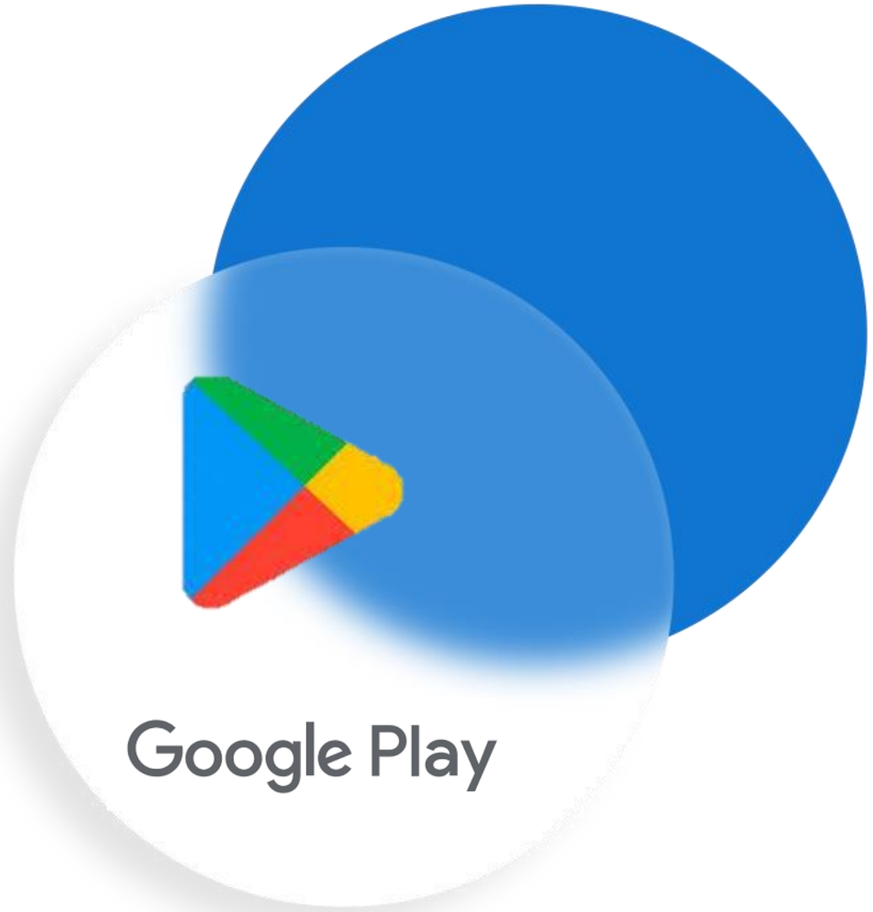
데이터덕후들

Team C2

2025.11.24

다운로드 수, 평점, 업데이트, 가격정책을 중심으로 본

앱의 성공 공식



- 팀장 : 박의진
- 팀원 : 고명희, 박정우, 이윤화
- 그 외 도움 주신 분들 : Claude Gemini

모두의연구소

데이터덕후들

Contents

발표 순서

1 프로젝트 개요

2 데이터 및 분석 과정 설명

3 EDA 테마별 분석 결과

(이윤화)
카테고리별 성공요인
평점 및 리뷰참여율 분석

(박정우)
다운로드 패턴
업데이트주기와 앱나이 분석

(고명희)
가격정책 영향 분석
앱사이즈 분석

4 모델링 결과와 종합 결론

5 프로젝트 회고

데이터덕후들

Information

프로젝트 개요

📊 프로젝트 소개

- 활용한 데이터셋은 2021년 6월 기준, 웹 크롤링을 통해 수집된 **Google Play Store 앱들의 정보**임.
- 본 분석팀은 약 230만개의 해당 데이터로부터 **앱의 성공 요인을 파악하기 위해 “앱의 다운로드 수”, “평점”, “업데이트 및 가격 정책” 등의 측면에서 분석을 시행**하였고, 이를 통해 비즈니스에 유용한 인사이트를 도출하고자 함.

✅ 왜 우리는 이 데이터셋을 뽑았는가?

- 일반인들에게도 **친숙한 도메인**이기에 초보자 수준에서 첫번째 팀 프로젝트로의 도전이 수월할 것으로 판단함.
- 비전공자들의 모임이라 **역지사지의 정신으로 협동이 가능**하며, 각기 다른 경험을 기반으로 참신한 관점에서 시장 구조를 들여다보는 힘이 있을 것으로 기대함.

데이터덕후들

Problem

문제 도출 배경

"경쟁은 치열한데, 성공의 기준은 없다."

성공 공식?

Problem. 1

치열한 경쟁구도의 시장

- 기술의 발달로 시장 진입 장벽이 낮아짐.
- 현재 수백만 개의 앱이 존재하며 유사 앱 만연.
- 누구나 앱을 만들 수 있지만 실제 의미있는 다운로드를 얻는 앱은 극소수임.

선택 기준?

Problem. 2

사용자 판단 요소 증가

- 공개되는 데이터가 많아지면서 사용자 행동에 영향을 미치는 요소들이 증가함.
- 유사 기능 모델들이 쏟아지면서 소비자의 사용 및 구매를 결정하는 선택 기준들이 명확해짐.

판단 근거?

Problem. 3

비즈니스 의사 결정의 지표 필요

- 신제품 개발 또는 개선 활동과 관련하여 객관적이면서, 실행 시 결과 예측 가능한 근거 자료가 필요함.
- 비즈니스 과정에서 투자 및 전략 판단 시 의사결정 지표 요구됨.

데이터덕후들

GOAL

분석 목표

문제 정의

도전하는 사람은 많지만 실질적으로 성공하는 앱은 극소임

분석 목표

성공하는 앱들의 공통 패턴을 찾아
비즈니스에 적용 가능한 핵심 인사이트를 도출

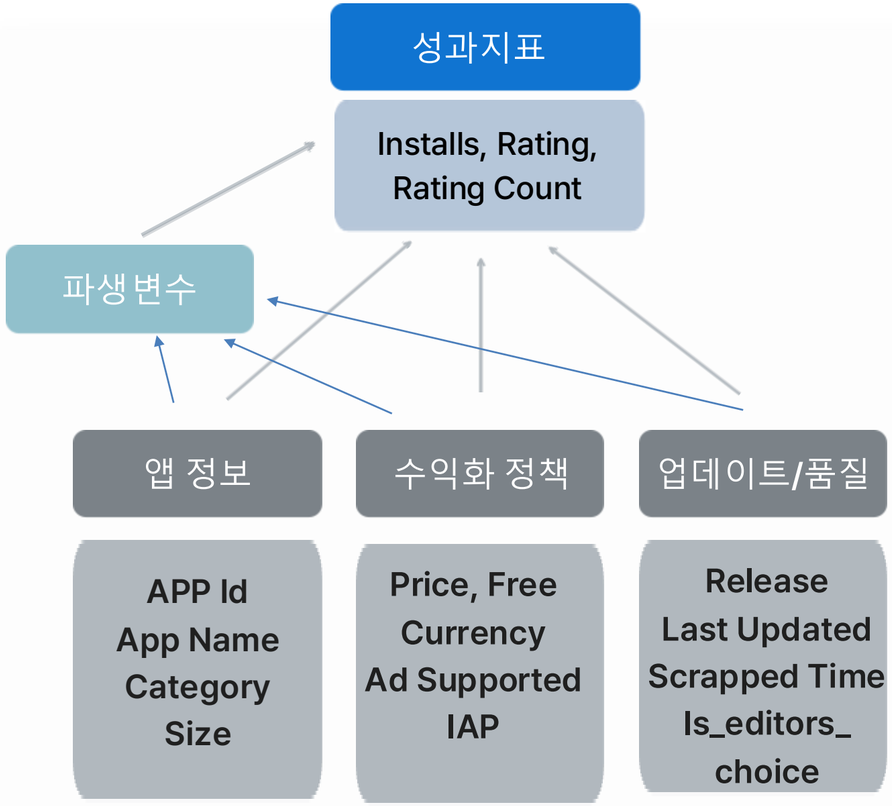
가설 설정

1. 카테고리별로 성공 가능성이 더 높은 그룹이 존재할 것이다.
2. 평점이 높으면 다운로드 수도 높을 것이다.
3. 자주 업데이트 하는 앱이 평점이나 다운로드가 높을 것이다.
4. 가격, 인앱결제와 같은 수익화 방식이 다운로드수, 평점에 영향을 준다.
5. 용량은 앱의 성공과는 큰 연관성이 없고 카테고리에 영향을 받을 것이다.

데이터덕후들

DATA

데이터 설명



데이터 구조 파악

- 데이터 개수 : 2,312,944 개
- 컬럼수 : 24개
- 결측 컬럼 :
  Developer Website (33%), Privacy Policy(18%), Released (3%)
  Rating (1%), Rating Count(1%) ...

