한국(아시아)의 NETFLIX 흥행 원인 분석

멋쟁이사자처럼 인공지능 통합과정 2nd 1차 프로젝트 4팀

CONTENTS

1. 분석 취지
2. 가설
3. 활용 데이터
4. 분석방법
5. 분석결과

분석취지


머니투데이

로그인

뉴스 증권 정치 법률 TOM 칼럼 자동차 연예 이슈 MT리포트 | VIP 속보 이상복 교수, 자본시장법 출간 금융소

501만명 넷플릭스에 725억원 긁었다...또 '역대 최고'

머니투데이 | 김수현 기자

2021.03.17 09:02

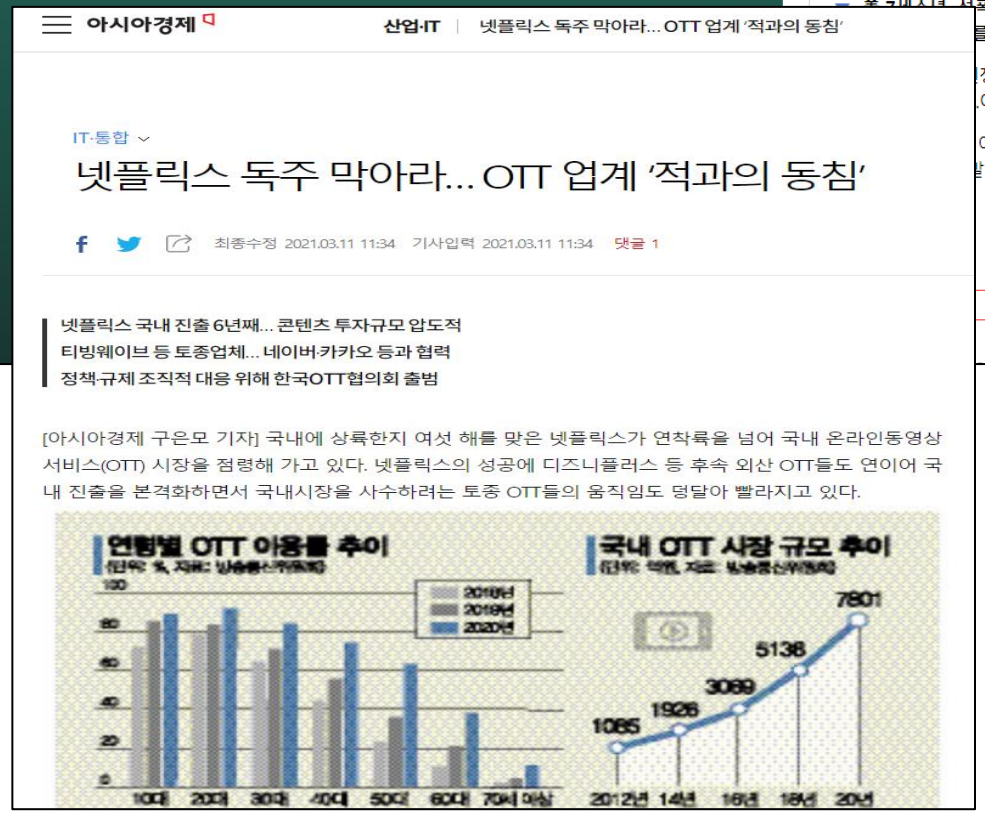
의견 남기기

2월 국내 넷플릭스 결제액 725억원 1년만에 3배
국내 유료 이용자 501만명 넷플릭스 '독주' 지속

실시간 급상승

'구미 친모' 알리바이 입
172 생아 몸무게...출산 증거

"박수홍 형이 재산 늘리
89 에...손헌수 "문제는 횡


아시아경제

산업·IT | 넷플릭스 독주 막아라... OTT 업계 '적과의 동침'

IT·통합

넷플릭스 독주 막아라... OTT 업계 '적과의 동침'

최종수정 2021.03.11 11:34 기사입력 2021.03.11 11:34 댓글 1

넷플릭스 국내 진출 6년째... 콘텐츠 투자규모 압도적
티빙웨이브 등 토종업체... 네이버·카카오 등과 협력
정책·규제 조직적 대응 위해 한국OTT협의회 출범

[아시아경제 구은모 기자] 국내에 상륙한지 여섯 해를 맞은 넷플릭스가 연착륙을 넘어 국내 온라인동영상 서비스(OTT) 시장을 점령해 가고 있다. 넷플릭스의 성공에 디즈니플러스 등 후속 외산 OTT들도 연이어 국내 진출을 본격화하면서 국내시장을 사수하려는 토종 OTT들의 움직임도 덩달아 빨라지고 있다.


ChosunBiz

증권 부동산 산업 IT 유통 금융 정책 국제 시사 오피니언

IT
ICT/미디어

넷플릭스, 3분기 성장세 아시아가 주도... 한국 가입자 330만명 돌파

조선비즈 | 황민규 기자

입력 2020.10.21 09:35 | 수정 2020.10.21 10:21

"한국·일본, 브로드밴드 쓰는 가정 두자릿수 점유율 달성"

한국과 일본이 세계 최대 온라인 동영상 스트리밍 업체 넷플릭스의 올해 3분기 성장에 일등공신이 됐다는 분석이다. 특히 3분기 전 세계 신규 가입자의 46%가 한국, 일본을 포함한 아태 지역인 것으로 나타났다.

20일(현지시각) 로이터통신에 따르면 넷플릭스는 이날 3분기 실적을 발표하면서 주주들에게 보낸 편지에서 한국과 일본을 포함한 아시아·태평양 지역 가입자가 처음으로 이 회사의 전체 유료 가입자 증가를 이끈 최대 공신이 됐다고 밝혔다. 아태 지역 가입자는 전 세계 신규 가입자의 46%를 차지했고, 아태 시장의 매출액은 작년 동기보다 66% 상승했다.

IT 주요뉴스

'IPO 大魚' 크래프톤 최대주주 노리는 텐센트..."상장 과실은 中 기업에"

美 마이크론·WD, 낸드 3위 日 키옥시아 인수 노려...4조 투자 SK하이닉스 돈 뺄까?

