[R] 예제를 통한 데이터 전처리 작업 — 나무늘보의 개발 블로그

노트북:	blog
만든 날짜:	2020-10-04 오후 7:20
URL:	https://continuous-development.tistory.com/49?category=793392

R

[R] 예제를 통한 데이터 전처리 작업

2020. 8. 3. 01:38 수정 삭제 공개

예제를 통한 데이터 전처리

```r
# 1. 데이터 전처리

# select와 filter를 통해 아래 컬럼만 뽑고
# 주소지가 서울특별시인 데이터만 추출하여 확인해보자
# 번호, 사업장명, 소재지전체주소, 업태구분명, 시설총규모, 인허가일자, 폐업일자,
# 소재지면적, 상세영업상태명, 영업상태구분코드

str(coffee)
seoul_coffee_select<-coffee %>%
  select(번호, 사업장명, 소재지전체주소, 업태구분명, 시설총규모, 인허가일자, 폐업일자, 소재지면적, 상세영업상태명, 영업상태구분코드)%>%
  filter( str_detect(소재지전체주소,"서울특별시")) # filter(str_detect(속성명,value)) 은 value의 패턴에 맞춰서 true값을 반환해준다. 고로
                                                  # true값에 해당하는 것들을 filter로 걸러서 가져온다.
```

처음에 str로 데이터 구조 , 변수 개수, 변수 명, 관찰지 개수 , 관찰치 보기

그다음 요구조건에 맞춰서 필요한 데이터만 추출한다. select으로 원하는 데이터를 가져온 뒤에 filter로 조건에 맞는 데이터를 추출해 seoul_coffee_select에 넣어준다.

```r

# 커피숍 업태만 선택하기

seoul_coffee_select<-seoul_coffee_select %>%
  filter(업태구분명=='커피숍')
```

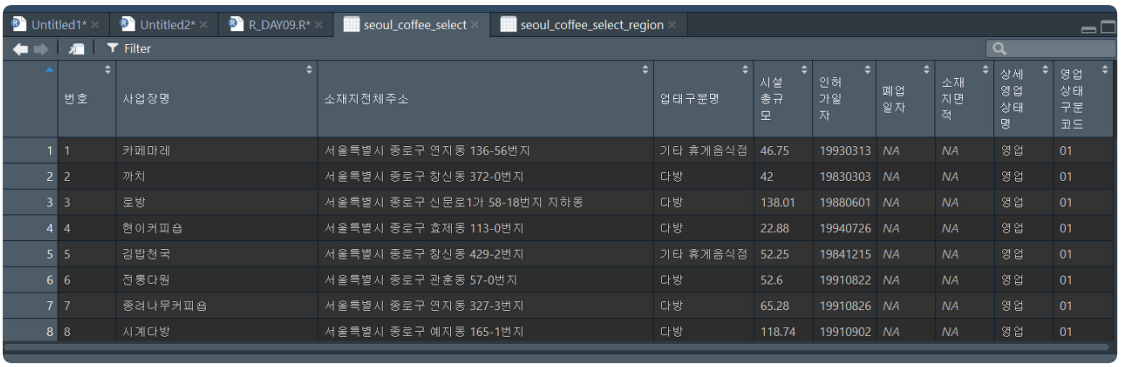
업태구분명 중에 커피숍인 데이터만 넣기 위해 filter를 넣었다.

```
> head(seoul_coffee_select) # 앞에 6개의 데이터를 출력
# A tibble: 6 x 10
  번호 사업장명 소재지전체주소                          업태구분명 시설총규모 인허가일자 폐업일자 소재지면적 상세영업상태명 영업상태구분코드
  <chr> <chr>  <chr>                                  <chr>      <chr>      <chr>    <lgl>    <chr>      <chr>          <chr>
1 1    카페마레 서울특별시 종로구 연지동 136-56번지      기타 휴게음식점~ 46.75      19930313 NA       NA         영업           01
2 2    까치   서울특별시 종로구 창신동 372-0번지        다방           42         19830303 NA       NA         영업           01
3 3    로방   서울특별시 종로구 신문로1가 58-18번지 지하동~ 다방          138.01     19880601 NA       NA         영업           01
4 4    현이커피숍 서울특별시 종로구 효제동 113-0번지       다방           22.88      19940726 NA       NA         영업           01
5 5    김밥천국 서울특별시 종로구 창신동 429-2번지        기타 휴게음식점~ 52.25      19841215 NA       NA         영업           01
6 6    전통다원 서울특별시 종로구 관훈동 57-0번지         다방           52.6       19910822 NA       NA         영업           01
>
```

그다음 데이터를 head를 통해 확인했다. head는 앞에 6개의 데이터를 보여준다.

