14주차:
구글 플레이 앱 스토어를 이용한 실전 프로젝트

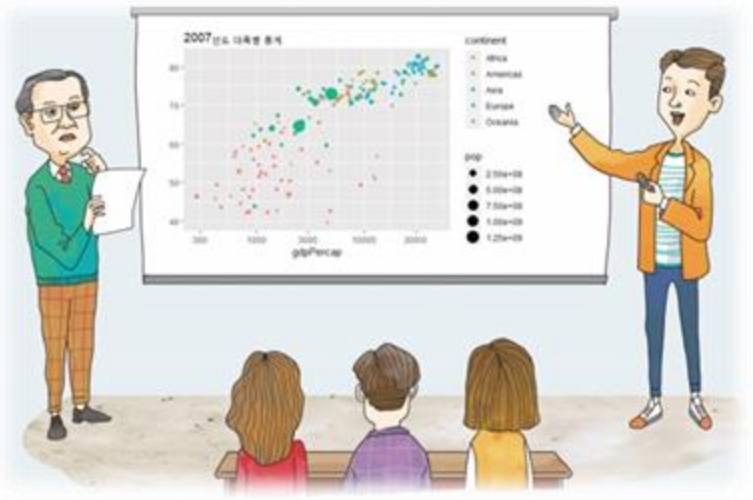
ChulSoo Park

School of Computer Engineering & Information Technology

Korea National University of Transportation

12 CHAPTER

실전 프로젝트



CONTENTS

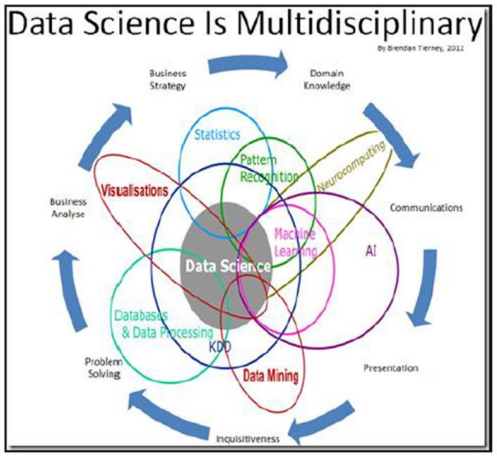
12.3 탐색적 data 분석

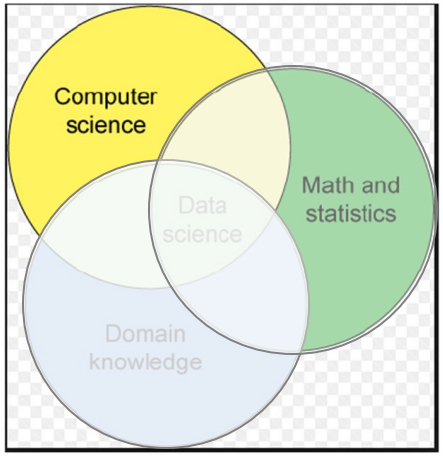
- 앱 개발자들의 바램(기대)
  - 앱 개발자들은 자신이 개발한 앱이 많은 다운로드와 높은 평점 기대
  - 구글 플레이 스토어 앱 기록된 데이터를 살피는 것은 매우 중요함
  - 기존 data를 활용한 평점과 다운로드 정보 분석
  - 다양한 앱들의 평점 분포 확인
  - 평점과 관련된 공통 특성 확인
  - 평점과 상관 관계가 높은 요인 확인

12.3 탐색적 data 분석

인터넷, IoT ,스마트, 웨어러블 시대의 도래로 빅데이터 시대에서

1. 데이터과학 분야의 전문가로서 갖추어야 할 다양한 학문적 이론과 실무지식을 익히고
2. 데이터 과학자로서 요구되는 창의성, 통찰력, 올바른 윤리의식 등을 갖추며
3. 데이터의 가치를 창출하기 위한 융합적 사고를 기반으로 데이터 분석, 설계 및 예측 능력을 갖춘 데이터 시대를 선도하는 전문가의 능력 함양





12.3 탐색적 data 분석

KBO리그 기록 및 순위

팀순위

2021.

순위	팀	경기수	승	패	무	승률	게임차	연속	출루율	장타율	최근 10경기
1	SSG	40	23	17	0	0.575	0.0	5승	0.348	0.415	7승-3패-0무
2	삼성	42	24	18	0	0.571	0.0	1승	0.347	0.404	5승-5패-0무
3	KT	40	22	18	0	0.550	1.0	1승	0.378	0.393	6승-4패-0무
4	키움	42	23	19	0	0.548	1.0	7승	0.357	0.394	9승-1패-0무
5	두산	40	21	19	0	0.525	2.0	2승	0.367	0.395	5승-5패-0무
6	LG	42	22	20	0	0.524	2.0	4패	0.344	0.386	5승-5패-0무
7	NC	41	21	20	0	0.512	2.5	3패	0.368	0.459	5승-5패-0무
8	한화	41	17	24	0	0.415	6.5	1패	0.331	0.335	4승-6패-0무
9	KIA	40	16	24	0	0.400	7.0	1패	0.349	0.340	3승-7패-0무
10	롯데	40	15	25	0	0.375	8.0	2패	0.363	0.396	3승-7패-0무

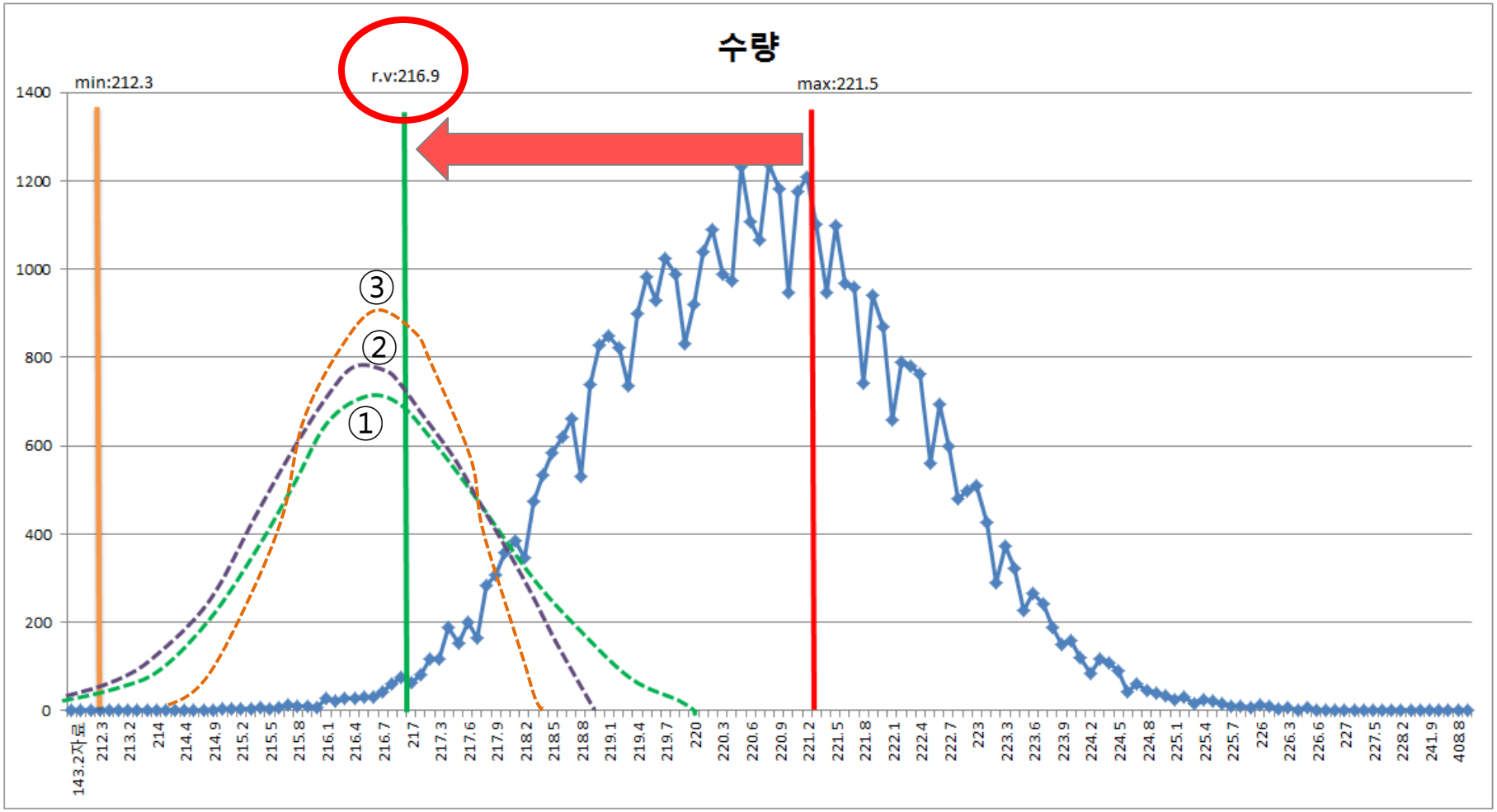
투수 순위 | 타자 순위

순위	선수	평균자책	경기수	이닝	승	패	세이브	홀드	탈삼진	피안타	피홈런	실점	볼넷	사구	승률
1	데스파이네 (KT)	1.66	10	59 2/3	5	3	0	0	51	41	2	16	26	1	0.625
2	카펜터 (한화)	1.69	9	53 1/3	2	3	0	0	56	35	3	15	24	5	0.400
3	수아레즈 (LG)	1.93	9	51 1/3	5	2	0	0	57	37	2	15	21	1	0.714
4	로켓 (두산)	1.99	9	54 1/3	4	3	0	0	43	57	1	12	18	4	0.571
5	뷰캐넌 (삼성)	2.10	9	55 2/3	5	1	0	0	55	46	2	16	16	0	0.833
6	원태인 (삼성)	2.13	8	50 2/3	6	2	0	0	52	43	3	12	15	1	0.750
7	박종훈 (SSG)	2.72	8	49 2/3	4	2	0	0	37	38	2	17	16	5	0.667
8	스트레일리 (롯데)	2.74	9	49 1/3	3	4	0	0	53	47	4	19	18	1	0.429
9	문승원 (SSG)	3.05	8	44 1/3	1	2	0	0	29	35	1	19	15	2	0.333
10	최원준 (두산)	3.07	8	44	4	0	0	0	33	38	5	15	11	4	1.000
11	요키시 (키움)	3.11	9	55	5	3	0	0	39	58	8	20	14	0	0.625
12	배제성 (KT)	3.19	8	42 1/3	4	3	0	0	40	31	2	18	29	1	0.571
13	루친스키 (NC)	3.53	9	51	3	3	0	0	53	51	1	25	23	2	0.500
14	브룩스 (KIA)	3.54	9	56	1	4	0	0	35	72	2	25	13	0	0.200
15	킹험 (한화)	3.77	8	45 1/3	4	3	0	0	39	36	5	21	12	2	0.571
16	켈리 (LG)	3.81	9	52	2	3	0	0	36	47	6	22	20	4	0.400
17	김민우 (한화)	3.83	9	47	5	2	0	0	42	35	5	21	23	3	0.714
18	정인 (KIA)	4.03	8	44 2/3	2	2	0	0	41	41	5	21	19	0	0.500
19	고영표 (KT)	4.40	7	43	3	2	0	0	33	45	3	22	6	7	0.600
20	박세웅 (롯데)	5.02	8	43	2	2	0	0	33	38	8	24	16	3	0.500

투수 순위 | 타자 순위

순위	선수	타율	경기수	타수	안타	2루타	3루타	홈런	타점	득점	도루	볼넷	삼진	출루율	장타율
1	강백호 (KT)	0.394	40	155	61	8	1	5	42	18	2	22	24	0.461	0.555
2	이정후 (키움)	0.364	42	162	59	17	3	1	31	35	5	26	13	0.454	0.525
3	알칸지 (NC)	0.351	40	131	46	10	1	9	37	28	1	22	13	0.459	0.649
4	피렐라 (삼성)	0.347	42	170	59	9	1	12	37	30	4	11	21	0.389	0.624
5	최원준 (KIA)	0.347	40	173	60	4	5	1	17	31	11	13	24	0.398	0.445
6	허경민 (두산)	0.342	39	155	53	7	0	3	17	25	0	11	12	0.391	0.445
7	강민호 (삼성)	0.341	37	123	42	7	0	5	26	16	0	13	17	0.401	0.520
8	박건우 (두산)	0.338	39	142	48	9	0	2	25	23	2	14	20	0.404	0.444
9	페르난데스 (두산)	0.331	38	154	51	6	0	6	22	29	0	16	12	0.391	0.487
10	이대호 (롯데)	0.328	35	134	44	3	0	8	28	16	0	15	15	0.400	0.530
11	알테어 (NC)	0.319	41	141	45	9	0	13	32	28	5	14	37	0.394	0.660
12	전준우 (롯데)	0.318	40	148	47	9	0	2	23	23	2	25	17	0.416	0.419
13	배정대 (KT)	0.315	40	146	46	8	0	2	18	30	6	26	31	0.429	0.411
13	김현수 (LG)	0.315	41	146	46	7	0	7	31	19	0	25	15	0.407	0.507
15	최정 (SSG)	0.311	39	132	41	7	0	11	30	29	5	26	36	0.434	0.614
16	박해민 (삼성)	0.308	42	146	45	7	0	2	21	25	14	21	22	0.386	0.397
17	홍창기 (LG)	0.308	42	156	48	8	1	2	21	32	9	35	23	0.447	0.410
18	안치홍 (롯데)	0.301	39	156	47	10	2	3	30	17	2	20	23	0.376	0.449
18	구자욱 (삼성)	0.301	42	156	47	9	4	5	25	34	9	22	28	0.380	0.506
20	심우준 (KT)	0.299	40	117	35	7	1	2	20	21	6	9	25	0.356	0.427

