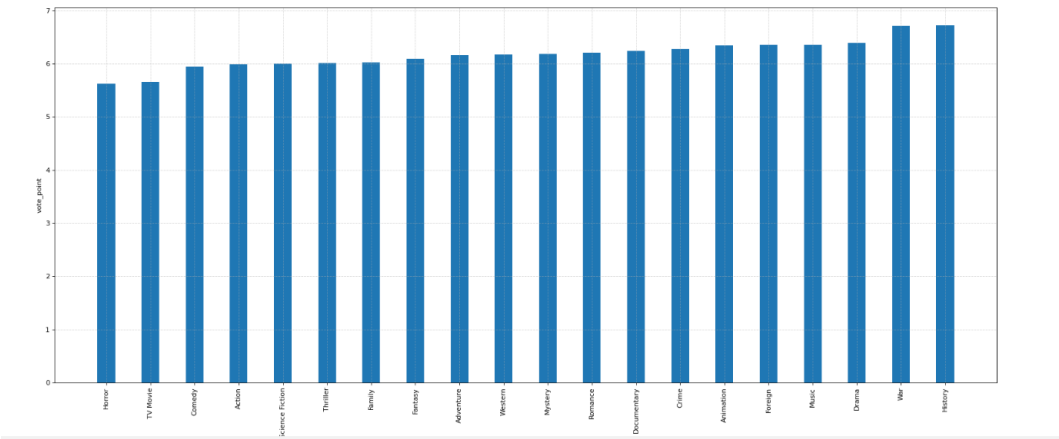
通过 python，对数据进行关系判断。

(1) 电影类型与评分



图七：电影类型分布

通过上图，可以看出 5000 部电影的类型分布占比，发现电影类型众多，以戏剧、喜剧占比最多。

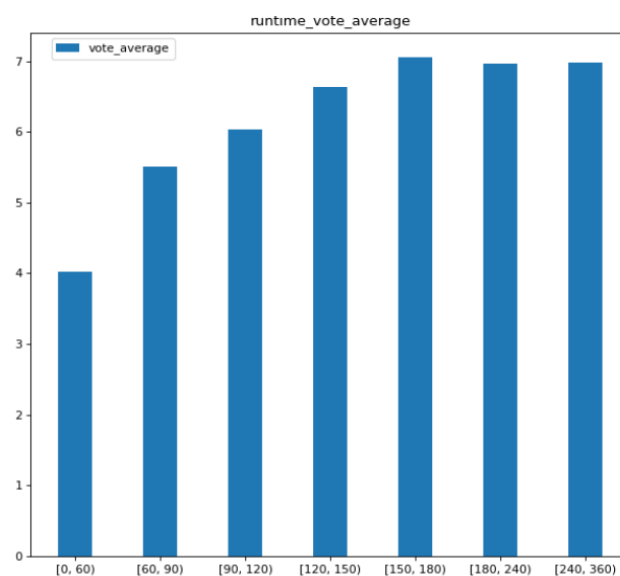


图八：电影类型与平均评分关系图

可以发现电影的类型与电影最终平均评分有较大关系，随着电影类型的切换，最终得分有 1 分多的差异（评分为 5000 部电影的平均得分，因此 1 分多的差距已经非常大了）。