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

View는 R에서 View 형태로 데이터를 볼 수 있게끔한다.



Untitled1* | Untitled2* | R_DAY09.R* | seoul_coffee_select | seoul_coffee_select_region

Filter

	번호	사업장명	소재지전체주소	업태구분명	시설총규모	인허가일자	폐업일자	소재지면적	상세영업상태명	영업상태구분코드
1	1	카페마레	서울특별시 종로구 연지동 136-56번지	기타 휴게음식점	46.75	19930313	NA	NA	영업	01
2	2	까치	서울특별시 종로구 창신동 372-0번지	다방	42	19830303	NA	NA	영업	01
3	3	로방	서울특별시 종로구 신문로1가 58-18번지 지하동	다방	138.01	19880601	NA	NA	영업	01
4	4	현이커피숍	서울특별시 종로구 효제동 113-0번지	다방	22.88	19940726	NA	NA	영업	01
5	5	김밥천국	서울특별시 종로구 창신동 429-2번지	기타 휴게음식점	52.25	19841215	NA	NA	영업	01
6	6	전통다원	서울특별시 종로구 관훈동 57-0번지	다방	52.6	19910822	NA	NA	영업	01
7	7	중려나무커피숍	서울특별시 종로구 연지동 327-3번지	다방	65.28	19910826	NA	NA	영업	01
8	8	시계다방	서울특별시 종로구 예지동 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 확인

했다.

```
# 지역구별로 데이터 나누기(서대문, 영등포, 동대문) 3개의 구만
# 추출(시각화로 사용할 예정)
seoul_coffee_select$소재지전체주소

head(seoul_coffee_select$소재지전체주소)
```

```
> head(seoul_coffee_select$소재지전체주소)
[1] "서울특별시 종로구 연지동 136-56번지"        "서울특별시 종로구 창신동 372-0번지"         "서울특별시 종로구 신문로1가 58-18번지 지하동"
[4] "서울특별시 종로구 효제동 113-0번지"         "서울특별시 종로구 창신동 429-2번지"         "서울특별시 종로구 관훈동 57-0번지"
> 
```

데이터가 위에 처럼 나와있었고 내가 필요한 건 구 였다. 그래서

substr를 통해서 소재지 전체 주소를 잘랐다. "서울특별시 "까지 6글자이다. 이 여섯 번째부터 10번째 글자까지 자르고 str_extract(매칭 문자열 추출)을 통해서 한글인 2-3글자에 구에 매칭 되는 데이터를 가져온다.

```
seoul_coffee_select$지역구 <- substr(seoul_coffee_select_live$소재지전체주소,6,10)
seoul_coffee_select$지역구 <-str_extract(seoul_coffee_select_live$소재지전체주소, '[가-힣]{2,3}구')


nrow(seoul_coffee_select %>%
       filter(str_detect(지역구, c("서대문구","영등포구","동대문구"))) )
```

그다음 nrow(행의 개수)를 통해 몇 개인지 확인한다.

