|  |  |
| --- | --- |
| 学号 | 10214518312 |



课 程 设 计

课程名称 python程序设计课程设计

|  |  |
| --- | --- |
| 题 目 | 菜某坤粉丝数及转发数据真假状况分析 |
| 专 业 | 软件工程 |
| 班 级 | 1182 |
| 姓 名 | 张湖川 |
| 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

# **1.分析目标**

蔡徐坤的微博转发是否存在假流量？

真假流量所占的比例各有多少？

假流量粉丝是如何生产出来的？

真流量粉的粉丝画像？

# **2.分析内容**

读取转发数据的全部信息，进行数据清洗，然后对转发数据进行统计分析，最后实现数据可视化分析。

**3.分析实施**

**3.1 数据获取**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import matplotlib as mpl

import seaborn as sns

from pyecharts.charts.bar import Bar

%matplotlib inline

plt.style.use('ggplot')

mpl.rcParams['font.sans-serif'] = ['SimHei'] #解决seaborn中文字体显示问题

plt.rc('figure', figsize=(10, 10)) #把plt默认的图片size调大一点

plt.rcParams["figure.dpi"] =mpl.rcParams['axes.unicode\_minus'] = False # 解决保存图像是负号'-'显示为方块的问题

**3.2 数据预处理**

对抓取完的数据，首先进行的是数据清洗。然后进行重复项删除，最后只留下唯一的数据。对job\_detailed.csv文件的数据清洗如图3.1所示，对com\_detailed.csv文件的数据清洗如图3.2所示。

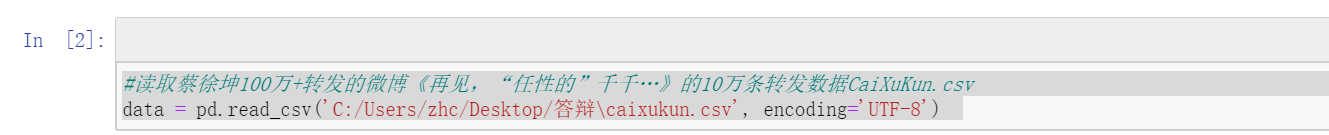


图3.2.1

因部分数据内容与分析无任何关系，故将不需要的数据不做筛选。我的判定方式是将testjob.csv文件中包含"city"、"cname"、"day"、"degree"、"iname"、"industry"、"maxsal"、"minsal"、"month"、"attraction"中的的字段保留,然后导入Mysql。处理之后数据如图3.3所示。

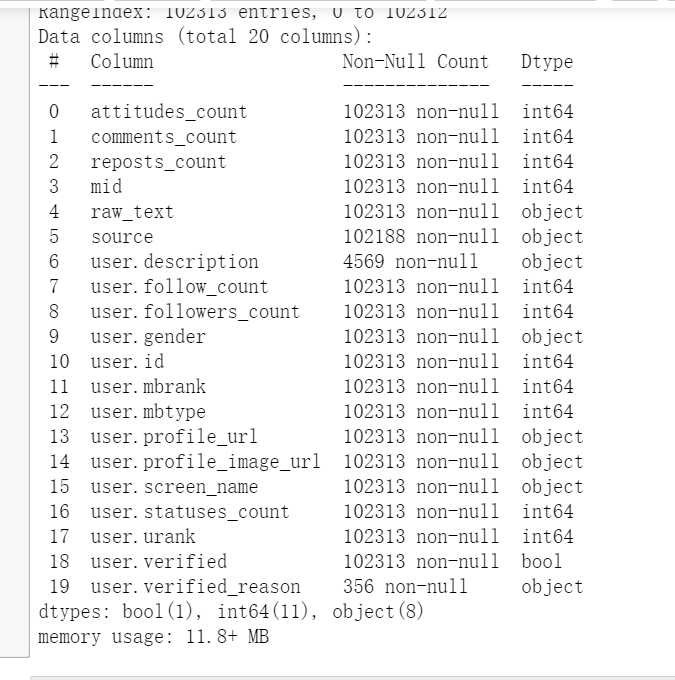


图3.2.2

由于数据入库的时候没有进行清洗，所以数据多出了很多没用的字段，需要先清洗掉

返回列

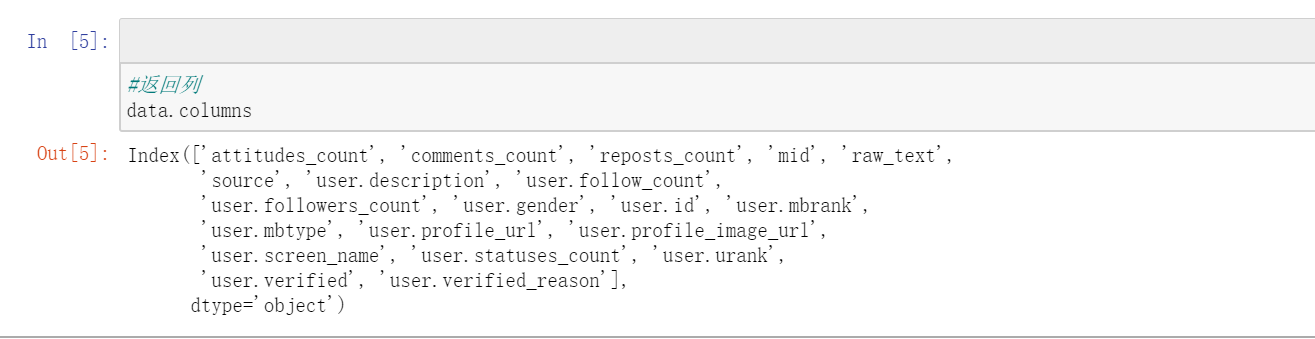
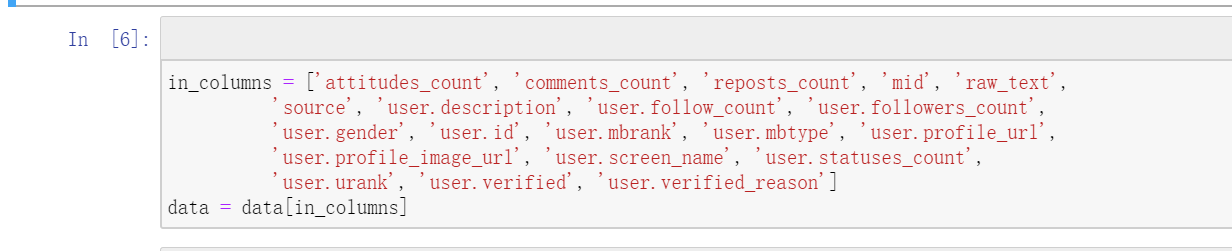


