|  |  |
| --- | --- |
| 学号 | 10214518336 |



课 程 设 计

课程名称 python程序设计课程设计

|  |  |
| --- | --- |
| 题 目 | 电影《流浪地球》的观众评价的数据分析 |
| 专 业 | 软件工程 |
| 班 级 | 软件1182 |
| 姓 名 | 魏静 |
| 成 绩 | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| 指 导 老 师 | \_\_\_\_\_\_\_\_\_\_徐勇\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |

2021 年 1 月 11 日至 2021 年 1 月 22 日

武汉华夏理工学院信息工程学院

**课 程 设 计 任 务 书**

课程名称：python程序设计课程设计 指导教师：徐勇

班级名称： 软件1181-1182 开课院、系：计算机与网络工程系

**一、课程设计目的与任务**

“python程序设计课程设计”是一个综合性的学习实践型实验教学环节，将在“python程序设计”课程的授课基础上，对python的基础语法、python 对文件的操作、python 对数据的操作、数据处理基础、数据可视化等若干个知识点进行综合运用。

python是一种具有天然开源基因的编程语言，了解开源社区和广泛的使用开源工具，也是Python学习实践的重要环节。因此在本次课程设计中，要求掌握jupyter notebook、Git等常用工具，以及github等重要开源社区的使用。

**二、课程设计的内容与基本要求**

数据分析的基本技术和方法在“python程序设计”课程中已经进行完整的讲授，由于数据分析本身是基于业务场景的，因此本次课程设计环节更加偏重于实际的业务场景的实践。通过对近期互联网热点的调查，准备了15个具体的业务场景，用于本次课程设计的具体任务场景。具体包括如下：

1. 北上广深租房状况分析；
2. 蔡某坤粉丝数及转发数据真假状况分析；
3. 地震的数据分析；
4. 英文名字的数据分析；
5. 外籍英文老师收入虚高情况数据分析；
6. 我国城市空气污染和烟花燃放的关系分析；
7. 针对996工作，程序员群体的看法的分析；
8. 吴某凡微博热点的分析；
9. 节假日长假景点人满为患的数据分析；
10. 针对荔枝的品种、销售地等维度，进行价格数据分析；
11. 分析芒果TV《我是大侦探》的观众评论数据；
12. 针对当前儿科医生的缺乏，对相关数据进行分析；
13. 著名网游《绝地求生》的数据分析；
14. 实习岗位状况的数据分析；
15. 电影《流浪地球》的观众评价的数据分析。

每三个学生组建一个课程设计小组，最后的任务输出包括代码每小组一份、课程设计报告每人一份、答辩ppt每小组一份，并进行课程设计成果答辩。小组成员均参与前述工作，但是每个人的侧重点不同。

每个课程设计小组可以从上述15个场景中选取一个作为课程设计的选题，选题中提供了待分析的数据，和现有的分析方法。各小组，通过学习和实践现有的分析方法，理解实战分析的思维过程并锻炼实际动手能力，再此基础上可以扩展更多维度的分析和数据展现形式。

每个班每个选题最多只能被两个小组选中，先选先得。同时，如果各小组发现更有意思的场景，并能够获取到相关待分析的数据，也可以申请作为选题方向。

本次课程设计的目标是培养学生的团队协作能力、对python知识点的综合运用、对实际场景的理解和适应能力、针对答辩的表达能力等。注重过程，期待成果，但不强求结果的尽善尽美。

**三、学时分配进度安排**

|  |  |  |
| --- | --- | --- |
| 序号 | 设计内容 | 所用时间 |
| 1 | 下发任务书，学生查阅资料 | 1天 |
| 2 | 组建团队，并进行选题和团队匹配 | 1天 |
| 3 | 各小组对自己的选题开始进行研究 | 2天 |
| 4 | 各小组开始准备数据并形成初步处理意见 | 1天 |
| 5 | 代码和实现分析的动手实践 | 2天 |
| 6 | 形成初步的报告书和ppt | 1天 |
| 7 | 答辩并完成报告书 | 2天 |
| 合 计 | | 2周 |

