

## [R] 예제를 통한 데이터 전처리 작업

노트북: [TIL-MY]

만든 날짜: 2020-08-05 오전 8:15

URL: <https://continuous-development.tistory.com/49?category=793392>

# 나무늘보의 개발 블로그

홈

태그

방명록

R

## [R] 예제를 통한 데이터 전처리 작업

· by 꾸까꾸 · 2020. 8. 3. · 수정 · 삭제

예제를 통한 데이터 전처리

분류 전체보기 

Python

Database

ASP.NET

Algorithm

Deep learning

```

28 # 1. 데이터 전처리
29
30 # select와 filter를 통해 아래 컬럼만 뽑고
31 # 주소지가 서울특별시인 데이터만 추출하여 확인해보자
32 # 번호, 사업장명, 소재지전체주소, 업태구분명, 시설종류명, 인허가일자, 폐업일자,
33 # 소재지면적, 상세영업상태명, 영업상태구분코드
34
35 |
36 str(coffee)
37 seoul_coffee_select<-coffee %>%
38 select(번호, 사업장명, 소재지전체주소, 업태구분명, 시설종류명, 인허가일자, 폐업일자, 소재지면적, 상세영업상태명, 영업상태구분코드)%>%
39 filter(str_detect(소재지전체주소,"서울특별시")) # filter(str_detect(속성명,value)) 은 value의 패턴에 맞춰서 true값을 반환해준다. 고로
40 # true값에 해당하는 것들을 filter로 걸러서 가져온다.
41

```

처음에 str로 데이터 구조, 변수 개수, 변수 명, 관찰  
치 개수, 관찰치 보기

그다음 요구조건에 맞춰서 필요한 데이터만 추출한  
다. select으로 원하는 데이터를 가져온 뒤에 filter로  
조건에 맞는 데이터를 추출해 seoul\_coffee\_select에  
넣어준다.

```

55
56 # 커피숍 업태만 선택하기
57
58 seoul_coffee_select<-seoul_coffee_select %>%
59 filter(업태구분명=='커피숍')
60

```

업태구분명 중에 커피숍인 데이터만 넣기 위해 filter  
를 넣었다.

```

> head(seoul_coffee_select) # 앞에 6개의 데이터를 출력
# A tibble: 6 x 10
  번호 사업장명 소재지전체주소 업태구분명 시설종류명 인허가일자 폐업일자 소재지면적 상세영업상태명 영업상태구분코드
  <dbl> <chr> <chr> <chr> <chr> <dbl> <dbl> <dbl> <chr> <dbl>
1 1 카페미레 서울특별시 중로구 연지동 136-56번지 기타 유통음식점 46.75 19930313 NA NA 영업 01
2 2 피자 서울특별시 중로구 왕선동 372-4번지 다방 42 19830303 NA NA 영업 01
3 3 로반 서울특별시 중로구 선문로1가 56-10번지 지하음-다방 138.01 19880601 NA NA 영업 01
4 4 원대커피숍 서울특별시 중로구 룡계동 113-4번지 다방 22.88 19940726 NA NA 영업 01
5 5 김밥전국 서울특별시 중로구 왕선동 429-2번지 기타 유통음식점 52.25 19841215 NA NA 영업 01
6 6 전통다원 서울특별시 중로구 관문동 57-4번지 다방 52.0 19910822 NA NA 영업 01

```

그다음 데이터를 head를 통해 확인했다. head는 앞  
에 6개의 데이터를 보여준다.

```

67
68 View(seoul_coffee_select) # View 형태로 데이터를 본다.
69

```

AWS

ETC..

R

공지사항

글 보실 때 주의사  
항

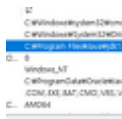
: 최근글 : 인  
기글

[R] R  
로 ...



2020.08.03

[R] R  
에...



2020.08.03

[R]  
예...

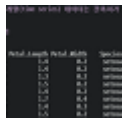


2020.08.03

[Algorithm] 파이선  
을 파..

