|  |
| --- |
| **프 로 젝 트 최 종 보 고 서**  보고서 및 논문 윤리 서약  1. 나는 보고서 및 논문의 내용을 조작하지 않겠습니다.  2. 나는 다른 사람의 보고서 및 논문의 내용을 내 것처럼 무단으로 복사하지 않겠습니다.  3. 나는 다른 사람의 보고서 및 논문의 내용을 참고하거나 인용할 시 참고 및 인용 형식을 갖추고 출처를 반드시 밝히겠습니다.  4. 나는 보고서 및 논문을 대신하여 작성하도록 청탁하지도 청탁받지도 않겠습니다.  나는 보고서 및 논문 작성 시 위법 행위를 하지 않고, 명지인으로서 또한 공학인으로서 나의 양심과 명예를 지킬 것을 약속합니다.    과 목 : 통계적 데이터 분석  담당 교수 : 오 민 식  학번 및 이름 :  60170559 김 다 진  60171615 강 수 민  60171635 박 종 혁 |

목 차

**1장 문제 정의3**

1. 주제 및 선정 배경3
2. 목적3
3. 일정 및 역할 분담3

**2장 데이터4**

1. 데이터 출처4
2. 데이터 설명4

**3장 데이터 분석 방법** **5**

1. 분석 프로세스5
2. 데이터 전처리5
3. 데이터 EDA9

**4장 데이터 분석 결과11**

1. 분석 (1)11
2. 분석 (2)13
3. 분석 (3)14
4. 결론16

**1장 문제 정의**

1. 주제 및 선정 배경

✓ 주제: Google Play Store 앱 Install 및 Rating에 영향을 미치는 요소 파악

✓ 주제 선정 배경

◼ 추후 앱 서비스 제작 시, 해당 분석 결과가 인 사이트 도출에 도움이 될 수 있다 고 생각.

◼ 충분한 크기의 데이터 (10.840개 rows), 다양한 변수 타입(int, object, float, category)이 존재하여 다양한 분석 방법을 적용하기 용이

2. 목적

✓ 통계적 데이터 분석 강의 수강으로 습득한 다양한 통계 분석 방법을 실제 raw data를

통해 분석하여 분석 역량을 강화

✓ Google Play Store 앱 관련 데이터를 이용하여 앱 Install 횟수 및 Rating에 영향을 미치

는 요소를 분석을 통해 파악

3. 일정 및 역할 분담

✓ 일정

◼ 22.11.23 주제 선정 회의, 분석 아이디어 도출

◼ 22.11.26 프로젝트 계획서 제출

◼ 22.11.28 데이터 확인 및 전처리 관련 회의

◼ ~22.12.2 데이터 1차 전처리

◼ ~22.12.10 데이터 분석

◼ 22.12.11 프로젝트 PPT 제작

◼ 22.12.12 프로젝트 발표

◼ 22.12.14 프로젝트 최종 보고서 제작

✓ 역할 분담

◼ 강수민: 데이터 1차 전처리, 분석, 발표

◼ 김다진: 분석, 발표 PPT 제작

◼ 박종혁: 분석, 최종 보고서 제작

**2장 데이터**

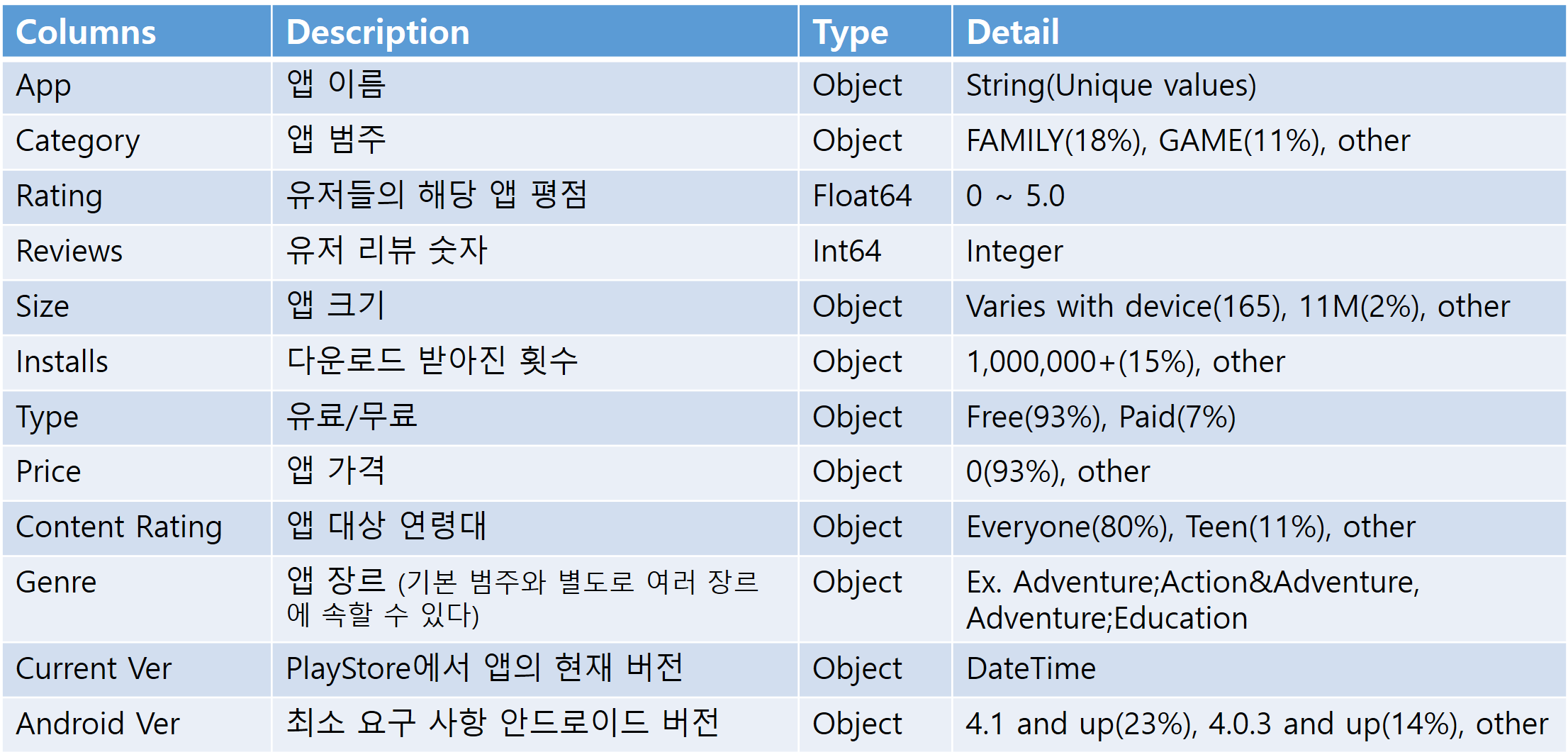
1. 데이터 출처

Kaggle “Google Play Store Apps”

