



14주차: 구글 플레이 앱 스토어를 이용한 실전 프로젝트

ChulSoo Park

School of Computer Engineering & Information Technology
Korea National University of Transportation



학습목표 (14주차)

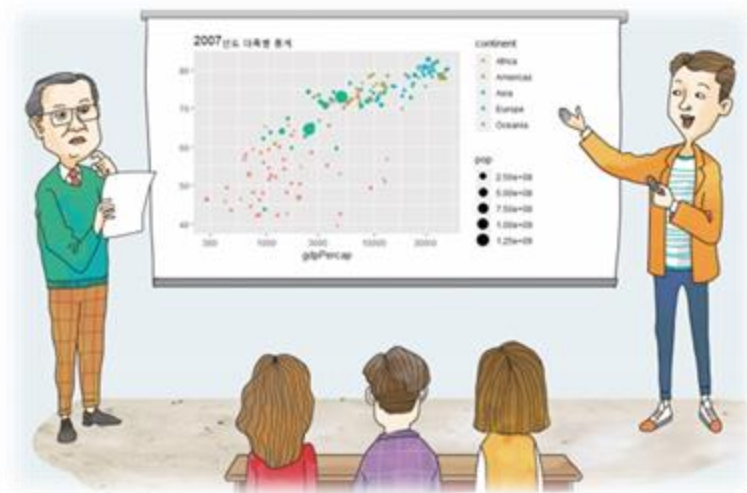
- ❖ 데이터를 이용한 실전 프로젝트 수행
- ❖ 데이터와 친해지기
- ❖ 전략적 통찰과 비즈니스에 집중한 분석
- ❖ 데이터 사이언스 메인 Process(pipeline) 정리



12

CHAPTER

실전 프로젝트



CONTENTS

12.1 프로젝트 소개

12.2 데이터 정제

12.3 탐색적 데이터 분석

12.4 모델링과 예측

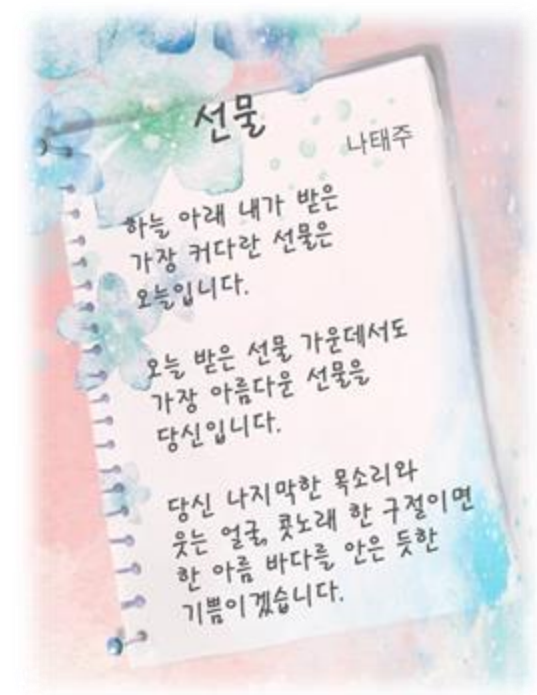
요약





■ 텍스트 데이터는 다음과 같은 독특한 성질을 가짐

- 1) 비정형의 data이다. 길이, 숫자, 특수 기호, 표현 방법 등등
- 2) 잡음이 많은 data이다. "하다", "위해" 등과 같이 불용어가 많고 구두점도 자주 나타난다.
- 3) 애매성이 많다. 사슴 같은 목, 화살 같은 세월, 거북 같은 행동 등등
- 4) 텍스트 분석에는 구문론(syntax)과 의미론(semantic)이 있다. 의미론은 단어의 의미(문맥)를 파악해서 문서를 해석해야 하므로 훨씬 어렵다. (HOT?)
- 5) 언어가 다양하다.



Hot

위키백과, 우리 모두의 백과사전.

Hot 또는 **HOT**은 다음과 같은 뜻이 있다.

- Hot는 뜨거운, 뜨겁다라는 뜻을 가진 영어 단어다.
- H.O.T.: 대한민국의 5인조 보이 그룹
- Hot (태양의 음반)
- Hot (에이브릴 라빈의 노래)



■ 영화평 분류: movie_review 데이터 모델링

- 6:4 비율로 훈련 집합과 테스트 집합으로 분할

Console C:/Rsources/ ↗

```
> # 데이터 나눔 훈련 집합(mtrain), 테스트 집합(mtest)
> train_list = createDataPartition(y= movie_review$sentiment, p = 0.6, list = FALSE)
> mtrain = movie_review[train_list, ]
> mtest = movie_review[-train_list, ]
```

- 훈련 집합에 대해 DTM 구축

```
> # 데이터 나눔 훈련 집합(mtrain), 테스트 집합(mtest)
> train_list = createDataPartition(y= movie_review$sentiment, p = 0.6, list = FALSE)
> mtrain = movie_review[train_list, ]
> mtest = movie_review[-train_list, ]
> doc = Corpus(VectorSource(mtrain$review))
> doc = tm_map(doc, content_transformer(tolower))
> doc = tm_map(doc, removeNumbers)
> doc = tm_map(doc, removeWords, stopwords('english'))
> doc = tm_map(doc, removePunctuation)
> doc = tm_map(doc, stripWhitespace)
> dtm = DocumentTermMatrix(doc)
> dim(dtm)
```

[1] 3000 36967

사전의 크기는 36871 (3000개 문서에서 36967개의 단어가 추출됨)



