


통계패키지활용 자료분석(2024년 겨울학기)			
담 당 교 수 : 김 태 수			
강좌 번호	100961-31001	본인의 과제 자체 평가	9/10 점
과제명 : 제 1차 Data 가공 with R			

이름	이지원
	

제 출 일	2024년 01월 14일
학 과	산업공학과 ITM전공
학 번	22102009

목차

I. mpg dataset을 바탕으로 한 데이터 가공

- 사용한 자료 설명
- 제시된 문제의 답안
 1. 조건에 맞는 데이터만 추출하기
 2. 필요한 변수만 추출하기
 3. 순서대로 정렬하기
 4. 파생 변수 추가하기
 5. 집단별로 요약하기
 6. 데이터 합치기

II. 분석 도전 과제: midwest dataset을 바탕으로한 데이터 가공

- 사용한 자료 설명
- 제시된 문제의 답안
- 결론 도출

III. 부록(분석 도전 과제의 R script)

I. mpg dataset을 바탕으로 한 데이터 가공

사용한 자료 설명

R ggplot2 패키지에 기본적으로 포함된 dataset인 mpg를 사용하였다. 1999년과 2008년에 출시된 여러 자동차 모델의 연비에 대한 정보들을 담고 있다. audi, chevrolet, ford, hyundai 등 총 15개의 제조사 자동차 모델들이 있고 실린더 수, 변속기 종류, 구동방식 등에 대한 정보를 가지고 있다.

제시된 문제의 답안

1. 조건에 맞는 데이터만 추출하기

Q1. 자동차 배기량에 따른 고속도로 연비?

- displ(배기량)이 4이하인 자동차와 5이상인 자동차 중 어떤 자동차의 hwy(고속도로 연비)가 평균적으로 더 높은가?

A1. displ(배기량)이 4이하인 자동차가 displ(배기량)이 5이상인 자동차보다 hwy(고속도로 연비)가 평균적으로 더 높다.

- 배기량이 4이하인 자동차의 고속도로 연비의 평균은 25.96319이고 배기량이 5이상인 자동차의 고속도로 연비의 평균은 18.07895이다.

Q2. 자동차 제조회사에 따른 도시연비?

- “audi”와 “toyota” 중 어느 manufacturer(자동차 제조 회사)의 cty(도시 연비)가 평균적으로 더 높은가?

A2. “toyota”의 cty(도시 연비)가 “au야”의 cty(도시 연비)보다 평균적으로 높다.

- “audi”의 도시 연비 평균은 17.61111이고 “toyota”의 도시 연비 평균은 18.52941이다.

Q3. “chevrolet”, “ford”, “honda” 자동차의 고속도로 연비 평균은?

- 이 회사들의 자동차를 추출한 뒤 hwy의 전체 평균?

A3. “chevrolet”, “ford”, “honda” 자동차의 hwy(고속도로 연비) 전체 평균은 22.50943이다.

2. 필요한 변수만 추출하기

Q1. class(자동차 종류), cty(도시 연비) 변수를 추출해 새로운 데이터를 만들고, 새로 만든 데이터의 일부를 출력해서 두 변수로만 구성되어 있는지 확인하시오.

A1. 다음과 같이 class와 cty 두 변수로만 구성되어있는 것을 확인할 수 있다.

<code>> head(Df)</code>	<code>> tail(Df)</code>
<pre> class cty 1 compact 18 2 compact 21 3 compact 20 4 compact 21 5 compact 16 6 compact 18 </pre>	<pre> class cty 229 midsize 18 230 midsize 19 231 midsize 21 232 midsize 16 233 midsize 18 234 midsize 17 </pre>

Q2. class(자동차 종류)가 "suv"인 자동차와 "compact"인 자동차 중 어떤 자동차의 cty(도시 연비)가 더 높은가?

A2. class(자동차 종류)가 "compact"인 자동차가 "suv"인 자동차보다 cty(도시 연비)가 높다.