<https://www.kaggle.com/datasets/whenamancodes/play-store-apps>

2. 데이터 설명

총 10,840 rows / 13 columns



**3장 데이터 분석 방법**

1. 분석 프로세스

데이터 수집: Kaggle.com

데이터 전처리 및 시각화: Jupyter Notebook, Python, Seaborn, matplotlib

데이터 분석: Numpy, Pandas, Scipy



2. 데이터 전처리

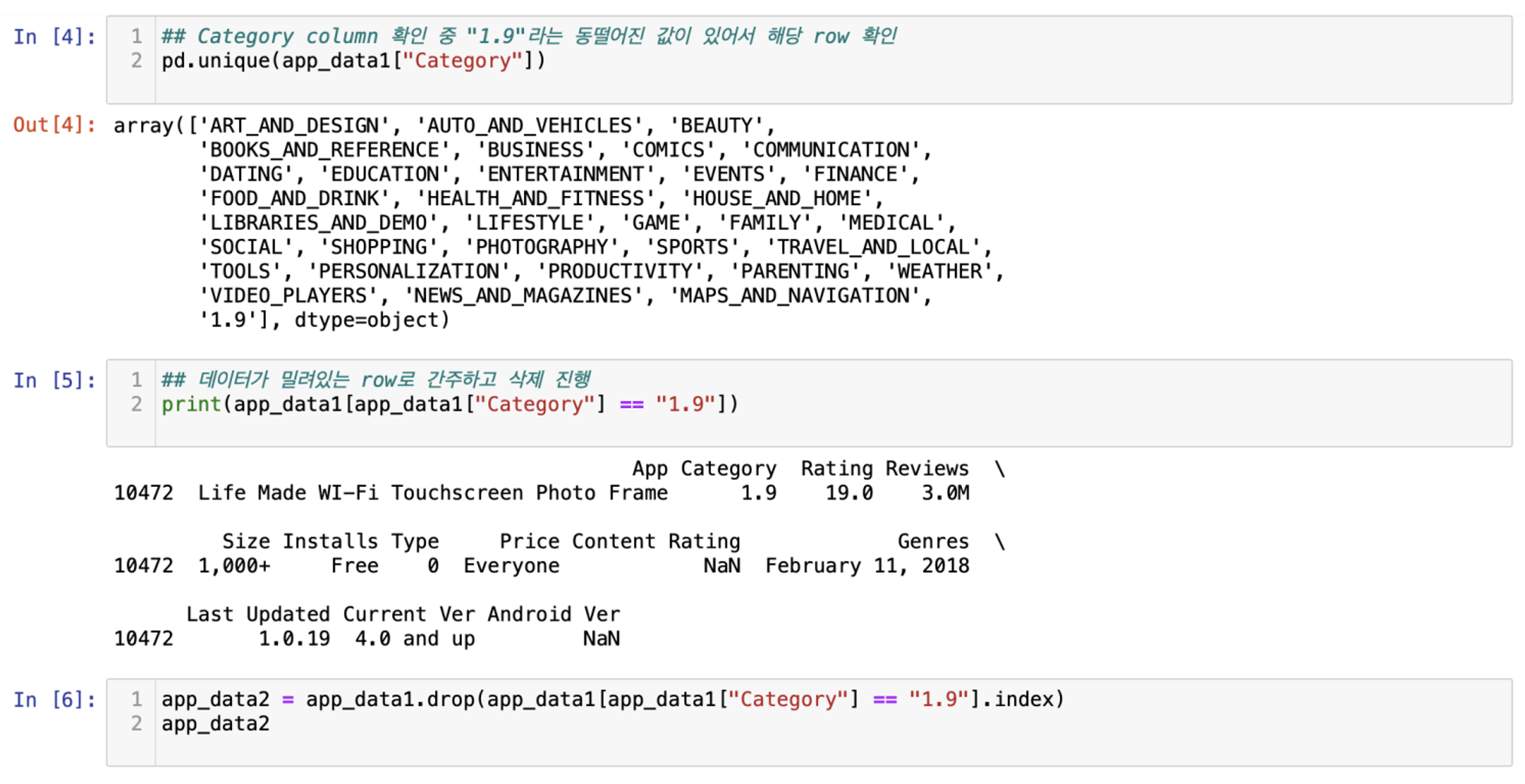
1차 전처리)

✓ NaN값 Drop



1차 전처리)