이다음에는 ymd함수를 통해 char데이터를 dated의 데이터 형식으로 바꿔준다.

```
# 인허가일자와 폐업일자의 데이터 형식이
# chr와 logic으로 되어있는 것을 확인할 수 있다.
# ymd함수를 통해 chr와 logi형식으로 되어있는 데이터형식을 Date로 바꾼다.ymd 함수는 문자형 데이터를 date 타입으로 바꿔준다,
install.packages("anytime")
library(anytime)
install.packages("lubridate")
library(lubridate)

seoul_coffee_select$인허가일자 <- ymd(seoul_coffee_select$인허가일자)
seoul_coffee_select$폐업일자 <- ymd(seoul_coffee_select$폐업일자)
```

```
> str(seoul_coffee_select$인허가일자)
 chr [1:25561] "20070827" "20070830" "20070907" "20071002" "20170316" "20170320" "20030214" "20170323" "20170223" "20170223" "20170223" ...
> str(seoul_coffee_select$폐업일자)
 logi [1:25561] NA NA NA NA NA NA ...
> seoul_coffee_select$인허가일자 <- ymd(seoul_coffee_select$인허가일자)
> seoul_coffee_select$폐업일자 <- ymd(seoul_coffee_select$폐업일자)
> str(seoul_coffee_select$인허가일자)
 Date[1:25561], format: "2007-08-27" "2007-08-30" "2007-09-07" "2007-10-02" "2017-03-16" "2017-03-20" "2003-02-14" "2017-03-23" "2017-02-23" "2017-02-23"
 ...
> str(seoul_coffee_select$폐업일자)
 Date[1:25561], format: NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
NA NA NA ...
```

date 형식의 데이터를 yeaar / month / day를 추출해낸다.

```
# Date로 바꾼 인허가 일자 데이터를 바탕으로 인허가일자
# year, month, day을 각각 추출해 가변수를 만들어보자

seoul_coffee_select$년도 <- year(seoul_coffee_select$인허가일자)
seoul_coffee_select$월 <- month(seoul_coffee_select$인허가일자)
seoul_coffee_select$일 <- day(seoul_coffee_select$인허가일자)
```

여기서 시설 총규모 타입을 as.numeric를 통해 수치형으로 바꿔준다.

```
# 데이터 형식 전처리(규모변수 추가)
# 시설총규모 타입 확인 후 문자형 -> 수치형
str(seoul_coffee_select$시설총규모)
seoul_coffee_select$시설총규모 <-as.numeric(seoul_coffee_select$시설총규모)
```

```
# 시설총규모에 따라 이를 구분지어
# 초소형, 초형, 중형, 대형, 초대형으로 구분지어볼려고 한다면
# 구분은 다음코드와 같이 임의로 지정
# 3제곱미터 이하는 초소형,
# 30제곱미터 이하는 소형,
# 70제곱미터이하는 중형
# 300제곱미터 이하는 대형 그 이상은 초대형

#mutate(data,newcol= value)

seoul_coffee_select<-seoul_coffee_select %>%
  mutate(규모=ifelse(시설총규모<=3,"초소형",
                ifelse(시설총규모>3 & 시설총규모<=30,"소형",
                ifelse(시설총규모>30 & 시설총규모<=70,"중형",
                ifelse(시설총규모>70 & 시설총규모 <=300,"대형",
                ifelse(시설총규모>300,"초대형",""))))))
```

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 summrise(n=n())을 통해 센다

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

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 셌다.

```
200  # 영업중이면서 인허가일자가 2000년 이후 인 커피숍 수를 규모별로 확인해 본다면
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
```