**四、课程设计考核及评分标准**

**1.设计报告要求**

课程设计报告要求逻辑清晰、层次分明、书写整洁。课程设计报告为每人一份，同一个小组的各成员的整体报告内容结构一致，但个人侧重点不同，个人着重撰写自己的工作内容，其他人的内容只要体现文档结构的完整性即可。

课程设计考核将综合考虑学生考勤和参与度、团队协作能力，过程管理能力、成果达成情况等。

**2.过程要求**

整个过程要求通过github来进行过程化发布，即阶段性地在github上提交过程结果。

**3.评分标准**

|  |  |
| --- | --- |
| **评分依据** | **评分成绩** |
| 1．团队协作能力 | 25分 |
| 2．python综合运用能力 | 25分 |
| 3．态度认真、刻苦钻研、遵守纪律 | 10分 |
| 4．过程完成、对工具的使用、对github的运用 | 20分 |
| 5．课程设计答辩逻辑清晰，内容正确 | 10分 |
| 6. 课程设计期间的课堂考勤、创新能力 | 10分 |
| 总分 | 100分 |

注：按上述六项分别记分后求和，根据小组成员贡献率综合评定，记载个人最后成绩。

成绩等级：优（90分—100分）、良（80分—89分）、中（70分—79分）、及格（60分—69分）、60分以下为不及格。

**五、指导地点与时间**

本课程设计将安排在第20-21周，采用腾讯会议和QQ群的方式，以在线形式进行。具体安排如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 周次 | 星期一 | 星期二 | 星期三 | 星期四 | 星期五 |
| 第20周 | 第5-8节 | 第5-8节 |  | 第5-8节 |  |
| 第21周 |  | 第5-8节 |  | 第5-8节 |  |

执笔:徐 勇 日期：2021-1-8

审阅:钱小红 日期：2021-1-8

**目录**

[1.选题背景 6](#_Toc62243030)

[2.需求开发环境 6](#_Toc62243031)

[3.分析目标 6](#_Toc62243032)

[4.分析内容 7](#_Toc62243033)

[5. 分析步骤 7](#_Toc62243034)

[5.1 csv文件信息读取 7](#_Toc62243035)

[5.2 数据清洗 7](#_Toc62243036)

[5.3 数据分析 8](#_Toc62243037)

[5.3.1对评分数量进行统计 8](#_Toc62243038)

[5.3.2对各演职员总体提及次数进行统计 10](#_Toc62243039)

[5.3.3对各演职员平均分进行统计 11](#_Toc62243040)

[5.3.4对各演职员低分提及百分比进行统计 12](#_Toc62243041)

[5.3.4评论来源城市分析 13](#_Toc62243042)

[5.3.4评论来源城市分析 14](#_Toc62243043)

[5.总体分析 15](#_Toc62243044)

[6.总结 15](#_Toc62243045)

[参考文献 15](#_Toc62243046)

# 1.选题背景

2019大年初一《流浪地球》全国上映。在豆瓣评分上，首日开分站稳8分以上，延续了之前点映的高口碑。微博上跟着出现吴京客串31天与投资6000万的热搜。知乎上关于“如何评价刘慈欣小说改编的同名电影《流浪地球》”的回答引起了众多人关注，包括该片导演郭帆的最高赞回答。

本篇文章爬取了豆瓣网上《流浪地球》的部分影评，并进行数据分析及可视化处理。

# 2.需求开发环境

Python3.9，jupyter notebook，GitHub等工具

# 3.分析目标

1.电影流浪地球中的各评分数量

2.高低分跟那些演员有关

### 4.分析内容

文件（data.csv）中包含“content”，“gender”，“id”，“nick”，“replyCount”，“score”，“time”等信息。首先读取全部信息去重；然后提前需要 的信息并进行去空处理；接着对提取的干净信息进行统计分析；最后生成相应的图。

