경상국립대학교
Gyeongsang National University

게임 데이터 분석 및 예측

데이터수집과관리 프로젝트 결과 발표

Steam 플랫폼 데이터 기반
게임 시장 수요 분석

2024. 06. 13.

프리조 | 정수용, 윤종인, 이준학

게임 데이터
분석 및 예측

목차

경상국립대학교
Gyeongsang National University

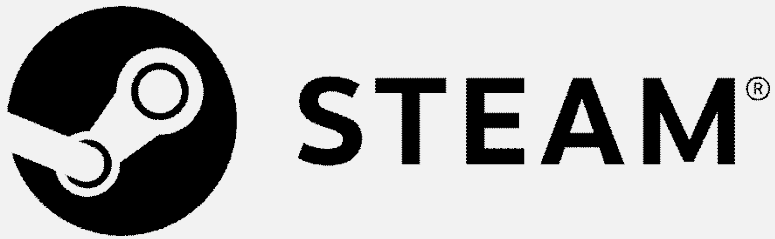
경상국립대학교
Gyeongsang National University

게임 데이터 분석 및 예측

프로젝트 개요

게임 데이터 분석 및 예측 | 목적

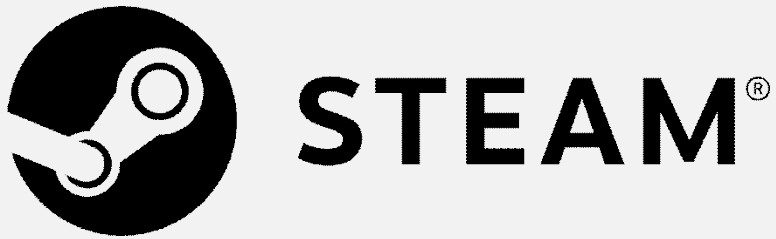
경상국립대학교 Gyeongsang National University



- Steam 플랫폼에서 제공하는 게임 관련 데이터 분석을 통해 게임을 추천하는 모델을 개발한다.
- 게임 개발사 및 배급사가 시장 수요를 정확히 파악하고 소비자들에게 개인 취향에 맞는 게임을 효율적으로 찾을 수 있도록 돕는다.

게임 데이터 분석 및 예측 | 배경

경상국립대학교 Gyeongsang National University



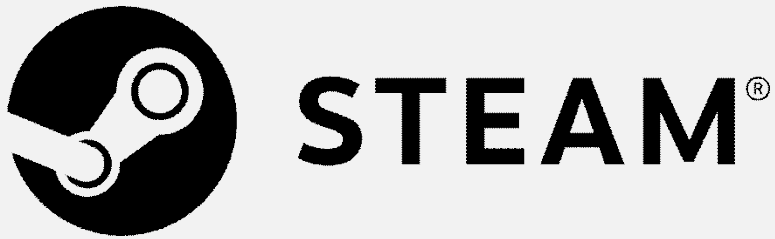
1. Steam
   - 세계 최대 규모의 전자 소프트웨어 유통망
   - 다양한 장르의 게임과 방대한 사용자 데이터를 보유

2. Steam Games Dataset
   - Kaggle에 있는 Steam API를 통해 수집된 데이터 셋
   - 게임 데이터 85,000개
   - 판매량, 사용자 리뷰, 플레이 타임, 장르 등

게임 데이터 분석 및 예측

연구 및 프로젝트의 필요성

경상국립대학교 Gyeongsang National University



- 사용자 리뷰 뿐만 아니라 게임 장르, 지원 언어, 긍정적 및 부정적 평가 등을 포함한 다양한 데이터를 활용하여 게임의 판매량을
  예측하고 장르별 게임을 추천하는 모델을 개발하여 게임의 인기도를 더 정확하게 예측하고 사용자에게 개인화된 게임 추천을 제공하여 사용자와 개발자에게 유용한 인사이트를 제공할 것이다.

게임 데이터
분석 및 예측

팀원 역할 및 진행 스케줄

경상국립대학교 Gyeongsang National University

주요 내용 \ 일정(주차)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
주제 선정	■	■	■	■	■										
데이터 수집				■	■										
데이터 전처리						■	■	■	■	■					
데이터 모델링					■	■	■	■			■	■	■		
데이터 시각화										■	■			■	

정수용(팀장): 프로젝트 총괄, 데이터 전처리
윤종인(팀원): 데이터 수집, 데이터 모델링
이준학(팀원): 데이터 시각화

경상국립대학교
Gyeongsang National University

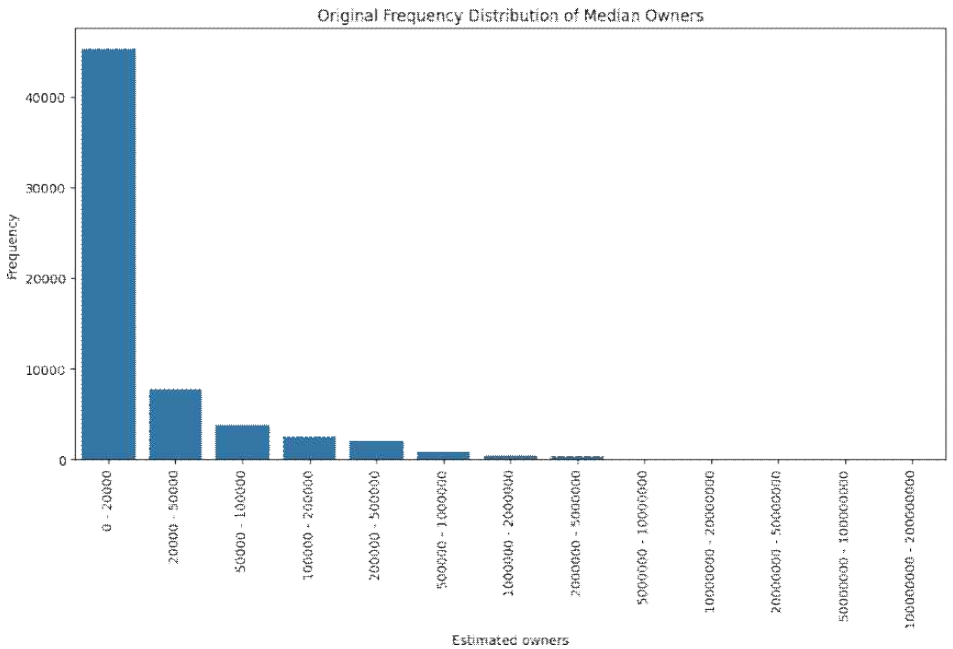
게임 데이터
분석 및 예측

분석

사용이 어려운 Feature 제거

필요 없음	정형화가 어려움	결측치가 매우 많음
Header Image *Website* *Support url* *Support Email* *Metacritic url* *Screenshots* *Movie*	*About the game* *Reviews* *Notes*	*User score (25)* *Score rank (20)*

게임 데이터 분석 및 예측 | 분석

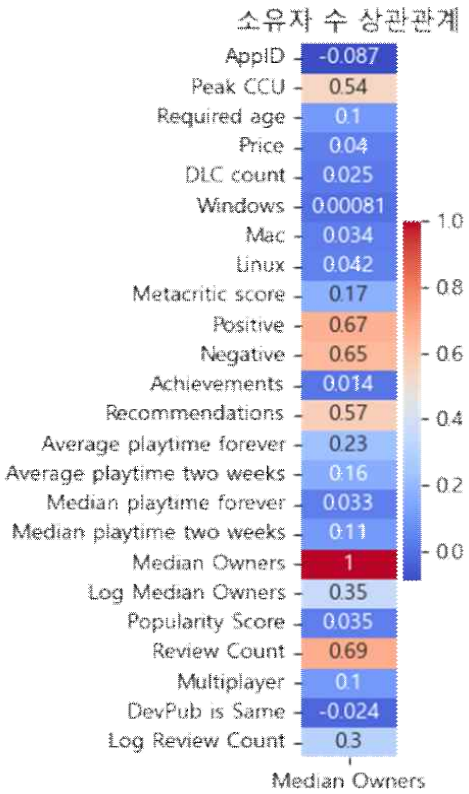


1. 게임 소유자 수 비율
   - 범주형 데이터
   - 대부분 0 – 20000
   - 데이터 분포가 급격히 감소
   - 매우 불균형한 데이터

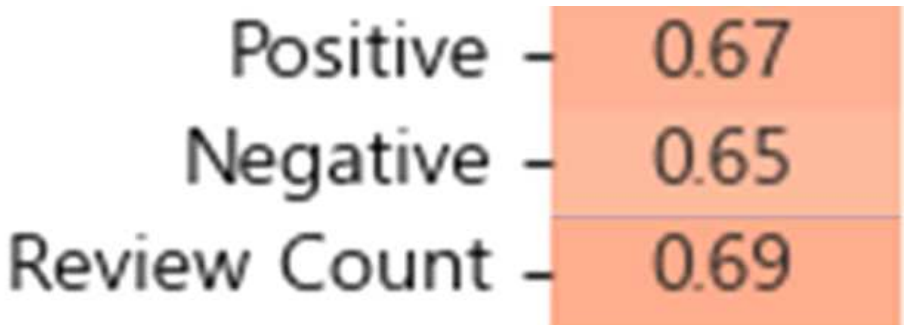
   - 몇몇 게임이 매우 높은 소유자를 가지는 반면, 대다수는 상대적으로 소유자가 적다.

   - 판매량의 지표가 되는 Owners를 대신할 다른 Feature가 있을 것이다.

