

Section1 Project

AI 18기 김보미

Our Process

다음 분기에 어떤 게임을 설계해야 할까?



01
EDA 수행



03
가설 검정



02
데이터 분석



04
결론 도출

About GAME DATA

초기 데이터의 구조를 먼저 살펴보자!

	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	
0	Candace Kane's Candy Factory	DS	2008.0	Action	Destineer	0.04	0	0	0	
1	The Munchables	Wii	2009.0	Action	Namco Bandai Games	0.17	0	0	0.01	
2	Otome wa Oanesama Boku ni Koi Shiteru Portable	PSP	2010.0	Adventure	Alchemist	0	0	0.02	0	
3	Deal or No Deal: Special Edition	DS	2010.0	Misc	Zoo Games	0.04	0	0	0	
4	Ben 10 Ultimate Alien: Cosmic Destruction	PS3	2010.0	Platform	D3Publisher	0.12	0.09	0	0.04	
...	
16593	16594	Ice Age 2: The Meltdown	GC	2006.0	Platform	Vivendi Games	0.15	0.04	0	0.01
16594	16595	Rainbow Islands: Revolution	PSP	2005.0	Action	Rising Star Games	0.01	0	0	0
16595	16596	NBA 2K16	PS3	2015.0	Sports	Take-Two Interactive	0.44	0.19	0.03	0.13
16596	16597	Toukiden: The Age of Demons	PSV	2013.0	Action	Tecmo Koei	0.05	0.05	0.25	0.03
16597	16598	The King of Fighters '95	PS	1996.0	Fighting	Sony Computer Entertainment	0	0	0.16	0.01

16598 rows × 10 columns

EDA

Data columns (total 10 columns):				
#	Column	Non-Null Count	Dtype	
0	Unnamed: 0	16598	non-null	int64
1	Name	16598	non-null	object
2	Platform	16598	non-null	object
3	Year	16327	non-null	float64
4	Genre	16548	non-null	object
5	Publisher	16540	non-null	object
6	NA_Sales	16598	non-null	object
7	EU_Sales	16598	non-null	object
8	JP_Sales	16598	non-null	object
9	Other_Sales	16598	non-null	object
dtypes: float64(1), int64(1), object(8)				

- 결측치 제거
16598개의 데이터 중 379개의 데이터는 무의미한 것으로 판단

EDA - Changing Data Type

- 'Year' column
 - float -> int
- 'oo_Sales' column
 - 단위는 millions
 - object -> float

NA_Sales	480K
EU_Sales	0.33M
JP_Sales	0K
Other_Sales	0.06



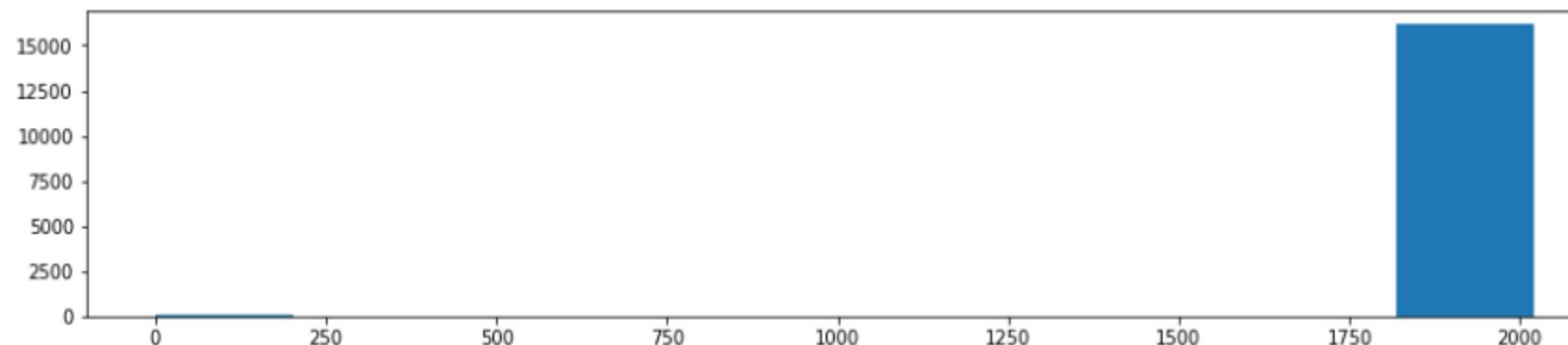
NA_Sales	480
EU_Sales	0
JP_Sales	0
Other_Sales	0.06

2	Year	16241	non-null	int64
3	Genre	16241	non-null	object
4	Publisher	16241	non-null	object
5	NA_Sales	16241	non-null	float64
6	EU_Sales	16241	non-null	float64
7	JP_Sales	16241	non-null	float64
8	Other_Sales	16241	non-null	float64
9	NaN	16241	non-null	float64

```
new_df[['NA_Sales','EU_Sales', 'JP_Sales', 'Other_Sales']] =  
df[['NA_Sales','EU_Sales', 'JP_Sales', 'Other_Sales']].apply(lambda x : x.str.replace('K', '').replace(r'\d*\.\d*M|0M', '0', regex=True))
```

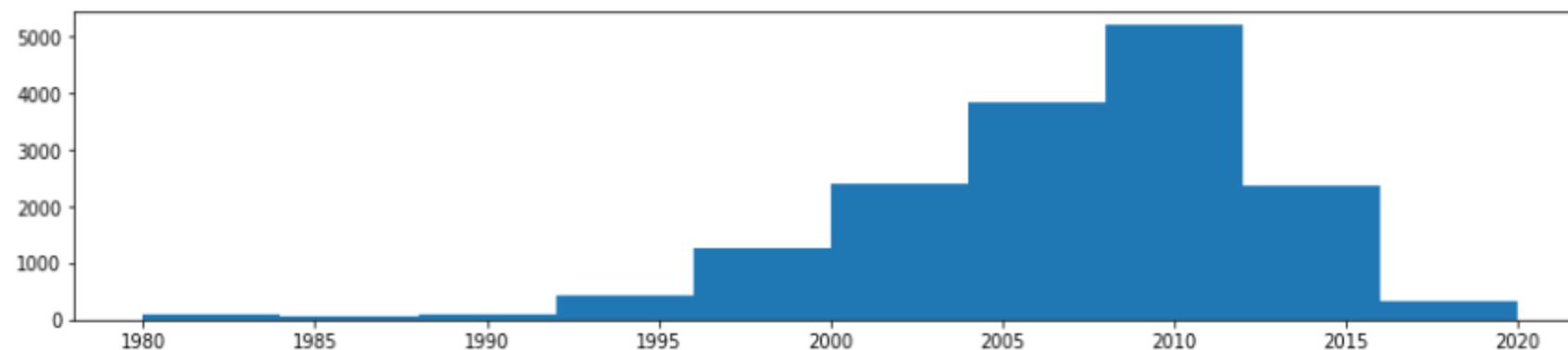
EDA - Remove Outlier

Year	
count	16241.000000
mean	1994.178437
std	155.484265
min	0.000000
25%	2003.000000
50%	2007.000000
75%	2010.000000
max	2020.000000