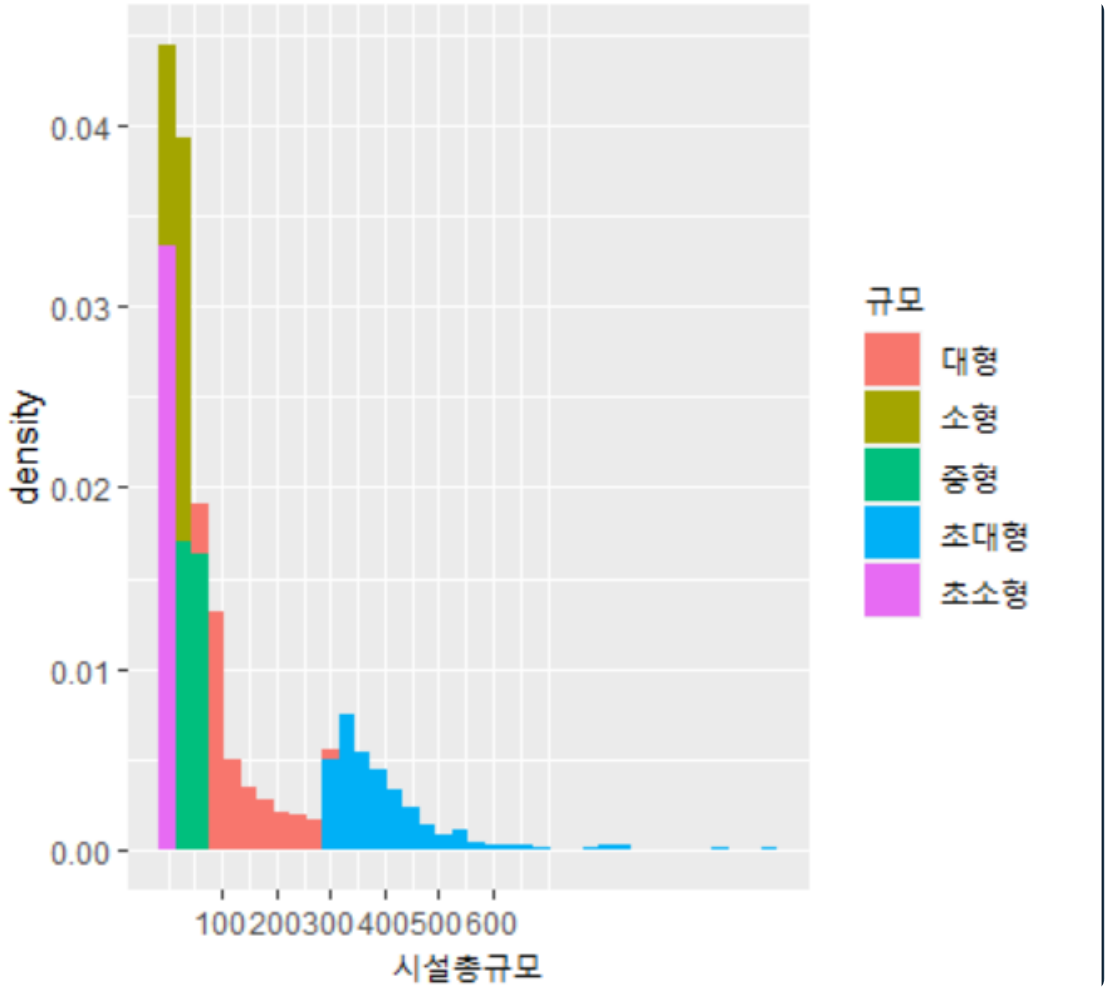
가장 큰 규모의 카페를 찾기 위해서 which 함수를 사용했다. which를 통해 max나 min 을통해 제일 크거나 작은 값을 찾을 수 있다.
그렇게 값이면 해당 행을 반환하는데 그 행의 값을 cafe2000 [ ]에 넣어줌으로써 해당 행의 모든 값을 볼 수 있게 한다.