图3.2.3

选取需要的字段：



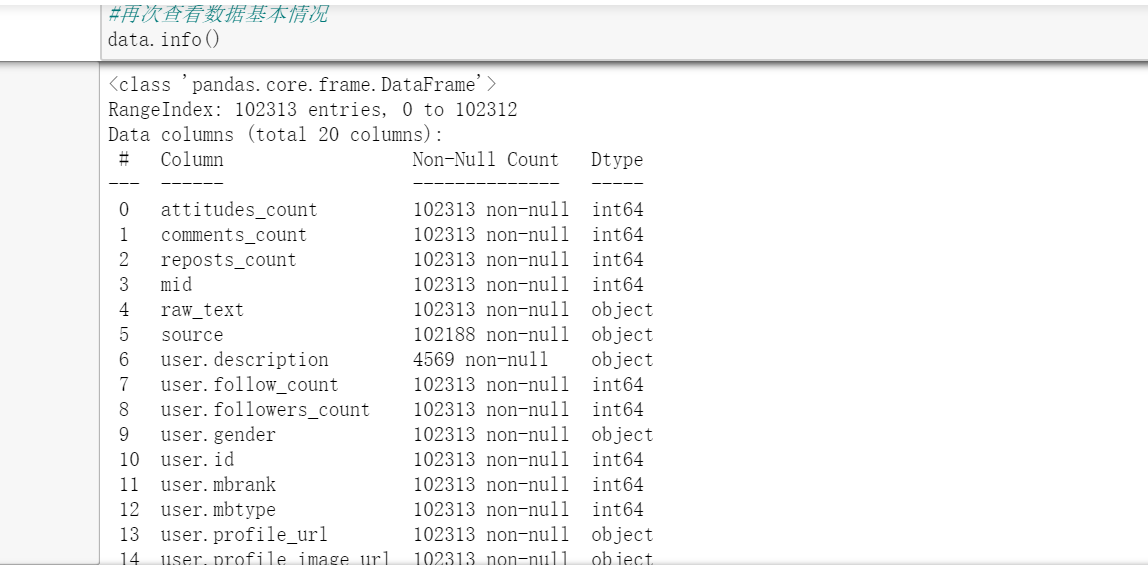


图3.2.4

# 数据分析

# 4.1**.1. 蔡徐坤的微博转发是否存在假流量？**

先来看看蔡徐坤的粉丝性别比例

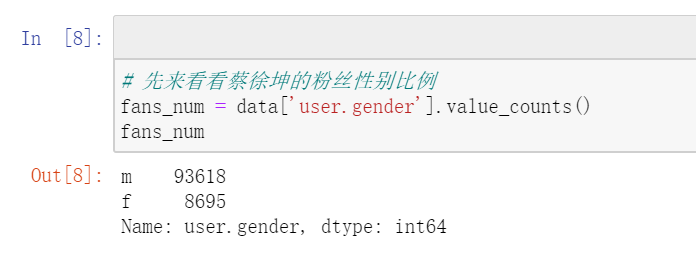


图4.1.1

利用Bar显示性别比例可视化



图4.1.2

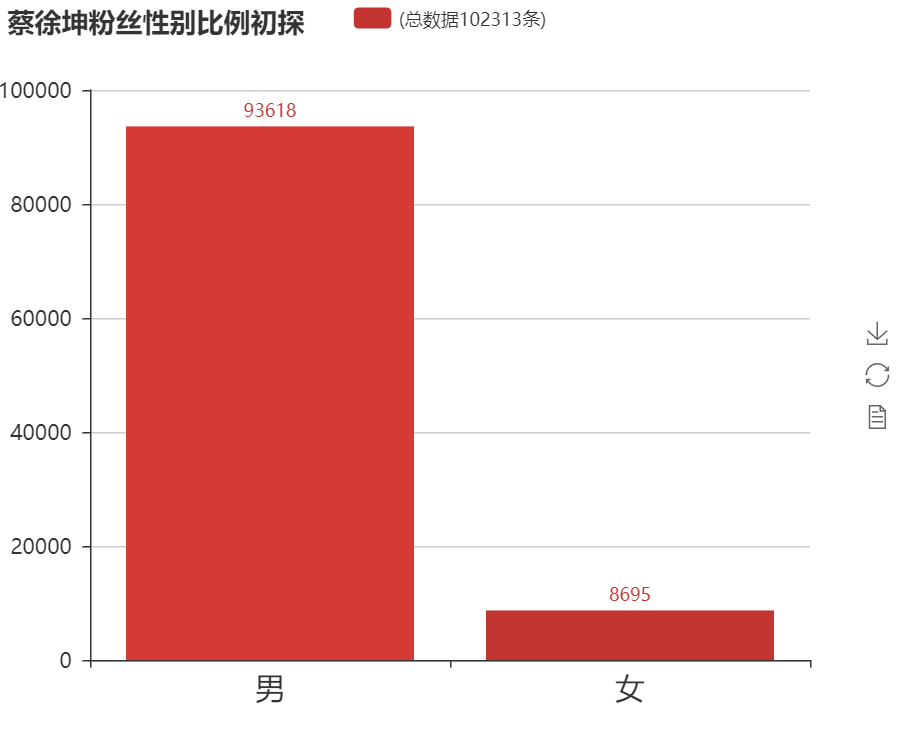


表4.1.1

男女比例

np.round(fans\_num/fans\_num.sum()\*100, 2)

