1. **데이터 이해**

* 데이터 선정 기준: 첫 주차이기 때문에 이해하기 쉽고 구분 행렬이 많지 않은 것으로, 흥미 유발 용으로 선정함.
* **Dataset**: Netflix Movies and TV Shows

<https://www.kaggle.com/shivamb/netflix-shows>

* **데이터 설명**

|  |
| --- |
| TV Shows and Movies listed on Netflix This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flixable which is a third-party Netflix search engine.  In 2018, they released an interesting report which shows that the number of TV shows on Netflix has nearly tripled since 2010. The streaming service’s number of movies has decreased by more than 2,000 titles since 2010, while its number of TV shows has nearly tripled. It will be interesting to explore what all other insights can be obtained from the same dataset.  Integrating this dataset with other external datasets such as IMDB ratings, rotten tomatoes can also provide many interesting findings. Inspiration Some of the interesting questions (tasks) which can be performed on this dataset -   1. Understanding what content is available in different countries 2. Identifying similar content by matching text-based features 3. Network analysis of Actors / Directors and find interesting insights 4. Is Netflix has increasingly focusing on TV rather than movies in recent years. |
| **넷플릭스에 등재된 TV 프로그램 및 영화**  이 데이터세트는 2019년 현재 넷플릭스(Netflix)에서 볼 수 있는 TV 쇼와 영화로 구성된다. 데이터 세트는 제3자 넷플릭스 검색 엔진인 플렉서블에서 수집된다.  2018년에는 넷플릭스의 TV 프로그램 수가 2010년 이후 3배 가까이 증가했다는 흥미로운 보고서를 발표했다. 스트리밍 서비스의 영화 수는 2010년 이후 2000편 이상 줄어든 반면 TV 쇼 수는 3배 가까이 증가했다. 동일한 데이터 집합에서 다른 모든 통찰력을 얻을 수 있는 것을 탐구하는 것은 흥미로울 것이다.  이 데이터 집합을 IMDB 등급과 같은 다른 외부 데이터 집합과 통합하면 썩은 토마토도 많은 흥미로운 발견을 제공할 수 있다.  **영감**  이 데이터 집합에서 수행할 수 있는 몇 가지 흥미로운 질문(태스크) 1. 다른 국가에서 사용할 수 있는 컨텐츠 이해 2. 텍스트 기반 기능을 일치시켜 유사한 내용 식별 3. 행위자/감독의 네트워크 분석 및 흥미로운 통찰력 찾기 4. 넷플릭스가 최근 몇 년 사이 영화보다 TV에 점점 더 집중하고 있는가. |

* **데이터 칼럼 분석**

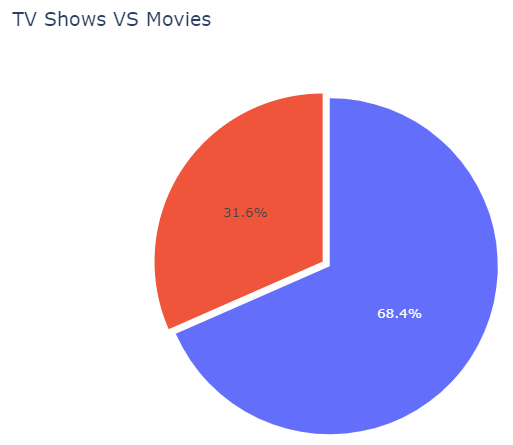
|  |  |
| --- | --- |
|  | **칼럼은 총 12개 (중요 10개)**   * show\_id: key값   (식별 키로 같은 값을 가질 수 없음)   * type: 영화/TV쇼 로 나눠짐 * title: 프로그램 이름 * director: 감독, 제작진 * cast: 배우 * country: 제작된 국가 * date\_added: 넷플릭스에 추가된 날짜 * release\_year: 출시 년도 * rating: 관람 등급 * duration: 프로그램의 길이 |