✓ 값이 밀려있는 행 처리



✓ Reviews/Price/Size 타입 변경

Reviews: object → float

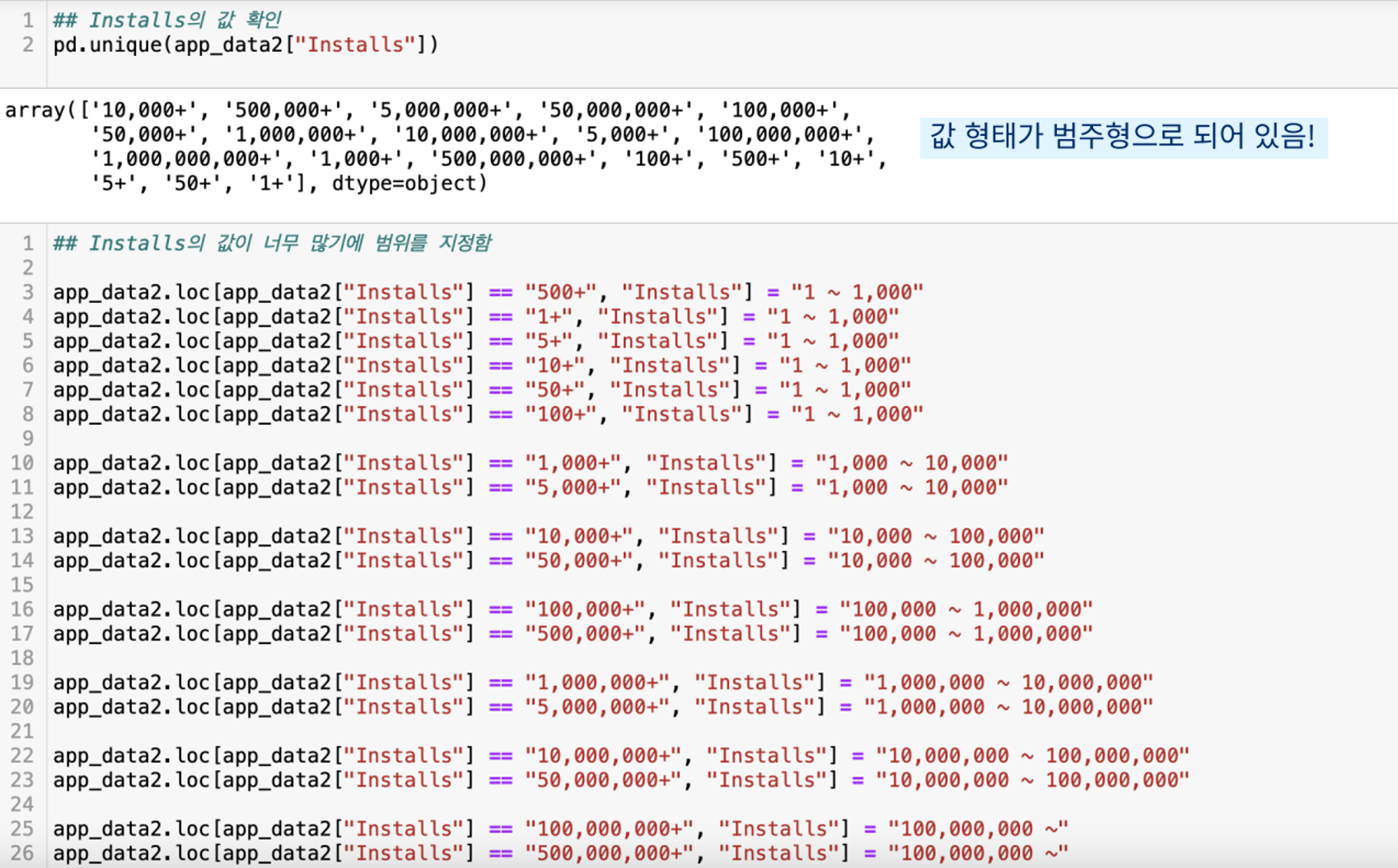
Price: object → float

Size: object → float

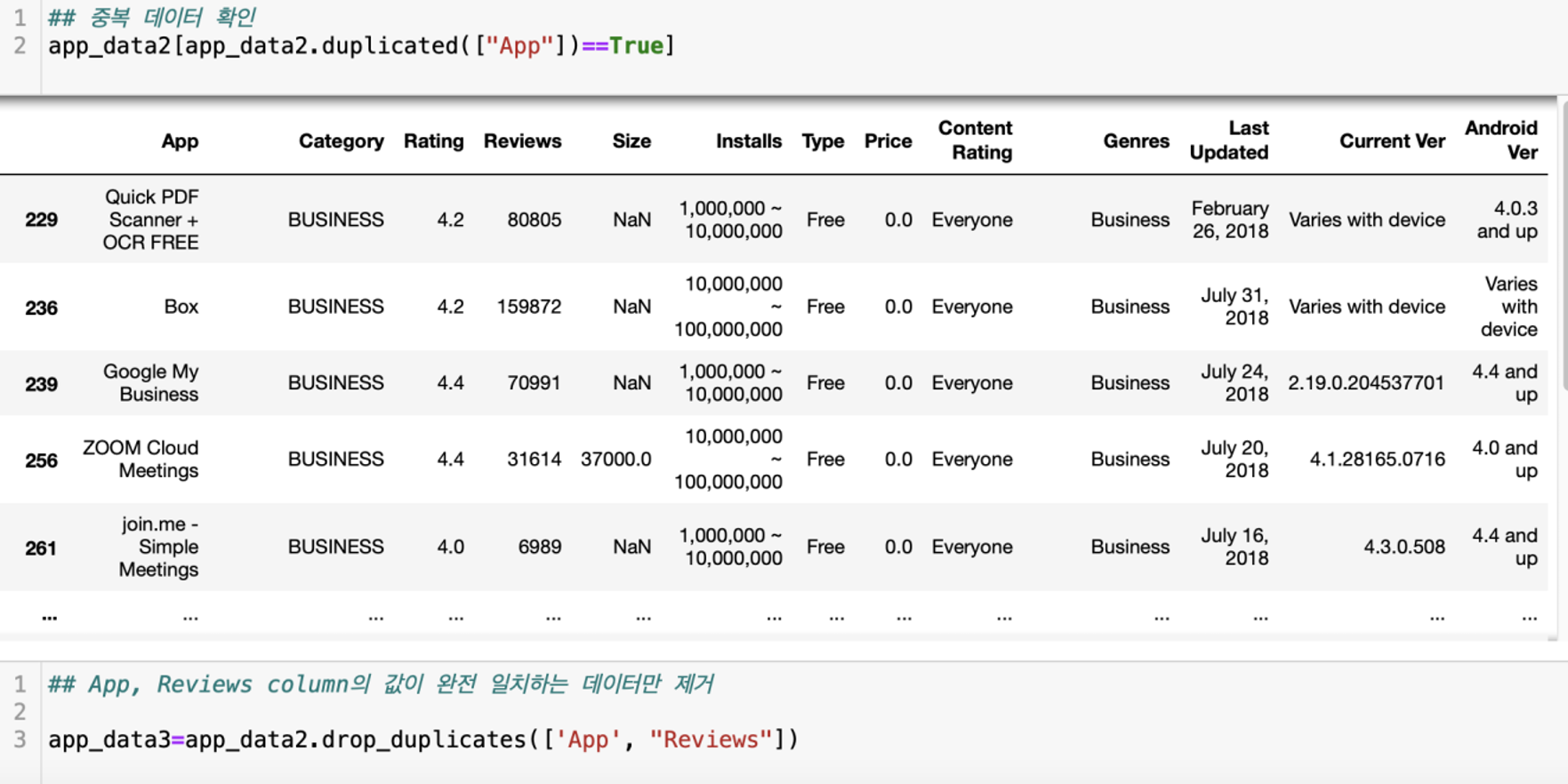


1차 전처리)