随机显示5组性别为m的用户的信息

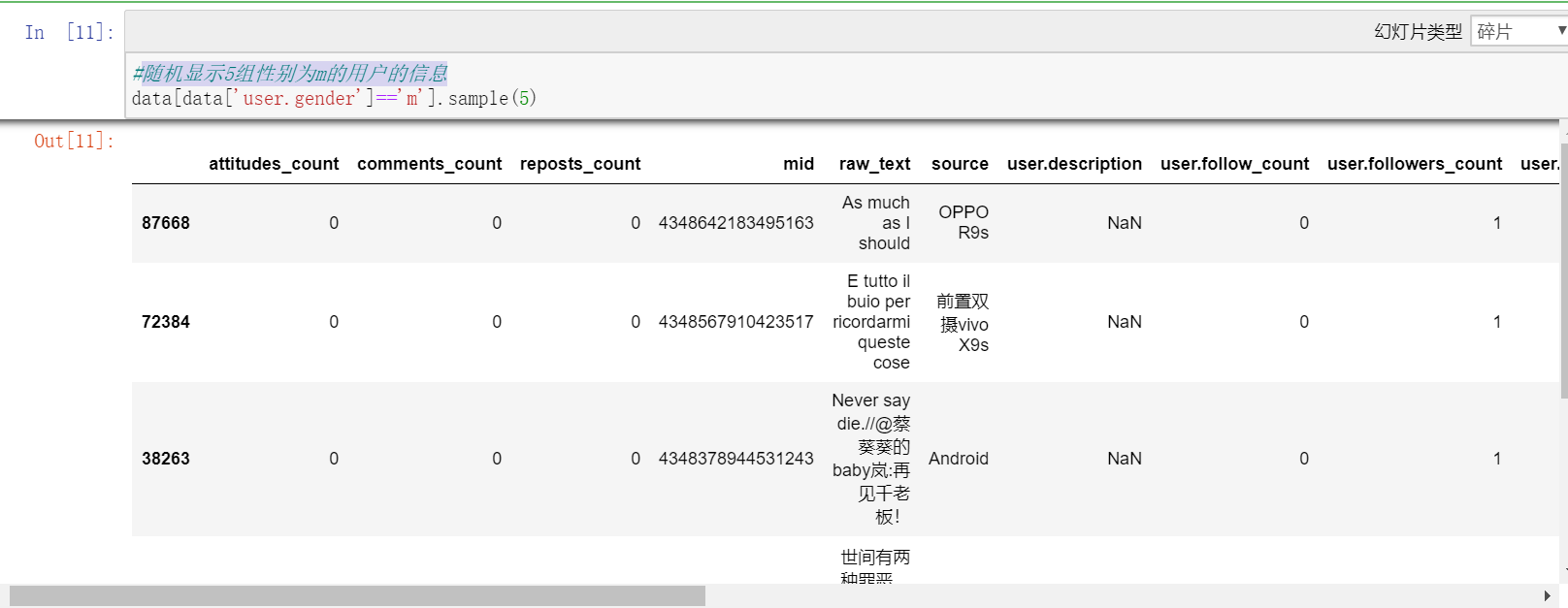


图4.1.3

### **4.2. 真假流量所占的比例各有多少？**

筛选出用户描述‘wu’