■ 데이터 프레임으로 변환하고 단어 구름 작성

Console C:/RSources/ ➔

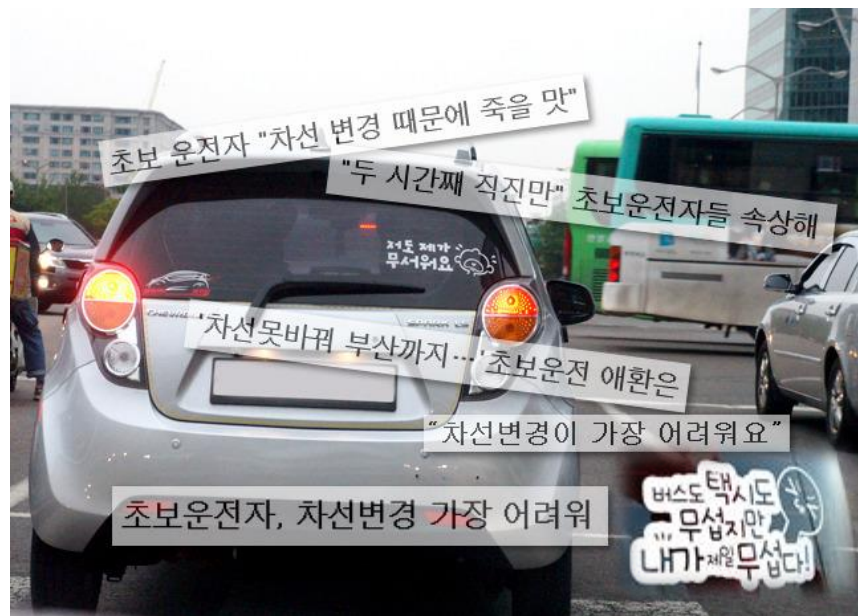
```
> m = as.matrix(dtm)
> v = sort(colsums(m), decreasing = TRUE)
> d = data.frame(word = names(v), freq = v)
> d1 = d[1:500, ] # 500개 단어 표시
> wordcloud2(d1)
```



- ✓ '있다', '통해', '및', '수' 등이 중요 자리 차지
- ✓ '데이터', '빅 데이터', '분석' 이 다른 단어로 간주되어 중요 자리 차지
- ✓ 영어 텍스트 마이닝을 한글에 적용한 한계



Preview



- “바다에 뛰어들지 않는 자는 바다를 건널 수 없다.”
- 이 책의 학습을 시작할 때 여러분은 이미 데이터 과학의 **바다에 풍덩 빠
지기로 하였다**. 차근차근 각 장의 내용을 이해하며 연습해왔다면 이제 용
기를 내어 바다를 건널 차례다.
- 실제 현장에서는 데이터로부터 무엇을 추출하고 어떻게 가공해야 하는지
, 어떤 결과가 나와야 하는지 그 누구도 말해주지 않으며 정해진 답도 존
재하지 않는다. 여러분 스스로 데이터 분석의 방향과 기법을 결정하고 결
과에 대한 객관적인 평가도 내려야 한다. **데이터를 끈기 있게 파헤쳐 정
보를 취득하고, 얻어진 지식을 최대한 활용할 방법을 모색**해야 한다.
- 데이터의 취득과 가공부터 시각화와 모델링, 그리고 예측에 이르기까지
여러분이 배운 데이터 과학의 기본기를 모두 발휘하여 데이터 과학의 넓
은 바다를 건너보자.



12.1 프로젝트 소개

- 캐글Kaggle에는 좋은 데이터들이 많이 올라와 있을 뿐 아니라 데이터 분석 사례들도 공유되어 있으므로 언제든지 참고할 수 있다. 그중 사용자들의 리뷰 Review를 많이 받은 데이터 하나를 사용하여 우리 나름대로 분석을 시도해보자. 사용할 데이터는 구글 플레이 스토어 앱 데이터다.

- <https://www.kaggle.com/lava18/google-play-store-apps>에서 googleplaystore.csv를 다운로드한다

K-FOOD 빅데이터 활용 창업경진대회

농식품

신청기간
5월 21일 (금)
- 6월 30일 (수)

참가 자격

- 만 15세 이상 대한민국 국민 누구나 (개인 또는 5인 이하 팀)
- 제출 서류 : 참가신청서, 참가서약서, 개인정보이용동의서, 서비스 개발 및 아이디어 기획서
- 참가 신청 : 온라인 접수 (대회 모집 홈페이지 www.k-foodcontest.kr 내 제출 서류 다운로드 온라인 신청)

공모 주제 : K-Food(농식품) 공공데이터를 활용한 창의적 아이디어 및 신규 비즈니스모델 발굴

1 서비스 개발

- 농식품 빅데이터를 활용한 서비스(앱 또는 웹)
- 실제 산출물 (프로토타입)

2 아이디어 기획

- 농식품 빅데이터를 분석·활용하여 제공 가능한 서비스 아이디어 또는 사업화 가능한 비즈니스 모델 기획 보고서 (PPT 20쪽 이내, 자유타입)

지원 혜택 총 상금 1,750만원

1) 시상내역 : 농림축산식품부 장관상(1점), aT 사장상(8점)

구분	서비스 개발	아이디어 기획	도상 훈격
대상	1점 / 500만원	1점 / 200만원	농림축산식품부 장관상
최우수상	1점 / 300만원	1점 / 200만원	aT 사장상
우수상	3점 / 150만원	3점 / 100만원	aT 사장상

2) 상위 2개 팀(본인명 1명)을 대상으로 9월 중 행정안전부 주관 통합 본선 참여 기회 제공

3) 경진대회 우수 수상 3개 팀에 전문가 컨설팅, 클라우드 펀딩 등 사업화 과정 지원

대회일정

신청 기간	1차 심사 발표	2차 심의 발표	본선 대회	본선 대회
5.21(목) ~6.30(수)	7.16(금)	7.22(목)	7.29(목)	8.5(목)

*시상식 8.12(목)

심사 기준

- 1) 서면 심사 (예선) : 주제 적합성, 추진 구체성, 독창성, 실현 가능성 등
- 2) 경진 대회 (본선) : 주제 적합성(30), 효과성(30), 데이터 활용도(20), 완성도(20)

문의처

esing@intw.co.kr/02-6954-1340

*대회 상세 내용은 대회 모집 페이지 www.k-foodcontest.kr 참고

■ 캐글에 공유된 구글 플레이 스토어 앱 데이터 셋

← → ↺

kaggle.com/lava18/google-play-store-apps

🔍 ☆

☰ kaggle

🏠 Home

🏆 Compete

📁 Data

🔗 Code

💬 Communities

🎓 Courses


⌵ More

Recently Viewed

🎮 Google Play Store Apps

🔍 Search

Dataset



Google Play Store Apps
Web scraped data of 10k Play Store apps for analysing the Android market.
Lavanya Gupta • updated 2 years ago (Version 6)

3603

Data

Tasks (18)

Code (690)

Discussion (74)

Activity

Metadata

Download (9 MB)

New Notebook

📊 Usability 7.1

🏷️ Tags business, computer science, internet, video games, mobile and wireless

Data Explorer

8.61 MB

📁 googleplaystore.csv

📁 googleplaystore_user_revie...

