

# Assignment 2

姓名：王昱

学号：18210980068

专业：应用统计

1. Find/design 5 sets of different data, and use 5 different types of plots to visualize the data using Python and matplotlib; please take a few sentences to describe the data information, background, and visualization effects for analysis. Submit your 5 data sets and code.

## 1) 饼图

本题选取直播数据集（1\_dataset.csv），共有 1626 个样本，指标为直播类型。直播类型数据共有四个类别，分别为：网游竞技、娱乐综艺、手游休闲和单机热游。各个类别数量如图 1.1。

直播类型	
网游竞技	792
娱乐综艺	373
手游休闲	311
单机热游	150

图 1.1 直播类型数量统计

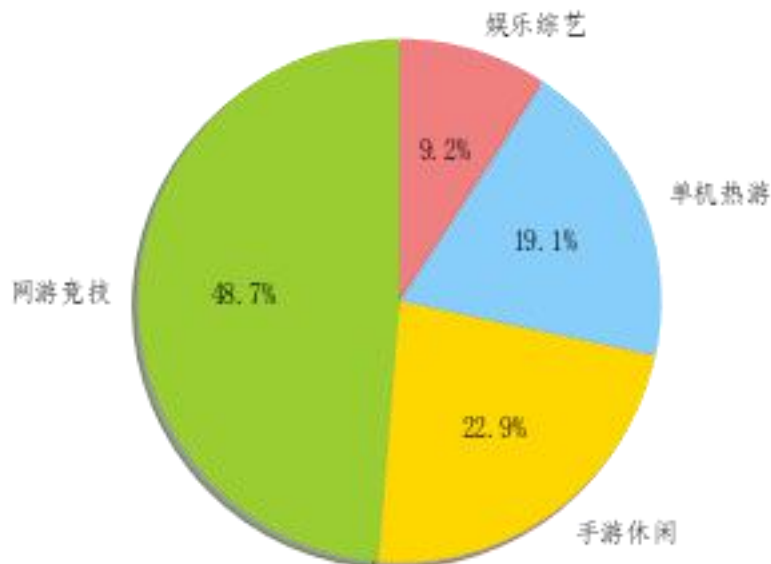


图 1.2 直播类型统计饼图

由图 1.2 可知，网游竞技类型的直播数量最多，占 48.7%。娱乐综艺类型的直播数量最少，占 9.2%。单机热游和手游休闲类型占比相差不大，均在 20%左

右。在直播平台上，网游竞技类更加具有竞争力。

2) 堆叠条形图

本题选取安卓手机游戏数据（2\_dataset.csv），共有 527 个样本，变量为：类别、评论数、评分。将评分大于 7.4 的赋值为 1，将评分小于等于 7.4 的赋值为 0。前 5 行数据如下表：

	类别	评论数	评分	评分分类
0	益智休闲	6654	7.6	1
1	益智休闲	7187	5.9	0
2	动作游戏	9372	5.1	0
3	益智休闲	8804	8.4	1
4	动作游戏	7371	8.2	1