## (2) 电影时长与评分的关系

同样使用 python 进行画图分析，得出电影时长和评分之间的关系图

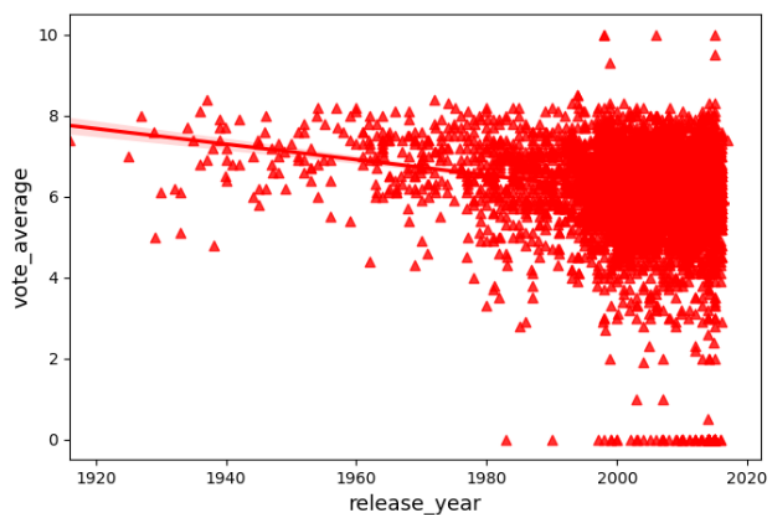


图九：电影时长与评分的关系

可以看出，0~60 分钟的短篇电影的评分普遍不高而 2 小时以上的电影评分较高。

## (3) 上映时间与评分关系

通过 python 做图，可以观察出电影的上映时间与评分之间的关系：



图十：电影上映时间与评分关系

由图可以看出，评分与上映时间存在一定关系，由于上文所说的，近年电影良莠不齐，导致了超低分评价的出现。

基于以上两部分内容，最终决定将电影上映时间、类型、评价、时长、上映国家、评分作为个性化推荐标准的指标。



四、 个性化电影推荐

本部分将使用按照上述指标从豆瓣电影网站上爬取到本人的 500 条观影数据进行综合分析处理。

4.1 数据爬取与处理

4.1.1 数据爬取

使用八爪鱼对个人豆瓣电影网站进行数据爬取。具体见附录。



图十一：豆瓣电影网

#	A	B	C	D	E	F	G
1	电影名	上映时间	导演	编剧	主演	类型	时长
2	健听女孩 / C 2021-01-29(圣丹斯电影节)	夏安·海德	夏安·海德 / 维多利亚·贝多斯 / 斯坦尼	艾米莉亚·琼斯 / 特洛伊·科特纳尔 / 玛丽·	剧情 / 音乐		111分钟
3	反贪风暴5 / 2021-12-31(中国大陆)	林德禄	黄智华	古天乐 / 张智霖 / 郑嘉颖 / 宣萱 / 曾志伟	动作 / 犯罪		95分钟
4	科学怪狗 / P 2012-10-05(美国)	蒂姆·波顿	约翰·奥古斯特 / 蒂姆·波顿	马丁·肖特 / 凯瑟琳·欧哈拉 / 马丁·兰道	喜剧 / 科幻 / 动画 / 恐怖		87分钟
5	误杀2 2021-12-17(中国大陆)	戴墨	李鹏 / 刘吾驷 / 杨梅媛	肖央 / 任达华 / 文咏珊 / 陈雨锶 / 宋洋	剧情 / 犯罪		118分钟
6	狙击手 2022-02-01(中国大陆)	张艺谋 / 张末	陈宇 / 张艺谋	陈永胜 / 章宇 / 张译 / 刘奕铁 / 曹炎 / 王	剧情 / 历史 / 战争		96分钟
7	七宗罪 / Se7 1995-09-22(美国)	大卫·芬奇	安德鲁·凯文·沃克	摩根·弗里曼 / 布拉德·皮特 / 凯文·史派	剧情 / 悬疑 / 惊悚 / 犯罪		127分钟
8	少林足球 2001-07-12(中国香港)	周星驰	周星驰 / 曾谨昌 / 冯勉恒 / 冯志强 / 卢	周星驰 / 赵薇 Wei Zhao / 吴孟达 / 谢贤 /	喜剧 / 动作 / 运动		113分钟 /
9	独立日 / Independence Day 1996-07-02(加拿大)	罗兰·艾默里奇	罗兰·艾默里奇 / 迪安·德夫林	威尔·史密斯 / 杰夫·高布伦 / 比尔·普尔	动作 / 科幻 / 冒险		145分钟 /
10	师父 2015-11-11(中国大陆)	徐浩峰	徐浩峰	廖凡 / 宋佳 / 蒋雯丽 / 金士杰 / 宋洋 / 廖	剧情 / 动作 / 武侠		109分钟
11	空天猎 2017-09-29(中国大陆)	李晨	李晨 / 张力 / 刘毅 / 高岩 / 褚翔宇	李晨 / 范冰冰 / 王千源 / 李佳航 / 赵达 /	剧情 / 动作 / 战争		115分钟
12	爱宠大机密 / 2016-07-08(美国)	克里斯·雷纳德 / 亚罗·切尼	辛科·保罗 / 肯·道里欧 / 布莱恩·林奇	路易·C·K / 艾瑞克·斯通斯崔特 / 凯文·哈	喜剧 / 动画		87分钟
13	一个叫欧维的男人决定去死 2015-12-25(瑞典)	汉内斯·赫尔姆	汉内斯·赫尔姆 / 弗雷德里克·巴克曼	罗夫·拉斯加德 / 巴哈·帕斯 / 托比亚斯·	剧情 / 犯罪		116分钟
14	小人物 / Not 2021-03-26(美国)	伊利亚·奈舒勒	德里克·科尔塔	约翰·塞纳 / 阿列克谢·谢列布罗夫	动作 / 犯罪		92分钟
15	白蛇传·情 2019-10-18(平遥电影节)	张险峰	莫菲	曾小敏 / 文汝倩 / 朱红星 / 王燕飞	剧情 / 爱情 / 戏曲		101分钟
16	俄罗斯方块 / Tetris 2020-02-22(柏林电影节)	瓦迪姆·佩尔曼	伊尔佳·佐芬 / 沃尔夫冈·科尔哈泽	纳威尔·佩雷兹·毕斯卡亚特 / 拉斯·米	剧情		127分钟
17	寄生虫 / Parasite 2019-05-21(戛纳电影节)	奉俊昊	奉俊昊 / 韩进元	宋康昊 / 李善均 / 裴斗娜 / 李东旭 / 朴	剧情		132分钟
18	荒岛余生 / Castaway 2000-12-22(美国)	罗伯特·泽米吉斯	小威廉·保尔斯	汤姆·汉克斯 / 海伦·亨特 / 克里斯·诺	剧情 / 冒险		143分钟
19	成事在人 / The Man Who Would Be King 2009-12-11(美国)	克林特·伊斯特伍德	安东尼·佩卡姆 / 约翰·卡林	摩根·弗里曼 / 马特·达蒙 / 托妮·戈	剧情 / 传记 / 历史 / 运动		143分钟
20	八恶人 / Hateful Eight 2015-12-07(洛杉矶首映)	昆汀·塔伦蒂诺	昆汀·塔伦蒂诺	塞缪尔·杰克逊 / 库尔特·拉塞尔 / 詹妮	剧情 / 犯罪 / 西部		168分钟
21	扬名立万 2021-11-11(中国大陆)	刘循子墨	里八神 / 刘循子墨 / 张本煜 / 柯达	尹正 / 邓家佳 / 喻恩泰 / 杨皓宇 / 陈明	剧情 / 喜剧 / 悬疑		123分钟
22	长津湖 2021-09-30(中国大陆)	陈凯歌 / 徐克 / 林超贤	兰晓龙 / 曹建新	吴京 / 易烊千玺 / 陈伟霆 / 朱亚文 / 李	剧情 / 历史 / 战争		176分钟
23	低俗小说 / Pulp Fiction 1994-05-12(戛纳电影节)	昆汀·塔伦蒂诺	昆汀·塔伦蒂诺 / 罗杰·阿夫瑞	约翰·特拉沃尔塔 / 乌玛·瑟曼 / 阿曼	剧情 / 喜剧 / 犯罪		154分钟
24	送你一朵小红花 2020-12-31(中国大陆)	韩延	韩延 / 韩今谅 / 贾佳磊 / 于勇 / 李晗	易烊千玺 / 刘浩存 / 朱媛媛 / 高亚麟 / 夏	剧情		128分钟
25	人之怒 / Wrath 2021-04-02(俄罗斯)	盖·里奇	尼古拉斯·布赫里夫 / 艾瑞克·贝斯纳 / 盖	杰森·斯坦森 / 戴夫·弗兰科 / 罗奇 / 威	动作 / 犯罪		118分钟 /
26	白蛇2：青蛇劫起 2021-07-23(中国大陆)	李豪	辛晓 / 吴宇森 / 梁淑华	唐小喜 / 董璇 / 魏超 / 赵乾 / 郑小璞 /	动画 / 奇幻 / 冒险		131分钟
27	英雄本色 1986-02-02(中国香港)	吴宇森	陈庆嘉 / 吴宇森 / 梁淑华	周润发 / 狄龙 / 张国荣 / 朱宝意 / 李子	剧情 / 动作 / 犯罪		95分钟
28	何以为家 / Howl 2019-05-17(戛纳电影节)	娜丁·拉巴基	娜丁·拉巴基 / 吉哈德·霍加里 / 米歇尔·	赞恩·阿尔·拉菲亚 / 约丹诺斯·希费	剧情 / 动作 / 犯罪		126分钟 /
29	拉方证人 / Witness 1985-12-12(美国)	比利·怀尔德	阿加莎·克里斯蒂 / 比利·怀尔德 / 哈里	哈里·贝瑞 / 安妮·贝内特 / 查理·辛 / 乔	剧情 / 悬疑 / 犯罪		116分钟