- 자동차 종류가 "suv"인 자동차의 도시 연비 평균은 13.5이고 "compact"인 자동차의 도시 연비 평균은 20.12766이다.

4. 순서대로 정렬하기

cyl(실린더 수)를 기준으로 mpg 데이터 셋을 정렬함

```

> Mpg %>% arrange(desc(cyl)) %>% head(10)
  manufacturer      model displ year  cyl  trans drv  cty   hwy fl   class
1         audi      a6 quattro  4.2  2008    8   auto(s6)  4   16   23  p midsize
2   chevrolet c1500 suburban 2wd  5.3  2008    8   auto(l4)  r   14   20  r    suv
3   chevrolet c1500 suburban 2wd  5.3  2008    8   auto(l4)  r   11   15  e    suv
4   chevrolet c1500 suburban 2wd  5.3  2008    8   auto(l4)  r   14   20  r    suv
5   chevrolet c1500 suburban 2wd  5.7  1999    8   auto(l4)  r   13   17  r    suv
6   chevrolet c1500 suburban 2wd  6.0  2008    8   auto(l4)  r   12   17  r    suv
7   chevrolet      corvette  5.7  1999    8 manual(m6)  r   16   26  p 2seater
8   chevrolet      corvette  5.7  1999    8   auto(l4)  r   15   23  p 2seater
9   chevrolet      corvette  6.2  2008    8 manual(m6)  r   16   26  p 2seater
10  chevrolet      corvette  6.2  2008    8   auto(s6)  r   15   25  p 2seater

```

5. 파생 변수 추가하기

Q1. Mpg 데이터 복사본을 만들고, cty와 hwy를 더한 '합산 연비 변수'를 추가

A1.

```

> Mpg_New <- Mpg
> Mpg_New <- Mpg_New %>% mutate(Total= cty + hwy)
> Mpg_New
  manufacturer      model displ year  cyl  trans drv  cty   hwy fl   class Total
1         audi          a4  1.8  1999    4   auto(l5)  f   18   29  p compact   47
2         audi          a4  1.8  1999    4 manual(m5)  f   21   29  p compact   50
3         audi          a4  2.0  2008    4 manual(m6)  f   20   31  p compact   51
4         audi          a4  2.0  2008    4   auto(av)  f   21   30  p compact   51
5         audi          a4  2.8  1999    6   auto(l5)  f   16   26  p compact   42
6         audi          a4  2.8  1999    6 manual(m5)  f   18   26  p compact   44
7         audi          a4  3.1  2008    6   auto(av)  f   18   27  p compact   45
8         audi      a4 quattro  1.8  1999    4 manual(m5)  4   18   26  p compact   44
9         audi      a4 quattro  1.8  1999    4   auto(l5)  4   16   25  p compact   41
10        audi      a4 quattro  2.0  2008    4 manual(m6)  4   20   28  p compact   48
11        audi      a4 quattro  2.0  2008    4   auto(s6)  4   19   27  p compact   46
12        audi      a4 quattro  2.8  1999    6   auto(l5)  4   15   25  p compact   40
13        audi      a4 quattro  2.8  1999    6 manual(m5)  4   17   25  p compact   42
14        audi      a4 quattro  3.1  2008    6   auto(s6)  4   17   25  p compact   42

```

Q2. 앞에서 만든 '합산 연비 변수'를 2로 나눠 '평균 연비 변수'를 추가
A2.

```
> Mpg_New <- Mpg_New %>% mutate(Avg= Total/2)
> Mpg_New
```

