2025-12-08

스팀(Steam) 게임 메타데이터 기반 흥행 요인 및 상관관계 분석

Steam Metadata-Based Success Factors and Correlation Analysis

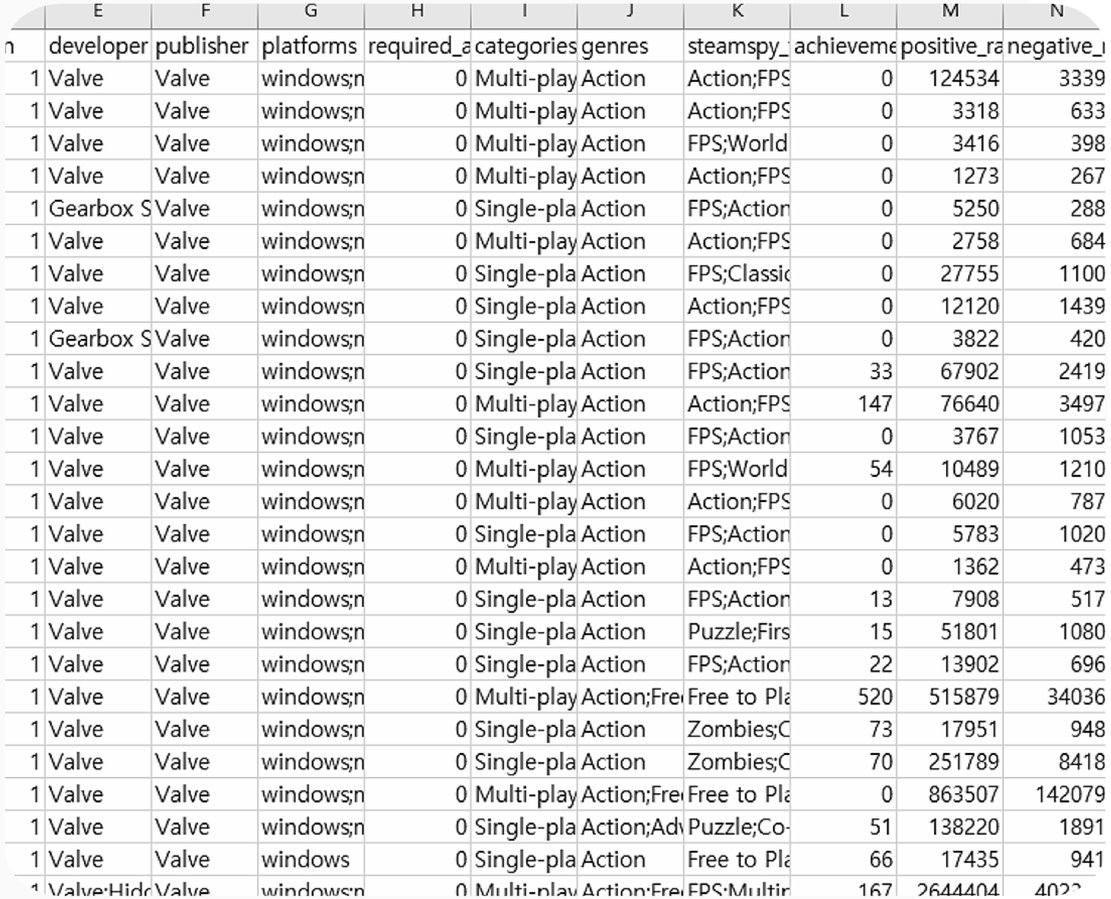
202344017 컴퓨터정보과 장예준

목차

Table of Contents

데이터 수집

Data Collection



	E	F	G	H	I	J	K	L	M	N
n	developer	publisher	platforms	required_a	categories	genres	steamspy_	achieveme	positive_ra	negative_
1	Valve	Valve	windows;n	0	Multi-play	Action	Action;FPS	0	124534	3339
1	Valve	Valve	windows;n	0	Multi-play	Action	Action;FPS	0	3318	633
1	Valve	Valve	windows;n	0	Multi-play	Action	FPS;World	0	3416	398
1	Valve	Valve	windows;n	0	Multi-play	Action	Action;FPS	0	1273	267
1	Gearbox S	Valve	windows;n	0	Single-pla	Action	FPS;Action	0	5250	288
1	Valve	Valve	windows;n	0	Multi-play	Action	Action;FPS	0	2758	684
1	Valve	Valve	windows;n	0	Single-pla	Action	FPS;Classic	0	27755	1100
1	Valve	Valve	windows;n	0	Single-pla	Action	Action;FPS	0	12120	1439
1	Gearbox S	Valve	windows;n	0	Single-pla	Action	FPS;Action	0	3822	420
1	Valve	Valve	windows;n	0	Single-pla	Action	FPS;Action	33	67902	2419
1	Valve	Valve	windows;n	0	Multi-play	Action	Action;FPS	147	76640	3497
1	Valve	Valve	windows;n	0	Single-pla	Action	FPS;Action	0	3767	1053
1	Valve	Valve	windows;n	0	Multi-play	Action	FPS;World	54	10489	1210
1	Valve	Valve	windows;n	0	Multi-play	Action	Action;FPS	0	6020	787
1	Valve	Valve	windows;n	0	Single-pla	Action	FPS;Action	0	5783	1020
1	Valve	Valve	windows;n	0	Multi-play	Action	Action;FPS	0	1362	473
1	Valve	Valve	windows;n	0	Single-pla	Action	FPS;Action	13	7908	517
1	Valve	Valve	windows;n	0	Single-pla	Action	Puzzle;Firs	15	51801	1080
1	Valve	Valve	windows;n	0	Single-pla	Action	FPS;Action	22	13902	696
1	Valve	Valve	windows;n	0	Multi-play	Action;Free	Free to Pla	520	515879	34036
1	Valve	Valve	windows;n	0	Single-pla	Action	Zombies;C	73	17951	948
1	Valve	Valve	windows;n	0	Single-pla	Action	Zombies;C	70	251789	8418
1	Valve	Valve	windows;n	0	Multi-play	Action;Free	Free to Pla	0	863507	142079
1	Valve	Valve	windows;n	0	Single-pla	Action;Adv	Puzzle;Co-	51	138220	1891
1	Valve	Valve	windows	0	Single-pla	Action	Free to Pla	66	17435	941
1	Valve;Hido	Valve	windows;n	0	Multi-play	Action;Free	FPS;Multip	167	2644404	4022

Kaggle 'Steam Store Games (Cleaned)' 데이터셋

총 데이터 수: 약 27,075개 (행)
주요 변수: 게임명, 가격, 장르, 긍정 평가 수, 부정 평가 수, 플레이 타임

선정 이유:
빅데이터(Big Data)로서 통계적 유의성을 가짐.
수치형(가격), 범주형(장르), 텍스트(이름) 데이터가 혼합되어 다양한 분석 기법 적용 가능.

스팀 상점의 실제 판매 데이터(이름, 가격, 장르, 평가 수 등)를 포함하고 있으며, 약 27,000개 이상의 방대한 표본을 확보하여 통계적 유의성을 가짐.

데이터 전처리

Preprocessing

--- 전처리 후 데이터 미리보기 ---

	name	genres	price	positive_ratings	negative_ratings	average_playtime	total_reviews	rating
0	Counter-Strike	Action	7.19	124534	3339	17612	127873	97.388815
1	Team Fortress Classic	Action	3.99	3318	633	277	3951	83.978740
2	Day of Defeat	Action	3.99	3416	398	187	3814	89.564761
3	Deathmatch Classic	Action	3.99	1273	267	258	1540	82.662338
4	Half-Life: Opposing Force	Action	3.99	5250	288	624	5538	94.799567

--- 기초 통계 요약 ---

	price	total_reviews	rating
count	27075.000000	2.707500e+04	27075.000000
mean	6.078193	1.211586e+03	71.447792
std	7.874922	2.242909e+04	23.359421
min	0.000000	1.000000e+00	0.000000
25%	1.690000	1.000000e+01	58.333333
50%	3.990000	3.600000e+01	76.033058
75%	7.190000	1.760000e+02	89.390531
max	421.990000	3.046717e+06	100.000000

1. 문제점

원본 데이터에는 비교 분석을 위한 **'객관적 지표'**가 부재함.
예: 긍정 리뷰 100개와 부정 리뷰 50개인 게임을, 긍정 10개/부정 0개인 게임과 어떻게 비교할 것인가?

