배우의 sns지수 와 매출액

1. 목적: 영화배우가 영화 제작에 앞서 고려해야 할 중요한 요소인지를 알아본다.
2. 데이터 사용:
   (1) 영진위 기간별(2010.01.01~2018.12.31) 박스오피스 순위 데이터
   : 최근 동향을 알아보기 위해 최근3년(2016~2018)간의 데이터만 사용.
   (2) 트위터 (온라인 부분 대표로 설정) 트윗수+좋아요수
   : 영화 개봉30일전~개봉후7일 기간 동안의 대표배우5명의 'sns지수' 도출
3. 가설 설정:
   가설: 배우의 sns 화제성은 영화 매출액에 영향을 미친다.

1. 영진위 데이터 전처리

: columns = (year, rank, 영화명, 개봉일, 매출액, 누적매출액, 관객수, 누적관객수, 스크린수, 상영횟수, 대표국적, 제작사, 배급사, 등급, 장르, 감독 배우)

1-1. 3년(2016~2018), 각 연도별 매출top30 영화

[Year, 영화명, 개봉일, 매출액, 배우] 총 5개의 columns사용

1) 2016년도 top30로 간추리기(iloc( ) 이용)

```
top30_2016 = top100_2016_step1.iloc[0:30]
```

2) 필요한 컬럼 미리 추가 (개봉30일전, 개봉7일후)

```
# 개봉30일전, 개봉7일후 컬럼 생성
df_2016_0507['개봉30일전'] = df_2016_0507['개봉일'] - timedelta(days=+30)
df_2016_0507['개봉7일후'] = df_2016_0507['개봉일'] + timedelta(days=+7)
```

3) 한 작품당 배우 5명씩 넣기 (한 행에 있는 여러values 쪼개기)

```
top30_2016_actor = top30_2016["배우"].str.split(',')
type(top30_2016_actor)

# series -> dataframe으로 변환
top30_2016_actor = top30_2016_actor.apply(lambda x: pd.Series(x))
type(top30_2016_actor)
```

```
pandas.core.frame.DataFrame
```

```
top30_2016_actor.stack().head()
```

```
0  0     공유
   1    김정희
   2    정유미
   3    김윤호
   4    마동석
dtype: object
```

```
top30_2016_actor = top30_2016_actor.stack().to_frame('배우명')
top30_2016_actor.head()
```

		배우명
0	0	공유
	1	김정희
	2	정유미
	3	김윤호
	4	마동석

위 그림과 같이 배우 구별 진행 -> 원래의 2016 top30 파일과 merge

->총 146개 결과값

나머지 2017년도 top30, 2018년도 top30 도 위와 같이 실행함.

2. 트위터 크롤링 (twitterscraper라는 오픈소스 이용)

전처리한 영진위 데이터를 이용하여 트윗수 및 트윗좋아요수 크롤링

```
name = low_2018_step2.배우명

date1 = low_2018_step2.개봉일
date2 = low_2018_step2.개봉7일후
date3 = low_2018_step2.개봉30일전

prediction3 = low_2018_step2.트윗수1
prediction4 = low_2018_step2.라잌수1
```

```
low_2018_step2.tail()
```

...

```
filename = 'low5_18.csv'

a = 0
while (a<=23):
    print(a)
    query = ""+name[a]+" since:"+str(date3[a])+" until:"+str(date2[a])   # 개봉 30일전 부터  개봉 7일 후
    counttweet = 0
    countlikes = 0
    for tweet in query_tweets_once(query):
        counttweet = counttweet + 1 + tweet.retweets
        countlikes = countlikes + tweet.likes
    prediction3[a] = counttweet
    prediction4[a] = countlikes

    a = a + 1

    time.sleep(2)

    low_2018_step2.to_csv(filename, index=False, encoding='cp949')
```

크롤링한 파일에 sns지수를 나타내는 sns언급도 컬럼추가

```
df_look['sns언급도'] = df_look['라잌수1'] + df_look['트윗수1']
```

중복되는 배우명 groupby로 정리

```
top30_sns_2016 = df_look_test.groupby(df_look_test['배우명']).mean()[['sns언급도', '매출액']]
```

```
top30_sns_2016.head()
```

배우명	sns언급도	상영수 대비 매출액	매출액
J.K. 시몬스	30.0	354491.193306	1.999614e+10
T.J. 밀러	6.0	334407.225446	2.759562e+10
강동원	91056.5	532826.045741	5.855252e+10

나머지 2017년도 top30, 2018년도 top30 도 위와 같이 실행함.

