基于 IMDb 电影数据的分析

刘嫣然 杨 杉

四川大学锦城学院 四川 成都 611731

【摘 要】IMDb(internetmoviedatabase 中国互联网在线电影信息资料库): 这是一个关于在线电影中的演员、电影、电视节目、电影明星、电子游戏以及其他电影电视制作的大型在线电影数据库，文中采用了 5 种方法进行数据分析: 得到有效数据、进行数据透视表分析、运用数据分析工具库、进行图表可视化分析、SPSS 探索频率分析数据; 从而得到了 IMDb 中各类数据之间的关系，为 IMDb 能够在以后的发展中提出一些建议。

【关键词】SPSS 探索频率分析；数据透视表，图表可视化

1 引言

随着现在互联网的发展，各种网络媒体也日新月异，在线观看电影的平台也越来越多，电影成为了大部分当代年轻人度过空闲时间的方式,不用花大量的时间，同时可以放松自己，让自己有充足的时间休息，因为电影的时长短，能在短时间内观看完毕，剧情相对紧凑精彩，能带给观看者更好的体验，不像电视剧那样需要花费大量时间、又或者是更新时间间隔太长。在现在这个网络时代，大家对于通过在线网上观看影视电影的各种平台的普及使用也越来越频繁，所以我对于 IMDb 这个平台电影中的演员、电影、电视节目、电影明星、电子游戏以及关于电影影片制作的各种在线电影数据库等都进行做了相应的统计分析，因为每个电影人都会拥有自己非常喜欢的各种电影演员类型、电影制作演员、电影制作导演等等，这样我可以更好地去了解到大家的喜好，知道对于哪一方面等感兴趣。

2 研究思路

IMDb 影明星、电子游戏以及其他在线电影相关视频内容制作的大型专业免费在线电影资讯数据库，它已经完全包括了海量有关在线影片的众多专业相关资讯信息，如：在线电影中的演员、片长、评论、内容简介等，有太多的人使用这个网站，同时我想了解一下国外的电影网站是怎么样的，他们喜欢什么类型的电影、喜欢什么类型的电影，本数据报告以 IMDb-Movie 为数据集，通过相关的指标对电影进行分析，具体指标包括：电影时长和评分分布，评分平均数，导演人数，演员人数等，来得到最后的分析结果。

3 数据说明

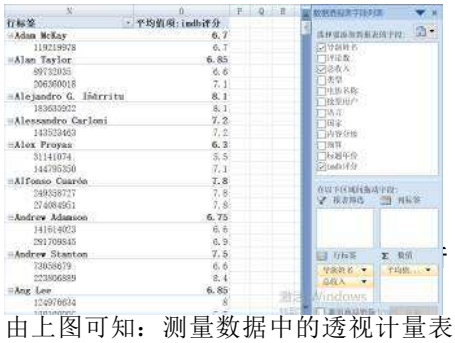
3.1 数据来源

数据来源于是来自从聚数力量（http://www.dataju.cn/Dataju/web/datasetInstanceDetail/226）中得到 IMDb 五千部电影数据。IMDb1990 年至 2016 年 17 年一部分数据，因为原始数据有 5000 行，我只选取了 300 行数据进行数据分析的演示。

3.2 数据清洗

在处理 IMDb 数据的时候，我会遇到很多空白单元格这样的无效数据，在我们后面的数据分析中会造成严重的误差和影响，所以我在分析数据之前就要把相应的空白单元格删除、将空白单元格这一行或者这一列删除，我们可以通过定位的方式现将空白单元格找出来，然后再进行统一删除；然后我们还要将重复值删除，在大量的数据中难免会有很多重复的数据，我们需要用到删除重复选项来进行数据规范。

4 IMDb 电影数据分析



由上图可知：测量数据中的透视计量表其实是一种交互式的计量表，可以用来进行某些量的计算，如数据求和与差的计数等。所需要进行的数据计算与处理数据跟整个数据库的透视或图表中的数据排列顺序有关。之所以它被称为一种数据级的透视式图表，是因为它们可以非常动态地随时改变它们的整体版面格局布置，以便按照不同显示方式重新分析相关数据，也就是可以随时重新安排诸如行号、列标和页标等字段。每一次用户改变新的版面网页布置时，数据分析透视和报表系统会立即按照新的网页布置重新开始计算相关数据。另外，如果你的原始数据没有发生任何更改，则用户可以手动更新原始数据回到透视器列表。我想要得到的数据是每一位导演从 1990 年至 2016 年 17 年的总收入有哪些，并且同时得到 IMDb 评分的平均值；这样我们就可以大概看出哪一位导演的总收入更多，他所拍出来的电影评分更高，作品更优秀。