2. 해결 방안

총 리뷰 수 (total_reviews): positive + negative → [게임의 인기도/화제성] 척도
평점 (rating): positive / total * 100 → [게임의 작품성/만족도] 척도 (0~100점)

데이터 전처리2

Preprocessing

Action;Free to Play
Action
Action
Action;Free to Play;Strategy
Action;Adventure
Action
Action;Free to Play
Indie
Action
Action
Action
Indie;Strategy
Indie;Strategy
Indie;Strategy

변환 전 데이터 개수: 27075
변환 후 데이터 개수: 76462

--- 장르 분리 후 데이터 미리보기 ---

	name	genres	price	rating
0	Counter-Strike	Action	7.19	97.388815
1	Team Fortress Classic	Action	3.99	83.978740
2	Day of Defeat	Action	3.99	89.564761
3	Deathmatch Classic	Action	3.99	82.662338
4	Half-Life: Opposing Force	Action	3.99	94.799567
5	Ricochet	Action	3.99	80.127833
6	Half-Life	Action	7.19	96.187836
7	Counter-Strike: Condition Zero	Action	7.19	89.387123
8	Half-Life: Blue Shift	Action	3.99	90.099010
9	Half-Life 2	Action	7.19	96.560060

1. 문제점

Multi-label(다중 장르)" 문제 해결
하나의 게임이 Action;Indie;RPG처럼 여러 장르를 동시에 가짐.
이 상태로는 'RPG 장르만의 평균 가격'을 산출할 수 없음
또한, 스팀 태그는 매우 작은 장르 요소만 포함되어있어도 태그에 포함시키기에, 장르 분할이 필수적

