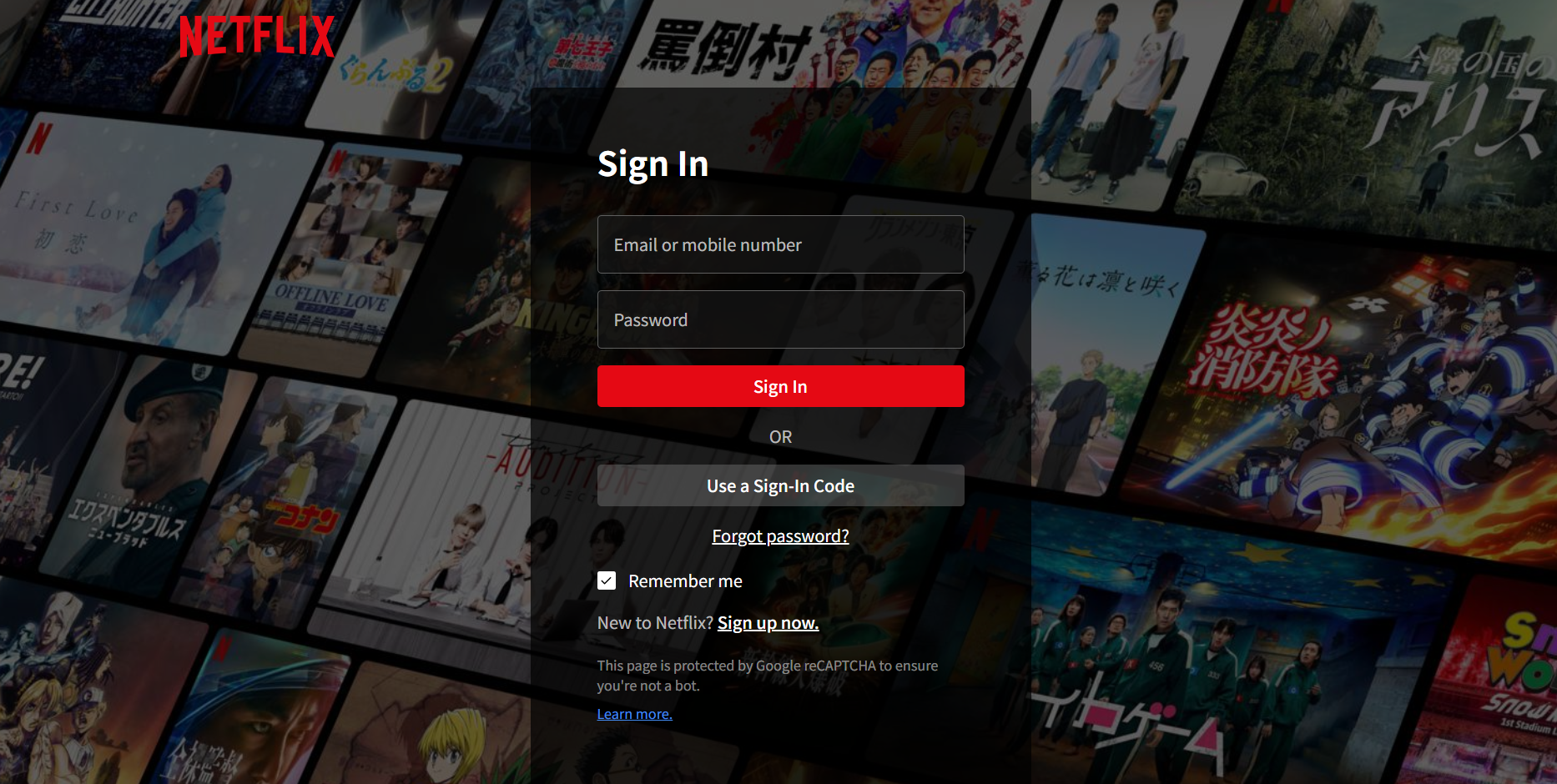
## Netflix TV Show 数据分析项目报告

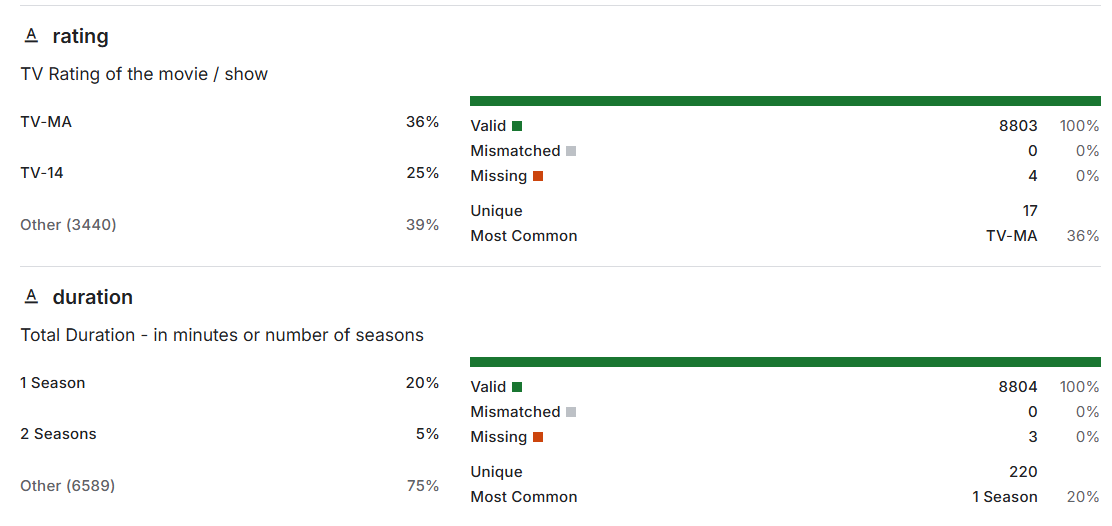