전처리 핵심 포인트

- 분석, 모델용 '핵심 변수 위주'로 정리
- 결측, 이상치 처리로 '신뢰 가능한 변수' 확보
- 의미있는 파생변수로 '설명력' 강화
- 중복 및 타입 정리로 모델링 준비
- 핵심 변수 처리 규칙을 정한 뒤 나머지는 자유롭게 처리함으로써 전처리 데이터 통합 효율화

데이터덕후들

GOAL

분석과정 요약

1차 산출물
Preprocessed Data
Preprocessed Notebook

2차 산출물
EDA Notebook
EDA PPT

3차 산출물
Final Notebook
Final PPT

Modeling input data
Modeling Notebook

STEP 1
데이터의 이해 및 준비

- ✓ 데이터 구조와 컬럼 파악
- ✓ 분석 방향 설정
- ✓ 불필요 컬럼 제거
- ✓ 결측, 중복 처리

STEP 2
전처리와 파생변수

- ✓ 자료형 변환, 날짜 파싱
- ✓ 로그, 구간 파생변수
- ✓ 날짜 기반 파생변수
- ✓ 이상치 점검 및 클리핑

STEP 3
테마별 EDA 진행

카테고리, 평점, 다운로드,
업데이트 주기, 가격 정책,
앱사이즈 관점에서 데이타 패턴
및 상관관계 분석, 시각화

STEP 4
인사이트 요약 및 모델링

- ✓ 성공요인 도출
- ✓ 머신러닝 후보 테스트
- ✓ 최종 모델링 및 결과 분석

DAY 1

DAY 2

DAY 3

데이터덕후들

GOAL

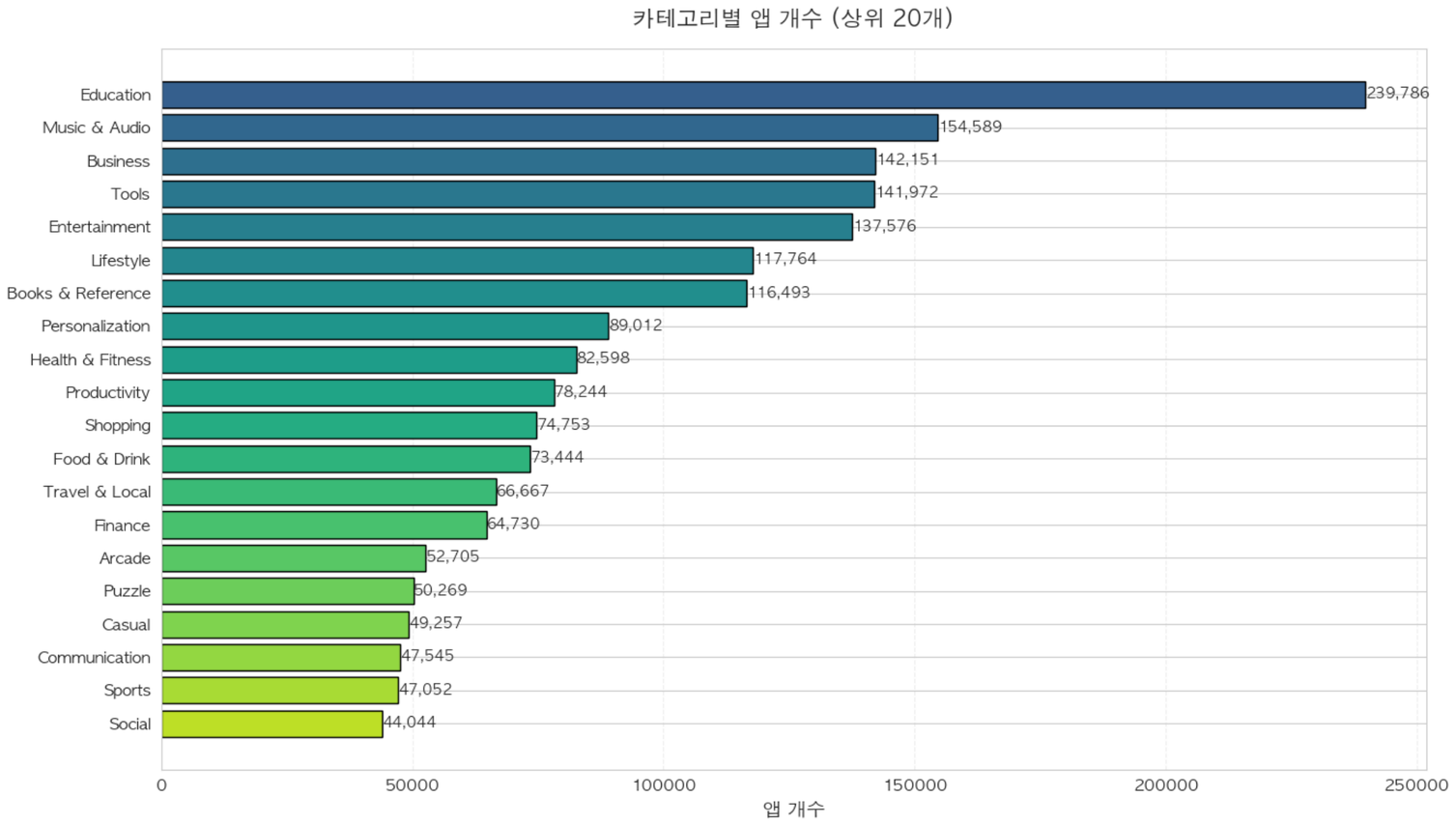
앱의 성공 요인 탐색을 위한 분석 흐름도

"앱 시장 구조를 카테고리 관점에서 이해"

데이터덕후들

EDA-1

시장 구조_카테고리별 앱 개수



핵심 메세지

앱 개수가 많은 카테고리
= 경쟁이 치열한 시장

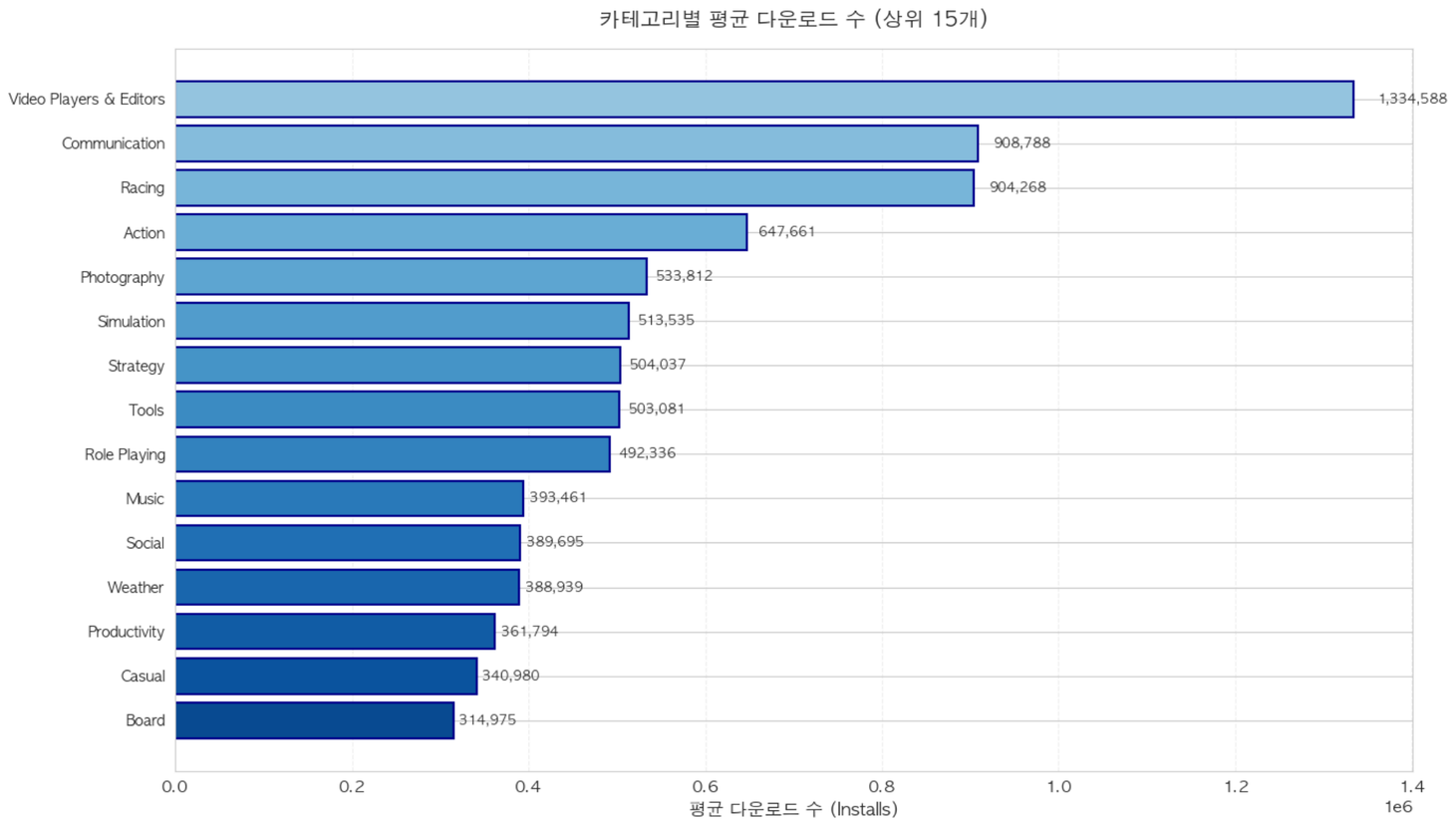
'구글플레이 앱시장'은 **극단적 불균형 구조**를 가짐.
총 48개 카테고리가 존재하며,
교육, 음악·오디오, 비즈니스, 툴, 엔터테인먼트 5개 정도가 전체의 절반이상을 차지함.
경쟁자수가 많기 때문에
다운로드를 얻기 위해서는 평점, 기능, 업데이트 빈도 등 차별화 요인이 있어야 할 것임.