12.3 탐색적 data 분석



12.3 탐색적 data 분석

Figure 5: A "Black Box" Co-operative

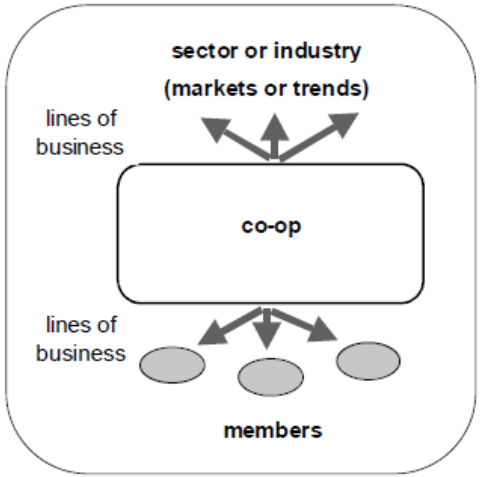
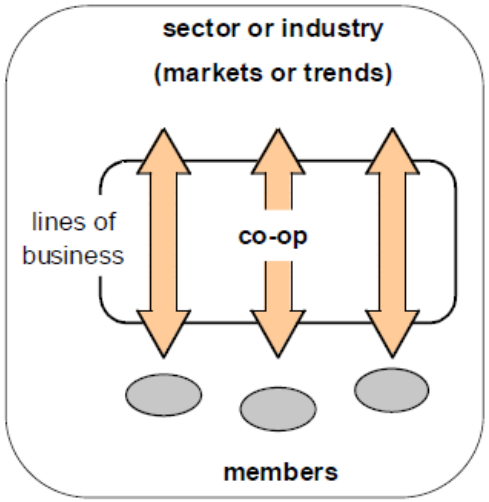
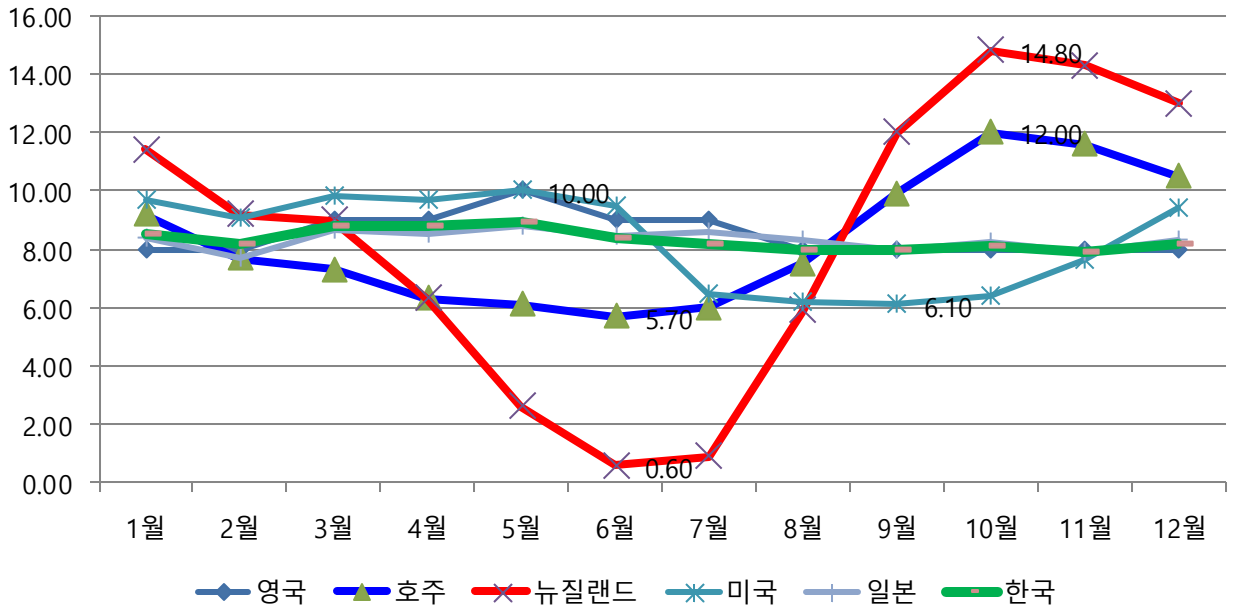


Figure 4: A Transparent Co-operative



Three Strategic Concepts for the Guidance of Co-operatives

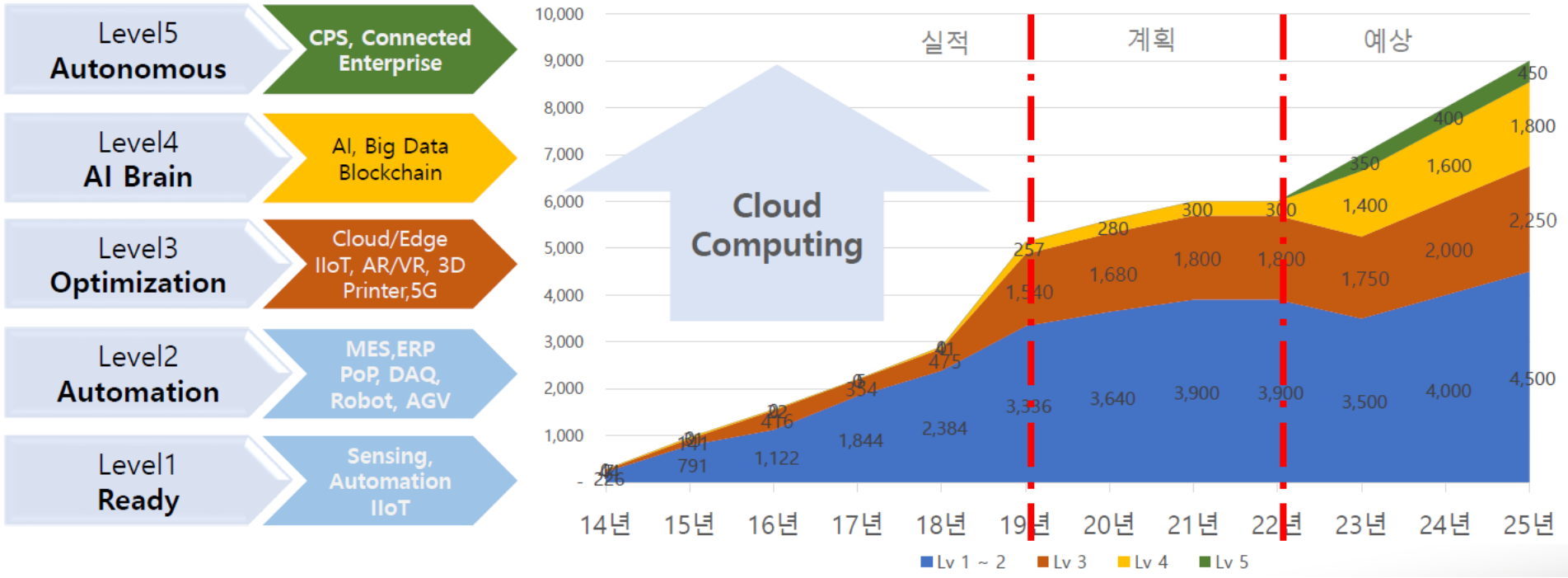
출처 : Centre for the Study of Co-operatives University of Saskatchewan



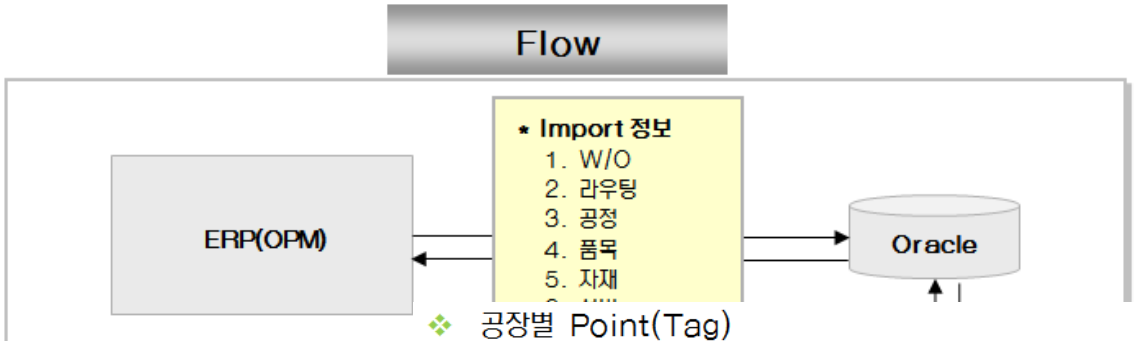
출처 : 낙농 선진국의 유대체계 연구(낙농육유협회 2012년)