📄 license.txt

< googleplaystore.csv (1.3 MB)

📄 🔄

Detail

Compact

Column

10 of 13 columns

About this file

details of the applications on Google Play. There are 13 features thstrong strong texttextat describe a given app.. Expilo. Ed

12.1 프로젝트 소개

- 구글 플레이 스토어에 등록된 앱의 종류를 비롯하여 사용자가 앱을 다운로드 해 설치한 횟수, 앱에 대해 사용자들이 남긴 평점과 리뷰 등의 데이터가 기록되어 있다.
- 앱을 구매할 때 흔히 발생하는 평범한 자료라고 생각할 수도 있지만, 앱을 개발하거나 앱 비즈니스를 하고자 하는 사람들에게는 매우 유용한 정보가 될 수 있다.
- 플레이 스토어에서 성공적인 앱이 되려면 어떻게 하면 좋을지 영감을 얻을 수 있는 자료가 될 수 있는 것이다.



12.2 데이터 획득 및 정제

표 12-1 프로젝트에 사용할 구글 플레이 스토어 앱 데이터의 구성

변수명(열 이름)	변수형	내용
App	character	앱의 이름
Category	레벨이 33개인 범주형	앱의 종류
Rating	double	사용자들이 매긴 평점
Reviews	double	사용자들이 적은 리뷰의 개수
Size	double	앱의 크기(byte)
Installs	double	앱이 스마트폰에 설치된 횟수
Type	레벨이 2개인 factor	유료와 무료의 구분
Price	double	유료 앱인 경우 가격, 무료 = 0
Content,Rating	레벨이 112개인 범주형	연령 등급
Genres	레벨이 112개인 범주형	앱의 장르, 카테고리과 유사하며, 다중 값을 갖는 것이 차이점
Last,Updated	Date	가장 최근에 업데이트된 날짜
Current,Ver	character	앱의 현재 버전
Android,Ver	character	앱이 구동되는 안드로이드의 버전



12.2 데이터 획득 및 정제

데이터 확인

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content R	Genres	Last Updated	Current V	Android Ver	
2	Photo Editor & Candy Camera &	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	0	Everyone	Art & Des	07-Jan-18	1.0.0	4.0.3 and up	
3	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Des	15-Jan-18	2.0.0	4.0.3 and up	
4	U Launcher Lite ??FREE Live Cool	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000	Free	0	Everyone	Art & Des	01-Aug-18	1.2.4	4.0.3 and up	
5	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000	Free	0	Teen	Art & Des	08-Jun-18	Varies with device	4.2 and up	
6	Pixel Draw - Number Art Coloring	ART_AND_DESIGN	4.3	967	2.8M	100,000+	Free	0	Everyone	Art & Des	20-Jun-18	1.1	4.4 and up	
7	Paper flowers instructions	ART_AND_DESIGN	4.4	167	5.6M	50,000+	Free	0	Everyone	Art & Des	26-Mar-17	1	2.3 and up	
8	Smoke Effect Photo Maker - Smo	ART_AND_DESIGN	3.8	178	19M	50,000+	Free	0	Everyone	Art & Des	26-Apr-18	1.1	4.0.3 and up	
9	Infinite Painter	ART_AND_DESIGN	4.1	36815	29M	1,000,000	Free	0	Everyone	Art & Des	14-Jun-18	6.1.61.1	4.2 and up	
10	Garden Coloring Book	ART_AND_DESIGN	4.4	13791	33M	1,000,000	Free	0	Everyone	Art & Des	20-Sep-17	2.9.2	3.0 and up	
11	Kids Paint Free - Drawing Fun	ART_AND_DESIGN	4.7	121	3.1M	10,000+	Free	0	Everyone	Art & Des	03-Jul-18	2.8	4.0.3 and up	
12	Text on Photo - Fontee	ART_AND_DESIGN	4.4	13880	28M	1,000,000	Free	0	Everyone	Art & Des	27-Oct-17	1.0.4	4.1 and up	
13	Name Art Photo Editor - Focus r	ART_AND_DESIGN	4.4	8788	12M	1,000,000	Free	0	Everyone	Art & Des	31-Jul-18	1.0.15	4.0 and up	
14	Tattoo Name On My Photo Editc	ART_AND_DESIGN	4.2	44829	20M	10,000,000	Free	0	Teen	Art & Des	02-Apr-18	3.8	4.1 and up	
15	Mandala Coloring Book	ART_AND_DESIGN	4.6	4326	21M	100,000+	Free	0	Everyone	Art & Des	26-Jun-18	1.0.4	4.4 and up	
16	3D Color Pixel by Number - Sanc	ART_AND_DESIGN	4.4	1518	37M	100,000+	Free	0	Everyone	Art & Des	03-Aug-18	1.2.3	2.3 and up	
17	Learn To Draw Kawaii Characters	ART_AND_DESIGN	3.2	55	2.7M	5,000+	Free	0	Everyone	Art & Des	06-Jun-18	NaN	4.2 and up	
18	Photo Designer - Write your nar	ART_AND_DESIGN	4.7	3632	5.5M	500,000+	Free	0	Everyone	Art & Des	31-Jul-18	3.1	4.1 and up	
10474	Life Made WI-Fi Touchscreen Ph		1.9	19	3.0M	1,000+	Free	0	Everyone	#####	1.0.19		4.0 and up	
1229	My CookBook Pro (Ad Free)	FOOD_AND_DRINK	4.6	2129	Varies with device	10,000+	Paid	\$3.49	Everyone	Food & D	28-Jun-18	Varies with device	Varies with device	
1230	Paprika Recipe Manager	FOOD_AND_DRINK	4.1	1268	2.3M	50,000+	Paid	\$4.99	Everyone	Food & D	03-Jun-18	1.4.4	4.0 and up	
10835	Chemin (fr)	BOOKS_AND_REFER	4.8	44	619k	1,000+	Free	0	Everyone	Books & F	23-Mar-14	0.8	2.2 and up	
10836	FR Calculator	FAMILY	4	7	2.6M	500+	Free	0	Everyone	Education	18-Jun-17	1.0.0	4.1 and up	
10837	FR Forms	BUSINESS	NaN	0	9.6M	10+	Free	0	Everyone	Business	29-Sep-16	1.1.5	4.0 and up	
10838	Sya9a Maroc - FR	FAMILY	4.5	38	53M	5,000+	Free	0	Everyone	Education	25-Jul-17	1.48	4.1 and up	
10839	Fr. Mike Schmitz Audio Teaching	FAMILY	5	4	3.6M	100+	Free	0	Everyone	Education	06-Jul-18	1	4.1 and up	
10840	Parkinson Exercices FR	MEDICAL	NaN	3	9.5M	1,000+	Free	0	Everyone	Medical	20-Jan-17	1	2.2 and up	
10841	The SCP Foundation DB fr nn5n	BOOKS_AND_REFER	4.5	114	Varies with device	1,000+	Free	0	Mature 17	Books & F	19-Jan-15	Varies with device	Varies with device	
10842	iHoroscope - 2018 Daily Horosc	LIFESTYLE	4.5	398307	19M	10,000,000	Free	0	Everyone	Lifestyle	25-Jul-18	Varies with device	Varies with device	