✓ Install 범주 축소

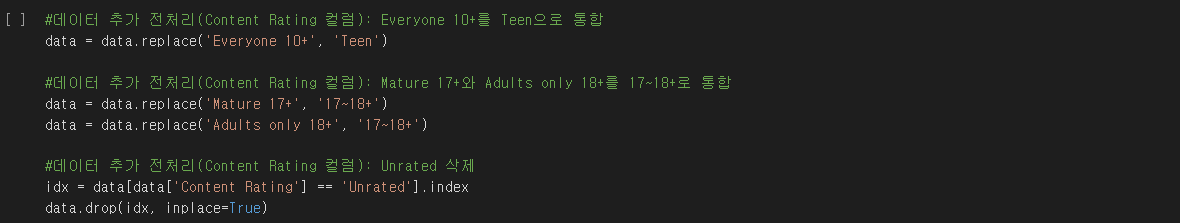


✓ 중복값 제거

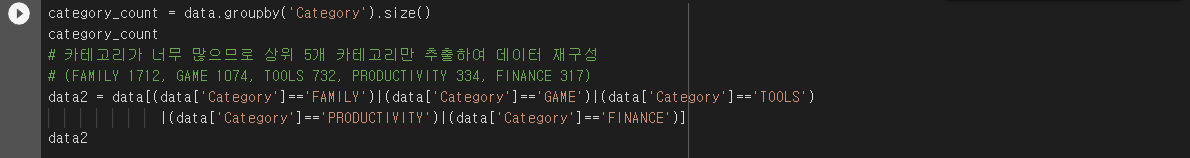


추가 전처리) 분석(1)에 활용됨

✓ Content Rating 범주 수정

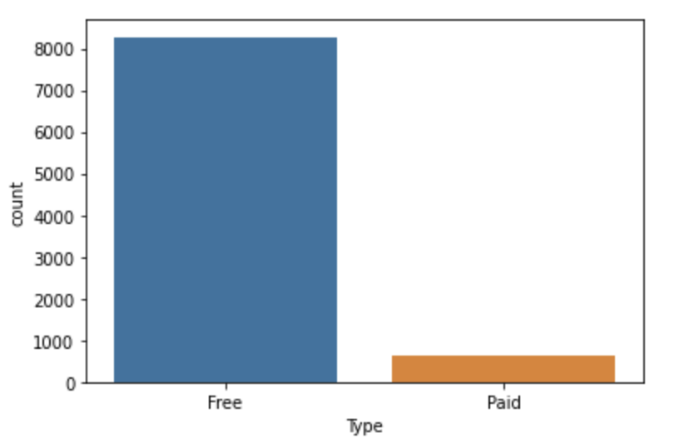


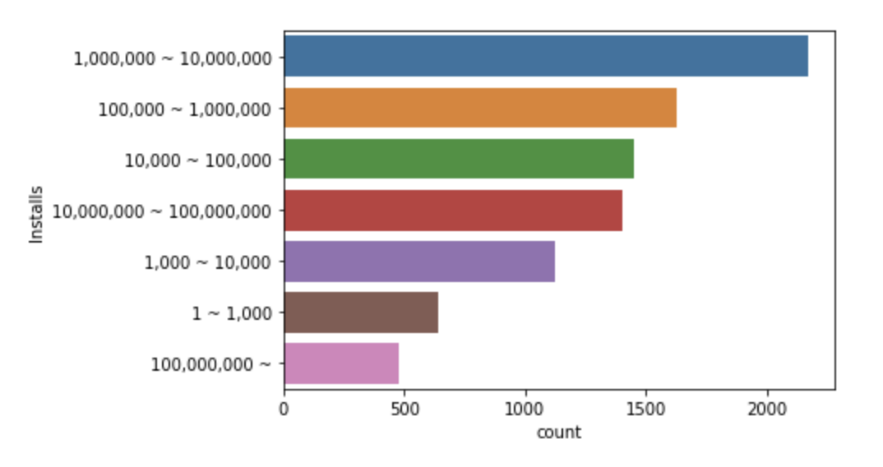
✓ Category 상위 5개만 추출하여 데이터 재구성 (Category 관련 분석에서만 활용)