"임금 더 올려달라" 삼성전자 노조, 7.5% 인상안 거부, 10% 요구 추진...LG 노조도 9%

구글, 카카오모빌리티에 560억 투자..."韓 IT 발전 노력의 일환"

한국 OTT기업들을 제치고 국내 OTT시장에서

폭발적인 상승을 하고 있는 NETFLIX의 흥행원인을 분석

가설

가설1. 단순 컨텐츠의 증가로 인한 접근성 확대

Updated 2021.4.2 금 10:24

MediaUs
우리가 미디어 ;미디어스

전체기사 | 뉴스 | 오피니언 | 인터뷰 | 미디어비평 | 미디

HOME > 뉴스 > 뉴스

넷플릭스 아시아 성장 배경은 '한국 콘텐츠'

지난해 유료가입자 2억 명 돌파...아시아태평양 지역 가입자 2549만 명

윤수현 기자 | 승인 2021.01.20 14:29

[미디어스=윤수현 기자] 넷플릭스 전 세계 가입자가 지난해 2억 명을 넘어섰다. 넷플릭스 가입자는 지난해 4분기 850만 명 증가했으며 이 중 아시아·태평양 지역 가입자 증가는 200만 명에 달했다. 넷플릭스는 아시아 지역 성장의 배경으로 한국 콘텐츠를 꼽았다.

가설

가설2. 코로나의 영향으로 인한 영화관의 부진에 따른 이용자 증가

ChosunBiz

증권 부동산 산업 IT 유통 금융 정책 국제 시사 오피니언

오피니언
기자수첩

[기자수첩] 두 번의 영화 관람료 인상과 넷플릭스의 잭팟

조선비즈 | 이선목 기자

입력 2021.03.24 06:30

"이제 2명이 극장에서 영화를 보려면 2만8000원, 거기에 팝콘, 음료까지 하면 거의 5만원 돈이 드네요."

지난 주말 영화 '미나리'를 관람했다는 직장인 박채은(30)씨는 불만을 터뜨렸다. 영화관 1위 CJ CGV가 4월 2일부터 영화 관람료를 1000원 올리기로 하면서다. 연인이 주말에 일반(2D) 영화를 보려면 둘이서 2만8000원을 내야한다. 3D 영화는 3만원이다.

CGV의 관람료 인상은 지난해 10월에 이어 6개월만이다. 롯데시네마, 메가박스도 동참할 가능성이 높다. 두 업체는 아직 관람료 인상과 관련해 '확정된 사항이 없다'는 입장이지만, 지난해에도 CGV가 10월 중순 관람료를 인상한 지 한 달여 만에 메가박스와 롯데시네마가 잇달아 관람료를 올렸다.

오피니언 주요뉴스

[기자수첩] 오락가락 재건축 안전진단, 정말 문제 없나요

가설

가설3. 1인 가구 증가에 따른 이용자수 증가

TV보는 2030 1인가구 줄어든다

입력 2018.04.17 13:40

0 0

KISDI 유료방송 서비스 가입추세 분석
케이블 · IPTV 해지율 젊은 1인가구 증가

게티이미지뱅크

젊은 층과 1인 가구를 중심으로 케이블TV, 인터넷(IP)TV 등 유료방송을 해지하는 사람들이 늘고 있다. 유료방송에 가입하지 않더라도 모바일 기기 등을 통해 쉽게 이용할 수 있는 온라인 동영상 서비스(오버더톱 · OTT) 등 대체 서비스가 다양해 굳이 TV를 시청하지 않는 것으로 보인다.

17일 정보통신정책연구원(KISDI)이 발표한 '유료방송 서비스 가입 추세 분석' 보고서에

많이 본 기사

1 막말에 부동산 '내로남불' 까지…겹악재 꼬이는 민주당
2 5.7일 만에 1600만 원으로 뚝딱…국제결혼쇼핑
3 배 한척이 불러온 물류대란…'제2 수에즈 운하' 개발과 …
4 극우와 갈라섰냐… 태극기 사라진 국민의힘 유세장
5 오세훈 '용산 참사' 발언… 망언과 오해 사이

경제

머스크 '만우절 트윗' 한 줄에 도지코인 가격 급등
'영혼 없는 서명'이 '영혼 없는 설명'으로…금소법 혼란
위원회 출범하고 MOU도…식품업계 'ESG 경영' 속도

활용 데이터

1. 넷플릭스 프로그램 정보

데이터 기본정보 ; 넷플릭스 프로그램 별 출시일자, 출시국가 및 제목, 출연진, 감독 등 기본정보를 포함한 자료

```
filename = '/content/drive/MyDrive/temp/netflix_titles.csv'
data = pd.read_csv(filename)
data.info
```

```
<bound method DataFrame.info of      show_id  ...                                        description
0         s1  ...  In a future where the elite inhabit an island ...
1         s2  ...  After a devastating earthquake hits Mexico Cit...
2         s3  ...  When an army recruit is found dead, his fellow...
3         s4  ...  In a postapocalyptic world, rag-doll robots hi...
4         s5  ...  A brilliant group of students become card-coun...
...      ...  ...                                                ...
7782   s7783  ...  When Lebanon's Civil War deprives Zozo of his ...
7783   s7784  ...  A scrappy but poor boy worms his way into a ty...
7784   s7785  ...  In this documentary, South African rapper Nast...
7785   s7786  ...  Dessert wizard Adriano Zumbo looks for the nex...
7786   s7787  ...  This documentary delves into the mystique behi...

[7787 rows x 12 columns]>
```

활용 데이터

2. 지역별 넷플릭스 구독자 및 수익 현황

데이터 기본정보 ; 2018년부터 2020년까지 4분기로 나누어 지역별로 구독자 및 수익규모를 포함한 자료

```
filename = '/content/drive/MyDrive/temp/NetflixSubscribersbyCountryfrom2018toQ2_2020.csv'
data = pd.read_csv(filename)
data.info
```

```
<bound method DataFrame.info of                                Area  Q1 - 2018  ...  Q1 - 2020  Q2 - 2020
0         United States and Canada   60909000  ...   69969000   72904000
1  Europe, Middle East, and Africa   29339000  ...   58734000   61483000
2                    Latin America   21260000  ...   34318000   36068000
3                     Asia-Pacific    7394000  ...   19835000   22492000

[4 rows x 11 columns]>
```

```
filename = '/content/drive/MyDrive/temp/NetflixsRevenue2018toQ2_2020.csv'
data = pd.read_csv(filename)
data.info
```

```
<bound method DataFrame.info of                                Area   Q1 - 2018  ...   Q1 - 2020   Q2 - 2020
0         United States and Canada  1976157000  ...  2702776000  2839670000
1  Europe, Middle East, and Africa   886649000  ...  1723474000  1892537000
2                    Latin America   540182000  ...   793453000   785368000
3                     Asia-Pacific   199117000  ...   483660000   569140000

[4 rows x 11 columns]>
```