게임 데이터 분석 및 예측 | 분석



2. 리뷰 수와 소유자 수의 상관관계
   - 대부분의 범주에 속하는 사례가 매우 적어 분석의 신뢰성을 저하시킬 수 있다.
     - 이를 해결하기 위해 판매 수와 상관관계가
       높은 리뷰 수를 이용

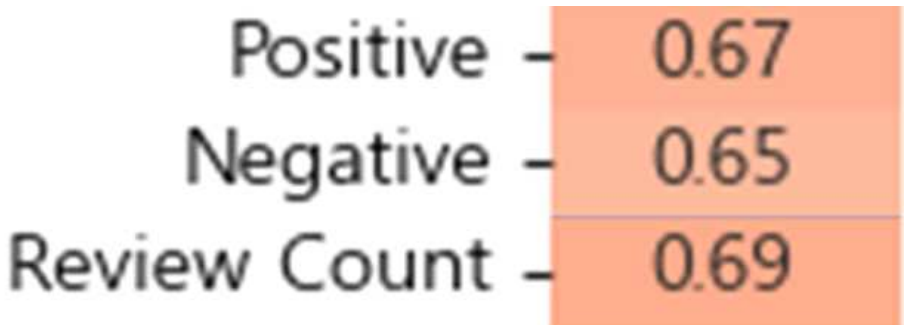


2. 리뷰 수와 소유자 수의 상관관계

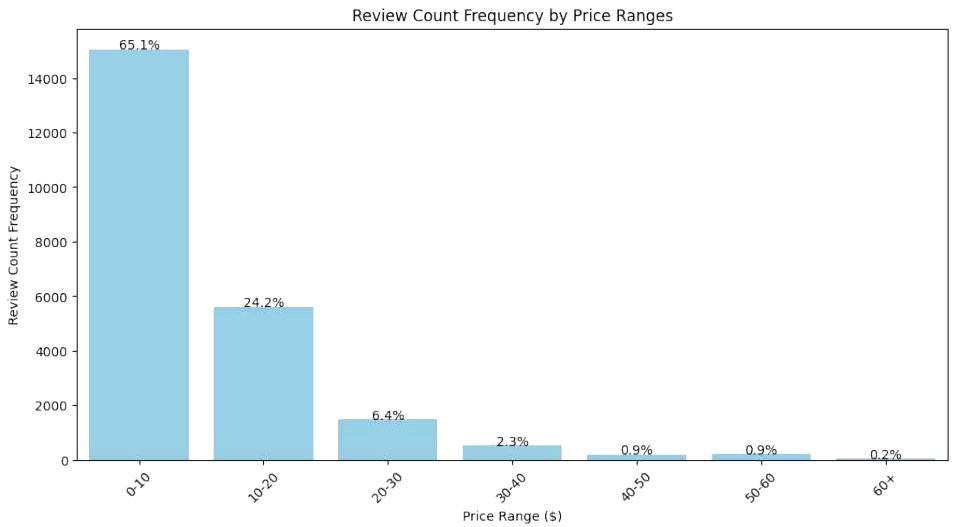
- Positive + Negative로 전체 평가 수(Reivew Count) 열 추가
  - Positive: 긍정 평가 수
  - Negative: 부정 평가 수
- 소유자 수 상관관계 분석 결과
  - Postive: 0.67
  - Negative: 0.65
  - Review Count: 0.69
- Review Count의 상관 관계수가 약 0.7로 높은 상관 관계

2.1. 판매량

- 리뷰 수를 판매량의 지표로 사용하기로 하였다.



게임 데이터 분석 및 예측 | 분석

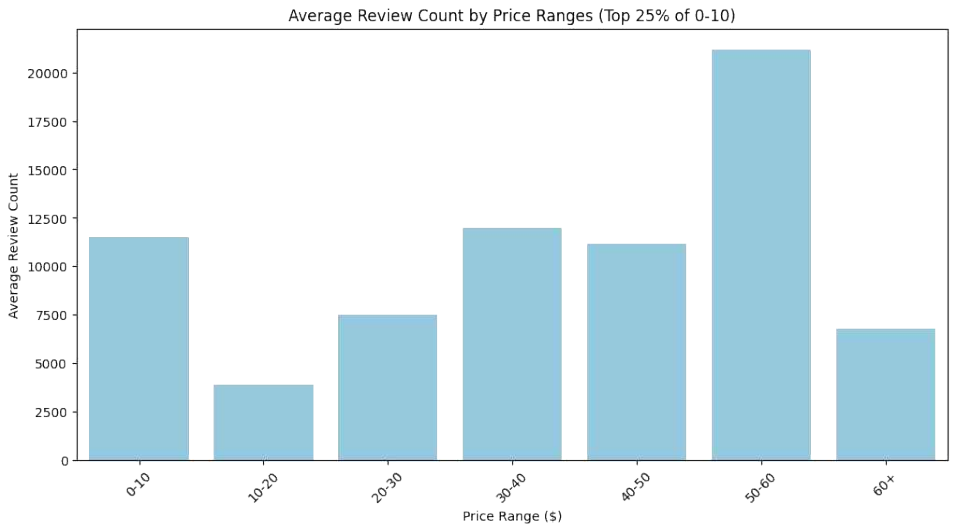


3.1. 가격 그룹 별 빈도수

- 가격 그룹별 판매량을 분석하기 위해 가격 그룹별 빈도수를 확인했다.

- 가격과 판매량을 알아보기 전에 먼저 가격 그룹별 빈도수를 확인했더니 0~10달러인 그룹이
지배적이다.

게임 데이터 분석 및 예측 | 분석

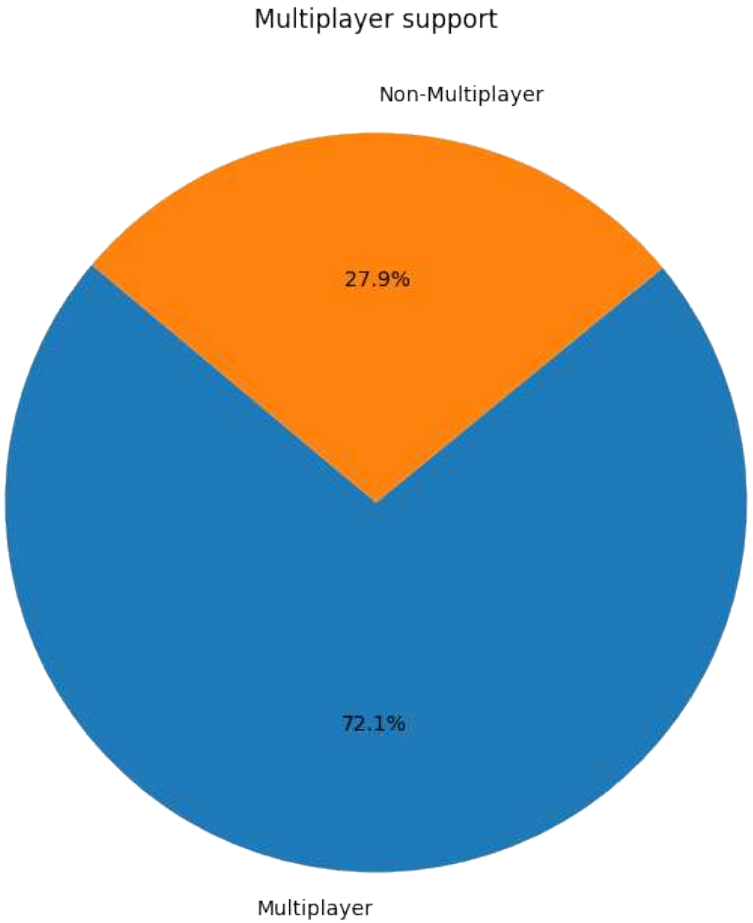


3.2. 가격 그룹 별 판매량

- 0~10달러인 그룹의 상위 25%만 반영하고 가격 그룹별 판매량의 평균을 분석했다.

- 소비자들은 0 ~ 60달러까지는 가격을 크게 신경쓰지 않는다. 하지만 60달러를 초과하면 판매량이 급감한다.

게임 데이터 분석 및 예측 | 분석

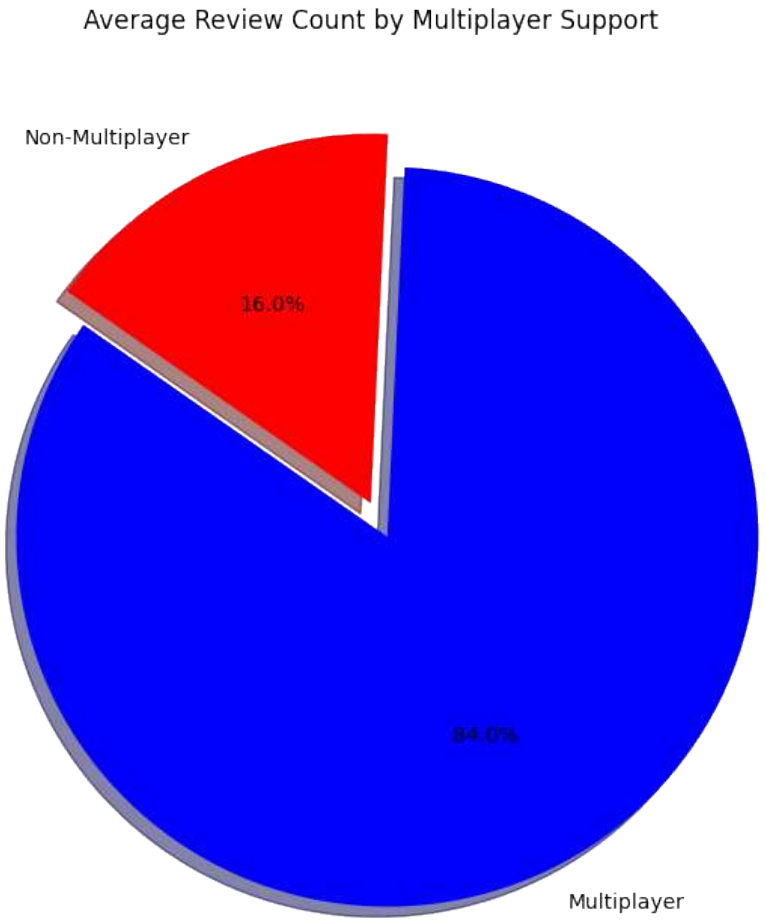


4.1. 멀티플레이 지원 여부 별 분포도

- 멀티플레이 기능이 있으면 커뮤니티가 형성되는 등 플레이어들의 관심을 끌어 판매량에 영향을 미친다고 생각했다.