데이터덕후들

"경쟁력 및 수익성 있는 카테고리는? "

EDA-1

시장 구조_카테고리별 앱 평균 다운로드 수



핵심 메세지

평균 다운로드 상위권 카테고리 = Video Editors / Communication / Racing / Action 등

평균 다운로드 수가 높은 상위 카테고리는 사용 목적이 명확하며 글로벌 수요가 큰 특징이 있음

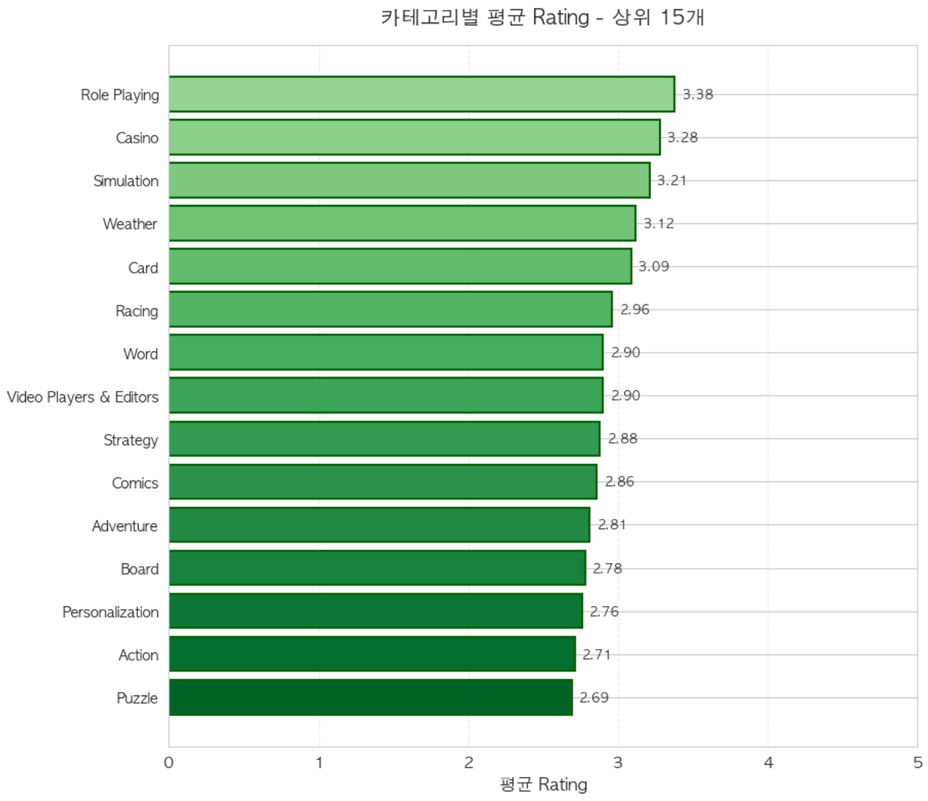
개발하려는 앱의 카테고리에 따라 예상 평균 다운로드 수를 예측할 수 있으며, 이는 해당 카테고리의 경쟁력과 수익성을 보여줌

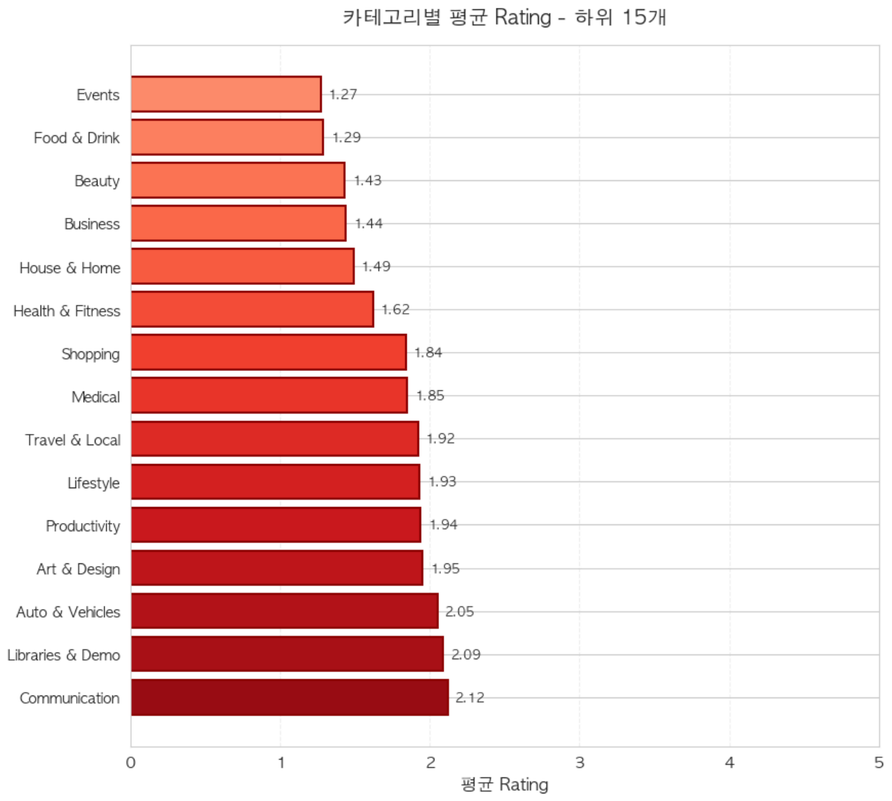
데이터덕후들

“사용자 만족도가 좋은 카테고리는? ”