3. 분석

- 상관분석을 이용하여 산포도 그리기

3-1. 2016년도

1) 상관분석1 (sns지수 전체 값 이용, min:10, 표본 118개)

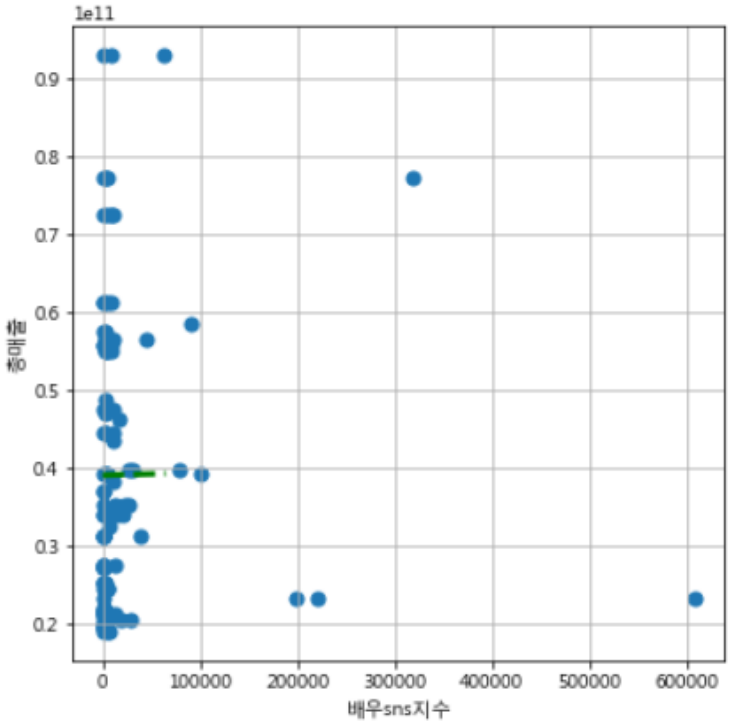
numpy 제공 corrcoef 메서드를 이용하여 두 변수 간의 관계가 있는지 보았다.

변수1: sns지수 / 변수2: 매출액

```
print(np.corrcoef(top30_sns_2016["sns지수"], top30_sns_2016["매출액"])) # 0.013
```

```
[[1.         0.01274545]
 [0.01274545 1.        ]]
```

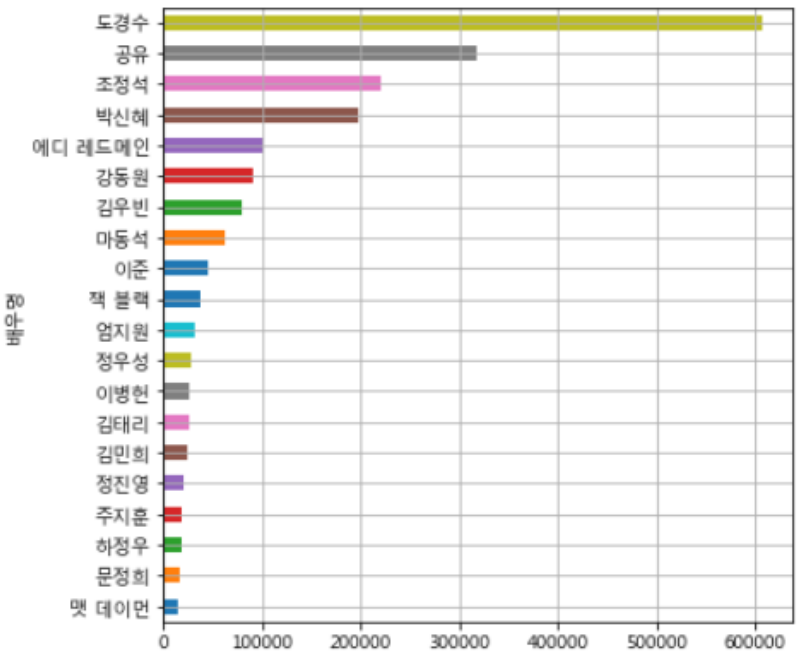
0.3 에도 한참을 못 미친다.