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

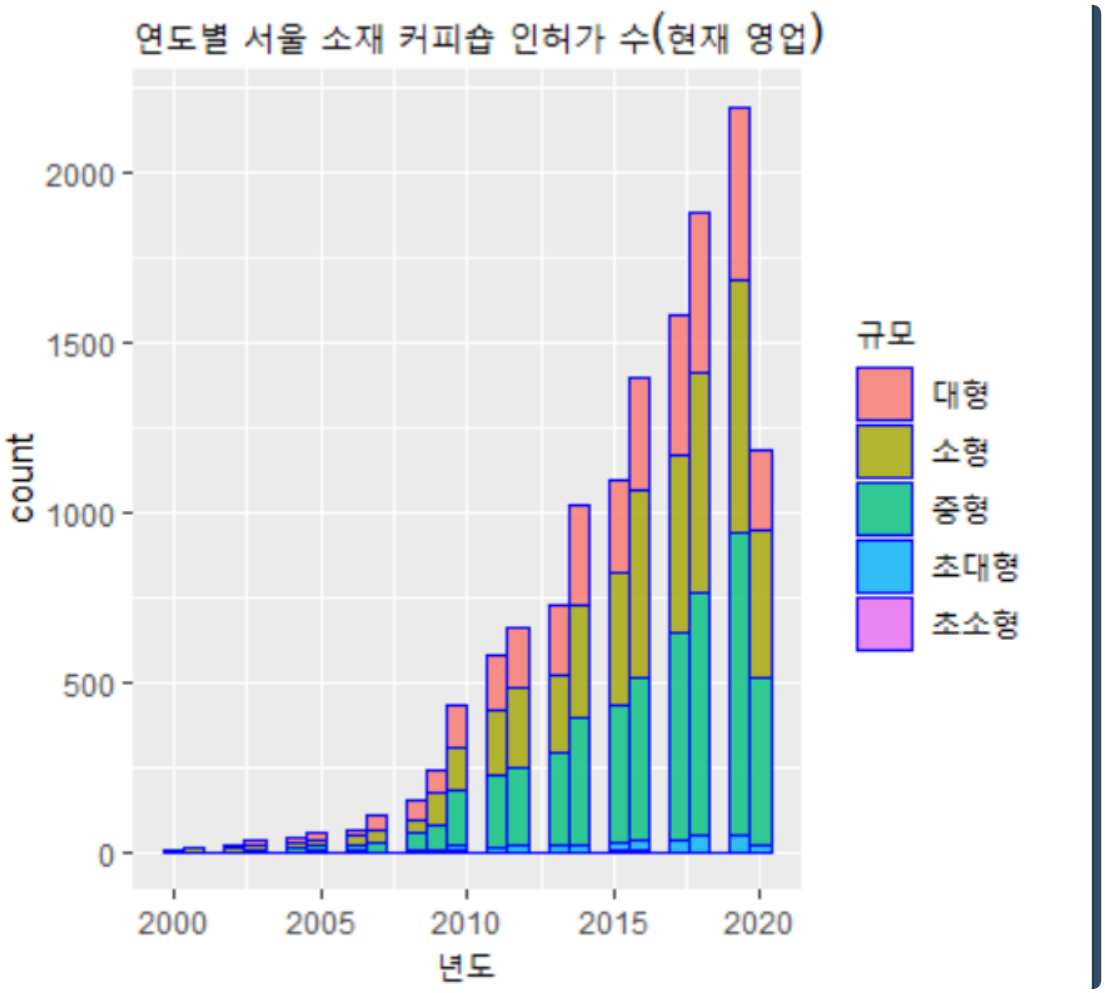
시설 총규모를 히스토그램으로 시각화하기 위해 ggplot을 사용했다. 시설 총규모를 히스토그램으로 나타내기 위해 x 축으로 넣고 desinty를 통해 밀도를 나타낸다.
히스토그램을 나타내는 명령어는 geom_histogram이다.

```
# 시설 총규모를 히스토그램으로 시각화한다면?
install.packages("ggplot2")
library(ggplot2)


cafe2000 %>% #desinty 는 밀도함수를 나타낸다. 밀도에
  ggplot(aes(x=시설총규모 , y=..density.., fill=규모))+ #
  geom_histogram(binwidth = 30 )+                      #바의 넓이
  scale_x_continuous(breaks = c(100,200,300,400,500,600))#간격을 나타낸다.
  geom_density(fill=NA, col="red", alpha=.8)
  geom_line(stat="density",size=1)
```