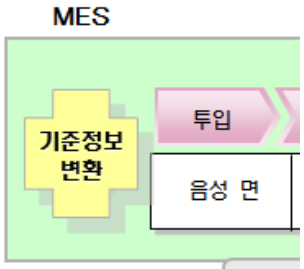
12.3 탐색적 data 분석

스마트팩토링의 5단계



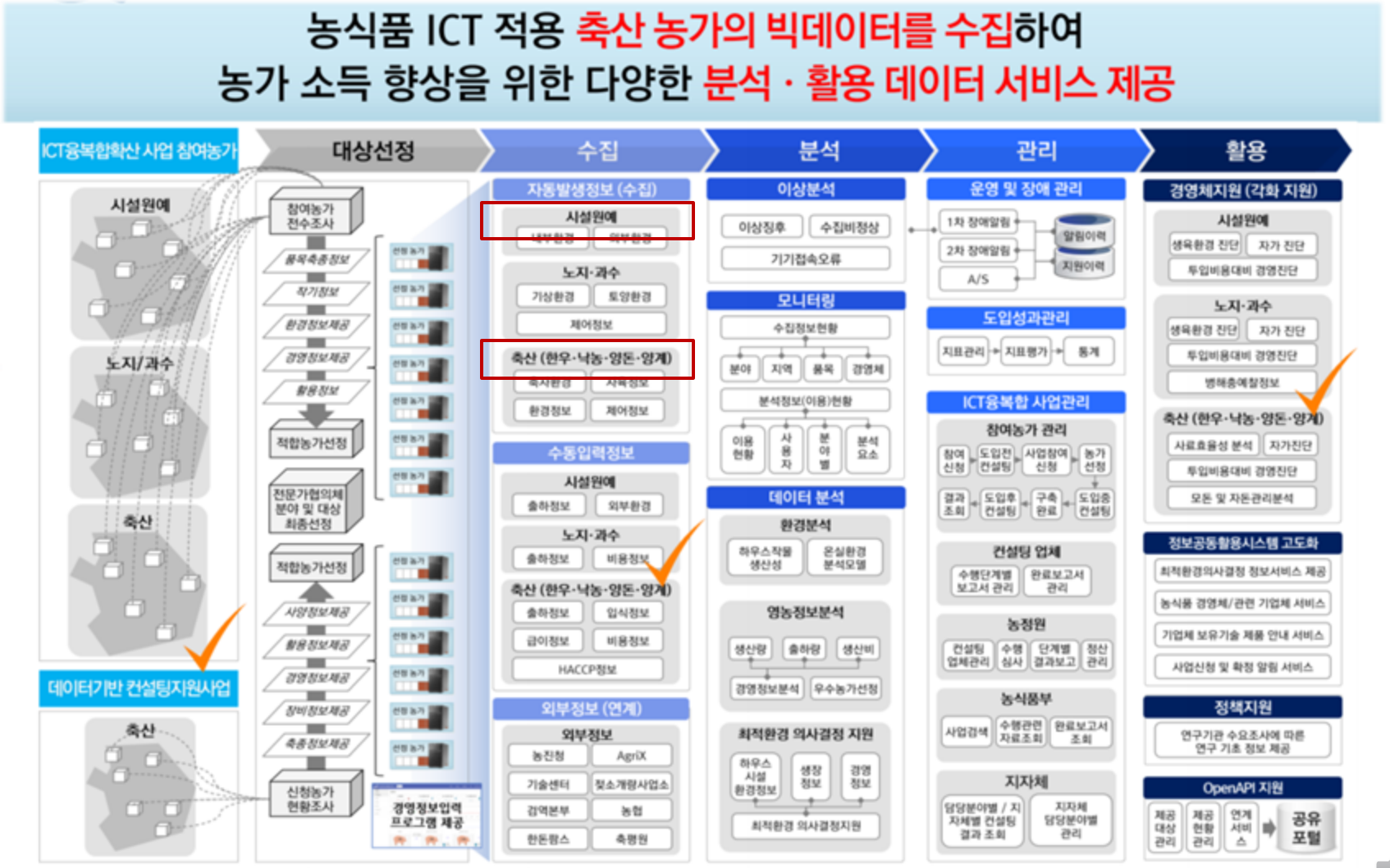
12.3 탐색적 data 분석





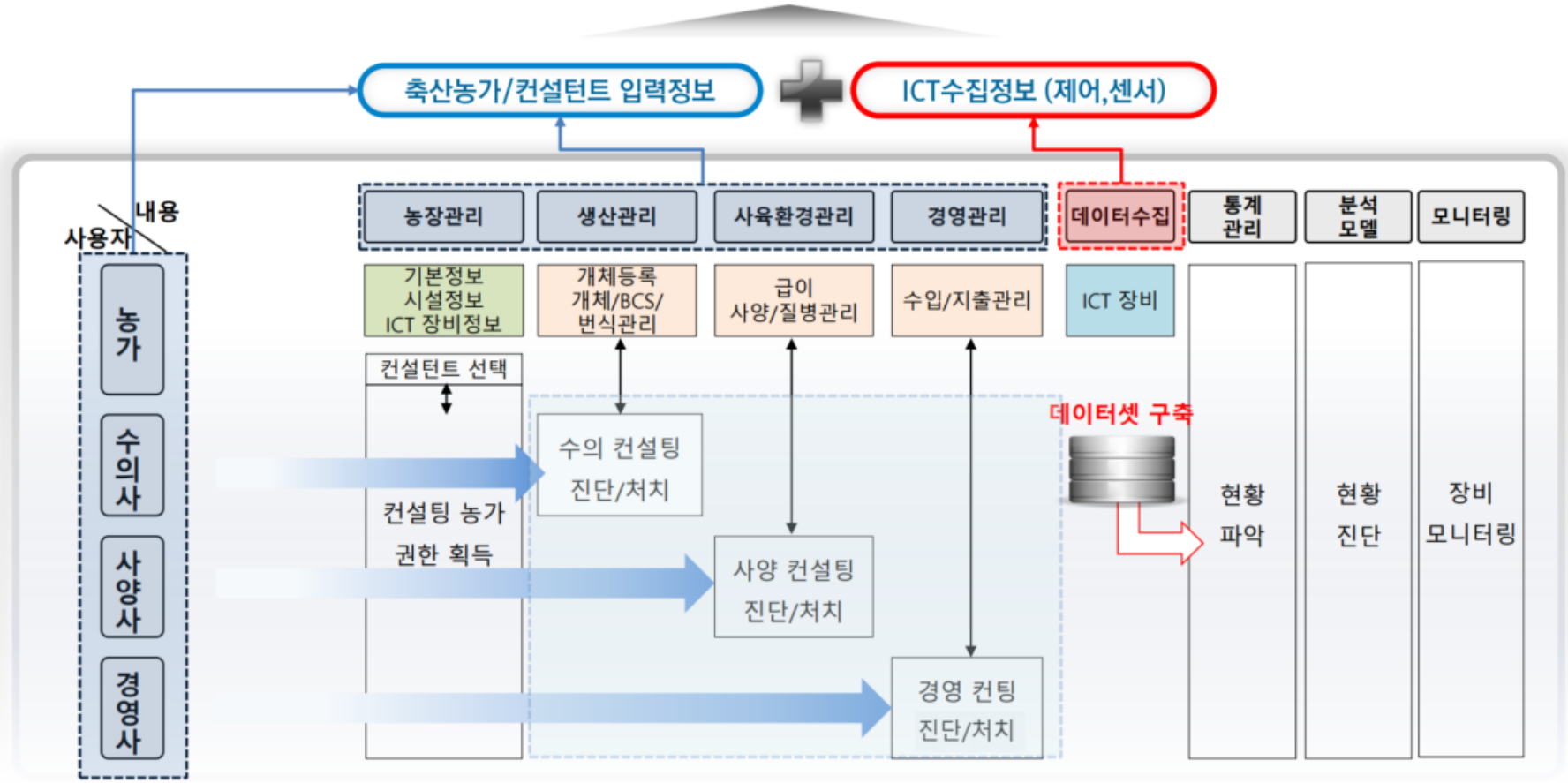
TAG 유형	TAG 유형 DESC	제일두부	제일생면	명가식품	제이생면
AI	산도, 약품농도, 염도, PH	-	15	-	-
AL	Alarm	13	-	-	-
BI	Brix	-	3	-	-
CT	Cycle Time	3	5	-	-
DE	두유농도	3	-	-	-
FT	가수량, 대두투입순간유량, 두유생산량, 응고제투입량	9	-	-	-
HI	습도	-	3	3	-
PB	Reset	3	-	-	-
PI	진공압	-	5	-	-
PN	제품 Selector	16	51	32	12
QT	적산량	28	47	38	28
SL	제품 Selector(TOPIA)	3	-	-	-
ST	가동상태(금속검출기)	3	4	8	8
TI	온도	45	47	18	4
WI	설정값	-	4	-	-
WT	On/Off, 비가동, Start/Stop/종료시간	16	42	30	23
합계		142	226	129	75

12.3 탐색적 data 분석



12.3 탐색적 data 분석

수동입력정보와 ICT수집정보를 활용한 데이터 기반 컨설팅 서비스 제공



12.3 탐색적 data 분석

- 평점 분포
  - 히스토그램을 이용하 플레이 스토어 앱에 매겨진 평점 분포 확인

```
> library(dplyr)
> # 03 탐색적 data  분석 #
> library(ggplot2)
> x%>%ggplot(aes(Rating)) + geom_histogram()
에러: StatBin requires a continuous x variable: the x variable is di
screte.Perhaps you want stat="count"?
Run `rlang::last_error()` to see where the error occurred.
```

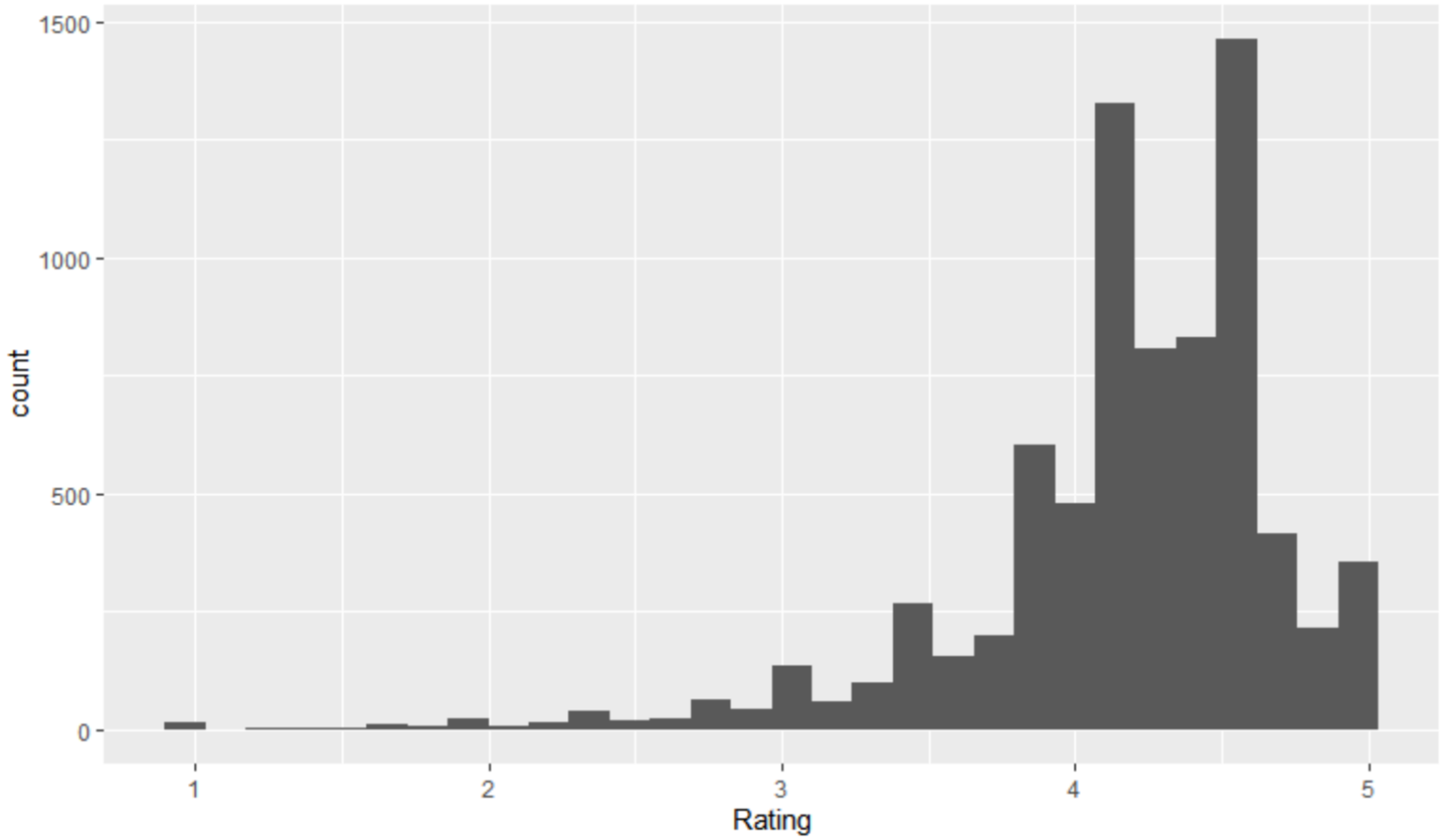
Console C:/RSources/

```
> glimpse(x)
Rows: 5,356
Columns: 13
$ App             <chr> "Coloring book mo
$ Category        <fct> ART_AND_DESIGN, A
$ Rating          <chr> "3.9", "4.4", "3.
$ Reviews         <dbl> 967, 167, 178, 36
$ Size            <dbl> 1.4e+07, 5.6e+06,
$ Installs        <dbl> 5e+05, 5e+04, 5e+
$ Type            <fct> Free, Free, Free,
$ Price           <dbl> 0, 0, 0, 0, 0, 0,
$ Content.Rating  <fct> Everyone, Everyone
$ Genres          <fct> Art & Design;Pret
$ Last.Updated    <date> NA, NA, NA, NA, N
$ Current.Ver     <chr> "2.0.0", "1", "1.
$ Android.Ver     <chr> "4.0.3 and up", "
```

```
> x$Rating = as.numeric(x$Rating)
> str(x)
'data.frame':   5356 obs. of  13 variables:
 $ App           : chr  "Coloring book moana" "Paper flowers instructions" "Smoke Effect Photo Make
r - Smoke Editor" "Infinite Painter" ...
 $ Category      : Factor w/ 33 levels "ART_AND_DESIGN",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Rating        : num  3.9 4.4 3.8 4.1 4.4 4.4 4.6 4.7 4.7 4.8 ...
 $ Reviews       : num  967 167 178 36815 13880 ...
 $ Size          : num  1.4e+07 5.6e+06 1.9e+07 2.9e+07 2.8e+07 1.2e+07 2.1e+07 5.5e+06 4.2e+06 6.0
e+06 ...
 $ Installs      : num  5e+05 5e+04 5e+04 1e+06 1e+06 1e+06 1e+05 5e+05 5e+05 1e+04 ...
 $ Type          : Factor w/ 2 levels "Free","Paid": 1 1 1 1 1 1 1 1 1 1 ...
 $ Price         : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Content.Rating: Factor w/ 6 levels "Adults only 18+",..: 2 2 2 2 2 2 2 2 3 2 ...
 $ Genres        : Factor w/ 116 levels "Action","Action;Action & Adventure",..: 13 10 10 10 10 10
10 10 10 10 ...
 $ Last.Updated  : Date, format: NA ...
 $ Current.Ver   : chr  "2.0.0" "1" "1.1" "6.1.61.1" ...
 $ Android.Ver   : chr  "4.0.3 and up" "2.3 and up" "4.0.3 and up" "4.2 and up" ...
 - attr(*, "na.action")= 'omit' Named int [1:3736] 1 3 4 5 9 10 13 15 16 18 ...
  ..- attr(*, "names")= chr [1:3736] "1" "3" "4" "5" ...
```