EDA-2

사용자 반응_카테고리별 평점





핵심 메세지

게임류 앱은 만족도가 상대적으로 높고 생활형 및 서비스형 앱은 낮음

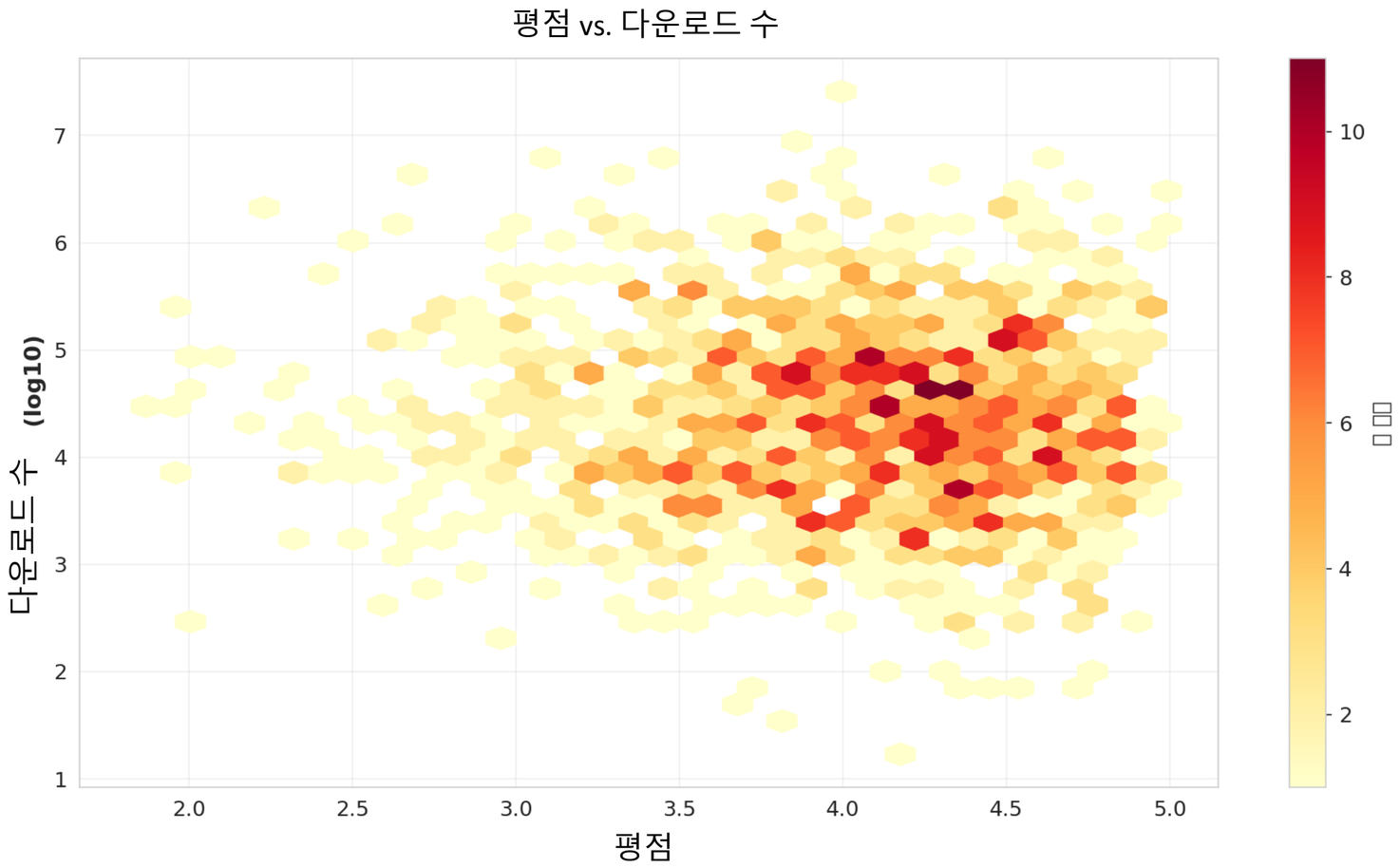
만족도 상위 카테고리는 ‘게임/엔터테인먼트 기반 카테고리(Role Playing, Casino, Simulation)가 차지하고 있음. 반면에 **Events, Food & Drink, Beauty 등의 현실 문제 해결형 앱들은 낮은 평점**을 보여 기대대비 품질 미달 가능성을 생각해볼 수 있음.

데이터덕후들

“평점과 다운로드간에는 약한 상관관계 존재”

EDA-2

사용자 반응_평점과 다운로드의 상관관계



핵심 메세지

평점과 다운로드 수 사이에는 약한 상관관계를 보임

대부분의 앱이 평점이 높음.

하지만, **높은 평점이 항상 많은 다운로드를 의미하는 것은 아님.**

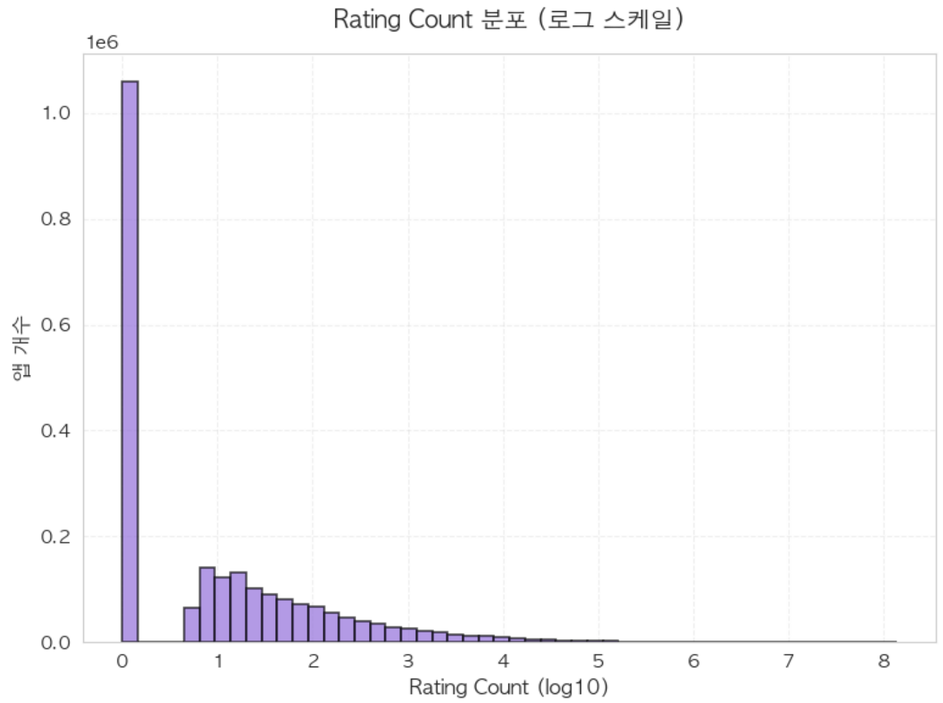
구글플레이 앱스토어 정책상 평점은 전체 기간 산술평균이 아닌 최신 평점에 더 큰 가중치를 두고 있는데 이는 앱의 최근 품질을 반영하기 위함이라고 함.

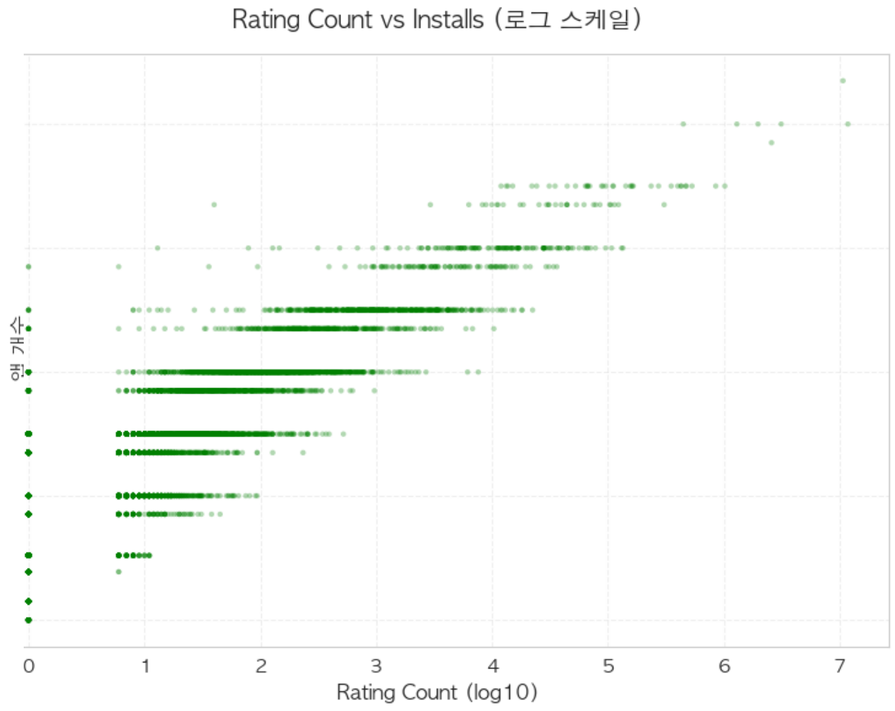
데이터덕후들

"사용자 반응 패턴에서 앱 선택 기본 조건 파악"