- 판매량의 멀티플레이 지원 여부 별 분포도를 분석해보니 지원하는 게임이 70%이상이므로 대부분의 게임이 멀티플레이를 지원한다.

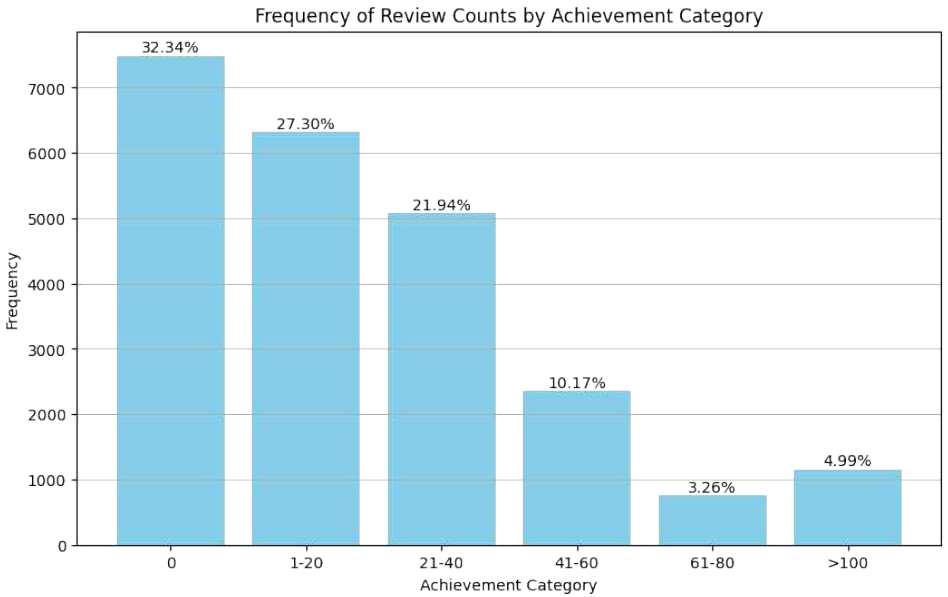
게임 데이터 분석 및 예측 | 분석



4.2. 멀티플레이 지원 여부 별 상위 40% 판매량

- 멀티플레이 지원 여부 별로 상위 40%씩 균일하게 잘라서 판매량을 분석했다.
- 멀티플레이 비지원 게임의 판매량: 16.0%
- 멀티플레이 지원 게임의 판매량: 84.0%
- 멀티플레이 기능이 없는 것보다 있는 것이 판매량이 압도적으로 높았다.

게임 데이터 분석 및 예측 | 분석

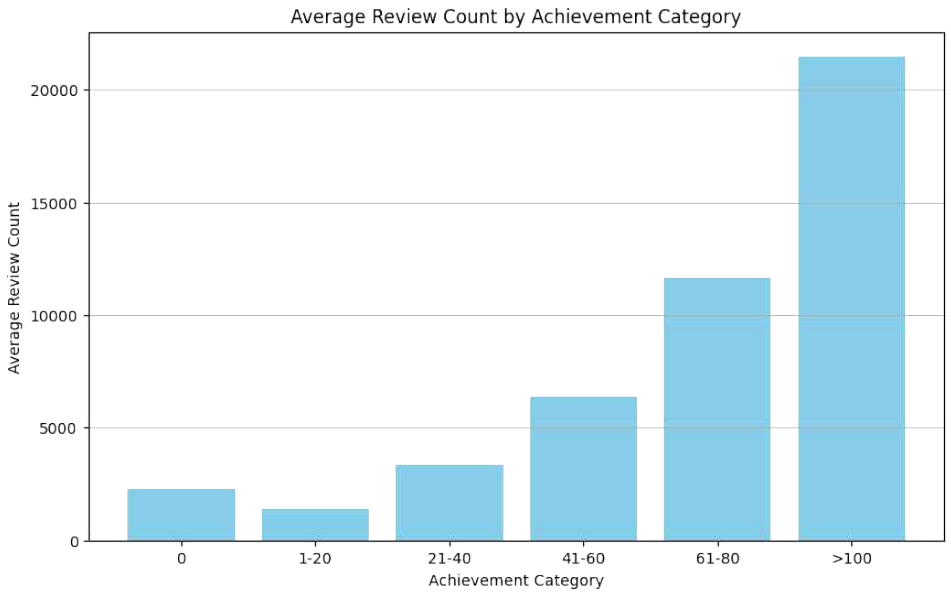


5.1. 업적수 그룹 별 빈도수

- 업적수가 높아질수록 사용자들은 컨텐츠가 많다고 인식하여 판매량이 높을 것이라고 가정하고
  업적수 그룹 별 빈도수를 분석하였다.

- 업적수의 빈도수가 낮은 순으로 빈도수가 높았다.

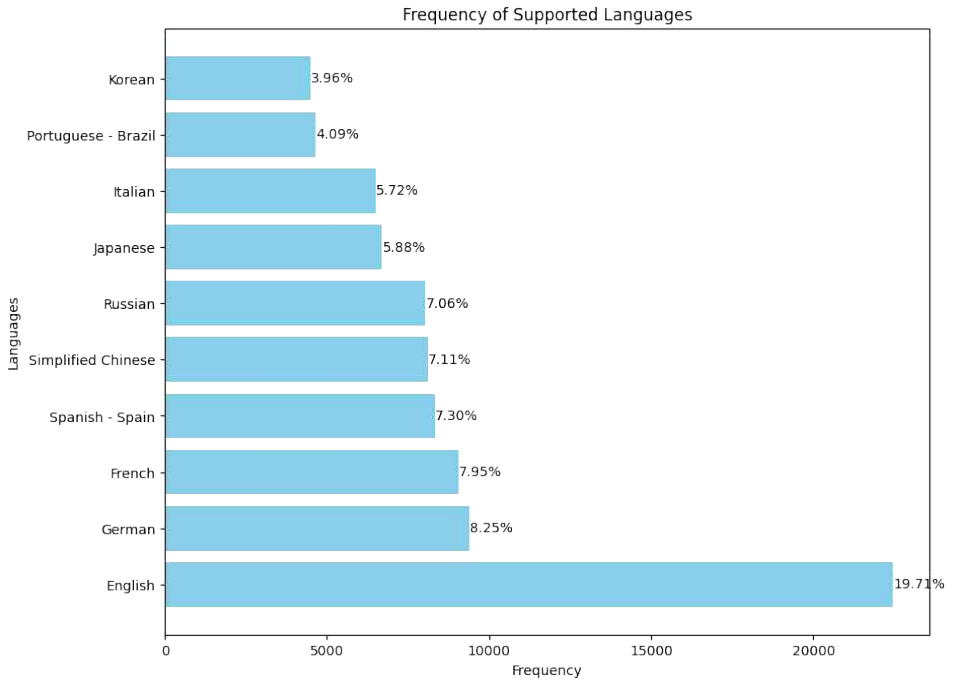
게임 데이터 분석 및 예측 | 분석



5.2. 업적수 별 판매량

- 업적수가 늘어날수록 판매량이 점진적으로 늘어난다.
- 업적수가 100개 이상일 때, 가장 많은 판매량을 기록했다.

게임 데이터 분석 및 예측 | 분석

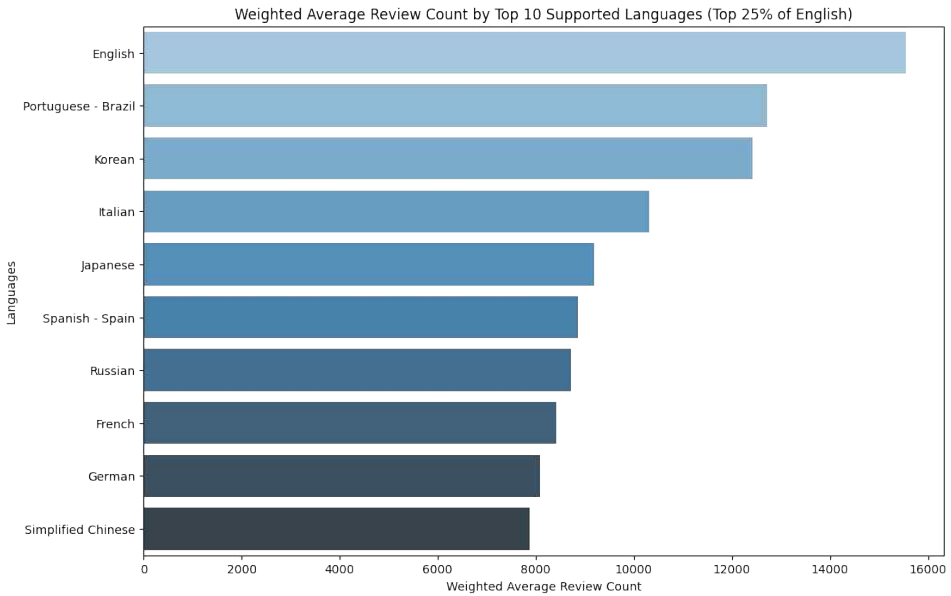


6.1. 지원 언어 빈도수 상위 10개

- 지원 언어 별 판매량을 분석하기 위해 지원 언어
  빈도수 별로 상위 10개를 시각화 하였다.