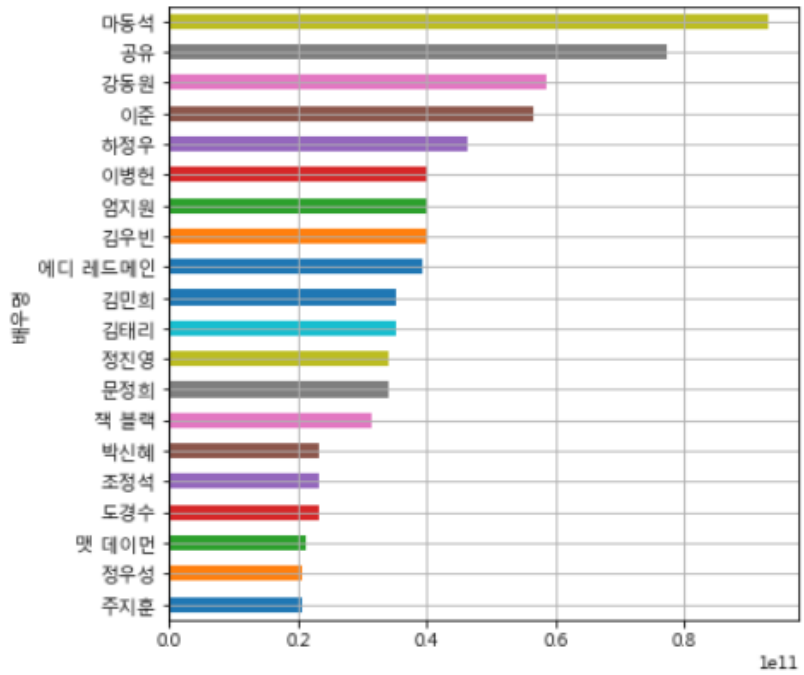
2016년도 sns지수1 (범위 10~)

결론: sns언급도가 영화의 매출에 영향을 미치지는 못함

2016년 TOP20 (sns지수, 매출액) 그래프 비교



2016년 sns지수 TOP20



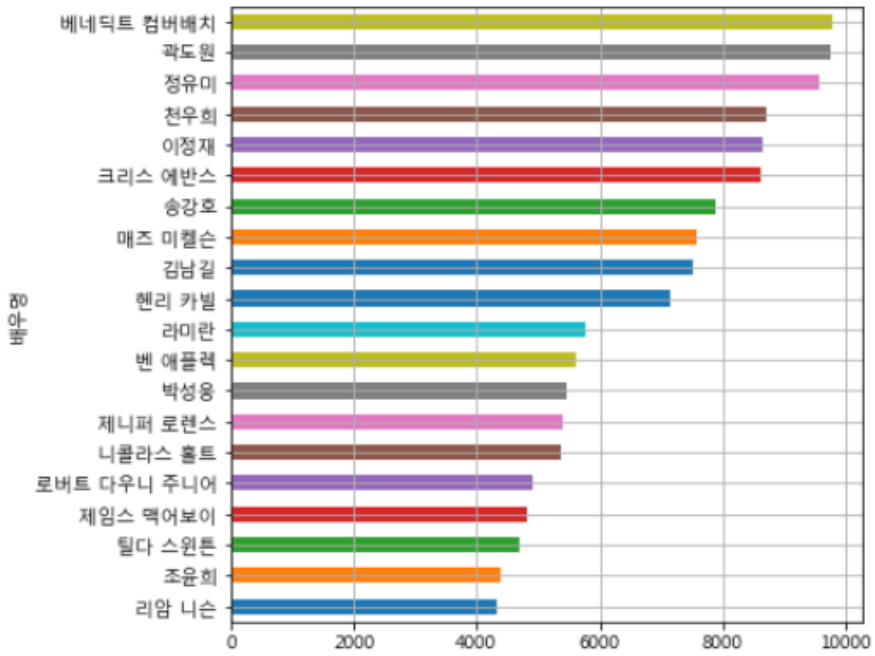
2016년 매출액 TOP20

2) 상관분석2 (sns지수 범위 조정)