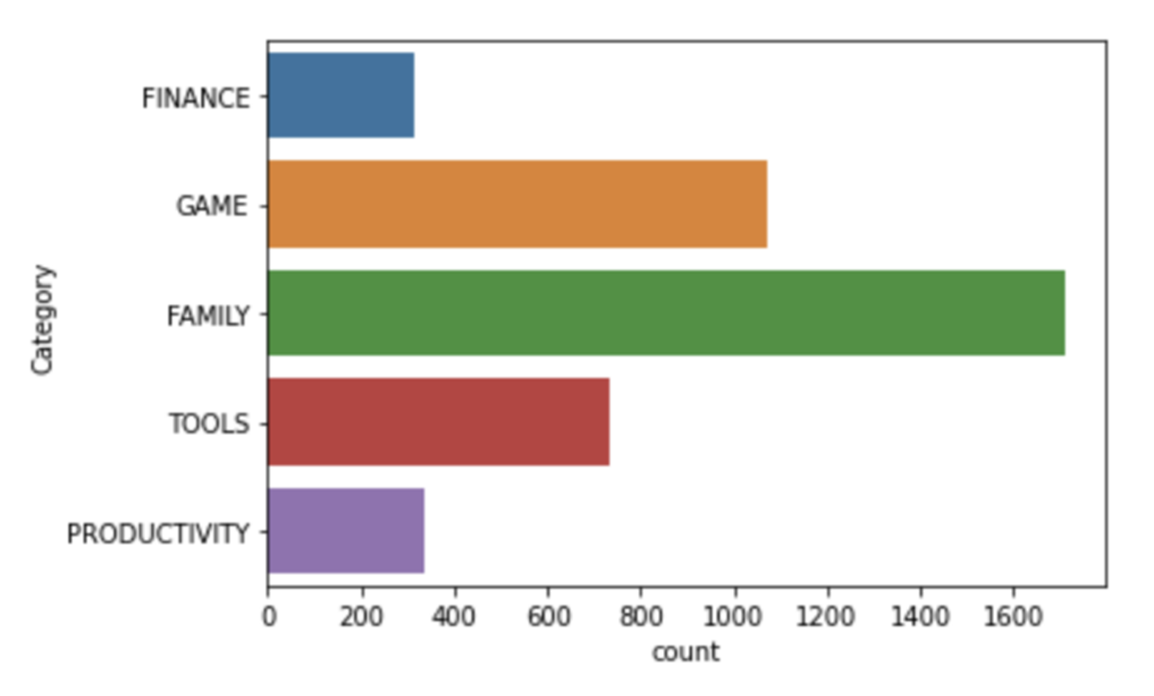


3. 데이터 EDA

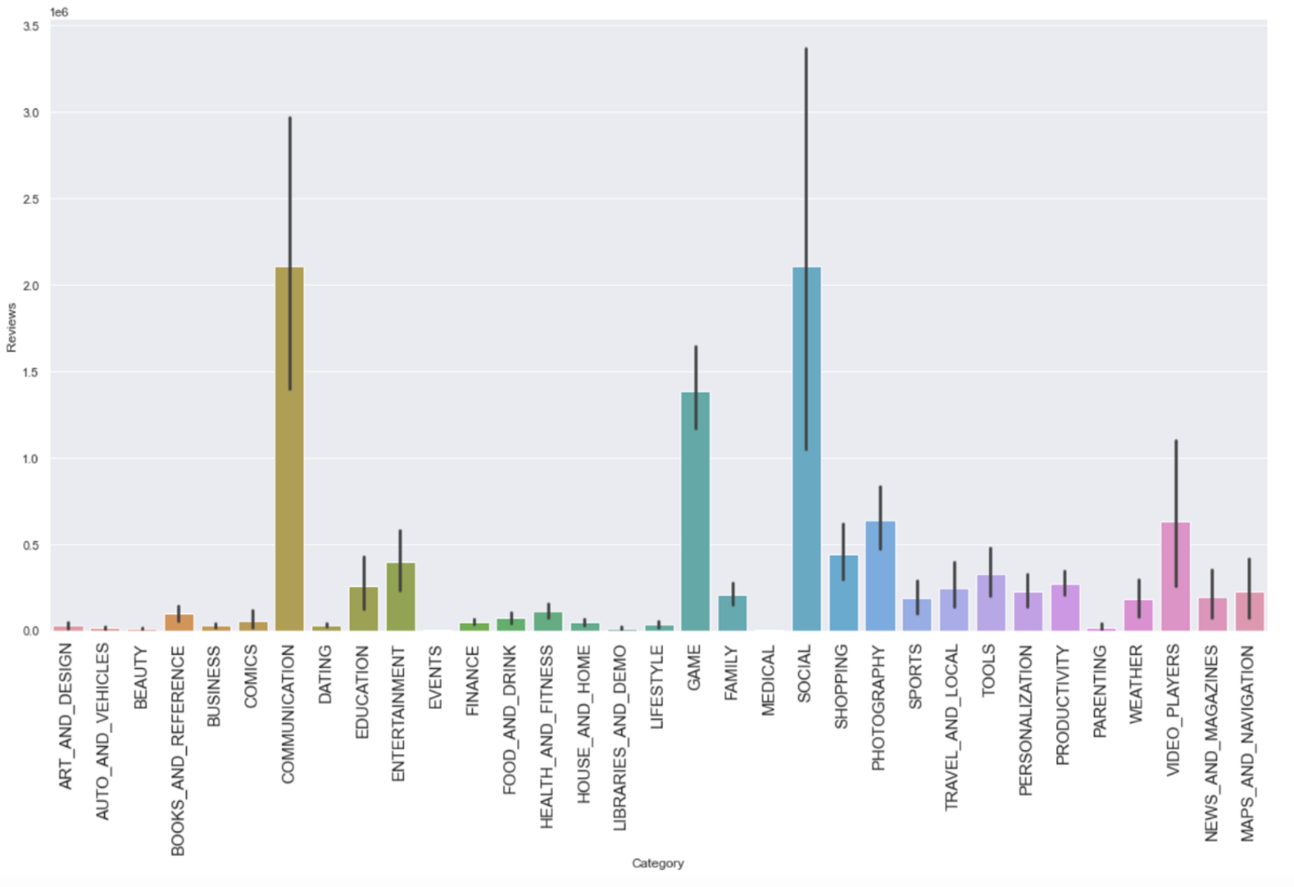
✓ 데이터의 형태를 파악하기 위한 범주형 데이터 분포 파악 (Type, Installs, Category)







✓ 데이터의 형태를 파악하기 위한 전반적인 데이터 EDA



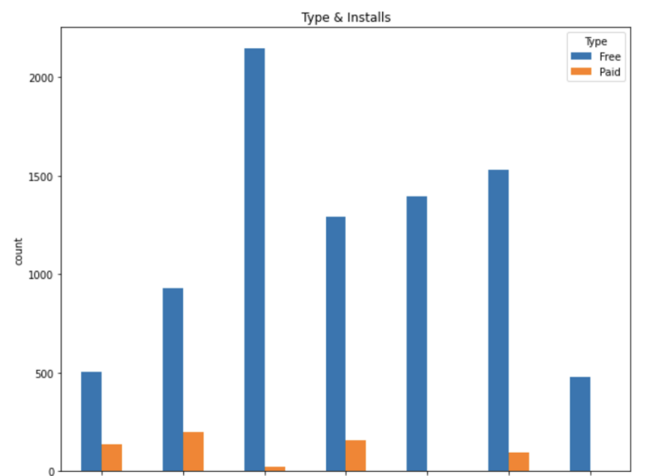
**4장 데이터 분석 결과**

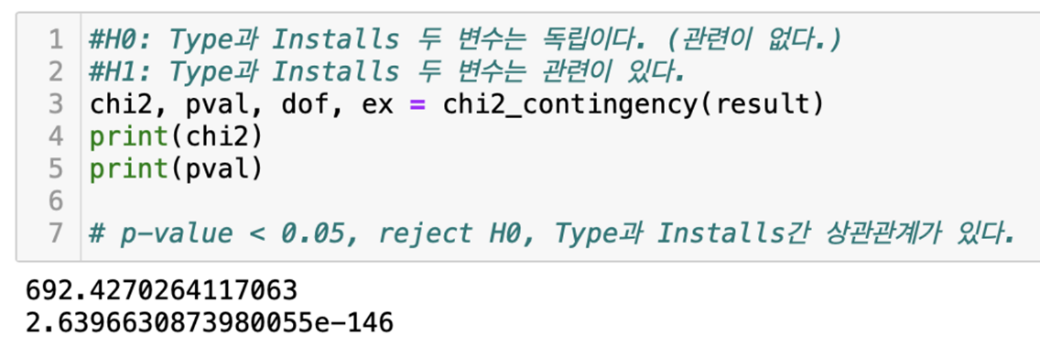
1. 분석 (1)

✓ 분석 내용: 앱의 유료/무료 여부(Type), 앱의 대상 연령대(Content Rating), 앱의 장르(Category)에 따라 설치 횟수(Installs)가 유의미한 차이가 있을까? 어떤 Column이 설치 횟수에 더 큰 영향을 줄까?

✓ 검정 방법: 범주형-범주형 변수 → 카이제곱 검정 (w/chisq)

✓ 1. Type – Installs 상관관계 분석

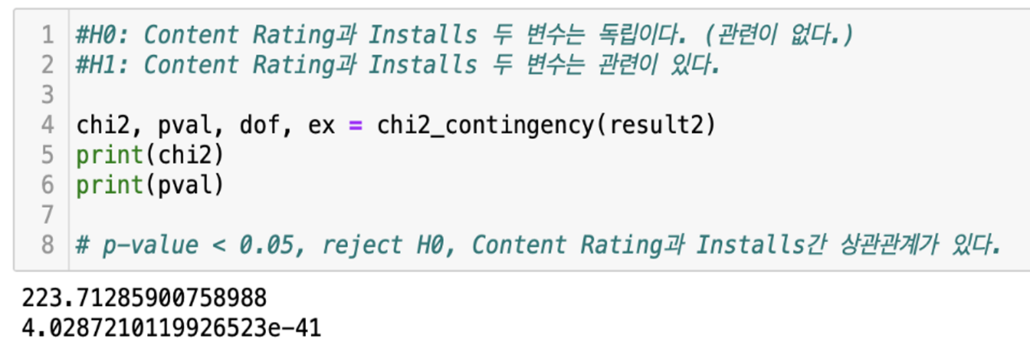




→ 결과: p-value가 0.05보다 작으므로 대립 가설 채택. 두 변수간 상관관계가 있다.