- 영어를 지원하는 경우가 19.71%로 가장 많았다.

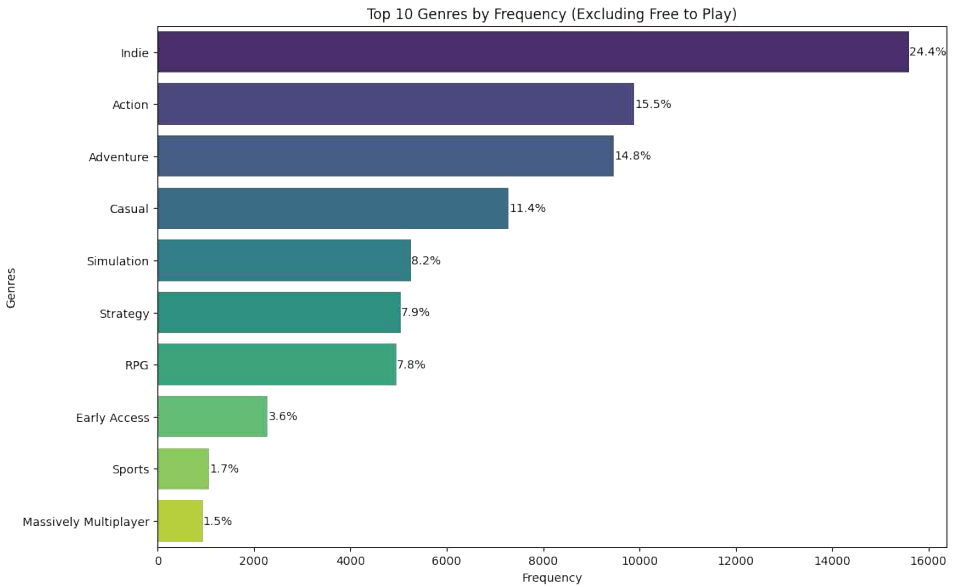
게임 데이터 분석 및 예측 | 분석



6.2. 지원 언어 별 판매량

- 영어만 상위 25%를 추출하고 나머지 그룹들의 판매량 평균을 계산하였다.
- 판매량은 영어(19.71%), 포르투갈어, 한국어, 이탈리아어, 일본어, 스페인어, 러시아어, 프랑스어, 독일어, 중국어 순서로 가장 많았다.
- 영어를 지원하는 경우가 19.71%로 가장 많았다.

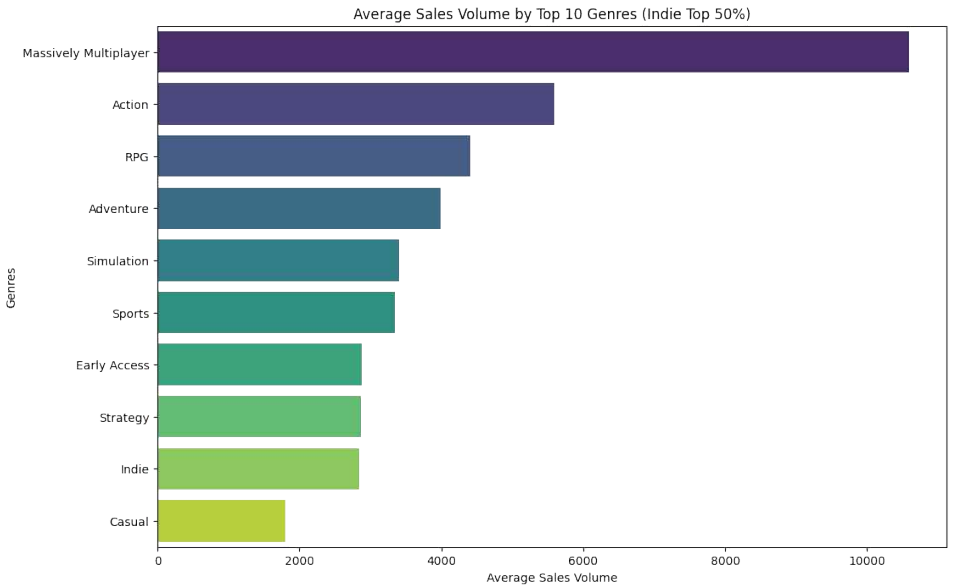
게임 데이터 분석 및 예측 | 분석



7.1. 장르 별 빈도수

- 장르 별 판매량을 분석하기 위해 장르 별 빈도수를 분석하였다.
- 장르 별 빈도수를 상위 10개를 뽑아 시각화 하였더니 ‘indie’ 장르가 24.4% 빈도수가 가장 높았다.

게임 데이터 분석 및 예측 | 분석

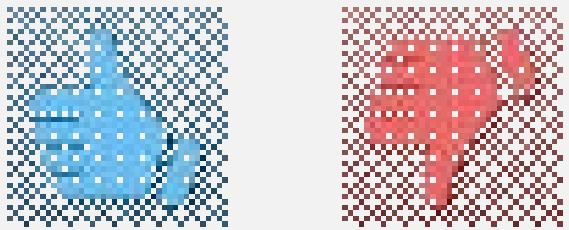


7.2. 장르 별 판매량

- ‘indie’ 장르의 상위 50%만 추출하고 나머지 장르의 판매량 평균을 계산하였다.

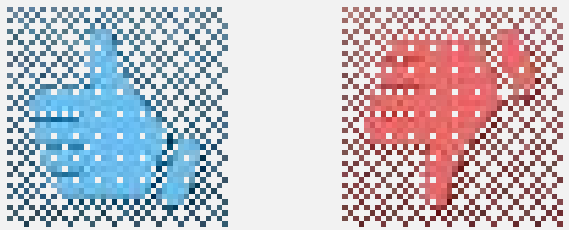
- ‘MMO(Massively Multiplayer Online)’ 장르가 빈도수에서는 꼴찌였는데 판매량에선 1등이었다.

게임 데이터 분석 및 예측 | 분석



8. 판매량과 feature

- 게임 제작 전 판매량(리뷰 수), 소유자 수를 예측할 수 있다고 가정했다.
  - 게임의 장르, 게임 구동 환경 등이 게임 구매 또는 리뷰 작성에 영향을 미칠 것이라고 생각.

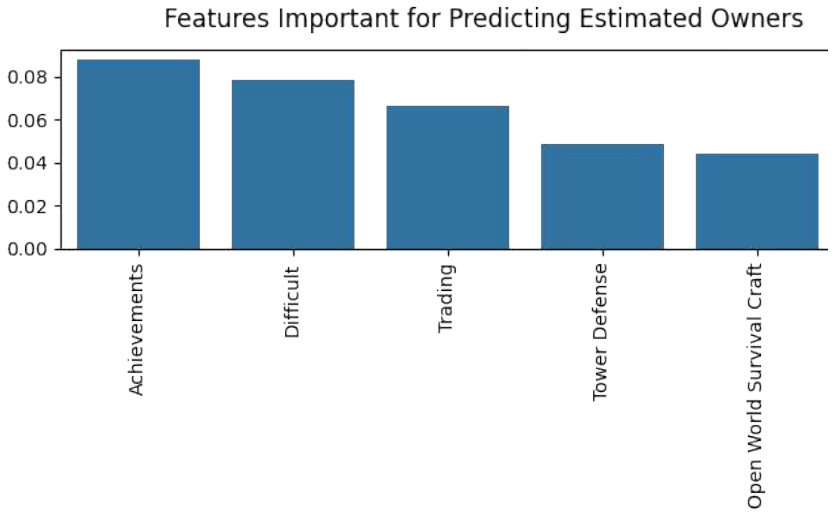


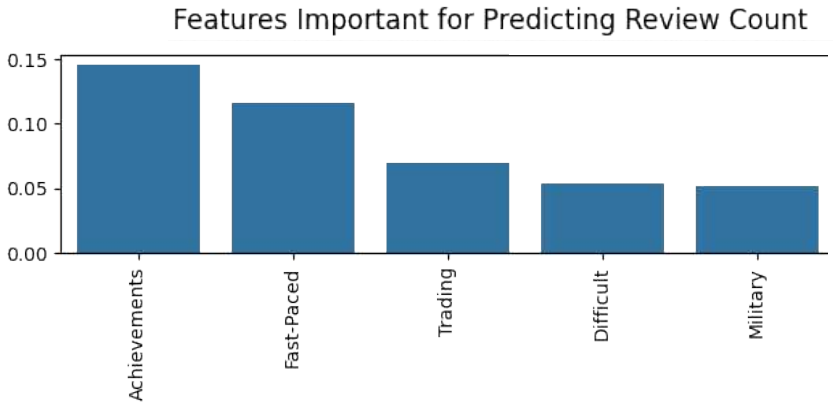
8.1. 판매량과 feature 분석을 위한 전처리

- Feature Selection
  - 소유자 수 및 판매량에 직접적인 영향을 미치거나, 게임 제작 계획 시 결정할 수 없는 요소(최고 동시접속자 수, 긍정적 평가, 부정적
    평가 등)에 대해 Feature 제거 후 분석