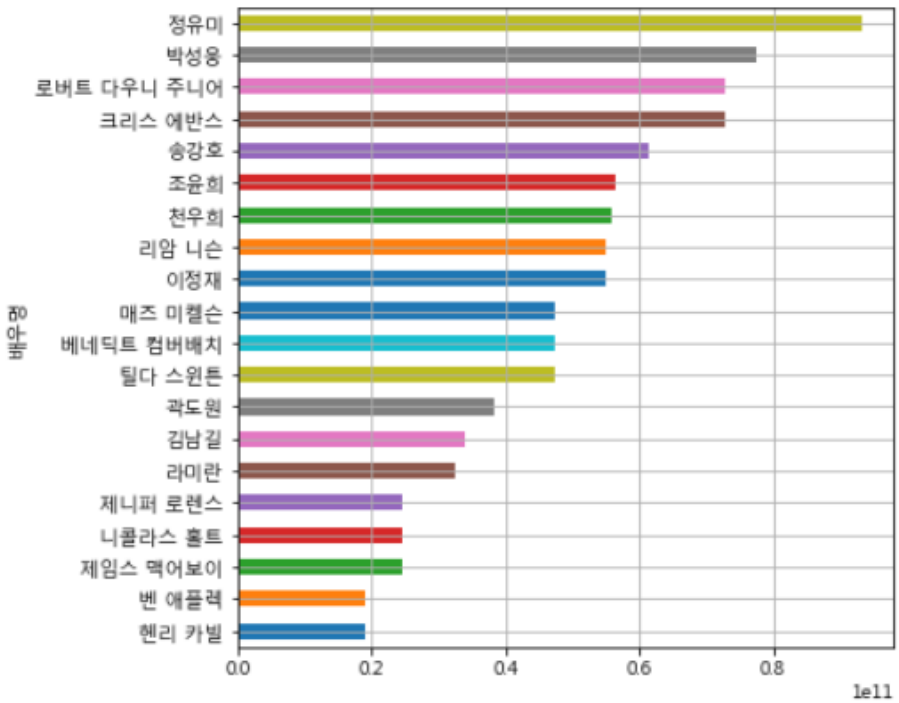
: sns언급도 null값 및 outlier로 보이는 값의 데이터 삭제 및 배우의 범위 조절 여러 차례

➔ sns지수 500이상, 1만 이하에서 아주 약한 양의 관계성을 띔 (표본 63개)

sns지수 중위권 배우 순위 (그래프로 시각화)



2016년 중위권 sns지수 순위_1



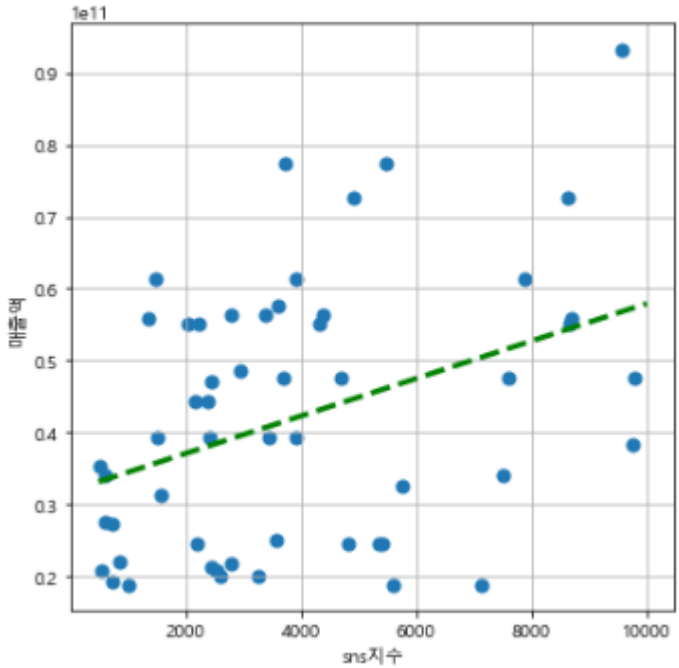
2016년 중위권 매출액 순위_1

```
print(np.corrcoef(top30_sns_2016_btw["sns지수"], top30_sns_2016_btw["매출액"]))
```

```
[[1.         0.37660472]
 [0.37660472 1.        ]]
```

```
fp1 = np.polyfit(top30_sns_2016_btw['sns지수'], top30_sns_2016_btw["매출액"], 1)
fp1
# array([2.60731662e+06, 3.18558869e+10]) = [기울기a, 절편b]  (y= ax+b)
```

```
array([2.60731662e+06, 3.18558869e+10])
```



2016년도 sns지수 영향2 (범위 500~10,000)

최종결론:

2016년도 매출 상위30개의 영화 대표배우 총 118명 중, sns지수 중위권인 배우 54명의 영화 매출 영향도가 상위 27명, 하위 37명의 배우보다 크다고 볼 수 있다.
(상위권 매출과의 상관관계가 없는 결과로 나왔고, 하위권의 경우 sns지수가 유의미하지 않다고 판단하였음)