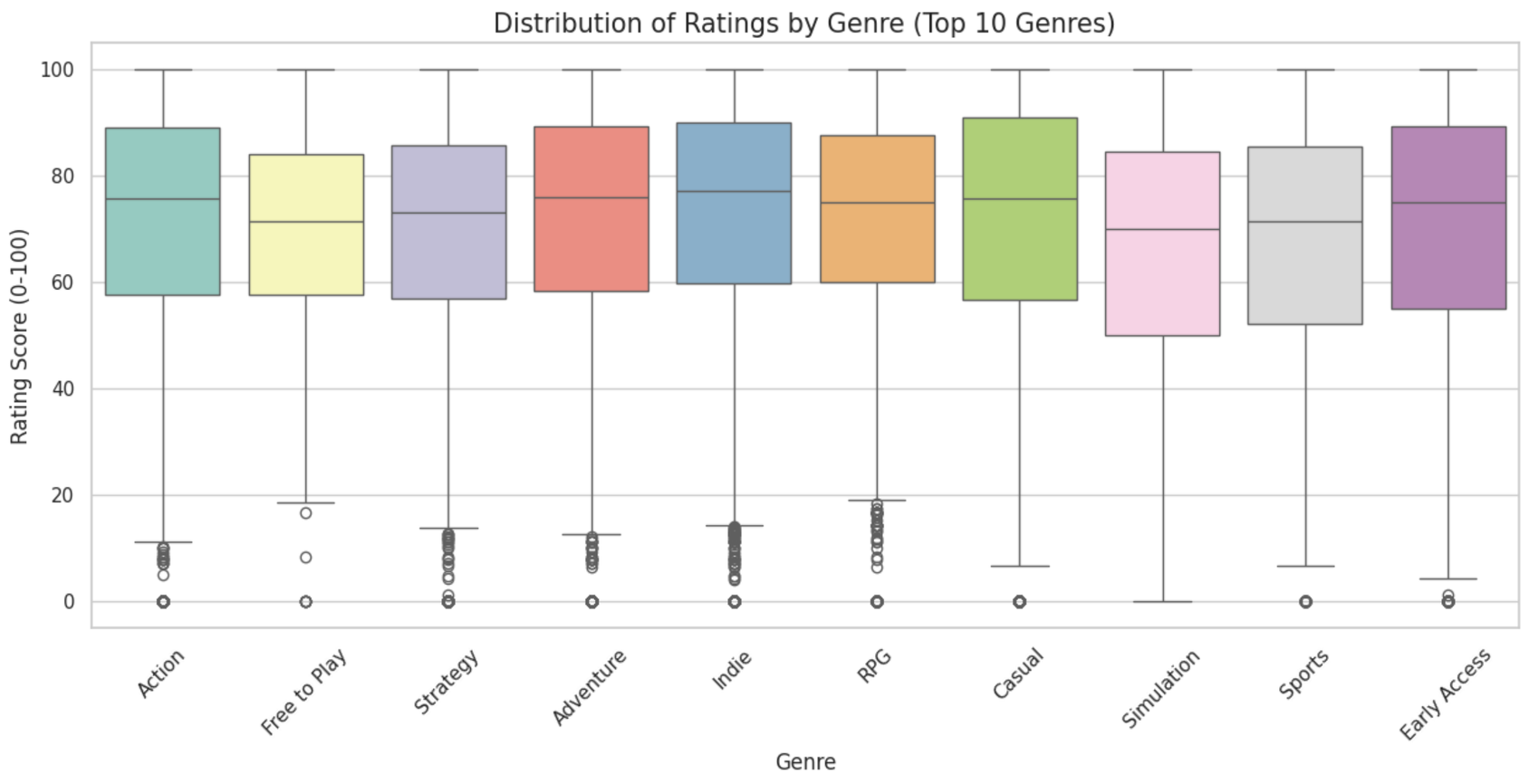
2. 해결 방안

Split: 세미콜론(;)을 기준으로 문자열 분리
Explode: 리스트 속 장르들을 개별 행(Row)으로 분리하여 데이터 확장

장르별 평점 분포 - 그래프

Boxplot

"어떤 장르가 유저들에게 후한 평가를 받는가?"



X축: 게임 장르 / Y축: 평점 (0~100점)

그래프: 상위 10개 장르(Action, RPG, Indie 등)에 따른 평점(Rating) 분포를 박스플롯(Boxplot)으로 시각화.

장르별 평점 분포 - 해설

Boxplot

상향 평준화된 시장

