

[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

```
28 # 1. 데이터 전처리
29
30 # select와 filter를 통해 아래 컬럼만 뽑고
31 # 주소지가 서울특별시인 데이터만 추출하여 확인해보자
32 # 번호, 사업장명, 소재지전체주소, 업태구분명, 시설총규모, 인허가일자, 폐업일자,
33 # 소재지면적, 상세영업상태명, 영업상태구분코드
34
35 |
36 str(coffee)
37 seoul_coffee_select<-coffee %>%
38   select(번호, 사업장명, 소재지전체주소, 업태구분명, 시설총규모, 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, 관찰지 개수, 관찰치 보기

그다음 요구조건에 맞춰서 필요한 데이터만 추출한다. 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 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> <chr> <chr> <chr> <chr>
1 1 카페마래 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.

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

```
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 372-0번지"      "서울특별시 중로구 신문로1가 58-18번지 지하동"
[4] "서울특별시 중로구 효제동 113-0번지"        "서울특별시 중로구 창신동 429-2번지"        "서울특별시 중로구 관훈동 57-0번지"
>
```

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.

```
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
```

이다음에는 ymd함수를 통해 char데이터를 dated의 데이터 형식으로 바꿔준다.

```

117 # 인허가일자과 폐업일자의 데이터 형식이
118 # chr와 logic으로 되어있는 것을 확인할 수 있다.
119 # ymd함수를 통해 chr와 logic형식으로 되어있는 데이터형식을 Date로 바꾼다.ymd 함수는 문자형 데이터를 date 타입으로 바꿔준다,
120 # install.packages("anytime")
121 library(anytime)
122 install.packages("lubridate")
123 library(lubridate)
124
125 seoul_coffee_select$인허가일자 <- ymd(seoul_coffee_select$인허가일자)
126 seoul_coffee_select$폐업일자 <- ymd(seoul_coffee_select$폐업일자)
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000

```

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.numeric를 통해 수치형으로 바꿔준다.

```
148
149 # 데이터 형식 전처리(규모변수 추가)
150 # 시설총규모 타입 확인 후 문자형 -> 수치형
151 str(seoul_coffee_select$시설총규모)
152 seoul_coffee_select$시설총규모 <-as.numeric(seoul_coffee_select$시설총규모)
153
```

```
154
155
157 # 시설총규모에 따라 이를 구분지어
158 # 초소형, 초형, 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

```
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 2000-01-01인 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 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
```