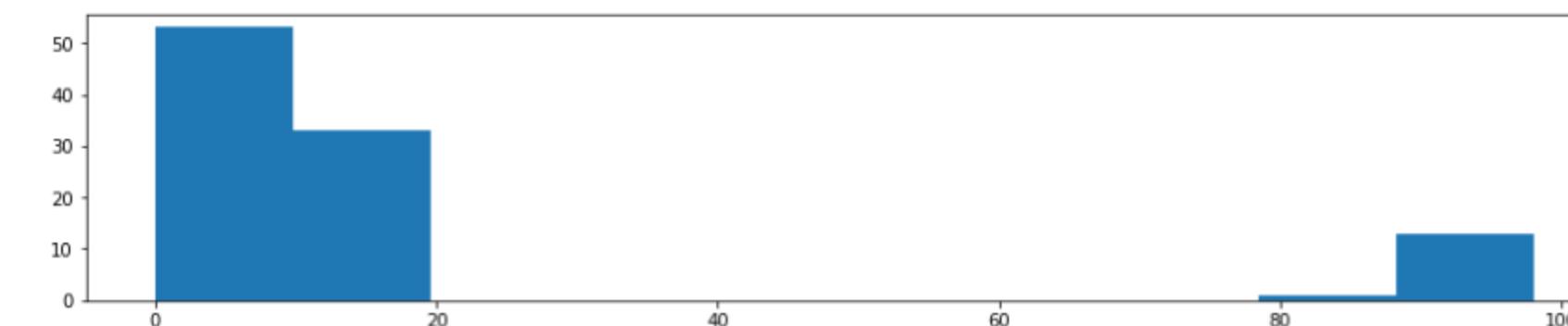


- Year 열
 - min 값 이상
-> 100년보다 작은 열 삭제

전체 데이터 분포

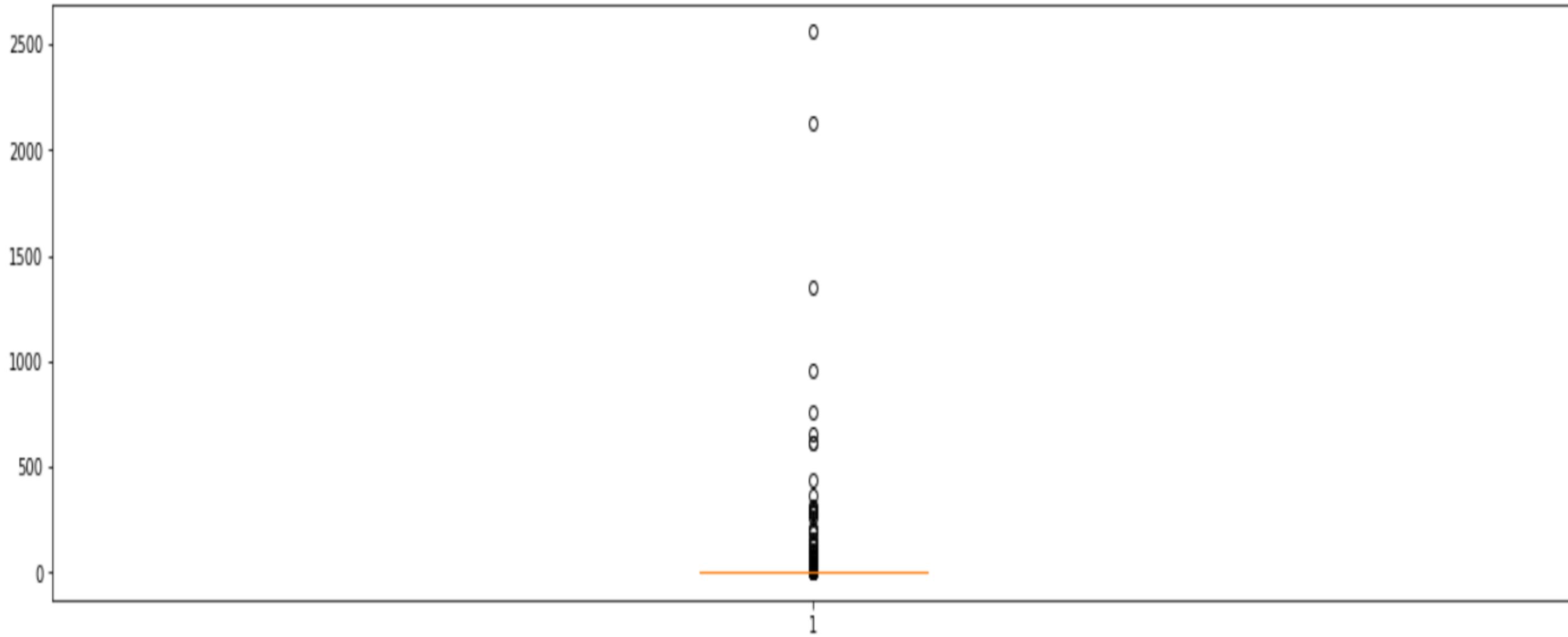


1000년 이상 데이터 분포



1000년 이하 데이터 분포

EDA - 'oo_Sales' column Analysis



- Na_Sales
 - 1000넘는 것 9개
 - 100넘는 것 41개
 - 10 넘는것 80개
- EU_Sales
 - 1000넘는 것 3개
 - 100넘는 것 23개
 - 10 넘는것 64개
- JP_Sales
 - 1000넘는 것 2개
 - 100넘는 것 16개
 - 10 넘는것 37개
- Other_Sales
 - 100넘는 것 9개
 - 10 넘는것 37개

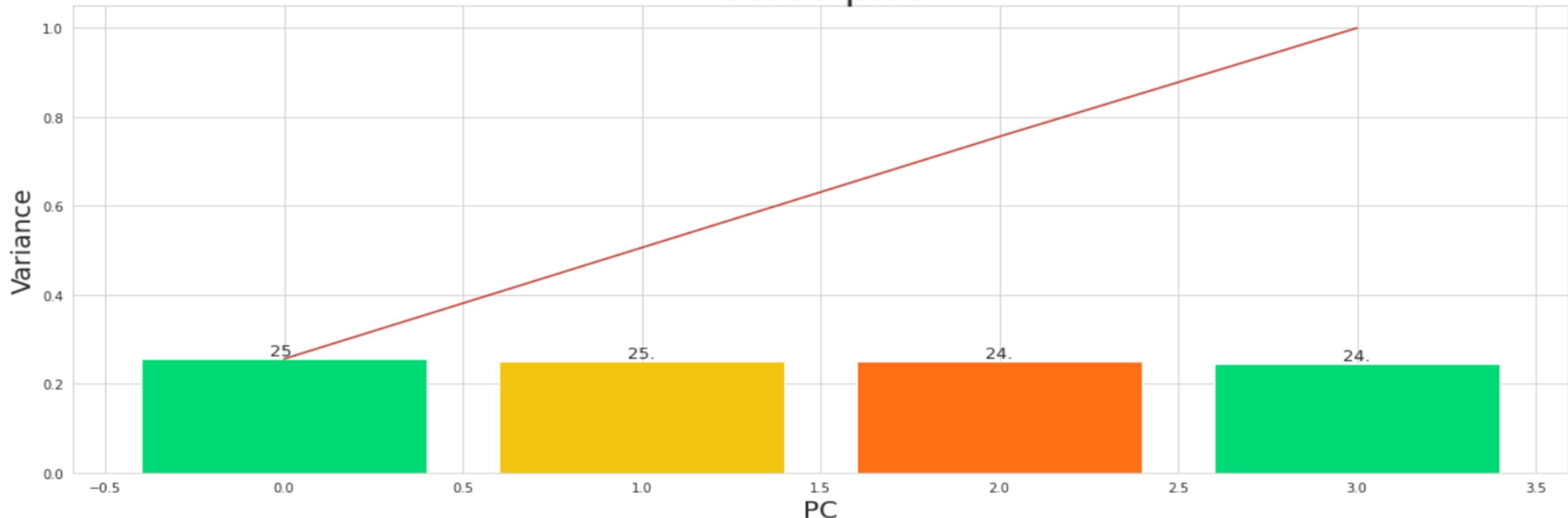
-> 출고량 열들은 모두 폭넓은 분포를 가지고 있다.

이상치가 있나 판단하려하였으나, 그저 다양한 분포를 가지는 출고량인 것으로 보여진다.

EDA - PCA를 통한 차원 축소

NA_Sales, EU_Sales, JP_Sales, Other_Sales
90% 내용을 나타내기 위해선, 주성분 4개가 모두 중요한 것으로 보여진다.
차원 축소할 필요성이 없다.
열을 그대로 유지할 것이다.