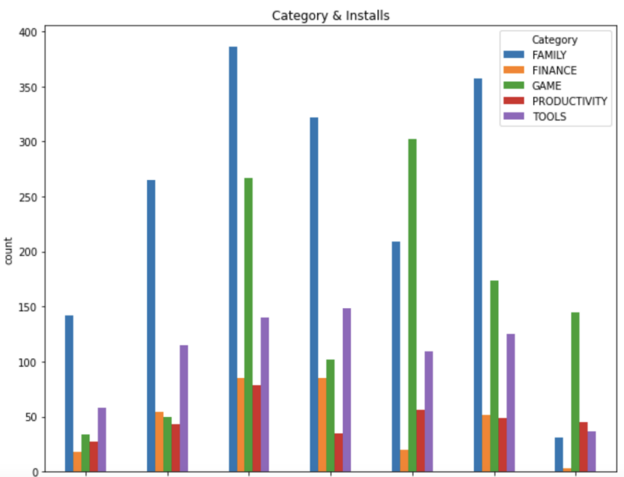
✓ 2. Content Rating – Installs 상관관계 분석

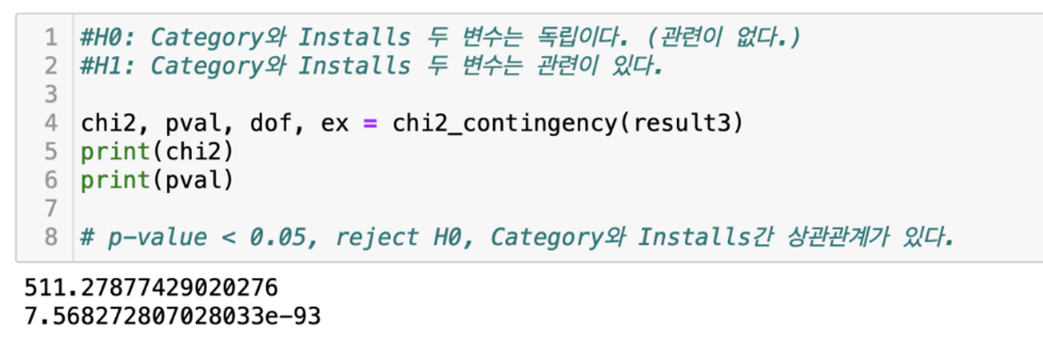




→ 결과: p-value가 0.05보다 작으므로 대립 가설 채택. 두 변수간 상관관계가 있다.

✓ 3. Category – Installs 상관관계 분석





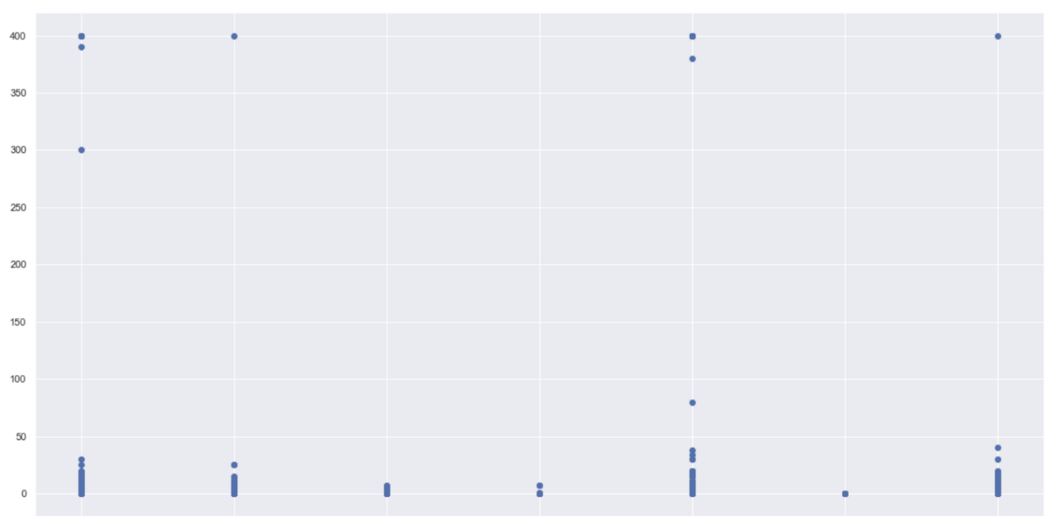
→ 결과: p-value가 0.05보다 작으므로 대립 가설 채택. 두 변수간 상관관계가 있다.

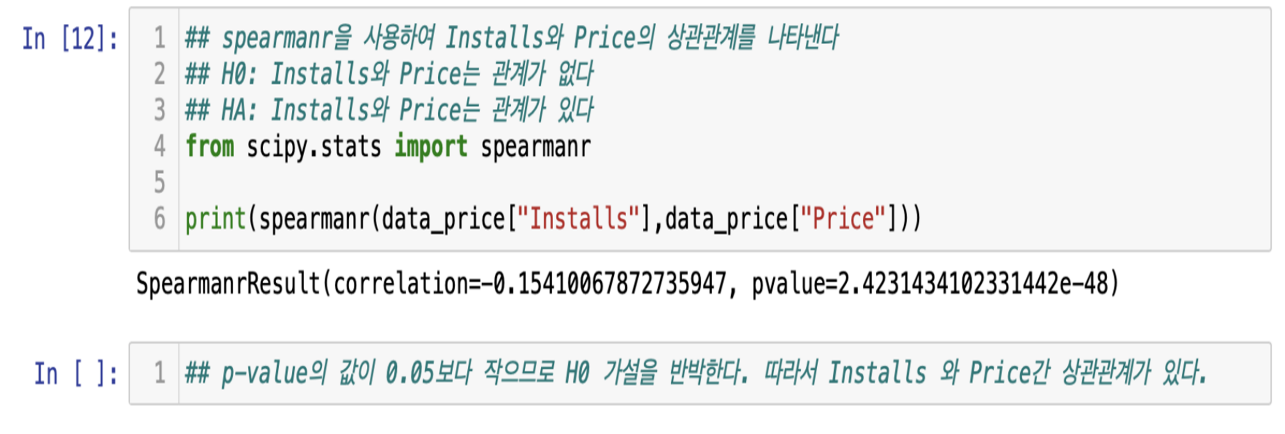
2. 분석 (2)

✓ 분석 내용: 앱의 가격(Price)에 따라 설치 횟수(Installs)가 유의미한 차이가 있을까?

