**模块B：数据分析（30分）**

**竞赛题目：电影数据分析**

**背景介绍**

在电影上映前是否能预先知道能否成功？ 是否有某些公司找到了一致的公式？电影公司制作一部新电影推向市场时，要想获得成功，通常要了解电影市场趋势，观众喜好的电影类型，电影的发行情况，改编电影和原创电影的收益情况，以及观众喜欢什么样的内容。接下来就让我们来分析一下数千部电影数据。

**数据介绍**

数据集包括美国地区一百年间（1916-2016）的电影作品。数据集(需要data/tmdb\_5000\_movies.csv和data/tmdb\_5000\_credits.csv连接)为movies.csv。

movies.csv文件： 记录着电影的类型、预算、标题、票房等相关信息。 该文件包含4803行和23列，每行对应一部电影的相关情况。

文件movies.csv中包含23个字段，具体信息如下：

| **No** | **属性** | **字段描述** |
| --- | --- | --- |
| 1 | budget | 电影预算 |
| 2 | genres | 电影类型 |
| 3 | homepage | 电影主页 |
| 4 | id | 电影ID |
| 5 | keywords | 电影关键词 |
| 6 | original\_language | 电影原始语言 |
| 7 | original\_title | 电影原始标题 |
| 8 | overview | 电影剧情摘要 |
| 9 | popularity | 电影流行度 |
| 10 | production\_companies | 电影出品公司 |
| 11 | production\_countries | 电影出品国家 |
| 12 | release\_date | 电影档期 |
| 13 | revenue | 电影票房 |
| 14 | runtime | 电影时长 |
| 15 | spoken\_languages | 电影语言 |
| 16 | status | 电影状态 |
| 17 | tagline | 电影标语 |
| 18 | title | 电影标题 |
| 19 | vote\_average | 电影平均评分 |
| 20 | vote\_count | 电影评分次数 |
| 21 | movie\_id | 编号 |
| 22 | cast | 演员阵容 |
| 23 | crew | 全体人员 |

**环境介绍**

在本竞赛模块环境中，已自动启动了 jupyter notebook 环境。环境中已安装 Python 中多种数据分析和可视化库，包括：

* Scipy
* Numpy
* Pandas
* Matplotlib
* Seaborn
* Pyecharts
* 其它

**任务列表**

1. 数据整合
2. 数据处理
3. 数据可视化

**注意事项**

每题完成之后必须保存该题结果，每题的保存代码在该题下的结果保存处，执行即可，禁止修改。

* 每题的最终结果变量名已预先设定，禁止修改
* 若没有保存结果则该题无分
* 若该题有修改需重新运行结果保存代码，否则视为上次提交

**任务启动**

**任务1：数据读取（1分）**

* 载入 movies.csv 文件，数据路径为 ../data/movies.csv
* 结果以变量data保存

**任务2：数据处理（9分）**

**任务2-1（1分）**

删除指定字段，具体要求如下：

* 删除homepage, original\_title, overview, spoken\_languages, status, tagline, movie\_id字段

**任务2-2（1分）**

增加profit字段，该字段为每部电影的收益，具体要求如下：

* 计算每部电影的收益，收益=电影票房-电影预算

**任务2-3（1分）**

处理runtime字段的缺失值，该字段有两个缺失值，具体要求如下：

* 将id为370980的缺失值填充为98
* 将id为459488的缺失值填充为81

**任务2-4（2分）**

处理release\_date字段的缺失值，具体要求如下：

* 将缺失值填充为2014-06-01
* 将release\_date字段只保留年份信息，如2014-06-01只保留2014
* 将release\_date列名改为year

**任务2-5（2分）**

分别对genres, keywords, production\_companies, production\_countries字段进行处理，具体要求如下：

* 每个字段只保留该字段内容中name键值对应的value值，并以|分隔

**提示**：例如[{"id": 28, "name": "Action"}, {"id": 12, "name": "Adventure"}, {"id": 14, "name": "Fantasy"}, {"id": 878, "name": "Science Fiction"}]，处理后变为Action|Adventure|Fantasy|Science Fiction

