|  |  |
| --- | --- |
| 学号 | 10212818110 |



课 程 设 计

课程名称 python程序设计课程设计

|  |  |
| --- | --- |
| 题 目 | 蔡徐坤粉丝数及转发数据真假状况分析 |
| 专 业 | 软件工程 |
| 班 级 | 软件1181 |
| 姓 名 | 全智勇 |
| 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、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 |

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.设计目的 1](#_Toc61701281)

[2.开发环境与开发技术](#_Toc61701282) [2](#_Toc61701282)

2.1 硬件平台…………………………………………………… 2

[2.2 Jupyter Netbook介绍 2](#_Toc61701283)

[2.3 python语言介绍 2](#_Toc61701284)

[2.4 GitHib介绍 3](#_Toc61701285)

[2.5 PyCharm开发工具介绍 4](#_Toc61701286)

[3.概要设计 6](#_Toc61701287)

[3.1函数库的导入 6](#_Toc61701288)

[3.2函数库的注解 7](#_Toc61701291)

[3.3CSV文件的读取 8](#_Toc61701294)

[3.4数据的清洗 9](#_Toc61701296)

[3.5 蔡徐坤转发微博是否存在假流量 10](#_Toc61701299)

[3.5.1 数据抽样 10](#_Toc61701300)

[3.5.2 蔡徐坤的粉丝性别比例 10](#_Toc61701299)

[3.5.3 蔡徐坤的粉丝性别比例 10](#_Toc61701299)

[3.6 真假流量所占比例各是多少 11](#_Toc61701301)

[3.6.1 蔡徐坤真假流量转发量 11](#_Toc61701302)

[3.7 假流量粉丝是如何生产出来的？ 12](#_Toc61701303)

[3.7.1蔡徐坤假粉丝Top10转发设备 12](#_Toc61701304)

4.[详细设计 16](#_Toc61701310)

# 1 设计目的

在2019年的时候，央视新闻曾公开披露流量明星数据造假。作为一名吃瓜群众，我秉着好奇心点开了 NBA 新春贺岁形象大使蔡徐坤的微博，发现他的微博转发量除了最新一条之外，其它的基本每条都是转发 100 万+。

为了解答这个问题，我们通过微博移动端随机抓取了蔡徐坤的一条微博《再见，“任性的”千千…》的10万条转发数据（时间节点2021年1月15日10时）。这条讲诉蔡徐坤关爱小动物的短视频微博，于2019年3月9日01:23发布，到2019年3月10日18:00的时候已经被转发100万+。

# 2 开发环境和开发技术

2.1 硬件平台

CPU：Intel(R) Core(TM) i7-8750H CPU @ 2.20GHz (8 CPUs) ~2.1GHz

内存：16G

硬盘空间：SSD-1T

2.2 Jupyter Netbook介绍

Jupyter Notebook（此前被称为 IPython notebook）是一个交互式笔记本，支持运行 40 多种编程语言。

Jupyter Notebook 的本质是一个 Web [应用程序](https://baike.baidu.com/item/%20%E5%BA%94%E7%94%A8%E7%A8%8B%E5%BA%8F/5985445" \t "_blank)，便于创建和共享文学化程序文档，支持实时代码，数学方程，可视化和 [markdown](https://baike.baidu.com/item/markdown/3245829" \t "_blank)。 用途包括：数据清理和转换，数值模拟，统计建模，机器学习等等

2.3 python语言介绍

Python由荷兰数学和计算机科学研究学会的Guido van Rossum 于1990 年代初设计，作为一门叫做[ABC语言](https://baike.baidu.com/item/ABC%E8%AF%AD%E8%A8%80/334996" \t "_blank)的替代品。Python提供了高效的高级[数据结构](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E7%BB%93%E6%9E%84/1450" \t "_blank)，还能简单有效地[面向对象](https://baike.baidu.com/item/%E9%9D%A2%E5%90%91%E5%AF%B9%E8%B1%A1/2262089" \t "_blank)编程。Python语法和动态类型，以及[解释型语言](https://baike.baidu.com/item/%E8%A7%A3%E9%87%8A%E5%9E%8B%E8%AF%AD%E8%A8%80/8888952" \t "_blank)的本质，使它成为多数平台上写脚本和快速开发应用的编程语言，随着版本的不断更新和语言新功能的添加，逐渐被用于独立的、大型项目的开发。