대부분의 장르에서 박스(Box, 중위수 50% 구간)가 60~80점 구간에 높게 형성되어 있습니다. 이는 스팀 유저들이 대체로 게임에 대해 관대하며, '기본적인 구동'만 되면 긍정적인 평가를 내리는 경향이 있음을 시사합니다.

이상치(Outlier)의 비대칭성

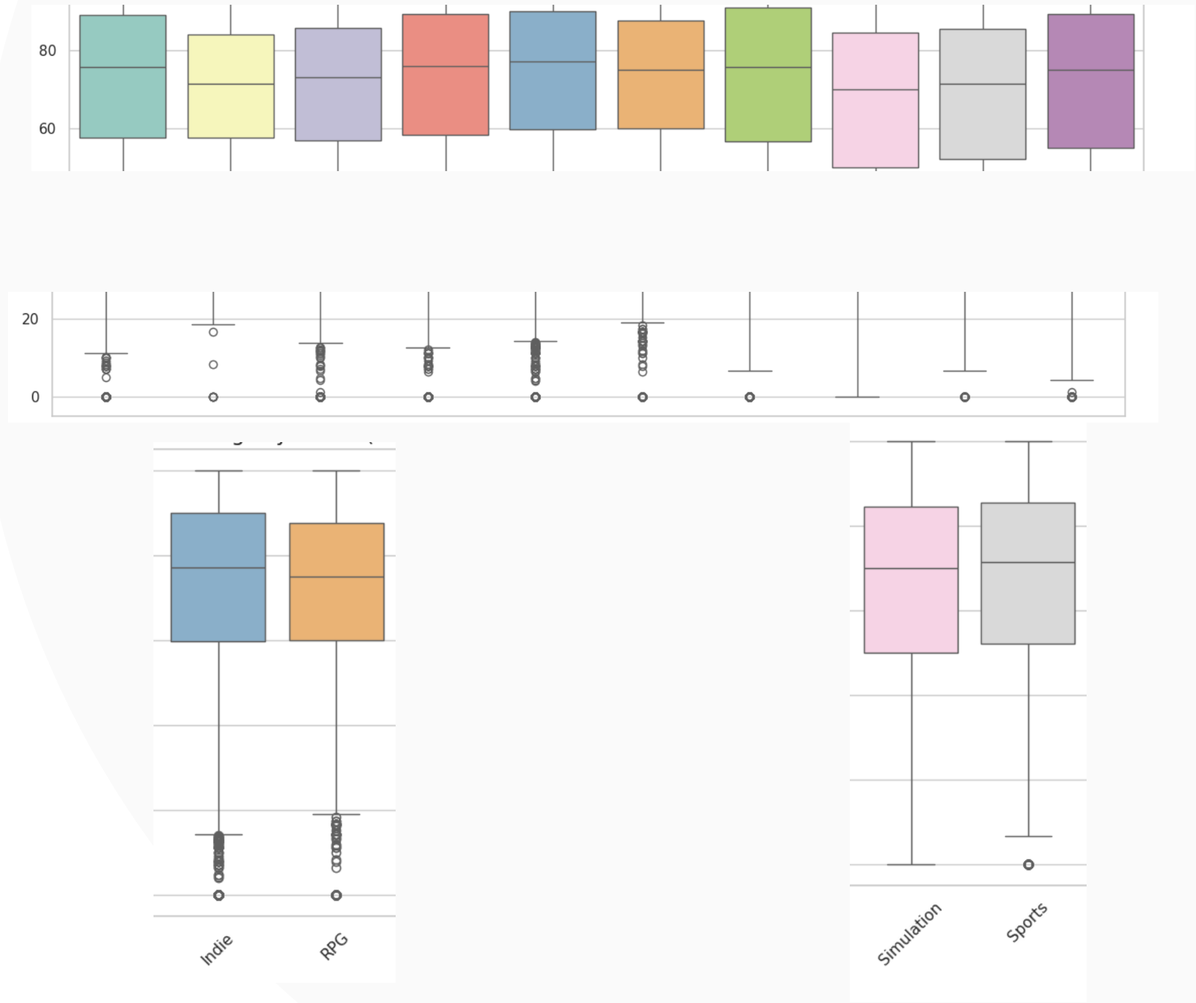
그래프 상단(100점)에는 이상치가 없으나, 하단(0~20점)에는 수많은 이상치가 존재합니다.
즉, 스팀에서 100점짜리 게임은 통계적으로 '드문 기적'이 아니지만, 0~20점대 게임은 통계적으로 매우 이질적인 '실패작'임을 의미합니다. 이는 시장에 품질 관리가 되지 않은 '저품질 양산형 게임'이 다수 존재함을 데이터로 증명합니다.