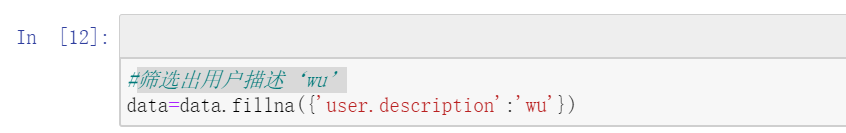


图4.2.1

通过一番探索分析，我们把转发数据中转发者的关注或者粉丝数少于等于5、筛选出用户描述为‘wu’，

#发表评论、转发之后被点赞数评论数再转发数都为0、微博会员等级为0级的数据，以及转发者的关注

#这部分数据，便是我们所说的假流量。



图4.2.2

然后粉丝数大于等于5但昵称包含“用户”的这样的数据抽取出来。基本上也可以断定是假粉丝



图4.2.3

最后把假的流量粉丝转发组合起来



图4.2.4

现在我们利用bar把蔡徐坤真假流量的转发量可视化！

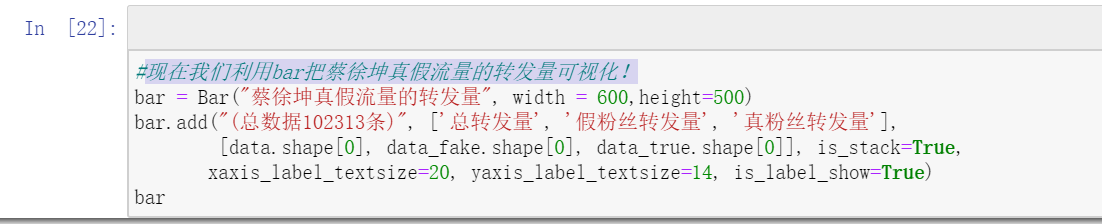


图4.2.5

# 蔡徐坤真假流量的转发量

表4.2.1

在102313条转发数据中，有95397条是由假粉丝转发的，占了总转发的93.24%，只有6916条是由真粉丝转发的，占6.76%。

那么6919条真粉丝转发的数据中，除去重复转发刷榜的数量，里面一共有多少个真粉丝在转发呢？我们把这部分数据按照粉丝微博ID进行去重。

### 蔡徐坤真假流量的转发量与真实转发粉丝量(总数据102313条)

表4.2.2

真实转发粉丝量占总转发数的3.84%