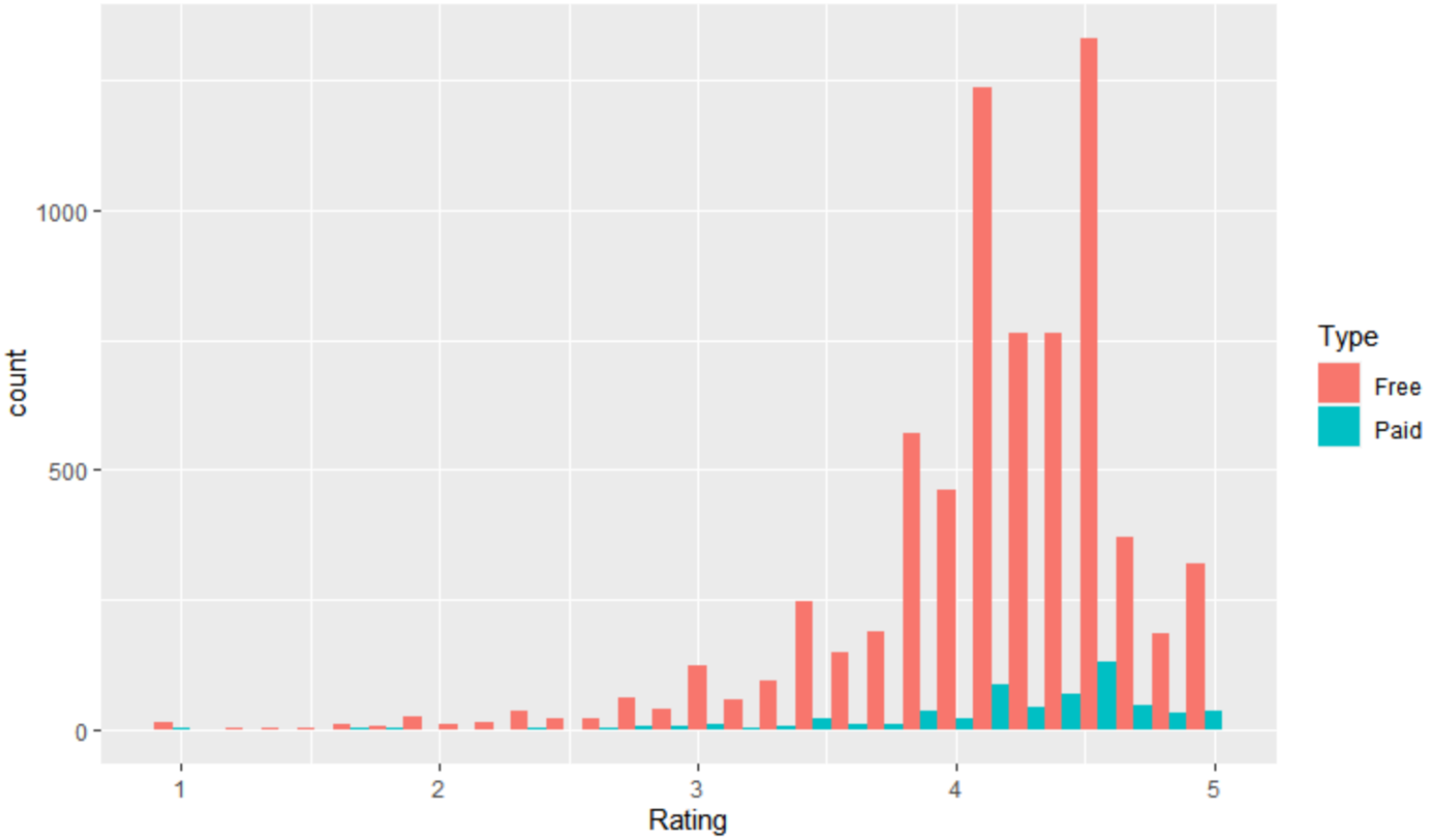
12.3 탐색적 data 분석

- 평점 분포
  - 히스토그램을 이용하 플레이 스토어 앱에 매겨진 평점 분포 확인
  - > x%>%ggplot(aes(Rating)) + geom_histogram()



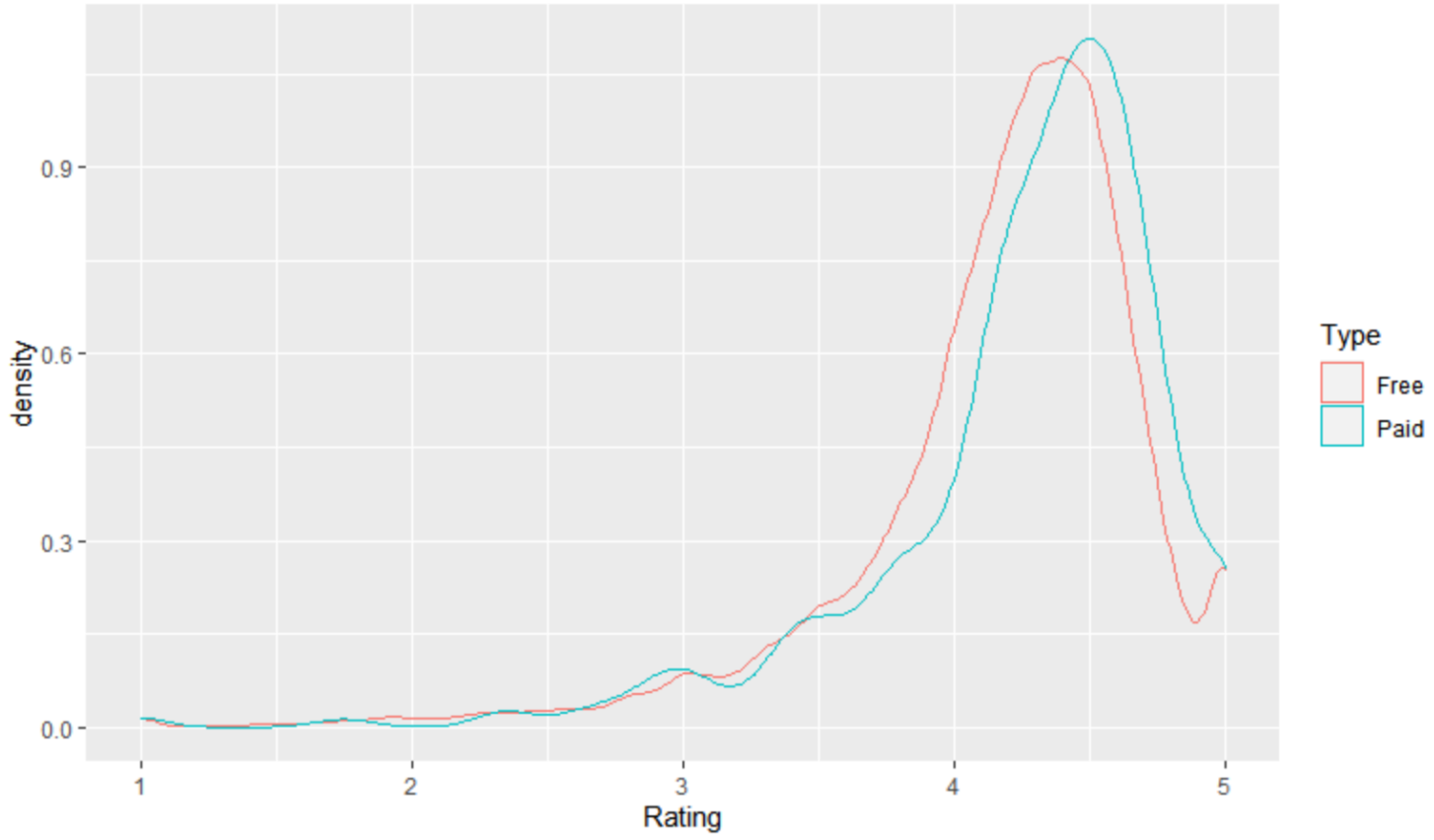
12.3 탐색적 data 분석

- 평점 분포
  - 히스토그램을 이용하 플레이 스토어 앱에 매겨진 평점 분포 확인
  - > x%>%ggplot(aes(Rating, fill = Type)) + geom_histogram(position = "dodge")



12.3 탐색적 data 분석

- 평점 분포
  - 히스토그램을 이용하 플레이 스토어 앱에 매겨진 평점 분포 확인
  - > x%>%ggplot(aes(Rating, col = Type)) + geom_density()



12.3 탐색적 data 분석

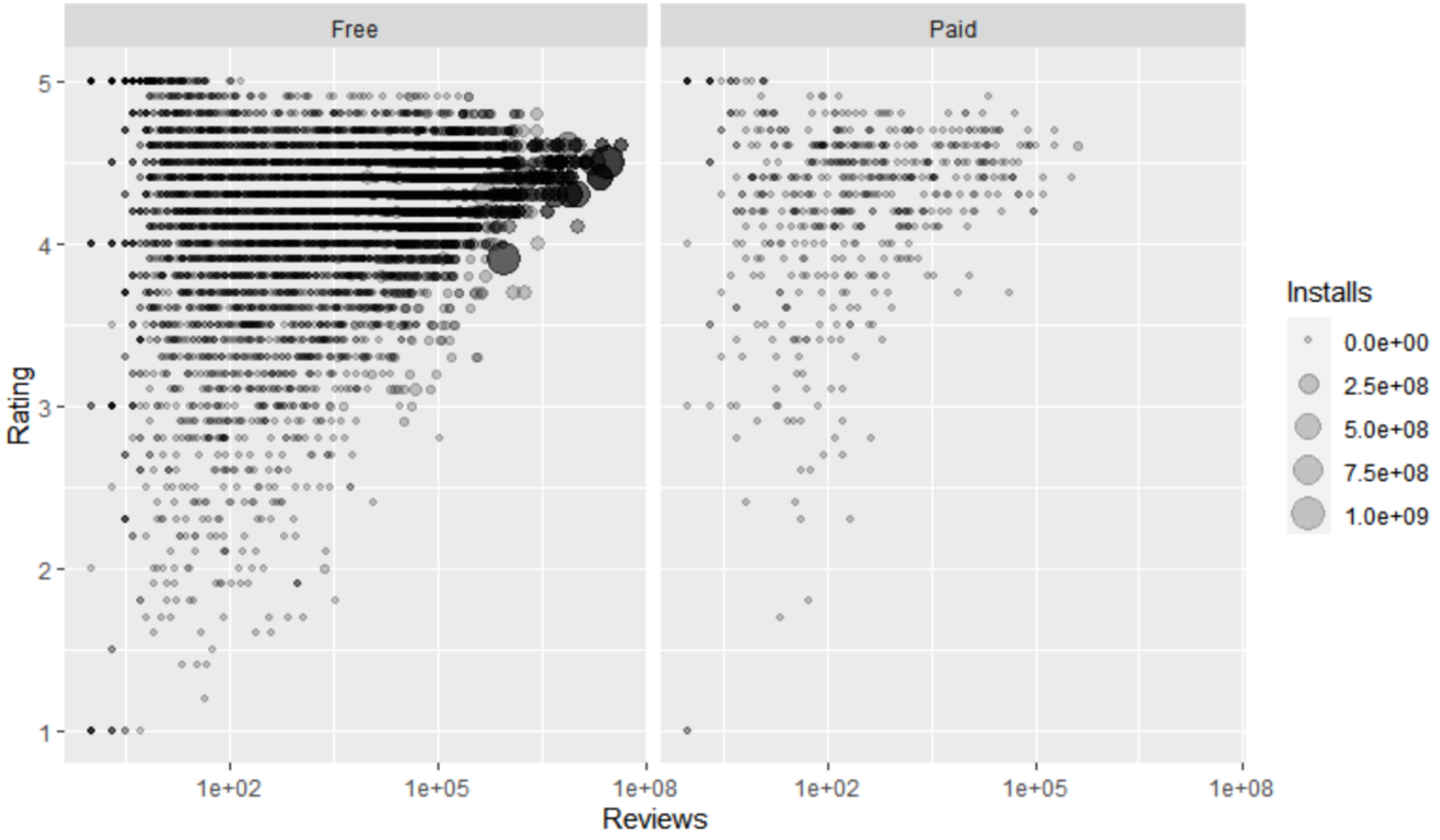
- 변수들의 상호 관계(평점과 리뷰의 관계)
  - 평점에 영향을 주는 변수, 상호 연관이 있는 속성은 무엇일까?
  - 이러한 것을 찾기 위한 시각화, 평점과 리뷰와의 관계
  - > x%>%ggplot(aes(Reviews, Rating, col = Type)) + geom_point(alpha = 0.2)
  - > x%>%ggplot(aes(Reviews, Rating, col = Type)) + geom_point(alpha = 0.8) + scale_x_log10()

12.3 탐색적 data 분석

- 변수들의 상호 관계(평점, 리뷰, 설치 횟수의 관계)
  - 평점과 설치 횟수의 관계
  - 설치 횟수는 Installs에 기록
  - 그래프 마커의 크기를 설치 횟수에 대응, 세 가지 변수의 연관성 찾아보기
  - > x%>%ggplot(aes(Reviews, Rating, col = Type, size = Installs)) + geom_point(alpha = 0.2) + scale_x_log10()