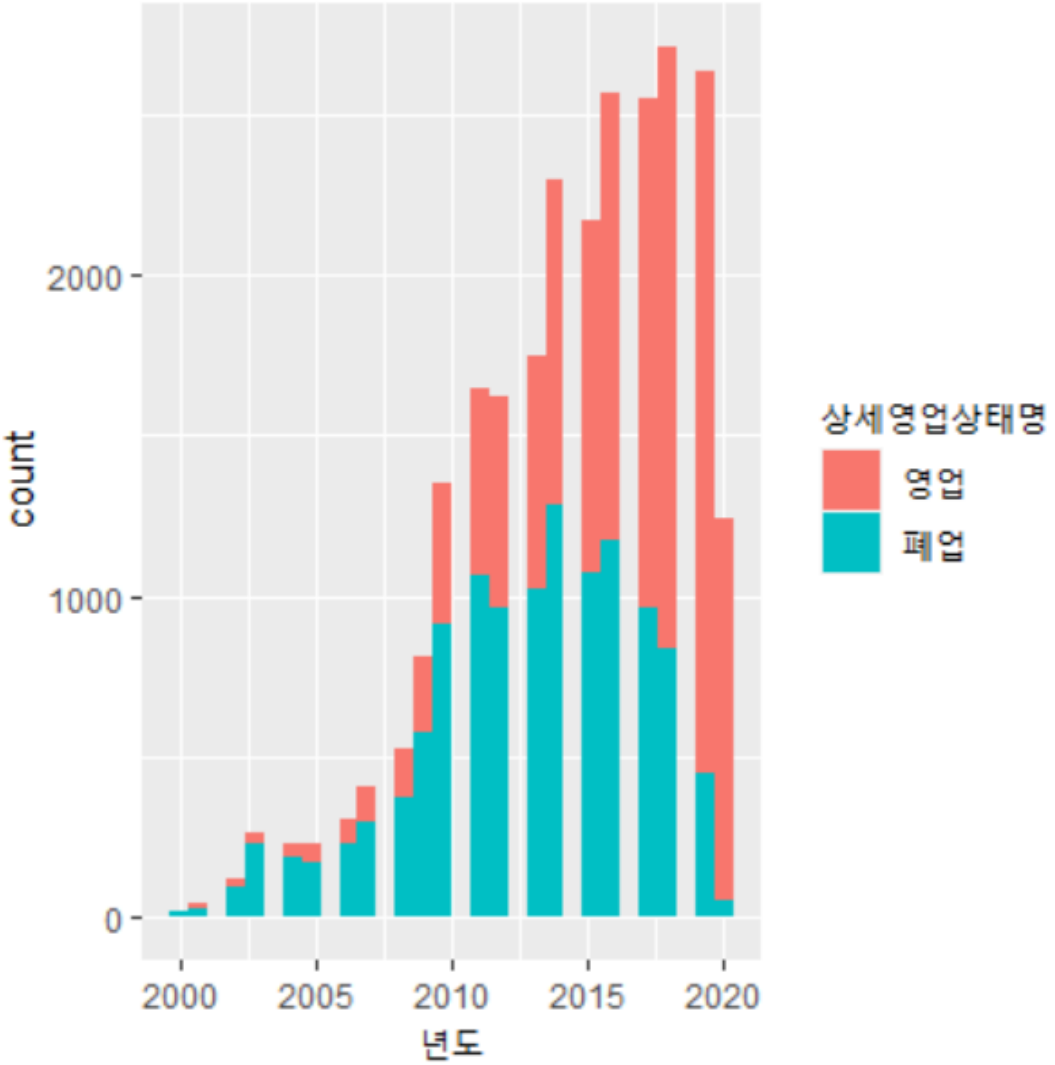


현재 영업 중인 카페의 인허가 연도를 히스토그램으로



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

```
263  # 영업과 폐업한 카페의 인허가 연도를 히스토 그램으로 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.frame을 사용하였다. 그리고 연도별 숫자를 확인하기 위해 group_by를 걸어줬다.

```
279
280  # 서울소재 커피숍의 인허가 년도별 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
```

```r
# 서울소재 커피숍의 인허가 년도별 숫자와 현재 영업중인 정보확인
# 정보확인 후 데이터 프레임으로 만드세요~
df2 <- as.data.frame(seoul_coffee_select %>%
                       filter(인허가일자 >= '2000-01-01' & 상세영업상태명=='영업')%>%
                       group_by(년도)%>%
                       summarise(n=n()))
```

```r
# 생존율 시각화
# 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.)

```r
d3 <- merge(d1, d2, by="년도")
```

```
> 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
```