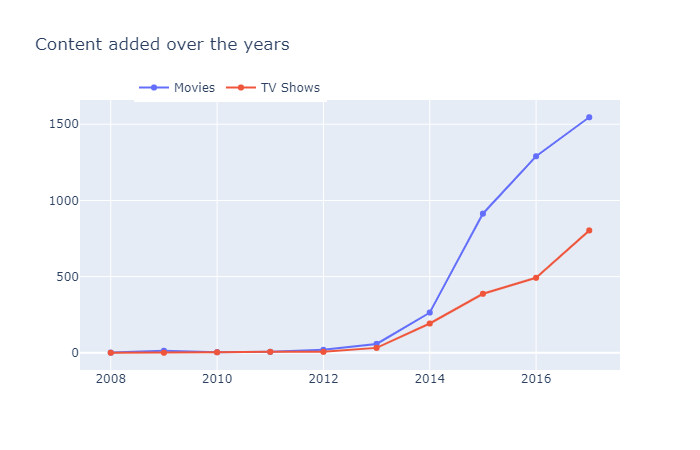
1. **데이터 활용 사례**

* **주제:** Netflix Consumption Analysis

<https://www.kaggle.com/sahib12/netflix-consumption-analysis>

**(기본 코드 외에 필요한 부분만 첨부하였습니다. 이 외에 부분은 각자 공부.)**

* **1. Content Type:**
  + In[5]: TV 쇼와 Movie의 비율을 측정 후 Pie 그래프로 표현
  + 비율은 [ 31.6% : 68.4% ] 로 보아 전체적으로 봤을 때 영화를 더 선호함을 알 수 있음.
  + 
* **2. Penetration Over The Years**



* + ln[6]: 데이터로 가져온 년도(2010~2019)와 그래프를 그리는 년도(2008~2019)가 다름. 그래프 볼 때 +2년씩 해야 맞음 유의!
  + Number of Movies that were added to Netflix platform in **Year 2018 and 2019** combined were more than sum of all movies in All the previous years from 2010 to 2017

넷플릭스 플랫폼에 2018년과 2019년을 합친 영화 수는 2010년부터 2017년까지 전년도 전체 영화 수를 합친 수보다 많았다.

* + Number of Movies on Netflix were almost 4 times Than number of TV Shows in 2018 year alone.

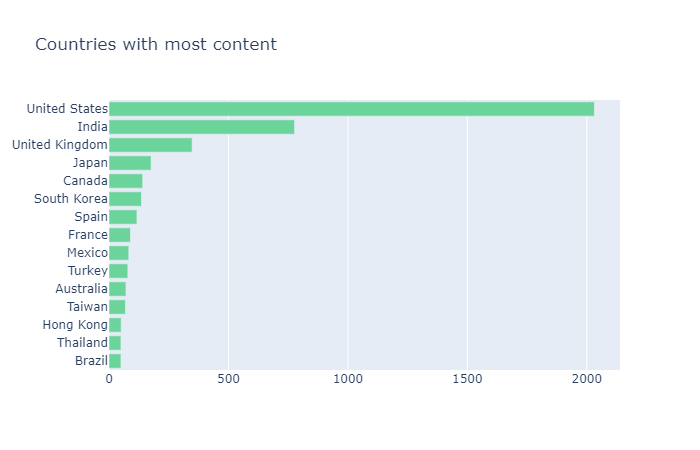
넷플릭스의 영화 수는 2018년 한 해 동안만 'TV쇼'의 4배에 가까웠다.

* + Session from 2016 to 2017 was the time when the number of Movies on Netflix increased about **4 times**.

2016년부터 2017년까지 세션은 넷플릭스의 영화 수가 약 4배 증가한 시기였다.

* **3. Countries With Most Content**
  + This data contains countries with Most content exclusively for that country