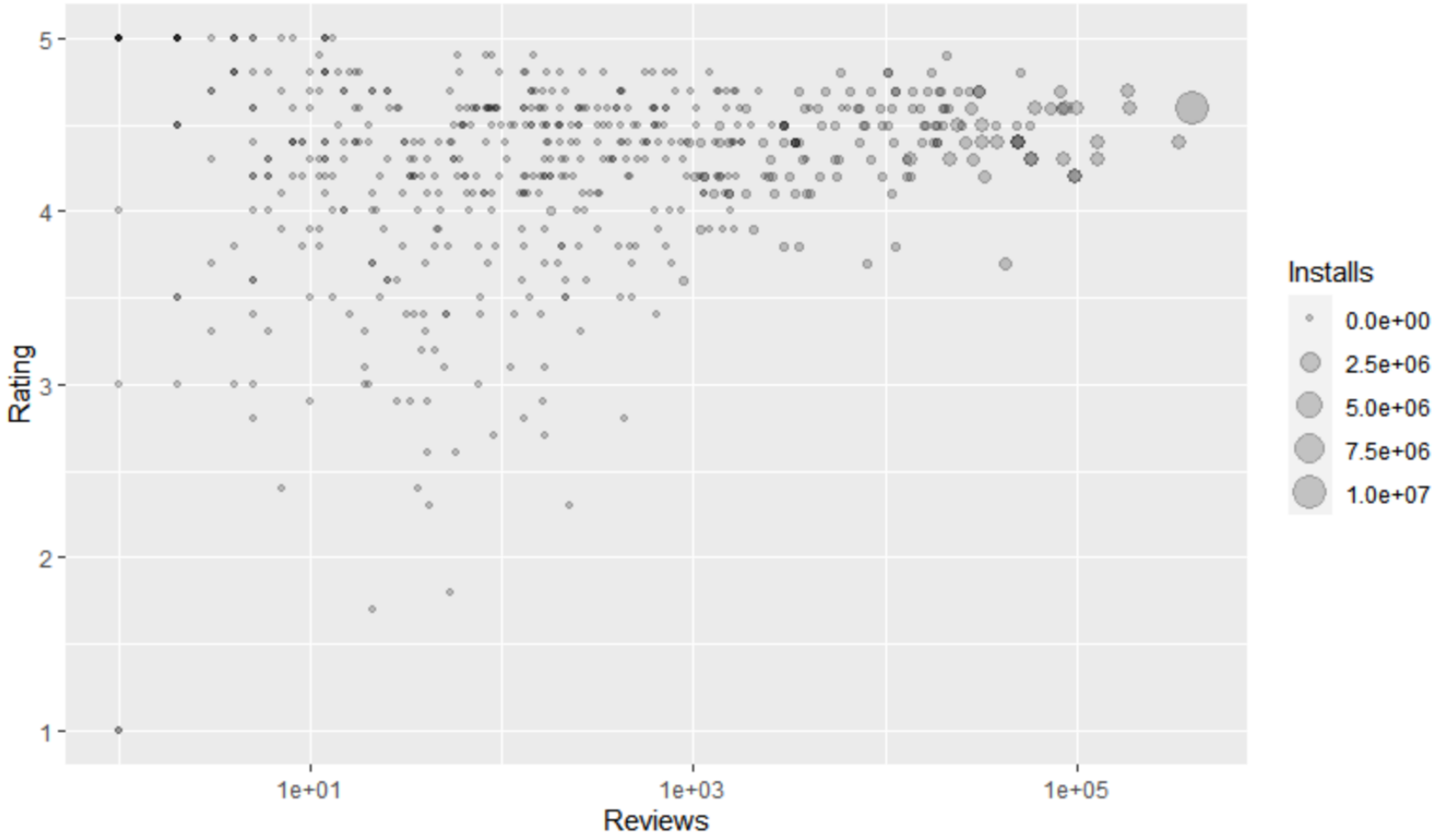
12.3 탐색적 data 분석

```
> x%>%ggplot(aes(Reviews, Rating, size = Installs)) +
geom_point(alpha = 0.2) +scale_x_log10() + facet_wrap(~Type)
```



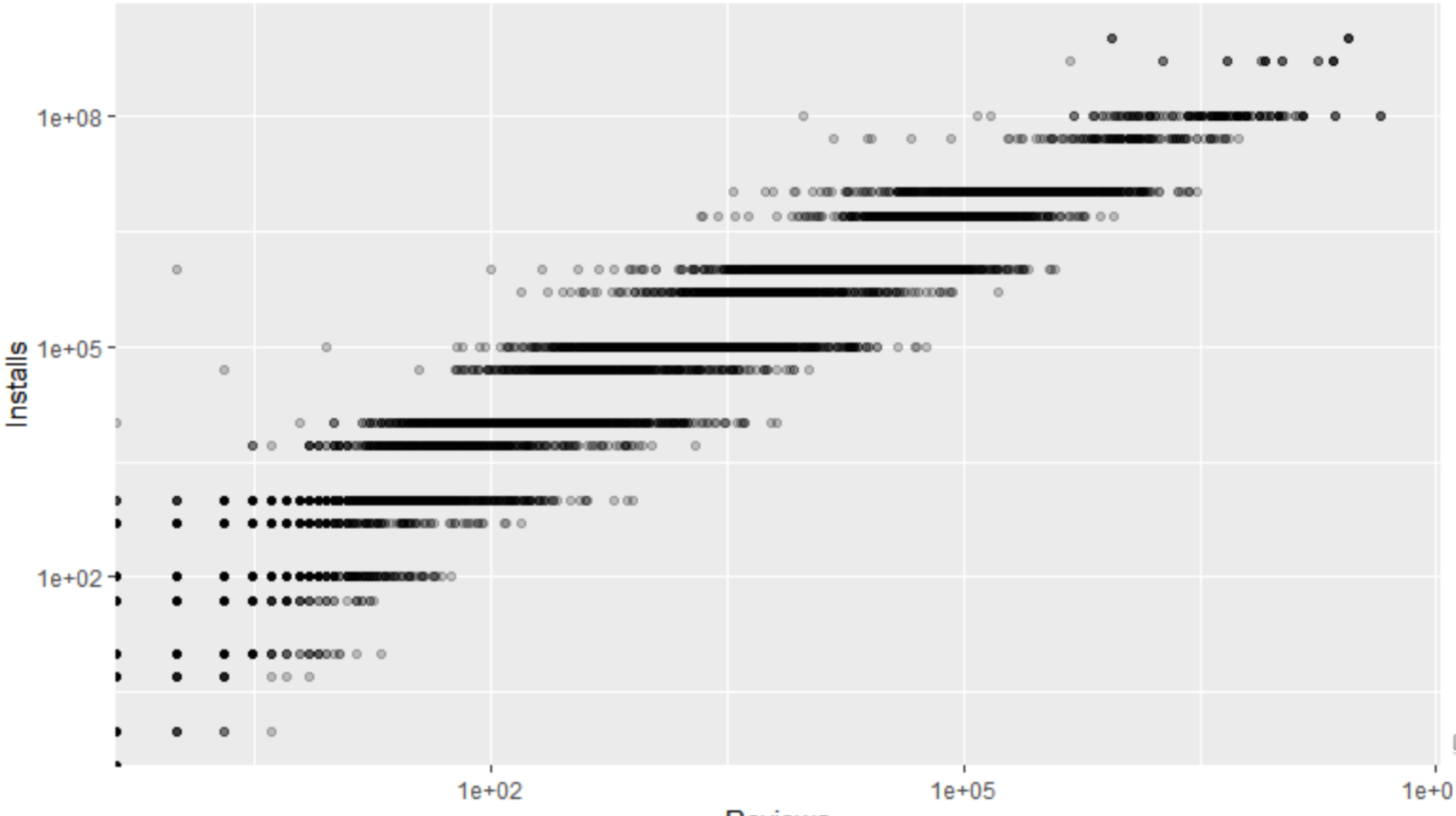
12.3 탐색적 data 분석

```
> x%>%filter(Type=="Paid")%>%ggplot(aes(Reviews, Rating, size = Installs)) + geom_point(alpha = 0.2) + scale_x_log10()
```



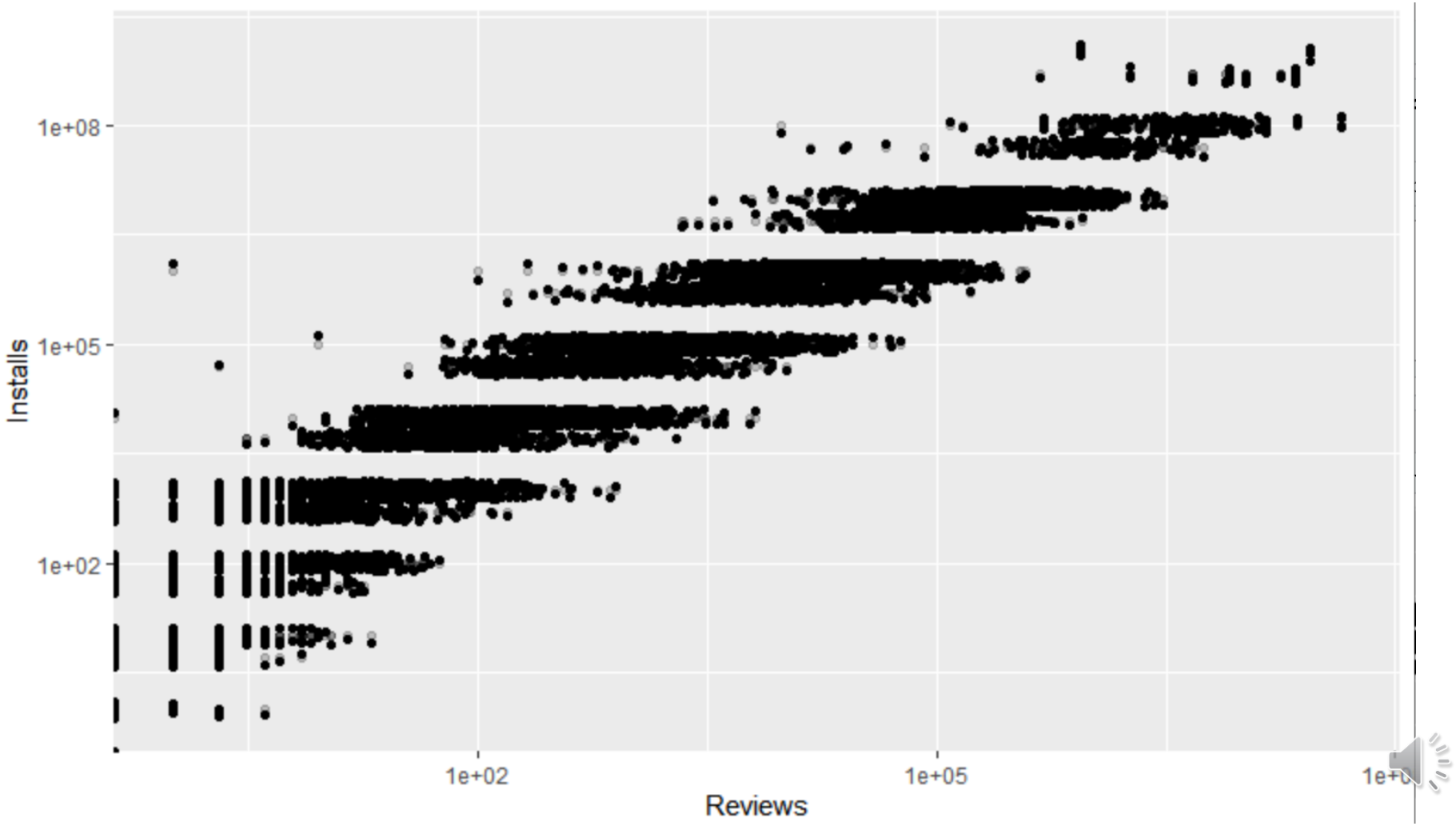
12.3 탐색적 data 분석

```
> x%>%ggplot(aes(Reviews, Installs)) + geom_point(alpha = 0.2) +
scale_x_log10() + scale_y_log10()
```



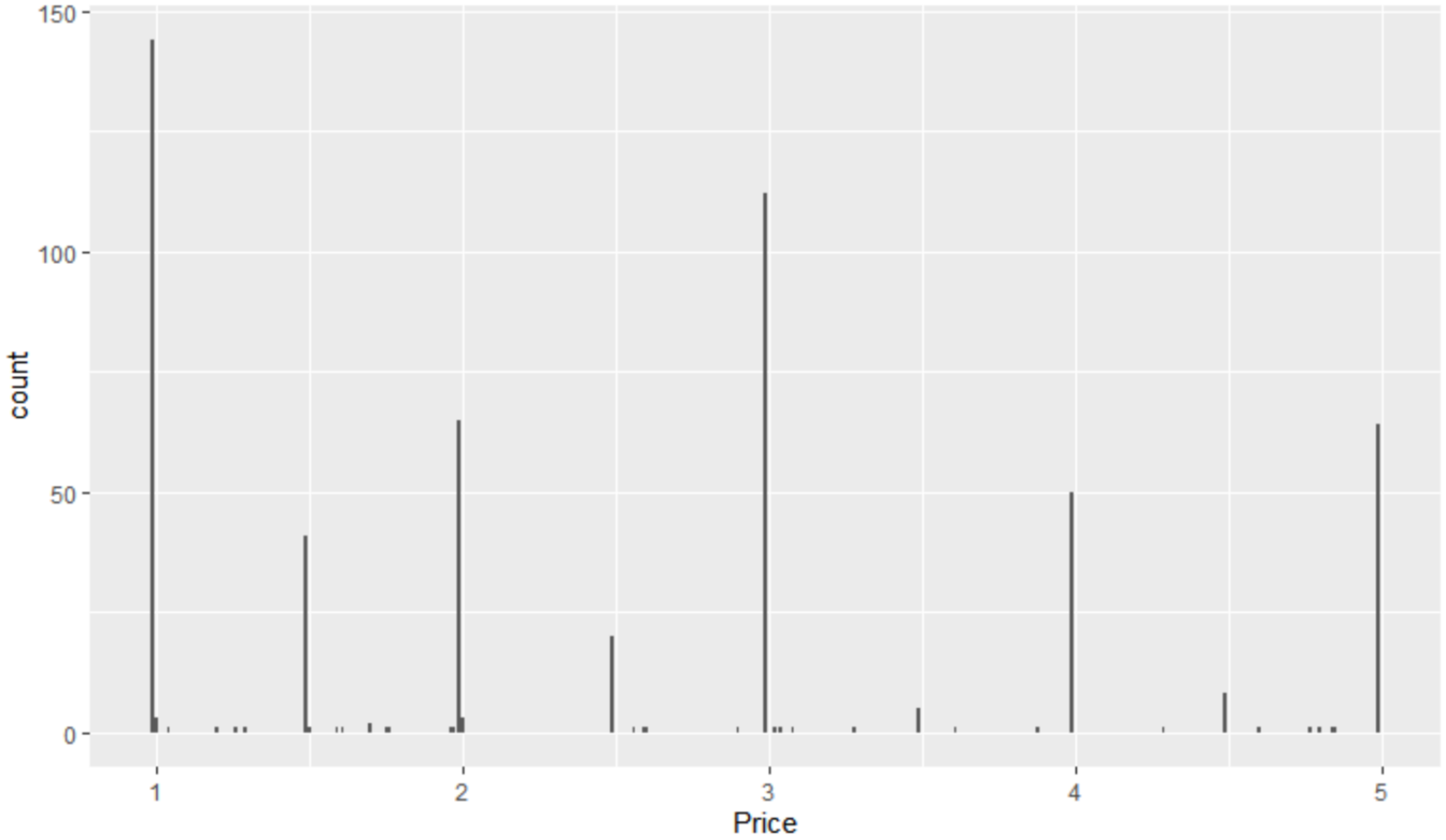
12.3 탐색적 data 분석

```
> x%>%ggplot(aes(Reviews, Installs)) + geom_point(alpha = 0.2) +
scale_x_log10() + scale_y_log10() + geom_jitter()
```