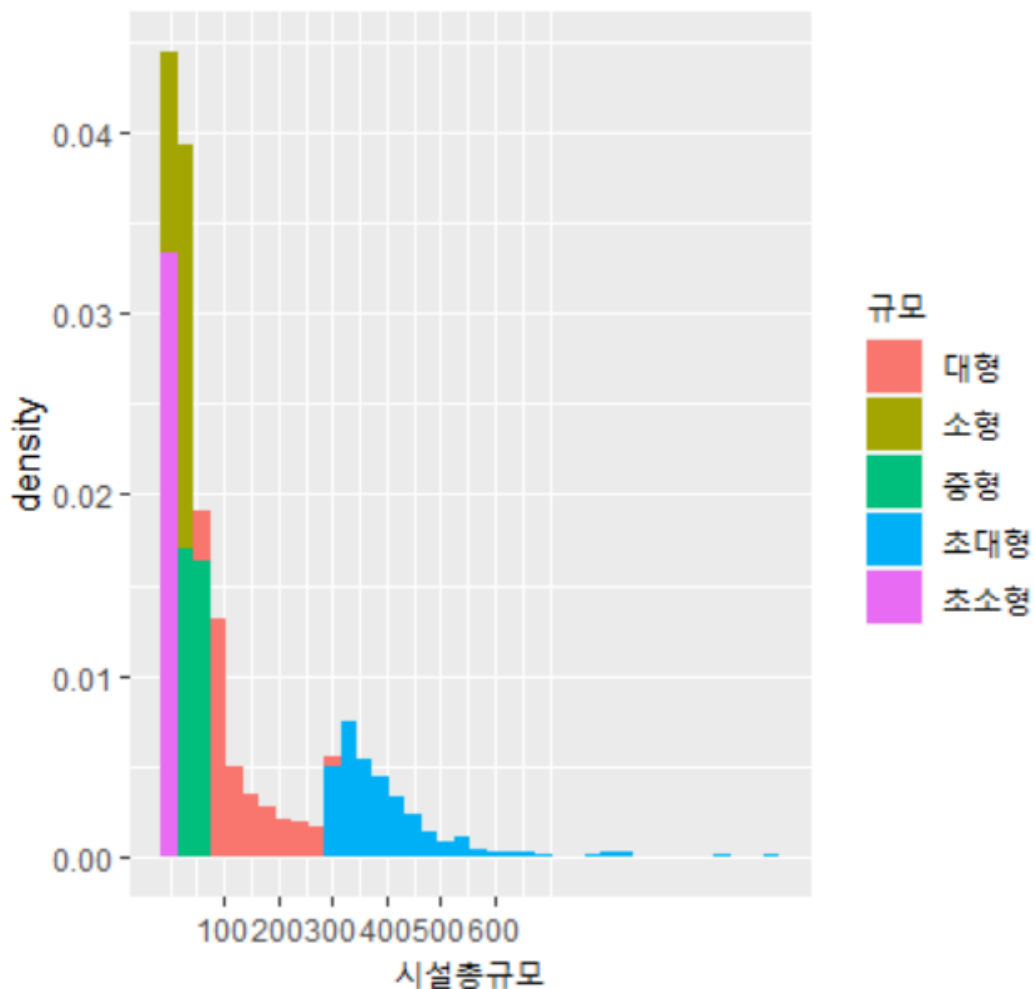
가장 큰 규모의 카페를 찾기 위해서 which 함수를 사용했다. which를 통해 max나 min 을통해 제일 크거나 작은 값을 찾을 수 있다.
그렇게 값이면 해당 행을 반환하는데 그 행의 값을 cafe2000 []에 넣어줌으로써 해당 행의 모든 값을 볼 수 있게 한다.

```