Python[解释器](https://baike.baidu.com/item/%E8%A7%A3%E9%87%8A%E5%99%A8/10418965" \t "_blank)易于扩展，可以使用C或[C++](https://baike.baidu.com/item/C%2B%2B/99272" \t "_blank)（或者其他可以通过C调用的语言）扩展新的功能和[数据类型](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E7%B1%BB%E5%9E%8B/10997964" \t "_blank)。Python 也可用于可定制化软件中的扩展程序语言。Python丰富的标准库，提供了适用于各个主要系统平台的源码或[机器码](https://baike.baidu.com/item/%E6%9C%BA%E5%99%A8%E7%A0%81/86125" \t "_blank)。

2.4 GitHub介绍

GitHub是一个面向[开源](https://baike.baidu.com/item/%E5%BC%80%E6%BA%90/20720669" \t "_blank)及私有[软件](https://baike.baidu.com/item/%E8%BD%AF%E4%BB%B6/12053" \t "_blank)项目的托管平台，因为只支持Git作为唯一的版本库格式进行托管，故名GitHub。

GitHub于2008年4月10日正式上线，除了[Git](https://baike.baidu.com/item/Git/12647237" \t "_blank)代码仓库托管及基本的Web管理界面以外，还提供了订阅、讨论组、文本渲染、在线文件编辑器、协作图谱（报表）、代码片段分享（Gist）等功能。目前，其注册用户已经超过350万，托管版本数量也是非常之多，其中不乏知名开源项目[Ruby](https://baike.baidu.com/item/Ruby/11419" \t "_blank) on Rails、[jQuery](https://baike.baidu.com/item/jQuery/5385065" \t "_blank)、[python](https://baike.baidu.com/item/python/407313" \t "_blank)等。

2018年6月4日，微软宣布，通过75亿美元的股票交易收购代码托管平台GitHub。

2.5 PyCharm开发工具介绍

PyCharm是一种[Python](https://baike.baidu.com/item/Python/407313" \t "_blank) IDE，带有一整套可以帮助用户在使用Python语言开发时提高其效率的工具，比如[调试](https://baike.baidu.com/item/%E8%B0%83%E8%AF%95/5852756" \t "_blank)、[语法高亮](https://baike.baidu.com/item/%E8%AF%AD%E6%B3%95%E9%AB%98%E4%BA%AE/9686751" \t "_blank)、Project管理、代码跳转、智能提示、自动完成、单元测试、版本控制。此外，该IDE提供了一些高级功能，以用于支持Django框架下的专业Web开发。

# 3 概要设计

### 3.1 函数库的导入

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

### 3.2 函数库的注解

pandas 是基于[NumPy](https://baike.baidu.com/item/NumPy/5678437" \t "_blank) 的一种[工具](https://baike.baidu.com/item/%E5%B7%A5%E5%85%B7/81891" \t "_blank)，该工具是为解决数据分析任务而创建的。Pandas 纳入了大量库和一些标准的[数据模型](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E6%A8%A1%E5%9E%8B/1305623" \t "_blank)，提供了高效地操作大型数据集所需的工具。pandas提供了大量能使我们快速便捷地处理数据的函数和方法。你很快就会发现，它是使Python成为强大而高效的数据分析环境的重要因素之一。

NumPy（Numerical Python）是Python的一种开源的数值计算扩展。这种工具可用来存储和处理大型矩阵，比Python自身的嵌套列表（nested list structure)结构要高效的多（该结构也可以用来表示矩阵（matrix）），支持大量的维度数组与矩阵运算，此外也针对数组运算提供大量的数学函数库。

Matplotlib 是一个 [Python](https://baike.baidu.com/item/Python/407313" \t "_blank) 的 2D绘图库，它以各种硬拷贝格式和跨平台的交互式环境生成出版质量级别的图形。通过 Matplotlib，开发者可以仅需要几行代码，便可以生成绘图，直方图，功率谱，条形图，错误图，散点图等。