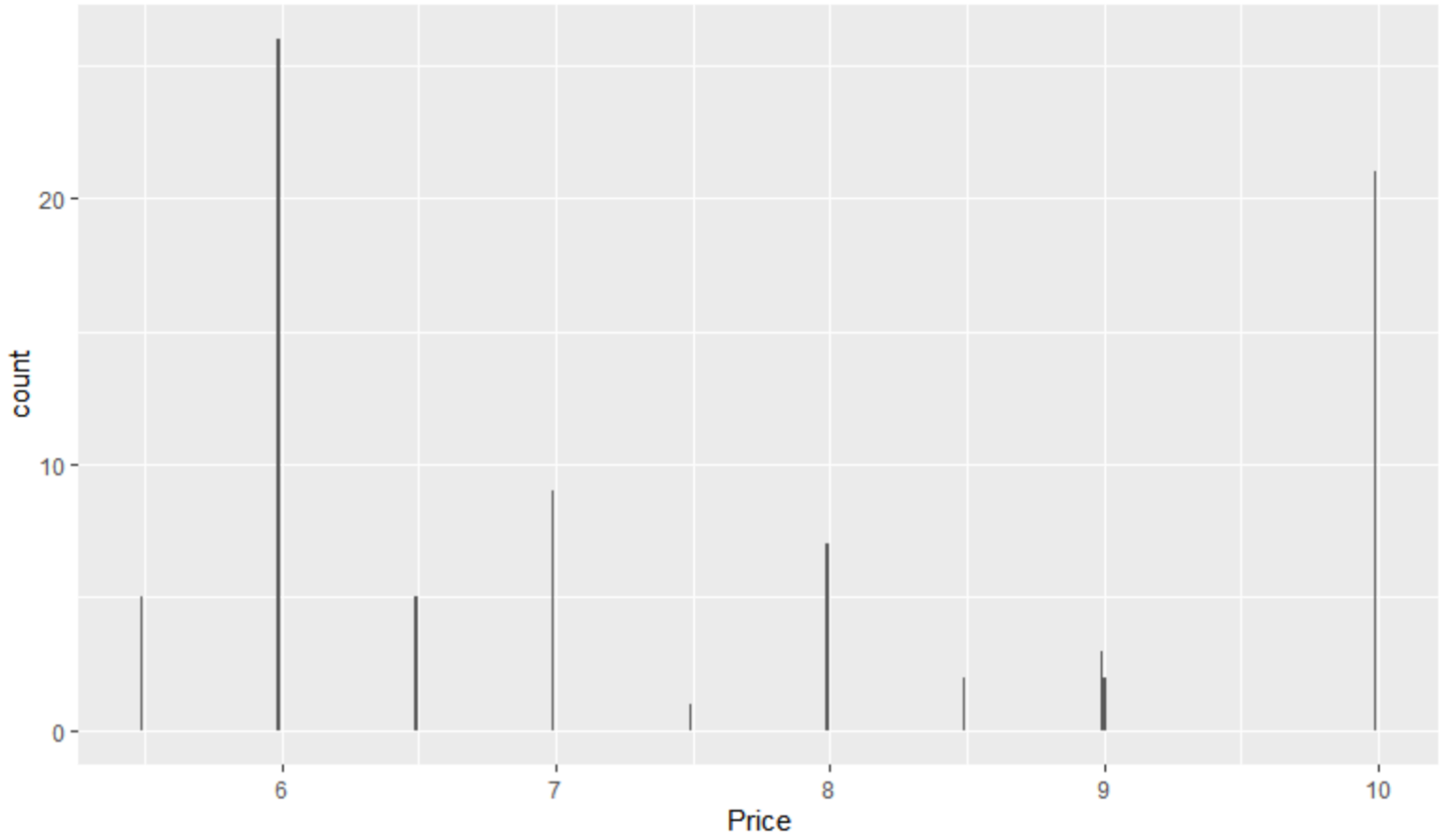
12.3 탐색적 data 분석

```
> x%>%filter(Type =="Paid" & Price < 5)%>%ggplot(aes(Price)) +
geom_histogram(binwidth = 0.01)
```



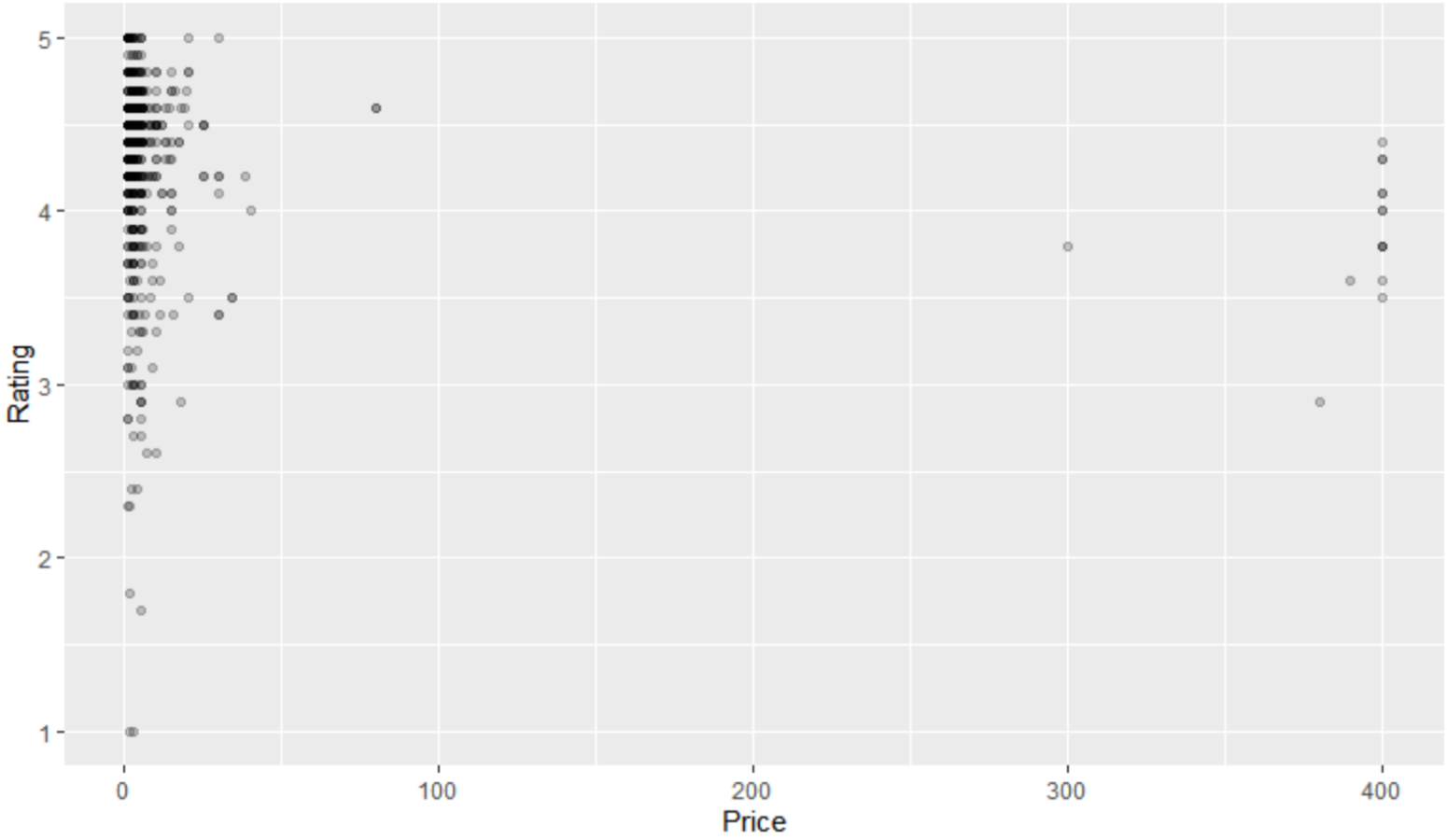
12.3 탐색적 data 분석

```
> x%>%filter(Type =="Paid" & Price < 10 & Price > 5)%>%ggplot(aes(Price)) + geom_histogram(binwidth = 0.01)
```



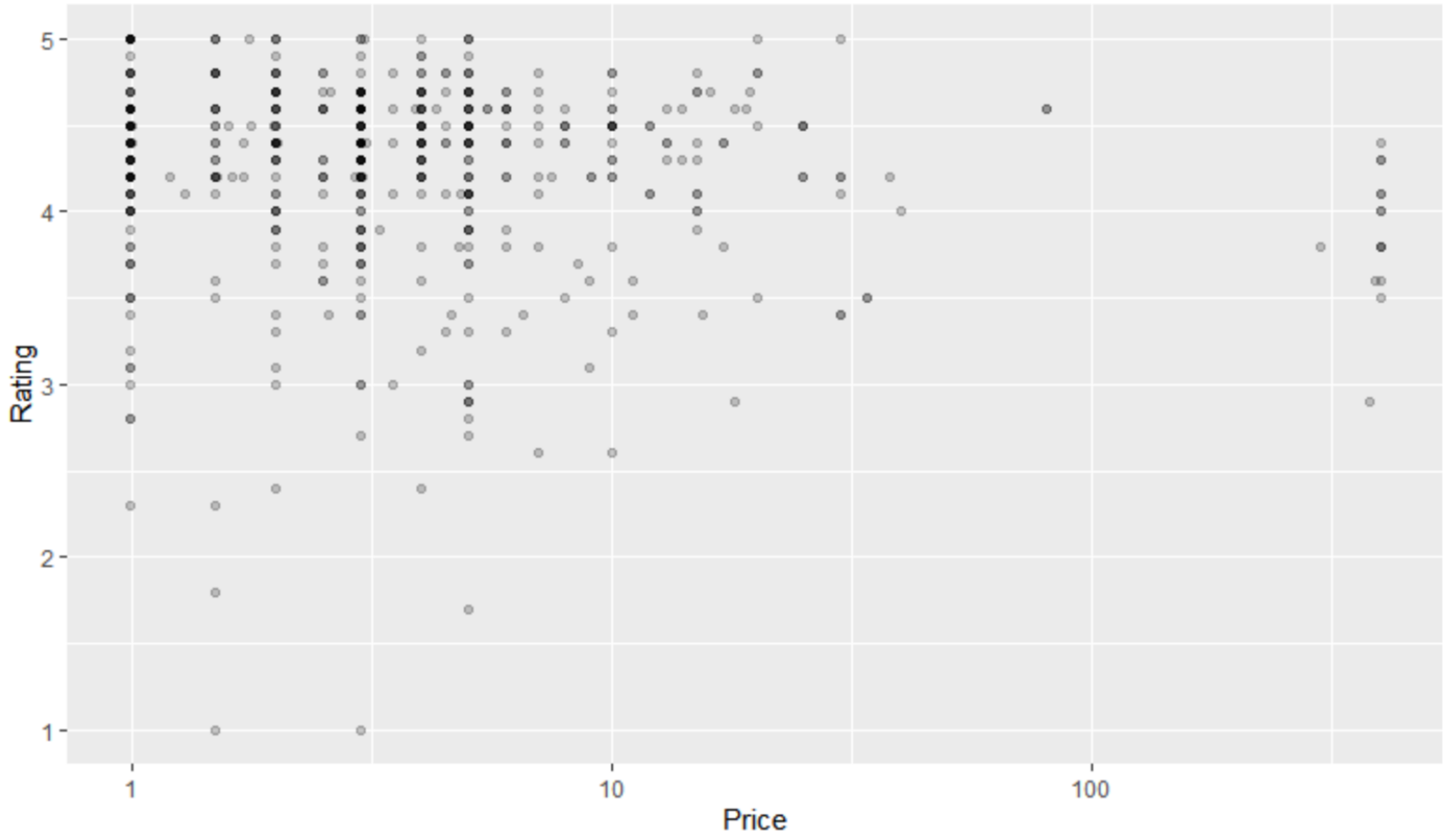
12.3 탐색적 data 분석

```
> x%>%filter(Type =="Paid")%>%ggplot(aes(Price, Rating)) +
geom_point(alpha = 0.2)
```