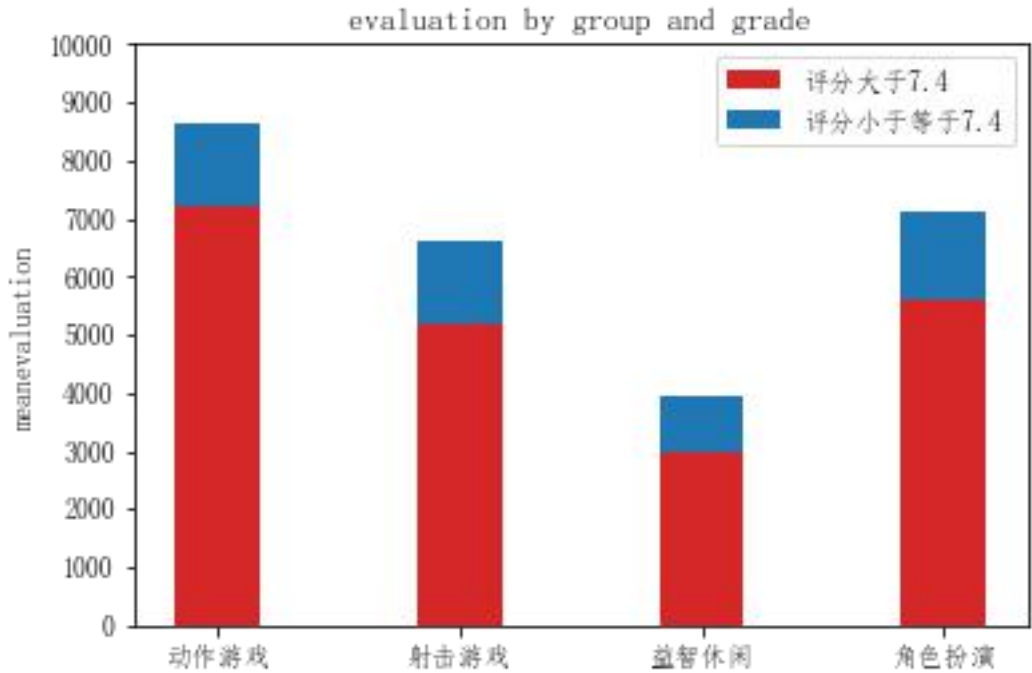


图 1.3 评论数堆叠条形图

由图 1.3 可知，评分大于 7.4 的游戏平均评论数远高于评分小于等于 7.4 的游戏评论数。这表明评论数对游戏的评分可能有正相关影响。评分小于等于 7.4 的各类游戏的平均评论数相差不大。动作类和角色扮演类的平均评论数整体较高，射击和益智休闲类的游戏平均评论数相对较少。

3) 核图

本题选取安卓手机游戏的热度为研究对象（3\_dataset.csv），共有 1140 个样本。前 5 行数据如下表：

热度	
0	95
1	94
2	93
3	90
4	89

绘制的核图如下：

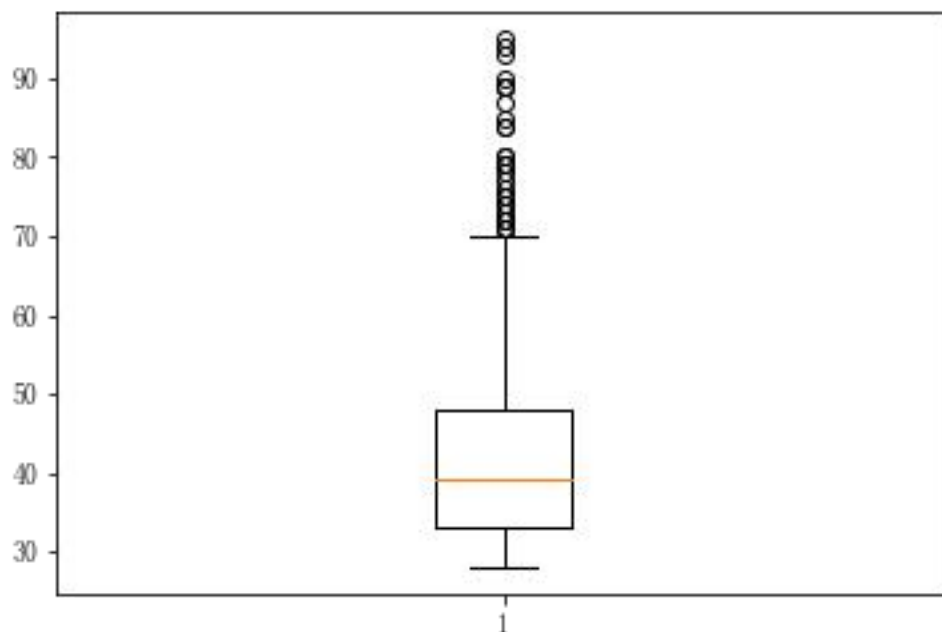


图 1.4 手机游戏热度核图

由 1.4 可知，手机游戏的热度值中位数为 39，四分之一分位点为 33，四分之三分位点为 48。手机游戏热度的最大值为 95，热度大于 70 的离群点较多。这说明，存在一部分高热度手机游戏，但整体来看热度值较低。

#### 4) 散点图

本题选取安卓手机游戏数据（4\_dataset.csv），共有 314 个样本。变量为：类别、评论数、喜欢数。游戏类别包括：策略塔防、冒险解谜、竞速游戏、模拟经营。前 5 行数据如下表：

	类别	评论数	喜欢数
0	竞速游戏	5680	3534
1	竞速游戏	14640	5039
2	竞速游戏	27449	7235
3	竞速游戏	6836	1428
4	竞速游戏	1423	249