图十二：爬取数据展示

通过 Excel 自带功能，对爬取到的数据进行初步的整理，得到需要数据的初步结果，如图十二。

4.1.2 上映时间转为时间戳

首先将上映时间转化为 yyyy-mm-dd hh:mm:ss 的统一格式，然后使用公式：“=(时间-70\*365-19)\*86400-83600” 进行时间戳格式转化。具体见附录。

4.1.3 上映国家和地区提取

从上图实例数据中可以看出，上映时间和上映国家及地区被绑定在了一起，本部分使用 Excel 自带的筛选、分列功能，将上映地区单独成列，并进行编号处理，是文本数据量化，后续可以继续分析。具体见附录。

#### 4.1.4 电影类型与时长

本部分处理方式雷同，时长方面，选择了首映时长，舍去了重映、点播删减版时长。电影类型方面，由于大多数电影类型不止 1 个类型，甚至有的电影有 5 个以上的类型，因此在类型选择方面，使用了最具代表性的两个类型作为电影的类型，对于第二类型不存在的电影，则将其设置为 0。最后，对所有的文本型类型数据进行编号处理，以便后续运算。具体见附录。

#### 4.1.5 评论数据

本文为了最大程度上还原评论数据对评分的影响，选取了豆瓣电影中给出的高分短评数据（每部电影选择五条），随后进行文本情感分析。



图十三：豆瓣电影短评数据

这里的高分短评并不意味着对电影全部是正面评价，而是评价本身被多数人认可，因此短评的数据褒贬不一，能够较好地体现他人的客观评价。

经过导入语料搜索上面的关键词，每个正面词加 1 分，每个负面词减 1 分。如果正面词或负面词前面有否定词，得分取反。如果正面词或负面词前面有程度词，得分乘以一个系数，2 或 0.5，具体参照词典。最后将总分相加，得到最终的情感得分。具体见附录。

#### 4.1.6 导演与演员数据

本部分数据本来准备将文本数据通过词袋模型转化为特征向量进行处理，但是由于每部电影演员的名字做完分词后词库太多，2000 多长的向量，实在是无法处理，就放弃了本部分数据。