활용 데이터

3. 코로나바이러스감염증-19(COVID-19) 현황 누적 데이터

데이터 기본정보 ; 코로나19감염증으로 인한 일별 확진자, 완치자, 치료중인 환자, 사망자 등에 대한 현황자료

```
filename = '/content/drive/MyDrive/temp/Covid19InfState.csv'
data = pd.read_csv(filename)
data.info
```

```
<bound method DataFrame.info of      seq   stateDt  ...                createDt                updateDt
0      1  20200101  ...   2020-01-31 17:47:33.33  2020-02-03 12:21:56.56
1      2  20200202  ...   2020-02-03 12:22:49.49                     NaN
2      3  20200203  ...   2020-02-03 14:41:17.17  2020-02-04 14:19:46.46
3      4  20200204  ...   2020-02-03 21:26:59.59                     NaN
4      5  20200205  ...   2020-02-04 23:56:31.31   2020-02-05 9:43:16.16
..   ...       ...  ...                      ...                     ...
452  459  20210325  ...   2021-03-25 09:34:48.91                     NaN
453  460  20210326  ...    2021-03-26 09:33:28.3                     NaN
454  461  20210327  ...  2021-03-27 09:53:35.032                     NaN
455  462  20210328  ...  2021-03-28 11:17:23.373                     NaN
456  463  20210329  ...  2021-03-29 09:41:57.983                     NaN

[457 rows x 14 columns]>
```

활용 데이터

4. 영화진흥위원회 개봉 일람

데이터 기본정보 ; 영화 개봉일, 전국 관객수, 전국 매출액 등 국내 상영영화에 관한 자료

```
movie.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 6070 entries, 1 to 6070
Data columns (total 17 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   영화명      6070 non-null   object
 1   감독       5429 non-null   object
 2   제작사      2282 non-null   object
 3   수입사      3481 non-null   object
 4   배급사      6054 non-null   object
 5   개봉일      6070 non-null   datetime64[ns]
 6   영화유형     6070 non-null   object
 7   영화형태     6070 non-null   object
 8   국적       6070 non-null   object
 9   전국
스크린수  6070 non-null   int64
 10  전국
매출액   6070 non-null   int64
 11  전국
관객수   6070 non-null   int64
 12  서울
매출액   6070 non-null   int64
 13  서울
관객수   6070 non-null   int64
 14  장르       6018 non-null   object
 15  등급       6070 non-null   object
 16  영화구분     6070 non-null   object
dtypes: datetime64[ns](1), int64(5), object(11)
memory usage: 853.6+ KB
```

- .info()로 기본정보를 첨부한 이유
- 통일성을 위해 .info로 첨부하려 했으나 하위항목에는 흥행성적이 저조한 작품들이 노출되는데, 이는 성인영화가 대부분이라 이와 같은 제목을 포함한 자료를 첨부하는 것 보다 info()로 첨부하는 것이 맞다고 판단

활용 데이터

5. 연령별_1인 가구 분석

데이터 기본정보 ; 행정구역별 2017년~2019년 까지의 1인가구를 연령 및 성별로 분류한 자료

```
filename = '/content/drive/MyDrive/temp/연령별_1인 가구 분석.xlsx'
data = pd.read_excel(filename)
data.info
```

```
<bound method DataFrame.info of      행정구역별(시군구)     연령별     2017   2017.1  ...   2018.2     2019   2019.1   2019.2
0    행정구역별(시군구)     연령별     1인가구       남자  ...       여자     1인가구       남자       여자
1              전국      합계  5618677  2791849  ...  2942274  6147516  3053733  3093783
2             NaN  20세 미만    61058    29108  ...    30531    59415    28591    30824
3             NaN   20~24   393503   183578  ...   219331   431750   194984   236766
4             NaN   25~29   568288   333886  ...   257832   685831   391896   293935
..            ...     ...      ...      ...  ...      ...      ...      ...      ...
332           NaN   65~69     3996     1656  ...     2439     4584     1995     2589
333           NaN   70~74     3289     1087  ...     2357     3757     1298     2459
334           NaN   75~79     3351      828  ...     2518     3373      812     2561
335           NaN   80~84     2605      472  ...     2141     2684      540     2144
336           NaN  85세 이상     2035      213  ...     1990     2434      330     2104

[337 rows x 11 columns]>
```

분석 방법

가설 1번 단순 컨텐츠의 증가로 인한 접근성 확대의 검증

사용자료 : 넷플릭스 프로그램 정보, 지역별 넷플릭스 구독자 및 수익 현황
사용 상관계수 : 피어슨 상관계수

넷플릭스 프로그램 정보
1. pandas를 통해 datetime으로 변경
2. country 목록에 포함된 띄어쓰기와 쉼표를 정리
3. 컨텐츠 출시일자를 Quater로 저장
4. country중 asia에 포함된 지역의 정보를 저장 후 컨텐츠 증가량 확인
5. 컨텐츠 증가량을 바탕으로 시각화. 단 자료의 비교를 위해 범위를 2018년 1분기부터 2020년 2분기까지로 한정

지역별 넷플릭스 구독자 및 수익 현황
1. asia지역의 수익자료를 추출하여 Quater 별로 나눠 시각화

=>컨텐츠 증가량과 asia 지역의 수익자료 병합

분석 방법

```
import pandas as pd

df = pd.read_csv(drive_path + 'netflix_titles.csv')

# 자료형 datetime으로 변경
df['date_added'] = pd.to_datetime(df['date_added'])

len(df[(df['date_added'].notnull())]) # 10개 제거
len(df[(df['country'].notnull())]) # 507개 제거

df = df[(df['date_added'].notnull()) & (df['country'].notnull())] # 7787 -> 7271

# unique한 country 목록
countries = set()

def parse_country(ctrs):
  ctrs = ctrs.split(',')
  for country in ctrs:
    country = country.strip()
    if country == '': continue
    countries.add(country)

# country 목록 확인용
df['country'].apply(parse_country)
# 각 row 별 country count 추가
df['country_cnt'] = df['country'].apply(lambda ctrs: len(ctrs.split(',')))

print(len(countries)) # 118

# 한국을 포함하는 netflix title 목록.
kr_dest = df[df['country'].str.contains('South Korea')].loc[:, ['title', 'country', 'country_cnt', 'date_added']].sort_values(by=['date_added','country_cnt'])

def date_to_quarter(date):
  return str(date.year) + ' - Q' + str(date.quarter)

# 날짜 분기 추가
kr_dest['date_quarter'] = kr_dest['date_added'].apply(date_to_quarter)
kr_dest.set_index('date_quarter', inplace=True)
kr_contents = kr_dest.groupby('date_quarter')['title'].count()
kr_contents.plot()

# 아시아 country 목록
asia_df = pd.read_csv(drive_path + 'Asia.csv')
asia_df.columns = ['a', 'b', 'c', 'd']
asia = asia_df['c'].unique()

def in_asia(ctrs):
  for a in asia:
    if a in ctrs:
      return True
  return False
```