- 삭제한 Feature
  - 출시일, 최고 동시접속자 수, 긍정적 평가, 부정적 평가, 게임 추천 수, 평균 플레이 시간, 개발사 및 배급사

게임 데이터 분석 및 예측 | 분석

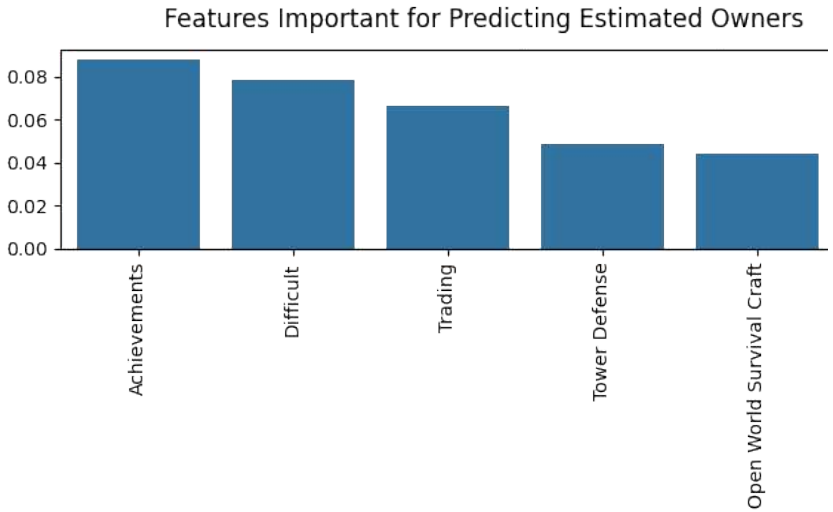


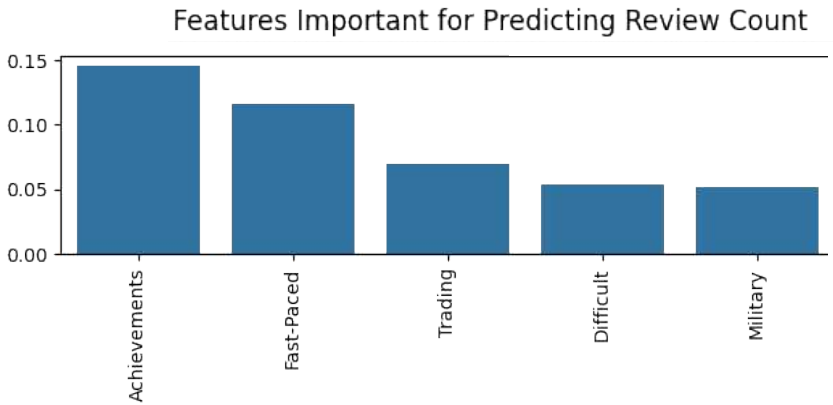


8.2. 판매량과 feature 분석 및 결과

- Random Forest를 이용해 변수 중요도 평가
- 대체로 모든 Feature가 판매량 및 리뷰 수에 낮은 상관관계를 보인다.
  - 상관계수 0.15 이하
- 따라서 판매량(리뷰 수), 소유자 수를 예측하는 모델 구축은 어려움.

게임 데이터 분석 및 예측 | 분석





8.2. 판매량과 feature 분석 및 결과

- 그래프를 보았을 때 눈에 띄는 요소가 있음.
  - 업적수 (Achievements)
- 게임 판매량(리뷰 수), 소유자 수에서 상대적으로 높은 중요도를 보임.
  - 약 0.09, 0.15로 1순위
- 게임 구매 시 업적수는 게임 구매 및 리뷰 작성에 적지 않은 영향을 미친다.

Tag - Tag

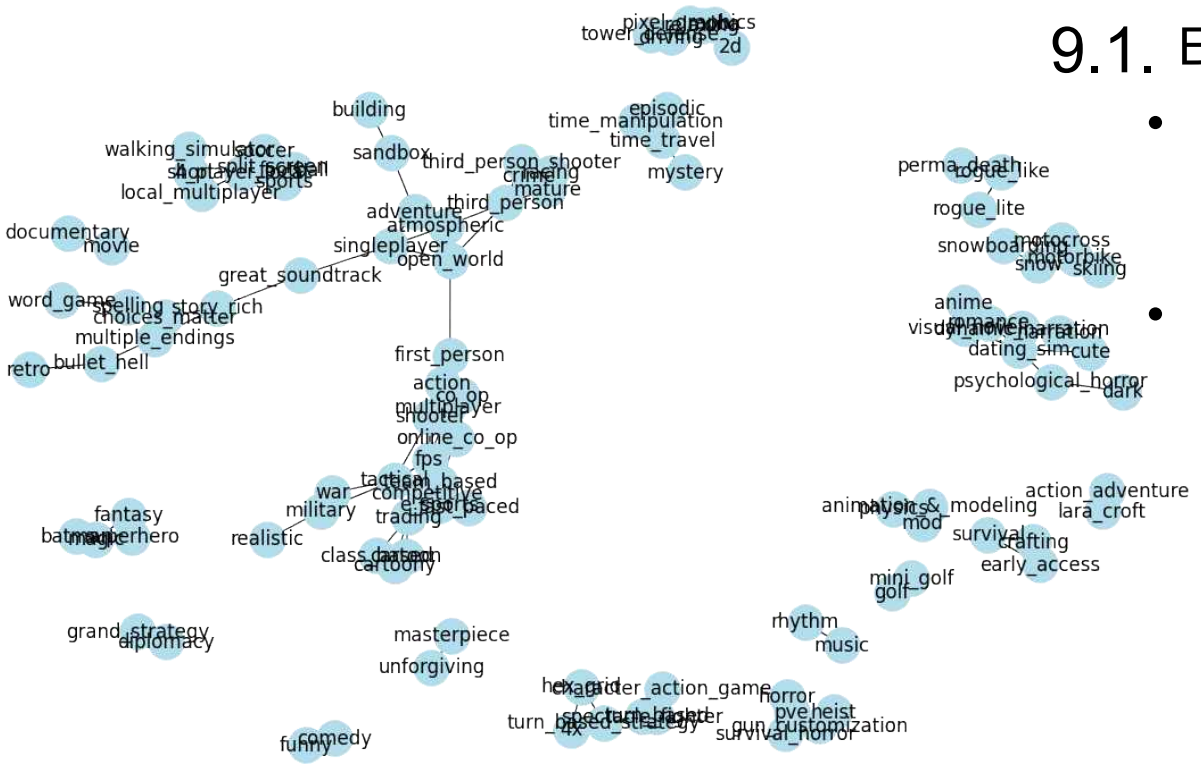
9. 태그와 태그 간의 상관관계
   - 태그
     - 게임에 적절하다고 생각하는 카테고리를 사용자가 분류 한 것

Tag - Tag

9. 태그와 태그 간의 상관관계
   - 태그들끼리의 상관관계가 있을 것이라고 가정했다.
     대부분 게임은 특정한 장르(예: RPG, FPS)와 테마(예: 판타지, 공상과학)를 공유하고 이런 요소들은 자연스럽게 함께 나타나는 경향이 있음.

     - *예) 판타지 테마는 종종 RPG 장르와 연결*
       공상과학 테마는 FPS 장르와 자주 연결

게임 데이터 분석 및 예측 | 분석



9.1. 태그와 태그 간의 상관관계 분석 및 결과

- 특성 간 상관관계를 네트워크 관계도로 시각화한다.
- 상관관계 강도에 따라 노드를 연결하고 간격을 결정한다.

Feature 와 Feature	상관계수
2D와 Pixel Graphics	0.863
4 Player Local과 Local Multiplayer	0.795
4X와 Hex Grid	0.801
4X와 Turn-Based Strategy	0.749
Action과 Co-op	0.831
Action과 First Person	0.853
Action과 Multiplayer	0.893
Action과 Shooter	0.863

9.1. 태그와 태그 간의 상관관계 분석 및 결과

- 태그 간 상관관계가 0.7 이상인 경우 head를 살펴본다.
  - Action과 다른 태그의 관계가 많음
  - Action과 관련 및 파생된 태그가 많을 것이다.

- 이와 같은 상관관계를 이용해 게임 추천 모델을
만들어본다.