通过以上步骤对数据进行的处理，将所有数据进行了量化处理，最终得到以下数据（仅显示前 10 条，详见附件）：

	A	B	C	D	E	F	G	H
1	电影名	Q4-豆瓣评分	Q7-类型1	Q2-时长	Q6-国家	Q5-总体情感得分	Q3-上映时间（时间戳格式）	Q8-类型2
2	健听女孩/CODA	8.6	1.0	111.0	1.0	20.0	1611794800.0	1.0
3	反贪风暴5最终	4.6	2.0	95.0	2.0	6.0	1640825200.0	2.0
4	科学怪狗/Fran	7.7	3.0	87.0	3.0	8.0	1349311600.0	3.0
5	误杀2	5.7	1.0	118.0	2.0	3.0	1639615600.0	2.0
6	狙击手	7.7	1.0	96.0	2.0	7.5	1643590000.0	4.0
7	七宗罪/Se7en	8.8	1.0	127.0	3.0	15.0	811644400.0	5.0
8	少林足球	8.1	3.0	113.0	4.0	22.0	994812400.0	6.0
9	独立日/Indepe	8.1	2.0	145.0	5.0	2.0	836182000.0	3.0
10	师父	8.2	1.0	109.0	6.0	55.5	1447116400.0	6.0

图十四：处理后数据

#### 4.2 因子分析

选取指标共 7 个，变量过多，不方便之后的分析，在此，使用因子分析的方式，对变量进行降维处理。

首先，通过对上述变量之间的相关系数矩阵进行分析，对变量进行选取。

```
> cor(movie) #相关系数阵
      Q4-豆瓣评分  Q7-类型1  Q2-时长  Q6-国家  Q5-总体情感得分  Q2-类型2
Q4-豆瓣评分  1.00000000 -0.060250074  0.24390575  0.235493478  0.43514576  0.14345169
Q7-类型1      -0.06025007  1.000000000  -0.13681729  0.004217153  -0.05236465  0.12647320
Q2-时长        0.24390575 -0.136817290  1.00000000  0.127011911  0.18100175  -0.09188079
Q6-国家        0.23549348  0.004217153  0.12701191  1.000000000  0.13807993  0.06365257
Q5-总体情感得分 0.43514576 -0.052364651  0.18100175  0.138079927  1.00000000  0.13276954
Q2-类型2       0.14345169  0.126473198  -0.09188079  0.063652571  0.13276954  1.00000000
```

图十五：基于 R 的变量间的相关系数矩阵

根据变量间的相关性，最终选取了豆瓣评分、类型 1、类型 2、时长、国家、总体情感得分这 6 个变量进行因子分析，而上映时间变量与其他变量的相关性都较弱，所以舍去。

随后，对剩余变量进行 KMO 检验与 Bartlett 检验，观察其数据是否适合做因子分析。

```
> KMO(movie)#KMO检验
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = movie)
Overall MSA = 0.61
MSA for each item =
      Q4-豆瓣评分  Q7-类型1  Q2-时长  Q6-国家  Q5-总体情感得分  Q2-类型2
      0.60          0.58          0.65          0.70          0.61          0.55
> cor.test(cor(movie),n=nrow(movie))#Bartlett检验
$chisq
[1] 211.7893

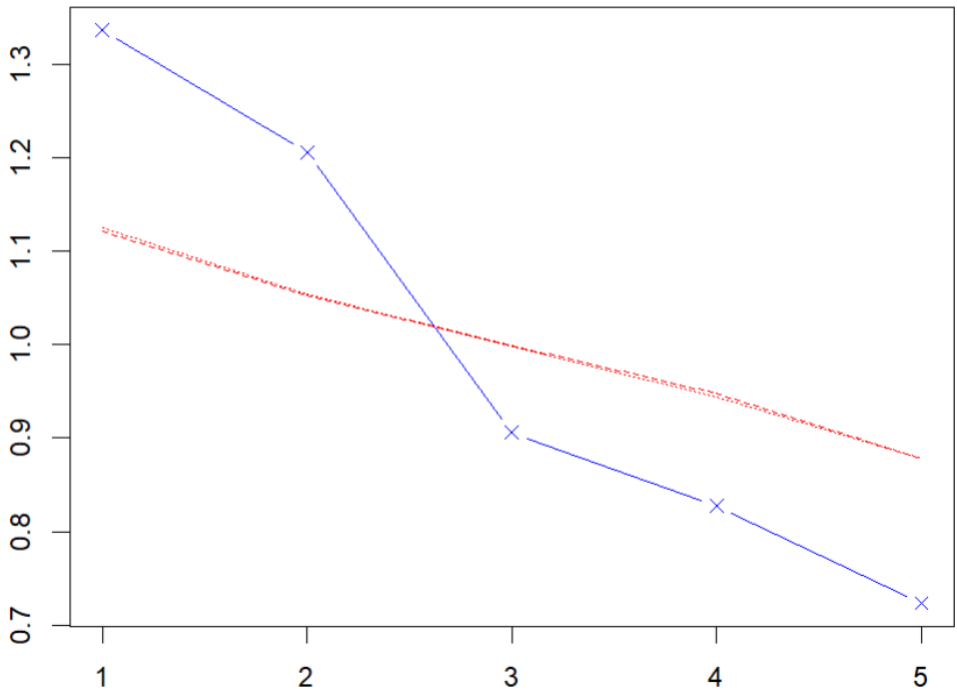
$p.value
[1] 8.459022e-37

$df
[1] 15
```