12.2 데이터 획득 및 정제

- data 취득 : 엑셀로 읽어 다른 이름으로 저장
 - C:/rdata/googleplaystore.csv → C:/rdata/googleplaystore1.csv

Console C:/RSources/ 

```
> # data를 있는 그대로 읽음(as.is = TRUE 옵션은 범주형 변환 비활성화)
> x = read.csv("C:/rdata/googleplaystore.csv", header = TRUE, sep = ",", as.is = TRUE)
경고메시지(들):
In scan(file = file, what = what, sep = sep, quote = quote, dec = dec, :
  따옴표로 묶인 문자열내에 EOF가 있습니다
> # data를 있는 그대로 읽음(as.is = TRUE 옵션은 범주형 변환 비활성화)
> x = read.csv("C:/rdata/googleplaystore1.csv", header = TRUE, sep = ",", as.is = TRUE)
```



12.2 데이터 획득 및 정제

■ 취득한 data 구조 등 확인

```
> str(x)
'data.frame': 10841 obs. of 13 variables:
 $ App      : chr "Photo Editor & Candy Camera & Grid & ScrapBook" "Coloring book moana" "U Laun
cher Lite ??FREE Live Cool Themes, Hide Apps" "Sketch - Draw & Paint" ...
 $ Category : chr "ART_AND_DESIGN" "ART_AND_DESIGN" "ART_AND_DESIGN" "ART_AND_DESIGN" ...
 $ Rating   : chr "4.1" "3.9" "4.7" "4.5" ...
 $ Reviews  : chr "159" "967" "87510" "215644" ...
 $ Size     : chr "19M" "14M" "8.7M" "25M" ...
 $ Installs : chr "10,000+" "500,000+" "5,000,000+" "50,000,000+" ...
 $ Type     : chr "Free" "Free" "Free" "Free" ...
 $ Price    : chr "0" "0" "0" "0" ...
 $ Content.Rating: chr "Everyone" "Everyone" "Everyone" "Teen" ...
 $ Genres    : chr "Art & Design" "Art & Design;Pretend Play" "Art & Design" "Art & Design" ...
 $ Last.Updated : chr "07-Jan-18" "15-Jan-18" "01-Aug-18" "08-Jun-18" ...
 $ Current.Ver : chr "1.0.0" "2.0.0" "1.2.4" "Varies with device" ...
 $ Android.Ver : chr "4.0.3 and up" "4.0.3 and up" "4.0.3 and up" "4.2 and up" ...
```

```
> head(x,10)
```

	App	Category	Rating	Reviews	Size	Installs
1	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10,000+
2	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+
3	U Launcher Lite ??FREE Live Cool Themes, Hide Apps	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+
4	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+
5	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8M	100,000+
6	Paper flowers instructions	ART_AND_DESIGN	4.4	167	5.6M	50,000+
7	Smoke Effect Photo Maker - Smoke Editor	ART_AND_DESIGN	3.8	178	19M	50,000+
8	Infinite Painter	ART_AND_DESIGN	4.1	36815	29M	1,000,000+
9	Garden Coloring Book	ART_AND_DESIGN	4.4	13791	33M	1,000,000+
10	Kids Paint Free - Drawing Fun	ART_AND_DESIGN	4.7	121	3.1M	10,000+

	Type	Price	Content.Rating	Genres	Last.Updated	Current.Ver	Android.Ver
1	Free	0	Everyone	Art & Design	07-Jan-18	1.0.0	4.0.3 and up
2	Free	0	Everyone	Art & Design;Pretend Play	15-Jan-18	2.0.0	4.0.3 and up
3	Free	0	Everyone	Art & Design	01-Aug-18	1.2.4	4.0.3 and up
4	Free	0	Teen	Art & Design	08-Jun-18	Varies with device	4.2 and up
5	Free	0	Everyone	Art & Design;Creativity	20-Jun-18	1.1	4.4 and up
6	Free	0	Everyone	Art & Design	26-Mar-17	1	2.3 and up
7	Free	0	Everyone	Art & Design	26-Apr-18	1.1	4.0.3 and up
8	Free	0	Everyone	Art & Design	14-Jun-18	6.1.61.1	4.2 and up
9	Free	0	Everyone	Art & Design	20-Sep-17	2.9.2	3.0 and up
10	Free	0	Everyone	Art & Design;Creativity	03-Jul-18	2.8	4.0.3 and up



12.2 데이터 획득 및 정제

■ data 정제

- 가격에 화폐 단위 확인 및 삭제 (화폐 단위가 여러 개 경우 ?)

```
Console C:/RSources/
> # 가격(price)에 화폐 단위($) 확인 및 제거
> x[1227:1230,]
```

	App	Category	Rating	Reviews	Size	Installs	Type
1227	Allrecipes Dinner Spinner	FOOD_AND_DRINK	4.5	61881	Varies with device	5,000,000+	Free
1228	My CookBook Pro (Ad Free)	FOOD_AND_DRINK	4.6	2129	Varies with device	10,000+	Paid
1229	Paprika Recipe Manager	FOOD_AND_DRINK	4.1	1268	2.3M	50,000+	Paid
1230	Yummly Recipes & Shopping List	FOOD_AND_DRINK	4.5	91359	27M	1,000,000+	Free

	Price	Content.Rating	Genres	Last.Updated	Current.Ver	Android.Ver
1227	0	Everyone	Food & Drink	10-Apr-18	Varies with device	Varies with device
1228	\$3.49	Everyone	Food & Drink	28-Jun-18	Varies with device	Varies with device
1229	\$4.99	Everyone	Food & Drink	03-Jun-18	1.4.4	4.0 and up
1230	0	Everyone	Food & Drink	05-Jul-18	2.0.3	4.4 and up

```
Console C:/RSources/
> x$Price = str_replace(x$Price, '\\$', '')
> x[1227:1230,]
```