경상국립대학교
Gyeongsang National University

게임 데이터 분석 및 예측

모델링

게임 데이터 분석 및 예측 | 모델링

경상국립대학교 Gyeongsang National University

AppID	Name	action	action_rpg	action_adv	addictive	adventure	agriculture	aliens
10	Counter-S	2681	0	0	0	0	0	0
20	Team Fort	208	0	0	0	15	0	0
30	Day of De	99	0	0	0	0	0	0
40	Deathmat	85	0	0	0	0	0	0
50	Half-Life:	211	0	0	0	87	0	122
60	Ricochet	108	0	0	0	0	0	0
70	Half-Life	766	0	0	0	306	0	424
80	Counter-S	377	0	0	0	40	0	0
130	Half-Life:	187	0	0	0	66	0	103
220	Half-Life 2	1761	0	0	0	937	0	555
240	Counter-S	1785	0	0	0	0	0	0
280	Half-Life:	186	0	0	0	68	0	108
300	Day of De	224	0	0	0	0	0	0
320	Half-Life 2	199	0	0	0	14	0	0
340	Half-Life 2	208	0	0	0	67	0	69
360	Half-Life D	101	0	0	0	0	0	0
380	Half-Life 2	354	0	0	0	165	0	156
400	Portal	674	0	0	0	377	0	0
420	Half-Life 2	444	0	0	0	206	0	202
440	Team Fort	8188	0	0	0	0	0	0
500	Left 4 Dea	633	0	0	0	184	0	0
550	Left 4 Dea	2804	0	0	0	896	0	0
570	Dota 2	5168	1664	0	0	0	0	0
620	Portal 2	1225	0	0	0	1644	0	0
630	Alien Swa	505	0	0	0	90	0	378
730	Counter-S	12973	0	0	0	0	0	0

1. 데이터 셋
   - 사용자들이 게임에 붙인 태그 데이터
   - 태그에 대한 빈도수를 크롤링하여 데이터 셋을 확보
   - 태그 빈도가 수치형으로 표현됨
   - 게임 추천 모델을 설계하기에 적합

게임 데이터 분석 및 예측 | 모델링

경상국립대학교 Gyeongsang National University

Normalized Tag Data Sample

	1980s	1990s	2.5d	2d	2d_fighter	360_video	3d	\
0	12.799760	77.659506	-0.054102	-0.094166	-0.044547	-0.014073	-0.052744	
1	-0.029548	9.737427	-0.054102	-0.094166	-0.044547	-0.014073	-0.052744	
2	-0.029548	-0.044455	-0.054102	-0.094166	-0.044547	-0.014073	-0.052744	
3	-0.029548	-0.044455	-0.054102	-0.094166	-0.044547	-0.014073	-0.052744	
4	-0.029548	10.564065	-0.054102	-0.094166	-0.044547	-0.014073	-0.052744	

	3d_platformer	3d_vision	4_player_local	...	warhammer_40k	\
0	-0.056544	-0.038616	-0.057471	...	-0.033268	
1	-0.056544	-0.038616	-0.057471	...	-0.033268	
2	-0.056544	-0.038616	-0.057471	...	-0.033268	
3	-0.056544	-0.038616	-0.057471	...	-0.033268	
4	-0.056544	-0.038616	-0.057471	...	-0.033268	

	web_publishing	werewolves	western	word_game	world_war_i	world_war_ii	\
0	-0.043102	-0.022867	-0.028896	-0.026508	-0.022218	-0.042029	
1	-0.043102	-0.022867	-0.028896	-0.026508	-0.022218	-0.042029	
2	-0.043102	-0.022867	-0.028896	-0.026508	0.590845	3.404246	
3	-0.043102	-0.022867	-0.028896	-0.026508	-0.022218	-0.042029	
4	-0.043102	-0.022867	-0.028896	-0.026508	-0.022218	-0.042029	

	wrestling	zombies	e_sports
0	-0.025062	-0.041819	6.779347
1	-0.025062	-0.041819	-0.014138
2	-0.025062	-0.041819	-0.014138
3	-0.025062	-0.041819	-0.014138
4	-0.025062	-0.041819	-0.014138

2. 전처리

- 태그 데이터 정규화
  - 태그 값은 빈도수를 나타냄.
  - 학습을 위해 빈도수 값 정규화

- 태그 벡터화
  - 게임 태그를 벡터로 변환
  - 게임 간 유사성 계산에 유용함

게임 데이터 분석 및 예측 | 모델링

경상국립대학교 Gyeongsang National University

```python
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity

# 태그 데이터 벡터화
tag_columns = data.columns[2:]
tag_data = data[tag_columns].astype(str).agg(' '.join, axis=1)

# TF-IDF 벡터화
vectorizer = TfidfVectorizer()
tfidf_matrix = vectorizer.fit_transform(tag_data)

# 코사인 유사도 계산
cosine_sim = cosine_similarity(tfidf_matrix, tfidf_matrix)

# 추천 함수
def recommend_games(game_name, cosine_sim=cosine_sim):
    idx = data.index[data['Name'] == game_name].tolist()[0]
    sim_scores = list(enumerate(cosine_sim[idx]))
    sim_scores = sorted(sim_scores, key=lambda x: x[1], reverse=True)
    sim_scores = sim_scores[1:11]  # 상위 10개 추천
    game_indices = [i[0] for i in sim_scores]
    return data['Name'].iloc[game_indices]

# 예시 게임 추천
recommended_games = recommend_games('Counter-Strike: Global Offensive')
print(recommended_games)
```

2. 추천 시스템 설계

- 협업 필터링 (Collaborative Filtering)
  - 사용자와 항목 간 상호작용 데이터를 기반으로 추천

- 사용자 기반 협업 필터링
  - 유사한 취향을 가진 사용자가 선호하는 게임 추천

- 아이템 기반 협업 필터링
  - 유사한 특성을 가진 게임 추천

게임 데이터 분석 및 예측 | 모델링

경상국립대학교 Gyeongsang National University

```python
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity

# 태그 데이터 벡터화
tag_columns = data.columns[2:]
tag_data = data[tag_columns].astype(str).agg(' '.join, axis=1)

# TF-IDF 벡터화
vectorizer = TfidfVectorizer()
tfidf_matrix = vectorizer.fit_transform(tag_data)

# 코사인 유사도 계산
cosine_sim = cosine_similarity(tfidf_matrix, tfidf_matrix)

# 추천 함수
def recommend_games(game_name, cosine_sim=cosine_sim):
    idx = data.index[data['Name'] == game_name].tolist()[0]
    sim_scores = list(enumerate(cosine_sim[idx]))
    sim_scores = sorted(sim_scores, key=lambda x: x[1], reverse=True)
    sim_scores = sim_scores[1:11]  # 상위 10개 추천
    game_indices = [i[0] for i in sim_scores]
    return data['Name'].iloc[game_indices]

# 예시 게임 추천
recommended_games = recommend_games('Counter-Strike: Global Offensive')
print(recommended_games)
```

2. 추천 시스템 설계

- 게임 태그 벡터화 (TF-IDF)

- 코사인 유사도로 유사한 게임 추천
  - 벡터 방향이 완전히 동일할 경우 1
  - 1에 가까울 경우 유사도가 높다.

게임 데이터 분석 및 예측 | 모델링

경상국립대학교 Gyeongsang National University

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity

# 태그 데이터 벡터화
tag_columns = data.columns[2:]
tag_data = data[tag_columns].astype(str).agg(' '.join, axis=1)

# TF-IDF 벡터화
vectorizer = TfidfVectorizer()
tfidf_matrix = vectorizer.fit_transform(tag_data)

# 코사인 유사도 계산
cosine_sim = cosine_similarity(tfidf_matrix, tfidf_matrix)

# 추천 함수
def recommend_games(game_name, cosine_sim=cosine_sim):
    idx = data.index[data['Name'] == game_name].tolist()[0]
    sim_scores = list(enumerate(cosine_sim[idx]))
    sim_scores = sorted(sim_scores, key=lambda x: x[1], reverse=True)
    sim_scores = sim_scores[1:11]  # 상위 10개 추천
    game_indices = [i[0] for i in sim_scores]
    return data['Name'].iloc[game_indices]