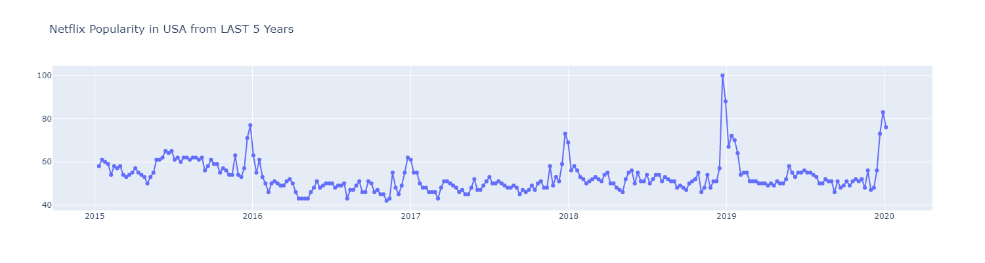
이 데이터에는 해당 국가만을 위한 콘텐츠가 가장 많은 국가가 포함됨

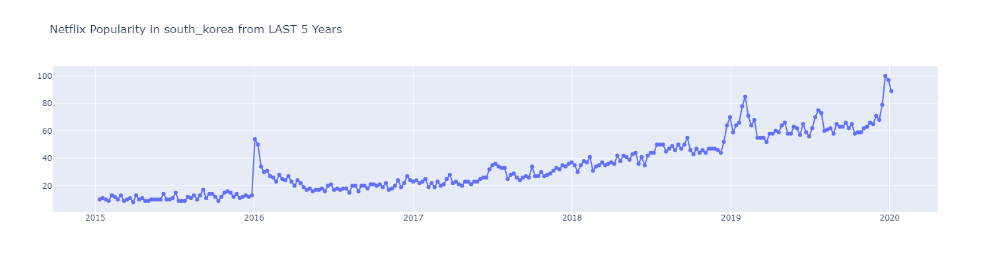


* + Most of the content is released exclusively for USA (even more than sum of top 15 countries) This might be because Netflix has been Very Popular in USA from a Long time whereas in other countries like India it's Popularity started to Increase from 2018

대부분의 콘텐츠는 미국 전용으로 출시된다(최상위 15개국 합계를 훨씬 웃도는) 이는 넷플릭스가 미국에서 오랜 시간 동안 매우 인기 있었던 반면 인도와 같은 다른 나라에서는 2018년부터 인기가 높아지기 시작했기 때문일 것이다.

* + **ln[8-9]: Netflix Popularity in Last 5 Years** in USA/Australia/Turkey/Korea/…

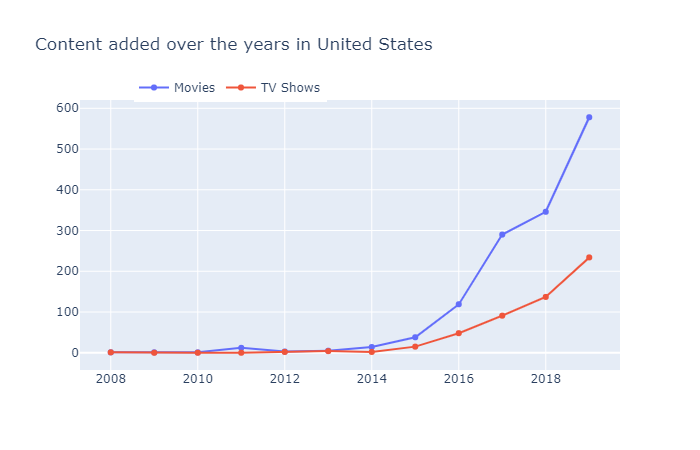
: USA

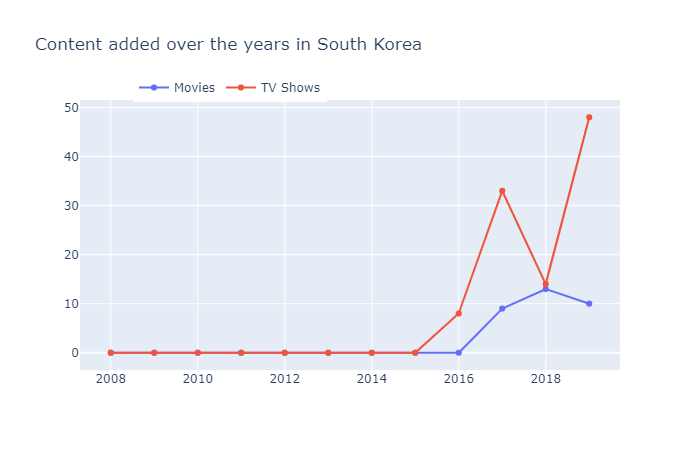
: Korea

* + <https://trends.google.com/trends/explore?date=today%205-y&geo=US&q=netflix>
  + from this website you can see: 이 웹사이트에서 볼 수 있다.
  + Netflix's popularity has never been less than 40% percent that can be the reason why Netflix release more content exclusively for USA

넷플릭스가 미국만을 위한 콘텐츠를 더 많이 출시하는 이유가 넷플릭스의 인기도 40% 이하인 적이 없다.