2020.07.31

[R] R  
을 ...



2020.07.30

최근댓글

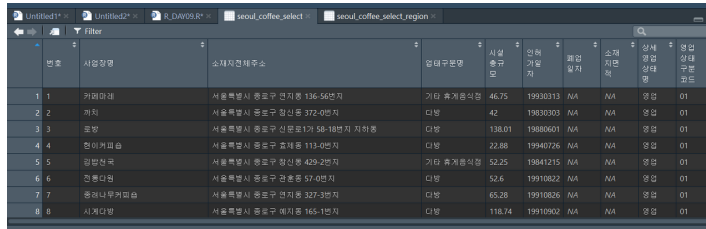
태그

cbind, SQL,

DDL, 날짜함수,

사용법,

View는 R에서 View 형태로 데이터를 볼 수 있게끔한다.



번호	사업장명	소재지전체주소	임대구분명	시설종류명	인허가일자	폐업일자	소재지명	상세영업상태명	영업구분코드
1	커피마래	서울특별시 중랑구 면지동 136-56번지	가다 휴게음식점	46.75	19910313	NA	NA	영업	01
2	커피	서울특별시 중랑구 장신동 372-0번지	다방	42	19830303	NA	NA	영업	01
3	로망	서울특별시 중랑구 신원로1가 58-18번지 지하층	다방	138.01	19880601	NA	NA	영업	01
4	현이커피숍	서울특별시 중랑구 고척동 113-0번지	다방	22.88	19940726	NA	NA	영업	01
5	김밥전국	서울특별시 중랑구 장신동 429-2번지	가다 휴게음식점	52.25	19841215	NA	NA	영업	01
6	전통다방	서울특별시 중랑구 관훈동 57-0번지	다방	52.6	19910822	NA	NA	영업	01
7	종려나무커피숍	서울특별시 중랑구 면지동 327-3번지	다방	65.28	19910826	NA	NA	영업	01
8	시계다방	서울특별시 중랑구 면지동 165-1번지	다방	118.74	19910902	NA	NA	영업	01

조건에서 폐업하지 않고 현재 영업 중인 카페였다. 그래서 기존의 데이터에서 filter를 걸어서 해당 조건에 맞는 영업이라고 적혀있는 데이터만 추출해 왔다.

```
69
70 # 폐업하지않고 현재 영업중인 카페찾기
71 seoul_coffee_select_live<-seoul_coffee_select %>%
72   filter(상세영업상태명 == "영업")
73
```

문제에서 원하는 것이 지역구별로 데이터를 나누는 것이었다. 서대문, 영등포, 동대문 이 3개를 찾기 위해 처음에 주소가 어떻게 나와있는지 확인했다.

시계열 그래프,  
R에서 DB 사용,  
R DB 사용,  
R 비정형 데이터 처리,  
R과 DB,  
전처리 예제,  
Oracle, rbind,  
테이블 생성,  
설정,  
파이썬을 파이썬  
답게,  
AWS,  
R 데이터 전처리  
예제,  
R 데이터베이스,  
데이터 전처리 예제,  
R로 하는 비정형  
데이터 처리,  
Oracle SQL,  
비정형 데이터 처리,  
파이썬, R DB,  
substr, 인스턴스,  
ggplot, Python,  
행렬

전체 방문자

129

Today : 1

Yesterday : 1

```

78
79 # 지역구별로 데이터 나누기(서대문, 영등포, 동대문) 3개의 구만
80 # 추출(시각화로 사용할 예정)
81 seoul_coffee_select$소재지전체주소
82
83 head(seoul_coffee_select$소재지전체주소)
84

```

```

> head(seoul_coffee_select$소재지전체주소)
[1] "서울특별시 중로구 연지동 136-56번지" "서울특별시 중로구 창신동 372-8번지" "서울특별시 중로구 신문로1가 58-18번지 지하동"
[4] "서울특별시 중로구 효제동 113-8번지" "서울특별시 중로구 창신동 429-2번지" "서울특별시 중로구 관훈동 57-8번지"