# 分析步骤

## 5.1 csv文件信息读取

读取文件信息并去重，保存成DataFrame格式

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import matplotlib as mpl

%matplotlib inline

#设置中文标签的显示

plt.rcParams['font.sans-serif'] = ['SimHei'] #设置显示中文字体

#读取csv中的数据，以“;”作为分隔符来添加字段名

film\_data = pd.read\_csv('data.csv', delimiter=',', encoding='utf8')

film\_data

输出如图5.1.1



图5.1.1

## 5.2 数据清洗

提取信息并进行去空，规范化处理

#提取有效的列

dyxx = film\_data.loc[:,['content','score','time']]

dyxx.head()

输出如图5.2.1

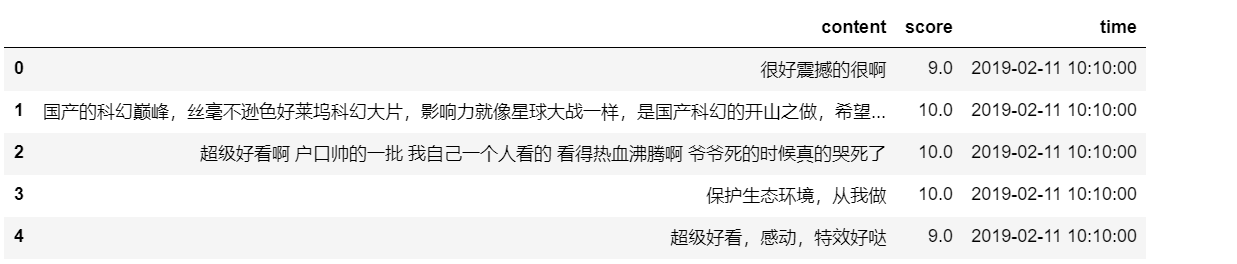


图5.2.1

#提取需要的列并去空

film\_data = film\_data[['content','score','time']].dropna()

film\_data.head()

输出如图5.2.2

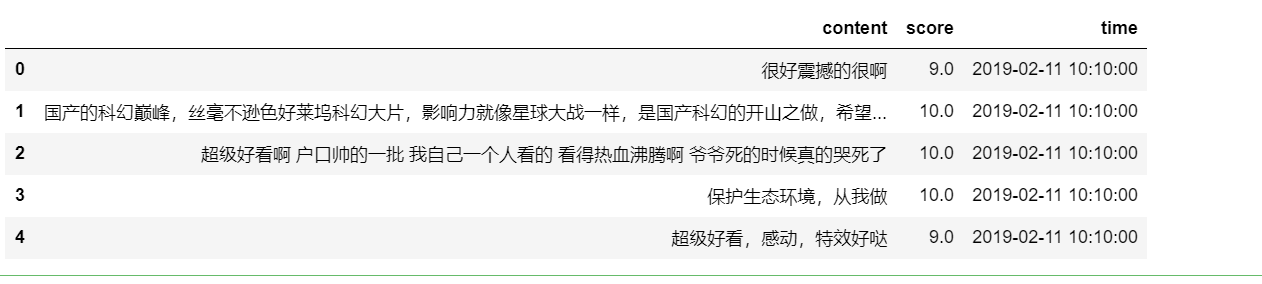


图5.2.2

## 5.3 数据分析

### 5.3.1对评分数量进行统计

film\_type = film\_data.groupby(film\_data['score'])['time'].count().reset\_index()

film\_type

输出如图5.3.1

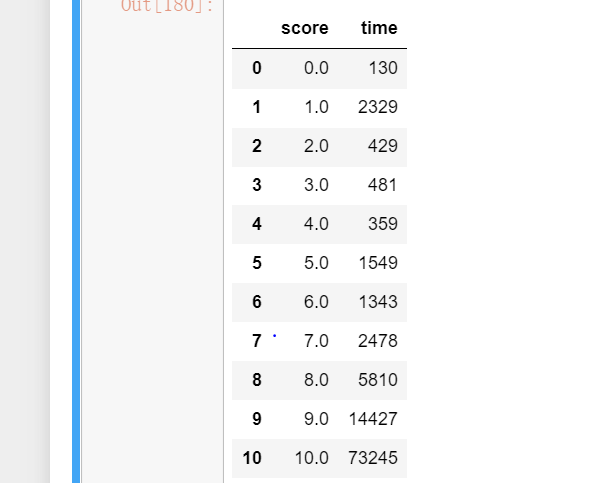


图5.3.1

#将“time”改为“小计”