* + **ln[10]: Content added over the Years** in USA/Australia/Turkey/ Korea/…

: USA

: Korea

* + OBSERVATIONS: 관측치
  + In ASIAN continent Netflix increased it's content once they saw an increase in Interest of people towards Netflix examples are:- Turkey, Thailand, Taiwan, Japan, South Korea ,India

아시아 대륙의 넷플릭스는 넷플릭스에 대한 사람들의 관심이 증가하자 콘텐츠를 늘렸다:-터키,태국,대만,일본,한국,인도

* + In North America (Mexico,USA and CANADA) netflix was always popular hence Netflix's always Made more content for North America

북아메리카(멕시코)에서는미국과 캐나다) 넷플릭스는 항상 인기가 있었고, 따라서 넷플릭스는 북미에서 항상 더 많은 콘텐츠를 만들었다.

* + In European countries (France and Spain) Netflix's Popularity started to increase after 2015.

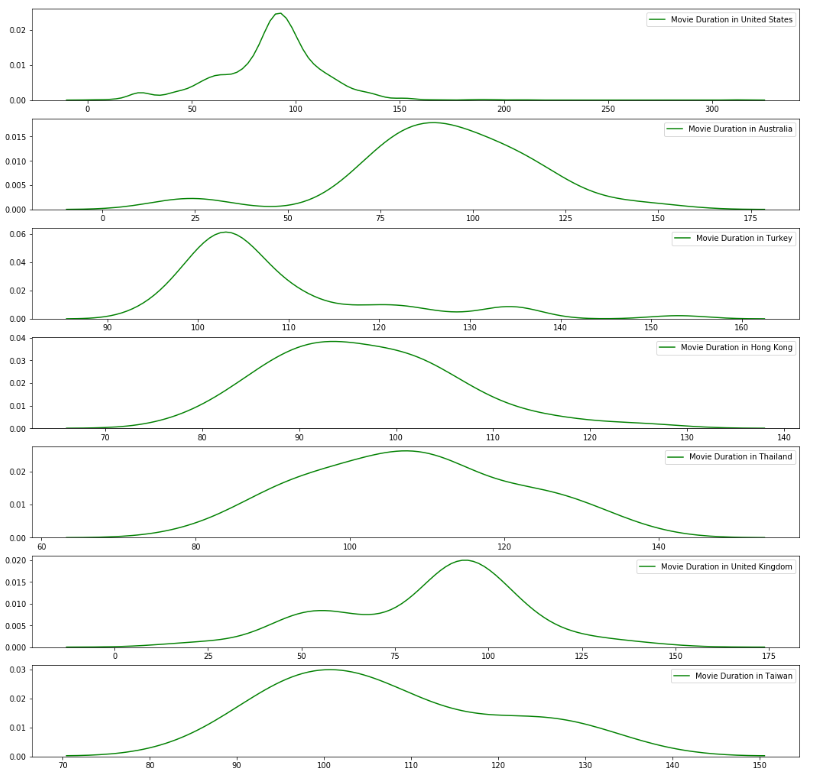
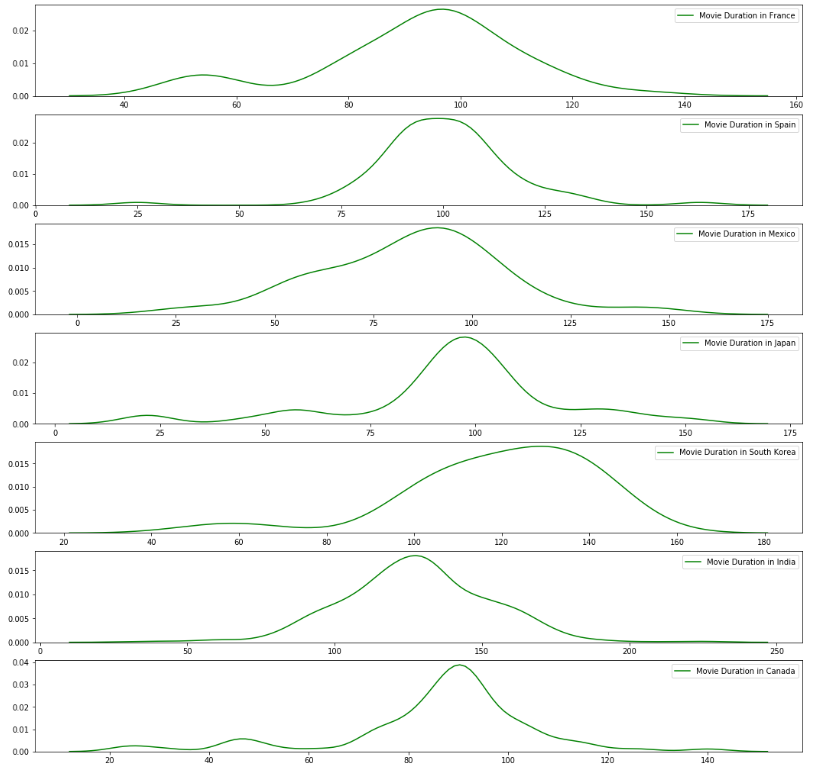
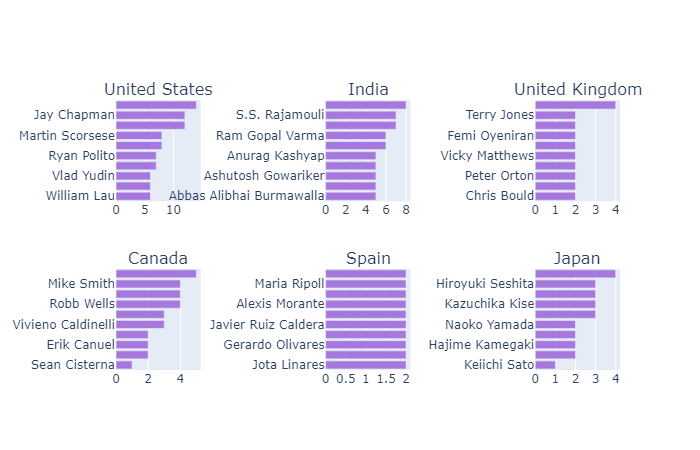
유럽 국가(프랑스·스페인)에서는 2015년 이후 넷플릭스의 인기도가 상승하기 시작했다.