# 예시 게임 추천
recommended_games = recommend_games('Counter-Strike: Global Offensive')
print(recommended_games)
```

3. TF-IDF 모델 결과 한계

- 해석이 가능할 정도의 간단한 모델을 구현할 수 있었음.
- 일부 관련 없는 게임을 추천하는 경우가 있음
- 태그 간의 상관관계를 학습하지 않아 낮은 성능을 보이는 것으로 보임

게임 데이터 분석 및 예측 | 모델링

경상국립대학교 Gyeongsang National University

Autoencoder Vectorized Tag Data Sample

	0	1	2	3	4	5	\
0	0.000000	0.000000	62.025379	0.000000	114.039665	0.000000	
1	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
2	1.653300	1.464302	3.636649	17.305664	6.612454	18.020454	
3	10.714424	25.503443	0.000000	38.663921	0.038289	47.374165	
4	0.000000	47.524311	15.276979	0.000000	22.086117	0.000000	

	6	7	8	9	...	54	55	\
0	0.000000	0.000000	0.000000	230.547287	...	0.000000	251.030899	
1	0.000000	0.221083	19.347893	37.335075	...	0.000000	35.981030	
2	21.186541	7.089534	26.332651	7.597995	...	15.650640	15.587583	
3	41.054882	0.000000	48.074509	5.546671	...	36.831913	12.671486	
4	0.000000	16.701242	18.171078	41.328613	...	0.000000	34.152683	

	56	57	58	59	60	61	\
0	0.000000	163.558655	0.000000	0.000000	0.000000	0.000000	
1	0.000000	24.526339	17.430275	0.000000	0.000000	0.000000	
2	9.177119	16.919405	26.712786	10.078756	15.742235	22.067097	
3	8.118855	10.324480	10.852064	28.191288	41.150150	20.067602	
4	17.688955	38.514629	47.070747	0.000000	0.000000	0.000000	

	62	63
0	0.000000	0.000000
1	0.000000	0.000000
2	15.107454	18.576069
3	40.133358	0.553476
4	0.000000	0.000000

[5 rows x 64 columns]

4. 모델 변경

- 오토인코더 방식
  - 태그 데이터를 인코더와 디코더로 구성된 오토인코더 모델에 학습
  - 인코더 중간 층에서 저차원 임베딩 벡터 추출
  - 임베딩 벡터 간 코사인 유사도로 유사한 게임 검색

게임 데이터 분석 및 예측 | 모델링

경상국립대학교 Gyeongsang National University

Autoencoder Vectorized Tag Data Sample

	0	1	2	3	4	5	\
0	0.000000	0.000000	62.025379	0.000000	114.039665	0.000000	
1	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
2	1.653300	1.464302	3.636649	17.305664	6.612454	18.020454	
3	10.714424	25.503443	0.000000	38.663921	0.038289	47.374165	
4	0.000000	47.524311	15.276979	0.000000	22.086117	0.000000	

	6	7	8	9	...	54	55	\
0	0.000000	0.000000	0.000000	230.547287	...	0.000000	251.030899	
1	0.000000	0.221083	19.347893	37.335075	...	0.000000	35.981030	
2	21.186541	7.089534	26.332651	7.597995	...	15.650640	15.587583	
3	41.054882	0.000000	48.074509	5.546671	...	36.831913	12.671486	
4	0.000000	16.701242	18.171078	41.328613	...	0.000000	34.152683	

	56	57	58	59	60	61	\
0	0.000000	163.558655	0.000000	0.000000	0.000000	0.000000	
1	0.000000	24.526339	17.430275	0.000000	0.000000	0.000000	