Scree plot

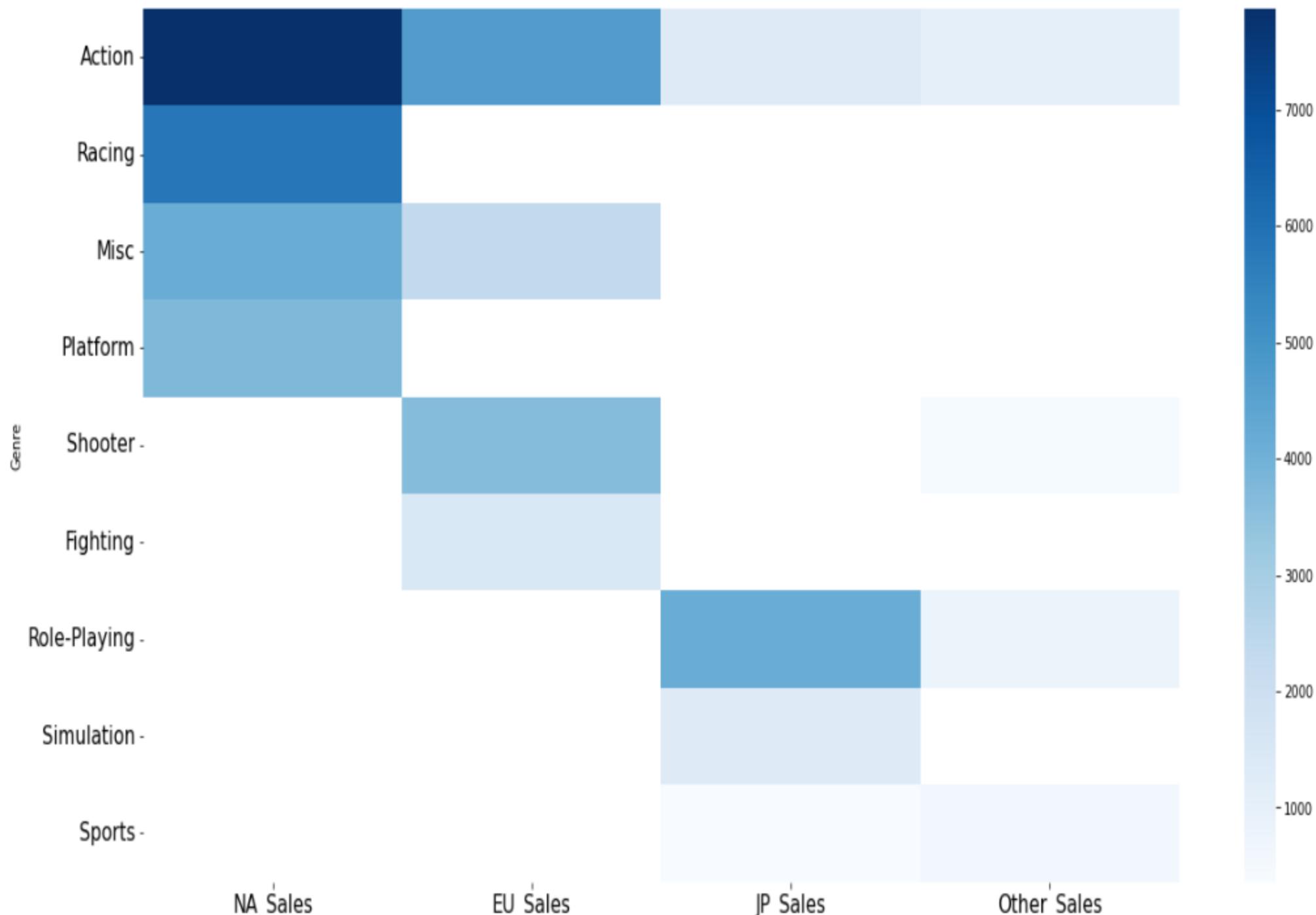


EDA Finish!

	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Sales
0	Candace Kane's Candy Factory	DS	2008	Action	Destineer	0.04	0.00	0.00	0.00	0.04
1	The Munchables	Wii	2009	Action	Namco Bandai Games	0.17	0.00	0.00	0.01	0.18
2	Otome wa Oanesama Boku ni Koi Shiteru Portable	PSP	2010	Adventure	Alchemist	0.00	0.00	0.02	0.00	0.02
3	Deal or No Deal: Special Edition	DS	2010	Misc	Zoo Games	0.04	0.00	0.00	0.00	0.04
4	Ben 10 Ultimate Alien: Cosmic Destruction	PS3	2010	Platform	D3Publisher	0.12	0.09	0.00	0.04	0.25
...
16593	Ice Age 2: The Meltdown	GC	2006	Platform	Vivendi Games	0.15	0.04	0.00	0.01	0.20
16594	Rainbow Islands: Revolution	PSP	2005	Action	Rising Star Games	0.01	0.00	0.00	0.00	0.01
16595	NBA 2K16	PS3	2015	Sports	Take-Two Interactive	0.44	0.19	0.03	0.13	0.79
16596	Toukiden: The Age of Demons	PSV	2013	Action	Tecmo Koei	0.05	0.05	0.25	0.03	0.38
16597	The King of Fighters '95	PS	1996	Fighting	Sony Computer Entertainment	0.00	0.00	0.16	0.01	0.17

16141 rows × 10 columns

Data Analysis 1 - 지역에 따라서 선호하는 게임 장르가 다를까



각 지역별 Top4 Genre

- 북미 : ['Action' 'Racing' 'Misc' 'Platform']
- 유럽 : ['Action' 'Shooter' 'Misc' 'Fighting']
- 일본 : ['Role-Playing' 'Action' 'Simulation' 'Sports']
- 기타 : ['Action' 'Role-Playing' 'Sports' 'Shooter']

대체적으로 Action게임이 선호도가 제일 크다.

하지만 일본만 유일하게,
캐릭터와 스토리를 즐기는 문화적 특성에 의해
Role-Playing의 선호도가 가장 크다.

Data Analysis 1 - 지역에 따라서 선호하는 게임 장르가 다를까

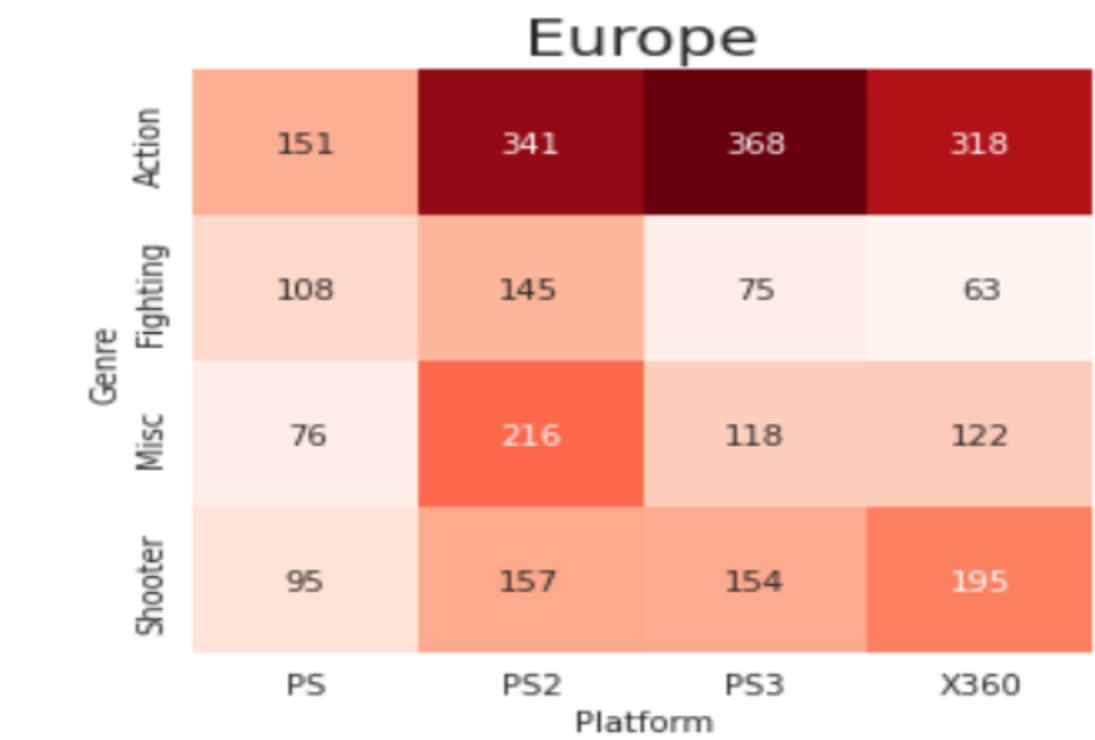
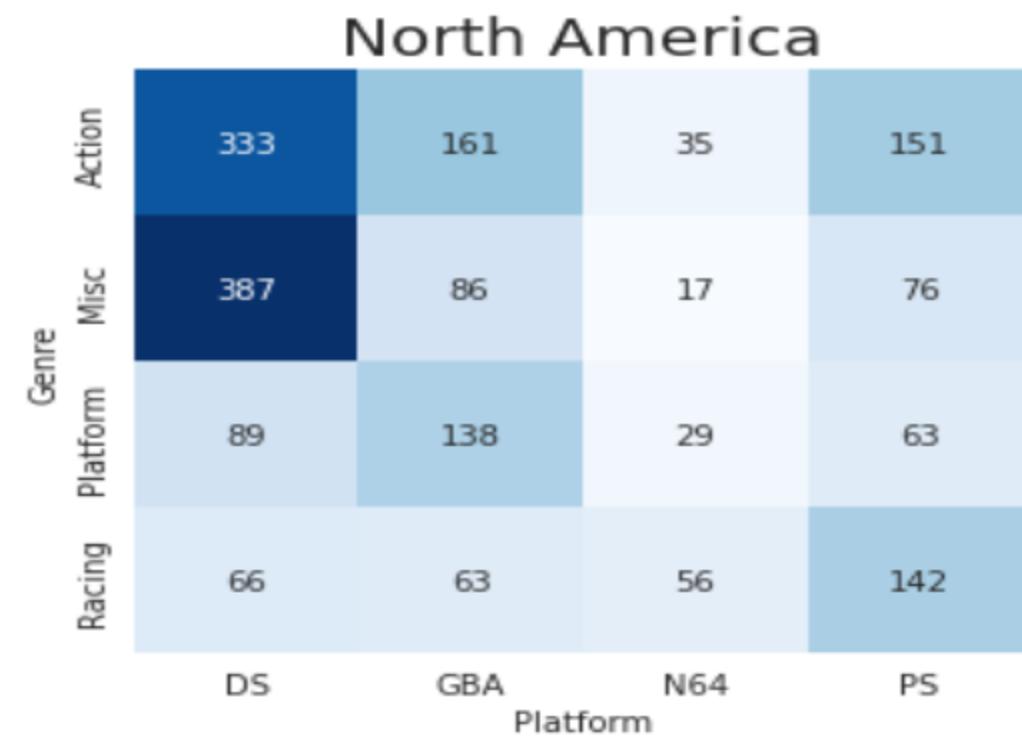
-> 방금 구한 지역별 게임 장르와 지역별 플랫폼 관의 관계성이 있을까?