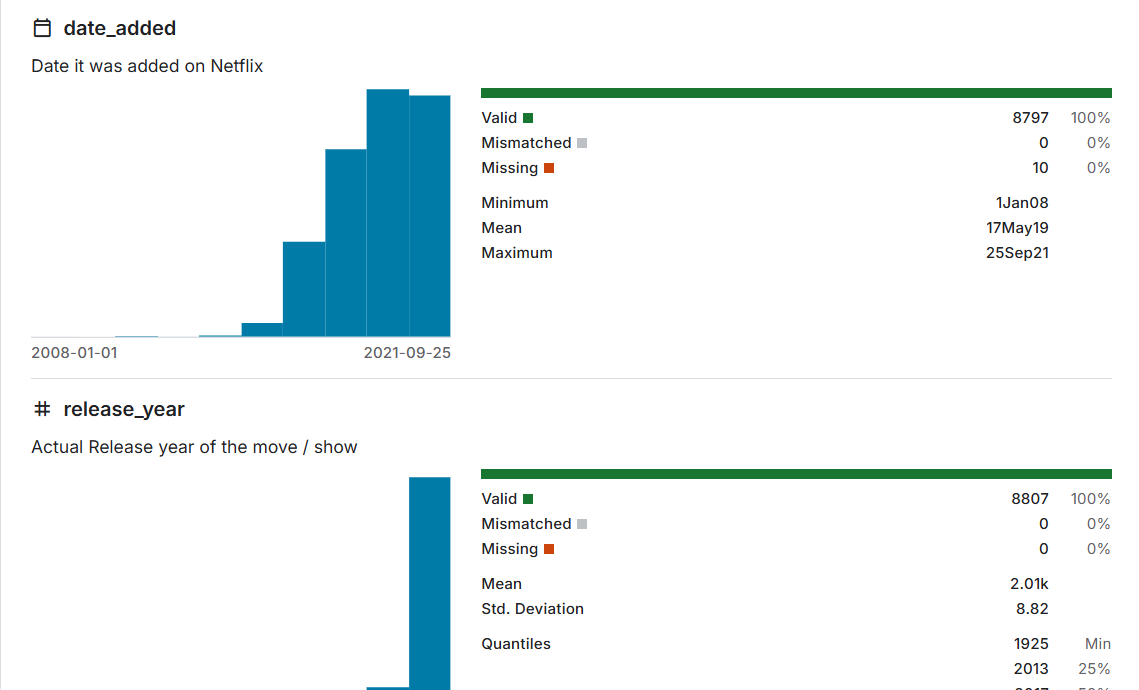
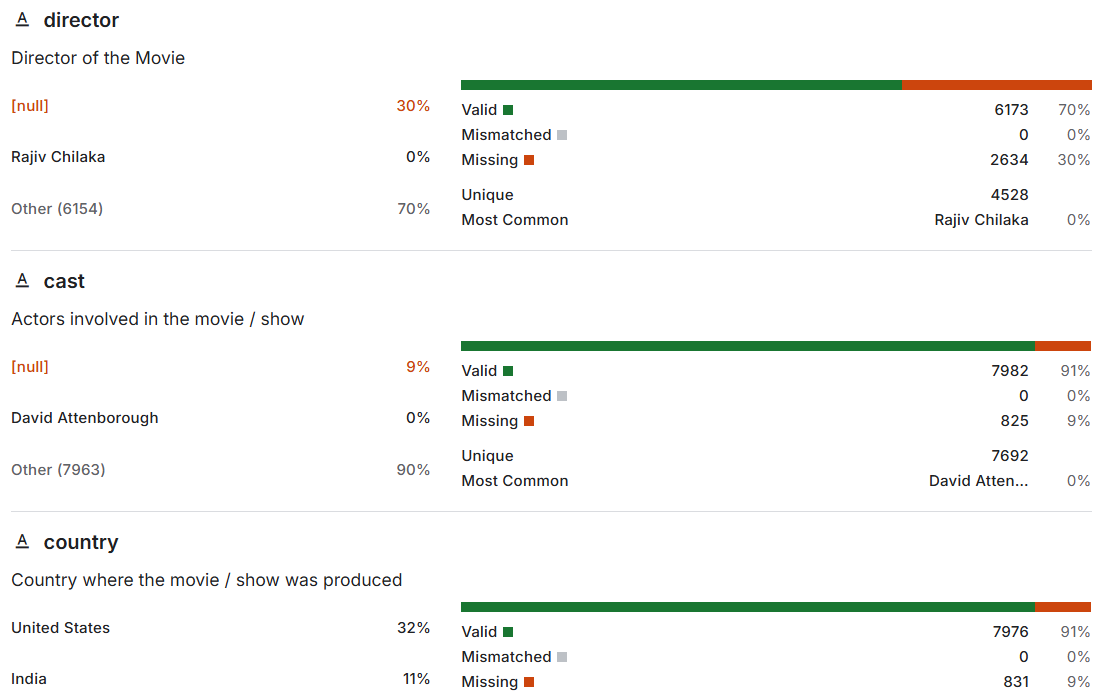