	manufacturer	model	displ	year	cyl	trans	drv	cty	hwy	fl	class	Total	Avg
1	audi	a4	1.8	1999	4	auto(l5)	f	18	29	p	compact	47	23.5
2	audi	a4	1.8	1999	4	manual(m5)	f	21	29	p	compact	50	25.0
3	audi	a4	2.0	2008	4	manual(m6)	f	20	31	p	compact	51	25.5
4	audi	a4	2.0	2008	4	auto(av)	f	21	30	p	compact	51	25.5
5	audi	a4	2.8	1999	6	auto(l5)	f	16	26	p	compact	42	21.0
6	audi	a4	2.8	1999	6	manual(m5)	f	18	26	p	compact	44	22.0
7	audi	a4	3.1	2008	6	auto(av)	f	18	27	p	compact	45	22.5
8	audi	a4 quattro	1.8	1999	4	manual(m5)	4	18	26	p	compact	44	22.0
9	audi	a4 quattro	1.8	1999	4	auto(l5)	4	16	25	p	compact	41	20.5
10	audi	a4 quattro	2.0	2008	4	manual(m6)	4	20	28	p	compact	48	24.0
11	audi	a4 quattro	2.0	2008	4	auto(s6)	4	19	27	p	compact	46	23.0
12	audi	a4 quattro	2.8	1999	6	auto(l5)	4	15	25	p	compact	40	20.0
13	audi	a4 quattro	2.8	1999	6	manual(m5)	4	17	25	p	compact	42	21.0
14	audi	a4 quattro	3.1	2008	6	auto(s6)	4	17	25	p	compact	42	21.0

Q3. '평균 연비 변수'가 가장 높은 자동차 3종의 데이터를 출력
A3.

```
> Mpg_New %>% arrange(desc(Avg)) %>% head(3)
```

	manufacturer	model	displ	year	cyl	trans	drv	cty	hwy	fl	class	Total	Avg
1	volkswagen	new beetle	1.9	1999	4	manual(m5)	f	35	44	d	subcompact	79	39.5
2	volkswagen	jetta	1.9	1999	4	manual(m5)	f	33	44	d	compact	77	38.5
3	volkswagen	new beetle	1.9	1999	4	auto(l4)	f	29	41	d	subcompact	70	35.0

Q4. Q1~Q3번 문제를 해결할 수 있는 하나로 연결된 dplyr구문을 만들어 출력
데이터는 복사본 대신 Mpg원본 이용.

A4.

- Q1~Q3번 문제를 해결할 수 있는 하나로 연결된 dplyr 구문:

```
Mpg %>% mutate(Total = cty + hwy, Avg = Total / 2) %>% arrange(desc(Avg))
%>% head(3)
```

- 실행 결과:

```
> Mpg %>% mutate(Total = cty + hwy, Avg = Total / 2) %>% arrange(desc(Avg)) %>% head(3)
```

	manufacturer	model	displ	year	cyl	trans	drv	cty	hwy	fl	class	Total	Avg
1	volkswagen	new beetle	1.9	1999	4	manual(m5)	f	35	44	d	subcompact	79	39.5
2	volkswagen	jetta	1.9	1999	4	manual(m5)	f	33	44	d	compact	77	38.5
3	volkswagen	new beetle	1.9	1999	4	auto(l4)	f	29	41	d	subcompact	70	35.0

6. 집단별로 요약하기

Q1. class는 "suv", "compact" 등 자동차를 특징에 따라 일곱 종류로 분류한 변수.

어떤 차종의 연비가 높은지 비교!! class별 cty의평균?

A1. "subcompact", "compact", "midsize", "minivan", "2seater", "suv",
"pickup" 순으로 cty(도시 연비)가 높다.

- 각 class 별 cty의 평균은 "subcompact"가 20.4, "compact"가 20.1,
"midsize"가 18.8, "minivan"이 15.8, "2seater"가 15.4, "suv"가 13.5,
"pickup"이 13이다.

Q2. 앞 문제의 출력 결과는 class 값 알파벳 순으로 정렬.

어떤 차종의 도시연비가 높은지 쉽게 알아볼 수 있도록, cty평균이 높은 순 정렬 & 출력.