그렇다면, 각 지역별로 선호하는 platform의 종류를 보자.

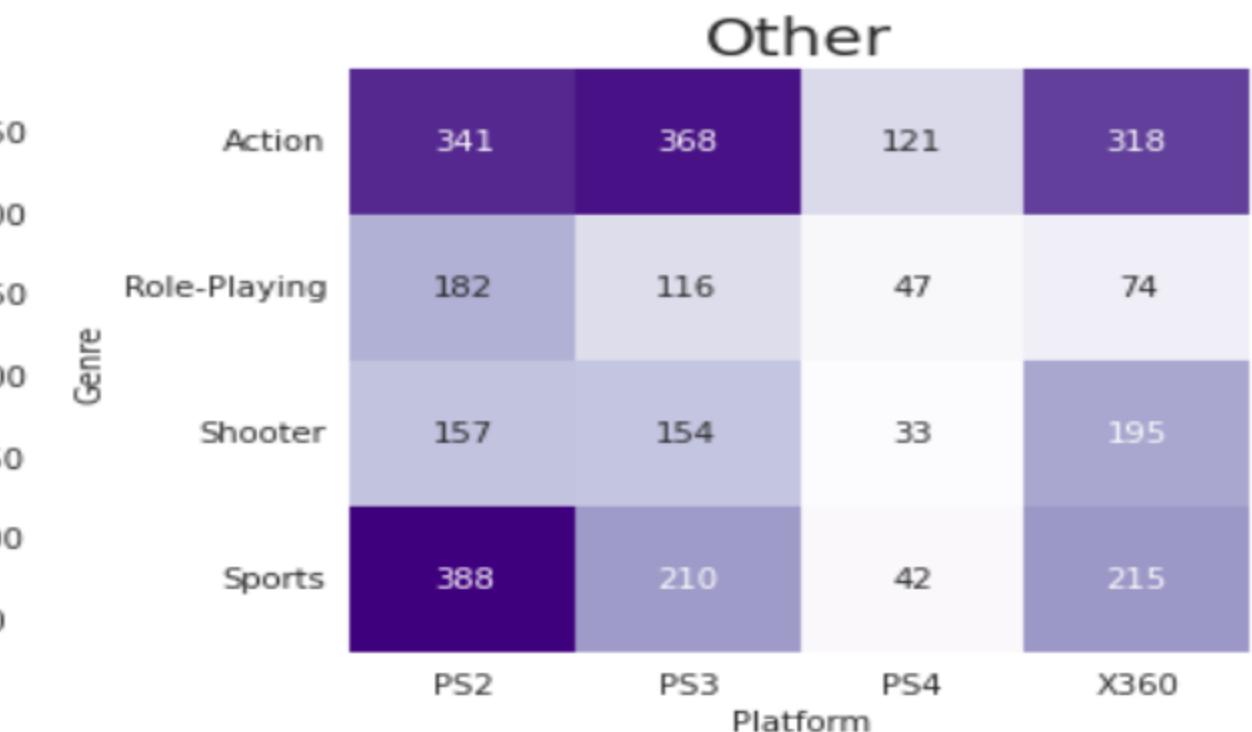
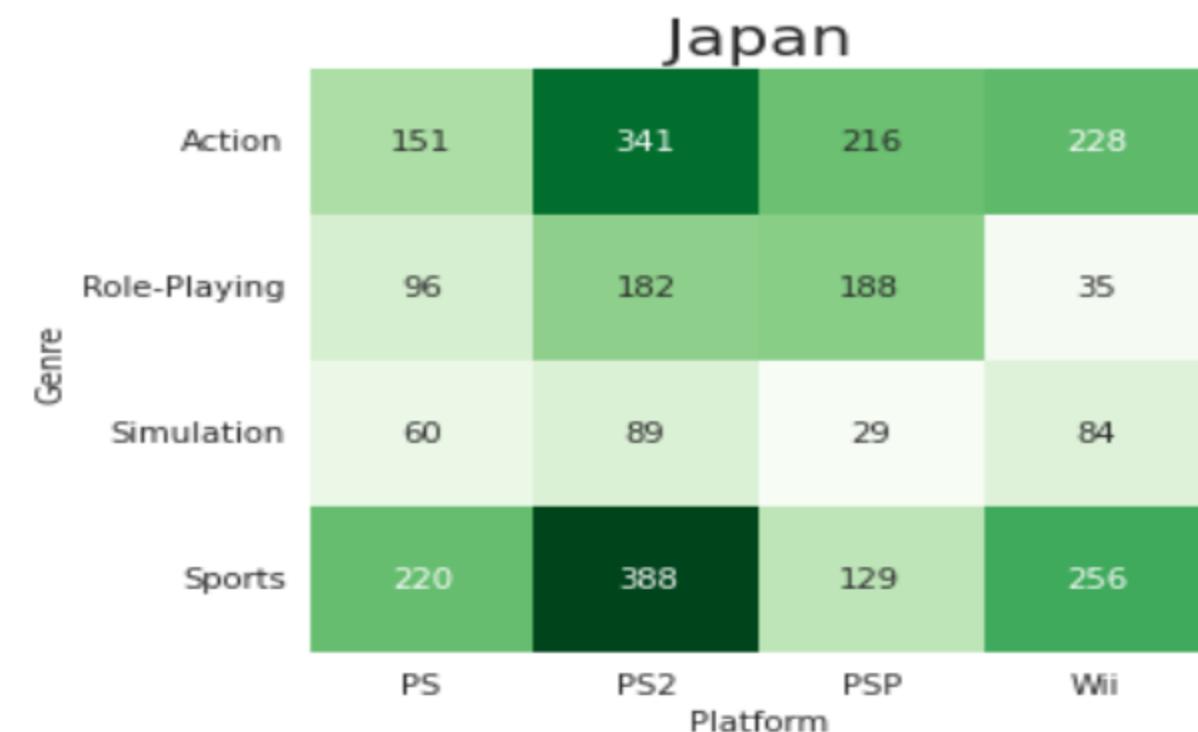
각 지역별 Top 4 Platform

- 북미 : ['PS' 'GBA' 'N64' 'DS']
- 유럽 : ['PS3' 'X360' 'PS2' 'PS']
- 일본 : ['PS' 'PS2' 'PSP' 'Wii']
- 기타 : ['X360' 'PS2' 'PS3' 'PS4']

Data Analysis 1 - 지역에 따라서 선호하는 게임 장르가 다를까



-> 방금 구한 지역별 게임 장르와
지역별 플랫폼 간 관계성이 있을까?
딱히 상관관계가 보이지 않음.