2	9.177119	16.919405	26.712786	10.078756	15.742235	22.067097	
3	8.118855	10.324480	10.852064	28.191288	41.150150	20.067602	
4	17.688955	38.514629	47.070747	0.000000	0.000000	0.000000	

	62	63
0	0.000000	0.000000
1	0.000000	0.000000
2	15.107454	18.576069
3	40.133358	0.553476
4	0.000000	0.000000

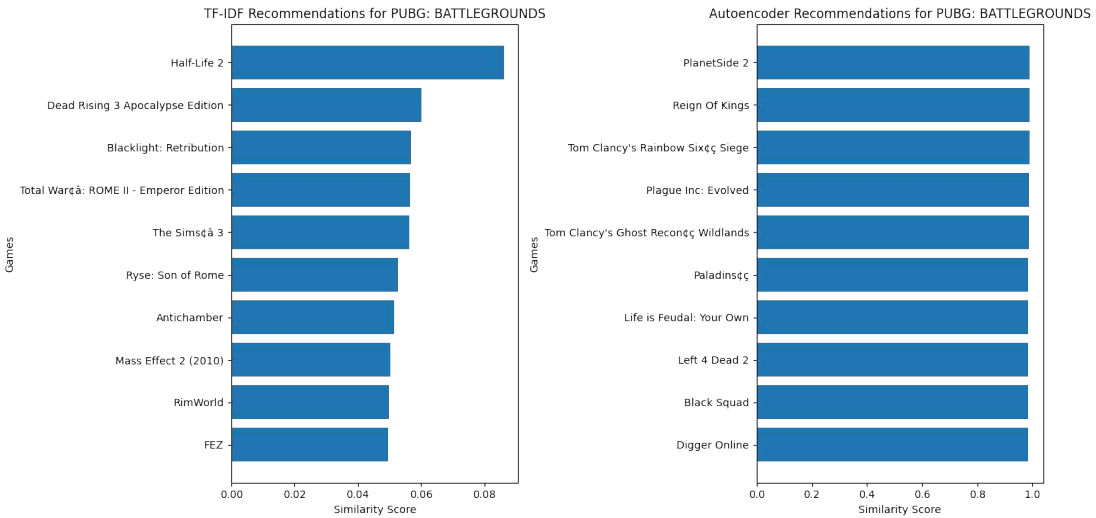
[5 rows x 64 columns]

4. 모델 변경

- 오토인코더 방식
  - 장점: 태그 간 관계를 학습해 정교한 추천 가능
  - 비선형 활성 함수를 사용해 기존 선형 학습법에 비해 성능 향상
  - 빈도수 뿐만 아니라 태그 간 비선형적 관계를 학습해 특성 반영

게임 데이터 분석 및 예측 | 모델링

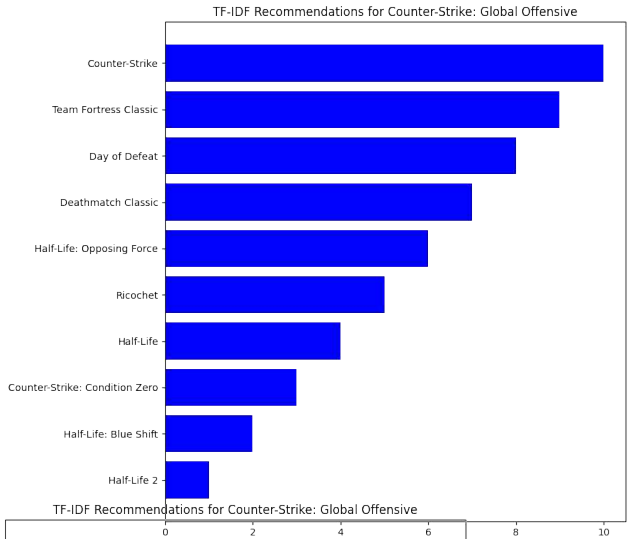
경상국립대학교 Gyeongsang National University

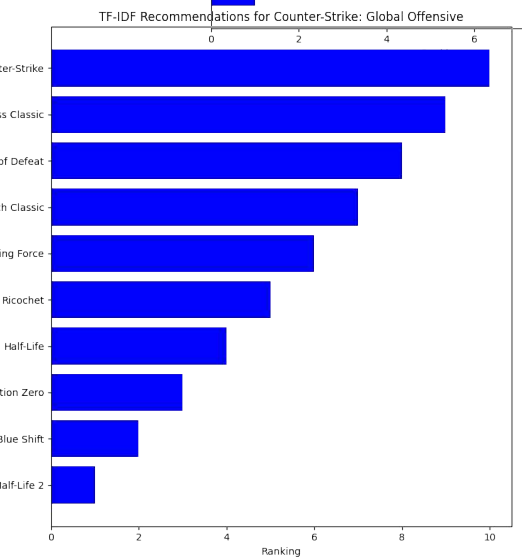


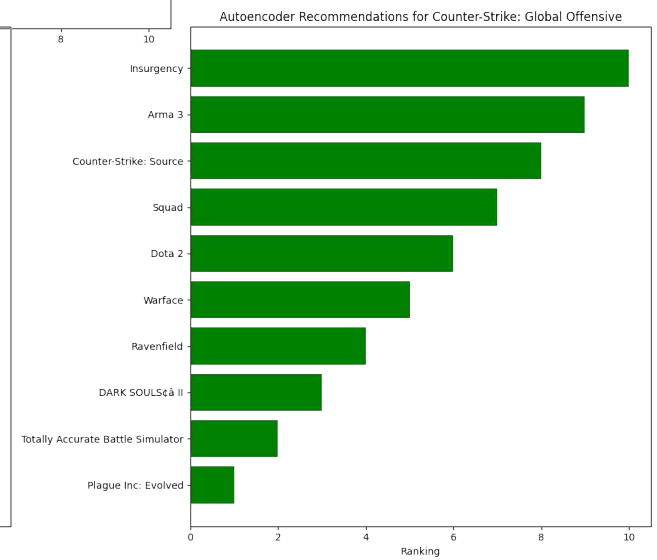
5. 추천 시스템 평가

- 성능 비교 방법
  - 상위 추천 결과 중 실제로 사용자가 선호한 게임의 비율을 확인해야 함
- 그러나 데이터 셋에 사용자 별 평가 데이터가 없음
  - 주관적인 평가가 필요함
  - 추천 결과의 직관적인 비교를 통해 성능 평가
- 대표 게임 2개에 대해 출력된 결과로 어떤 모델이 더 나은 추천을 제공하는지 평가

게임 데이터 분석 및 예측 | 모델링



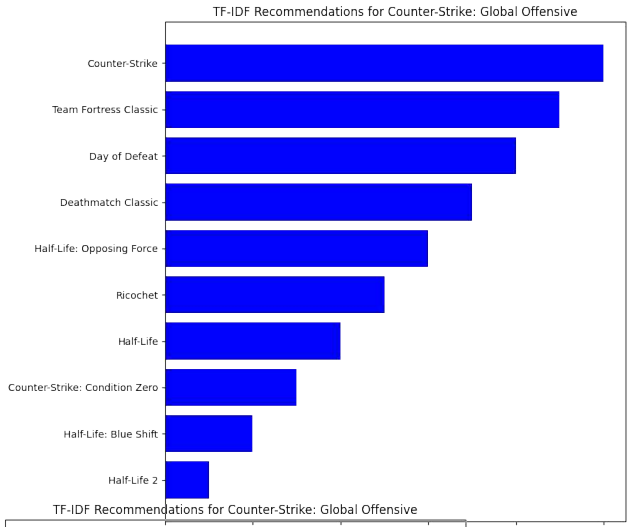


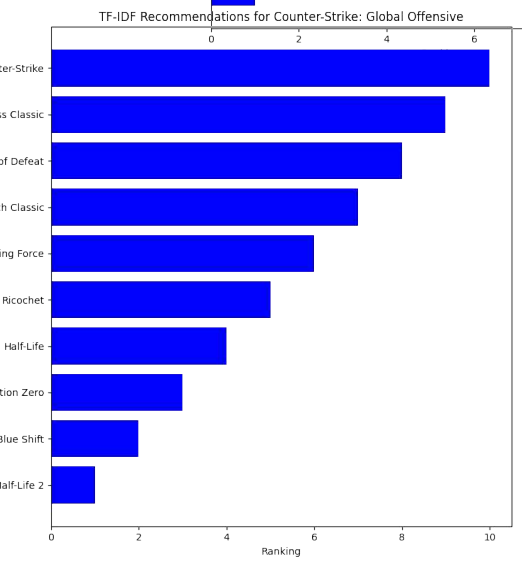


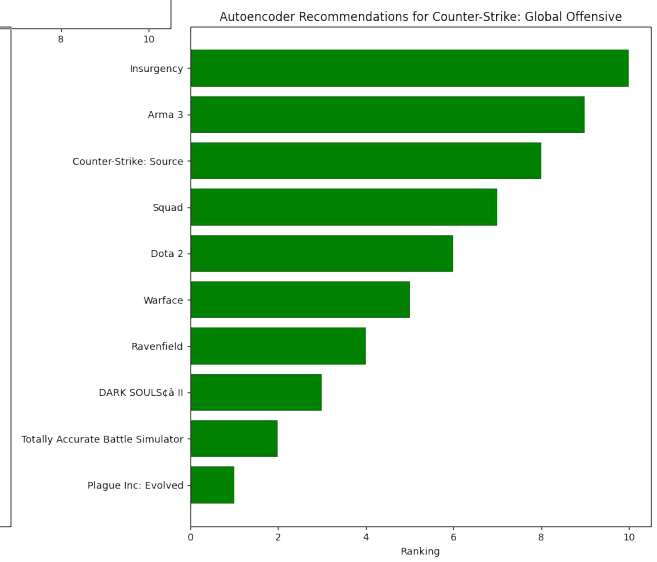
5.1. 추천 시스템 평가: CS:GO

- TF-IDF
  - 태그와 유사함
  - 고전 게임 추천
    - 대부분 2004년 이전
- 오토인코더
  - 태그와 유사함
  - 최신 게임 추천
    - 대부분 2013년 이후
- 게임의 중요 요소인 전략을 요하는 게임이 많이 추천됨

게임 데이터 분석 및 예측 | 모델링





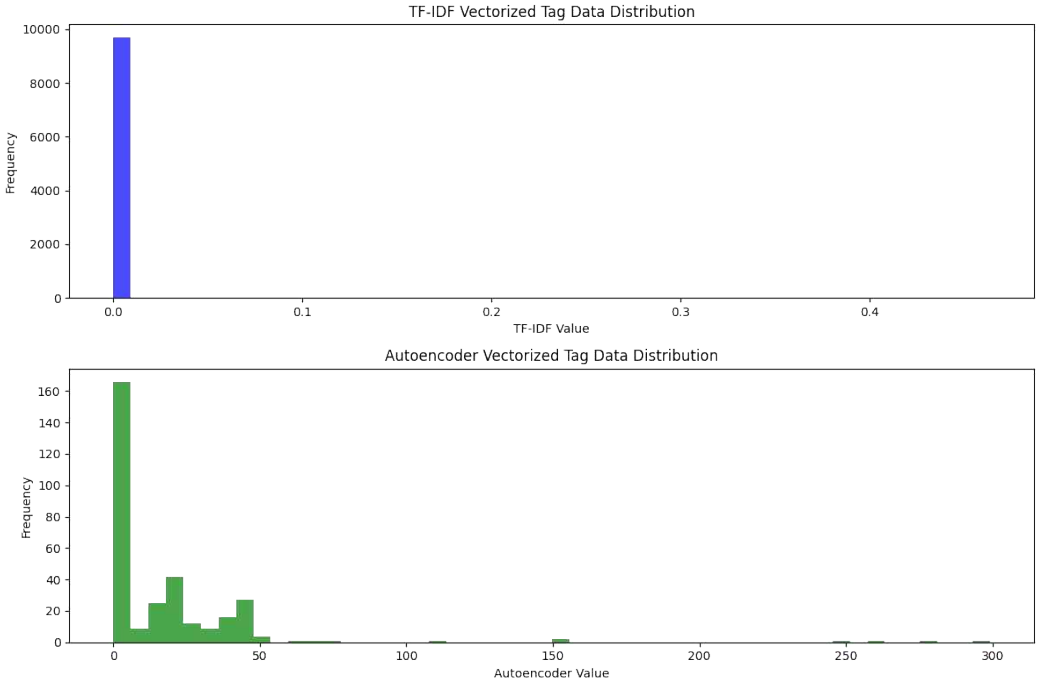


5.2. 추천 시스템 평가: BATTLEGROUNDS

- TF-IDF
  - 주요 태그와 유사하지 않음
    - 특정 태그에 치중됨
    - 생존, 시뮬레이션 슈팅, 전략 중 하나
- 오토인코더
  - 태그와 유사함
    - 생존, 배틀로얄
  - 게임의 중요 요소인 배틀로얄을 요하는 게임이 많이 추천됨

게임 데이터 분석 및 예측 | 모델링

경상국립대학교 Gyeongsang National University



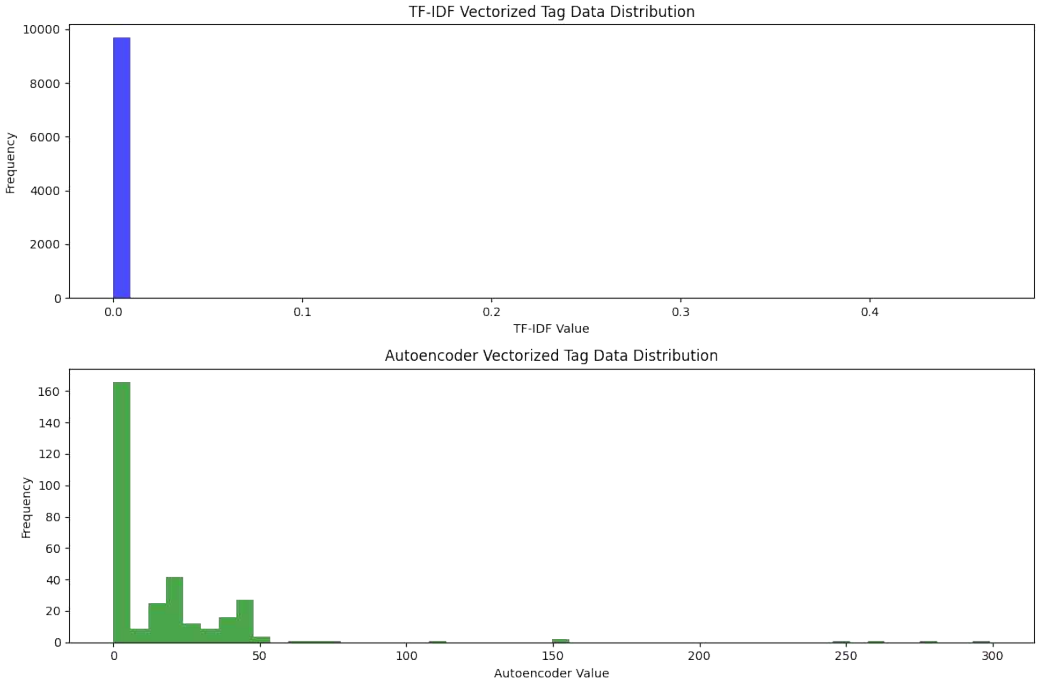
6. 추천 시스템 결론

- TF-IDF 벡터화된 데이터 히스토그램
  - 대부분의 값이 0에 가까운 것을 보여줌.
  - 데이터가 매우 희소하게 분포함.

- 오토인코더 벡터화된 데이터 히스토그램
  - 값이 0에서 멀리 떨어진 데이터 포인트가 더 많아, 태그 간의 상관관계를 더 잘 반영함.
  - 데이터가 더 균일하게 분포함.

게임 데이터 분석 및 예측 | 모델링

경상국립대학교 Gyeongsang National University



6. 추천 시스템 결론

- 오토인코더 모델
  - 비선형적 관계를 학습하여 더 정교한 임베딩을 제공해 추천 시스템 성능 향상에 유리함

- TF-IDF 모델
  - 간단하고 빠름
  - 태그 간 상관관계를 다소 반영하지 못해 성능이 다소 떨어짐

경상국립대학교
Gyeongsang National University

게임 데이터 분석 및 예측

결과

게임 데이터 분석 및 예측 | 결론

경상국립대학교 Gyeongsang National University

1. 게임 가격이 20달러가 넘어가면 사람들이 잘 안 사려고 한다.

2. 멀티플레이를 지원하면 커뮤니티가 형성되는 등 플레이어들의 관심을 끌어 판매량에 영향을 미친다.

3. 지원 언어 빈도수와 판매량은 다르다.
최적의 지원 언어를 선정하려면 판매량 순위를 기준으로 해야한다.

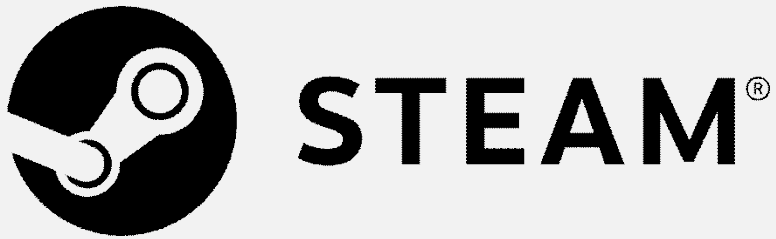
4. 게임 업적수는 게임 구매 및 리뷰 작성에 적지 않은
영향을 미친다.

5. 게임 태그 데이터를 이용해 오토인코더 기반 게임 추천 시스템을 개발할 수 있었다.

게임 데이터 분석 및 예측

추후 발전 가능성 및 후속 연구 / 프로젝트

경상국립대학교 Gyeongsang National University



1. 출시일에 따른 가격 추세 분석
2. 게임 구매 데이터를 이용해 협업 필터링을 이용한 게임 추천 시스템을 개발