由上图可知：数据可视化，对于数据的视觉表现形式，它是关于对数据的视觉表现形式进行研究，其中，这种对于数据的视觉表现形式被明确地定义为，一种以特殊的概要方法或者简称描述方式进行抽提得到的信息，包含了相应的信息单元的各种属性和变量，数据的视觉表现形式与信息图型、信息视觉表现形式、科学视觉表现形式密切联系。我想要得到 IMDb1990 年至 2016 年 17 年每一年的总收入与评论平均分的情况，想了解到年份与评分平均分是否有关系、年份与收入又是否有关系。我使用的方法是先运用数据透视表，将我需要的数据中整理统计出来，然后在五一组合图的方式得到相应的图表，进行图表可视化数据分析。

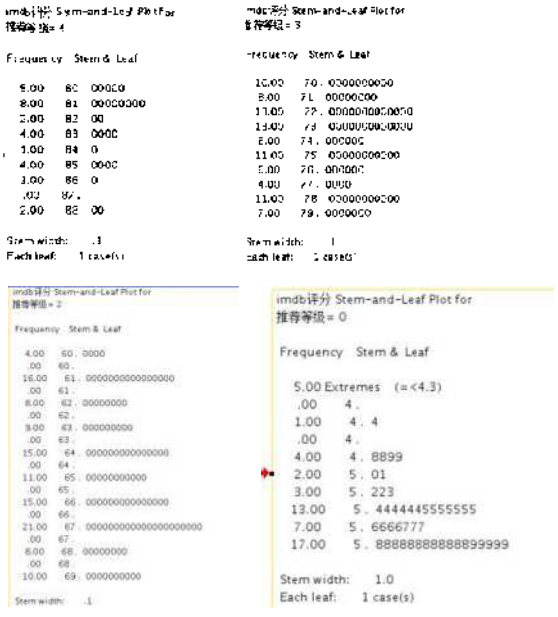
4.2 不同年份中的总收入与评论平均分分析

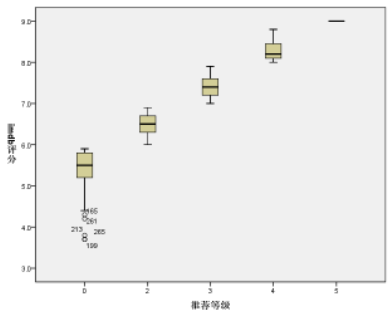
4.3 评论数、总收入、预算、标题年份、IMDb 评分数据分析

评论数		总收入		预算		标题年份		imdb评分	
平均	333.7412587	平均	176875551.5	平均	150934265.7	平均	2008.657343	平均	6.744755245
标准误差	9.339069377	标准误差	6897388.974	标准误差	2399173.538	标准误差	0.320761951	标准误差	0.055169198
中位数	301	中位数	145547603.5	中位数	145000000	中位数	2010	中位数	6.7
众数	284	众数	#N/A	众数	150000000	众数	2013	众数	6.7
标准差	157.9379942	标准差	116645431.8	标准差	40573706.13	标准差	5.424576814	标准差	0.932995803
方差	24944.41001	方差	1.36062E+16	方差	1.64623E+15	方差	29.42603362	方差	0.870481168
峰度	0.155956265	峰度	3.823224391	峰度	0.553174368	峰度	0.020127722	峰度	0.436036082
偏度	0.742222317	偏度	1.564756428	偏度	0.769266588	偏度	-0.7455839	偏度	-0.344071116
区域	751	区域	759840421	区域	262000000	区域	26	区域	5.3
最小值	62	最小值	665426	最小值	38000000	最小值	1990	最小值	3.7
最大值	813	最大值	760505847	最大值	300000000	最大值	2016	最大值	9
求和	95450	求和	50586407737	求和	43167200000	求和	574476	求和	1929
观测数	286	观测数	286	观测数	286	观测数	286	观测数	286
置信度(95.0%)	18.38230077	置信度(95.0%)	13576286.19	置信度(95.0%)	4722347.356	置信度(95.0%)	0.631362979	置信度(95.0%)	0.108590777