**项目目标：**

本项目的目标是对Netflix TV Show数据集进行探索性数据分析 (EDA)，通过数据清洗、预处理和可视化，深入了解Netflix电视节目的特征，包括内容评级、上架时间、节目时长、国家分布和类别等，从而挖掘潜在的趋势和洞察。

**数据介绍：**

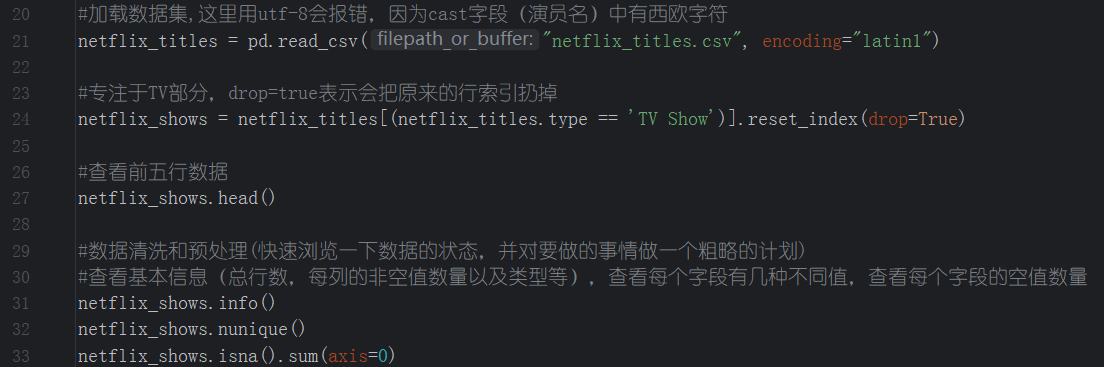
总数据共8807行，12列（截至2025年8月）。字段包括show\_id，每部电影和电视剧都有自己独特的id；type，说明该剧是电影还是电视剧；title，电影或电视剧的名字等。剩下几个字段如图所示



#### **一、**数据概览和预处理

数据预处理步骤：

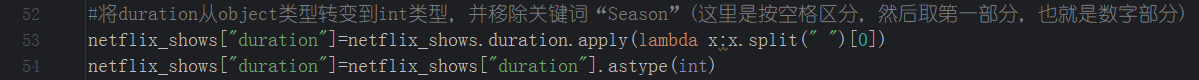
**1.1数据加载与筛选**：导入原始数据集netflix\_titles.csv，并筛选出type列为'TV Show'的数据，生成新的DataFrame netflix\_shows。



**1.2缺失值处理**：对于director,cast,country这种明确的特征用Unknown值填充，date\_added用虚拟日期1800-1-1填充；rating用已有NR（无评级）填充。

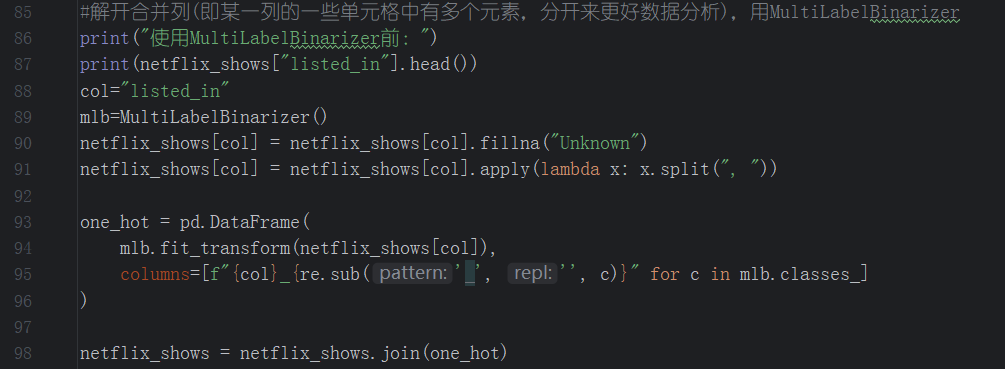


**1.3修正数据类型：**将duration从object类型转变到int类型，并移除关键词“Season”，表示季数；将date\_added分成年月日格式，先转成str类型，再分别提取年月日信息，最终整合为datetime类型。





**1.4解开合并列：**用MultiLabelBinarizer进行独热编码(One-Hot Encoding)，为每个类别创建独立的列（如listed\_in\_Dramas），值为0或1，以便于后续分析。在这里主要对listed\_in这一列进行拆分，这一列代表的含义是这部电视剧属于哪几种类别的电视剧。