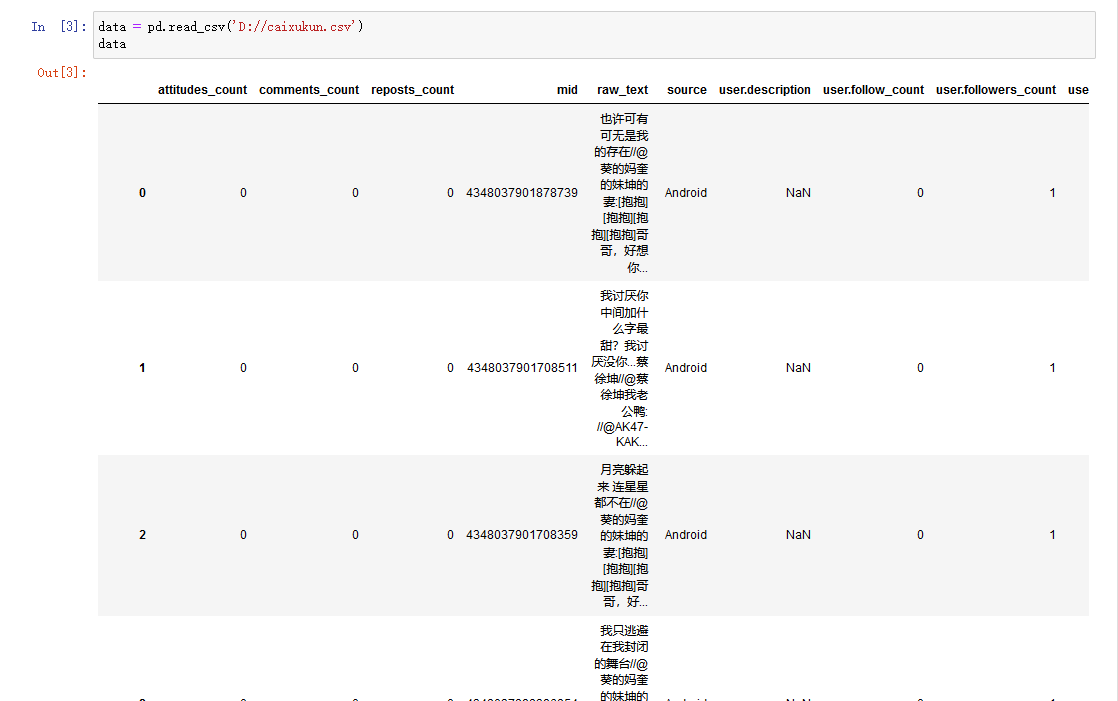
Seaborn其实是在matplotlib的基础上进行了更高级的API封装，从而使得作图更加容易，在大多数情况下使用seaborn就能做出很具有吸引力的图，而使matplotlib就能制作具有更多特色的图。