	App	Category	Rating	Reviews	Size	Installs
1227	Allrecipes Dinner Spinner	FOOD_AND_DRINK	4.5	61881	Varies with device	5,000,000+
1228	My CookBook Pro (Ad Free)	FOOD_AND_DRINK	4.6	2129	Varies with device	10,000+
1229	Paprika Recipe Manager	FOOD_AND_DRINK	4.1	1268	2.3M	50,000+
1230	Yummly Recipes & Shopping List	FOOD_AND_DRINK	4.5	91359	27M	1,000,000+

	Type	Price	Content.Rating	Genres	Last.Updated	Current.Ver	Android.Ver
1227	Free	0	Everyone	Food & Drink	10-Apr-18	Varies with device	Varies with device
1228	Paid	3.49	Everyone	Food & Drink	28-Jun-18	Varies with device	Varies with device
1229	Paid	4.99	Everyone	Food & Drink	03-Jun-18	1.4.4	4.0 and up
1230	Free	0	Everyone	Food & Drink	05-Jul-18	2.0.3	4.4 and up



12.2 데이터 획득 및 정제

■ data 정제

- 데이터 중 일부 누락으로 삭제 처리 (많은 경우는, 사용하려면 ?)

```

Console C:/RSources/
> x[10472:10474,]

```

	App	Category	Rating	Reviews	Size	Installs
10472	Xposed Wi-Fi-Pwd	PERSONALIZATION	3.5	1042	404k	100,000+
10473	Life Made WI-Fi Touchscreen Photo Frame		1.9	19	3.0M	1,000+ Free
10474	osmino Wi-Fi: free WiFi	TOOLS	4.2	134203	4.1M	10,000,000+

```

Type Price Content.Rating Genres Last.Updated Current.Ver Android.Ver
10472 Free 0 Everyone Personalization 05-Aug-14 3.0.0 4.0.3 and up
10473 0 Everyone 11-Feb-18 1.0.19 4.0 and up
10474 Free 0 Everyone Tools 07-Aug-18 6.06.14 4.4 and up
> x[310:312,]

```

	App	Category	Rating	Reviews
310	Truy 沼 넷 vui T책 Qu 梳 畚	COMICS	4.5	144
311	Comic Es - Shoyo manga / love comics free of charge	COMICS	3.9	2181
312	comico Popular Original Cartoon Updated Everyday	Comico COMICS	3.2	93965

```

Size Installs Type Price Content.Rating Genres Last.Updated Current.Ver
310 4.7M 10,000+ Free 0 Everyone Comics 19-Jul-18 3
311 100,000+ Free 0 Teen Comics 05-Mar-18 1.2.12 4.0.3 and up
312 15M 5,000,000+ Free 0 Teen Comics 03-Jul-18 6.3.0
Android.Ver
310 4.0.3 and up
311
312 4.0.3 and up

```

```

Console C:/RSources/
> x = x[-10473, ]
> x[10472:10474,]

```

	App	Category	Rating	Reviews	Size	Installs	Type	Price
10472	Xposed Wi-Fi-Pwd	PERSONALIZATION	3.5	1042	404k	100,000+	Free	0
10474	osmino Wi-Fi: free WiFi	TOOLS	4.2	134203	4.1M	10,000,000+	Free	0
10475	Sat-Fi Voice	COMMUNICATION	3.4	37	14M	1,000+	Free	0

```

Content.Rating Genres Last.Updated Current.Ver Android.Ver
10472 Everyone Personalization 05-Aug-14 3.0.0 4.0.3 and up
10474 Everyone Tools 07-Aug-18 6.06.14 4.4 and up
10475 Everyone Communication 21-Nov-14 2.2.1.5 2.2 and up

```



12.2 데이터 획득 및 정제

■ data 정제

- size에 문자가 입력된 경우 NA로 표시
- size 단위 통일 Mb, Kb byte로 통일 (Gb, Tb 이면 ?)

```
> x[207:212,]
      App Category Rating Reviews Size Installs Type
207      Call Blocker BUSINESS      4.6  188841      3.2M 5,000,000+ Free
208 Jobs in Alabama - Jobs in Alba BUSINESS      4.1  11622 Varies with device 5,000,000+ Free
209      Square Point of Sale - POS BUSINESS      4.6  95912 Varies with device 5,000,000+ Free
210      Plugin:AOT v5.0 BUSINESS      3.1   4034      23k  100,000+ Free
211      Kariyer.net BUSINESS      3.9  45964      16M 1,000,000+ Free
212      SEEK Job Search BUSINESS      4.3  14955 Varies with device 1,000,000+ Free
  Price Content.Rating Genres Last.Updated Current.Ver Android.Ver
207      0      Everyone Business 21-Jun-18      1.1.13      4.0 and up
208      0      Everyone Business 26-Jul-18 Varies with device Varies with device
209      0      Everyone Business 30-Jul-18 Varies with device Varies with device
210      0      Everyone Business 11-Sep-15 3.0.1.11 (Build 311)      2.2 and up
211      0      Everyone Business 18-Jul-18      5.1.5      4.1 and up
212      0      Everyone Business 30-Jul-18 Varies with device Varies with device
```

```
> # size에 이상 data "Varies with device" NA로 처리
> # Mega byte M은 x 1,000,000
> # Kilo byte k은 x 1,000
> x$Size = sub("Varies with device", NA, x$Size)
> x$Size = sub("M", "e6", x$Size, fixed = TRUE)
> x$Size = sub("k", "e3", x$Size, fixed = TRUE)
> x[207:212,]
```