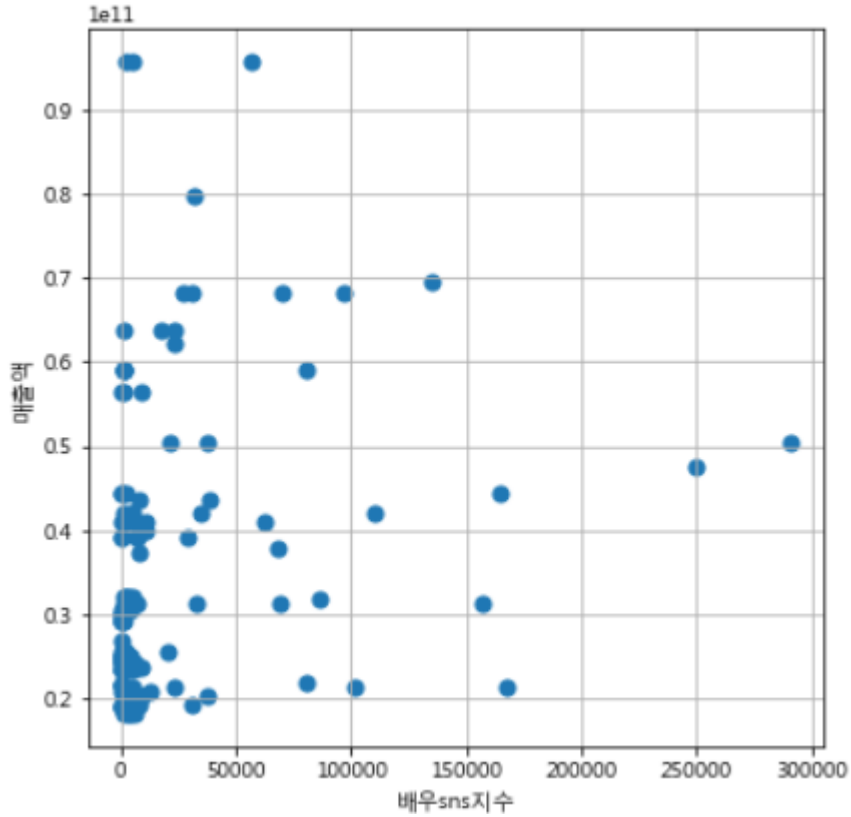
제안: 영화 제작 시 배우 캐스팅에 있어서 무조건 sns화제성이 높은 배우보다는 어느 정도의 sns화제성 +α (연기력, 성별 등) 를 가진 배우를 캐스팅하는 것을 추천한다.

3-2. 2017년도

1) 상관분석1 (sns지수 전체 값 이용, min:10, 표본 131개)

```
print(np.corrcoef(top30_sns_2017["sns지수"], top30_sns_2017["매출액"]))  # 0.23
```

```
[[1.         0.22709997]
 [0.22709997 1.        ]]
```



2017년 sns지수-매출액 관계도

2) sns지수 범위 조정

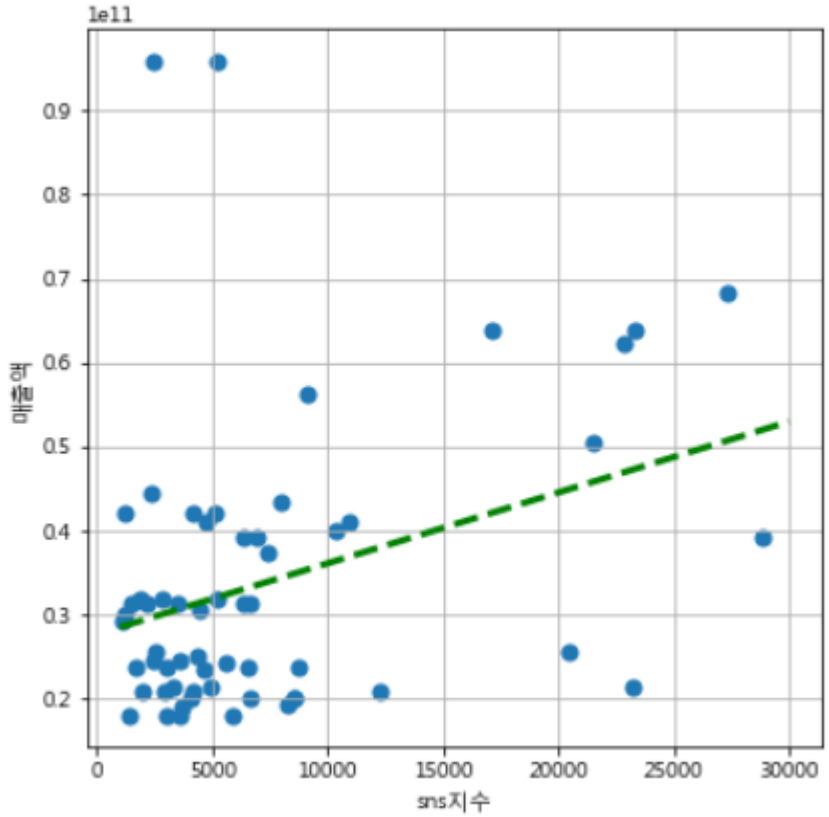
- sns 1000이상 3만 이하 (표본 59개)

```
print(np.corrcoef(top30_sns_2017_btw2["sns지수"], top30_sns_2017_btw2["매출액"]))   # 0.34
# 약하지만 상관성이 생긴 것이 확인가능하다!
```

```
[[1.        0.3377981]
 [0.3377981 1.       ]]
```

```
fp1 = np.polyfit(top30_sns_2017_btw2["sns지수"], top30_sns_2017_btw2["매출액"], 1)
# array([8.44208156e+05, 2.76639534e+10]) = [기울기a, 절편b]  (y= ax+b)
```

```
array([8.44208156e+05, 2.76639534e+10])
```