A2.

```
> Mpg %>% group_by(class) %>% summarise(Mean_cty= mean(cty)) %>% arrange(desc(Mean_cty))
# A tibble: 7 x 2
  class      Mean_cty
  <chr>      <dbl>
1 subcompact 20.4
2 compact   20.1
3 midsize   18.8
4 minivan   15.8
5 2seater   15.4
6 suv       13.5
7 pickup    13
```

Q3. 어떤 회사 자동차의 hwy(고속도로 연비)가 가장 높은가?

hwy 평균이 가장 높은 회사 세 곳을 출력.

A3. “honda”의 hwy(고속도로 연비)의 평균이 높다.

```
> Mpg %>% group_by(manufacturer) %>% summarise(Mean_hwy= mean(hwy)) %>% arrange(desc(Mean_hwy)) %>%
head(3)
# A tibble: 3 x 2
  manufacturer Mean_hwy
  <chr>         <dbl>
1 honda        32.6
2 volkswagen   29.2
3 hyundai      26.9
```

Q4. 어떤 회사에서 "compact"(경차) 차종을 가장 많이 생산하는가?

각 회사별 "compact" 차종 수를 내림차순으로 정렬 & 출력.

A4. “audi”에서 “compact”(경차) 차종을 가장 많이 생산한다.

```
> Mpg %>% filter(class== "compact") %>% group_by(manufacturer) %>% summarise(count = n()) %>% arrange(desc(count))
# A tibble: 5 x 2
  manufacturer count
  <chr>          <int>
1 audi           15
2 volkswagen     14
3 toyota         12
4 subaru         4
5 nissan         2
```

7. 데이터 합치기

자동차 연료별 가격을 나타낸 표

fl	연료종류	가격(갤런당 USD)
c	CNG	2.35
d	diesel	2.38
e	ethanol E85	2.11
p	premium	2.76
r	regular	2.22

fuel 데이터 만들기 (연료 가격을 담은 데이터 만들기)

```
Fuel <- data.frame(fl= c("c", "d", "e", "p", "r"),  
                  Price_Fl = c(2.35, 2.38, 2.11, 2.76, 2.22),  
                  stringsAsFactors = F)
```

Q1. 연료 가격을 나타낸 변수 생성!! 위에서 만든 fuel 데이터를 이용,
mpg 데이터에 price_fl(연료 가격) 변수를 추가!!

A1.

```
> Mpg <- left_join(Mpg, Fuel, by = "fl")  
> Mpg  
  manufacturer      model displ year cyl  trans drv  cty   hwy fl   class Price_Fl  
1         audi          a4   1.8 1999   4  auto(l5) f   18   29 p compact    2.76  
2         audi          a4   1.8 1999   4 manual(m5) f   21   29 p compact    2.76  
3         audi          a4   2.0 2008   4 manual(m6) f   20   31 p compact    2.76  
4         audi          a4   2.0 2008   4  auto(av) f   21   30 p compact    2.76  
5         audi          a4   2.8 1999   6  auto(l5) f   16   26 p compact    2.76  
6         audi          a4   2.8 1999   6 manual(m5) f   18   26 p compact    2.76  
7         audi          a4   3.1 2008   6  auto(av) f   18   27 p compact    2.76  
8         audi    a4 quattro   1.8 1999   4 manual(m5) f   18   26 p compact    2.76  
9         audi    a4 quattro   1.8 1999   4  auto(l5) f   16   25 p compact    2.76  
10        audi    a4 quattro   2.0 2008   4 manual(m6) f   20   28 p compact    2.76  
11        audi    a4 quattro   2.0 2008   4  auto(s6) f   19   27 p compact    2.76  
12        audi    a4 quattro   2.8 1999   6  auto(l5) f   15   25 p compact    2.76
```

Q2. 연료 가격 변수가 잘 추가됐는지 확인!!

model, fl, price_fl 변수를 추출 & 앞부분 5행 출력!!

A2.

```
> Mpg %>% select(model, fl, Price_Fl) %>% head(5)  
  model fl Price_Fl  
1    a4 p    2.76  
2    a4 p    2.76  
3    a4 p    2.76  
4    a4 p    2.76  
5    a4 p    2.76
```

II. 분석 도전 과제: midwest dataset을 바탕으로한 데이터 가공

사용한 자료 설명

midwest dataset은 R의 내장 데이터셋 중 하나로 2000년 미국 중서부 437개 지역의 인구통계학적 정보를 담고 있다.