**1.5数据清理：**Show\_id不包含任何相关信息，type只有一个唯一的值（“TV Shows”）。除此之外，对listed\_in进行拆分合并之后，原listed\_in列也没有了存在意义。因此删除这三列数据。





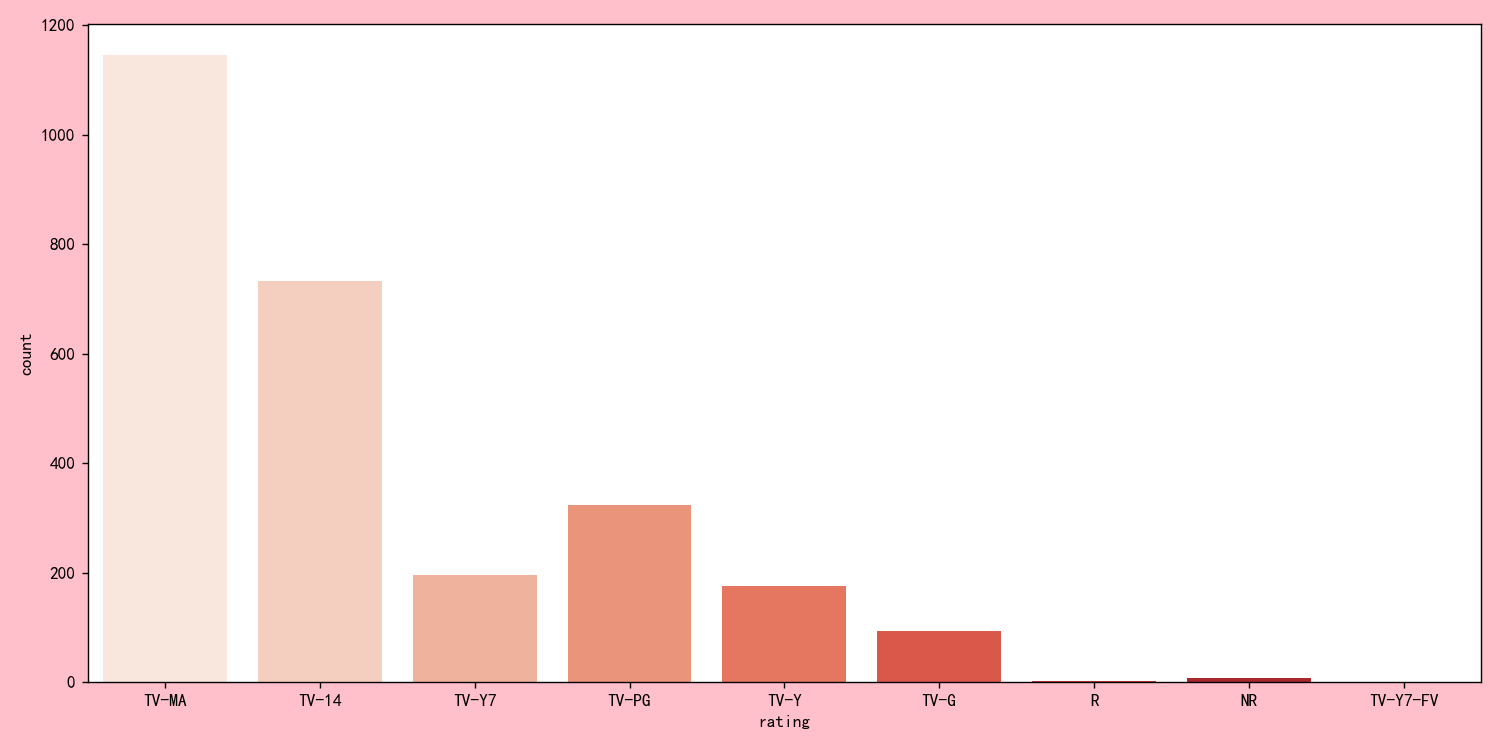
**1.6数据预处理：**为方便后续可视化分析，这里新建三个特征值，分别是1.星期几上架,结果是一个0-6的整数：0表示星期一,6表示星期日,可以用来分析Netflix喜欢在周几上架新剧；2.首季发行年份，例如release\_year为2021，duration为2，那这里结果就是2019；3.从首季发行到Netflix上架的间隔年数



以上就是数据清理和预处理的全部过程。

#### **二、**可视化分析与结论

**2.1内容评级（Rating）分析**



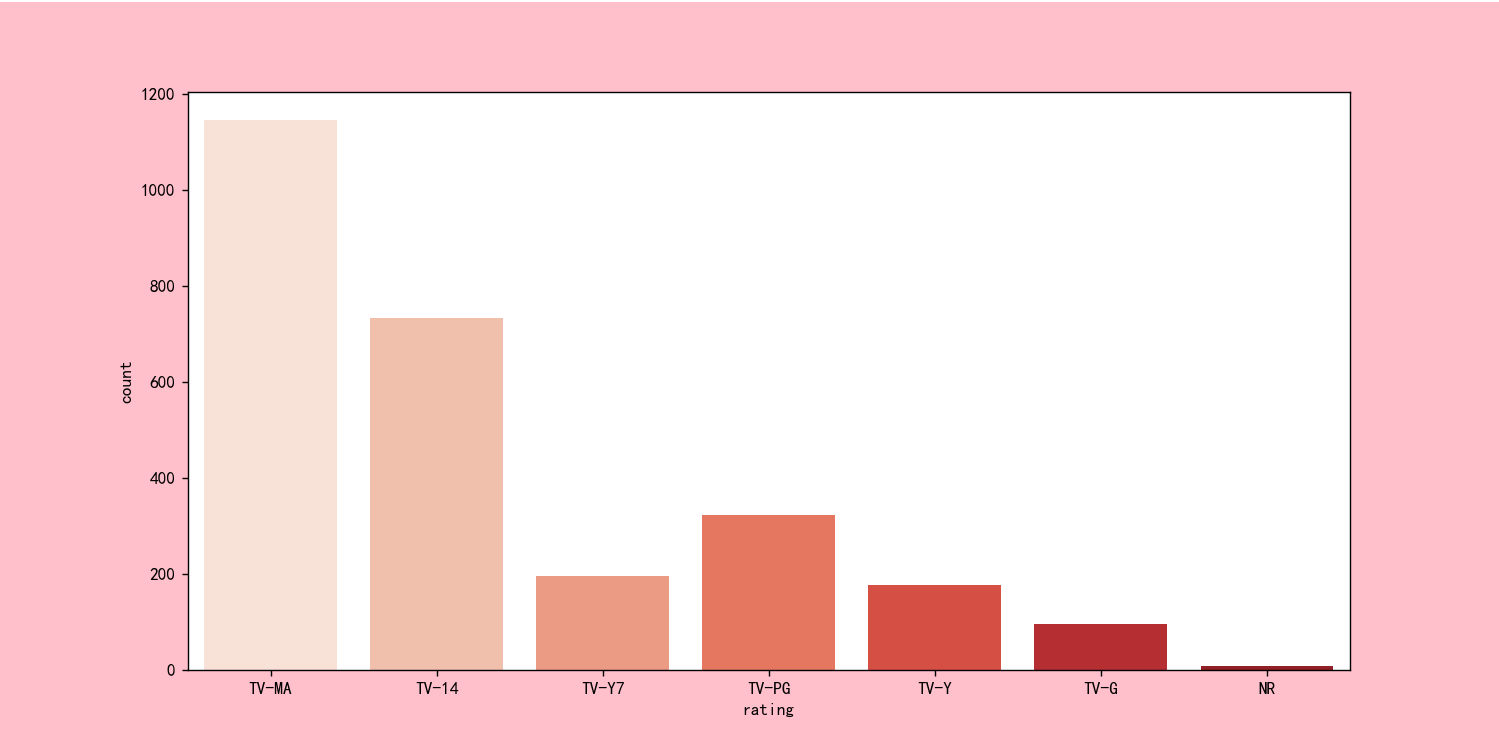
先简单介绍一下横坐标代表的含义：TV-Y: 2-6 years；TV-Y7: >= 7 year；TV-G: all ages；TV-PG: parental guidanceTV-14: >= 14 years；TV-MA: mature adult。