**任务2-6（2分）**

对crew字段进行处理，具体要求如下：

* 只保留导演的名字，并以|分隔
* 将crew字段列名改为director

提示：导演的job为Director

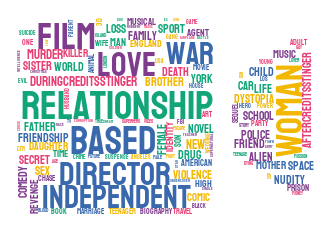
**任务3：数据可视化（20分）**

**任务3-1（2分）**

对电影的关键词制作词云图

* 成功制作词云图（1分）
* 词云图图形为video（1分）

参考图形如下：

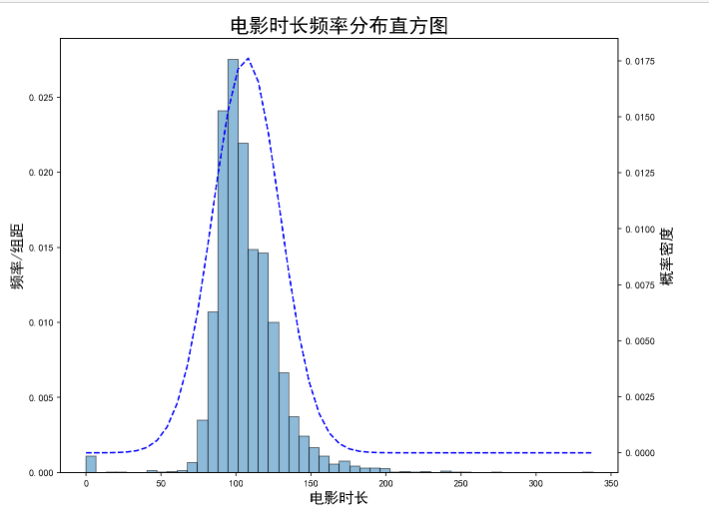


**任务3-2（2分）**

制作电影的时长频率分布直方图

* 直方图图形类型正确（0.5分）
* 直方图颜色为蓝色，透明度为 0.5，框线为黑色（0.5分）
* 概率密度曲线以虚线表示（0.5分）
* X轴和Y轴标签正确，左轴为直方图Y轴，右侧为概率密度曲线轴（0.5分）

参考图形如下：



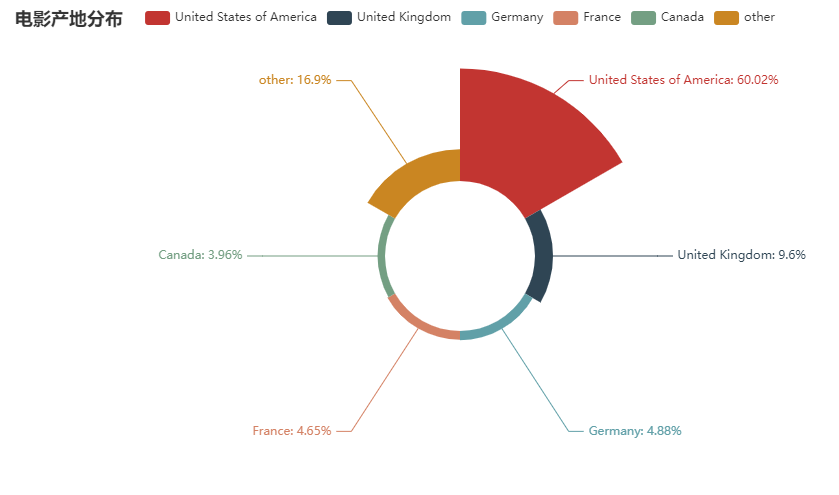
**任务3-3（2分）**

制作电影产地分布玫瑰图

* 统计不同产地的电影数量，电影数量低于前五的产地，全部归为others
* 成功制作玫瑰图（0.5分）
* 标题为电影产地分布（0.5分）
* 标签数值正确（0.5分）
* 标签为国家：百分比，百分比保留两位小数，例如：Canada: 2.53%（0.5分）

Canada: 2.53%，此数据为虚构的，真实数值需要自行计算

参考图形如下：

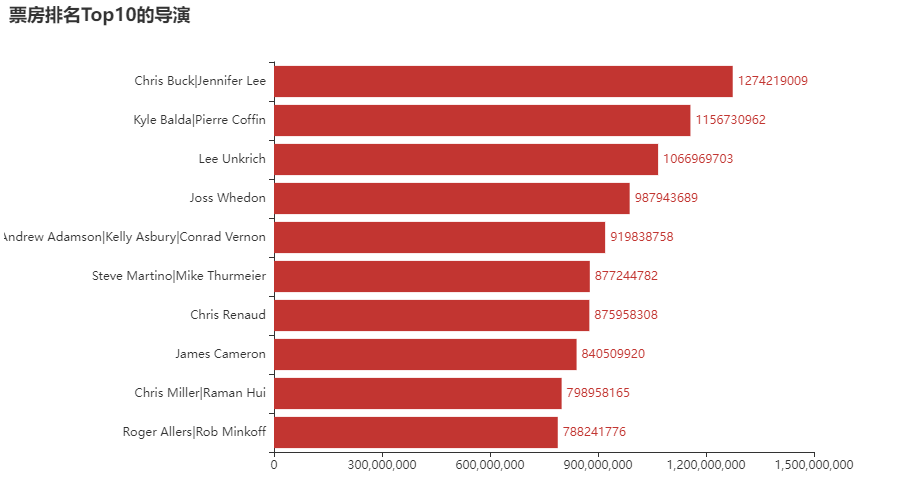


**任务3-4（2分）**

统计票房均值Top10的导演，并以条形图展示

* 票房均值以整数表示，四舍五入
* 标题为票房排名Top10的导演（0.5分）
* 标签数值正确（0.5分）
* 标签位置放置右侧（0.5分）
* 纵坐标导演名字显示完全（0.5分）