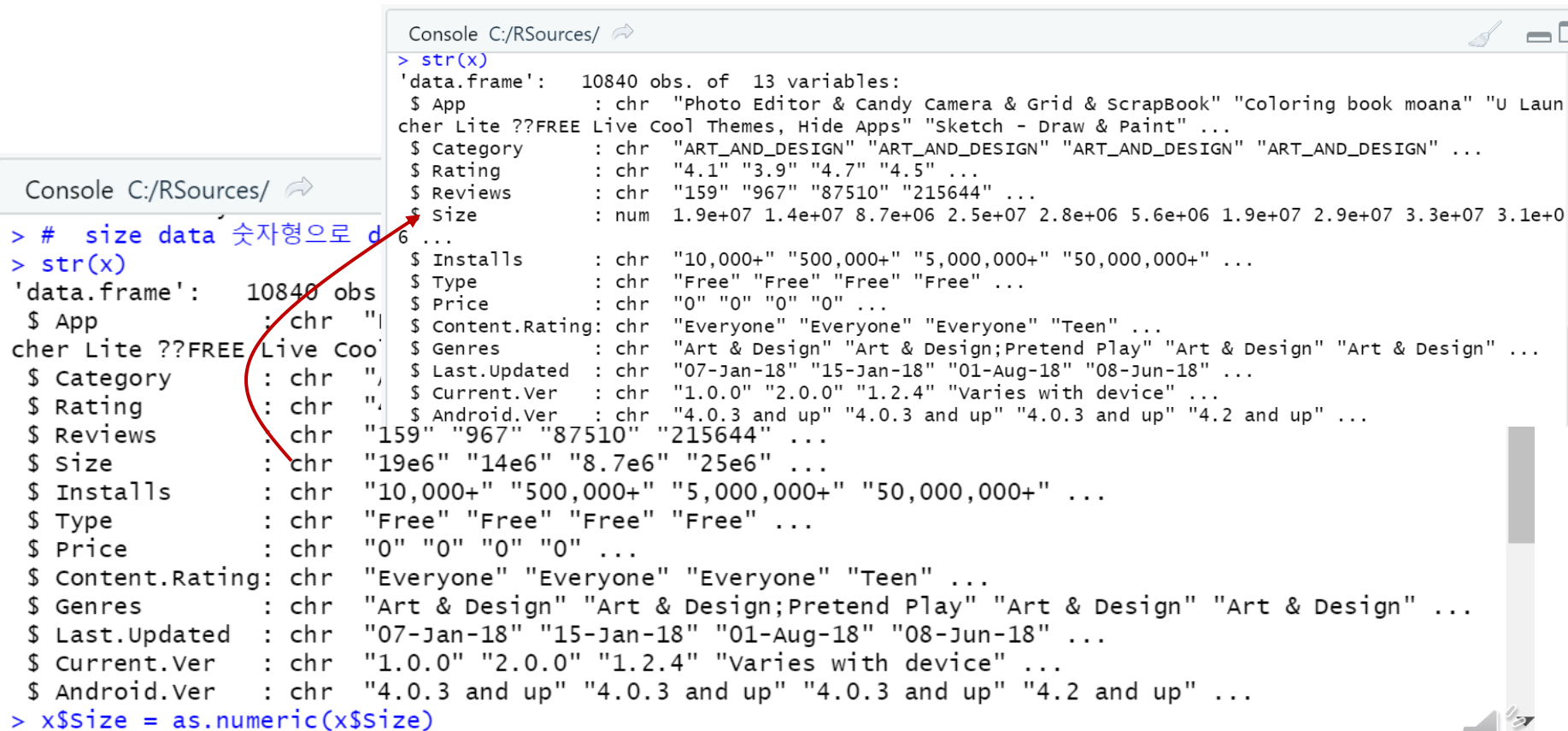
```
      App Category Rating Reviews Size Installs Type Price
207      Call Blocker BUSINESS      4.6  188841 3.2e6 5,000,000+ Free      0
208 Jobs in Alabama - Jobs in Alba BUSINESS      4.1  11622 <NA> 5,000,000+ Free      0
209      Square Point of Sale - POS BUSINESS      4.6  95912 <NA> 5,000,000+ Free      0
210      Plugin:AOT v5.0 BUSINESS      3.1   4034 23e3  100,000+ Free      0
211      Kariyer.net BUSINESS      3.9  45964 16e6 1,000,000+ Free      0
212      SEEK Job Search BUSINESS      4.3  14955 <NA> 1,000,000+ Free      0
  Content.Rating Genres Last.Updated Current.Ver Android.Ver
207      Everyone Business 21-Jun-18      1.1.13      4.0 and up
208      Everyone Business 26-Jul-18 Varies with device Varies with device
209      Everyone Business 30-Jul-18 Varies with device Varies with device
210      Everyone Business 11-Sep-15 3.0.1.11 (Build 311)      2.2 and up
211      Everyone Business 18-Jul-18      5.1.5      4.1 and up
212      Everyone Business 30-Jul-18 Varies with device Varies with device
```



12.2 데이터 획득 및 정제

■ data 정제

- size를 모두 정제하고 최종적으로 숫자형 data로 변환



```

Console C:/RSources/
> # size data 숫자형으로 d
> str(x)
'data.frame': 10840 obs. of 13 variables:
 $ App      : chr "Photo Editor & Candy Camera & Grid & ScrapBook" "Coloring book moana" "U Laun
cher Lite ??FREE Live Cool Themes, Hide Apps" "Sketch - Draw & Paint" ...
 $ Category : chr "ART_AND_DESIGN" "ART_AND_DESIGN" "ART_AND_DESIGN" "ART_AND_DESIGN" ...
 $ Rating   : chr "4.1" "3.9" "4.7" "4.5" ...
 $ Reviews  : chr "159" "967" "87510" "215644" ...
 $ Size     : num 1.9e+07 1.4e+07 8.7e+06 2.5e+07 2.8e+06 5.6e+06 1.9e+07 2.9e+07 3.3e+07 3.1e+0
6 ...
 $ Installs : chr "10,000+" "500,000+" "5,000,000+" "50,000,000+" ...
 $ Type     : chr "Free" "Free" "Free" "Free" ...
 $ Price    : chr "0" "0" "0" "0" ...
 $ Content.Rating: chr "Everyone" "Everyone" "Everyone" "Teen" ...
 $ Genres   : chr "Art & Design" "Art & Design;Pretend Play" "Art & Design" "Art & Design" ...
 $ Last.Updated : chr "07-Jan-18" "15-Jan-18" "01-Aug-18" "08-Jun-18" ...
 $ Current.Ver : chr "1.0.0" "2.0.0" "1.2.4" "Varies with device" ...
 $ Android.Ver : chr "4.0.3 and up" "4.0.3 and up" "4.0.3 and up" "4.2 and up" ...

Console C:/RSources/
> x$Size = as.numeric(x$Size)
  
```

12.2 데이터 획득 및 정제

■ data 정제

- installs의 일정 이상을 뜻하는 + 기호와 천 단위 구분 “,” 등의 처리
- str_replace 함수를 사용하여 빈 문자로 치환

Console C:/Rsources/ ↗

```
> x[208:210,]
```

	App	Category	Rating	Reviews	Size	Installs	Type	Price
208	Jobs in Alabama - Jobs in Alba	BUSINESS	4.1	11622	NA	5,000,000+	Free	0
209	Square Point of Sale - POS	BUSINESS	4.6	95912	NA	5,000,000+	Free	0
210	Plugin:AOT v5.0	BUSINESS	3.1	4034	23000	100,000+	Free	0