### 假流量粉丝是如何生产出来的？【对data\_fake进行讨论】

假流量粉丝性别比例



图4.2.6

利用Bar实现假粉丝性别可视化

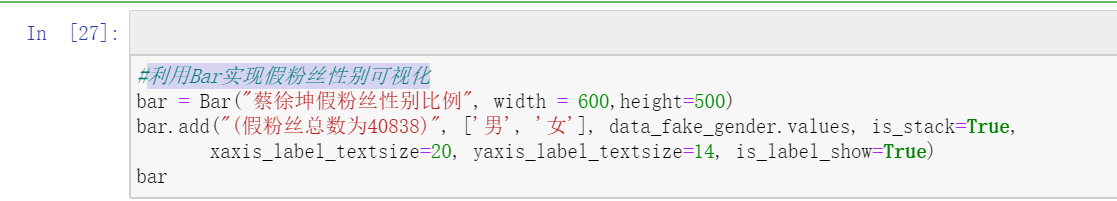


图4.2.7

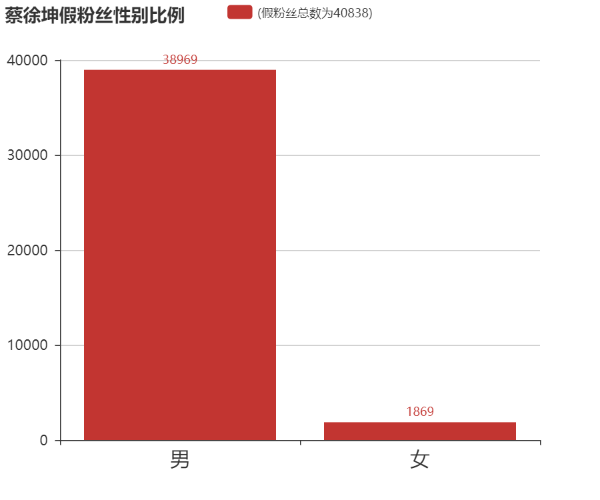


表4.2.3

从上图可见在95397条假转发中，有40838个假粉丝。其中男性的比例高达95.42%！

再看一下假粉丝转发微博里都是些啥玩意！

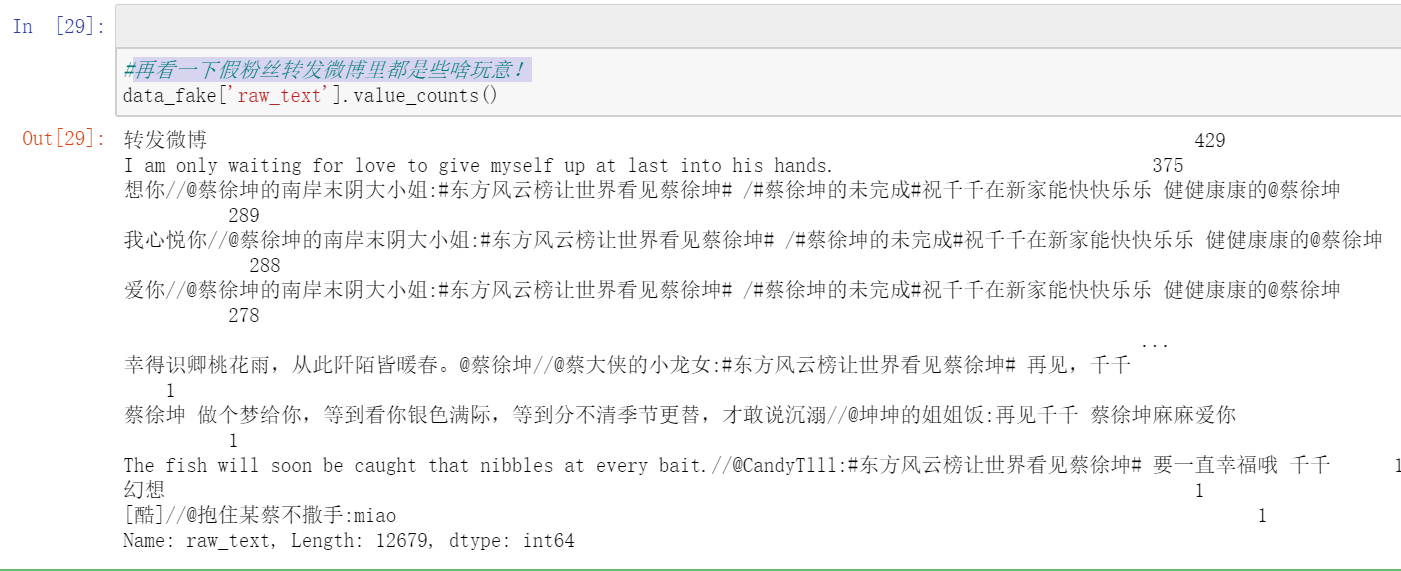


图4.2.8

假粉丝使用的前10设备

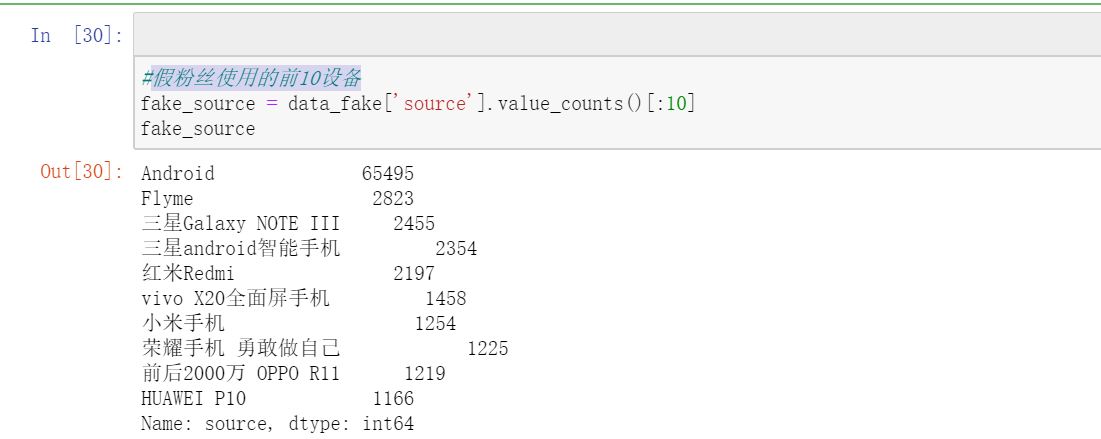


图4.2.9

利用Bar实现假粉丝使用设备比例可视化

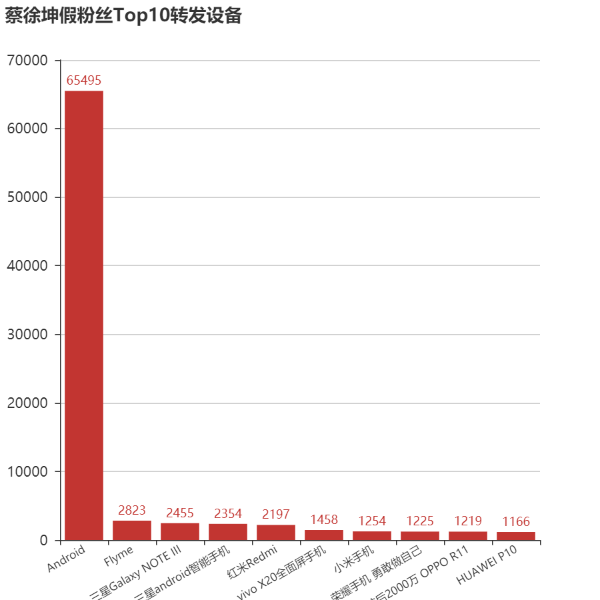


表4.2.4

在假粉丝使用的Top10转发设备中，Android高居榜首，这也再次证明了这些粉丝是假粉丝。

另外还有一些有趣的发现（请接着看下面代码）：假粉丝的平均关注是3.44，平均粉丝数是1.04，没有简介，昵称基本上都是“中文+英文和数字”这个格式，很多假粉丝的昵称都带有“坤”、“蔡”、“葵”、“kun”等字，头像都是蔡徐坤（说明很多都是定制粉啊）。