분석 방법

```
# 아시아 country를 포함하는 netflix title 목록
asia_dest = df[df['country'].apply(in_asia)].loc[:, ['title', 'country', 'country_cnt', 'date_added']].sort_values(by=['date_added','country_cnt'])
asia_dest['date_quarter'] = asia_dest['date_added'].apply(date_to_quarter)
asia_dest.set_index('date_quarter')
asia_contents = asia_dest.groupby('date_quarter')['title'].count()
asia_contents.plot()

# 아시아 country의 수익자료를 추출하여 Quater 별로 나눠 시각화
revenue = pd.read_csv(drive_path + 'NetflixsRevenue2018toQ2_2020.csv', index_col='Area')
revenue = revenue.T
asia_rev = revenue['Asia-Pacific']

# Q1 - 2018 -> 2018 - Q1
def to_quarter(date):
  q = date.split(' - ')[0]
  y = date.split(' - ')[1]
  return y + ' - ' + q
asia_rev.index = pd.Series(asia_rev.index).apply(to_quarter)
asia_rev2 = asia_rev[range(0, len(asia_rev) - 1)]
asia_rev2.index = asia_rev.index[range(1, len(asia_rev))]
asia_diff = asia_rev - asia_rev2

asia_diff['2018 - Q1'] = 20135000
asia_diff.plot()

con_x, con_y = asia_contents.index, asia_contents.values # 2016-2021 total_contents
rev_x, rev_y = asia_diff.index, asia_diff.values

asia_contents = asia_contents[rev_x]
con_x, con_y = asia_contents.index, asia_contents.values # 2018-2020 contents

fig, ax1 = plt.subplots()
ax2 = ax1.twinx()

line_contents = ax1.plot(con_x, con_y, label="Contents Increase")
line_revenue = ax2.plot(rev_x, rev_y, color='r', label="Revenue Increase")

ax1.set_ylim(0, 220)
ax2.set_ylim(0, 90000000)

ax1.set_xlabel('Quarter')
ax1.set_ylabel('contents')
ax2.set_ylabel('revenue')

lines = line_contents + line_revenue
labels = [l.get_label() for l in lines]
plt.legend(lines, labels, loc=2)
plt.title('넷플릭스 매출 증가량과 contents 증가량 관계 (아시아)')

fig.autofmt_xdate(rotation=45)
plt.rcParams["figure.figsize"] = (15, 5)
plt.show()
```

분석 방법

가설 2번 코로나의 영향으로 인한 영화관의 부진에 따른 이용자 증가를 검증

사용자료 : 코로나바이러스감염증-19(COVID-19) 현황 누적 데이터, 영화진흥위원회 개봉 일람, 지역별 넷플릭스 구독자 및 수익 현황
사용 상관계수 : 피어슨 상관계수
이 가설은 크게
1) 코로나 확진자 수와 영화관 관객수간의 상관관계를 분석 후
2) 영화관정보와 넷플릭스 구독자수와의 상관관계 분석함

1) 코로나 확진자 수와 영화관 관객수간의 상관관계를 분석

코로나 자료
1. 2020년 자료를 추출하고 중복값 제외, 결측치 채우기
2. 확진자 증가량을 위해
당월 누적확진자 - 전월 누적확진자를 하고 시각화

영화관자료
1. 2020년 자료를 추출하고 그룹으로 월별 관객수 집계
2. 시각화

=> 코로나 자료와 영화관 자료 병합

2) 영화관자료와 넷플릭스 구독자수와의 상관관계 분석

영화관 자료
1. 데이터를 분기로 나누고 분기별 개봉영화 갯수 , 영화수익, 영화 관객 추출
2. 데이터 비교를 위해 기간을 넷플릭스 구독자자료 내 기간으로 한정하고 시각화

넷플릭스 구독자 자료
1. 분기별 평균을 구함
2. 시각화
=>영화관 자료와 넷플릭스 구독자 자료 병합

분석 방법

코로나 확진자 수와 영화관 관객수 간의 상관관계 분석

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from pandas import DataFrame, Series
from numpy.random import randn
import seaborn as sns
import datetime
%matplotlib inline
%config InlineBackend.figure_format = 'retina'

!apt -qq -y install fonts-nanum

import matplotlib.pyplot as plt
import matplotlib.font_manager as fm

fontpath = '/usr/share/fonts/truetype/nanum/NanumBarunGothic.ttf'
font = fm.FontProperties(fname=fontpath, size=10)
fm._rebuild()

# 그래프에 retina display 적용
%config InlineBackend.figure_format = 'retina'

# Colab 의 한글 폰트 설정
plt.rc('font', family='NanumBarunGothic')

#코로나 데이터셋 열기
covid = pd.read_csv(drive_path + "Covid19InfState.csv", index_col = 'seq',encoding='euc-kr')

# stateDt 컬럼을 보면 하루에 두번 이상 체크된 것들을 확인하고, last 값만 삭제
covid = covid.drop_duplicates(["stateDt"],keep='last')

covid["stateDt"] = covid["stateDt"].astype(str)
covid["stateDt"] = pd.to_datetime(covid["stateDt"])

covid["stateDt_year"] = covid["stateDt"].dt.year
covid["stateDt_month"]= covid["stateDt"].dt.month
covid["stateDt_day"]= covid["stateDt"].dt.day

# covid 데이터셋은 2020년 자료와 2021년 자료가 함께 있기 때문에, 2020년 데이터만 따로 추출
covid_20 = covid[covid["stateDt_year"]==2020]

# decideCnt(확진자 수)의 결측치를 0으로 채워줌.
covid_20 = covid_20.fillna(0)
covid_20[["stateDt_month","decideCnt"]]

# 증가량을 구하기 위해 월별 누적확진자수를 구하기
# 당월 누적 확진자 수 - 전월 누적 확진자수 = 당월 신규 확진자 수

# 매월 마지막 데이터만 남기고 모두 drop. -> 매월의 마지막날 누적확진자 수만 남게 됨.
covid_20_month = covid_20.drop_duplicates(["stateDt_month"], keep = 'last')
```

분석 방법

코로나 확진자 수와 영화관 관객수 간의 상관관계 분석