✓ 검정 방법: 연속형-범주형 변수 → 상관 분석 (w/Correlation Test)

✓ Price – Installs 상관관계 분석





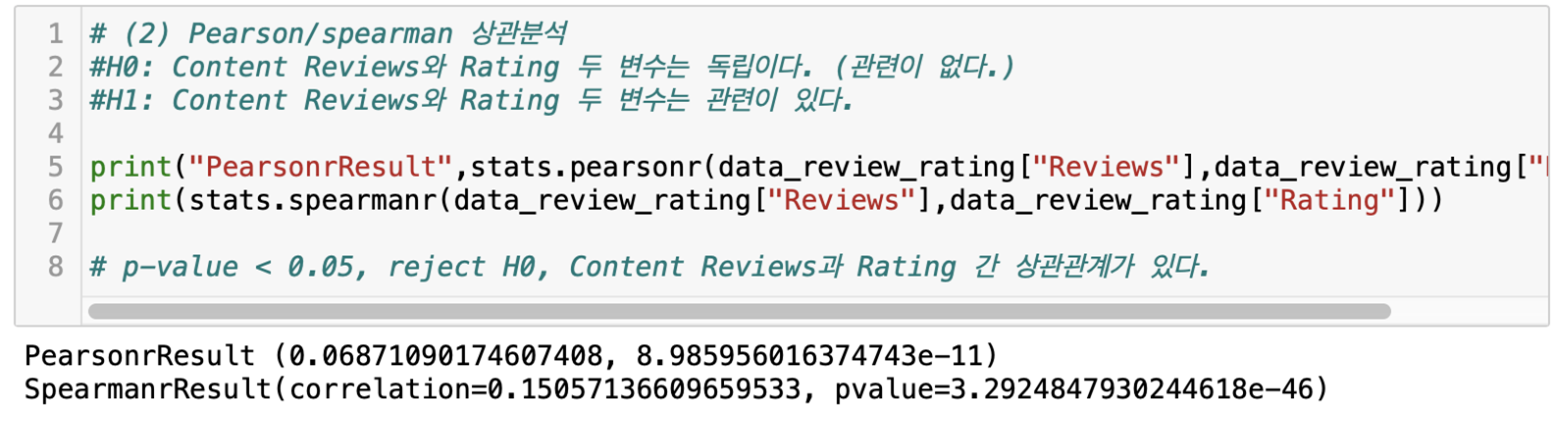
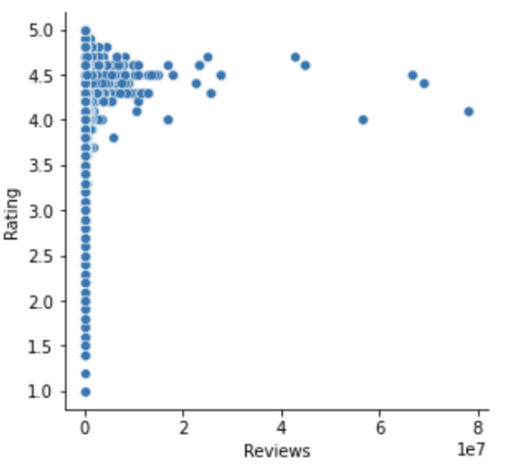
→ 결과: p-value가 0.05보다 작으므로 대립 가설 채택. 두 변수간 상관관계가 있다.

3. 분석 (3)

✓ 분석 내용: 리뷰 수(Reviews), 앱의 크기(Size), 앱의 가격(Price) 중 평점(Rating)과 가장 큰 상관관계를 지니는 변수는 무엇일까?

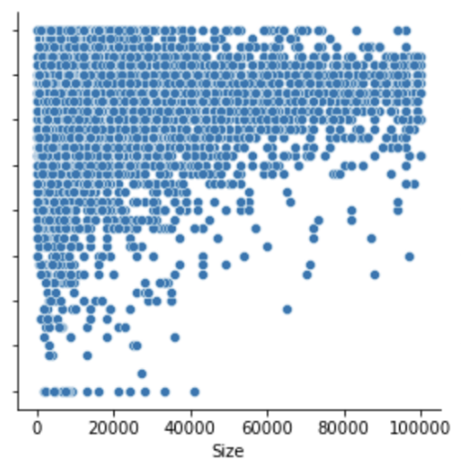
✓ 검정 방법: 연속형-연속형 변수 → 상관 분석 (w/Correlation Test)

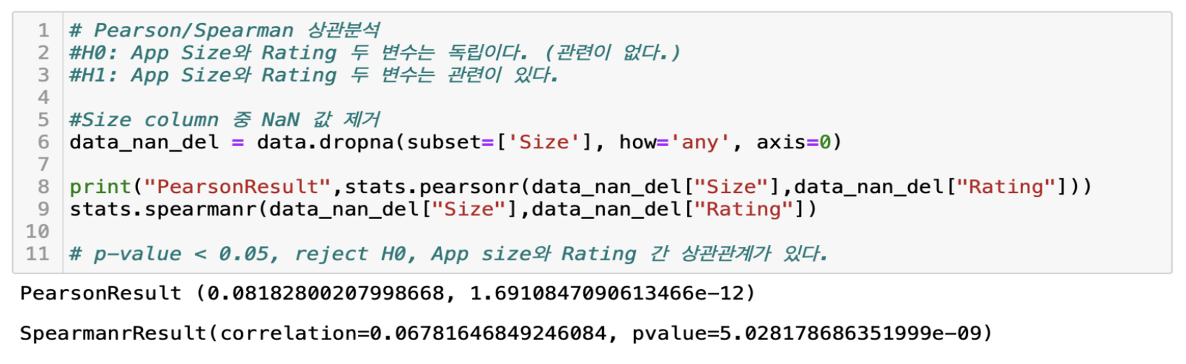
✓ 1. Reviews – Rating 상관관계 분석



→ 결과: pearson, spearman 상관분석 결과 모두 상관관계가 있는 것으로 나오나, 상관계수 값이 0.069로 굉장히 작고 시각화 결과도 상관관계가 유의미하게 있어 보이지 않는다.