参考图形如下：

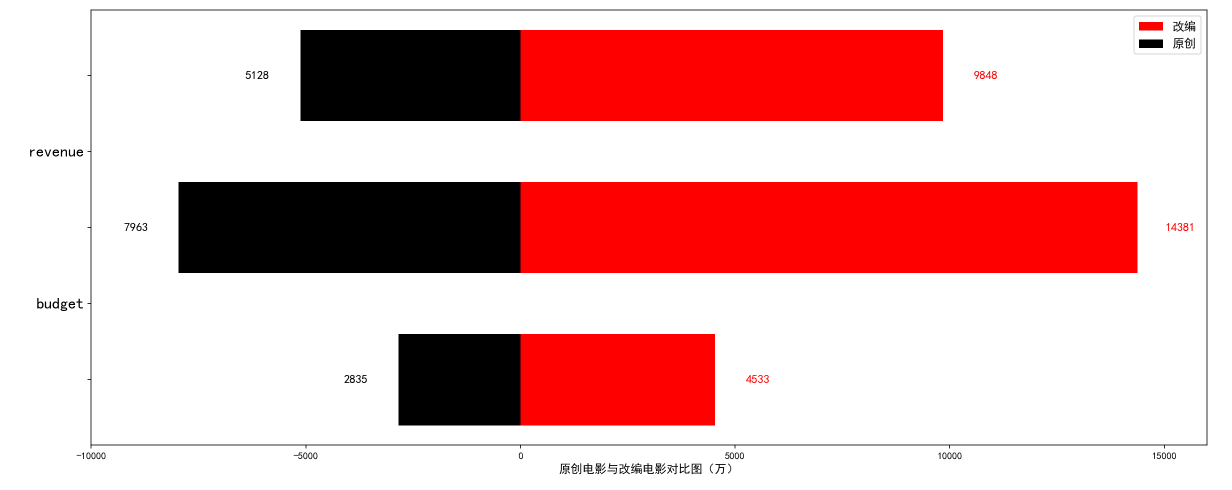


**任务3-5（2分）**

分析原创电影与改编电影在收益、票房、预算方面的表现，以旋风图展示：

* 成功制作旋风图（0.5分）
* 旋风图标签数值正确，以万为单位，取整数（四舍五入）（0.5分）
* 标题为原创电影与改编电影对比图（万）（0.5分）
* 添加图例（0.5分）

参考图形如下：

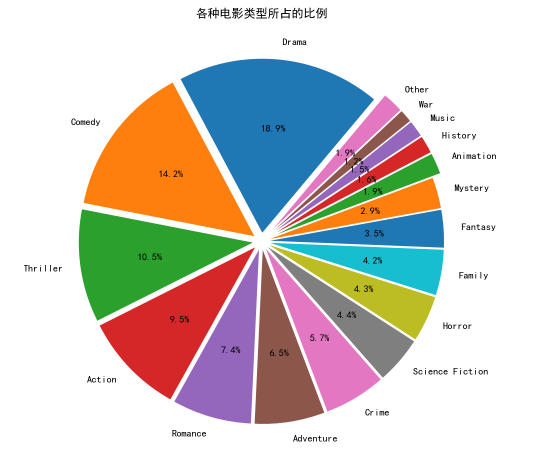


**任务3-6（2分）**

统计各种电影类型所占的比例，以饼图展示

* 设置other类，当电影类型所占比例小于%1时，全部归到other类中（0.5分）
* 所占比例小于或等于%2时，对应的饼状图往外突出一截（0.5分）
* 数值正确，百分比保留一位小数（0.5分）
* 标题为各种电影类型所占的比例（0.5分）