图十六：KMO 检验与 Bartlett 检验

可以看出，KMO 值大于 0.6，Bartlett 检验的 p 值符合要求，上述变量可以进行因子分析。

随后，通过对数据进行碎石图绘制，来确定因子个数。



图十七：碎石图绘制

根据图中折线变化趋势，可以看到，在 3 之后，变化幅度大大减小，因此因子数选择 3 较为合理。同时，考虑到特征值大于 1，因子数代表的特征值大于均值，因子数选 2 较为合理。

为进一步确定因子个数，本文通过比较方差解释率进行因子选取发现，在因子数为 2 时，对于原始 6 个变量的解释度过低，而因子数为 3 时，对原始变量的解释率较好，因此最终选取的因子个数为 3。

总方差解释						
成分	旋转前方差解释率			旋转后方差解释率		
	特征根	方差解释率(%)	累积方差解释率(%)	特征根	方差解释率(%)	累积方差解释率(%)
1	1.753	29.219	29.219	174.158	29.026	29.026
2	1.221	20.351	49.571	123.265	20.544	49.571

图十八：方差解释率

最后，通过 R 语言分析，进行因子的载荷分析，以对每个因子再次进行细致划分，因子的方差贡献度如图十九所示，可以发现，在没有进行因子旋转前，因子的含义难以解释，所以本文对因子使用最大方差法进行因子旋转。

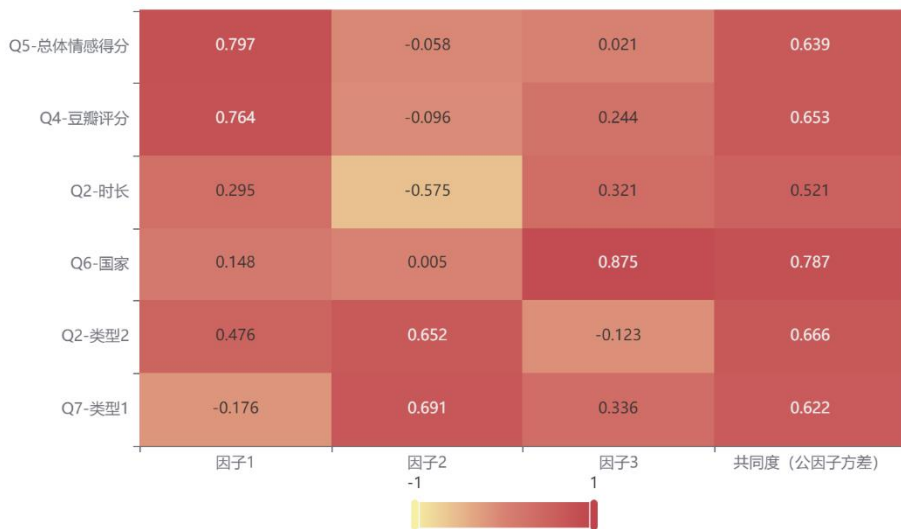
```
> round(Fa1$loadings[1:6,],4)#4 : 保留小数
               Factor1 Factor2 Factor3
Q4-豆瓣评分    0.7271  0.0189 -0.0064
Q7-类型1       -0.0904  0.3178  0.0879
Q2-时长         0.3449 -0.3455  0.0474
Q6-国家         0.3269  0.0076  0.3565
Q5-总体情感得分 0.5958  0.0489 -0.1599
Q2-类型2        0.1855  0.4510 -0.0011
```

图十九：因子旋转前的因子载荷

```
> round(Fa2$loadings[1:6,],4)
               Factor1 Factor2 Factor3
Q4-豆瓣评分    0.6581 -0.0496  0.3058
Q7-类型1       -0.0756  0.3329  0.0196
Q2-时长         0.2436 -0.3682  0.2136
Q6-国家         0.1450  0.0185  0.4611
Q5-总体情感得分 0.6086 -0.0256  0.1092
Q2-类型2        0.2273  0.4286  0.0494
```

图二十：旋转后因子载荷图

对上图进行热力图绘制，以更好地发掘因子含义。



图二十一：热力图

从图中可以发现，因子1中，豆瓣评分与评论的情感得分占比较大，该因子可以解释为“外界评价”；因子2中，类型1与类型2占比较大，因此将因子2命名为“电影类型”；因子3中，电影上映国家和时长比重较大，因此将因子3命名为“电影属性”。至此因子分析降维结束，具体因子得分见附件。