```

데이터가 위에 처럼 나와있었고 내가 필요한 건 구  
였다. 그래서

substr를 통해서 소재지 전체 주소를 잘랐다. "서울특  
별시 "까지 6글자이다. 이 여섯 번째부터 10번째 글  
자까지 자르고

str\_extract(매칭 문자열 추출)을 통해서 한글인 2-3글  
자에 구에 매칭 되는 데이터를 가져온다.

```

88 seoul_coffee_select$지역구 <- substr(seoul_coffee_select_live$소재지전체주소,6,10)
89 seoul_coffee_select$지역구 <- str_extract(seoul_coffee_select_live$소재지전체주소, '[가-힣]{2,3}구')
90
91
92 nrow(seoul_coffee_select %>%
93   filter(str_detect(지역구, c("서대문구","영등포구","동대문구"))))
94

```

그다음 nrow(행의 개수)를 통해 몇 개인지 확인한다.

이다음에는 ymd함수를 통해 char데이터를 dated의  
데이터 형식으로 바꿔준다.

```

117
118 # 인스타그램자와 폐업일자의 데이터 형식이
119 # chr와 logic으로 되어있는 것을 확인할 수 있다.
120 # ymd함수를 통해 chr와 logic형식으로 되어있는 데이터형식을 Date로 바꾸고, ymd 함수는 문자형 데이터를 date 타입으로 바꿔준다,
121 install.packages("anytime")
122 library(anytime)
123 install.packages("lubridate")
124 library(lubridate)
125
126 seoul_coffee_select$인스타그램지 <- ymd(seoul_coffee_select$인스타그램지)
127 seoul_coffee_select$폐업일자 <- ymd(seoul_coffee_select$폐업일자)
128

```

[illegible]

date 형식의 데이터를 year / month / day를 추출해낸다.

```
136 # Date로 바꾼 인허가 일자 데이터를 바탕으로 인허가일자
137 # year, month, day을 각각 추출해 가변수를 만들어보자
138
139 seoul_coffee_select$년도 <- year(seoul_coffee_select$인허가일자)
140 seoul_coffee_select$월 <- month(seoul_coffee_select$인허가일자)
141 seoul_coffee_select$일 <- day(seoul_coffee_select$인허가일자)
142
```

여기서 시설 총규모 타입을 `as.numeric`를 통해 수치  
형으로 바꿔준다.

```
149 # 데이터 형식 전처리(규모변수 추가)
150 # 시설종규모 타입 확인 후 수치형 -> 수치형
151 str(seoul_coffee_select$시설종규모)
152 seoul_coffee_select$시설종규모 <- as.numeric(seoul_coffee_select$시설종규모)
153
```

```

157 # 시설총규모에 따라 이를 구분지어
158 # 초소형, 소형, 중형, 대형, 초대형으로 구분지어볼려고 한다면
159 # 구분은 다음코드와 같이 임의로 지정
160 # 3제곱미터 이하는 초소형,
161 # 30제곱미터 이하는 소형,
162 # 70제곱미터이하는 중형
163 # 300제곱미터 이하는 대형 그 이상은 초대형
164
165 #mutate(data,newcol= value)
166
167 seoul_coffee_select<-seoul_coffee_select %>%
168   mutate(규모=ifelse(시설총규모<=3,"초소형",
169     ifelse(시설총규모>3 & 시설총규모<=30,"소형",
170       ifelse(시설총규모>30 & 시설총규모<=70,"중형",
171         ifelse(시설총규모>70 & 시설총규모 <=300,"대형",
172           ifelse(시설총규모>300,"초대형",""))))))
173
174

```

해당 조건에 맞춰서 mutate를 사용해서 열을 추가한다. 규모라는 열을 추가하는데 시설 총규모에 따라 값을 넣어준다. ifelse를 통해 if 조건일 때 값을 넣고 아니면

else로 넘어가는데 이 부분에 ifelse를 넣어줘서 해당 조건이 아니면 넘어가게끔 만든다.