film\_type.rename(columns={'time':"小计"},inplace=True)

film\_typ

输出如图5.3.2

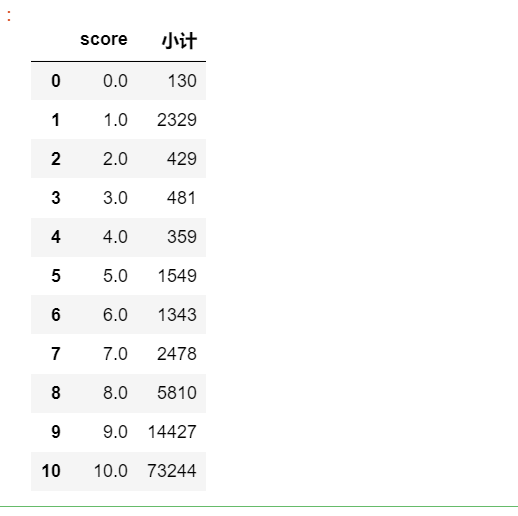


图5.3.2

# 绘图-评分总计

fig = plt.figure(figsize=(6,6))

ax = fig.add\_subplot(111)

ax.set\_title('评分总计')

ax.set\_ylabel('数量')

ax.set\_xticks(film\_type['score'])

ax.bar(film\_type['score'],film\_type['小计'])

plt.show()

输出如图5.3.3

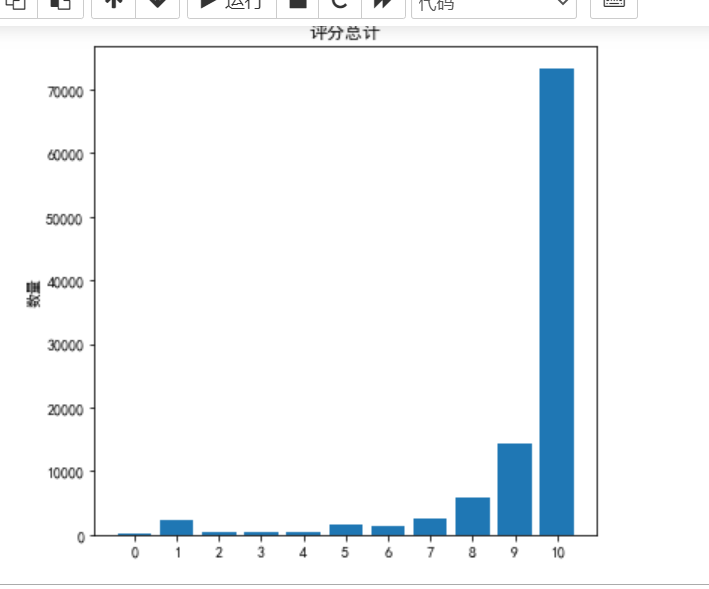


图5.3.3

## 5.3.2对各演职员总体提及次数进行统计

mapping = {'liucixin':'刘慈欣|大刘', 'guofan':'郭帆', 'quchuxiao':'屈楚萧|刘启|户口', 'wujing':'吴京|刘培强',

'liguangjie':'李光洁|王磊', 'wumengda':'吴孟达|达叔|韩子昂', 'zhaojinmai':'赵今麦|韩朵朵'}

for key, value in mapping.items():

film\_data[key] = film\_data['content'].str.contains(value)

# 总体提及次数

staff\_count = pd.Series({key: film\_data.loc[film\_data[key], 'score'].count() for key in mapping.keys()}).sort\_values()

staff\_count

输出如图5.3.4

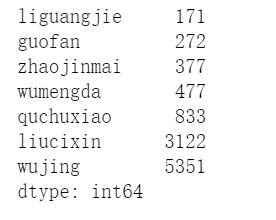


图5.3.4

# 绘图-总体提及次数

fig = plt.figure(figsize=(6,6))

ax = fig.add\_subplot(111)

x = ('李光洁', '郭帆', '赵今麦', '吴孟达', '屈楚萧', '刘慈欣', '吴京')

ax.set\_title('总体提及次数')

ax.pie(staff\_count.values,labels=x,textprops={'fontsize':15,'color':'black'}, autopct="%1.2f%%", shadow=True)

plt.show()