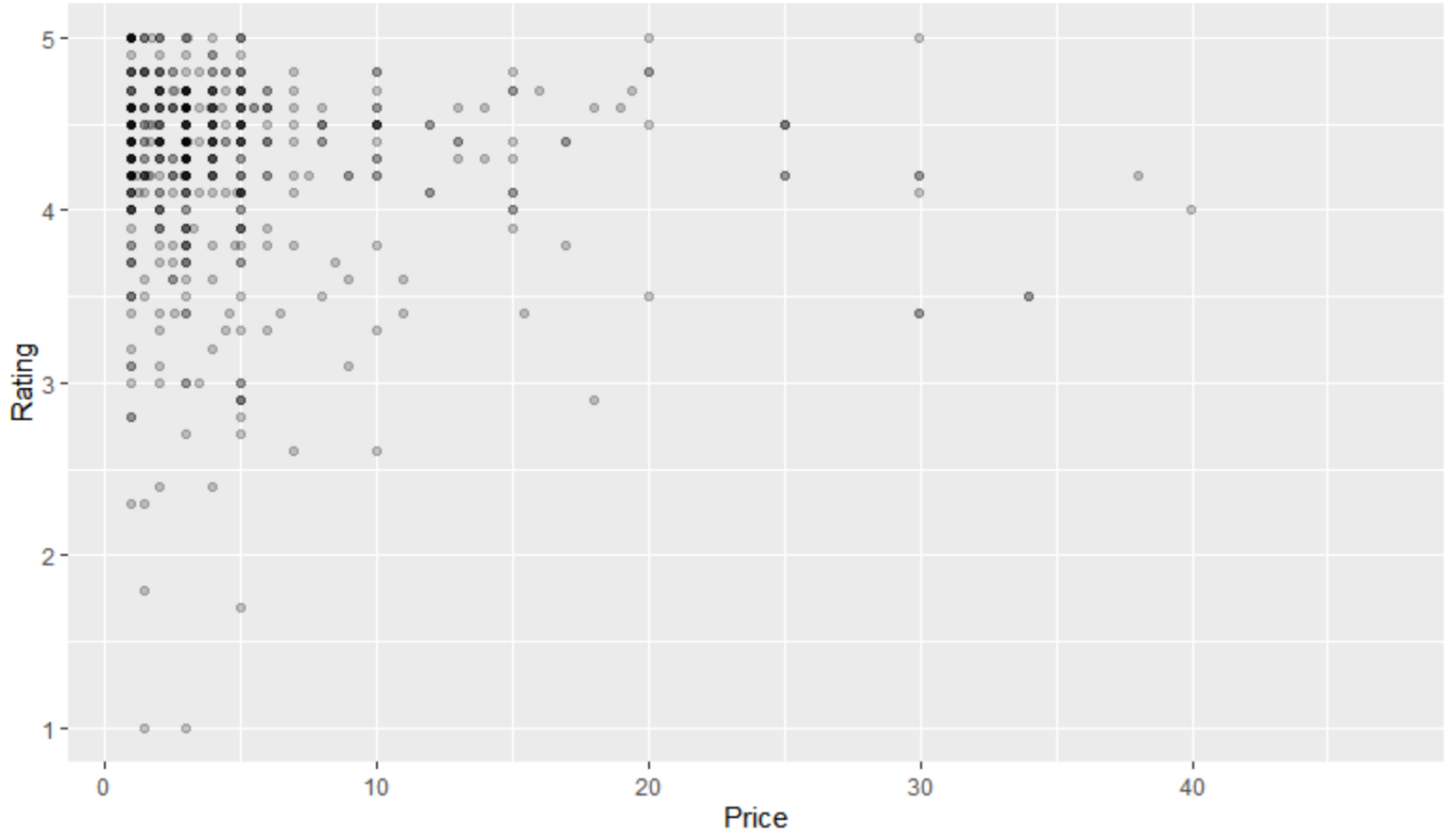
12.3 탐색적 data 분석

```
> x%>%filter(Type=="Paid")%>%ggplot(aes(Price, Rating)) +
geom_point(alpha = 0.2) + scale_x_log10()
```



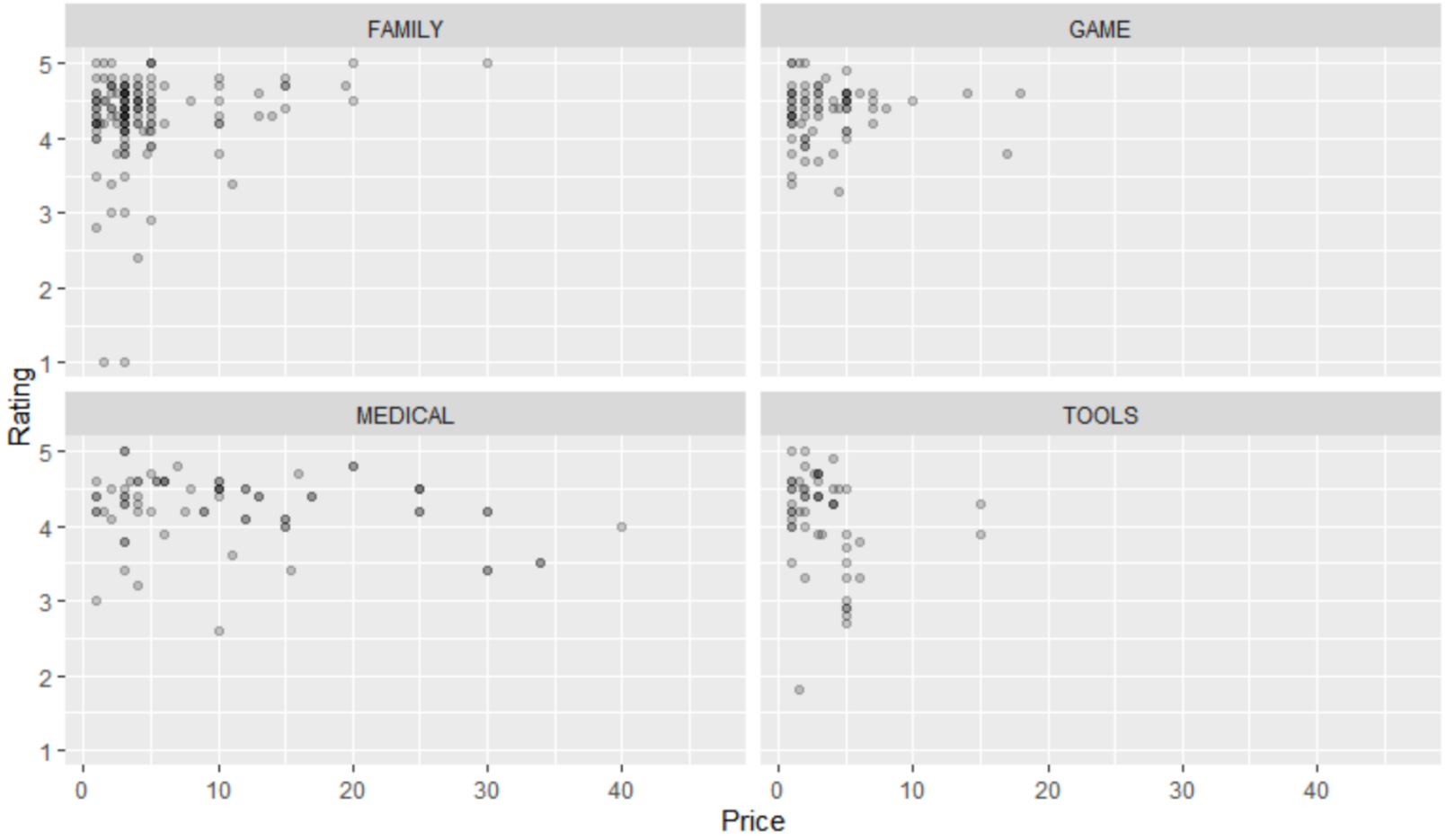
12.3 탐색적 data 분석

```
> x%>%filter(Type=="Paid" & Price < 50)%>%ggplot(aes(Price, Rating)) + geom_point(alpha = 0.2)
```



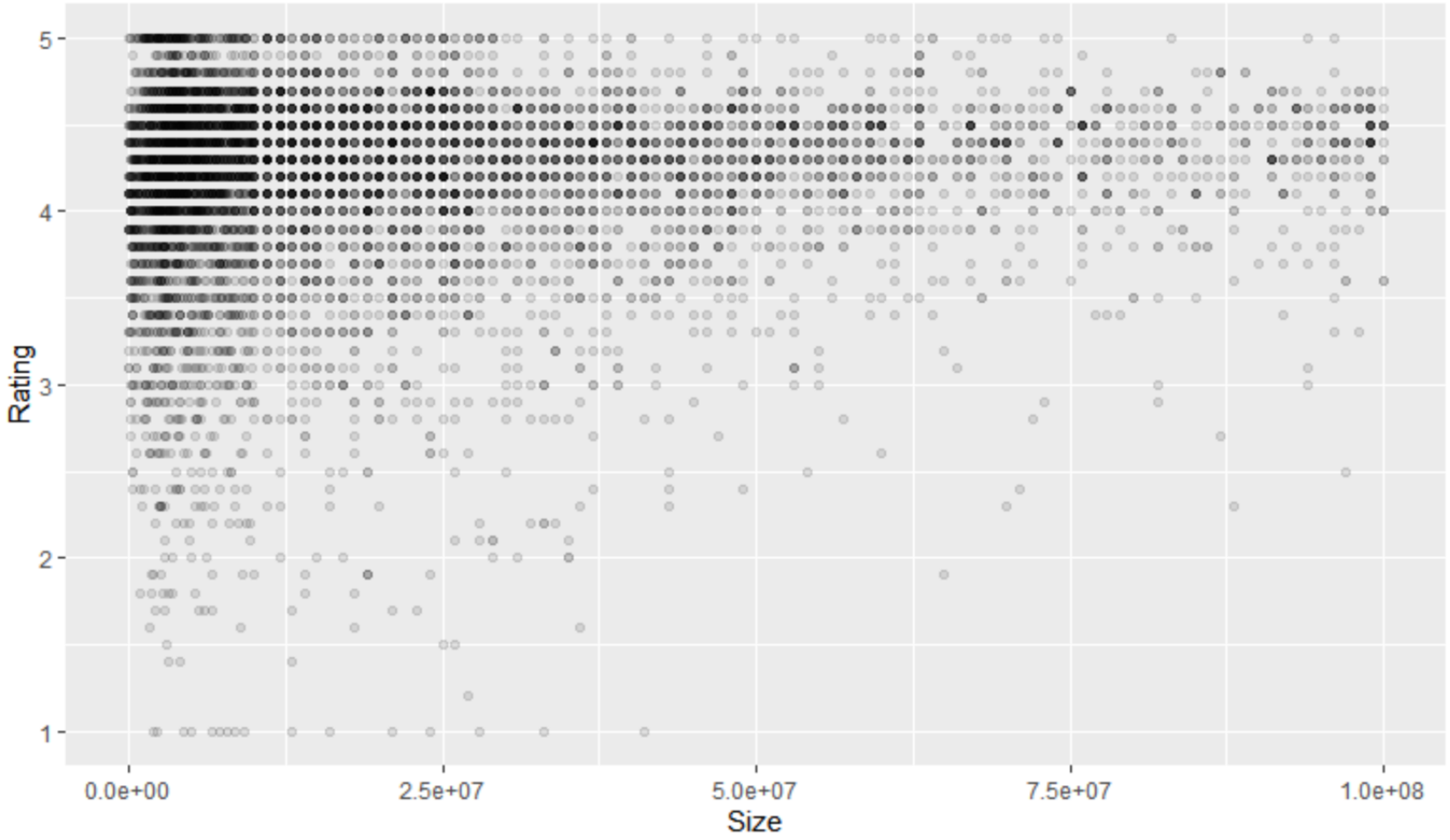
12.3 탐색적 data 분석

```
> x%>%filter(Type=="Paid" & Price < 50 & Category %in%
top4$Category)%>%ggplot(aes(Price, Rating)) + geom_point(alpha = 0.2)
+ facet_wrap(~Category)
```