EDA-2

사용자 반응 _리뷰 수와 다운로드 수의 상관관계





핵심 메세지

리뷰 수와 다운로드 수는 명확한 우상향 관계

대부분 앱에서 리뷰가 없고, 많은 리뷰를 받는 앱은 소수인 상황임

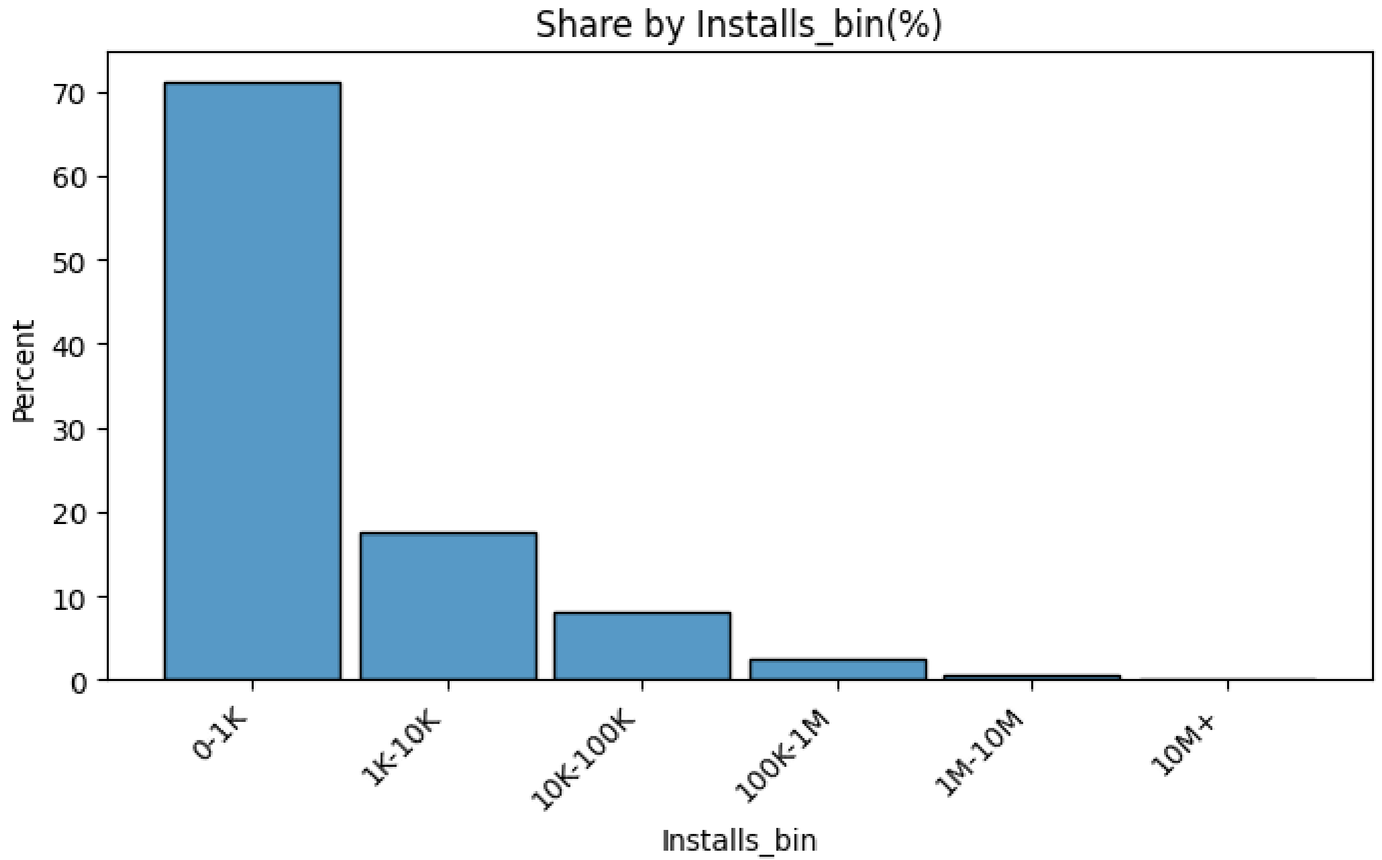
앞에서 보다시피 대부분의 평점이 3.5-4점 대이며 리뷰 개수와 다운로드 간에는 양의 상관관계를 지니므로 **리뷰가 많으면 다운로드할 확률이 높을 것**으로 예상됨.

데이터덕후들

"전체 다운로드 시장의 구조적 특징은?"

EDA-3

다운로드 분포



핵심 메세지

앱 다운로드 분포는
극단적 롱테일 구조임

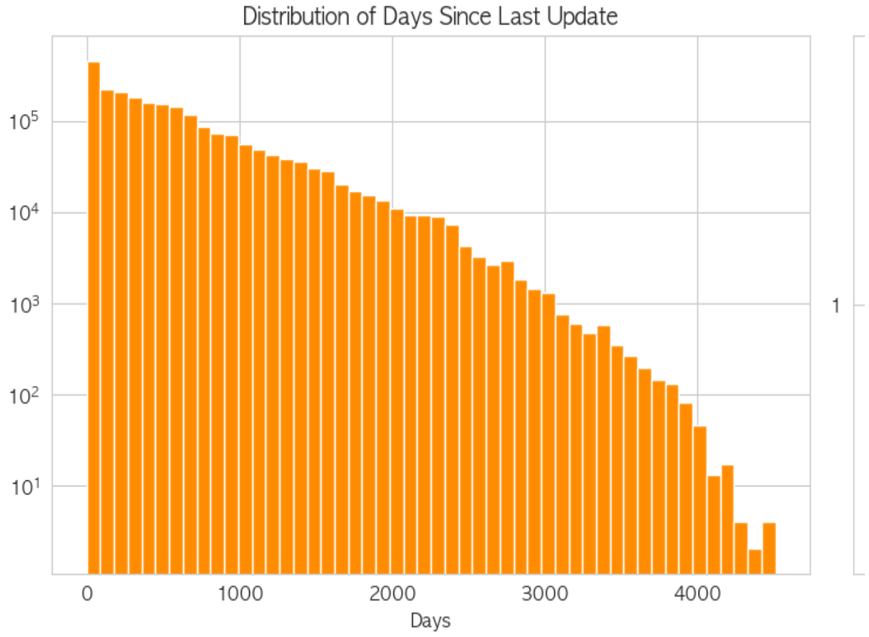
전체 앱의 70% 이상이 다운로드 1000개 이하 구간에 몰려 있음.
다운로드가 많이 되는 상위 앱은 극소수임.
즉, **대부분의 앱은 사용되지 않고, 상위 앱에 수요가 집중**되는 winner-take – all 구조가 앱 시장의 본질임.
결국 시장진입 장벽은 낮지만 성공 가능성은 희박함을 극명히 보여줌.

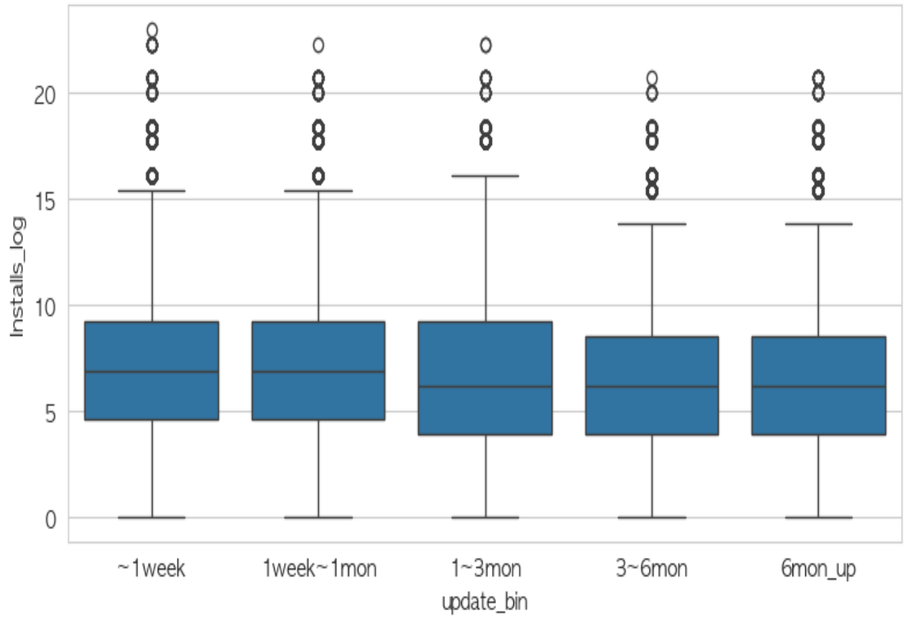
데이터덕후들

EDA-4

"부지런한 관리가 다운로드와 평점에 미치는 영향은? "

운영(업데이트) 전략





핵심 메세지

업데이트 주기가 커지면 다운로드 소폭 감소 경향

대다수의 앱들은 최근 6개월 이내 업데이트 하며, 가장 업데이트가 활발한 카테고리는 Weather임.

구글 플레이스토어 **정책상 정기적인 업데이트를 권장하며 장기간 방치 시 정책 위반으로 간주**될 수 있다고 함.