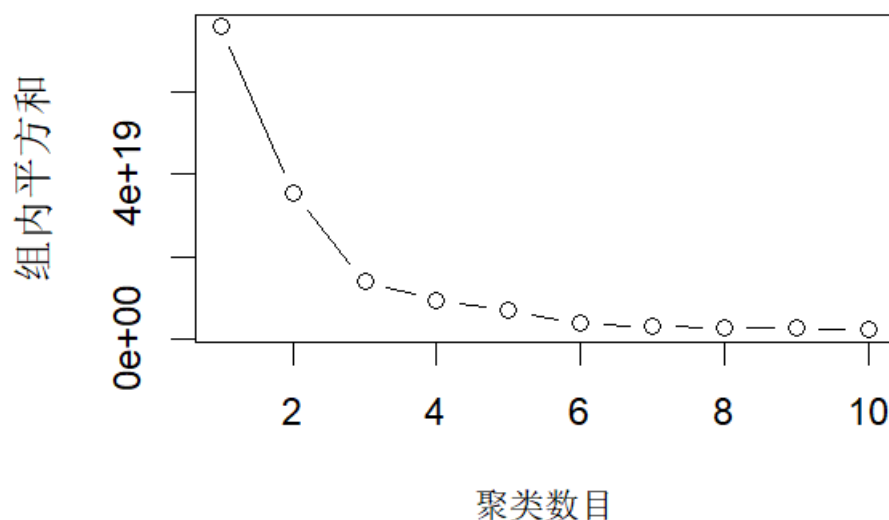
### 4.3 聚类分析

本文之后的思路为，使用多元回归分析，以上述得到的三个因子以及电影上映时间作为变量，以本人对于电影的评分作为因变量，构建多元线性回归模型，算出模型中的变量前的系数，构建个人电影评分模型。随后，可以使用模型，将其余的电影数据代入，算出得分，进行个性化推荐。

但在使用多元回归分析的过程中，发现，由于电影数据有 500 条，且各个类型、高分低分都有，且分布没有规律，构建出的线性回归模型十分不理想，远远达不到实现个性化推荐的效果。

因此，在进行多元线性回归构建个性化推荐模型前，先对 500 条电影数据进行聚类分析，随后针对每一个类，进行模型构建，这样可以优化模型的精确度，并对本人的观影数据进行进一步的分析、了解和划分。

本文通过组内平方和的方法，以组内平方和基本不再明显变化为标准，从而确定聚类的数目。

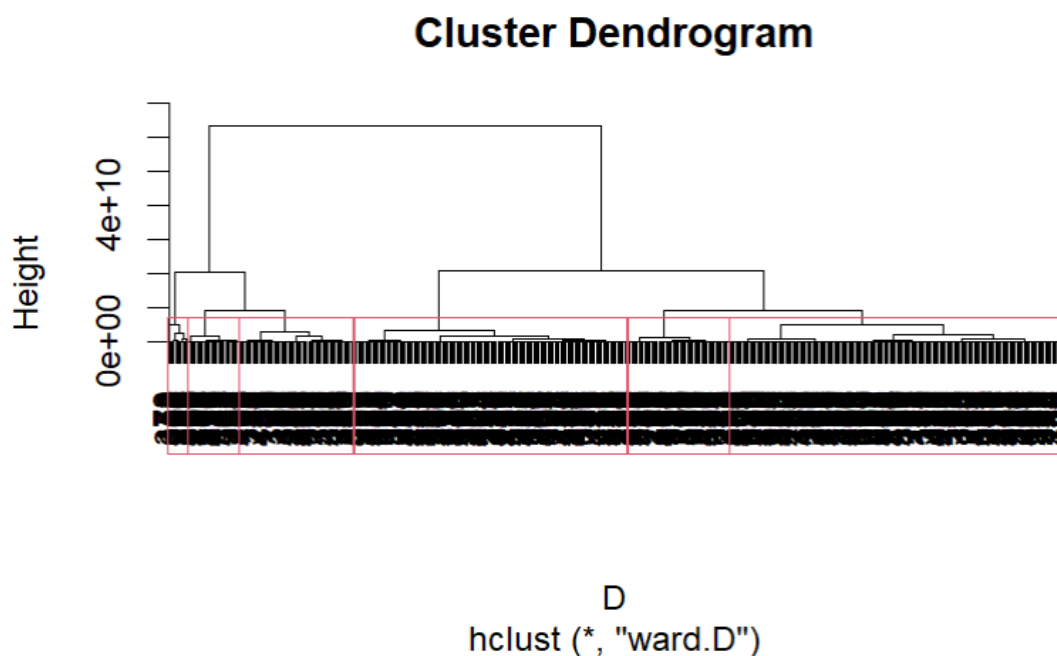


图二十二：组内平方和法确定聚类数目

如上图所示，当聚类数目为 6 时，组内平方和不再明显变化，因此聚类数目定为 6。同时使用 Medoids 周围分类法与贝叶斯信息准则指标进行二次判断。其中前者给出的聚类类别为 2，但描述变量仅 50%左右，极其不适合。BIC 指标显示，聚类数为 6 左右时最优，因此聚类数目最终确定为 6。

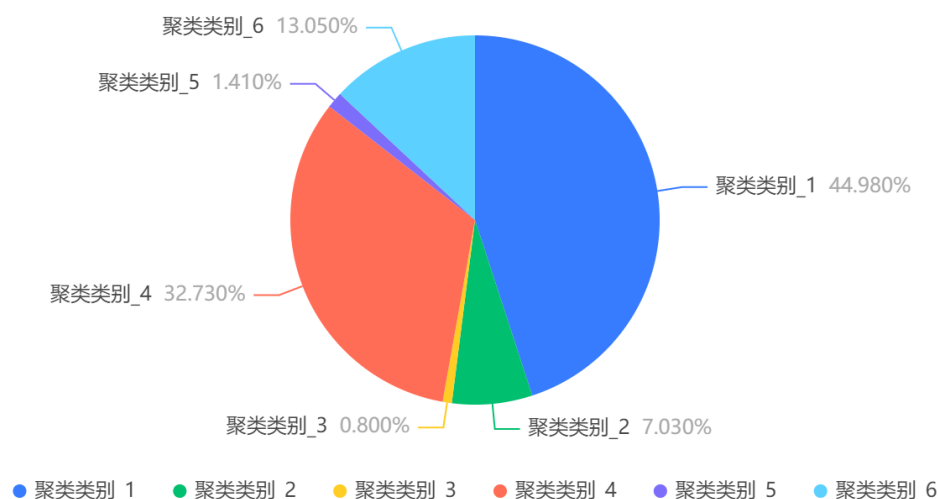
本文先后尝试了最短距离法、最长距离法、中间距离法、类平均法、ward.D2 法等 7 种聚类方法，发现聚类效果并不是很理想，所以最终直接使用 K-means 聚类法进行分析。