输出如图5.3.5

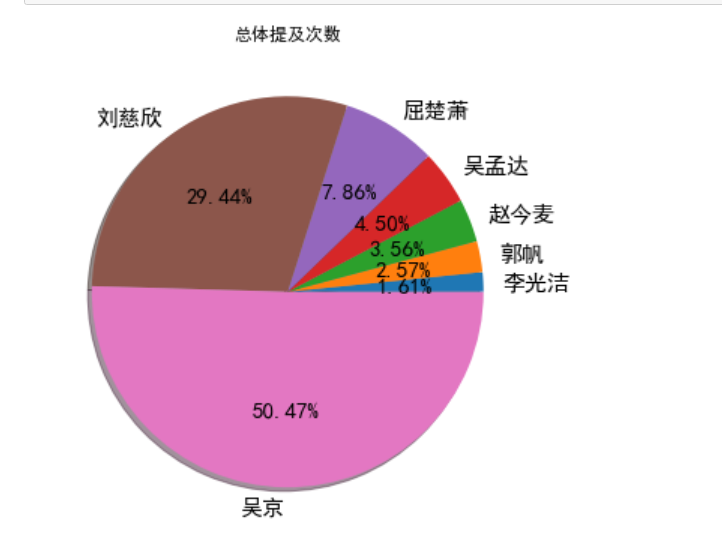


图5.3.5

### 5.3.3对各演职员平均分进行统计

#《流浪地球》演职员平均分

average\_score = pd.Series({key: film\_data.loc[film\_data[key], 'score'].mean() for key in mapping.keys()}).sort\_values()

average\_score

输出如图5.3.6

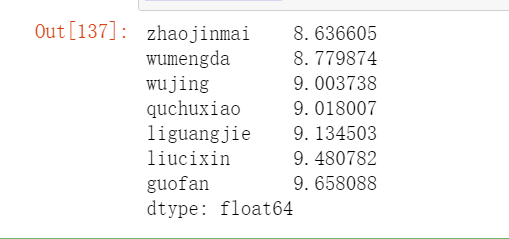


图5.3.6

# 绘图-流浪地球》演职员平均分

y = ('赵今麦', '吴孟达', '吴京', '屈楚萧', '李光洁', '刘慈欣', '郭帆')

plt.bar(y, average\_score.values)

plt.show()

输出如图5.3.7

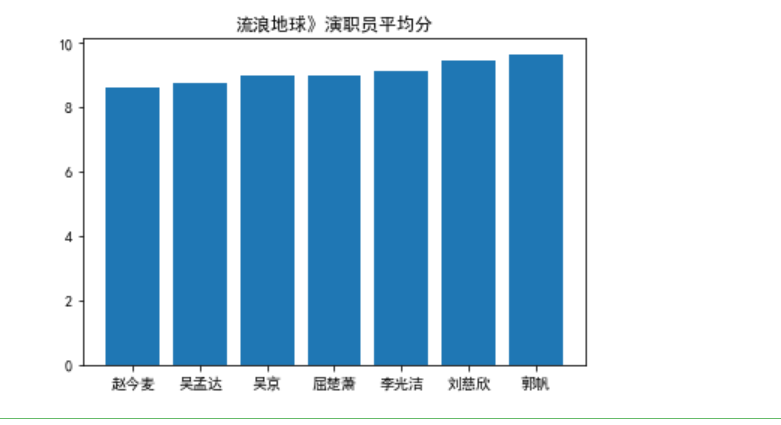


图5.3.7

### 5.3.4对各演职员低分提及百分比进行统计

#各演职员低分评论中被提及次数

staff\_count\_low = pd.Series({key: film\_data.loc[film\_data[key]&(film\_data['score']<5), 'score'].count() for key in mapping.keys()}).sort\_values()