观察表数据可以发现，R和TV-Y7-FV是不应该存在的数据，NR则是最初数据清洗时用来填充空值表示无评级no rating。

在数据分析中，数据清洗是一个迭代的过程而不是一个一劳永逸的过程，在这里需要把R和TV-Y7-FV这两列进行处理，我选择的处理方法是把R改到最普遍的TV-MA列，把TV-Y7-FV改到TV-Y7列。

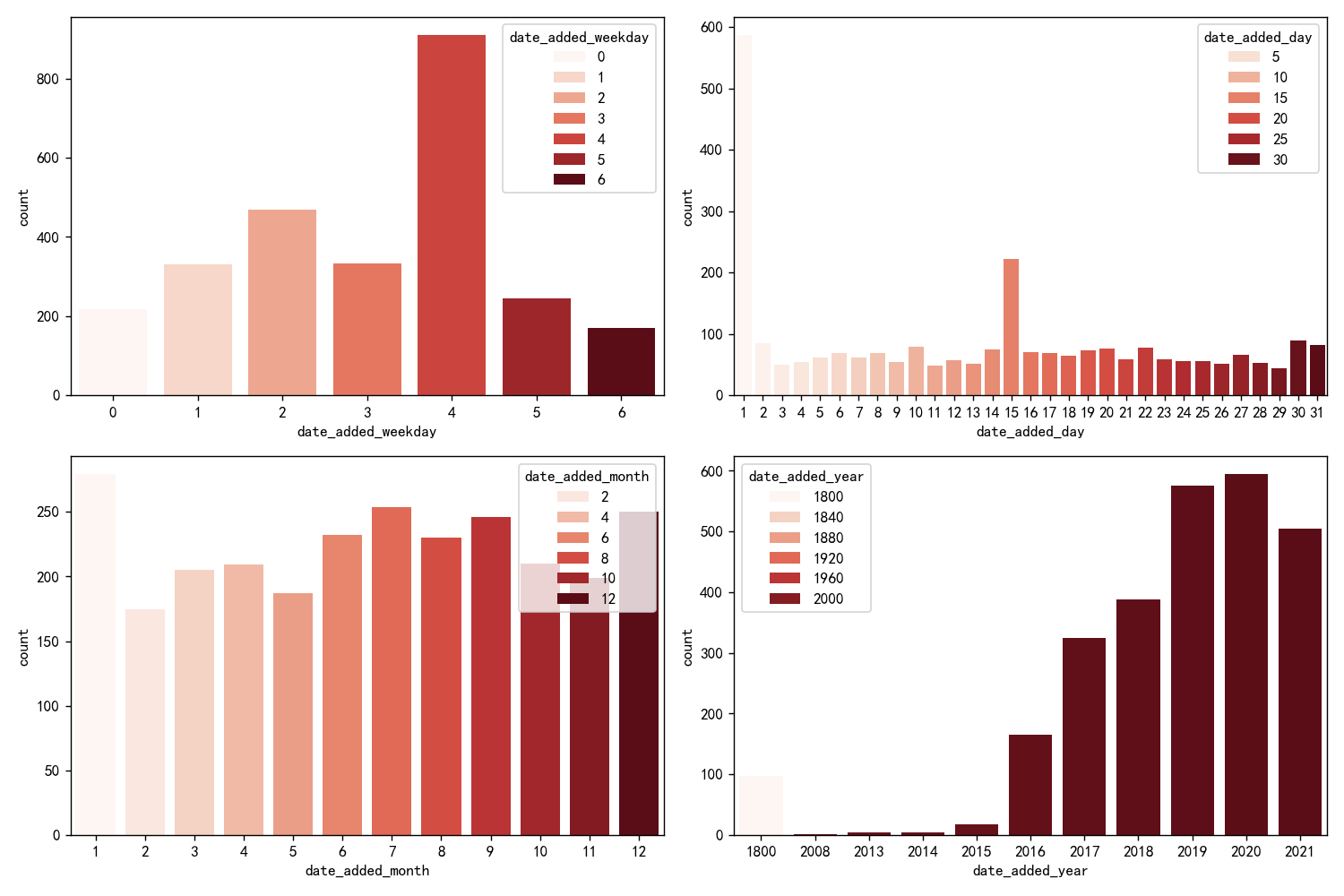


再运行一次结果如下



结论：1.TV-MA是最多的，其次是TV-14和TV-PG。2.最少的是TV-G，其次是TV-Y和TV-Y7，这些都是为年龄比较小的准备的电视节目。因此可以得出结论，Netflix上的电视节目更多是为了成年人而准备的。

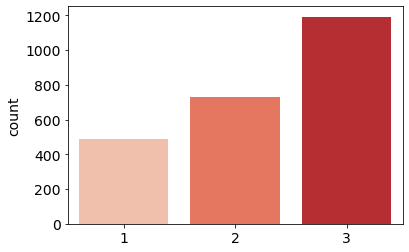
**2.2上架时间（date\_added）分析**



在数据清洗部分中，我把原先的年月日信息进行拆分并重组，并用feature engineering新建了一个字段date\_added\_weekday，其中0表示周一，6表示周日。

根据上图可以得出结论：1.周五加进来的电视节目是最多的。2.周一和周末加进来的电视节目较少。3.大部分电视节目都在每个月的1号或15号加入4.从加入月份来看，比较平均，但是下半年会更多一点。