绘制散点图如下：

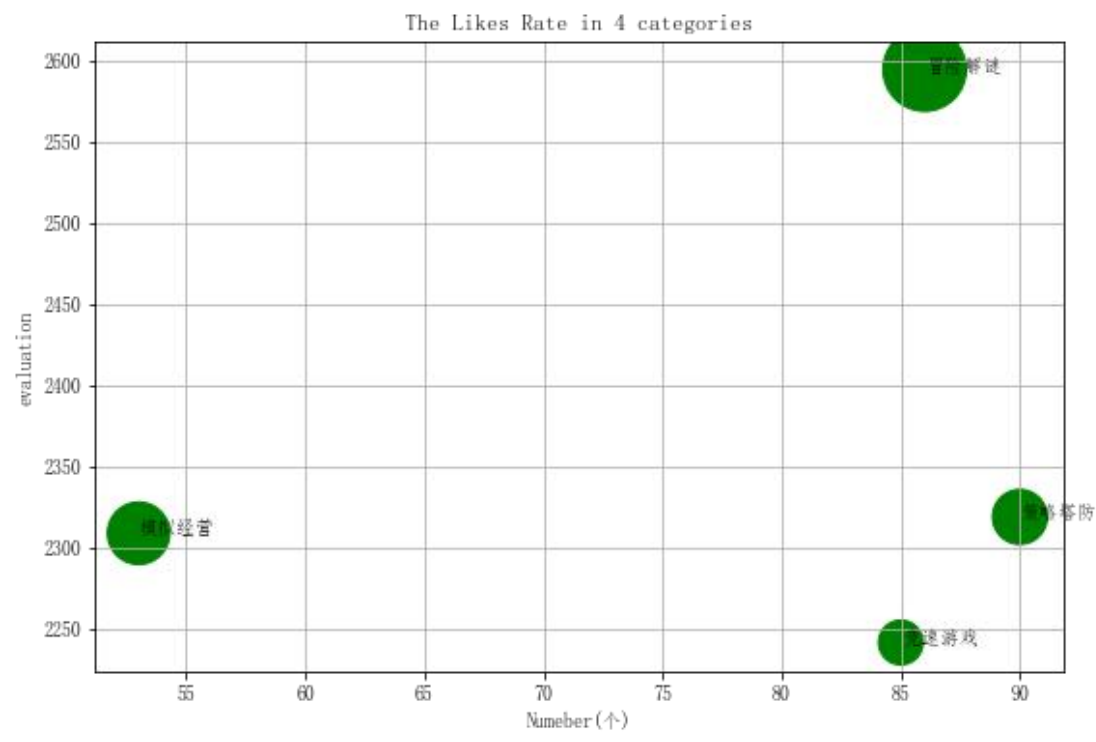


图 1.5 喜欢数的散点图

由图 1.5 可知，冒险解谜类游戏平均评论数以及平均喜欢数均为最大的。模拟经营类游戏个数最少，但评论数和喜欢数与策略塔防相差不大。竞速类游戏虽然个数多，但平均评论数和喜欢数均较低。由此可得，冒险解谜类游戏更受用户喜欢，策略塔防类游戏开发数量较多。

5) 折线图

本题选取安卓手机游戏数据（5\_dataset.csv），共有 56 个样本。变量为：最后更新时间、平均喜欢数。每个样本的最后更新时间减去所有样本中最早更新时间，得到时间差(天)。数据前 5 行如下：

	最后更新时间	平均喜欢数	时间差
0	2016-01-11	705	0.0
1	2016-01-12	380	1.0
2	2016-01-13	114	2.0
3	2016-01-14	437	3.0
4	2016-01-15	4320	4.0