### 3.3 csv文件的读取



### 3.4 数据的清洗

在使用数据之前，先对数据进行一下清洗。



### 3.5 蔡徐坤的微博转发是否存在假流量？

3.5.1数据的抽样

随机抽取10个蔡徐坤粉丝转发的数据进行分析。

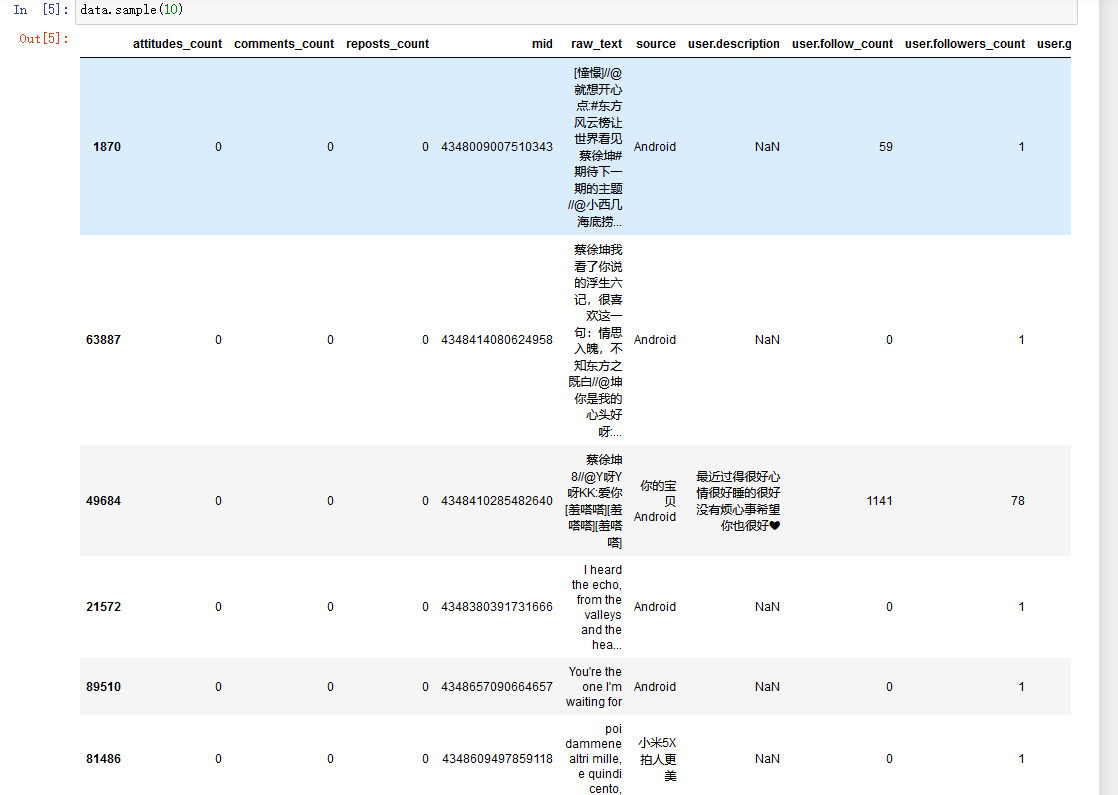


图3.5.1 数据的抽样

3.5.2蔡徐坤的粉丝性别比例

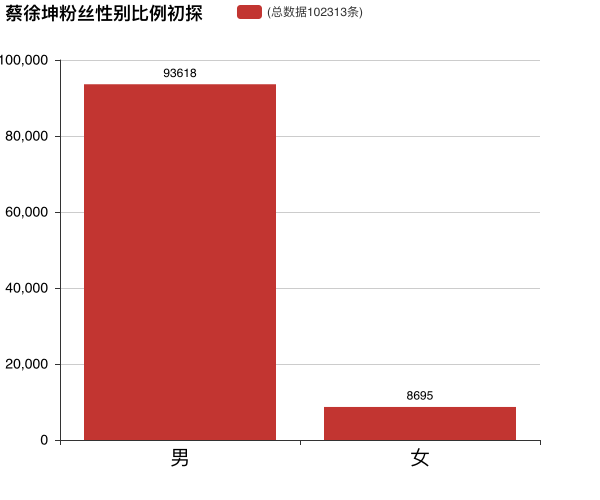


图3.5.2 蔡徐坤粉丝的性别比例

通过柱状图可以清晰的发现，蔡徐坤的男粉丝数跟女粉丝数比例差距过大，所以进一步怀疑，蔡徐坤存在流量造假行为。

3.5.3蔡徐坤男性粉丝数据

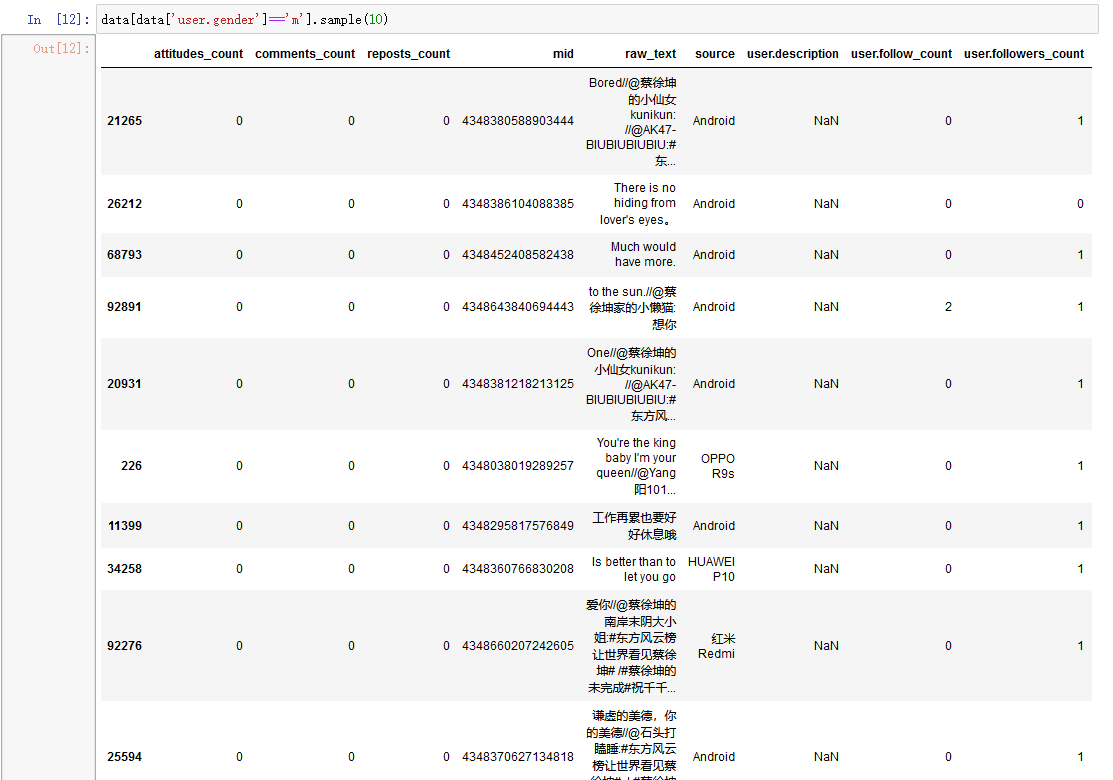


图3.5.3 蔡徐坤男性粉丝数据

随机抽取10个男性粉丝数据，发现大部分人的粉丝数为0，关注人为1。

### 3.6 真假流量所占的比例各有多少？

3.6.1蔡徐坤真假流量转发量



图3.6.1 蔡徐坤真假流量转发量

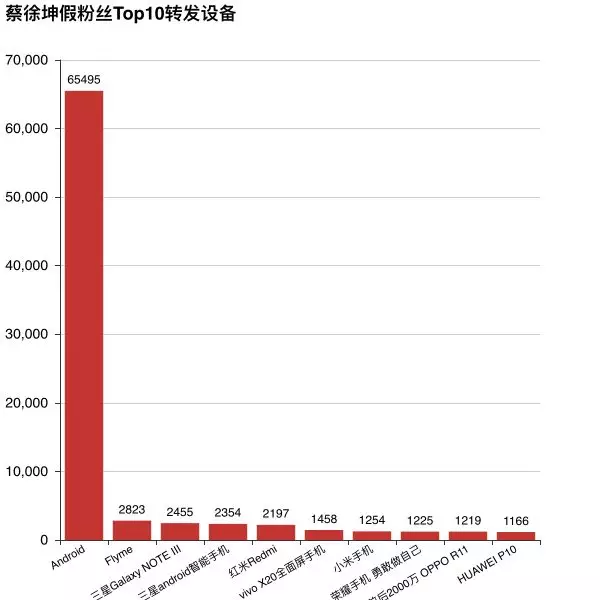
通过上图可以发现，蔡徐坤假粉丝转发量远远大于真粉丝的转发量。