장르별 특성

RPG & Indie: 중앙값이 타 장르 대비 높다. 이는 해당 장르 팬덤의 충성도가 높거나, 개발자들이 틈새 시장(Niche Market)을 잘 공략했음을 의미하기도 합니다.

Simulation & Sports: 박스의 위치가 상대적으로 낮고 퍼져 있습니다. 이는 유저들의 기대치가 까다롭거나, 구현의 현실감이 중요하기에 구현이 어려우며, 호불호가 극명하게 갈리는 장르임을 나타냅니다.

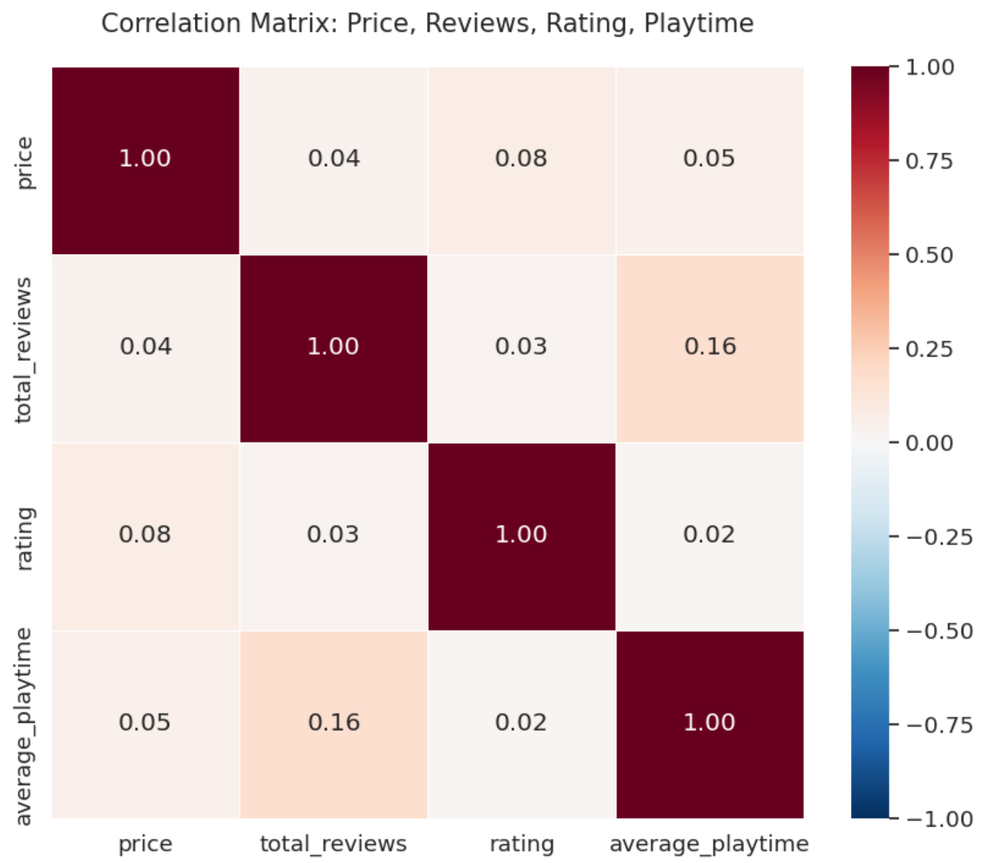
"어떤 장르가 유저들에게 후한 평가를 받는가?"