绘制手机游戏喜欢数和时间的折线图。

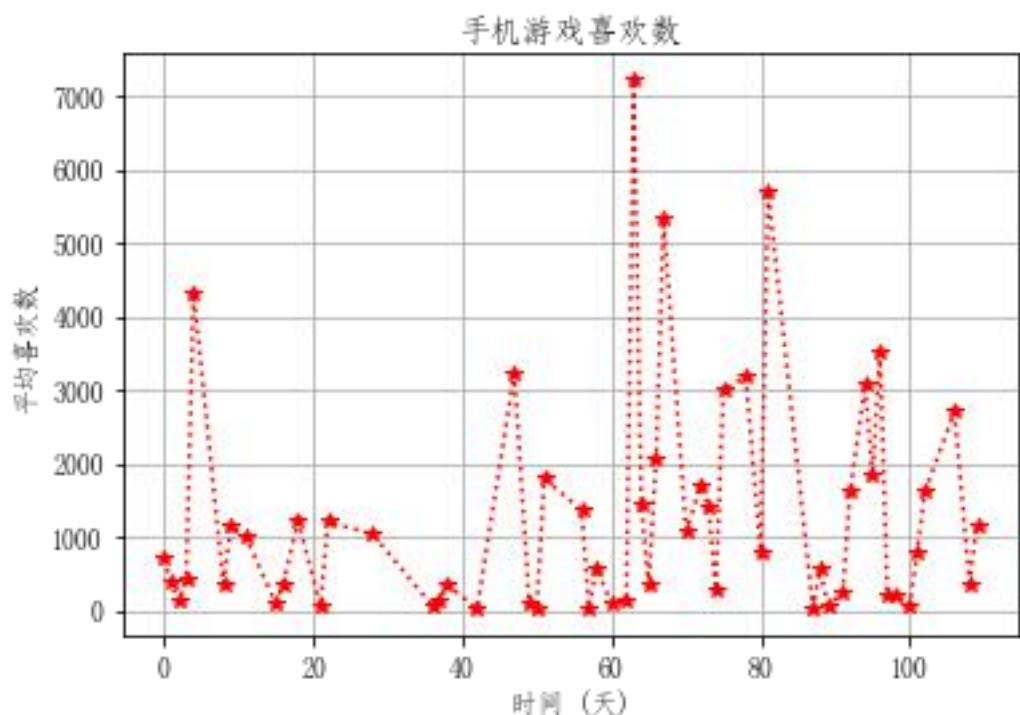


图 1.6 手机游戏喜欢数折线图

由图 1.6 可知，手机游戏喜欢数和最后更新时间没有明显的线性关系。最后更新时间较晚的手机游戏平均喜欢数较高。这说明了手机游戏的更新可能完善了用户的体验，使得用户满意度较好。

2.Find/design a dataset and visualize the data using either the techniques of animation or metaphor, or both of them. Please take a few sentences to describe the data information, background, and visualization effects for analysis. Submit your data and code if and only if you use any.（做一个画图用到动画功能或（和）隐喻可视化功能）