随机5个假粉丝的用户名



图4.2.10

看一下假粉丝昵称中带有“坤”、“蔡”、“葵”、“kun”的总和



图4.2.11

假粉丝中平均在线状态



图4.2.12

### 4.3**真流量粉的粉丝画像**

### 先来看看真粉丝的性别比例。可见3926个真粉丝中，女生（f）占了绝大部分，这才是符合逻辑的粉丝比例嘛。

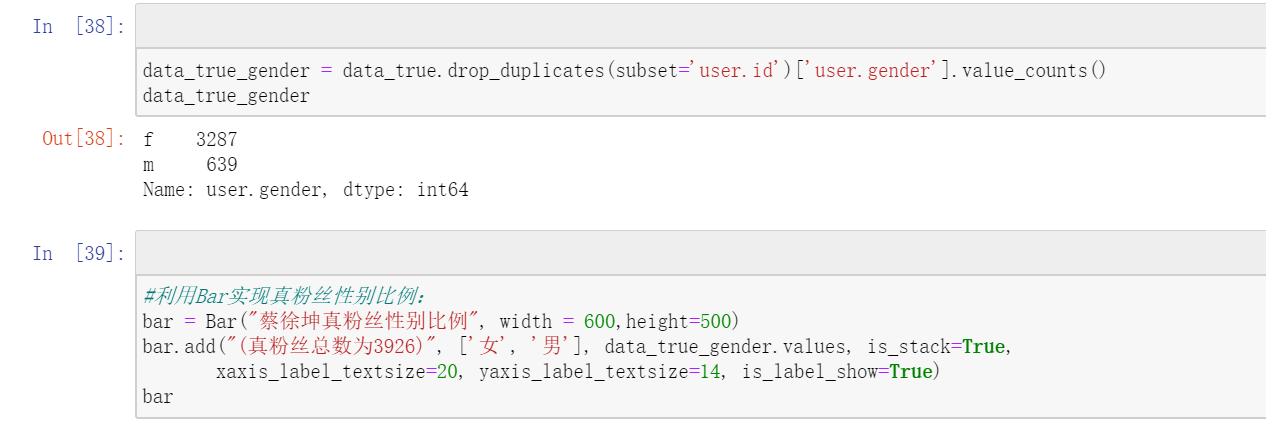


图4.3.1

利用Bar实现真粉丝性别比例

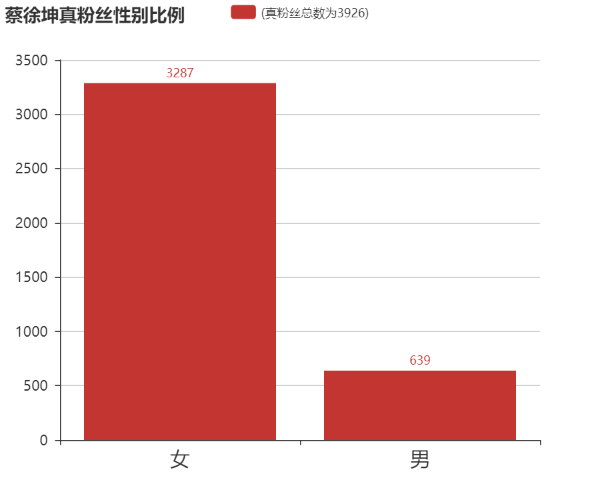


表4.3.1

在这些粉丝转发所携带的评论中可以看到，很多都是支持蔡徐坤拿下#明星势力榜#或者#东方风云榜#第一名而转发的。



图4.3.2

利用Bar实现真粉丝使用设备比例：

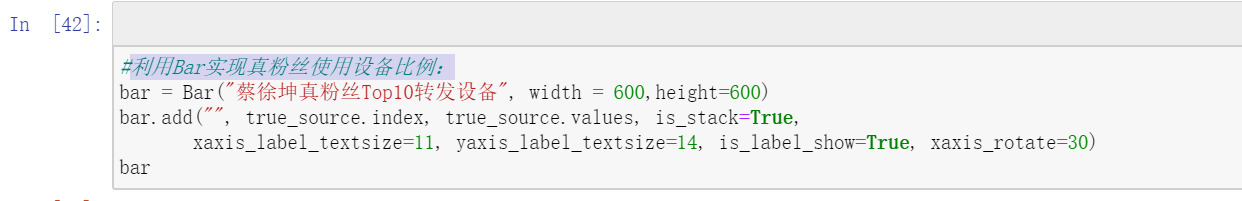


图4.3.3