만든 데이터를 통해 규모별 커피숍 수를 확인하기 위해 group\_by를 통해 규모를 묶고 이 총개수를 summarise(n=n())을 통해 센다

```

188
189 # 규모별 커피숍 수 확인하기
190 seoul_coffee_select %>%
191   group_by(규모) %>%
192   summarise(n=n())
193

```

문제의 조건을 맞추기 위해 영업 중 이면서 인허가  
일자가 2000-01-01 인 조건을 filter를 통해 걸고

그다음에 규모별로 확인하기 위해 group\_by를 통해  
규모를 묶고 그 개수를 셸다.

```
200 # 영업중이면서 인허가일자가 2000년 이후 인 커피숍 수를 규모별로 확인해 본다만
201 str(seoul_coffee_select)
202
203 seoul_coffee_select %>%
204   filter(상세영업상태명=="영업" & 인허가일자=="2000-01-01") %>%
205   group_by(규모) %>%
206   summarise(n=n())
207
```

가장 큰 규모의 카페를 찾기 위해서 which 함수를 사  
용했다. which를 통해 max나 min 을통해 제일 크거  
나 작은 값을 찾을 수 있다.

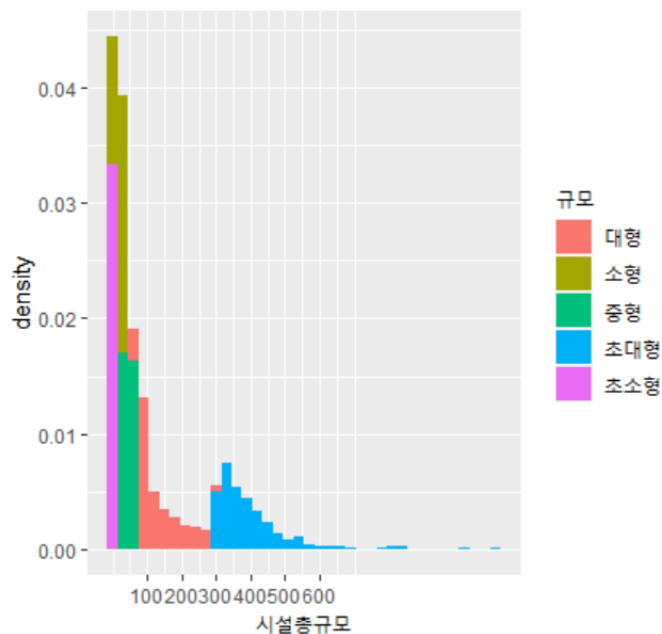
그렇게 값이면 해당 행을 반환하는데 그 행의 값을 c  
afe2000 [ ]에 넣어줌으로써 해당 행의 모든 값을 볼  
수 있게 한다.

```
223
224 # 가장 큰 규모의 카페는 어떨까요?
225 cafe2000
226 with.max(cafe2000$시설총규모)
227
228 cafe2000 <- seoul_coffee_select %>%
229   filter(상세영업상태명=="영업" & 인허가일자>="2000-01-01")
230
231 which
232
233 which.max(cafe2000$시설총규모) # which.max로 시설총규모중 최대값을 가진행을 가져온다.
234 which.min(cafe2000$시설총규모)
235
236 cafe2000[which.max(cafe2000$시설총규모) , ] # 그 행을 cafe2000에 행위치에 넣어서 해당행의 값을 가져온다.
237 [cafe2000[which.min(cafe2000$시설총규모) , ]
238
```

시설 총규모를 히스토그램으로 시각화하기 위해 ggplot을 사용했다. 시설 총규모를 히스토그램으로 나타내기 위해 x 축으로 넣고 desinty를 통해 밀도를 나타낸다.