Data Analysis 1 - 지역에 따라서 선호하는 게임 장르가 다를까

-> 방금 구한 지역별 게임 장르와 지역별 플랫폼 관의 관계성이 있을까?

- Genre

- 북미지역 : ['Action' 'Racing' 'Misc' 'Platform']
- 유럽지역 : ['Action' 'Shooter' 'Misc' 'Fighting']
- 일본지역 : ['Role-Playing' 'Action' 'Simulation' 'Sports']
- 기타지역 : ['Action' 'Role-Playing' 'Sports' 'Shooter']

- Platform

- 북미지역 : ['PS' 'GBA' 'N64' 'DS']
- 유럽지역 : ['PS3' 'X360' 'PS2' 'PS']
- 일본지역 : ['PS' 'PS2' 'PSP' 'Wii']
- 기타지역 : ['X360' 'PS2' 'PS3' 'PS4']

PS 플랫폼들이 가장 많아보인다.

- PS 플랫폼이 *Action* 장르만 주로 하는 것은 아니었다. *Action*, *Sports*가 비슷하게 많으며, 다양한 종류를 모두 다루고 있는 플랫폼이었다.

플랫폼, 장르는 상관관계가 크지 않아 보인다.

각 플랫폼마다 특정한 장르를 많이하기보다는, 각 플랫폼마다 다양한 장르를 가지고 있다.

Data Analysis 1 결론

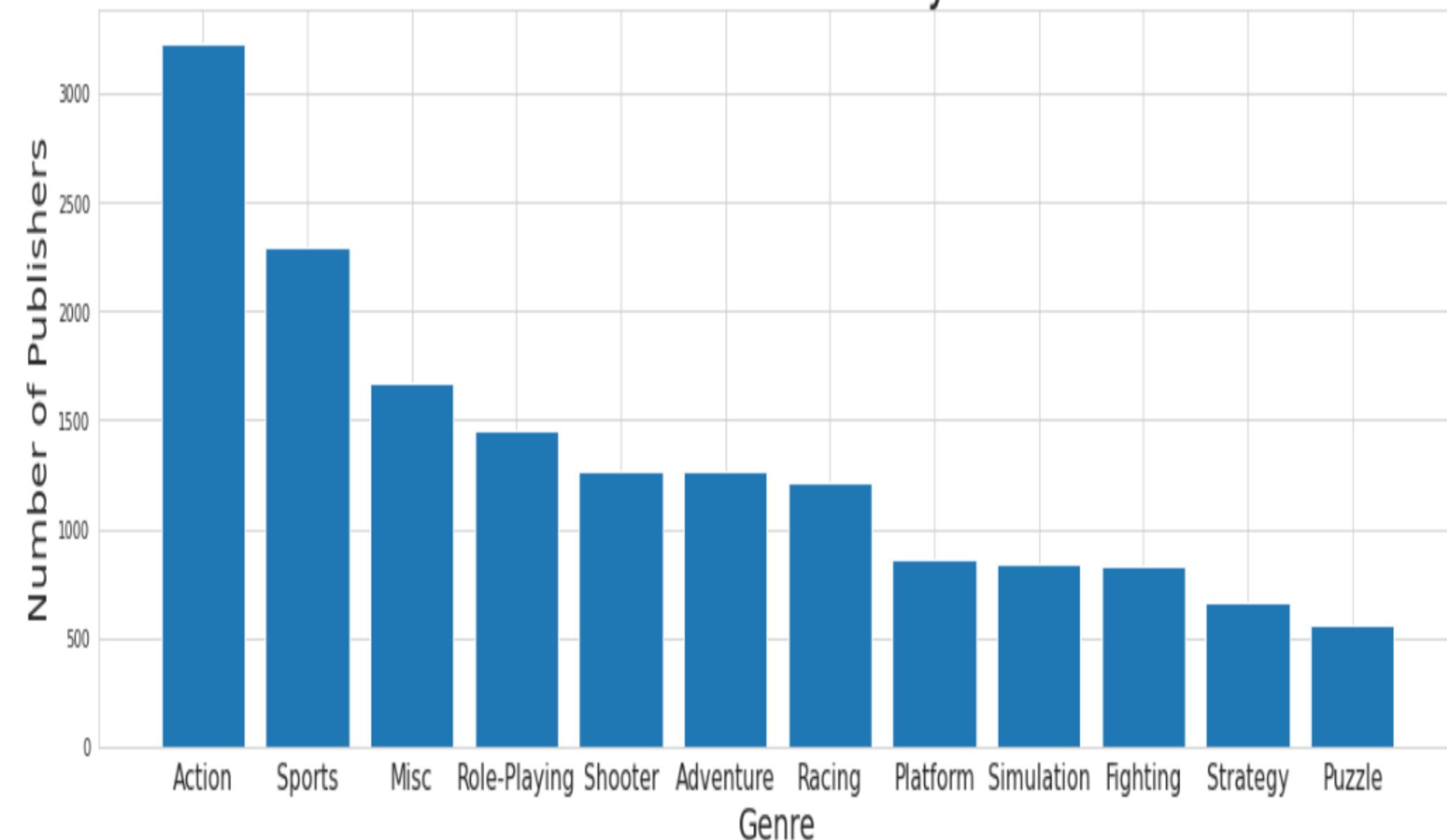
결론적으로, 지역별로 선호하는 게임의 장르는 다르지만, Action 게임의 선호도가 가장 높은 것으로 판단되었다.

그렇다면, 게임 장르 별 출판사 수는 어떠할까?

결론적으로, 여기서도, Action 게임이 출판사가 가장 많은 것으로 판단되었다.

-> Action 장르의 게임은 지역별 선호도가 가장 높고, 출판사 수도 가장 많다.