* + In South East Asia, Japan and Australian continent Tv Shows went popular after 2018.

동남아시아에서는 2018년 이후 일본과 호주 대륙 TV쇼가 인기를 끌었다.

* + In France and Spain Tv shows are also getting popular.

프랑스와 스페인에서도 TV 쇼가 인기를 끌고 있다.

* **4. Preferable Movie Duration:**
  + In **Asian** Countries(**India,South Korea,Taiwan,Hong Kong,Thailand**) and **Australia** Movies with Duration greater than 110 minutes have good number of Viewers
  + 아시아 국가들(인도, 한국, 대만, 홍콩)에서태국)과 호주영화는 110분 이상 관람객이 많다.
  + In **North America**(**Canada,United States,Mexico**) and **European** countries(**France,Spain United Kingdom**) people like movies between 50 minutes and 110 minutes.
  + 북미(캐나다, 미국, 멕시코)와 유럽 국가(프랑스, 스페인 영국)에서는 50분에서 110분 사이의 영화를 좋아한다.
  +  
* **5. Top 10 Movie Directors:**
  + 

1. **나라면**

* 위 Notebook은 사람들의 소비 **분석**만 한 것이다.
* 이 데이터 분석에는 listed\_in(장르), description이 빠졌는데 이를 추가하면 나중에는 **예측/추천 시스템**까지도 만들어 볼 수 있을 것이다. (실제로도 있음)
  + 데이터에 시청 날짜 또는 시간이 없는 것이 아쉽다.
  + release\_year, date\_added 연관성을 분석 사람들의 분명 사람들의 프로그램 시청 선택 기준에는 최근의 것인지도 포함될 것이기 때문
  + description에서도 같은 장르에서의 상위권에 있는 프로그램에서 공통으로 쓰인 단어를 추출하면 사람들의 흥미, 관심도 파악이 가능