	Content.Rating	Genres	Last.Updated	Current.Ver	Android.Ver
208	Everyone	Business	26-Jul-18	Varies with device	Varies with device
209	Everyone	Business	30-Jul-18	Varies with device	Varies with device
210	Everyone	Business	11-Sep-15	3.0.1.11 (Build 311)	2.2 and up

```
> x$Installs = str_replace(x$Installs, '\\\\+', '')
> x$Installs = str_replace_all(x$Installs, ',', '')
> x[208:210,]
```

	App	Category	Rating	Reviews	Size	Installs	Type	Price
208	Jobs in Alabama - Jobs in Alba	BUSINESS	4.1	11622	NA	5000000	Free	0
209	Square Point of Sale - POS	BUSINESS	4.6	95912	NA	5000000	Free	0
210	Plugin:AOT v5.0	BUSINESS	3.1	4034	23000	100000	Free	0

	Content.Rating	Genres	Last.Updated	Current.Ver	Android.Ver
208	Everyone	Business	26-Jul-18	Varies with device	Varies with device
209	Everyone	Business	30-Jul-18	Varies with device	Varies with device
210	Everyone	Business	11-Sep-15	3.0.1.11 (Build 311)	2.2 and up



■ data 정제

- installs 를 모두 정제하고 최종적으로 숫자형 data로 변환

```

Console C:/Rsources/
$ Category      : chr  "ART_AND_DESIGN" "ART_AND_DESIGN" "ART_AND_DESIGN" "ART_AND_DESIGN" ...
$ Rating        : chr  "4.1" "3.9" "4.7" "4.5" ...
$ Reviews       : chr  "159" "967" "87510" "215644" ...
$ Size          : num  1.9e+07 1.4e+07 8.7e+06 2.5e+07 2.8e+06 5.6e+06 1.9e+07 2.9e+07 3.3e+07 3.1e+0
6 ...
$ Installs      : chr  "10000" "500000" "5000000" "50000000" ...
$ Type          : chr  "Free" "Free" "Free" "Free" ...
$ Price         : chr  "0" "0" "0" "0" ...
$ Content.Rating: chr  "Everyone" "Everyone" "Everyone" "Teen" ...
$ Genres        : chr  "Art & Design" "Art & Design;Pretend Play" "Art & Design" "Art & Design" ...
$ Last.Updated  : chr  "07-Jan-18" "15-Jan-18" "01-Aug-18" "08-Jun-18" ...
$ Current.Ver   : chr  "1.0.0" "2.0.0" "1.2.4" "Varies with device" ...
$ Android.Ver   : chr  "4.0.3 and up" "4.0.3 and up" "4.0.3 and up" "4.2 and up" ...
> x$Installs = as.numeric(x$Installs)
경고메시지(들):
강제형변환에 의해 생성된 NA 입니다
> str(x)
'data.frame': 10840 obs. of 13 variables:
 $ App          : chr  "Photo Editor & Candy Camera & Grid & ScrapBook" "Coloring book moana" "U Laun
cher Lite ??FREE Live Cool Themes, Hide Apps" "Sketch - Draw & Paint" ...
 $ Category     : chr  "ART_AND_DESIGN" "ART_AND_DESIGN" "ART_AND_DESIGN" "ART_AND_DESIGN" ...
 $ Rating       : chr  "4.1" "3.9" "4.7" "4.5" ...
 $ Reviews      : chr  "159" "967" "87510" "215644" ...
 $ Size        : num  1.9e+07 1.4e+07 8.7e+06 2.5e+07 2.8e+06 5.6e+06 1.9e+07 2.9e+07 3.3e+07 3.1e+0
6 ...
$ Installs      : num  1e+04 5e+05 5e+06 5e+07 1e+05 5e+04 5e+04 1e+06 1e+06 1e+04 ...
$ Type          : chr  "Free" "Free" "Free" "Free" ...
$ Price         : chr  "0" "0" "0" "0" ...
$ Content.Rating: chr  "Everyone" "Everyone" "Everyone" "Teen" ...
$ Genres        : chr  "Art & Design" "Art & Design;Pretend Play" "Art & Design" "Art & Design" ...

```



■ data 정제

- 결측 값 제거
- 결측 값 제거 후 정제가 완료 되면 형 변환 시행
- data 형의 쉬운 변환으로 lubridata 라이브러리의 mdy 함수 이용

Console C:/RSources/ ↗

```
> # 결측 값 제거 (정제 과정에서 발생한)
> x = na.omit(x)
> # 정제 후 data 형을 고려하여 변환
> x$ > str(x)
'data.frame': 5356 obs. of 13 variables:
 $ App      : chr "Coloring book moana" "Paper flowers instructions" "Smoke Effect Photo Maker -
Smoke Editor" "Infinite Painter" ...
 $ Category : Factor w/ 33 levels "ART_AND_DESIGN",...: 1 1 1 1 1 1 1 1 1 1 ...
 $ Rating   : chr "3.9" "4.4" "3.8" "4.1" ...
 $ 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 6.0e+0
6 ...
 $ 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 NA 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" ...
```



■ data 정제

- 정제를 마친 data는 최종적으로 glimpse와 view를 사용하여 최종 확인

Console C:/RSources/ 

```
> glimpse(x)
```

```
Rows: 5,356
```

```
Columns: 13
```