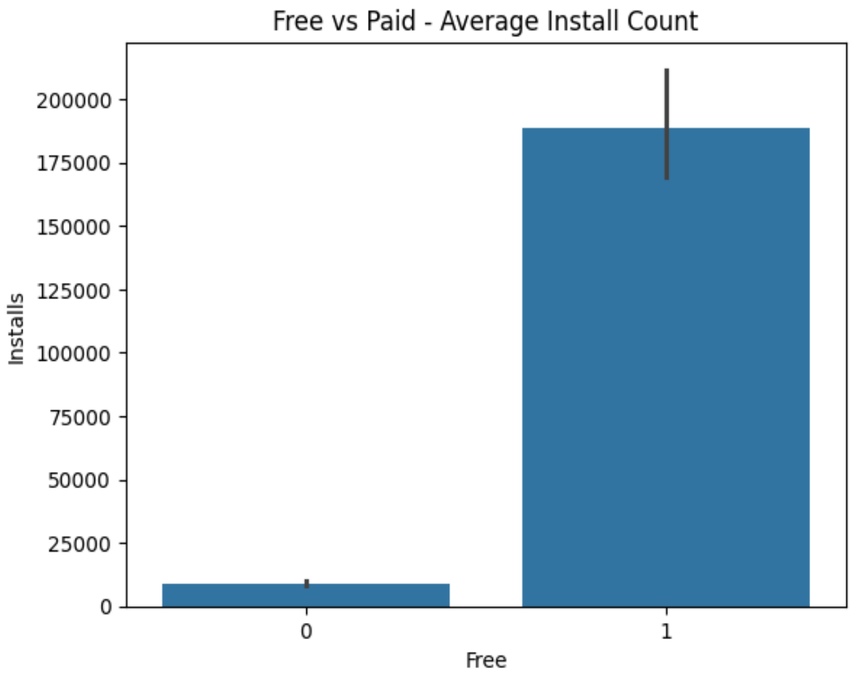
업데이트 주기가 짧을 수록 다소 높은 다운로드 경향을 나타내긴 하지만 1주와 6개월 비교 시 중앙값에 큰 차이는 없음

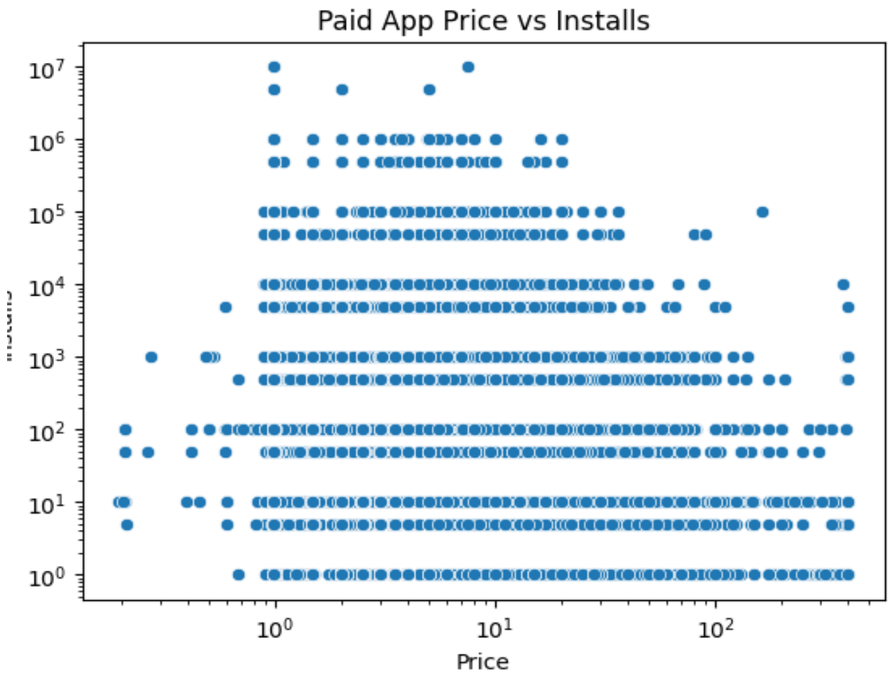
데이터덕후들

"수익화 전략이 성과 지표에 어떤 차이를 가져올까?"

EDA-5

수익화 전략_유료와 가격의 영향





핵심 메세지

무료앱은 다운로드수를 높이는 확실한 전략이지만 유료앱에서의 가격 차이는 다운로드 수에 영향이 없음

무료앱이 대다수이며 다운로드와 강력한 상관관계를 가짐.

하지만 유로앱에서의 가격이 높고 낮음의 정도는 다운로드 수에 영향을 미치지 않음.

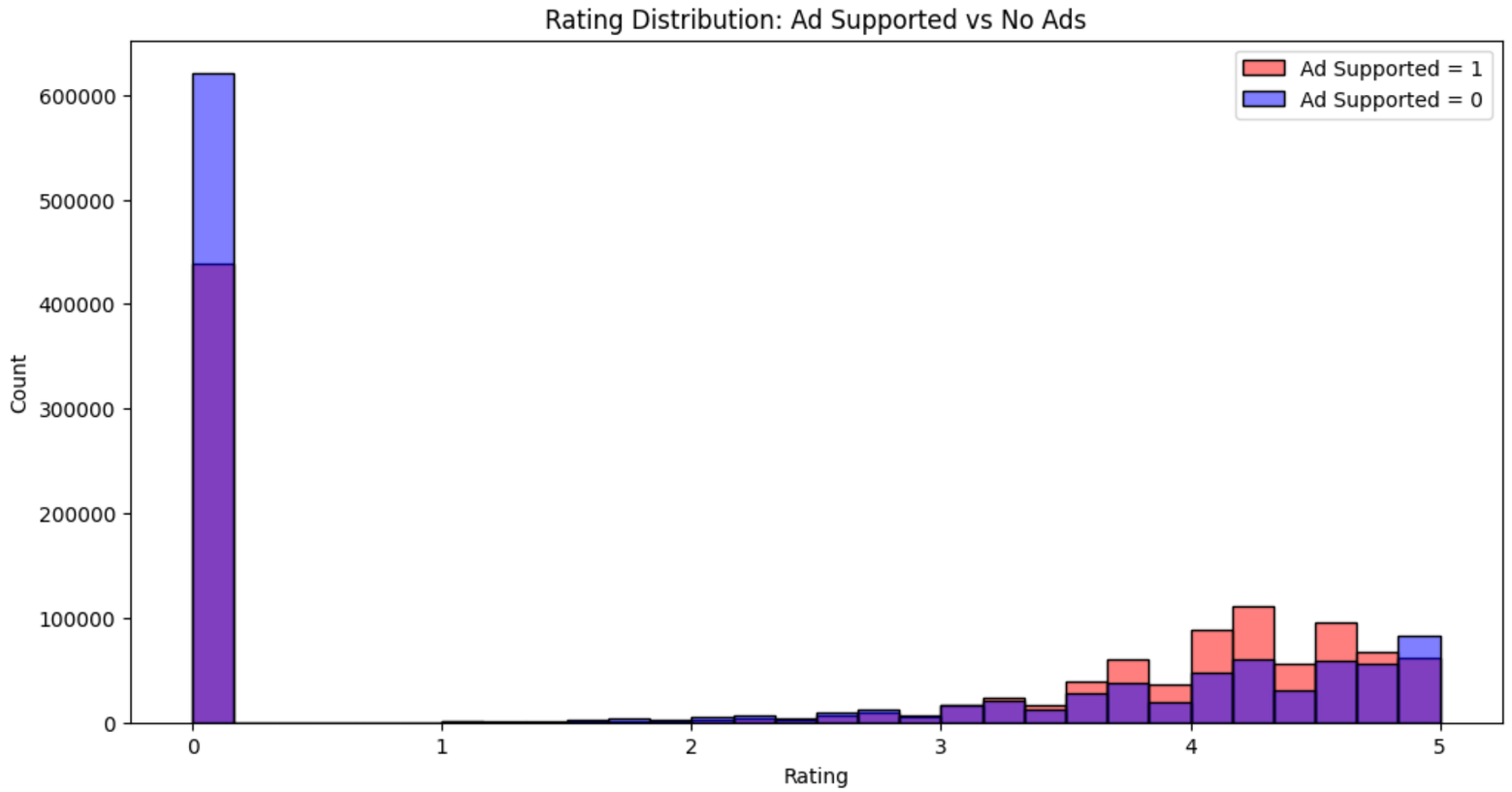
따라서 **유료앱의 성공 요인은 가격보다 다른 요인일 가능성 높음.**

데이터덕후들

"광고의 유무는 사용자 만족도를 떨어뜨릴까?"