由上图可知：以评论数为例子来进行数据分析说明：平均值反映了数据的平均水平，数据为 333.74；标准误差是指样本平均值的“抽样误差”，数据为 9.34；中位数是对数据趋中性的一种描述，是样本中数据从小到大排列后的中间值，数据为 301；众数角度是测量样本分布数据中平均出现次数频率最高的一个数值，数据为 284；数据标准峰度偏差系数是衡量所选一个样本的标准差，是用来衡量样本数值分布相对于其一个平均值的数据离散偏大程度的一个指标，数据方差为 157.94；方差系数是衡量标准峰度偏差的平方，同样它也是用来描述样本数据分布离散偏大程度的一个指标，数据方差为 24944.41；数据峰度偏差是用来刻画衡量测度样本数据分布陡缓方向程度的一个指标，数据偏度为 0.15596，表示是因为，若数据峰度 >0，则可能说明其数据分布较正的标准正态数据分布在该曲线更尖锐，也可能就是说明数据更向一个平均值下方聚集，属于一条尖峰角度分布；数据偏度也可能就是数据偏态系数，也可简称不对称度，是衡量测度样本数据分布的一条偏斜角度方向和偏大程度的一个指标，数据偏度为 0.74222，表示是因为，若数据偏度 >0，则可能说明其数据分布较正态数据分布在该曲线更向右向左偏，称为正偏或右向左偏，说明数据存在偏大的一个极端绝对值，有点像一条黑色长尾巴被拖在其上分布在该曲线的的最右端；偏大程度的一个绝对值越大，说明测度数据分布在该曲线的数据偏斜方向程度就越大，偏度 =0 是就是无偏斜的情况；最大值为整个数据系列中数值最大的一个，数据为 813；最小值为数据系列中数值最小的一个，数据为 62；求和表示数据之和；观测数表示分析数据的数量为 286；置信水平表示样本数据的数值落在某一区间的概率。

4.4SPSS 探索总收入与推荐等级分析





由上图可知：SPSS 中的探索分析是一个能够一次性地将数据结果呈现出来，其中茎叶图是一种以连续变量来进行描述的一种手法，主要包括了三个部分，分别是频率、茎和叶；茎和叶也都有不同的含义，茎代表着我们观测值的十位数、而叶则代表着我们观测值的个位数，每一个个位数都代表一个观测值，每一行左边的频率就是相对应的个案数，茎叶组成的数值结合起来乘以茎宽就叫做数据的值，茎叶图它不仅保留了数据的频率分布，同时还保留了我们原有的数据；箱型图，它可以显示出最小值、最大值、中位数、第一个四分位数和第三个四分位数这五个数值。通过网购茎叶图和箱型图可以得出：当推荐等级为 4 时，IMDb 的推荐分数都在 8-9 分左右；当推荐等级为 3 时，IMDb 的推荐分数都在 7-8 分左右；当推荐等级为 2 时，IMDb 的推荐分数都在 6-7 分左右；当推荐等级为不推荐时时，IMDb 的推荐分数都在 4-6 分左右。

5 结论及建议

5.1 结论

IMDb 拥有大量的数据，也有足够的顾客和消费者，参与评论的人也很多，并且 IMDb 评论平均分非常地准确，具有很强的参考价值，在内容方面也不错，详细具体，想要查询和了解的资料基本上都有，是一个优秀的网站。对于运用到的 Excel 内容进行总结：在数据分析之前一定要进行数据清洗与规范整理，这样才能确保数据的准确性，以免将无效数据加在最后结果中，进行错误分析,在进行数据分析的时候可以进行多方面的分析，不只是数据透视表，组合图的形式，还有更多其他的形式，比如通过 SPSS 来进行数据分析，节约时间，直接得到我们想要的结果，让我们能更加直观地得到结论进行结论分析。

5.2 建议

希望 IMDb 能够多上映一些其他国家的电影，每个国家都有优秀的代表作品，而不只是美国国内的电影，可以拓展延伸网站的涉及地域,使自己的网站更加全面，得到更多的数据支持，这样也能吸引更多的观看者，流量数据越来越多。 IMDb 不仅仅是一个电影和其他一些电子游戏的信息数据库，还让你可以在网络上提供每天更新的一些电影和电视新闻，还让你可以在网络上搜集每一位人对电影感兴趣的话题和方向，对每一位人进行智能化的推荐，以及针对不同的电影节庆活动，提高大家对这些电影节庆活动的认识和了解，同时也可以提高自己对电影的知名度；还让我们可以开展更多其他领域的活动，比如 IMDb 的一些论坛也很积极，除了每个数据库中的一些条目都配备了一块留言板之外，还配备了关于许多种类型和各种形式的综合性主题讨论版，可以引领大家去参与讨论，组织一些有奖回答的活动方式、或者匿名建议的方式，这样 IMDb 自己也可以收集大量观看者的真实建议信息，便于改进、不断发展进步。

【参考文献】

[1] 樊潮，秦娥．基于最新 Python 的 Excel 应用浅析 [J]. 计算机时代，2021(04):69-71+75.

[2] 初道忠，陈瑞鑫．Excel 规划求解在数据分析与处理中的应用 [J]. 福建电脑，2021,37(03):104-106.

[3] 杨丽芳．以“数说我的国”项目为载体的 Excel 电子表格与数据分析单元的教学设计 [J]. 科教文汇（上旬刊），2020(12):137-138.

[4] 张燕．基于 Excel 的成绩管理与分析 [J]. 信息与电脑（理论版），2020,32(20):26-28.

[5] 马道京，陈有源，宋海波．大数据时代 Excel 的基本应用概述 [J]. 科技风，2020(25):63-64.