staff\_count\_low

输出如图5.3.8

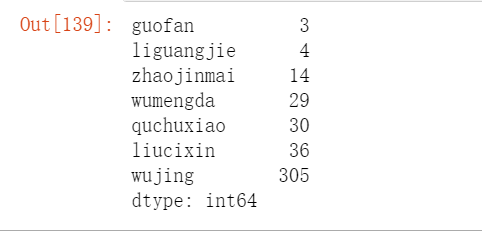


图5.3.8

#各演职员低分提及百分比

staff\_count\_pct = np.round(staff\_count\_low/staff\_count\*100, 2).sort\_values()

staff\_count\_pct

输出如图5.3.9

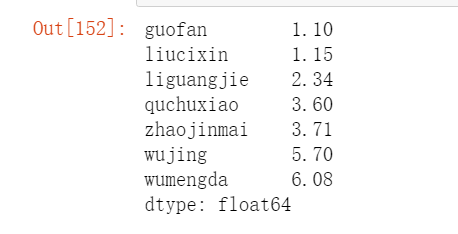


图5.3.9

# 绘图-各演员低分提及百分比

fig = plt.figure(figsize=(6,6))

ax = fig.add\_subplot(111)

z=('郭帆', '刘慈欣', '李光洁', '屈楚萧', '赵今麦', '吴京', '吴孟达')

ax.set\_title('各演员低分提及百分比')

ax.pie(staff\_count\_pct.values,labels=z,textprops={'fontsize':15,'color':'black'}, autopct="%1.2f%%", shadow=True)

plt.show()