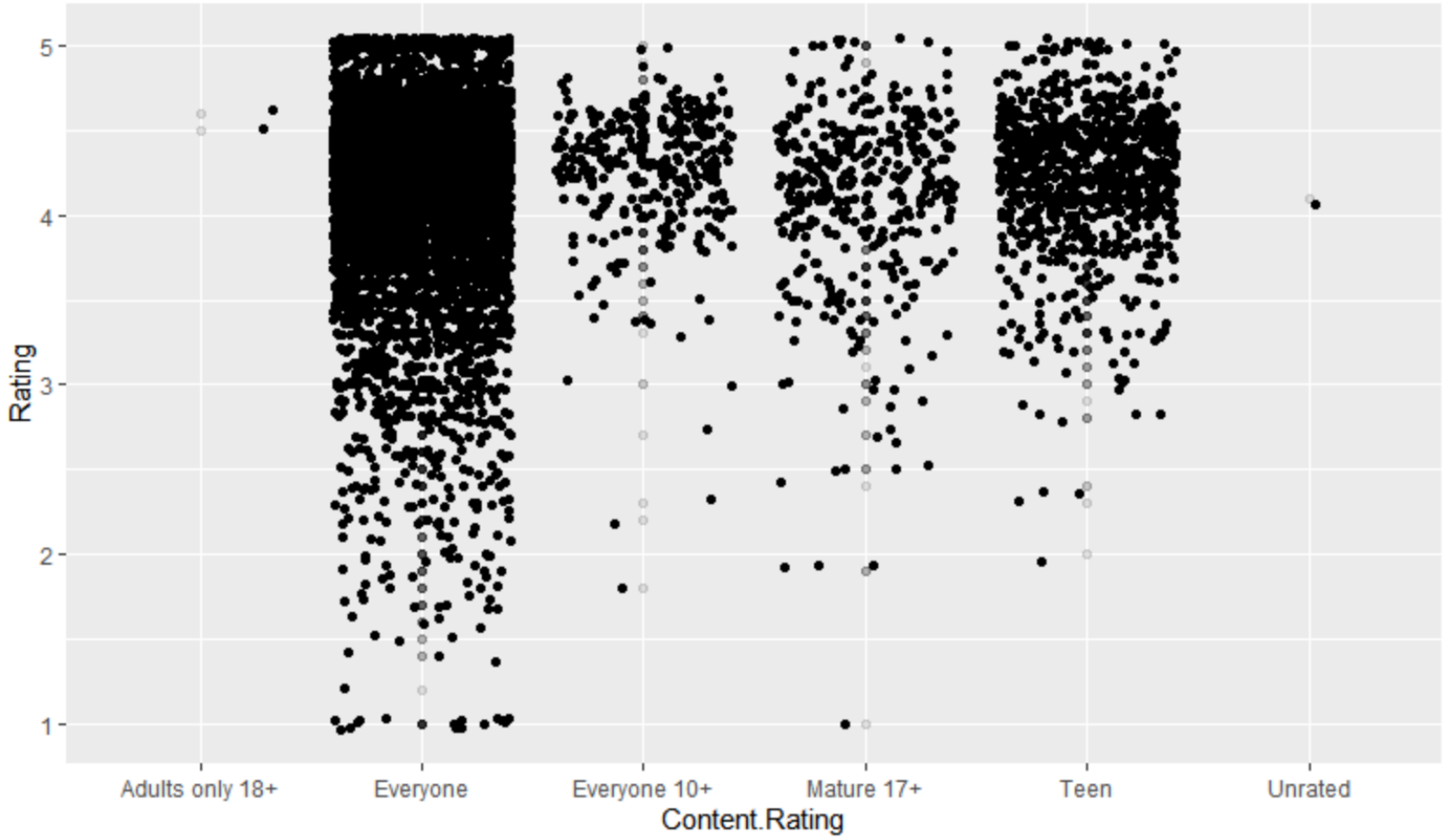
12.3 탐색적 data 분석

```
> x%>%ggplot(aes(Size, Rating)) + geom_point(alpha = 0.1)
```



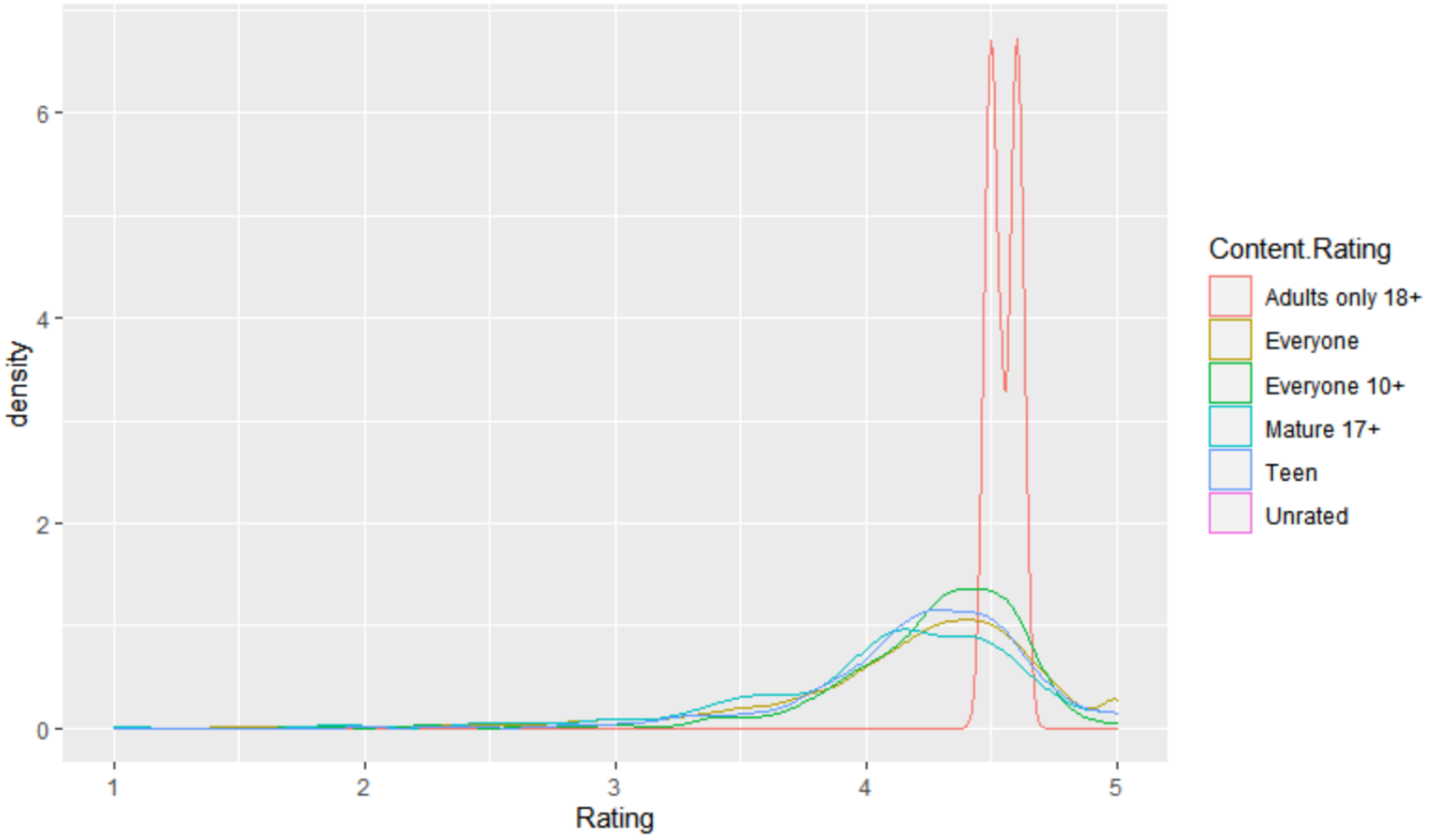
12.3 탐색적 data 분석

```
> x%>%ggplot(aes(Content.Rating, Rating)) + geom_point(alpha = 0.1) + geom_jitter()
```



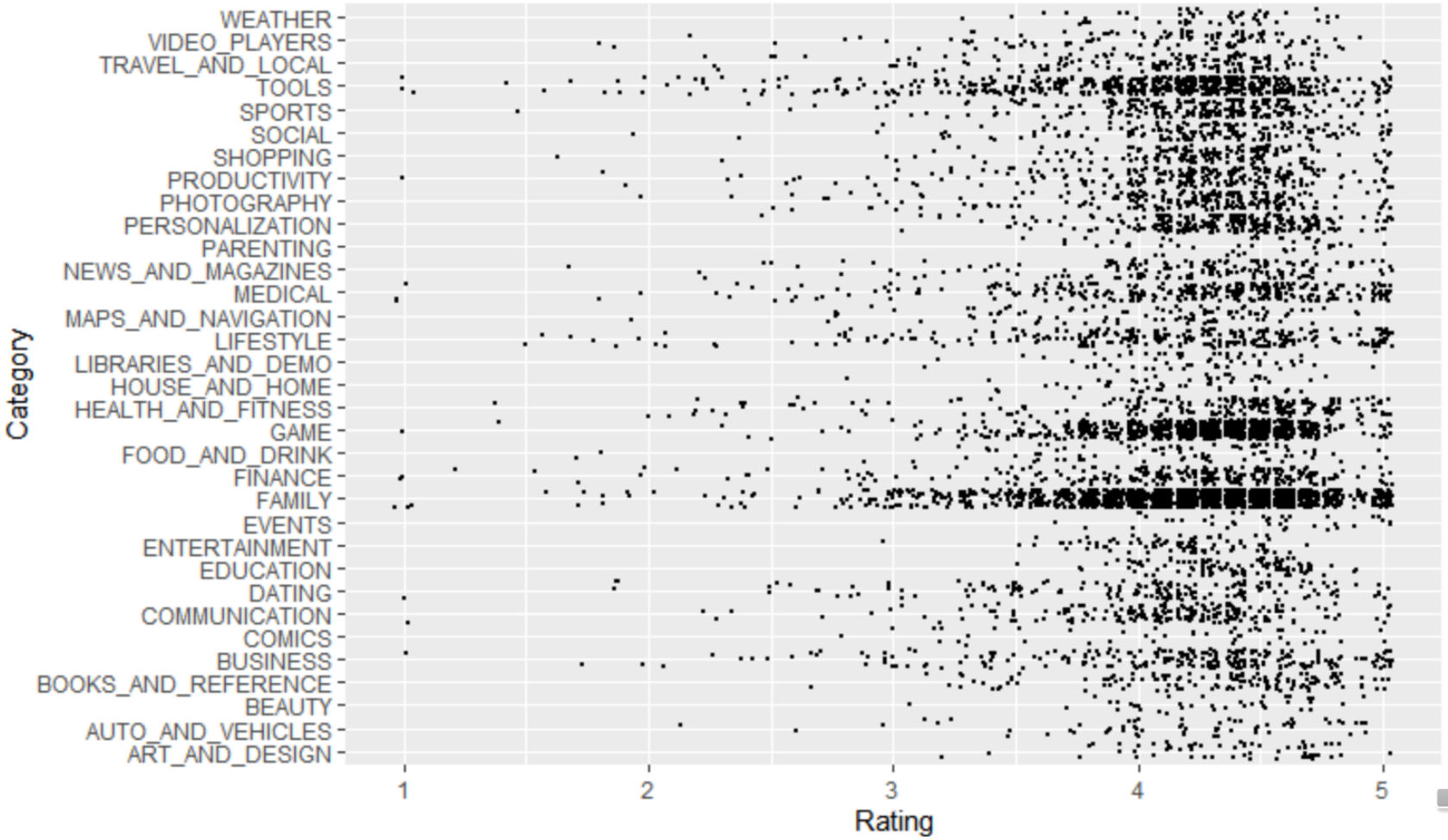
12.3 탐색적 data 분석

```
> x%>%filter(Content.Rating!= "Adults only 18 +")%>%ggplot(aes(Rating,
col = Content.Rating)) + geom_density()
```



12.3 탐색적 data 분석

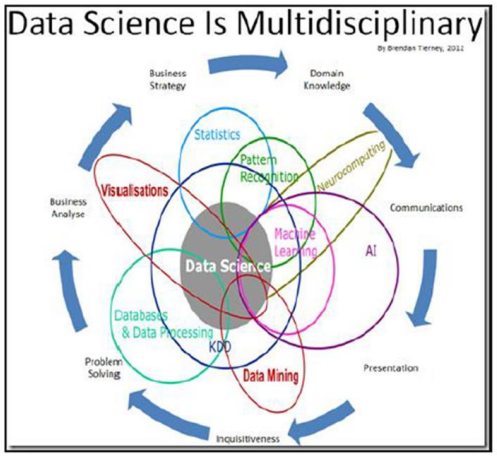
```
> x%>%ggplot(aes(Category, Rating)) + geom_point(size = 0.1,
position = "jitter") + coord_flip()
```

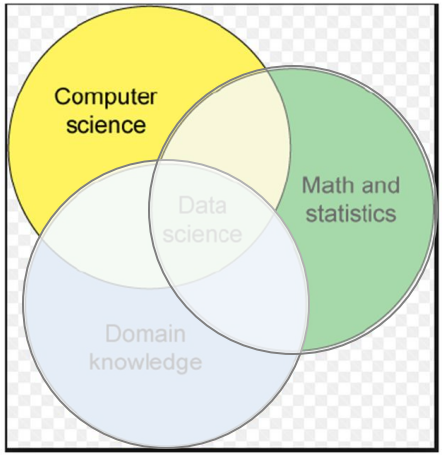


12.3 탐색적 data 분석

인터넷, IoT ,스마트, 웨어러블 시대의 도래로 빅데이터 시대에서

1. 데이터과학 분야의 전문가로서 갖추어야 할 다양한 학문적 이론과 실무지식을 익히고
2. 데이터 과학자로서 요구되는 창의성, 통찰력, 올바른 윤리의식 등을 갖추며
3. 데이터의 가치를 창출하기 위한 융합적 사고를 기반으로 데이터 분석, 설계 및 예측 능력을 갖춘 데이터 시대를 선도하는 전문가의 능력 함양





Thank you


continent
Africa
Americas
Asia
Europe
Oceania
pop
2.50e+08
5.00e+08
7.50e+08
1.00e+09
1.25e+09
90
70
60
50
40
300
1000
3000
10000
30000
gdpPercap