```
# shift 를 사용하여 당월 누적 확진자 수에서 전월 누적 확진자 수를 빼줌.
# fillna(0)을 사용하여 결측치를 0으로 메꿔줌.
covid_20_month["delta"] = covid_20_month['decideCnt'] - covid_20_month['decideCnt'].shift(1).fillna(0)
covid_20_month["delta"]

# 월별 신규 확진자 수 그래프 그리기
y = covid_20_month["delta"]
x = covid_20_month["stateDt_month"]
plt.xlabel('월 (2020)')
plt.ylabel('신규 확진자 수')
plt.title('월별 신규 확진자 수')
plt.bar(x,y)

# 영화 관객수/매출액 데이터셋 열기
movie = pd.read_excel(drive_path + "kobis.xlsx", index_col="순번")

# 개봉일 컬럼의 타입은 이미 datetime이기 때문에 바로 연월일 컬럼 만들기
movie["movie_released_year"] = movie["개봉일"].dt.year
movie["movie_released_month"] = movie["개봉일"].dt.month
movie["movie_released_day"] = movie["개봉일"].dt.day

# 2020년도 개봉 영화 데이터만 따로 만들기
movie_20 = movie[movie["movie_released_year"]==2020]

# 분기별이 아닌 월별 데이터를 보기 위해 groupby로 월별 관객 수 집계
movie_20_month_user = movie_20["전국\n관객수"].groupby(movie_20["movie_released_month"]).sum()

# 월별 영화관 전국 관객 수 그래프 그리기
x = movie_20_month_user.index
y = movie_20_month_user.values
plt.xlabel('월 (2020)')
plt.ylabel('전국 관객 수(단위:천만)')
plt.title('월별 전국 영화관 관객 수')
plt.bar(x,y)

# 확진자 수와 전국 영화관 관객 수와의 상관관계
fig, ax1 = plt.subplots()
ax2 = ax1.twinx()
line_covid = ax1.plot(x, covid_20_month["delta"], color='green',label="신규 확진자 수")
ax1.set_xlabel('월(2020)')

ax1.set_ylabel('신규 확진자 수')
line_user = ax2.plot(x, movie_20_month_user.values, color='deeppink',label="전국 관객 수")
ax2.set_ylabel('전국 관객 수 (단위:천만)')

lines = line_covid + line_user
labels = [l.get_label() for l in lines]
plt.legend(lines, labels, loc=2)
plt.title('코로나19 확진자 수와 전국 영화관 관객 수의 상관관계')
plt.rcParams["figure.figsize"] = (12, 5)

plt.show()
```

분석 방법

```
# 데이터 불러오기 및 기본 세팅 (영화관 자료)
movie = pd.read_excel(drive_path+'KOBIS_.xlsx', index_col = '순번')
moviedf = movie[['영화명','개봉일','전국\n매출액','전국\n관객수']]
moviedf.columns = ['name','releaseday','revenue','customers']

#timestamp
moviedf['releaseday'] = pd.to_datetime(moviedf['releaseday'])
moviedf['year'] = moviedf['releaseday'].dt.year
moviedf['year'] = moviedf['year'].astype('str')

#quarter switch
moviedf['q'] = moviedf['releaseday'].dt.quarter
moviedf['q'] = moviedf['q'].replace(1,'Q1')
moviedf['q'] = moviedf['q'].replace(2,'Q2')
moviedf['q'] = moviedf['q'].replace(3,'Q3')
moviedf['q'] = moviedf['q'].replace(4,'Q4')

#datematching
moviedf['Years'] = moviedf['q'] + ' - ' + moviedf['year']
moviedf['years'] = moviedf['year'] + ' - ' + moviedf['q']

# remove 2016 2017
moviedf = moviedf[moviedf['year']!='2016']
moviedf = moviedf[moviedf['year']!='2017']

#remove 2020 q3 q4
moviedf = moviedf[moviedf['Years'] != 'Q3 - 2020']
moviedf = moviedf[moviedf['Years'] != 'Q4 - 2020']

# 영화 개봉 개수
movie_count = pd.DataFrame(moviedf['years'].value_counts())
movie_visual = movie_count.sort_index()

count = moviedf.groupby('years').count()
lst = ['Q1 - 2018', 'Q2 - 2018','Q3 - 2018','Q4 - 2018','Q1 - 2019','Q2 - 2019','Q3 - 2019','Q4 - 2019','Q1 - 2020','Q2 - 2020']
count.index = lst

# 분기별 수익 합
sum_r_c = moviedf.groupby('years').sum('revenue')
sum_r_c.index = lst

# 분기별 개봉 영화 개수, 영화 수익, 영화 관객 dataframe
final_mv_df = sum_r_c.join(count.name)
asia_revenue= asia_scb.set_index('Years')

#넷플릭스 구독자 수 병합
final_mv_df= final_mv_df.join(asia_revenue.Subscribers)

final_mv_df.columns = ['movie_revenue','movie_customers','movie_count','neflixsubscriber']
display(final_mv_df.corr())

plt.figure(figsize=(8,6))
sns.heatmap(final_mv_df.corr())
```

```
# 데이터 불러오기 및 기본 세팅 (넷플릭스 구독자수)
scb_spread = pd.read_csv(drive_path + 'NetflixSubscribersbyCountryfrom2018toQ2_2020.csv')

asia_scb=scb[scb['Area']=='Asia-Pacific'][['Area','Years','Subscribers']]