输出如图5.3.0

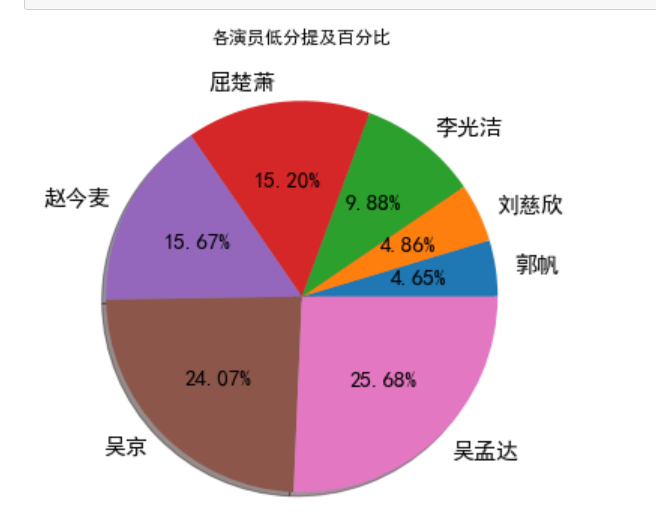


图5.3.0

### 5.3.4评论来源城市分析

调用pyecharts的page函数，可以在一个图像对象中创建多个chart，只需要对应的add即可。如图5.3.1

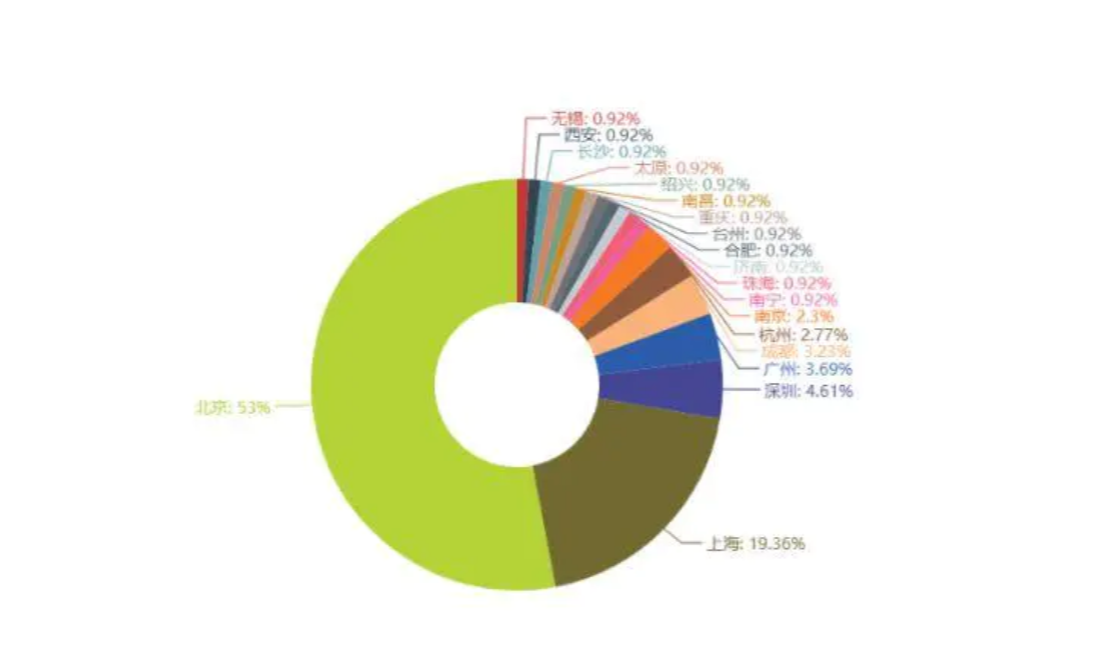


图5.3.1

### 5.3.4评论来源城市分析

读取csv文件，以dataframe(df)形式保存

遍历df行，保存到list

统计相同日期相同评分的个数

转换为df格式，设置列名

按日期排序

遍历新的df，每个日期的评分分为5种，因此需要插入补充缺失数值。如图5.3.2

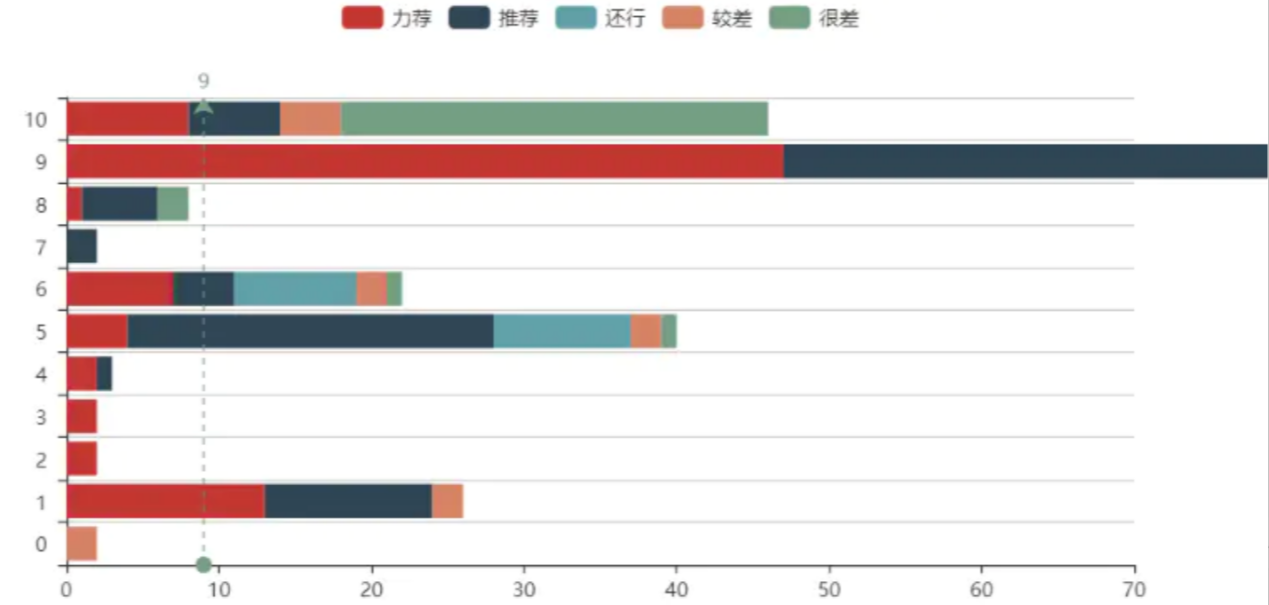


图5.3.2

# 5.总体分析

由于允许爬取的量少和时间问题，部分数据不是很明显。但依然可以得出一些发现。在影片上映开始的一周内，为评论高峰，尤其是上映3天内，这符合常识，但是也可能有偏差，因为爬虫获取的数据是经过豆瓣电影排序的，倘若数据量足够大得出的趋势可能更接近真实情况。