### 3.7 假流量粉丝是如何生产出来的？

3.7.1蔡徐坤假粉丝Top10转发设备

图3.6.1 蔡徐坤假粉丝Top10转发设备

通过这个可以看出，蔡徐坤假粉丝的转发设备大都为Android设备。



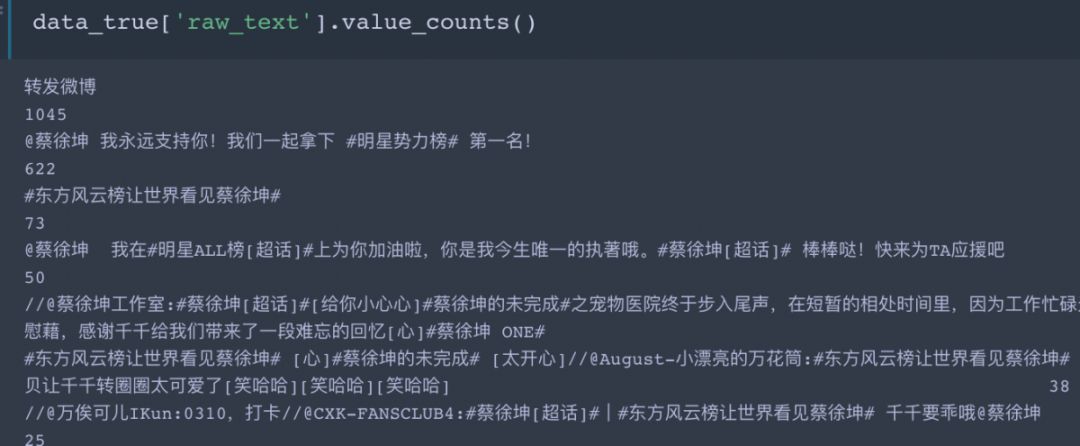
3.7.2其他分析结果

另外还有一些有趣的发现：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）。



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

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。



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。

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。

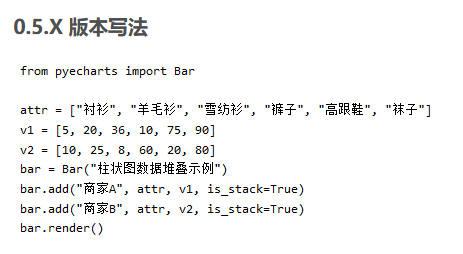
数据显示，蔡徐坤动辄100万+的微博转发中，确实存在绝大部分的假流量。这些假流量估计有两方面的来源：一是自身经纪公司购买的，二是忠实的粉丝自费购买的。