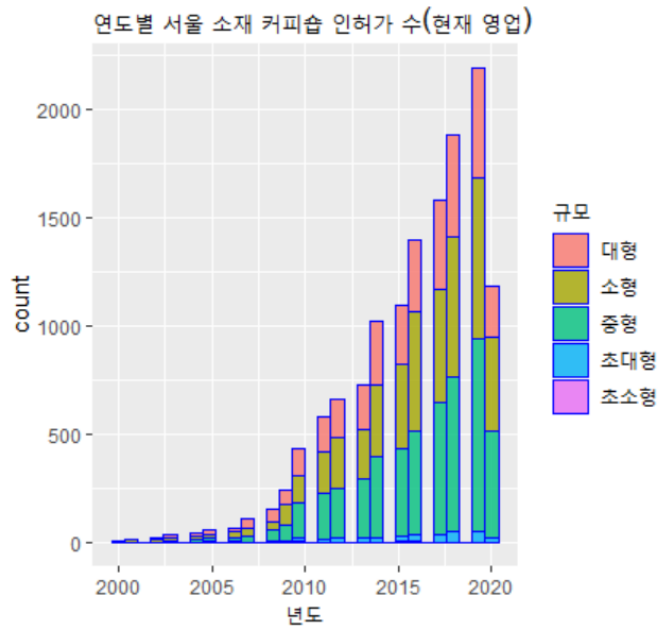
히스토그램을 나타내는 명령어는 geom\_histogram이다.

```
239 # 시설 총규모를 히스토그램으로 시각화한다면?
240 install.packages("ggplot2")
241 library(ggplot2)
242
243
244 cafe2000 %>% #desinty 는 밀도함수를 나타낸다. 밀도에
245 ggplot(aes(x=시설총규모 , y=..density.., fill=규모))+ #
246   geom_histogram(binwidth = 30 )+ #바의 넓이
247   scale_x_continuous(breaks = c(100,200,300,400,500,600))#간격을 나타낸다.
248   geom_density(fill=NA, col="red", alpha=.8)
249   geom_line(stat="density",size=1)
```



현재 영업 중인 카페의 인허가 연도를 히스토그램으로



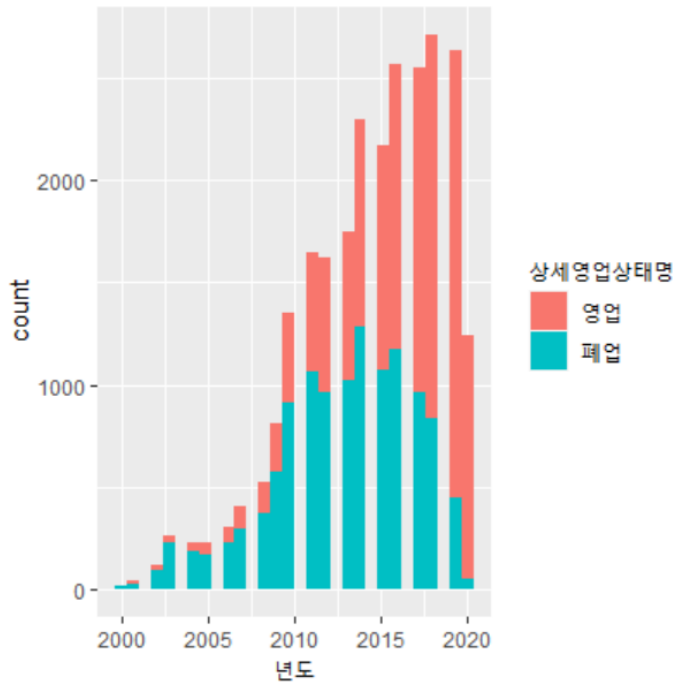


영업과 폐업한 카페의 인허가 연도를 보기 위해 x축에 연도를 fill로 통해 해당 값에서 영업과 폐업을 나타낸다.

```

263 # 영업과 폐업한 카페의 인허가 연도를 히스토그램으로 시각화
264
265 seoul_coffee_select %>%
266   filter(인허가일자 >= "2000-01-01") %>%
267   ggplot(aes(x=년도, fill=상세영업상태명))+
268   geom_histogram()