EDA-5

수익화 전략_광고 유무



핵심 메세지

고평가되는 구간에서는
광고가 없는 앱이 많음

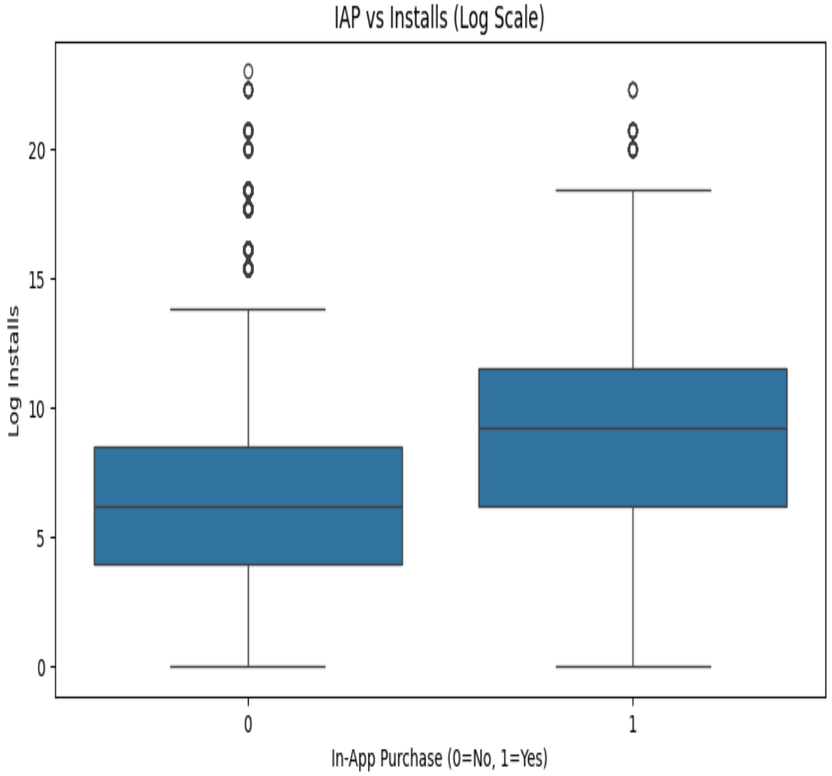
평점이 없는 0점 구간을 제외하고 3-5점 구간을 보면 광고있는 앱이 더 많이 분포하고 있음을 알 수 있음.
이는 **광고가 사용자 경험의 질을 떨어뜨릴 수 있다는 추측**을 할 수 있음.
다만 다운로드와의 상관관계는 EDA 과정에서 누락되어 확인하지 못함.

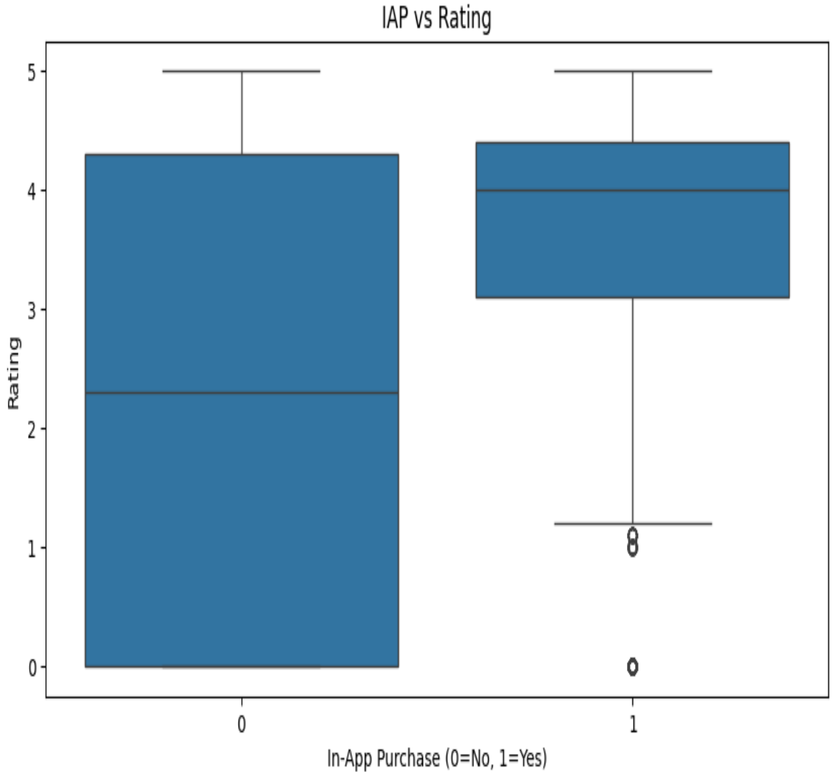
"수익화 전략이 성과 지표에 어떤 차이를 가져올까?"

데이터덕후들

EDA-6

수익화 전략_인앱결제 기능 포함





핵심 메세지

인앱결제 기능 유무는 앱 성공에 강한 영향을 미침

인앱결제 기능이 있는 앱은 다운로드 수에서도 더 높은 성과를 보이고, 평점에서도 안정적으로 높은 중앙값을 유지하는 것이 관찰됨.

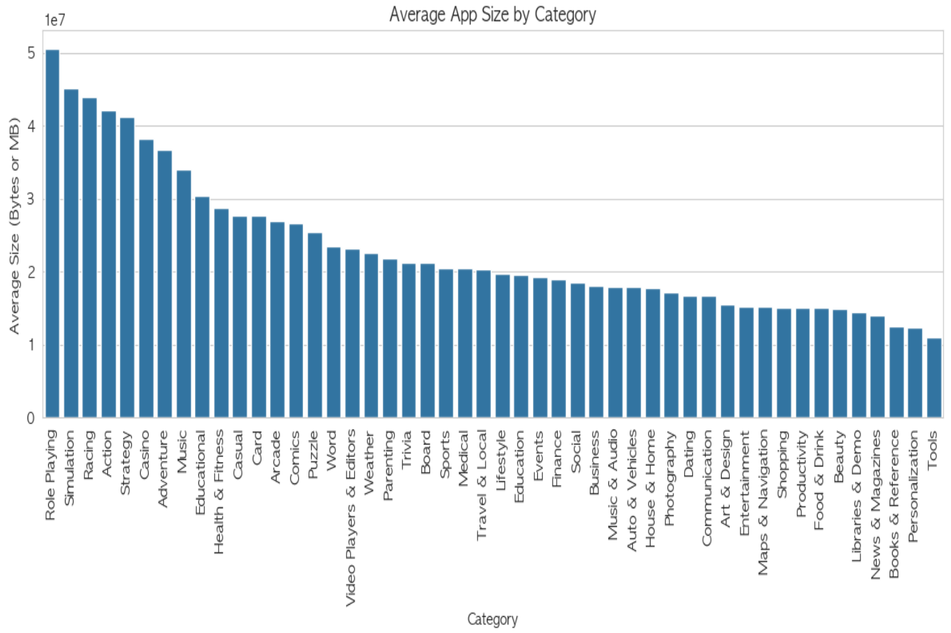
이는 **인앱결제 기능이 앱 성공요인에 강력한 영향 변수**이며, 이러한 시스템을 갖추려면 개발역량, 운영 투자, 콘텐츠 관리 기반을 갖춰야 하므로 **높은 품질 활동이 높은 성과로 이어질 가능성을 시사함.**

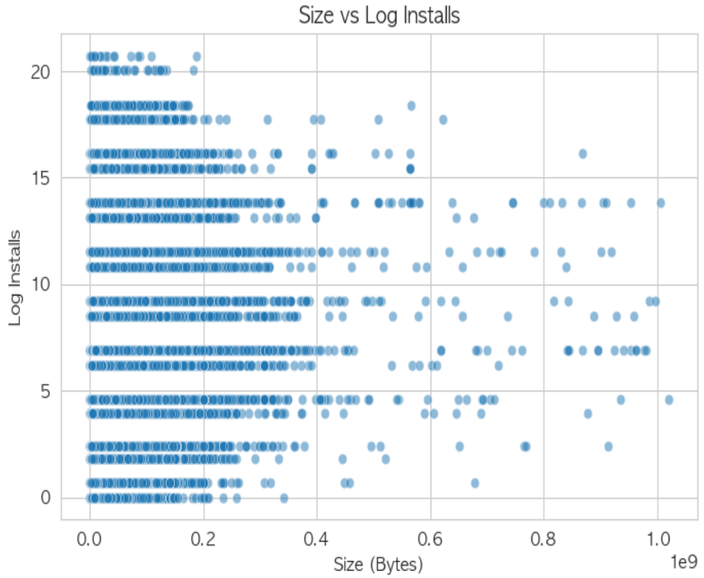
데이터덕후들

"앱 사이즈 같은 특성도 성과 지표랑 관계가 있을까?"

EDA-7

기술적(앱용량) 특성





핵심 메세지

앱 용량은 카테고리 성격에 따라 결정되며, 다운로드 수와는 직접적인 상관관계는 없음

앱 용량은 다운로드 수에 영향을 미치지 않으므로 용량을 줄이거나 늘리는 것이 **앱의 성공 여부에 큰 영향을 주지 않을 것으로 예상함.**

다만 **카테고리에 따라 용량이 달라짐.** 게임, 영상, 지도앱 같은 고사양의 경우 앱 용량이 크고 유틸리티 텍스트 기반은 가장 가벼움.

데이터덕후들

Modeling

모델링_트리기반모델(R/F)

다운로드 예측을 위한 모델링

앱의 설치 수가 극단적 롱테일 구조라서
install 값 대신 **installs_log를 예측**하기로 함

모델링 개요

목적 : 다운로드를 높이는 핵심 요인은 무엇인가?