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) , ]
238
```

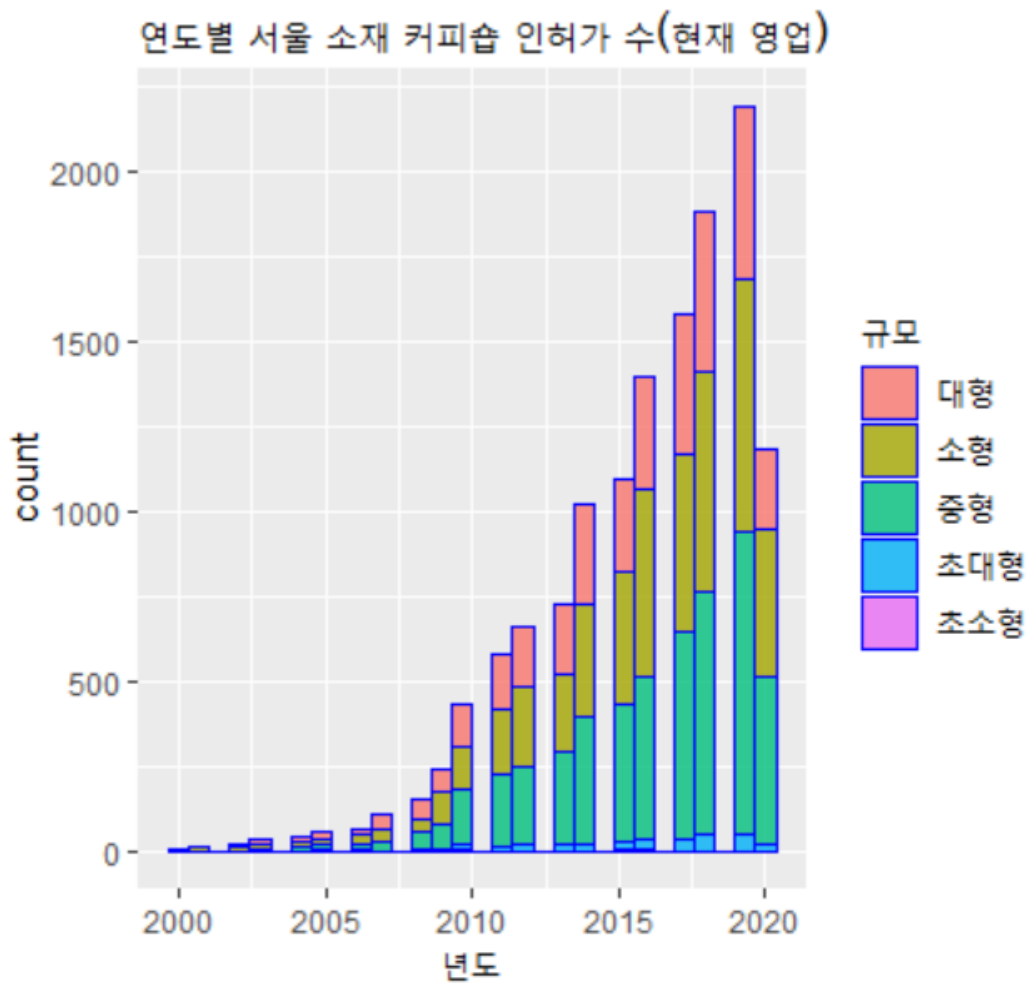
시설 총규모를 히스토그램으로 시각화하기 위해 ggplot을 사용했다. 시설
총규모를 히스토그램으로 나타내기 위해 x 축으로 넣고 density를 통해
밀도를 나타낸다.