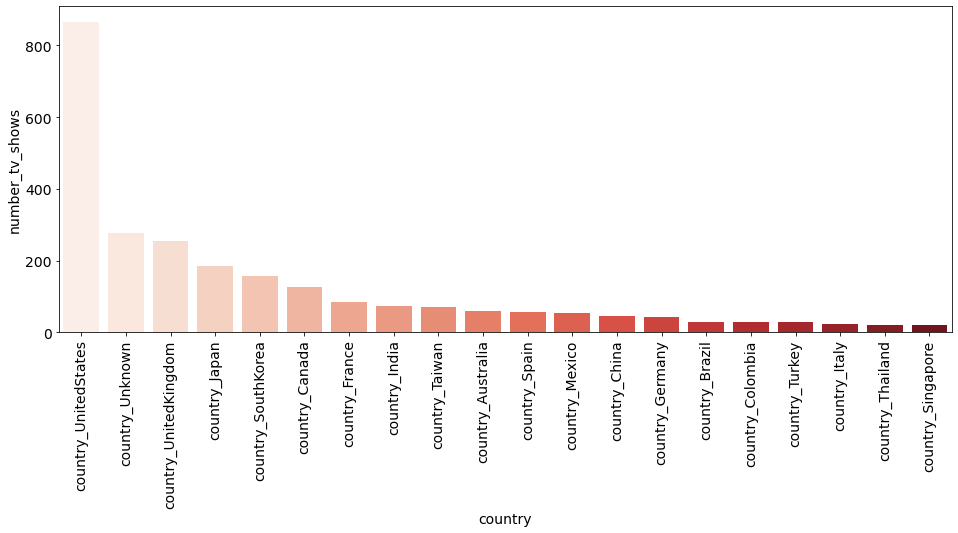
**2.3电视节目分类（listed\_in）分析**



这张图表示的是节目包含的类别数量分布图，也就是只属于一个种类的电视剧有400多种，属于两个种类的有700多种，属于三个种类有的将近1200种。

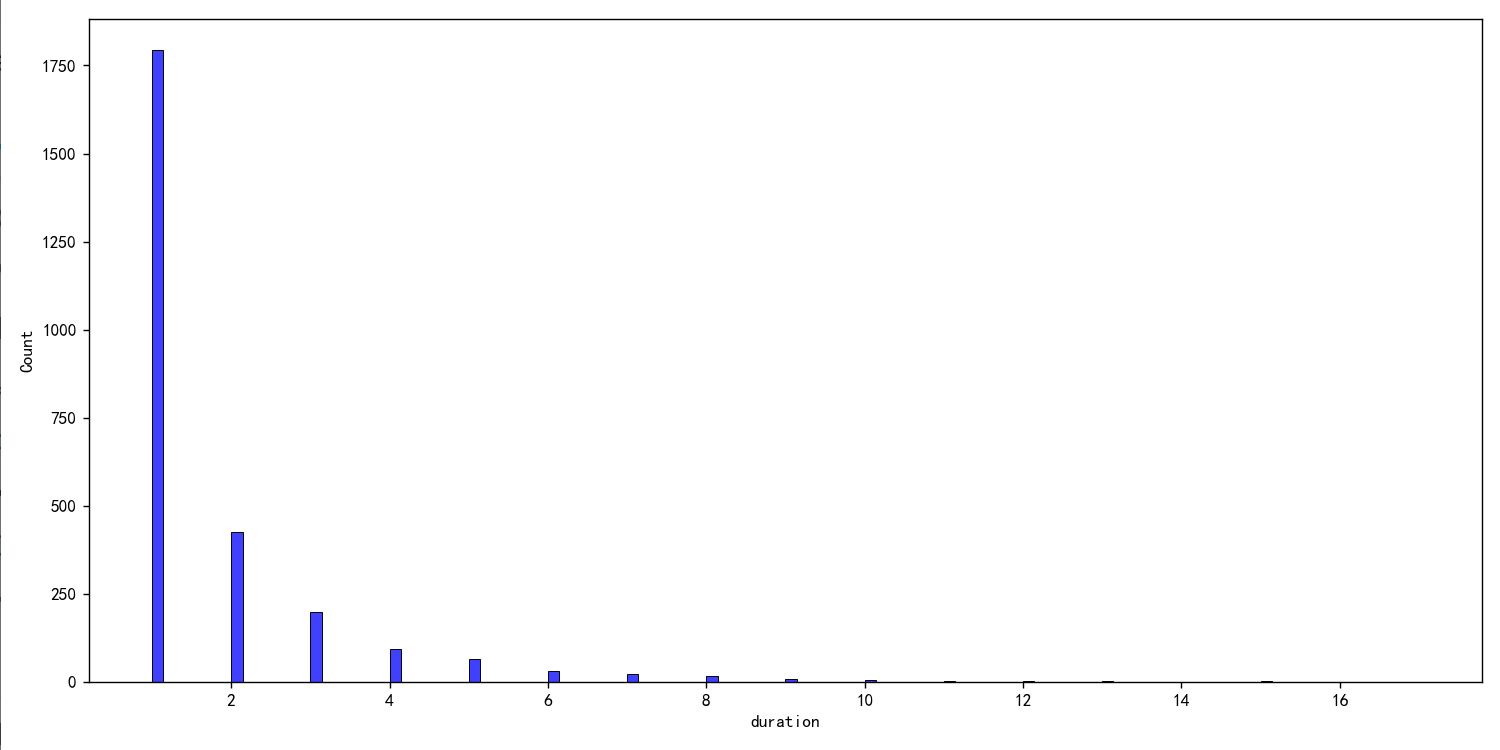
结论：一个电视节目至少被列入一个类别。Netflix上大约一半的电视节目被分为三类。