Number of Publishers by Genre



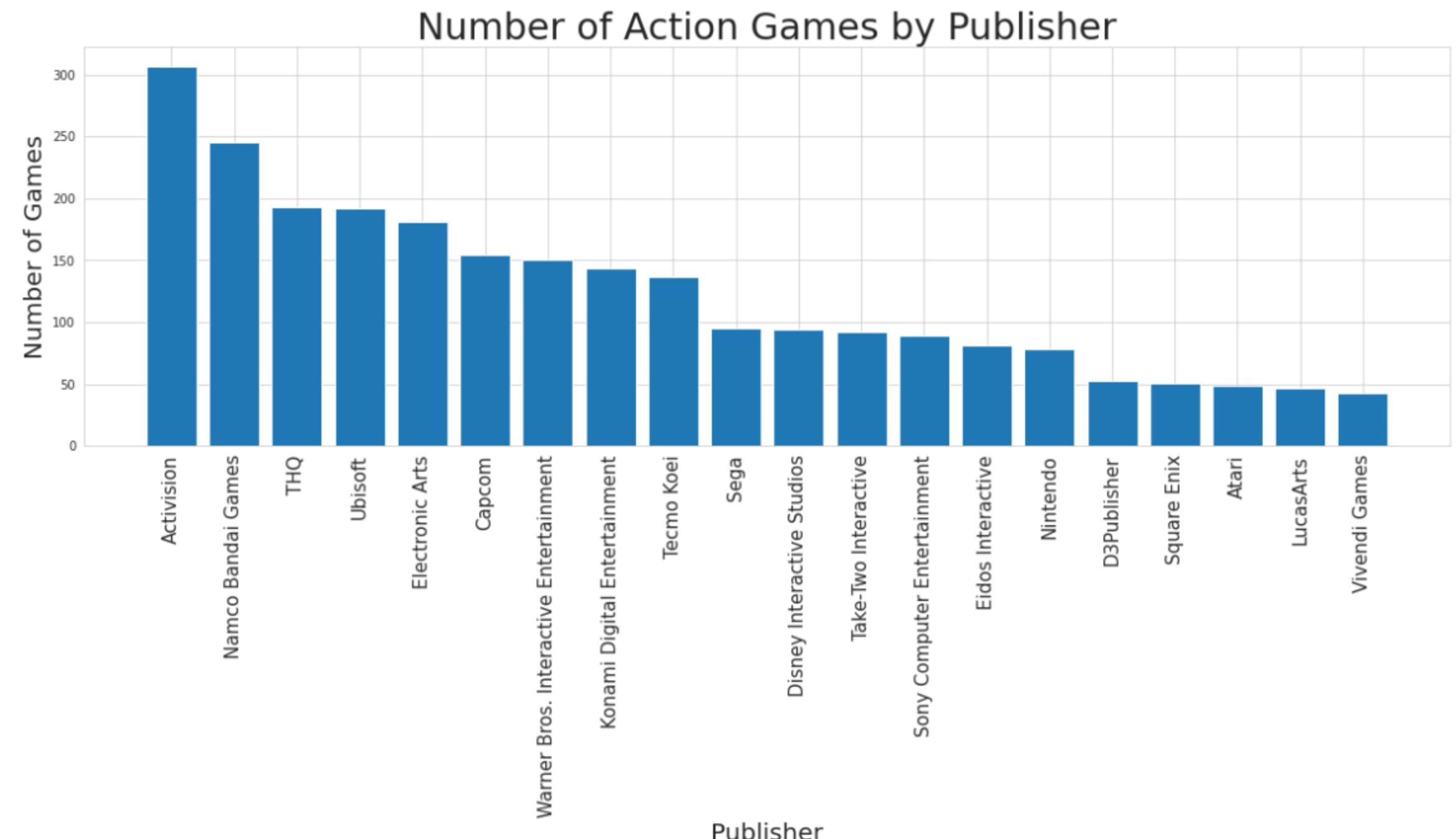
Data Analysis - Action Game

그렇다면, Action 장르가 다음 기수 게임의 장르가 될 가능성이 늘고 있는데,

Action 장르에 대해 추가적으로 분석해보자.

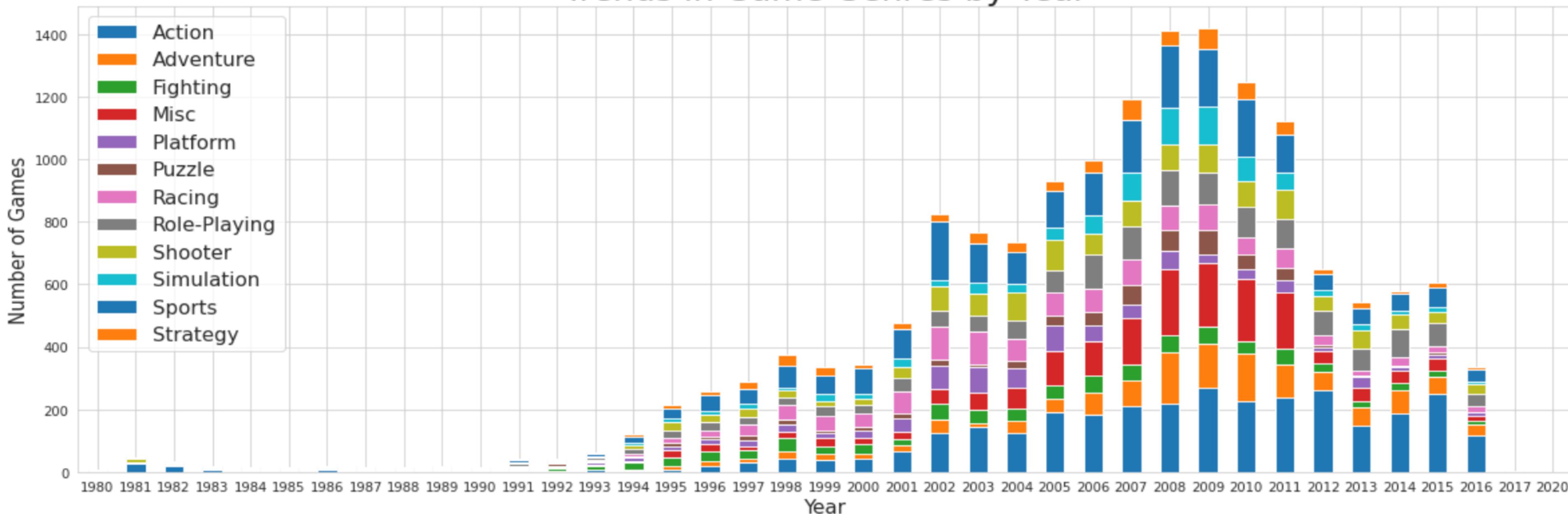
- Action 장르에서 어떤 출판사가 가장 많이 쓰였나?

'Activision', 'Namco Bandai Games', 'THQ', 'Ubisoft'가 주로 많이 쓰였다.