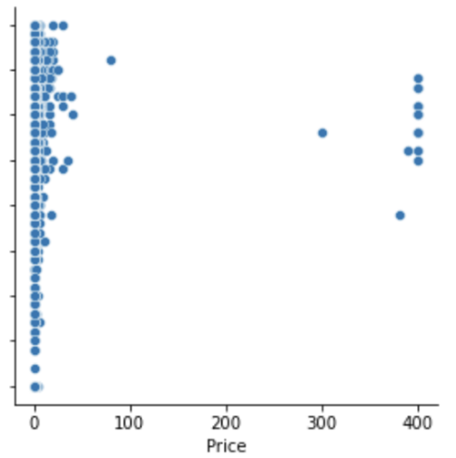
✓ 2. Size – Rating 상관관계 분석

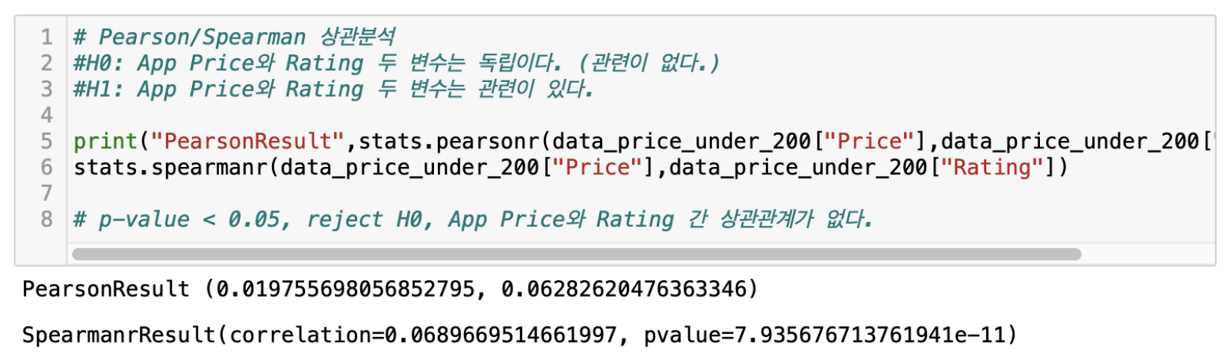




→ 결과: p-value가 0.05보다 작으므로 대립 가설 채택. 두 변수간 상관관계가 있다.

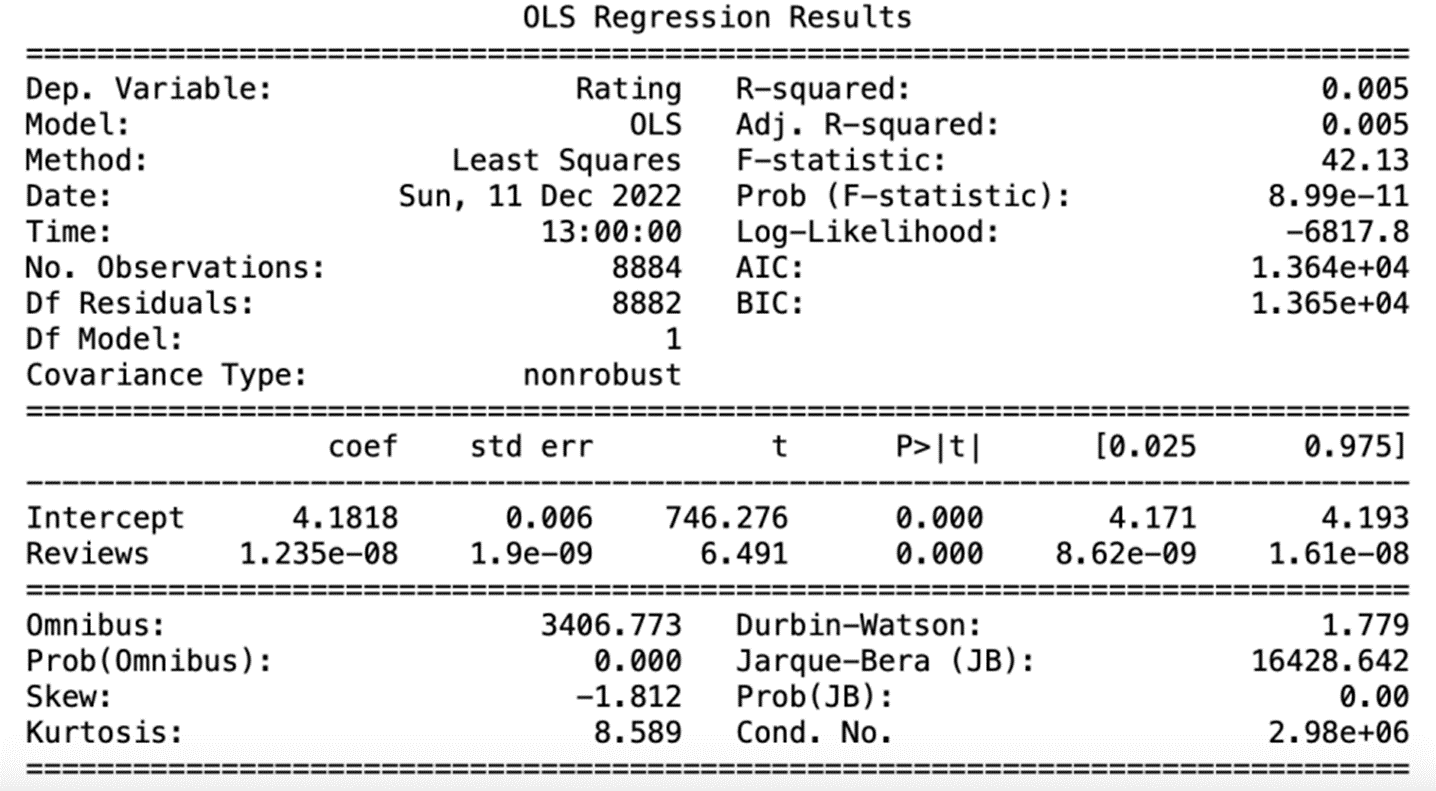
✓ 3. Price – Rating 상관관계 분석





→ 결과: p-value가 0.05보다 크므로 귀무 가설 채택. 두 변수간 상관관계가 없다. (Pearson)

✓ 4. 상관관계가 존재하는 Review-Rating/ Size-Rating 선형회귀분석 진행





4. 분석 결론

✓ 앱의 유료/무료 여부(Type), 앱의 대상 연령대(Content Rating), 앱의 장르(Category), 앱

가격(Price)와 설치 횟수(Installs)가 모두 유의미한 상관관계를 지니고 있다. 그 중에서도

Type이 가장 큰 상관관계를 지니고 있다고 볼 수 있다.

✓ 앱 가격(Price)은 평점(Rating)에 유의미한 영향을 끼치지 않는다.

리뷰 수(Reviews)와 앱 크기(Size)는 평점(Rating)에 유의미한 영향을 끼치는데, 그 중에

서도 Size가 Rating과 더 큰 상관관계를 지니고 있다.