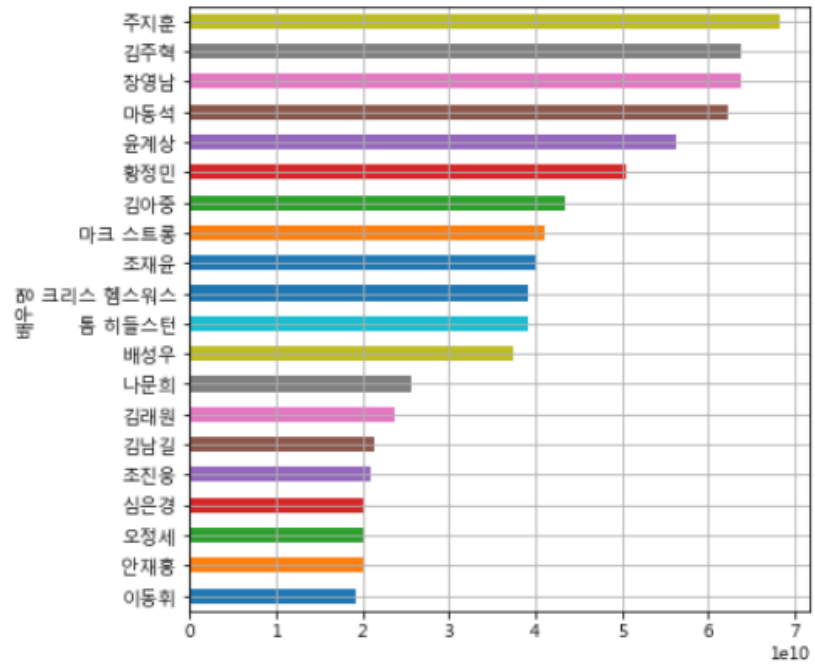


2017년 sns지수-매출액 2

그래프로 시각화

2017년 sns 지수 중위권 배우 순위

2017년 중위권 sns 지수 순위 1



2017년 중위권 매출액 순위 1

최종결론:

2016년도 매출 상위30개의 영화 대표배우 총 131명 중, sns지수 중위권인 배우 59명의 영화 매출 영향도가 상위 25명, 하위 47명의 배우보다 크다고 볼 수 있다.

(상위권 매출과의 상관관계가 없는 결과로 나왔고, 하위권의 경우 sns지수가 유의미하지 않다고 판단하였음)

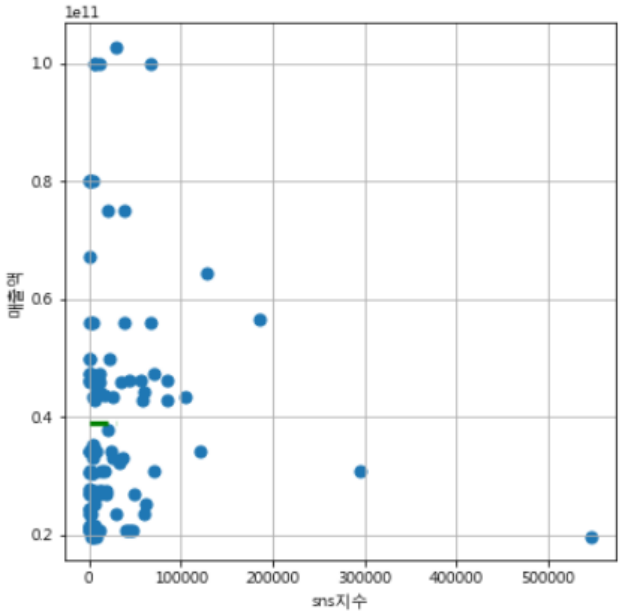
3-3. 2018년도

1) 상관분석1 (sns지수 전체 값 이용, min:10, 표본 121개)