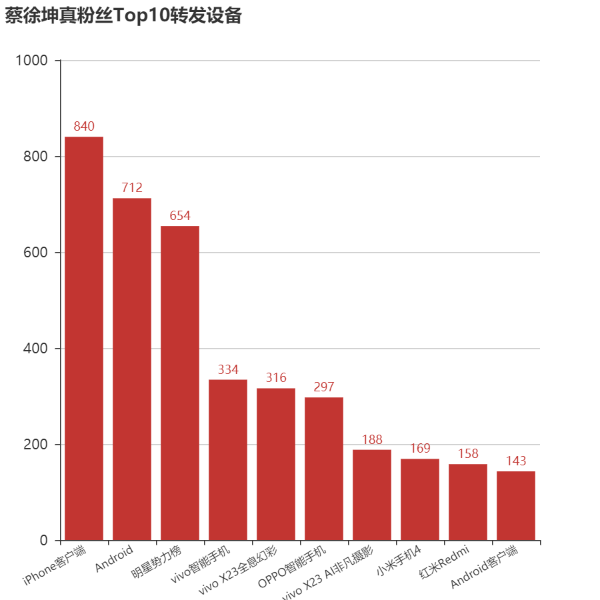


表4.3.2

真粉丝所使用的转发设备，各种设备分布都比较均匀，最受欢迎的是iPhone客户端。

下面是一些真粉丝中的信息，真粉丝的平均关注量为222，平均粉丝数是179，同样，很多粉丝名字中喜欢带有“坤”、“蔡”、“葵”、“kun”等字。

真粉丝平均关注的人



图4.3.4

随机取出5个真粉丝的名字



图4.3.5

我们把真粉丝的简介做成了词云图。

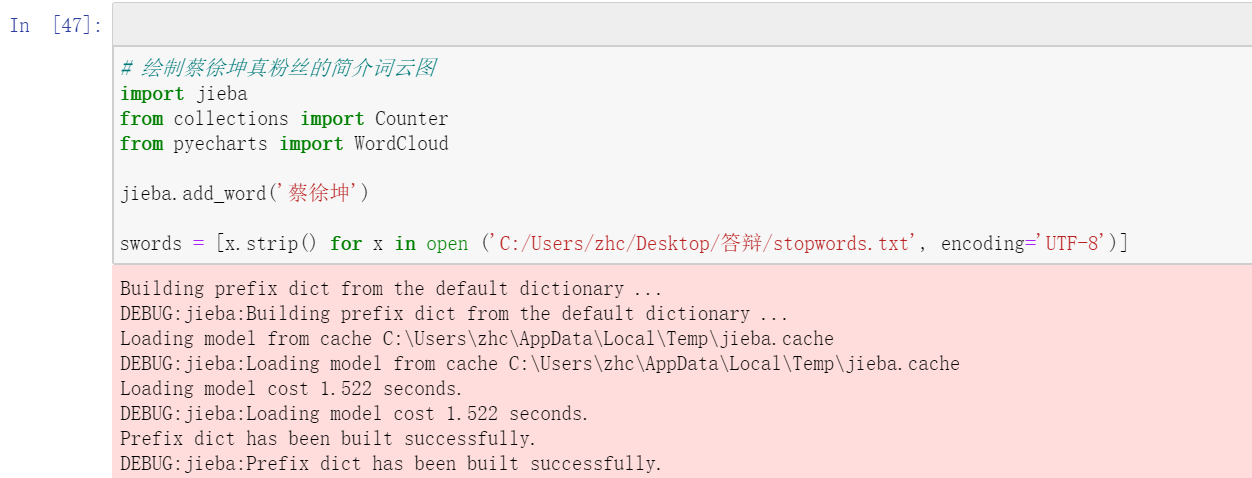


图4.3.6

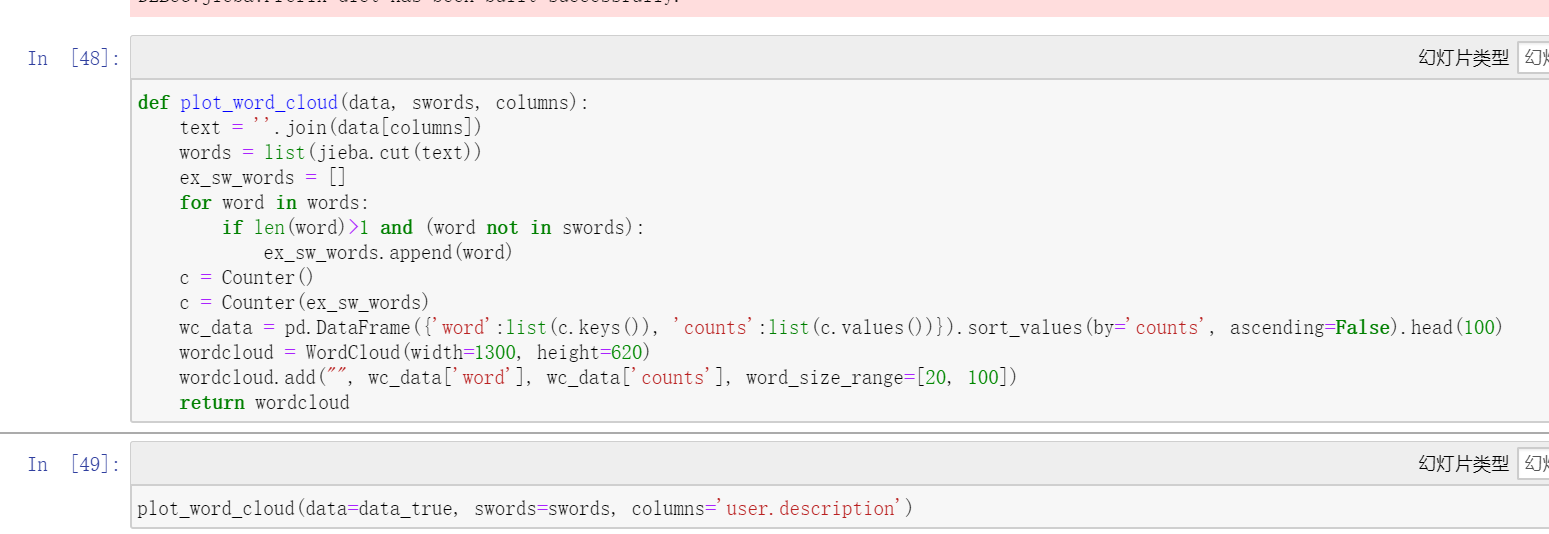


图4.3.7

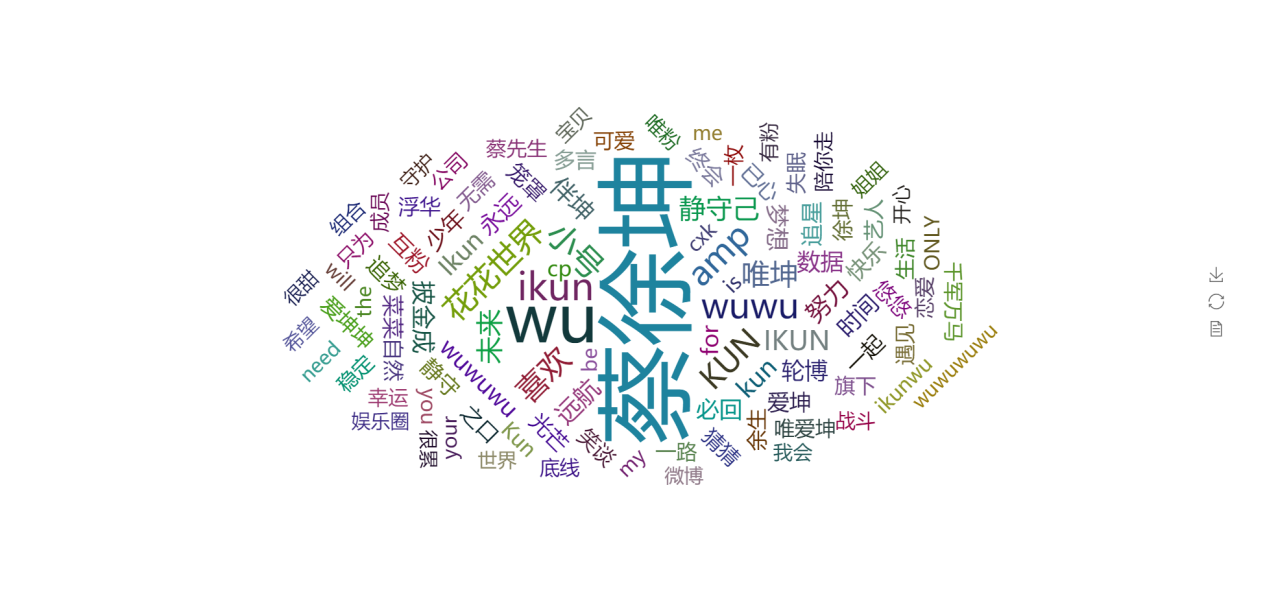


图4.3.8

可见，真粉丝们的简介都喜欢带上蔡徐坤的名字，很喜欢蔡徐坤，想要一直陪着他走下去。看着里面的“少年”、“努力”、“自由”“追梦”等字眼，发现这就是我们以前的青春呀。