另外发现，影片在上映前也有部分评论，分析可能是影院公映前的小规模试映，且这些提前批的用户的评分均值，差不多接近影评上映后的大规模评论的最终评分 ，从这些细节中，我们或许可以猜测，这些能提前观看影片的，可能是资深影迷或者影视从业人员，他们的评论有着十分不错的参考价值。

# 6.总结

通过本次课程项目的实践，收获还是特别大的。本次课程项目实践给了我一-个很好的锻炼机会，我把理论的知识用到了实践，使得我的能力有了一点的提高，jupyter,github等工具有了更进一一步的认识，使用的熟练程度也较以前有了很大的提升，对于python项目，我觉得在做时，-定要细心，许多问题都很重要，稍微错--点就没法运行，还有就是工具使用问题，确实收获了不少，在程序的调试过程中遇到了许许多多的错误，在不多的努力下，最终都解决了，这个收获是最大的-笔知识财富，为我以后的学习奠定了很大的基础，为以后的学习和工作铺好了-条我所要的路。在完成报告的过程中，自己也重新梳理了一遍流程 ，当今软件的开发已经逐步成熟,有着和传统工业一样的系统化开发流程，这样规范了开发的各个步骤，同时也有利于提高开发的效率。在开发的过程中，也遇到了很多以前没有见过的难题。确实，很多时候问题都是边做边发现的，平时觉得很不起眼的小功能，到了真正实现的时候，就会发现没有想象中的那么容易。也只有这样,你才能全方位的提高自己的技术水平,动手才是解决问题和提高自己的最好方法。学习是无止境的，对于计算机程序而言，我们只有不断的去学习理论知识，同时要不断的去做项目来巩固知识，只有这样，我们才会学有所成，我们才能真正的把程序学好，才能够达到我们想要的结果。

# 参考文献

1. 基于Python的数据可视化[J].罗博炜.信息记录材料.2019(12)

[2]基于Python的招聘网站信息爬取与数据分析[J].王芳.信息技术与网络安全.2019(08)

[3]刘凌霞，郝宁波，吴海涛，21天学通Python[M].北京:电子工业出版社，2016.

[4]刘宇宙.Python 3.5从零开始学[M].北京:清华大学出版社，2017.

[5]董付国.Python可以这样学[M].北京:清华大学出版社，2017.

**课程设计成绩评定表**

|  |  |  |  |
| --- | --- | --- | --- |
| 课程设计题目 | 电影《流浪地球》的观众评价的数据分析 | | |
| 课程设计学生答辩或质疑记录：   1. 为什么要使用reset\_index（）   在不指定任何参数的情况下使用reset\_index（），则序列号将成为新索引，而原始索引将保留l为新列。   1. jupyter notebook如何删除输出   找到顶部菜单栏,点击 “cell”, 选择 “current outputs”,点击“clear”即可。”。   1. 如何设置中文标签的显示   plt.rcParams['font.sans-serif'] = ['SimHei'] | | | |
| **评 分 依 据** | | **分 值** | **评分成绩** |
| 1．团队协作能力 | | 25分 |  |
| 2．python综合运用能力 | | 25分 |  |
| 3．态度认真、刻苦钻研、创新能力 | | 10分 |  |
| 4．过程完成、对工具的使用、对github的运用 | | 20分 |  |
| 5．课程设计答辩逻辑清晰，内容正确 | | 10分 |  |
| 6. 课程设计期间的课堂考勤、遵守纪律 | | 10分 |  |
| 总 分 | | 100分 |  |
| 最终评定等级为：  指导老师签字：  2021 年 1 月 15日 | | | |