역시나 관계가 없는 것으로 나왔다.

```
print(np.corrcoef(top30_sns_2018["sns지수"], top30_sns_2018["매출액"])) # 0.001
```

```
[[1.        0.0014513]
 [0.0014513 1.       ]]
```



2018년 sns지수-매출액 1

2) sns 지수 범위 조정

상위권 27명 / 중위권 53명 / 하위권 41명 (3만초과 / 2천~3만 / 2천 미만)

모두 매출액과 상관관계가 없었다. 범위를 여러 번 조정해봐도 상관관계 없는 것으로 확인

예) 중위권 53명

```
top30_sns_2018_btw2 = top30_sns_2018[(top30_sns_2018['sns지수'] >= 2000)&(top30_sns_2018['sns지수'] <= 30000)]
```

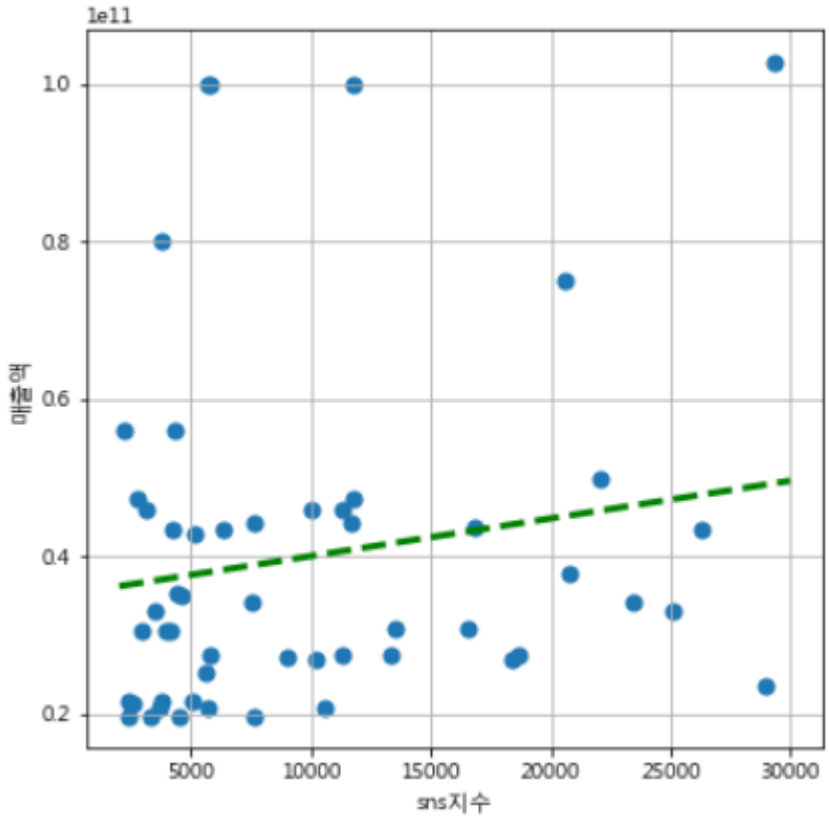
```
print(np.corrcoef(top30_sns_2018_btw2["sns지수"], top30_sns_2018_btw2["매출액"]))
# 상관성 없음
```

```
[[1.         0.16717467]
 [0.16717467 1.        ]]
```