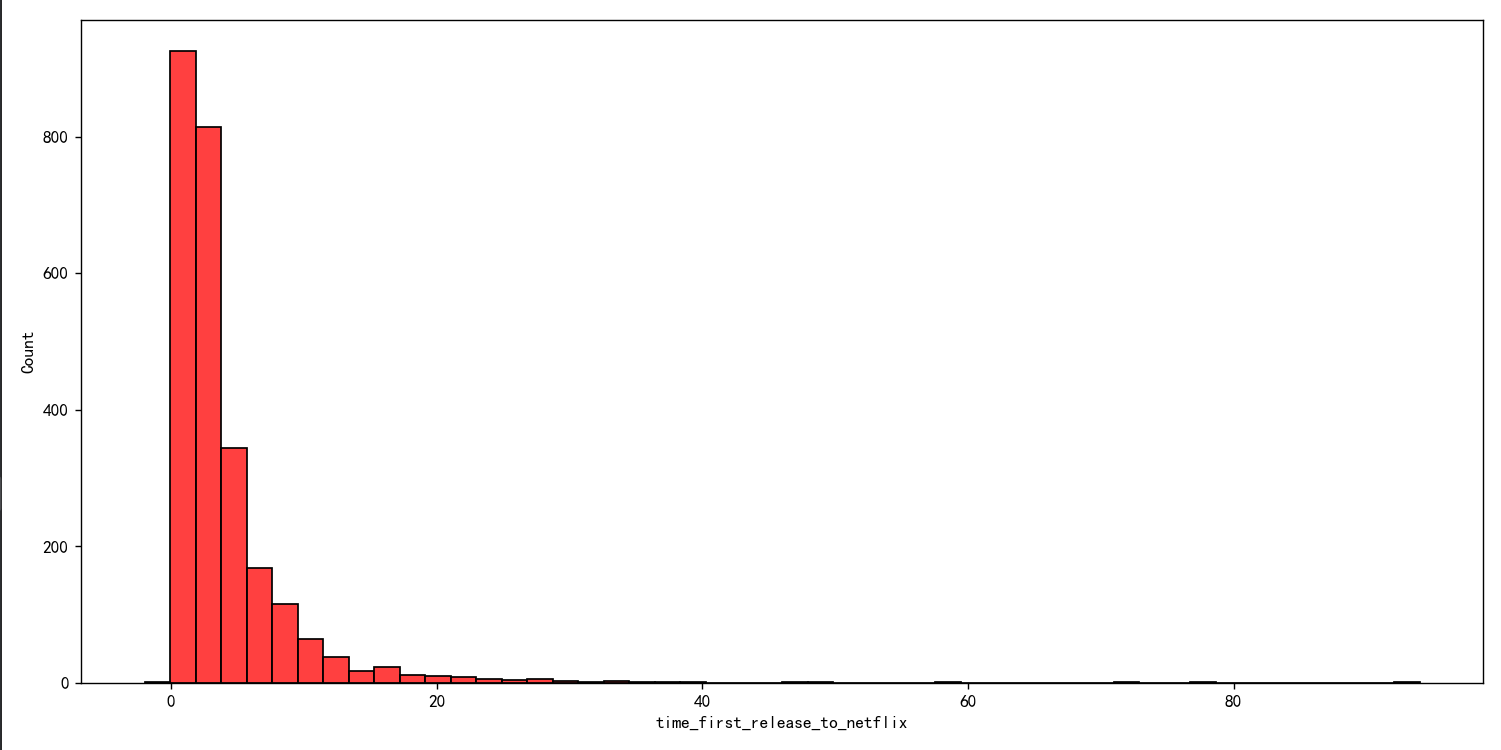
**2.4国家分布（country）分析**



结论：1.属于美国的电视剧最多，这符合Netflix是美国本土品牌的事实。英国和日本紧随其后2.unknown非常多，可以移除此列

**2.5节目季度数和首季上架间隔分析**

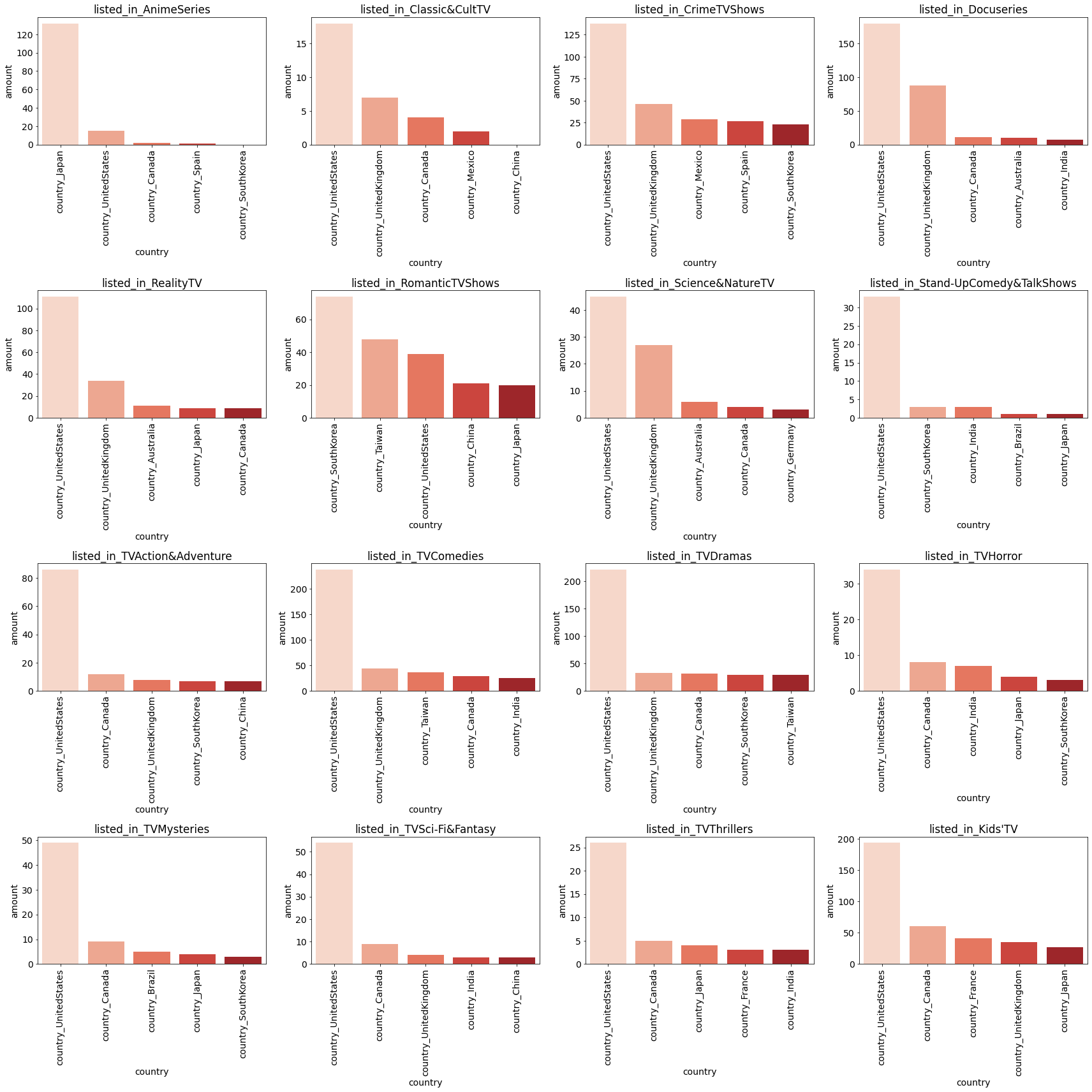




这里用到了seaborn的histplot作图（前面是countplot）。解释一下，duration字段的意义是该电视剧共有几季（数据清洗时的处理）；time\_first\_release\_to\_netflix字段表示的是该电视剧从首季发行到 Netflix 上架的间隔年数。

结论：1.大部分电视节目只有一季，随着季数的增加，节目数量显著减少。2.大部分电视节目在首季发行后很快就被添加到Netflix。

**2.6国家与分类的交叉分析**



结论：1.日本有最多的动漫节目。2.爱情片在亚洲地区（如韩国、中国等）更多

#### **三、**总结与展望

本项目通过一系列数据处理和可视化步骤，成功地从 Netflix TV Show 数据集中提取了有价值的洞察。我们了解了 Netflix 节目内容的主要受众（成年人）、上架时间的规律（周五和每月 1、15 号），以及不同国家在内容偏好上的差异。

未来的工作可以进一步深化分析：1.文本分析：对 description、title 等文本字段进行关键词提取和主题建模，以发现更多内容趋势。2.时间序列分析：分析不同评级或类别的节目在不同年份的上架趋势，预测未来的内容策略。3.国家和演员关联性：探索特定国家或演员与节目受欢迎程度之间的关系。