전처리 핵심 :

- 해석력 좋게 작동하도록 로그변환으로 정규성 완화 (극단값 완화)
- 카테고리 고유값 많아서 Target Encoding 적용
- 파생변수 생성 (날짜 기반, 수익화 구조 기반, 업데이트 패턴)

설계 시 주의사항 :

- 타깃누수 방지 (install 관련 파생 변수 제외), 환율 USD로 통일

모델 선정 이유

1. RandomForestRegressor

- 일단 우리는 어떤 앱이 성공, 실패인지 이진분류가 목적은 아니므로 분류모델은 최종 발표용에서 제외
- 선형회귀 (Linear Regression)을 혼합 분석하여 해석력을 높였으면 더 좋았겠지만 변수의 중요도 도출이 주목적이라 시간 관계상 1개의 모델만 선택 시행
- 본 데이타는 로그 변환 후 선형/비선형의 관계 모두 존재하고 상호작용이 강한 변수들이 존재함. 또한 범주형 변수가 많고 (48개 컬럼) 200만 이상의 대규모 데이터이므로 안정적 성능을 지닌 본 모델을 최종 선정

"평점, 리뷰, 앱 관리가 앱 다운로드의 승패를 가른다."

데이터덕후들

Modeling

모델링 결과_회귀 모델

지표	Baseline	모델	개선폭	의미
RMSE	3.109	1.361	**-56%**	큰 오차를 절반 이하로 감소
MAE	2.531	1.048	**-59%**	평균 오차 절반 이하로 감소
R^2		0.8		모델 설명력 우수한 편

[지표 의미]

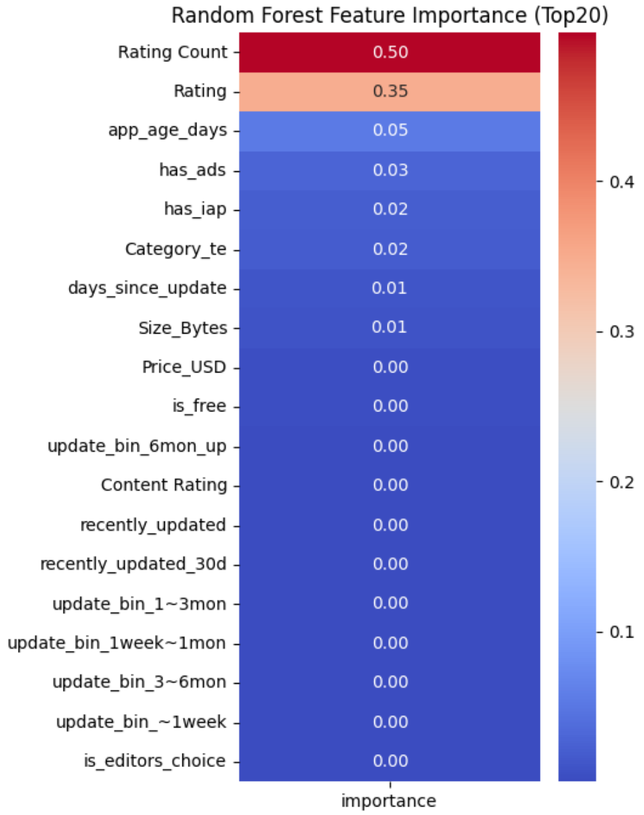
RMSE는 모델 평균 오차 크기이며 값이 낮을수록 정확하다는 뜻임

MAE는 예측값이 실제값에 얼마나 벗어났는지를 볼 수 있는 지표로 값이 작을수록 좋음

R^2 스코어는 해당 모델이 타깃 변동성을 얼마나 잘 설명하는가를 의미함

[성능 해석]

- 모델이 단순 평균 예측보다 더 뛰어남
- RMSE, MAE 모두 약 2.3 ~ 2.4배 성능 향상
- 다운로드를 증가시키는 요인을 효과적으로 학습한 모델이라는 의미



[영향력 변수 순위 해석]

R/F 모델로 도출한 영향력 변수들은 1위로는 리뷰 개수, 2위 평점, 3위는 앱나이 (출시된 지 경과일) 이었음.

해당 변수들은 앱의 신뢰성 및 검증 정도와 직접 연결된 요소들임. 결국 사용자들은 앱 다운로드 결정 시 이미 많은 사람들이 오랜 기간 사용해보고 평가해 준 앱을 더 신뢰하는 경향이 있음을 알 수 있음.

결론적으로 검증된 앱 (리뷰가 많고 고평점일수록) 다운로드 수는 증가한다는 강력한 의사결정 패턴이 존재함.

"앱 성공 공식 = 신뢰 + 관리 + 사용자 경험"

데이터덕후들

EDA-1

종합 결론 및 핵심 인사이트

분석을 통해 도출된 사실

◆ 앱 시장은 극단적 롱테일 구조로 상위 소수앱이 압도적 다운로드를 차지.

◆ 다운로드의 핵심 결정 요인은 평점, 리뷰수, 앱나이임.

◆ 무료전략과 인앱결제 기능은 다운로드 확보에 효과적인 전략이며, 광고는 평점을 떨어뜨리는 요인으로 작용.

비즈니스를 위한 핵심 메세지

◆ "신뢰가 다운로드를 만든다"→ 앱의 유지, 리뷰수, 평점 등 신뢰 확보가 성장의 핵심임.

◆ 무료 +인앱결제 기능 모델이 가장 큰 유저 기반 확보 전략임.

◆ 업데이트 주기 관리는 바로 매출로 연관되지 않더라도 평판유지, 품질관리 차원에서 필수임.

아쉬운 점과 개선 포인트

◆ 평점과 리뷰수의 조합, 업데이트와 앱나이의 조합과 같은 특징 (Featuer Engineering)을 추가하여 예측력 및 해석력 강화

◆ 랜덤포레스트 이외 선형회귀 기반의 계수 해석을 포함하여 설명력 강화

◆ 카테고리, 업데이트, 평점 간의 상호작용 변수를 좀 더 체계적으로 반영

◆ 극상위 소수앱만의 공통 특징 분석도 흥미로울 듯 (분류 모델 사용)

데이터덕후들

KPT

프로젝트 회고

희 (喜)

설렘 반 두려움 반으로 모인 사람들!
생각보다 빠른 결정력과 수월한 진행 과정 속에서
우리 쫌 잘하고 있다? 는 근거없는 우리만의 자신감이
싹트기 시작함.

노 (怒)

첫번째 당황은 바로 데이터 로드부터 시작됨.
각자 작업 환경이 다르고 결과물을 정리하는 방식이
달라서 계속되는 팀장의 좌절감, 그리고.. 속절없이
흐르는 시간에 다들 분노함.

애 (愛)

전처리 및 EDA까지 잘 흘러가서 모델링도 문제
없겠지라고 생각했으나 큰 오해였음.
알 수 없는 오류의 반복! 게다가 최종 취합에서 누락된
내용들이 발견되며 불안과 슬픔이 찾아옴.

락 (樂)

데이터톤의 여정을 거치며 우리 모두는 초보자 수준의
과제 수행이 아닌, 실제 현업에서 사용할 수 있는
인사이트를 만들어냈다는 점에서 미래에 대한 희망을
느끼며 즐겁게 마무리할 수 있었음.

데이터덕후들

KPT

프로젝트 회고

Keep

연습과 반복

- 그동안의 실습자료, 책, 생성형 AI를 총동원해 한 셀 한 셀 채워 나감.
- 오류 없는 코드 생성과 논리적인 분석 결과의 도출을 위해서는 연습과 반복 뿐임.

Problem

개인 vs. 단체

- 혼자 할 때 몰랐던 코드 병합의 어려움을 시행착오를 겪으며 뼈 저리게 체감함.
- 다양한 의견으로 갈렸을 때 가장 합리적인 판단을 위한 기준을 설정하는 것이 쉽지 않았음. 단순 다수결 처리에 의존하진 않았나?

Try

타산지석

- 남의 장점이나 단점도 나에게는 배울점으로 작용함.
- 타인의 관점과 방법론의 통해 나의 개선과 학습에도 도움이 되는 것을 경험함.
- 앞으로도 ‘남들은 어떻게 하는지’에 대한 관찰의 자세를 강화하면 좋을 듯함.

데이터덕후들

Team

팀원분들께 감사를 전하며...

노베이스지만 협동으로 이룰 수 있는 최고치를 보여주자!

- **데이터셋 분석 및 전처리 :** 전원 개별 진행 후 취합
- **추가 변수 생성 및 이상치와 패턴 탐색 :** 아이디어 회의 및 1차 전처리 결과로 도출
- **EDA :** 앱 시장 구조 관련 (이윤화), 앱 품질 관련 (박정우) , 운영 전략 관련 (고명희)
- **모델링 분석 :** 고명희, 박정우
- **취합 및 정리 :** 박의진, 이윤화

데이터덕후들

Come On!!

Q&A