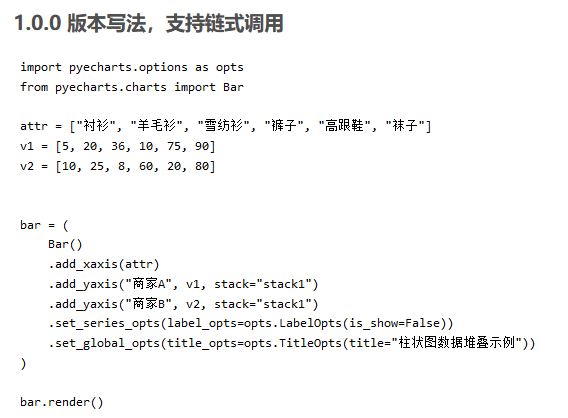
若是自身经纪公司购买的话，这确实扰乱了整个娱乐圈市场的运作，对于娱乐圈乃至整个社会风气都是不好的。若是忠实粉丝购买的，Alfred觉得，数据只是一个数据，这其中的资金，可以通过其它方面更好的方面去给自己的偶像添加影响力呀。我看最近蔡徐坤转发的#脱贫攻坚战星光行动# 这个话题就很好呀，通过自己的影响力，去做更多正能量的事情。

# 4详细设计

在这次的团队合作课设中，我主要是负责ppt的一些制作和数据清洗，其中关于GitHub也掌握了一些基本的操作，而对于python语句则有有一些一下的收获：

Pyecharts版本





Sample()函数

作用：数据抽样

Shape()函数

作用:返回数据的行数和列数

shape[0]返回的是数据的行数，有几行

shape[1]返回的是数据的列数，有几列

str.contains()函数

作用;查询每行是否包含指定的字符串

drop()函数

DataFrame.drop(labels=None, axis=0, index=None, columns=None, level=None, inplace=False, errors='raise')

其中常用的参数有：

labels：标签或列表

axis：按那个方向检索 0：按逐行索引向下走   1;按逐列 向右走

columns： 列名

作用：删除行，删除列

value\_counts()函数

作用:对数据里面的每个值进行计数并且排序

mean()函数

Series.mean(axis=None,skipna=None,level=None,numeric\_only=None, \*\*kwargs)

axis:要应用的功能的轴。

skipna:计算结果时排除NA /null值。

level:如果轴是MultiIndex(分层)，则沿特定级别计数，并折叠成标量。

numeric\_only:仅包括float，int，boolean列。

\*\*kwargs:要传递给函数的其他关键字参数。

作用:求平均值

str.contanis函数

Series.str.contains（pat，case = True，flags = 0，na = nan，regex = True)

Pat:str类型 字符序列或正则表达式。

case :bool，默认为True,如果为True，区分大小写。

Flags:int，默认为0（无标志） 标志传递到re模块，

Na:默认NaN 填写缺失值的值。

Regex:bool，默认为True

如果为True，则假定pat是正则表达式。

如果为False，则将pat视为文字字符串。

作用：筛选数据

**课程设计成绩评定表**

|  |  |  |  |
| --- | --- | --- | --- |
| 课程设计题目 | 蔡某坤粉丝数及转发数据真假状况分析 | | |
| 课程设计学生答辩或质疑记录：   1. 数据清洗的目的是什么？   答：数据清洗的目的有两个，第一是通过清洗让数据可用。第二是让数据变的更适合进行后续的分析工作。换句话说就是有”脏”数据要洗，干净的数据也要洗。  2．在做课设的过程中遇到了什么问题，是怎样解决的？  答：遇到了一个pyecharts版本问题，因为电脑下载的是1.9.0的版本，所以导致我的语句有问题，解决方法就是把pyecharts版本换成了0.5.5的版本。  3．数据表中的重复值怎么处理。  答：pandas中有两个函数是专门用来处理重复值的，第一个是duplicated函数。Duplicated函数用来查找并显示数据表中的重复值。第二个是drop\_duplicates函数用来删除数据表中的重复值，判断标准和逻辑与duplicated函数一样。 | | | |
| **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 | | | |