另外还把转发所带的评论做成了词云图。

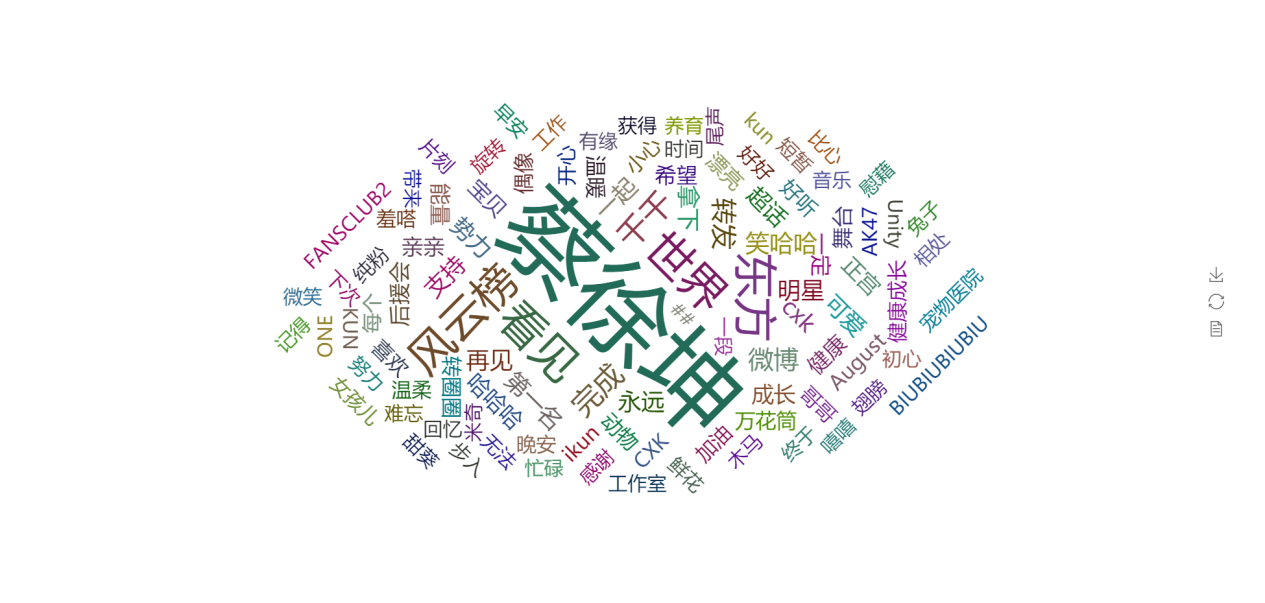


图4.3.9

可见，粉丝们都很在乎“风云榜”这件事，并且要帮助蔡徐坤拿第一名。里面有早安打卡的，有超级话题的，还有很多“开心”、“比心”、“温暖”等字眼。说明绝大部分真粉还是很温暖的。

**总结**

**[¶](http://localhost:8889/notebooks/CaiXuKun (1).ipynb" \l "%E6%80%BB%E7%BB%93)**数据显示，蔡徐坤的100万+的微博转发中，确实存在绝大部分的假流量。这些假流量估计有两方面的来源：一是自身经纪公司购买的，二是忠实的粉丝自费购买的。所以在这里提醒一下大家，追星需谨慎呀！

# 5.思考

此次课程设计主要是对之前学习的SQL及可视化图表的实践，真实的数据分析远不止这些，还需要扎实的教学、统计学基础，良好的数据敏感度，开拓但严谨的思维等，是一个终生学习的过程。

# 6.总结

仔细阅读有关Python的书籍，好记心不如烂不同，一定要做笔记，过后回顾笔记。书上给出的练习，一定要做，并且争取全部弄懂。书上有的练习代码，一定要自己调试一遍，知其所以然，进行本次Python课设的过程中，查阅了许多资料，在老师和同学的提点下，完成了这个实习岗位状况的数据分析。在这个过程中，我学到了很多，如怎样去全面的看待并分析一个问题等，同时也在总体上巩固了自己的Python知识，发现自己之前从来没有像现在一样将零散的知识点结合在一起。

每天的编码必不可少，既然选择学习编程，学习Python，坚持编码应该是必须做到的。没有代码积累，要写出高质量的代码，几乎不可能。我认为python虽然在许多方面相对于c语言比较方便，但也有其相对于弱一点的方面，比如说for循环等方面。不过也依然不会影响到python的强大希望自己能够在不断地学习中，将python学习的更加好。

这些知识就像是建筑工地随处可见的砖石，不管这些砖石怎样的不起眼，但是没有一幢建筑可以离开砖石，学习的过程是枯燥的，不过这也正符合非常现实的一条规律。也正是因为它枯燥苦闷，学习有难度，才保证了，我们学习了这些知识以后，可以靠它们为生，在这个知识时代闯出自己的一片天!

# **参考文献**

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

[2]刘宇宙.Python 3.5从零开始学[M].北京:清华大学出版社.

[3] 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)

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

[5] 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.

**课程设计成绩评定表**

|  |  |  |  |
| --- | --- | --- | --- |
| 课程设计题目 | 数据分析的学习与实践 | | |
| 课程设计学生答辩或质疑记录：  1.你们小组的分工是怎么样的? 答:章华鼎主要负责代码编写、数据清洗、ppt 制作，徐佳清主要负责Github管理及代码、报告编写;张湖川主要负责代码编写及代码调试，但每个人都参与了各个项目环节。  2.如何对数据进行处理? 首先读取并筛选有效的数据，然后进行数据清洗(包括检查数据的一致性、有无缺失项、有无重复)，再把清洗好的数据另存为.csv文件，再将其导入Mysgl数据库,观察数据，通过sgl语句，对有问题的数据进行处理。如 3.在做的过程中，应该注意什么? 在做的过程中，应尽可能多地观察数据，发现有问题的数据及时进行处理，否则←之后进行分析实现数据可视化的图形很容易产生歧义。 | | | |
| **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、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 |  |
| 最终评定等级为：  指导老师签字：  2021 年 1 月 15日 | | | |