히스토그램을 나타내는 명령어는 geom_histogram이다.

```
239 # 시설 총규모를 히스토그램으로 시각화한다면?  
240 install.packages("ggplot2")  
241 library(ggplot2)  
242  
243  
244 cafe2000 %>% #density 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

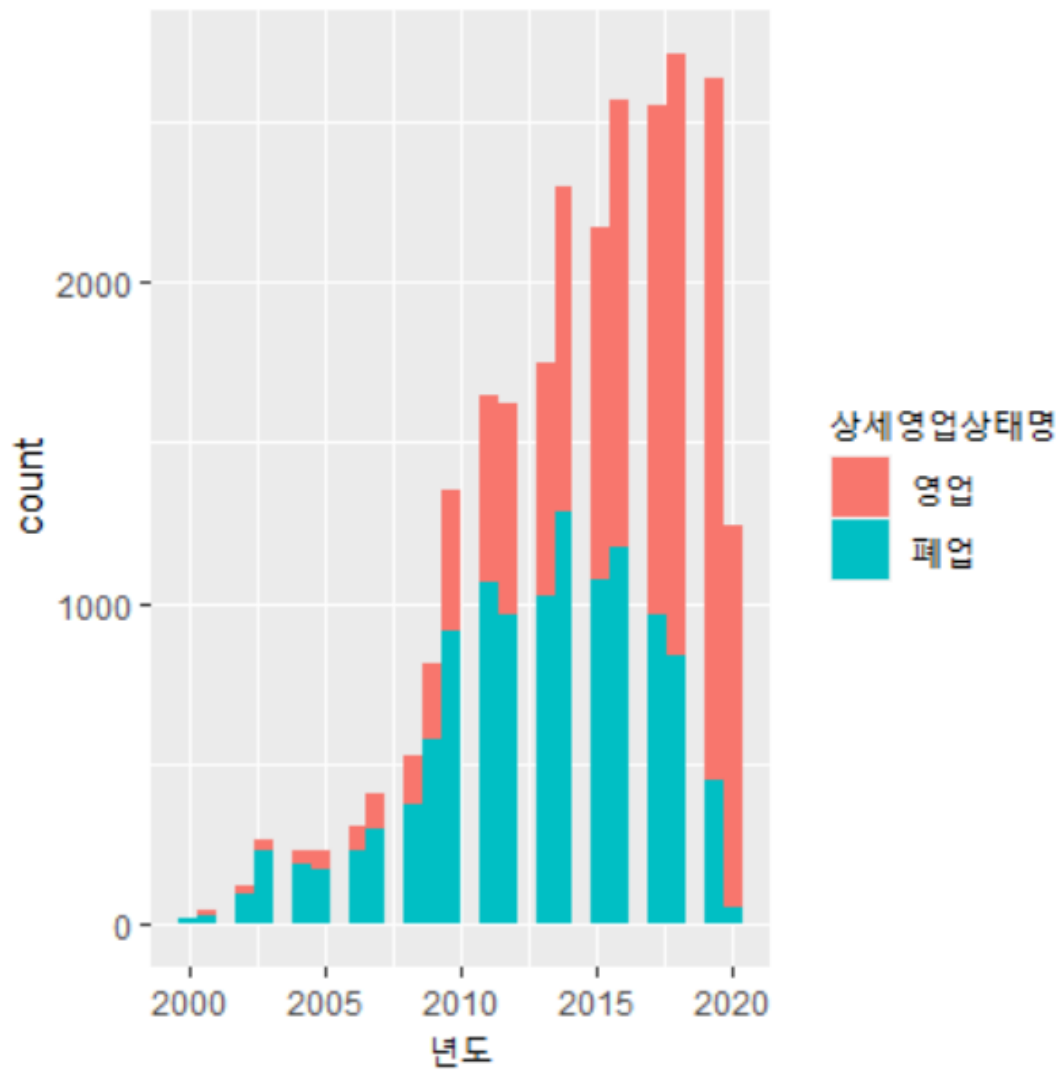


영업과 폐업한 카페의 인허가 연도를 보기 위해 x축에 연도를 fill로 통해 해당 값에서 영업과 폐업을 나타낸다.