选取安卓手机游戏数据（gamegrade.csv），共有 20 个样本，变量分别为：游戏名称、评分。  
数据前 5 行如下：

	游戏名称	评分
0	捕鱼达人	7.6
1	三国志无双战	5.1
2	地铁跑酷	8.4
3	火箭飞人	8.2
4	暴力摩托	5.1

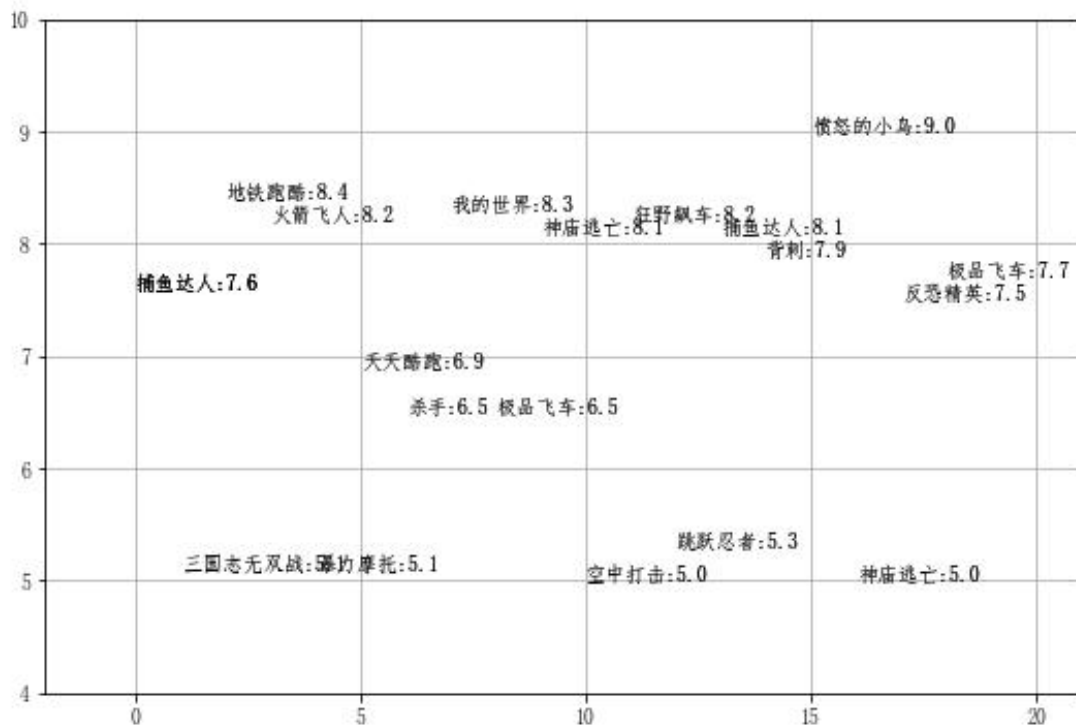


图 2.1 游戏评分分布

(PS: 此为静态图, 动图见压缩文件中)

由图 2.1 可知, 各游戏的分数差异较大。愤怒的小鸟游戏评分较高为 9.0 分, 而神庙逃亡、空中打击等游戏评分较低为 5.0 分。多数游戏的分数处于 7.5-8.4 之间。这一现象符合我们对于手机游戏评分的基本认知: 即存在少数分数较高的游戏, 但大多数游戏的评分处于中间水平。

3. Visualize the GDP values and changes of GDP of the countries (select 5-10 countries) for the past 20 years.

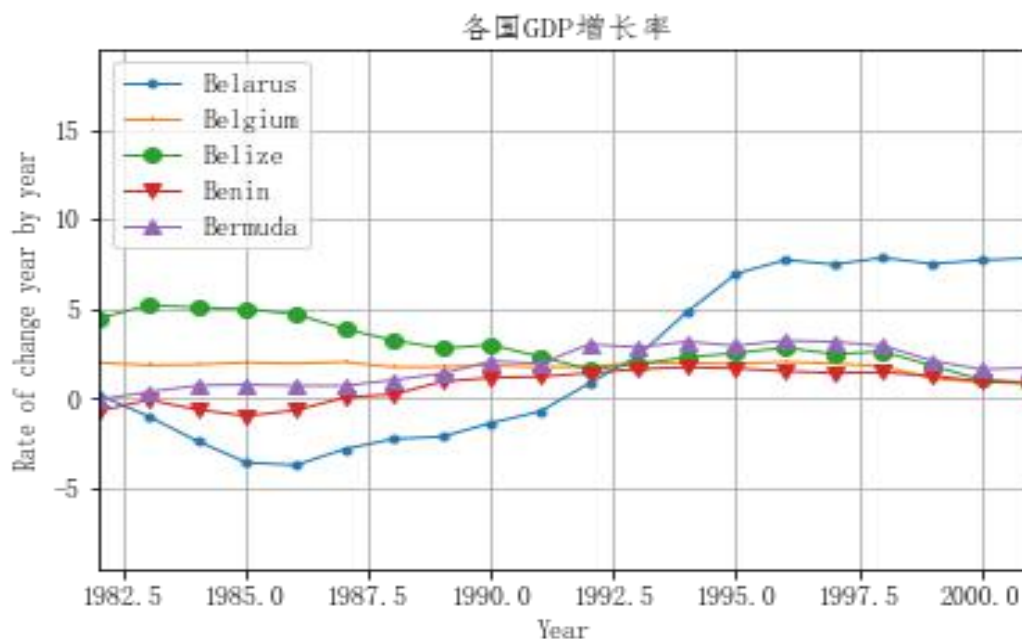


图 3.1 各国 GDP 值和变化曲线

本题数据集为 Elearning 资源（world\_history\_gdp\_data.csv）。选取了 Belarus、Belgium、Belize、Benin、Bermuda 这五个国家从 1982 年到 2002 年这 20 年的 GDP 值作为研究对象。由图 3.1 可得，Belarus 从最初 GDP 处于落后到 1992 年开始反超，成为这 5 个国家中 GDP 最高的国家。Belize 则是从一开始的 GDP 领先到之后 GDP 呈现下降趋势。在 1995 年之后，Belgium、Belize、Benin、Bermuda 这四个国家的 GDP 相差不大。