图二十三: ward.D 聚类结果展示

使用离差平方和 (ward.D) 进行聚类的结果较为合理, 但结果很难根据实际意义进行分组命名, 所以最终选择 K-means 进行聚类分析。



图二十四: K-means 聚类结果展示

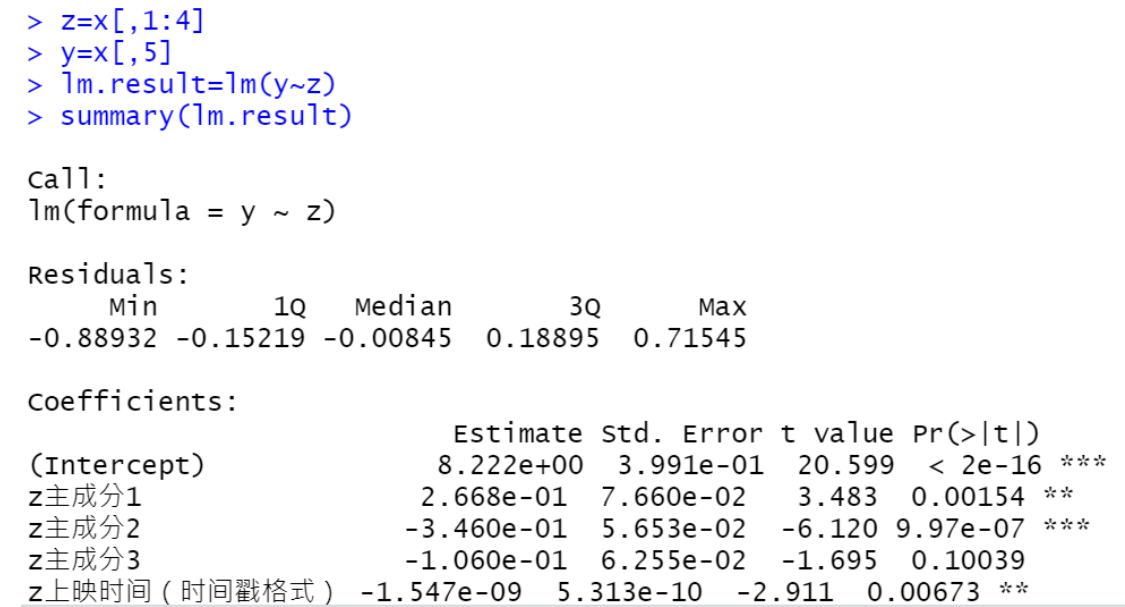
通过方差分析, 发现对于变量主成分 1、主成分 3、上映时间 (时间戳格式), 水平上呈现显著性, 拒绝原假设, 说明以上三个变量在聚类分析划分的类别之间存在显著性差异; 对于变量主成分 2, 显著性 P 值为 0.175, 水平上不呈现显著性, 不能拒绝原假设, 说明变量主成分 2 在聚类分析划分的类别之间不存在显著性差异。因此, 分类主要看主成分 1、主成分 3、上映时间 (时间戳格式) 变量。

聚类结果中的第三类、第五类，由于其包含数量少，所以将它们合并命名为类“其他”。第一类根据其呈显著性差异的变量得分区间，可将其命名为“低分低属性上映时间晚”；第四类可以概括为“低分高属性上映时间晚”；第六类为“高分低属性上映时间早”；第二类为“高分高属性上映时间早”。分类名称中的高低、早晚均是相对而言，可理解为较早、较晚。

至此，聚类分析完成，详细分类结果参附件。

#### 4.4 多元线性回归分析

本部分将着手对每个类别中的电影数据，进行构建个性化推荐模型（下文以类型 2（高分高属性上映时间早）类的数据进行展示）。



图二十五：基于 R 语言的回归分析

从上图中可以得出以下模型：

$Y$ （个人评分）=8.222（常数）+0.2668\*外界评价（主成分 1）-0.346\*电影类型（主成分 2）-0.106\*电影属性（主成分 3）-（1.547e-09）\*上映时间

Residual standard error: 0.3002 on 30 degrees of freedom  
Multiple R-squared: 0.7709, Adjusted R-squared: 0.7403  
F-statistic: 25.23 on 4 and 30 DF, p-value: 3.161e-09

图二十六：检验系数展示

可以看到， $R^2$ 的值为 0.7709，较为接近 1，即该模型的拟合度良好。即在当电影属于“高分高属性上映时间早”类型时，可以使用上述模型通过已有数据对个人电影评分进行预测分析。最终实现电影的个性化推荐。