#분기별 subscriber 평균
mean_dic={}
for i in range(len(scb_spread.columns)-1):
  mean_dic[scb_spread.columns[i+1]]= (scb_spread.iloc[:,i+1].mean())
mean_scb = pd.DataFrame(mean_dic.values(),index = mean_dic.keys(), columns = ['subscribers'])
```

영화관자료와 넷플릭스 구독자수와의 상관관계 분석

분석 방법

가설 3번 1인 가구 증가에 따른 이용자수 증가를 검증

사용자료 : 연령별 1인 가구 분석, 지역별 넷플릭스 구독자 및 수익 현황
사용 상관계수 : 피어슨 상관계수

이 가설 검증의 전제조건
1. 1인 가구 수는 20,30대로 한정. (한국 OTT서비스 매출의 60%는 20, 30대이다.)
2. 데이터간 상관관계 파악을 위해 자료가 없는 부분은 기사에서 수치를 발췌함.
2017년 1분기 구독자 수 = 4660000명
출처 - https://www.screendaily.com/news/netflix-reveals-giant-subscriber-surge-in-asia-pacific-region/5145716.article#:~:text=The%2014.5m%20subscriber%20level,through%20September%2030%20this%20year.

연령별 1인가구 분석
1. 년도별 전국 1인가구중 20대와 30대를 추출하여 합함

지역별 넷플릭스 구독자 및 수익 현황

1. 아시아 지역 넷플릭스 구독자 정보중 년도별 구독자수를 추출
2. 1인 가구와의 상관관계를 알아보기 위해 기존 데이터에 column으로 추가 및 시각화

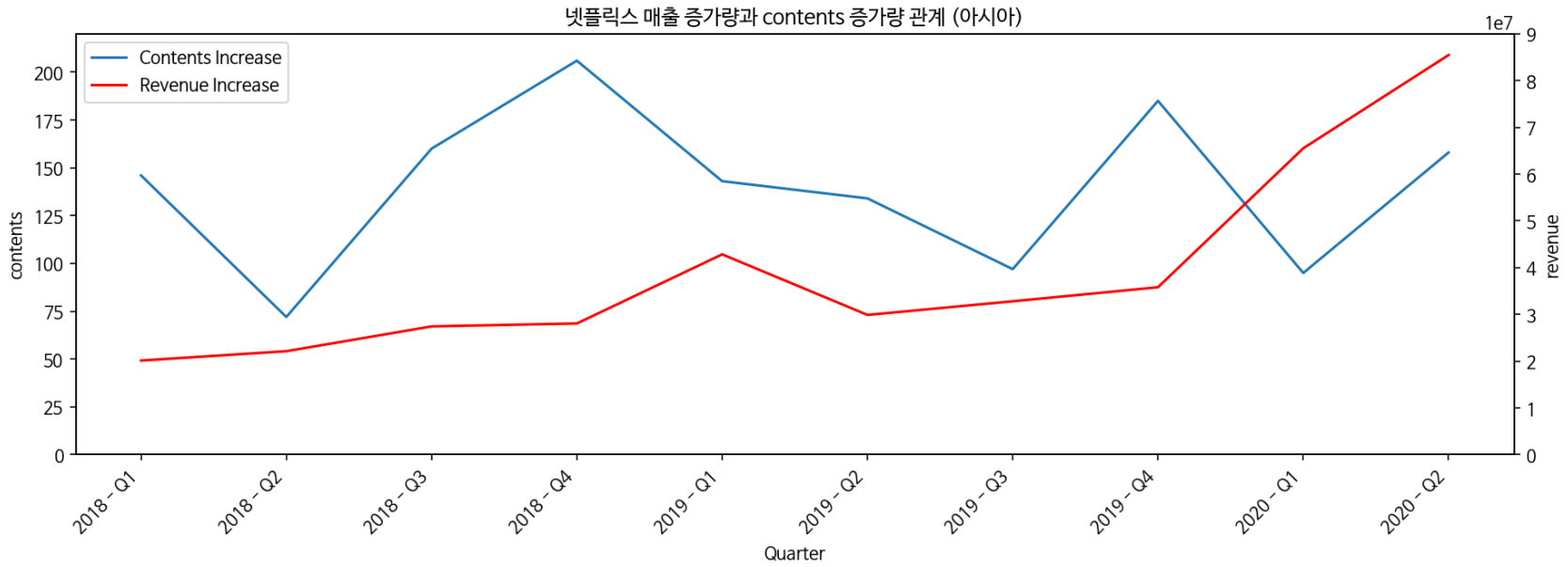
분석 방법

```
import numpy as np
import pandas as pd
from google.colab import drive
import matplotlib.pyplot as plt
import matplotlib.font_manager as fm

# 2017년 ~ 2019년 20,30대 1인가구 증가
df2 = pd.read_excel(drive_path + "ᄉᆞᄒᆞᆼᆞᄆᆞᄂᆞᆺᆞᄋᆞᄒᆞᄂᆞᄅᆞᄒᆞᆼᄇᆞᄒᆞᆯ_1ᄋᆞ ᄂᆞᄀᆞ ᄒᆞᄀᆞᄀᆞᄉᆞ ᄀᆞᄀᆞᄂᆞᄀᆞᄀᆞ_20210402102821_ᄇᆞᄀᆞᄂᆞᄉᆞ ᄒᆞᄀᆞ(ᄂᆞᄀᆞᄀᆡ).xlsx")
age20 = list(map(int,df2.loc[3:4][['2017','2018','2019']].sum()))
age30 = list(map(int,df2.loc[5:6][['2017','2018','2019']].sum()))
ind = ['2017년','2018년','2019년']
df3 = pd.DataFrame({'20s':age20,'30s':age30}, index = ind)
df3['20s+30s'] = list(df3.sum(axis=1))

# 넷플릭스 구독자 변화 데이터프레임
subs =  pd.read_csv(drive_path + "DataNetflixSubscriber2020_V2.csv")

#아시아 지역 넷플릭스 구독자
asia_subs = subs[subs['Area'] == 'Asia-Pacific']
asia_subs['year'] = asia_subs['Years'].astype(str).str.split('-').str.get(1)

subscriber = []
# 2017년 1분기 구독자 수 = 4660000명
subscriber.append(4660000)

# 2018년, 2019년도 1분기 기준 구독자
for i in [1,5]:
  subscriber.append(list(asia_subs['Subscribers'])[i-1])
asub = pd.DataFrame({'subs':subscriber}, index = ind)

# 1인 가구 데이터 column추가
asub['1인 가구(명)'] = list(df3['20s+30s'])

fontpath = '/usr/share/fonts/truetype/nanum/NanumBarunGothic.ttf'
font = fm.FontProperties(fname=fontpath, size=10)
fm._rebuild()

x = list(asub['subs'])
y = list(asub['1인 가구(명)'])

fig, ax = plt.subplots(2)
fig.suptitle('1인 가구 수 변화와 넷플릭스 구독자 수 변화')
ax[0].set_title('(1)넷플릭스 구독자 수 변화')
ax[0].set(ylabel = '구독자 수 (백만명)')
ax[0].plot(ind,x)
ax[1].set_title('(2) 1인 가구 수(명) 변화')
ax[1].set(ylabel='1인 가구 수 (백만명)')
ax[1].plot(ind,y, color = 'm')