상관관계 분석 - 그래프

Heatmap Analysis

가격이 비싸면 평점이 높을까? 리뷰가 많으면 '갓겜'일까?"



주요 변수(가격, 리뷰 수, 평점, 플레이 타임) 간의 피어슨 상관계수(Pearson Correlation Coefficient) 히트맵.

범위: -1(반비례) ~ 1(정비례), 0은 관계없음

상관관계 분석 - 해석

Heatmap Analysis

가격이 비싸면 평점이 높을까? 리뷰가 많으면 '갓겜'일까?"

가격(Price) vs 평점(Rating): 상관계수 0.08

통계적으로 '무의미'에 가깝습니다.
소비자가 비싼 돈을 지불한다고 해서 더 높은 만족감(평점)을 얻는 것은 아니다.
간단하게, 60달러짜리 AAA 게임과 10달러짜리 인디 게임의 재미는 가격에 비례하지 않습니다.

리뷰 수(Reviews) vs 평점(Rating): 상관계수 0.03

해석: 가장 직관과 어긋나는 결과다. '유명한 게임(리뷰 많음)'이 곧 '좋은 게임(평점 높음)'이라는 공식이 성립하지 않는다.

이유: 유명 게임은 대중의 기대치가 높아 작은 실수에도 '비추천 테러(Review Bombing)'를 당하기 쉽고, 반대로 인디 게임은 소수 매니아층의 지지로 높은 평점을 방어하기 쉽기 때문입니다.

플레이 타임 vs 리뷰 수: 상관계수 0.16

약한 양의 상관관계를 보입니다. 플레이 타임이 긴 게임(RPG, 멀티플레이)일수록 커뮤니티가 활성화되어 리뷰 작성이 활발함을 알 수 있습니다.

핵심 메시지

"흥행의 3요소(가격, 인기, 재미)는 서로 독립적으로 움직입니다. 비싼 가격이 재미를 보장하지 않으며, 인기가 많다고 해서 모두가 호평하는 것은 아닙니다."

가격 정책과 품질의 관계 - 그래프

Scatter Plot Analysis

가격 정책과 품질의 관계 - 해설

Scatter Plot Analysis

시각화 설명

그래프: 가격(X축)과 평점(Y축)의 산점도. (투명도 조절로 밀집도 표현)
보조선: $60 (AAA 게임 표준 가격선) 표시.

상세 분석 내용 (Key Findings)

가격의 표준화 (Barcode Pattern):
데이터가 $4.99, $9.99, $19.99, $59.99 등 특정 수직선에 몰려 있다. 이는 게임 시장에 강력한 심리적 가격 정책(Psychological Pricing)이 작용하고 있음을 시각적으로 보여줍니다

저가 시장의 고변동성 (High Variance)

$20 미만 구간에서는 점들이 0점부터 100점까지 꽉 차 있다. 이 구간은 '가성비 명작'과 '디지털 쓰레기'가 혼재된, 소비자에겐 '복불복(Gamble)'의 영역임을 뜻한다.