```



데이터 프레임으로 만들기 위해 `as.data.frame`을 사용하였다. 그리고 연도별 숫자를 확인하기 위해 `group_by`를 걸어줬다.

```

279
280 # 서울소재 커피숍의 인허가 연도별 숫자 확인
281 # 정보확인 후 데이터 프레임으로 만드세요~
282
283 df1<-as.data.frame(seoul_coffee_select %>%
284   filter(인허가일자>='2000-01-01') %>%
285   group_by(년도) %>%
286   summarise(n=n())
287 )
288

```

```

295 # 서울소재 커피숍의 인허가 년도별 숫자와 현재 영업중인 정보확인
296 # 정보확인 후 데이터 프레임으로 만드세요~
297 df2 <- as.data.frame(seoul_coffee %>%
298   filter(인허가일자 >= '2000-01-01' & 상세영업상태명=='영업')%>%
299   group_by(년도)%>%
300   summarise(n=n()))
301

```

```

313
314 # 생존율 시각화
315 # geom_line , geom_point

```

생존율을 시각화하기 위해서는 인허가 정보를 받은 커피숍 대비 현재도 영업 중인 커피숍이 필요합니다.

그래서 d2와 d1을 merge를 통해 하나의 데이터로 묶어줍니다.

이 두 값을 연도로 묶어준다, (원래는 df1, df2입니다.)

```

308
309 d3 <- merge(d1, d2, by="년도")
310

```

```

> d3
  년도  n.x  n.y
1 2000   17    5
2 2001   40   14
3 2002  119   25
4 2003  265   37
5 2004  228   41
6 2005  229   57
7 2006  200   70

```

여기서 나온 n.x값과 n.y값을 나눠서 인허가 일자를 받은 수에서 영업 중인 가게를 구한다.

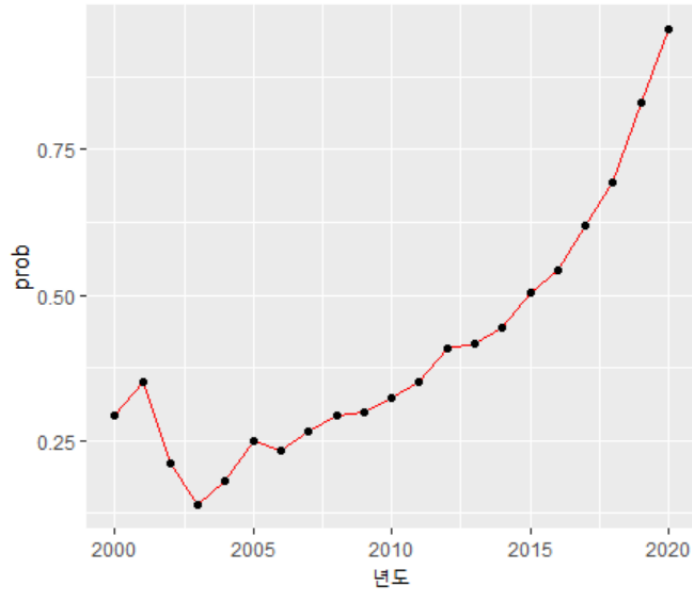
```
310  
311 d3 <- d3 %>%  
312   mutate(prob = (n.y)/(n.x))  
313
```

```
> d3  
  년도  n.x  n.y   prob  
1 2000   17    5 0.2941176  
2 2001   40   14 0.3500000  
3 2002  119   25 0.2100840  
4 2003  265   37 0.1396226  
5 2004  228   41 0.1798246  
6 2005  229   57 0.2489083  
7 2006  300   70 0.2333333  
8 2007  408  109 0.2671569  
9 2008  526  154 0.2927757  
10 2009  813  242 0.2976630
```

이 퍼센트를 그래프로 그린다.

```
320  
321 d3 %>%  
322   ggplot(aes(x=년도, y=prob))+  
323   geom_line(color="red")+  
324   geom_point()+  
325   ggtitle("서울소재 커피숍의 인허가 연도별 생존률")  
326
```