for ax in fig.get_axes():
    ax.label_outer()
```

분석 결과

가설1번 단순 컨텐츠의 증가로 인한 접근성 확대에 관한 결과

시각화 결과



분석결과

- 매출량과 contents량 모두 전반적으로 늘고 있음.
- contents양이 늘어나고 매출도 반응하여 늘어나 보임.
- contents량이 매출량에 직접적인 영향을 미쳤다고 확정짓기엔 데이터가 부족할 수 있음
- 매출이 2020년 들어 증가폭이 확연히 높아지는데 '코로나19' 의 영향으로 보임

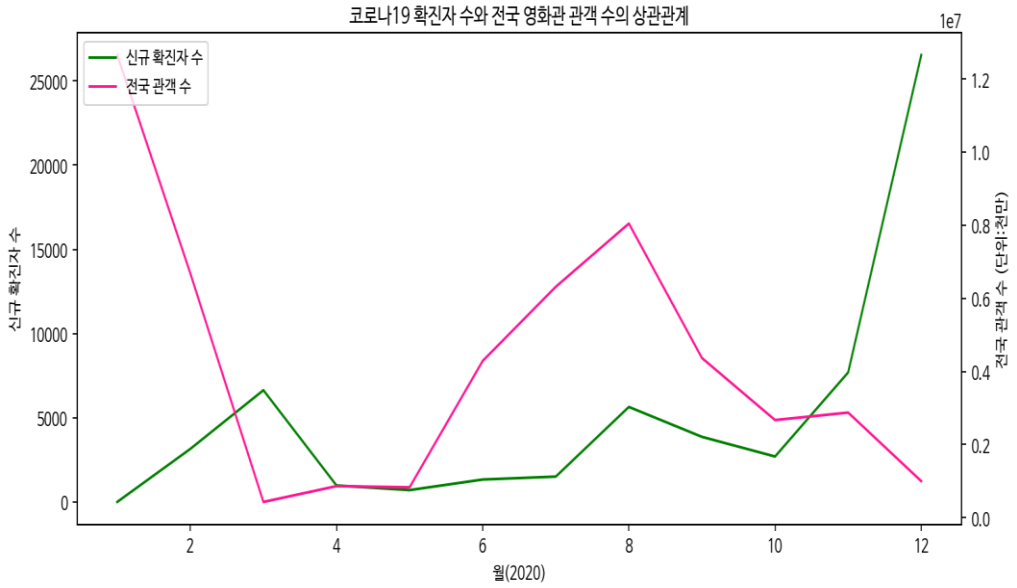
아쉬운점

- 2018년 이전 매출 데이터가 없어 장기간에 걸친 추이를 판단하지 못함
- contents 수보다는 아시아 contents에 투자한 제작비 등의 수치가 가설을 증명하기에 더욱 유용할 수 있음
- 넷플릭스 외 타 OTT와 비교하여 컨텐츠 투자 대비 매출 추이를 보면 관련성을 판단하기 더 좋았을 것 같음

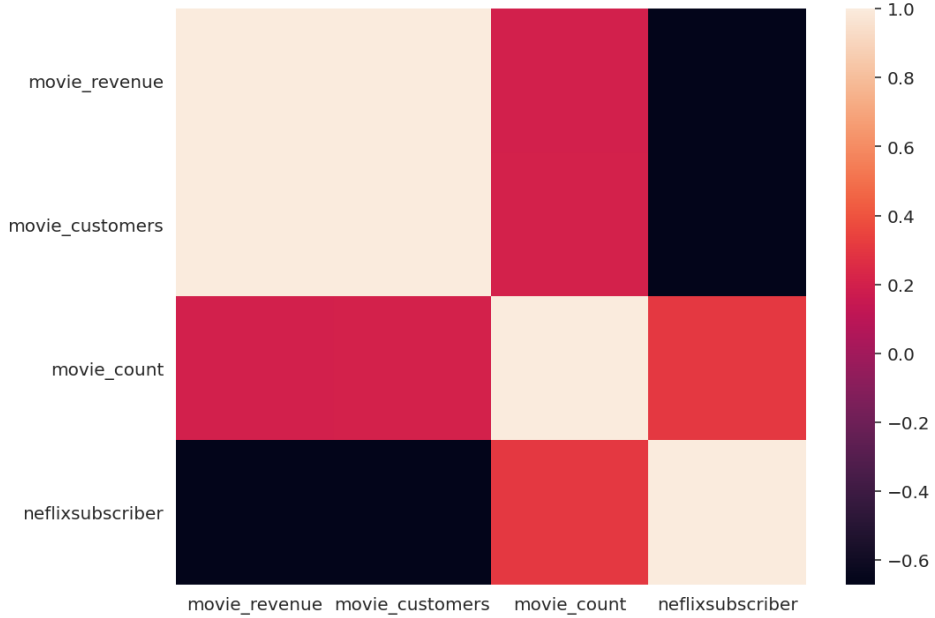
분석 결과

가설2번 코로나의 영향으로 인한 영화관의 부진에 따른 이용자 증가에 대한 결과

시각화 결과



	신규 확진자 수	전국 영화관 관객 수
신규 확진자 수	1.000000	-0.343074
전국 영화관 관객 수	-0.343074	1.000000



	movie_revenue	movie_customers	movie_count	neflixsubscriber
movie_revenue	1.000000	0.999227	0.202625	-0.664907
movie_customers	0.999227	1.000000	0.208344	-0.670818
movie_count	0.202625	0.208344	1.000000	0.303699
neflixsubscriber	-0.664907	-0.670818	0.303699	1.000000

분석결과

- 음의 상관관계를 띈다.
- 전체적인 흐름을 보면 코로나 확진자 수가 증가함에 따라 영화관 관객수가 감소함,
- 7,8월에 신규확진자수 대폭증가에도 영화관 관객 수는 증가 추세를 유지하다가 2020.08.23 사회적 거리두기 2단계 격상과 함께 2020년 9월 영화관 관객수 감소
- 2020.12.08 사회적 거리두기 2.5단계 격상과 함께 2020년 12월 관객수가 감소 함께 2020년 12월 관객수가 감소

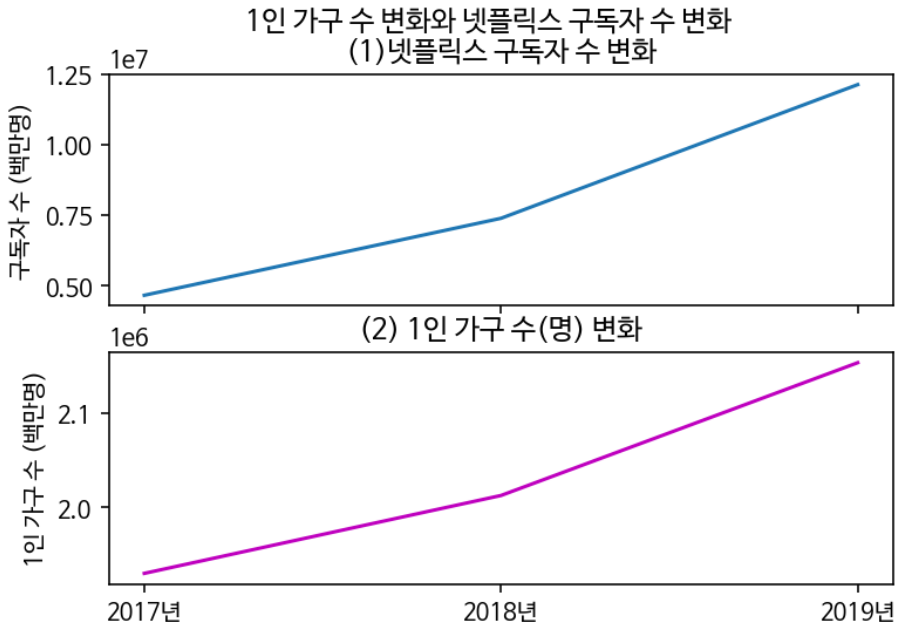
아쉬운점

- 넷플릭스 revenue,subscriber 데이터가 2020-2분기 까지만 제공되고 있다는 점, 국가별이 아닌 대륙별로 카테고리화가 되어있다는 점
  (한국 넷플릭스 수입,구독자의 2020년 전체 흐름과 코로나 확진자 수를 비교불가)
- 7,8월에 코로나 확진자 수가 늘었음에도 불구하고 즉각적으로 영화관 관객수가 줄지 않았다는 점에서, 해당 월의 한국의 revenue, subscriber 데이터의 흐름 분석이 추가적으로 필요했지만 데이터의 부재로 분석하지 못했다는 점.

분석 결과

가설3번 1인 가구 증가에 따른 이용자수 증가에 대한 결과

시각화 결과



	subs	1인 가구(명)
subs	1.000000	0.999988
1인 가구(명)	0.999988	1.000000

분석결과

- 1인 가구수의 증가와 넷플릭스 구독자 간의 양의 상관관계가 있을 것으로 보임

아쉬운점

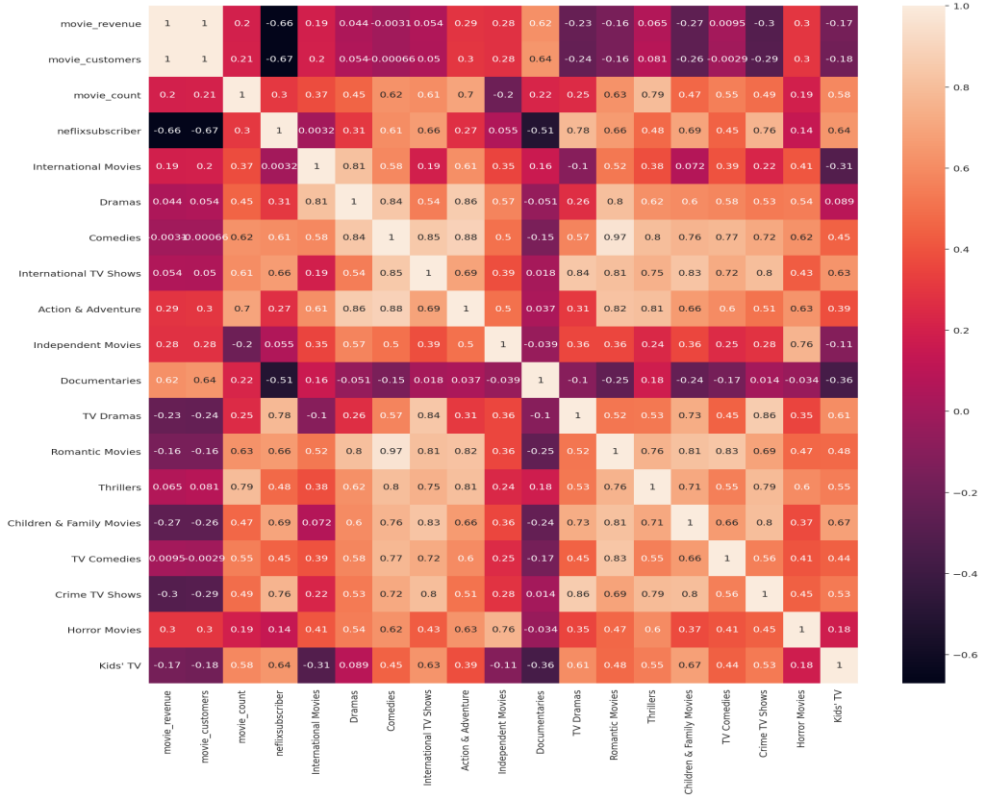
- 가설을 입증하기 위해서는 1인 가구수에 대한 더 많은 데이터가 필요함

분석 결과 (추가)

최종 상관 분석

넷플릭스 구독자수와 상관관계가 큰 요소는 무엇일지에 대해서 상관행렬을 시각화함
그 결과 상관계수 절대값 0.5를 넘는 것을 기준으로 영화관 수익(= 영화관 고객수), 넷플릭스 코미디 프로그램, international TV Show, 다큐멘터리, TV 드라마, 로맨틱 영화, 가족영화, 범죄 TV쇼, 어린이 TV 등이 있으며 그 순위는 다음과 같은 순으로 나타남.