同理的，对于类型 1，进行线性回归后，有以下模型：

$$y=10.712-0.055*\text{主成分 3}-0.305*\text{主成分 1}-0.101*\text{主成分 2}-(1.32\text{e-}09)*\text{上映时间}$$

对于类型 4，有以下模型：

$$y=9.907-0.105*\text{主成分 2}+0.097*\text{主成分 1}-(1.47\text{e-}09)*\text{上映时间}+0.118*\text{主成分 3}$$

对于类型 6，有以下模型：

$$y=4.411+(1.13\text{e-}09)*\text{上映时间}-0.181*\text{主成分 3}+0.116*\text{主成分 2}-0.059*\text{主成分 1}$$

对于其他类型，有以下模型：

$$y=8.222-(1.57\text{e-}09)*\text{上映时间}-1.32*\text{主成分 1}-0.353*\text{主成分 2}+0.972*\text{主成分 3}$$

至此，多元线性回归分析部分全部完成。

电影的个性化推荐，可以将待推荐电影的评论数据、上映时间数据等众多数据按照上文数据处理中的方法进行处理，然后通过因子分析中给出的各变量在因子中的成分占比，得出三种因子的得分，随后，根据各因子的数值和上映时间，判断该电影属于聚类分析中得出结果的哪种类型。最后，将相应的数值代入该电影所在类对应的线性模型，计算出预测评分，根据预测评分的高低进行电影推荐。

## 五、课程心得

这篇报告可能看起来还有很多的不足，数据处理方法还很稚嫩，但它已经是我上大学后，写的最为认真的、认为很有逻辑的一篇课程报告了。从选题到最后的处理数据、编写报告的整个过程中，我保守估计有超过一周的时间天天没事的时候就在想这个报告的事，做 PPT 和处理数据，也有三四天搞到了 23:00 多。以至于写到课程心得后，长舒了一口气，并想快点写完（本来想着在前面加一页前言和摘要页的，但现在一点想法也没有了），然后把打开的几十个网页、R、python、Excel 全部关掉，给电脑放个假，并且在很长一段时期内不再碰这些东西！

在本学期的课程中，确实学习到了很多内容。本着这个老师不好忽悠的心理，在写报告的时候也学到了很多知识，虽然遇到了很多问题，比如：数据的指标太少，因子分析没啥意义，然后费劲的看能把哪些指标加进来；文本处理与情感处理问题（几乎从零开始，现学现卖）；千辛万苦找到的很多指标，在计算 KMO 值时，发现相关度差；最后使用线性回归时，拟合度差  $R^2$  值小于 0.6（电影评分是一个很主观的东西，我就改了几个，优化了一下数据~~）。都通过百度、查资料一一解决，可以说是收获满满。

对于课程，上课时实操部分有些少，以至于我用 R 不是很熟练，希望下学期可以多些上机操作。还有，课程的理论部分较少，聚类分析理论简单些，主成分和因子有很多线代的知识，理解起来费劲，其实老师讲理论部分还是讲的很好的，希望下学期可以多涉及些原理方面的知识。

## 六、附录


### 6.1 Python 程序

本部分主要是 3.2 中对数据进行分析 and 处理的 python 项目，包括数据、源码和绘图。

 5000条电影数据分析	15/12/2022 下午9:43	文件夹
---	-------------------	-----

### 6.2 R 项目

本部分主要是个性化电影推荐部分使用的因子分析、聚类分析、多元线性回归分析的 R 语言程序，包括源码、数据和绘图。

 R-project	15/12/2022 下午9:48	文件夹
---	-------------------	-----



### 6.3 原始数据与编号

本部分是使用爬虫收集数据，并初步使用 Excel 进行数据整理，并对其中定性变量进行标号处理后的原始数据。

 电影数据 (暂定) .xlsx	15/12/2022 下午9:52	Microsoft Excel ...	481 KB
---	-------------------	---------------------	--------

### 6.4 情感分析结果

本部分是使用下列词典，对短评进行情感打分后的情感得分。

 情感分析结果.xlsx	15/12/2022 下午9:52	Microsoft Excel ...	368 KB
 自定义词典.xlsx	18/6/2022 下午5:00	Microsoft Excel ...	300 KB


### 6.5 因子分析结果

本文档存储因子得分情况。

 因子分析结果.xlsx	14/12/2022 下午11:28	Microsoft Excel ...	61 KB
---	--------------------	---------------------	-------

### 6.6 聚类分析结果

本文档存储聚类分析后的电影类别数据。

 聚类分析结果.xlsx	15/12/2022 下午8:57	Microsoft Excel ...	44 KB
---	-------------------	---------------------	-------

## 七、参考文献

- [1]张正风,强承魁,段素峰.个性化电影推荐算法综述[J].电脑知识与技术,2021,17(22):80-81+84.DOI:10.14004/j.cnki.ckt.2021.2187.
- [2]李浩.基于评论文本的个性化推荐算法研究[D].广西大学,2022.DOI:10.27034/d.cnki.ggxiu.2022.002080.
- [3]谢淳钰,万文君.二级市场股票价格影响因素辨析——基于 R 的多元线性回归[J].中国商论,2022(15):99-102.DOI:10.19699/j.cnki.issn2096-0298.2022.15.099.
- [4]焦子涵.基于 R 语言因子分析和聚类分析的市政建设水平综合评价研究[J].福建建材,2022(10):102-106.