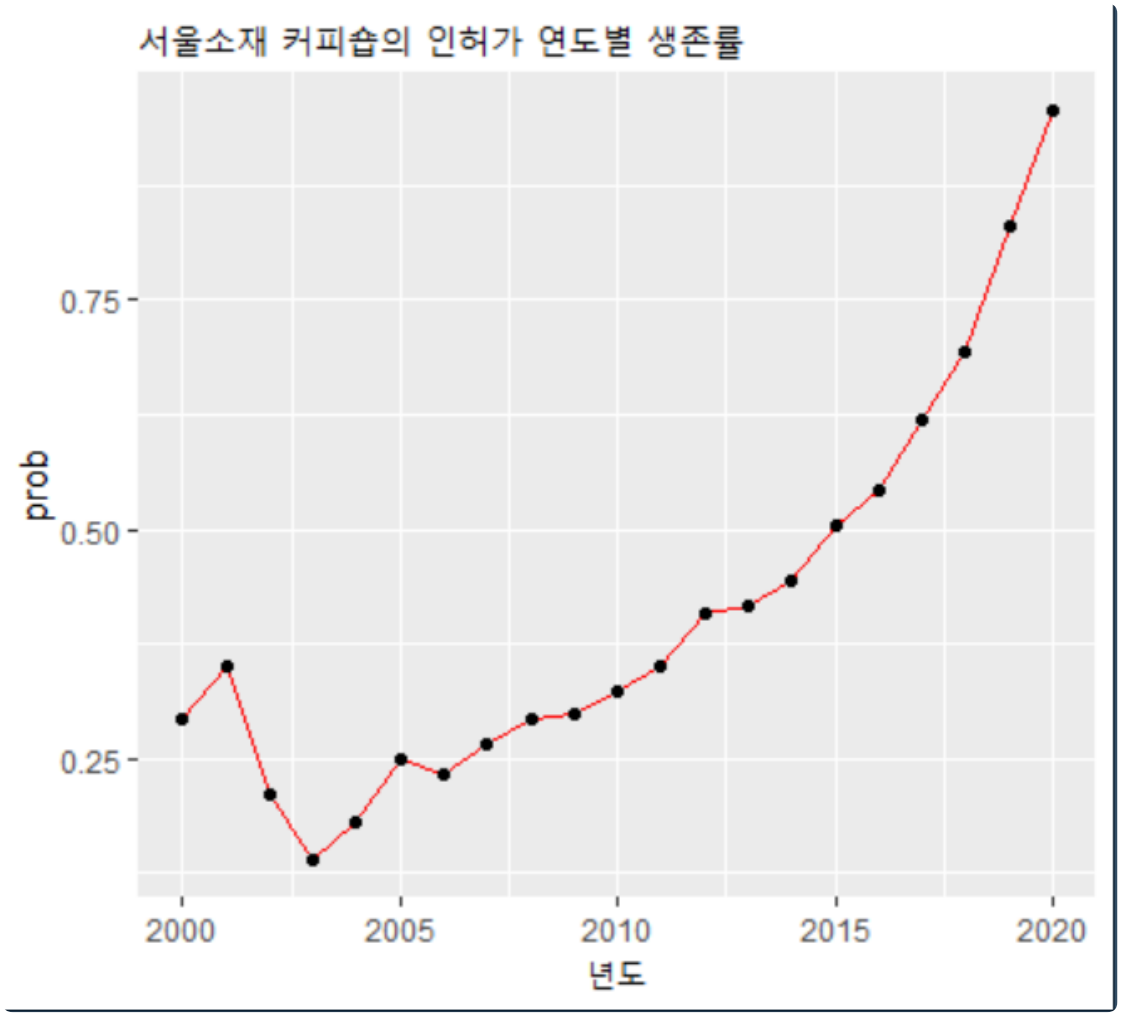
여기서 나온 n.x값과 n.y값을 나눠서 인허가 일자를 받은 수에서 영업 중인 가게를 구한다.

```r
d3 <- d3 %>%
  mutate(prob = (n.y)/(n.x))
```

```
> 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
```