```
#최종 상관 행렬 및 시각화
final_corr_df = final_mv_df.join(cat_df3)
display(final_corr_df.corr())
plt.figure(figsize=(15,15))
sns.heatmap(final_corr_df.corr(),annot=True)
```




```
#상관관계 순위
corr_rank_df = pd.DataFrame(final_corr_df.corr().loc[:,'neflixsubscriber'])

ranking = corr_rank_df.apply(abs)
ranking = ranking.sort_values(by='neflixsubscriber',ascending=False)
ranking['Rank'] = range(len(ranking))
ranking

corr_final_df = corr_rank_df.join(ranking['Rank'])
corr_final_df=corr_final_df[corr_final_df['Rank']!=0]
corr_final_df.sort_values(by='Rank')
```

	neflixsubscriber	Rank
TV Dramas	0.780931	1
Crime TV Shows	0.764317	2
Children & Family Movies	0.694781	3
movie_customers	-0.670818	4
movie_revenue	-0.664907	5
International TV Shows	0.661036	6
Romantic Movies	0.656836	7
Kids' TV	0.636971	8
Comedies	0.608305	9
Documentaries	-0.509526	10
Thrillers	0.475148	11
TV Comedies	0.450090	12
Dramas	0.306544	13
movie_count	0.303699	14
Action & Adventure	0.270889	15
Horror Movies	0.144157	16
Independent Movies	0.054905	17
International Movies	0.003158	18