参考图形如下：

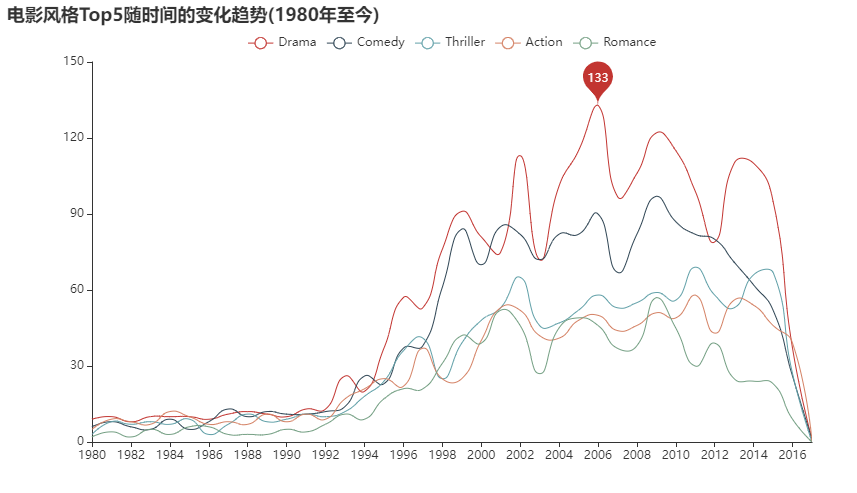
****

**任务3-7（2分）**

统计电影风格（类型）Top5随时间（1980年至今）的变化趋势，以折线图展示

* 折线图设置为顺滑模式（0.5分）
* 标题和图例不能重叠（0.5分）
* 数据起点和纵坐标没有间隙（0.5分）
* 为最大值做标记（0.5分）

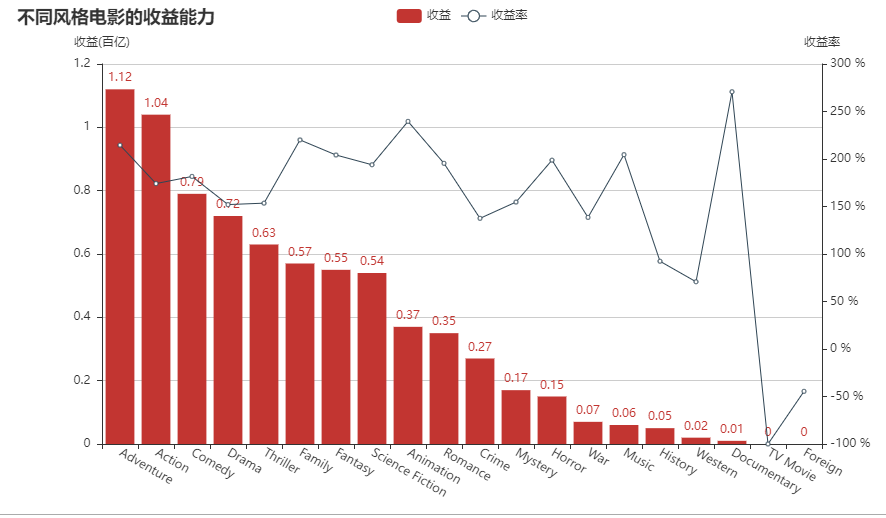
参考图形如下：



**任务3-8（2分）**

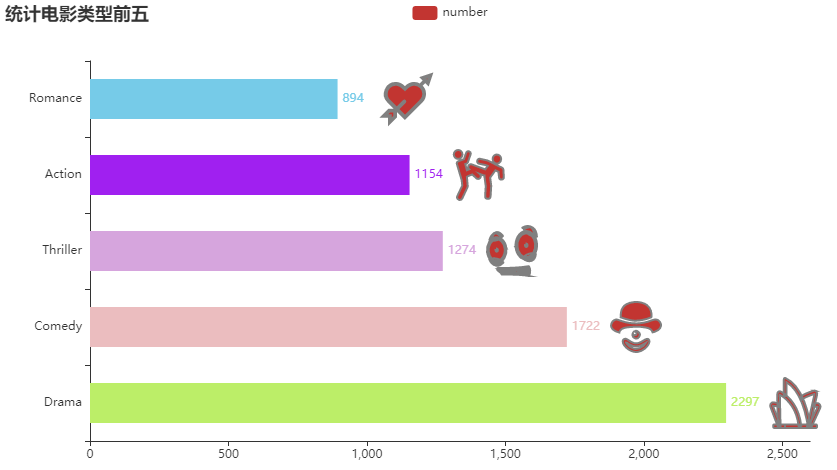
统计不同风格（类型）电影的收益能力，以条形图和折线图混合展示

* 条形图为不同电影类型的收益总和，以千亿为单位，保留两位小数（0.5分）
* 折线图为不同电影类型的收益率，收益率=收益/预算（0.5分）
* 条形图标签数值正确（0.5分）
* 折线图不显示标签数值（0.5分）

参考图形如下： 

**任务3-9（4分）**

统计电影类型Top5，以条形图展示

* 条形图颜色依次为#BCEE68, #EBBDBF, #D6A5DD, #A020F0, #76CBE8 （0.5分）
* 条形图右端添加象形图片，symbols已提前给出，象形图框线为灰色grey （2分）
* 标签数值正确（0.5分）
* 标签放置条形图右侧，显示完全，不能被遮挡（1分）
* 参考图形如下：  
  ​

**任务结束**

恭喜完成本次比赛，祝您取得好成绩。