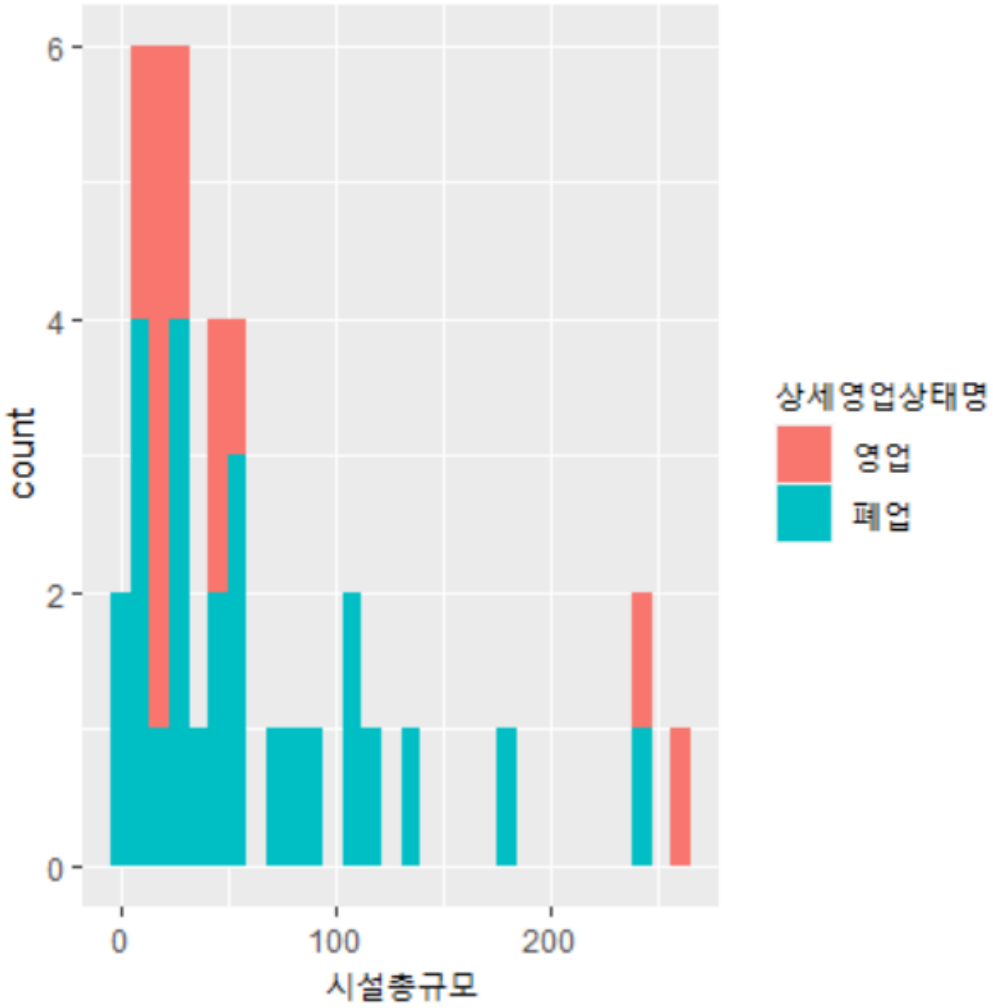
이 퍼센트를 그래프로 그린다.

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




```r

# 2001년도 시설총규모에 따른 영업구분을 히스토그램으로 시각화
str(seoul_coffee_select)
seoul_coffee_select %>%
  filter(년도==2001) %>%
  ggplot(aes(x=시설총규모,fill=상세영업상태명))+
  geom_histogram()
```




```r
# 2000년도 ~
# 지역구에 따른 년도별 커피숍 인허가 정보를 요약하고
# 데이터 프레임으로 만들어보자
str(seoul_coffee_select)
g1 <- as.data.frame(seoul_coffee_select %>%
                      filter(인허가일자 >= '2000-01-01') %>% # 인허가 정보라 했으니 인허
                      group_by(지역구,년도) %>%
                      summarise(n=n()) #
                      )
```

```r
# 2000년도 ~
# 지역구에 따른 년도별 커피숍 인허가 정보와
# 현재영업중인 정보를 요약하고
# 데이터 프레임으로 만들어보자

g2 <- as.data.frame(seoul_coffee_select %>%
                      filter(인허가일자>='2000-01-01' | 상세영업상태명=='영업') %>%
                      group_by(지역구,년도)%>%
                      summarise(n=n())
                    )

```


'R' 카테고리의 다른 글

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

[R] R에서 Database 사용하기 / DB 기본적인 구문 사용하기

[R] 예제를 통한 데이터 전처리 작업

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

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

[R] ggplot2 패키지 설치 에러시 해결 방법

R 데이터 전처리 예제 | 데이터 전처리 | 데이터 전처리 예제 | 전처리 예제

꾸까꾸

혼자 끄적끄적하는 블로그 입니다.