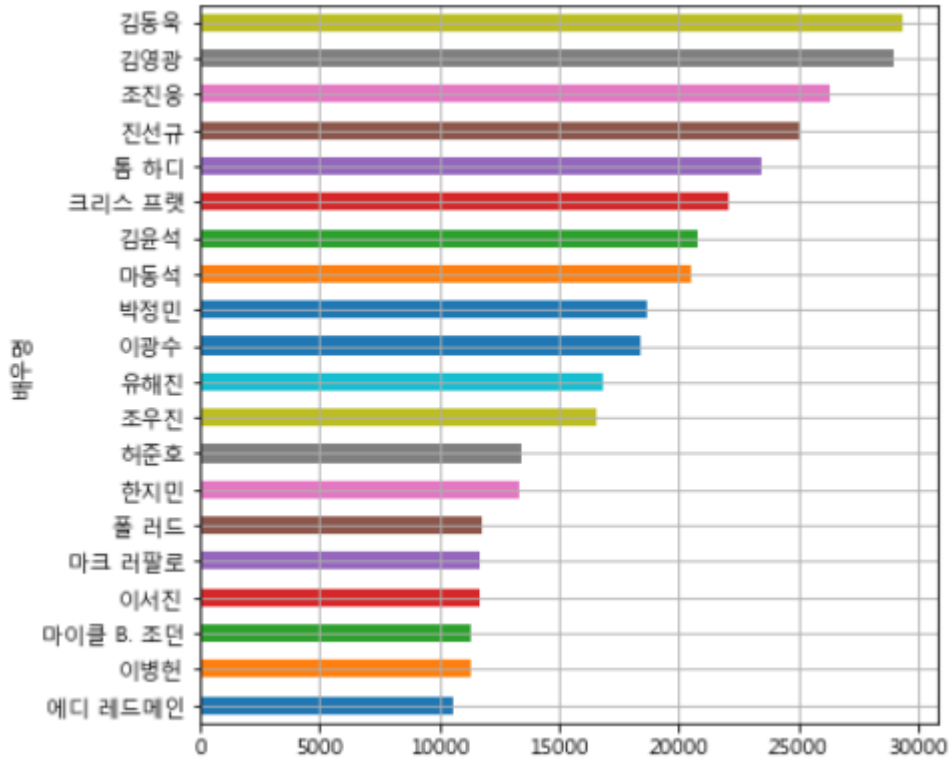
회귀선 그리기

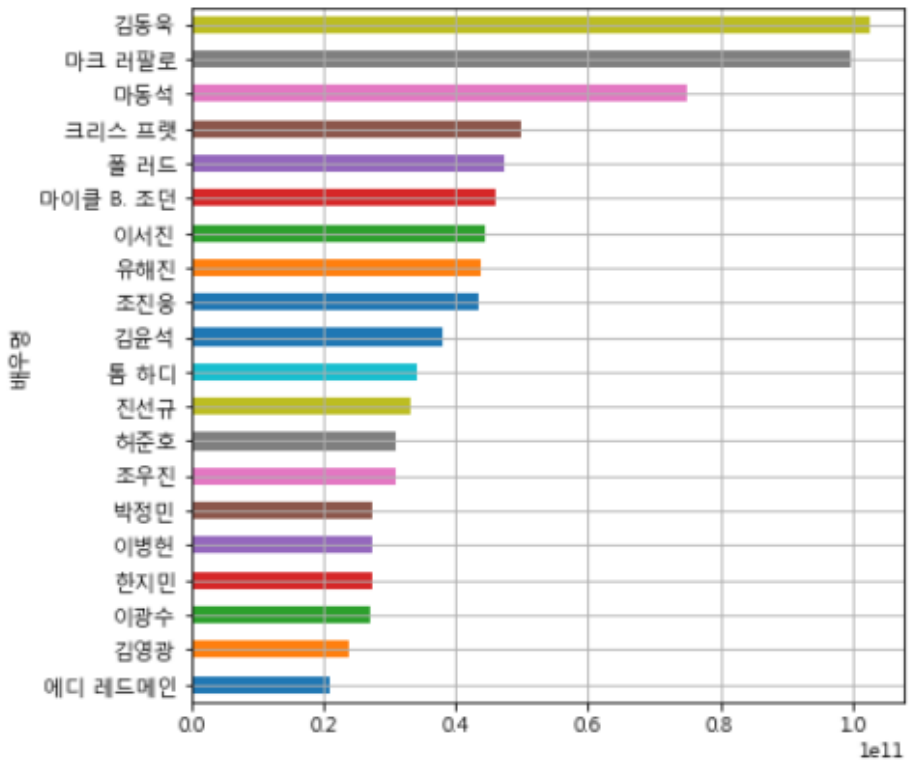
```
fp1 = np.polyfit(top30_sns_2018_btw2["sns지수"], top30_sns_2018_btw2["매출액"], 1)
fp1
```

```
array([4.77313861e+05, 3.52876081e+10])
```



그래프 시각화 (중위권 배우 sns 및 매출 순위 top20)


김동욱
김영광
조진웅
진선규
톰 하디
크리스 프랫
김윤석
마동석
박정민
이광수
유해진
조우진
허준호
한지민
폴 러드
마크 러팔로
이서진
마이클 B. 조던
이병헌
에디 레드메인
0
5000
10000
15000
20000
25000
30000


김동욱
마크 러팔로
마동석
크리스 프랫
폴 러드
마이클 B. 조던
이서진
유해진
조진웅
김윤석
톰 하디
진선규
허준호
조우진
박정민
이병헌
한지민
이광수
김영광
에디 레드메인
0.0
0.2
0.4
0.6
0.8
1.0
1e11