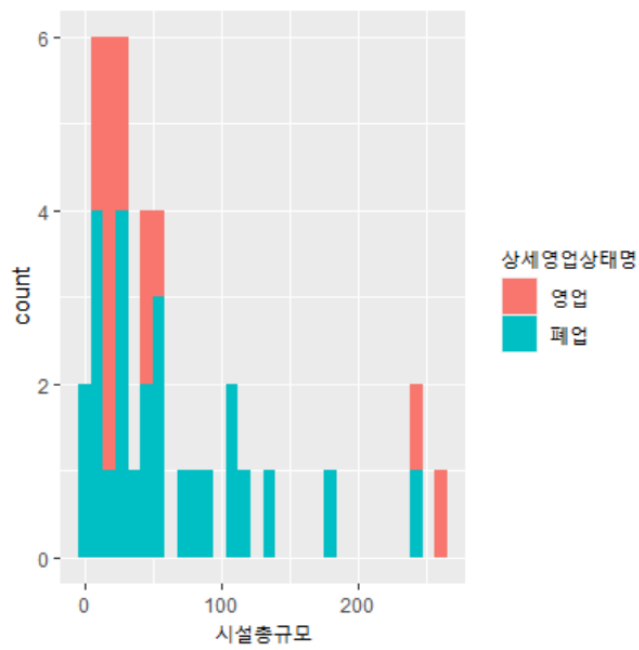
서울소재 커피숍의 인허가 연도별 생존률



```

330
331 # 2001년도 시설총규모에 따른 영업구분을 히스토그램으로 시각화
332 str(seoul_coffee_select)
333 seoul_coffee_select %>%
334   filter(년도==2001) %>%
335   ggplot(aes(x=시설총규모,fill=상세영업상태명))+
336   geom_histogram()
337

```



```

345 # 2000년도 ~
346 # 지역에 따른 년도별 커피숍 인허가 정보를 요약하고
347 # 데이터 프레임으로 만들어보자
348 str(seoul_coffee_select)
349 g1 <- as.data.frame(seoul_coffee_select %>%
350   filter(인허가일자 >= '2000-01-01') %>% # 인허가 정보라 했으니 인허
351   group_by(지역구, 년도) %>%
352   summarise(n=n()) #
353 )

```

```

362 # 2000년도 | ~
363 # 지역에 따른 년도별 커피숍 인허가 정보와
364 # 현재영업중인 정보를 요약하고
365 # 데이터 프레임으로 만들어보자
366
367 g2 <- as.data.frame(seoul_coffee_select %>%
368   filter(인허가일자>='2000-01-01' | 상세영업상태명=='영업') %>%
369   group_by(지역구, 년도)%>%
370   summarise(n=n())
371 )
372
373

```



### 'R' 카테고리의 다른 글

[R] R로 하는 비정형 데이터 처리 (facebook 데이터를 통한 긍정/부정 나누기) (0) 2020.08.03

[R] R에서 Database 사용하기 / DB 기본적인 구문 사용하기 (0) 2020.08.03

[R] 예제를 통한 데이터 전처리 작업 (0) 2020.08.03

[R] R을 통해 시계열 그래프 만들기 (자료 분석을 위한 시각화와 실습 예제) (0) 2020.07.30

[R] 같은 형태의 ggplot 과 barplot 만들기 (차이 비교) (0) 2020.07.30

[R] ggplot2 패키지 설치 에러시 해결 방법 (0) 2020.07.30

## 태그

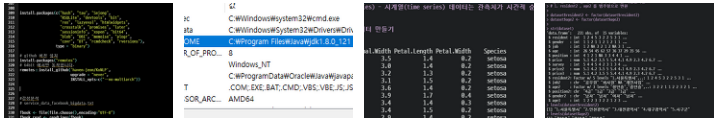
R 데이터 전처리 예제

데이터 전처리

데이터 전처리 예제

전처리 예제

## 관련글



[R] R로 하... [R] R에서 ... [R] R을 통... [R] 같은 형...

댓글 0

< 1 2 3 4 5 6 7 ...  
18 >