Data Analysis 2 - 연도별 게임의 트렌드

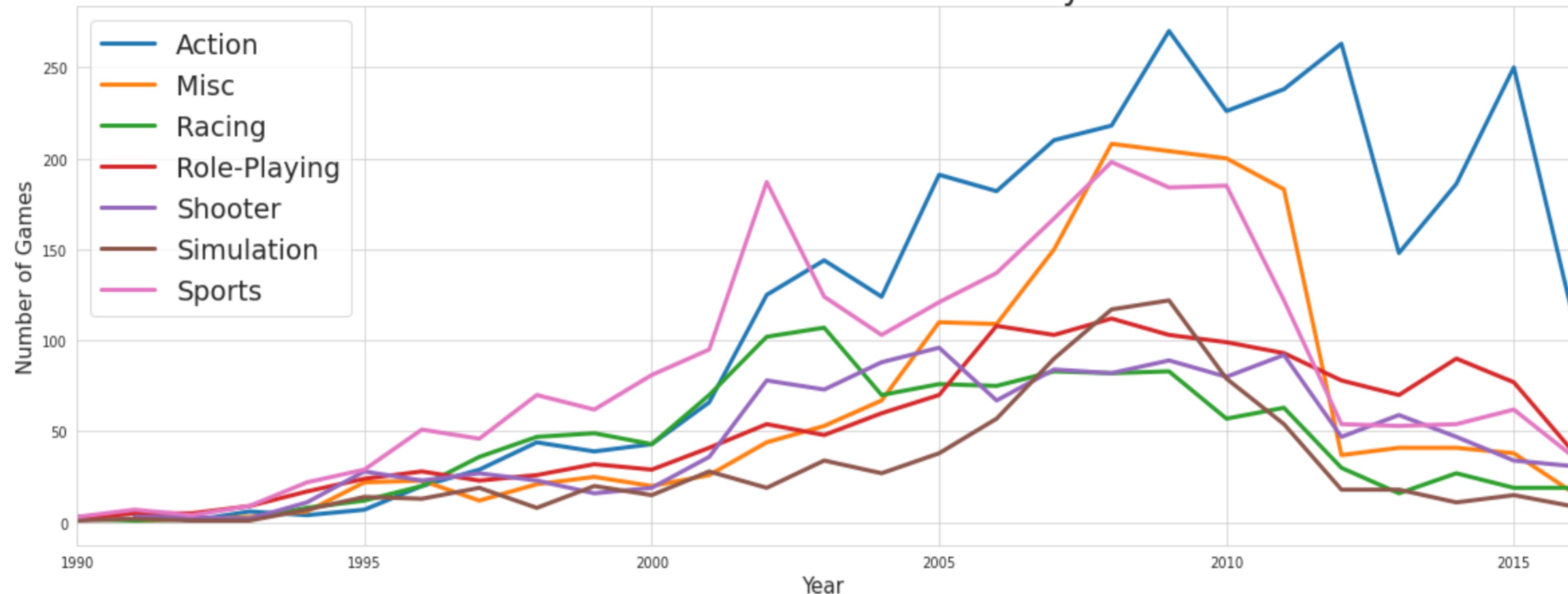
Trends in Game Genres by Year



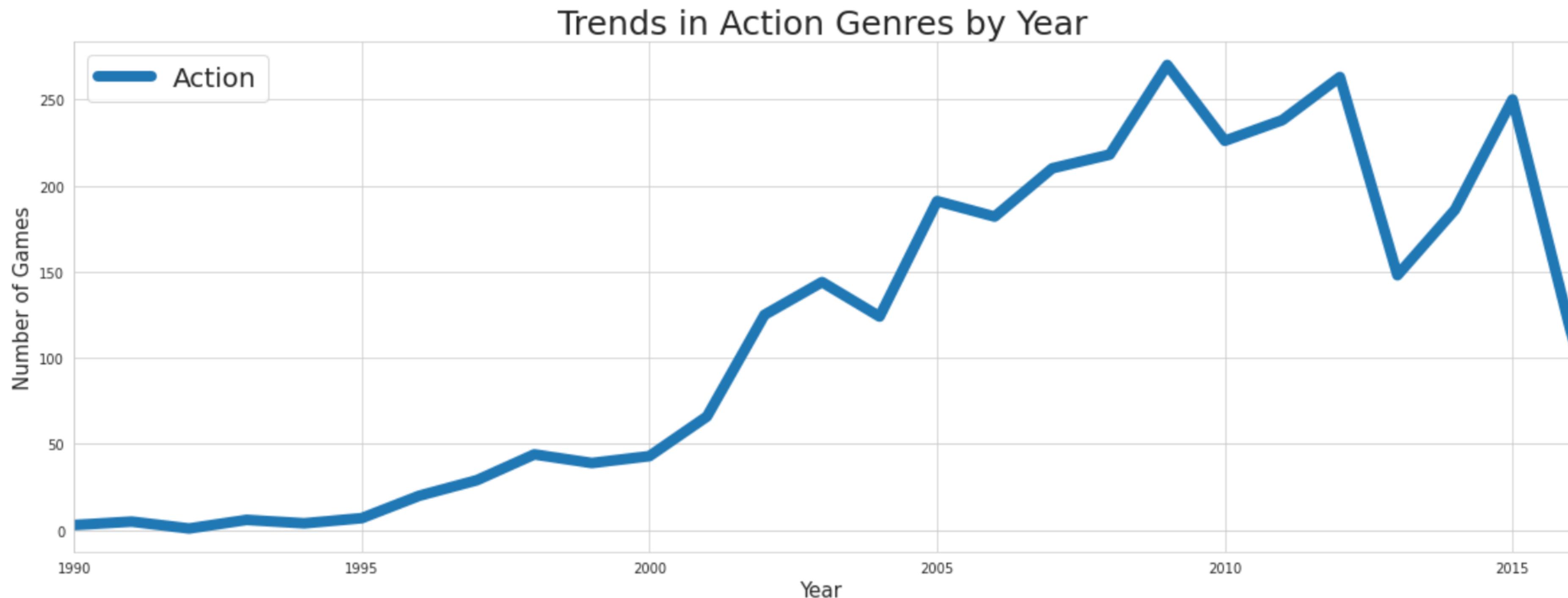
2008, 2009년에 게임 트렌드가 제일 활발함.
전체적으로 action 장르의 게임은 꾸준히 많이 나옴.

Data Analysis 2 - 연도별 게임의 트렌드

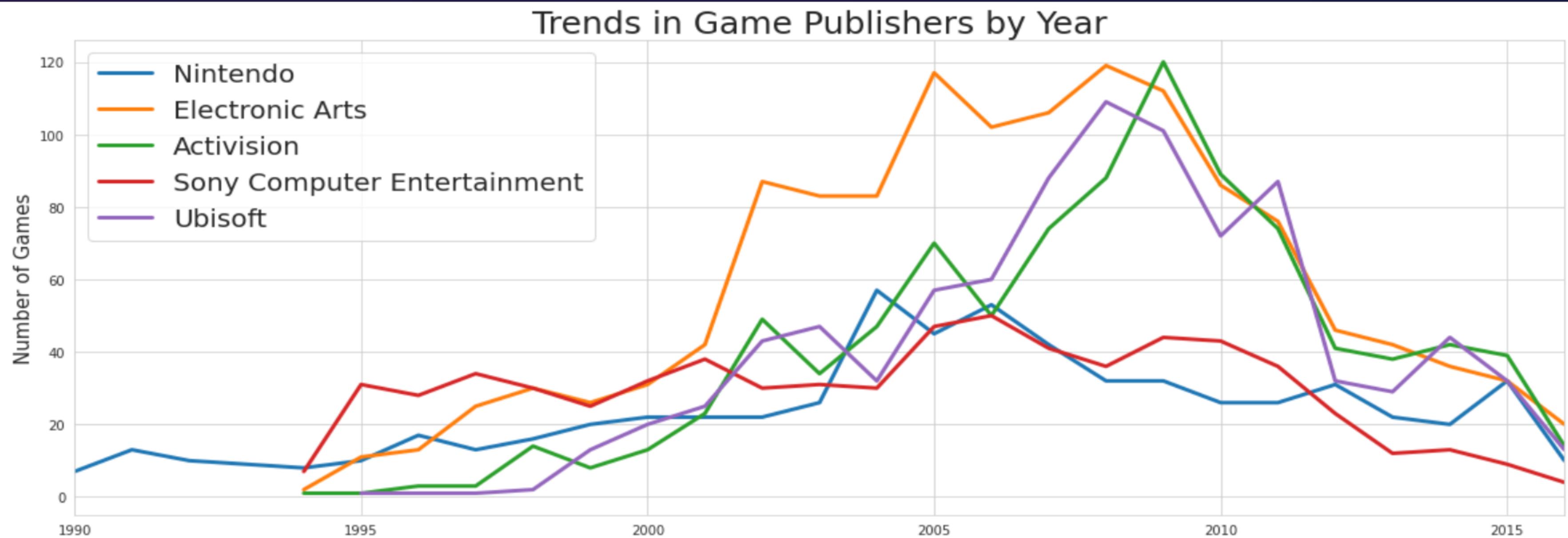
Trends in Selected Game Genres by Year



Data Analysis 2 - 연도별 게임의 트렌드



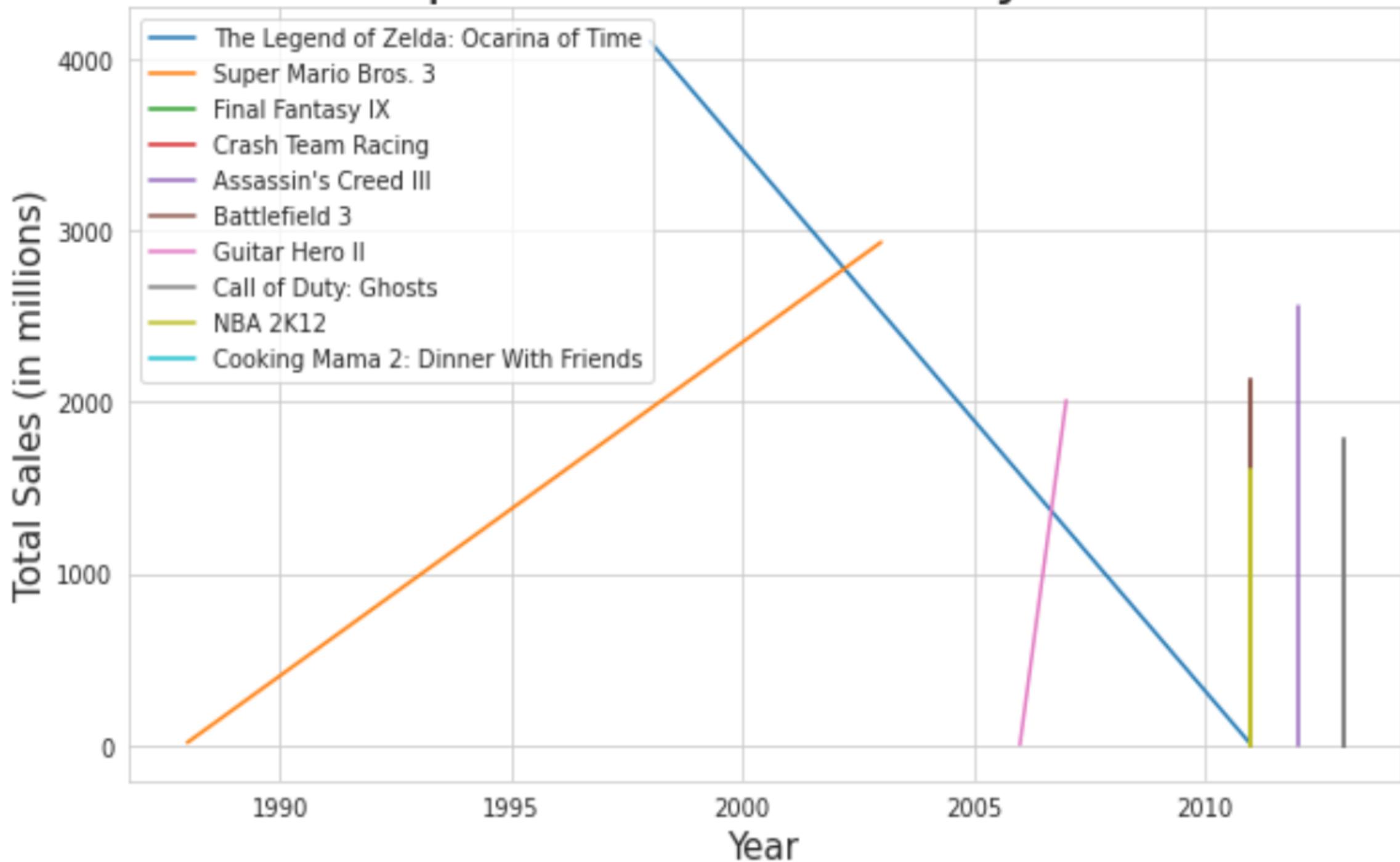
Data Analysis 2 - 연도별 게임의 트렌드



연도별 게임 장르, 출판사 확인 결과-> 연도별로 게임의 트렌드가 바뀐다.