```

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_select %>%
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.)

```

309 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.

```

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

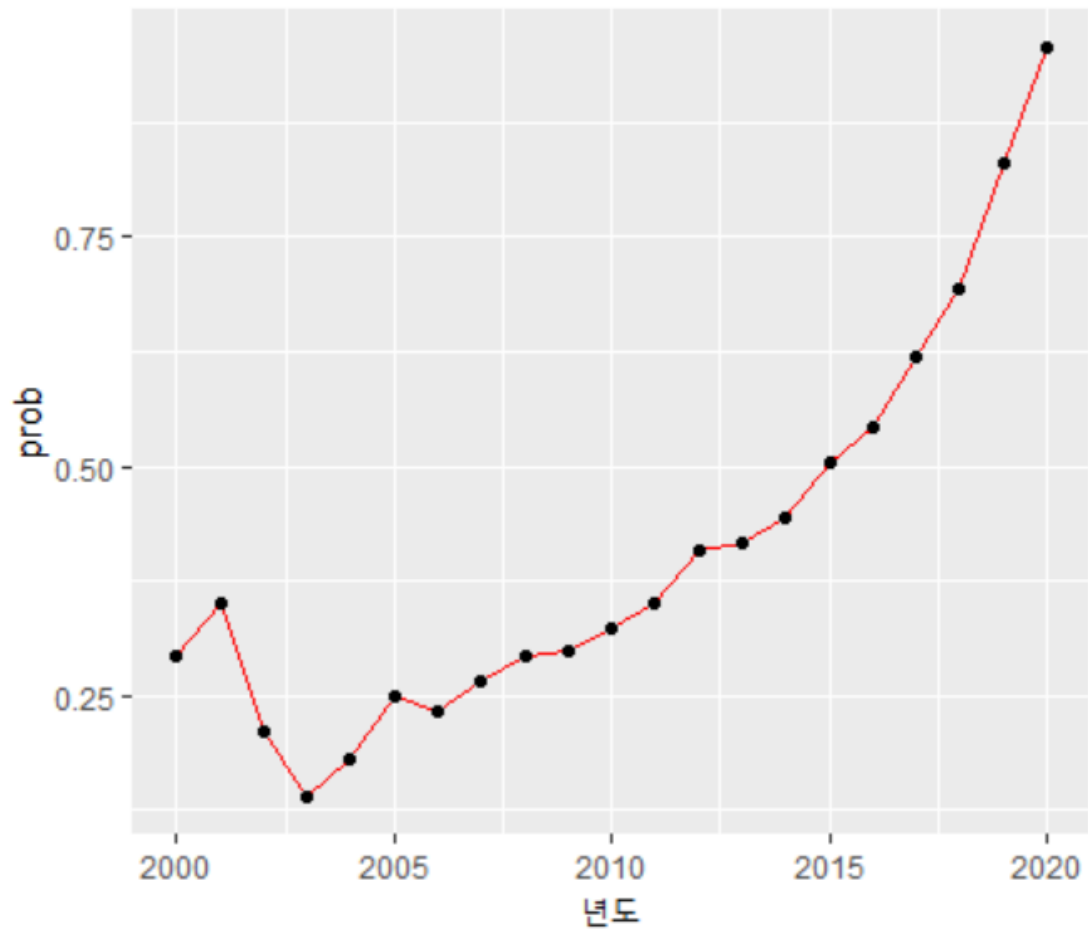
```

```
> 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
```

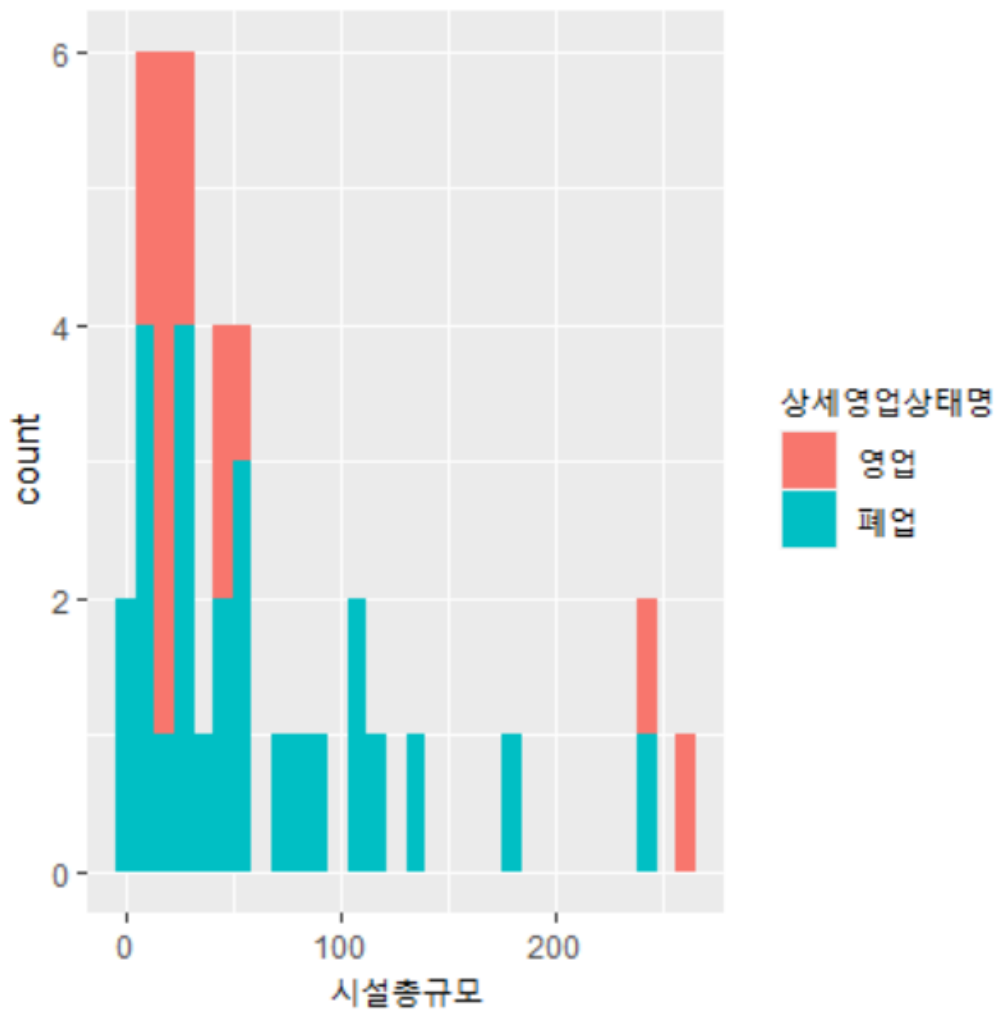
서울소재 커피숍의 인허가 연도별 생존률



```

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

```

'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)□

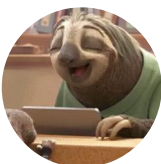
[R] ggplot2 패키지 설치 에러시 해결 방법□

R 데이터 전처리 예제

데이터 전처리

데이터 전처리 예제

전처리 예제



꾸까꾸

혼자 끄적끄적하는 블로그 입니다.