고가 시장의 안전망 (Low Variance)

$60 라인을 보면, 평점 0~40점대의 하위 구간에 점이 거의 없다.
해석: 비싼 게임이 무조건 90점을 받는 건 아니지만(상관계수가 낮은 이유), 적어도 망작이 되지는 않는다. 자본이 투입된 게임은 QA(품질보증)를 통해 '최소한의 품질(Quality Floor)'을 보장한다는 것이 증명되었습니다.

성공 요인 4분면 분석

Bubble Chart Analysis

성공 요인 4분면 분석

Bubble Chart Analysis

기초 설명

그래프: 평점(X)과 리뷰 수(Y, 로그 스케일)를 축으로 하는 버블 차트.
점의 속성: 크기 = 가격 / 색상 = 장르.
기준선: 평점 70점(수작 기준), 리뷰 1,000개(흥행 기준).

1사분면 (우상단, Blockbuster):

Action 장르(파란색)가 압도적으로 많습니다. 대중성과 작품성을 모두 잡으려면 액션 장르가 가장 유리하지만, 그만큼 경쟁이 치열한 레드오션입니다.
단, 액션이라는 장르 자체가 매우 흔히 사용되는 태그인것은 고려되야합니다.

4사분면 (우하단, Hidden Gems):

평점은 90점대인데 리뷰가 1,000개 미만인 구간에 Indie(빨간색)와 Adventure(초록색)가 집중되어 있습니다.
해석: 게임성은 훌륭하지만 마케팅 부족이나 장르적 한계로 대중에게 발견되지 못한 '비운의 명작들'입니다.

하단부 패턴 (Statistical Artifacts):

그래프 하단의 물결무늬는 리뷰 수가 극히 적을 때(1~10개) 나타나는 수학적 패턴입니다. 이 구간의 '평점 100점'은 표본 부족으로 인해 신뢰할 수 없는 데이터임을 시각적으로 확인시켜 줍니다.

핵심 메시지

액션 장르는 히트가 된다면 대박이지만 경쟁이 치열한 레드오션이고, 인디게임은 작품성은 뛰어나지만 '발견 가능성' 문제를 해결하는 것이 열쇠입니다.
다만 1사분명에서 서술하였듯, 액션 장르 자체가 매우 흔히 사용되는 태그인것은 고려되야합니다.

최종 결론

Conclusion

평점 인플레이션

발견: 대부분 장르의 평점 중앙값이 70점 이상이며, 이상치(Outlier)가 하위 구간(0~20점)에만 집중됨.
상세: 스팀 유저는 기본적으로 관대합니다. 따라서 평점 70~80점은 '성공'이 아니라 '평균'입니다.
결론: 스팀 시장에서의 경쟁력은 '좋은 평점'을 받는 것보다, '압도적으로 나쁜 평가(하위 이상치)'를 피하고 리뷰 수를 확보하는 것에서 갈립니다.

가격의 역할

발견: 가격과 평점의 상관계수는 0.08로 사실상 무관함.
상세: 고가(AAA급) 게임이 저가(Indie) 게임보다 반드시 재미있는 것은 아닙니다.
반전: 단, 산점도 분석 결과 $60 이상의 고가 게임은 평점 0~40점대의 실패작이 거의 존재하지 않습니다.
결론: 높은 가격은 '재미의 보증수표'가 아니라, '품질 실패를 막는 안전망(Safety Net)' 역할을 수행합니다. 소비자는 높은 가격을 지불함으로써 '망작을 피할 확률'을 구매하는 셈입니다.

장르별 성공 방정식의 차이

발견: Action: 평점과 리뷰 수가 모두 높은 '블록버스터' 구역 점유. (High Risk, High Return)
Indie: 평점은 최상위권이나 리뷰 수가 적은 '숨겨진 보석' 구역 점유.
결론: 액션 게임은 '경쟁'이 치열하고, 인디 게임은 '발견'이 어렵습니다.

데이터의 한계 (Data Limitation)

무료 게임(Free to Play)의 수익 모델 미반영: 본 분석은 판매 가격만 고려했기에, 무료 게임 내부의 인앱 결제의 수익성을 파악하지 못했습니다. 흥행을 '매출'이 아닌 '리뷰 수'로 추정한 점에 한계가 있습니다.

생존자 편향 (Survivorship Bias): 스팀 스토어에 등록된 게임만 분석했으므로, 개발 도중 취소되거나 스토어에서 내려간 게임들의 데이터는 누락되었습니다.