Data Analysis 3 - 인기가 많은 게임에 대한 분석 및 시각화

Top 10 Games Sales by Year



'인기가 많다' 의미는?

Name은 고유한 이름마저 11238개로 너
무 많아 인기 식별 불가능

Sales가 가장 많이 판매된 것이 인기
가 많겠지!

상위 10개의 게임 중, 연도에 따른 게임 순위를 알아보았다.

연도에 따라 증가, 감소하는 것이 확연히 보여진다.

추가적으로 연도축에 수직으로 되어있는 게임들은 판매량의 차이가 있
어도 연도가 추가적으로 다른 데이터가 없는 것으로 보여진다.

가설 검정 - 평균 인기 지속 기간

```
top_10000_games = new_df.sort_values(by='Sales', ascending=False).head(10000)

# 중복되는 게임 : 4814rows
duplicate_games = top_10000_games[top_10000_games.duplicated(subset=['Name'], keep=False)]
duplicate_games.shape

(4814, 10)

duplicate_games.query('Name == "The Legend of Zelda: Ocarina of Time"')

      Name Platform Year Genre Publisher NA_Sales EU_Sales JP_Sales Other_Sales Sales
1964 The Legend of Zelda: Ocarina of Time     N64 1998 Action   Nintendo    4100.00     1.89     1.45     0.16  4103.50
3148 The Legend of Zelda: Ocarina of Time     3DS 2011 Action   Nintendo     2.03     1.27     0.62     0.30     4.22

def get_game_years(duplicate_games):
    game_years = pd.DataFrame(columns=['Game', 'Year1', 'Year2', 'Year_Diff'])

    for game_name in duplicate_games.Name.unique():
        game = duplicate_games[duplicate_games.Name == game_name]
        year1 = game.Year.min()
        year2 = game.Year.max()
        year_diff = year2 - year1
        game_years = game_years.append({'Game': game_name, 'Year1': year1, 'Year2': year2, 'Year_Diff': year_diff}, ignore_index=True)

    return game_years

result = get_game_years(duplicate_games)
```

	Game	Year1	Year2	Year_Diff
0	The Legend of Zelda: Ocarina of Time	1998	2011	13
1	Super Mario Bros. 3	1988	2003	15
2	Assassin's Creed III	2012	2012	0
3	Battlefield 3	2011	2011	0
4	Guitar Hero II	2006	2007	1

가설 검정 - 평균 인기 지속 기간

```
# 무작위로 20개의 게임 추출
sample_games = result.sample(n=20, random_state=42, replace=True)

# 추출한 게임들의 평균 인기 지속 기간 계산
sample_mean = np.mean(sample_games.Year_Diff)

# 이전에 출시된 게임들의 평균 인기 지속 기간 계산
pop_mean = np.mean(result.Year_Diff)

sample_games['Year_Diff'] = sample_games['Year_Diff'].astype(int)

# 가설 검정 수행
t_stat, p_val = stats.ttest_1samp(sample_games.Year_Diff, pop_mean)
```

평균 인기 지속 기간에 대한 가설 검정

- H0 (귀무가설): 다음 분기의 게임들의 평균 인기 지속 기간은 현재까지 출시된 게임들의 평균 인기 지속 기간과 차이가 없다.
- H1 (대립가설): 다음 분기의 게임들의 평균 인기 지속 기간은 현재까지 출시된 게임들의 평균 인기 지속 기간과 차이가 있다.

Sample Mean: 1.50
Population Mean: 0.80
t-statistic: 0.78
p-value: 0.44629

즉, 이전에 출시된 게임들과 다음 분기에 출시될 게임들의 평균 인기 지속 기간 간에는 차이가 없었을 것으로 예측할 수 있습니다.

그렇다면, 다음 분기 게임의 지속성을 예측할 때
지속 기간이 어느 정도인 것으로 잡아야 할까요?

가설 검정 - 평균 인기 지속 기간

```
# 무작위로 20개의 게임 추출  
sample_games = result.sample(n=20, random_state=42, replace=True)  
  
# 추출한 게임들의 평균 인기 지속 기간 계산  
sample_mean = np.mean(sample_games.Year_Diff)  
  
# 이전에 출시된 게임들의 평균 인기 지속 기간 계산  
pop_mean = 2  
  
# 가설 검정 수행  
z_stat = (sample_mean - pop_mean) / (np.std(result.Year_Diff) / np.sqrt(len(sample_games)))  
p_val = norm.sf(abs(z_stat))
```

- H0 (귀무 가설): 다음 분기에 출시될 게임들의 평균 인기 지속 기간은 2 이하이다.
- H1 (대립 가설): 다음 분기에 출시될 게임들의 평균 인기 지속 기간은 2보다 크다.

Sample Mean: 1.50
Population Mean: 2.00
z-statistic: -0.85
p-value: 0.19719

다음 분기에 출시될 게임들의 평균 지속 기간은 2보다 작다.

가설 검정 - 평균 인기 지속 기간

Sample Mean: 1.50
Population Mean: 1.00
z-statistic: 0.85
p-value: 0.19719

다음 분기에 출시될 게임들의 평균 지속 기간
은 1보다 작다.

Sample Mean: 1.50
Population Mean: 0.50
z-statistic: 1.70
p-value: 0.04425

다음 분기에 출시될 게임들의 평균 지속 기간
은 0.5보다 크다.

Sample Mean: 1.50
Population Mean: 0.60
z-statistic: 1.53
p-value: 0.06263

다음 분기에 출시될 게임들의 평균 지속 기간
은 0.6보다 작다.

-> 평균 인기 지속 기간은 약 반년 정도로 보임.

결론 도출

- **Data Analysis**

- Action 장르의 게임은 지역별 선호도가 가장 높고, 출판사 수도 가장 많다.
- 트렌드 면에서도, action 장르가 가장 인기있음.
- Action 장르에서 출판사 'Activision'이 가장 많았음.
- PS 시리즈 플랫폼이 가장 빈번함.

- **Hypothesis Test**

- 평균인기 지속 기간 반년인 것

=> Action 장르, year_diff=반년정도되는것으로 다음 분기 게임 예측

- PS platform : "PlayStation"의 약어로, 소니(Sony)가 제작한 게임 콘솔 시리즈 중 하나
1994년 출시된 PS1 이후, PS2는 2000년에, PS3는 2006년에, PS4는 2013년에 출시되었으며, 가장 최근에는 2020년에 PS5가 출시
따라서, PS 시리즈가 높아질수록 최신 모델이고, 그에 따라 하드웨어 성능 및 게임 콘텐츠도 발전해 왔다.

결론 도출

비교적 최근에 나온 것 뽑기

```
result2 = result1[result1['Year'] == result1.Year.max()]
result2
```

	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Sales	Year1	Year2	Year_Diff
3801	Teenage Mutant Ninja Turtles: Mutants in Manha...	PS4	2016	Action	Activision	0.04	0.02	0.0	0.01	0.07	2016	2016	0
3802	Teenage Mutant Ninja Turtles: Mutants in Manha...	PS3	2016	Action	Activision	0.01	0.03	0.0	0.01	0.05	2016	2016	0
14997		Ghostbusters (2016)	PS4	Action	Activision	0.02	0.00	0.0	0.01	0.03	2016	2016	0

출고량이 가장 많은 것 뽑기 -> 다음 기수 게임 예측 완료

```
result2[result2['Sales'] == result2.Sales.max()]
```

	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Sales	Year1	Year2	Year_Diff
3801	Teenage Mutant Ninja Turtles: Mutants in Manha...	PS4	2016	Action	Activision	0.04	0.02	0.0	0.01	0.07	2016	2016	0