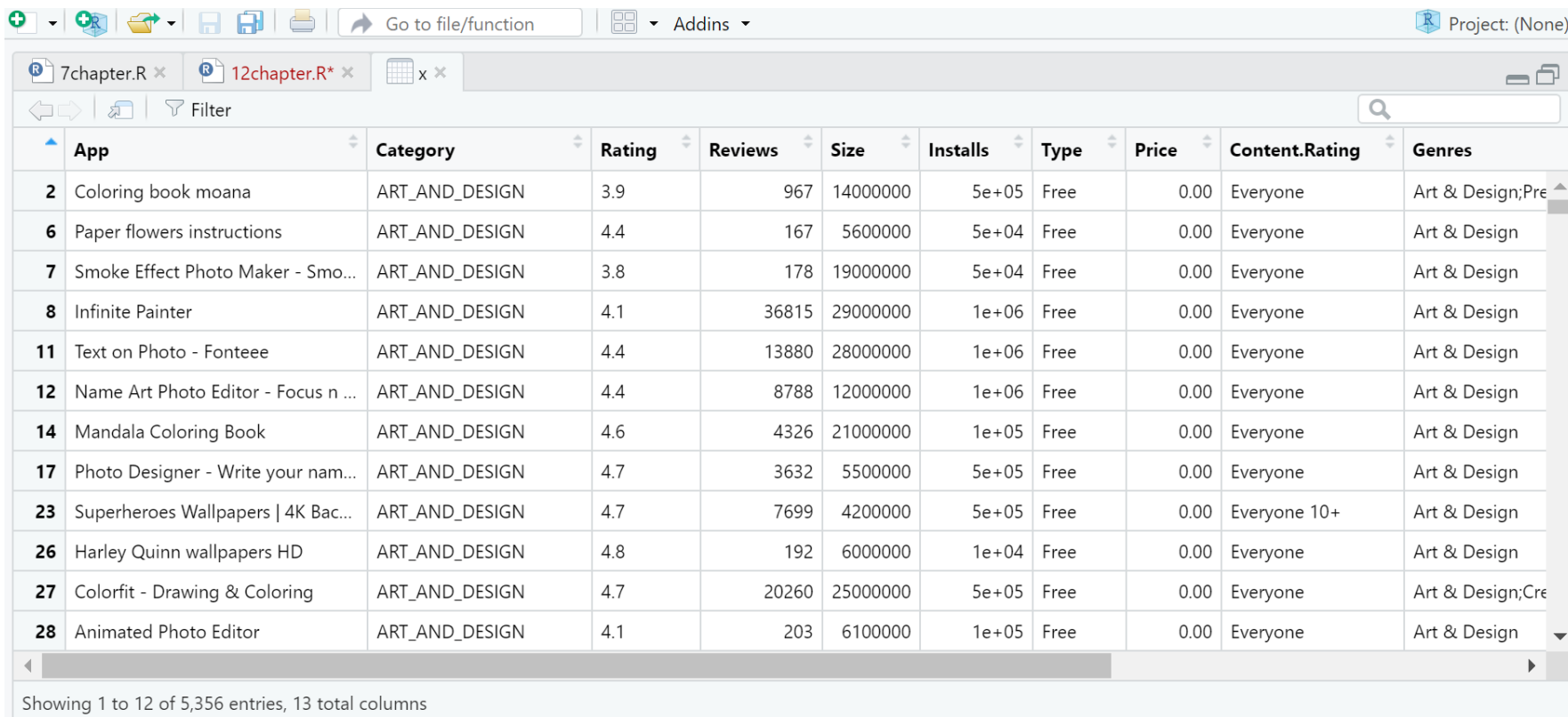
```
$ App      <chr> "Coloring book moana", "Paper flowers instructions", "Smoke Effect Pho...
$ Category <fct> ART_AND_DESIGN, ART_AND_DESIGN, ART_AND_DESIGN, ART_AND_DESIGN, ART_AN...
$ Rating   <chr> "3.9", "4.4", "3.8", "4.1", "4.4", "4.4", "4.6", "4.7", "4.7", "4.8", ...
$ Reviews  <dbl> 967, 167, 178, 36815, 13880, 8788, 4326, 3632, 7699, 192, 20260, 203, ...
$ Size     <dbl> 1.4e+07, 5.6e+06, 1.9e+07, 2.9e+07, 2.8e+07, 1.2e+07, 2.1e+07, 5.5e+06...
$ Installs <dbl> 5e+05, 5e+04, 5e+04, 1e+06, 1e+06, 1e+06, 1e+05, 5e+05, 5e+05, 1e+04, ...
$ Type     <fct> Free, Free, Free, Free, Free, Free, Free, Free, Free, Free, Free, Free...
$ Price    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
$ Content.Rating <fct> Everyone, Everyone, Everyone, Everyone, Everyone, Everyone, Everyone, ...
$ Genres    <fct> Art & Design;Pretend Play, Art & Design, Art & Design, Art & Design, A...
$ Last.Updated <date> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
$ Current.Ver <chr> "2.0.0", "1", "1.1", "6.1.61.1", "1.0.4", "1.0.15", "1.0.4", "3.1", "2...
$ Android.Ver <chr> "4.0.3 and up", "2.3 and up", "4.0.3 and up", "4.2 and up", "4.1 and u...
```



12.2 데이터 획득 및 정제

■ data 정제

- 정제를 마친 data는 최종적으로 glimpse와 view를 사용하여 최종 확인



	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content.Rating	Genres
2	Coloring book moana	ART_AND_DESIGN	3.9	967	14000000	5e+05	Free	0.00	Everyone	Art & Design;Pre
6	Paper flowers instructions	ART_AND_DESIGN	4.4	167	5600000	5e+04	Free	0.00	Everyone	Art & Design
7	Smoke Effect Photo Maker - Smo...	ART_AND_DESIGN	3.8	178	19000000	5e+04	Free	0.00	Everyone	Art & Design
8	Infinite Painter	ART_AND_DESIGN	4.1	36815	29000000	1e+06	Free	0.00	Everyone	Art & Design
11	Text on Photo - Fontee	ART_AND_DESIGN	4.4	13880	28000000	1e+06	Free	0.00	Everyone	Art & Design
12	Name Art Photo Editor - Focus n ...	ART_AND_DESIGN	4.4	8788	12000000	1e+06	Free	0.00	Everyone	Art & Design
14	Mandala Coloring Book	ART_AND_DESIGN	4.6	4326	21000000	1e+05	Free	0.00	Everyone	Art & Design
17	Photo Designer - Write your nam...	ART_AND_DESIGN	4.7	3632	5500000	5e+05	Free	0.00	Everyone	Art & Design
23	Superheroes Wallpapers 4K Bac...	ART_AND_DESIGN	4.7	7699	4200000	5e+05	Free	0.00	Everyone 10+	Art & Design
26	Harley Quinn wallpapers HD	ART_AND_DESIGN	4.8	192	6000000	1e+04	Free	0.00	Everyone	Art & Design
27	Colorfit - Drawing & Coloring	ART_AND_DESIGN	4.7	20260	25000000	5e+05	Free	0.00	Everyone	Art & Design;Cre
28	Animated Photo Editor	ART_AND_DESIGN	4.1	203	6100000	1e+05	Free	0.00	Everyone	Art & Design

Showing 1 to 12 of 5,356 entries, 13 total columns



12.2 데이터 획득 및 정제

요리의 기본은 식자재 손질!

고기 식감을 UP, 섬유질 자르기



샐러드 or 조림, 조리별 채소 손질



볶음 or 튀김, 조리 별 새우 손질



맛은 UP 시간은 DOWN, 효율성 높은 무 칼집



Thank you