28개의 변수를 가지고 있으며, 해당 지역의 카운티 이름, 주 이름, 지역의 면적, 전체 인구, 백인 인구, 흑인 인구, 아시아 인구 등의 정보가 변수로 나타내어져 있다.

제시된 문제의 답안

Q1. popadults(해당 지역의 성인 인구), poptotal(전체 인구)

midwest 데이터에 '전체 인구 대비 미성년 인구 백분율' 변수 추가!!

A1. `Midwest <- Midwest %>% mutate(underage = (poptotal-popadults)/poptotal*100)`

Q2. 미성년 인구 백분율이 가장 높은 상위 5개 county(지역)의 미성년 인구 백분율 출력

A2.

- 코드: `Midwest %>% arrange(desc(underage)) %>% select(county, underage) %>% head(5)`

- 실행결과:

```
> Midwest %>% arrange(desc(underage)) %>% select(county, underage) %>% head(5)
  county underage
1 ISABELLA  51.50117
2 MENOMINEE  50.59126
3   ATHENS   49.32073
4  MECOSTA   49.05918
5  MONROE    47.35818
```

Q3. 분류표의 기준에 따라 미성년 비율 등급 변수 추가,

분류	기준
Large	40% 이상
Middle	30% ~ 40% 미만
Small	30% 미만

각 등급에 몇 개의 지역이 있는가?

A3.

- 위 분류표의 기준에 따라 미성년 비율 등급(grade)을 추가하는 코드:

```
Midwest <- Midwest %>% mutate(grade= ifelse(underage >= 40, "Large",
ifelse(underage >= 30, "Middle", "Small")))
```

- 각 등급의 몇 개의 지역이 있는지 알아보기 위해 grade를 기준으로 표를 출력하는 코드: `table(Midwest$grade)`

- 실행결과

```
> table(Midwest$grade)
```

```
Large Middle Small
   32    396     9
```

- Large 등급에는 32개의 지역, Middle 등급에는 396개의 지역, Small 등급에는 9개의 지역이 있다.

Q4. popasian(해당 지역의 아시아인 인구)

'전체 인구 대비 아시아인 인구 백분율' 변수 추가, 하위 10개 지역의 state(주), county(지역명), 아시아인 인구 백분율 출력.

A4. Midwest %>% mutate(asian = (popasian/poptotal)*100) %>% arrange(asian) %>% select(state, county, asian) %>% head(10)

- 실행결과

```
> Midwest %>% mutate(asian = (popasian/poptotal)*100) %>% arrange(asian) %>% select(state, county, asian) %>% head(10)
```

	state	county	asian
1	WI	MENOMINEE	0.00000000
2	IN	BENTON	0.01059210
3	IN	CARROLL	0.01594981
4	OH	VINTON	0.02703190
5	WI	IRON	0.03250447
6	IL	SCOTT	0.05315379
7	IN	CLAY	0.06071645
8	MI	OSCODA	0.06375925
9	OH	PERRY	0.06654625
10	IL	PIATT	0.07074865

결론도출

1. 미성년 인구 비율 상위 5개 지역은 ISABELLA, MENOMINEE, ATHENS, MECOSTA, MONROE이다.
1-1. 이 중 ISABELLA와 MENOMINEE는 미성년 비율이 전체의 50% 이상이다.
2. 미성년 인구가 전체 40%이상일 때 Large, 30%이상 40%미만일 때 Middle, 30%미만일 때 Small이라 분류할 때 32개의 county가 Large, 396개의 county가 Middle, 9개의 county가 small에 해당된다.
3. 전체 인구 대비 아시아인 인구 백분율의 경우 MENOMINEE가 0%로 가장 낮았다.

부록: 분석 도전 과제의 R script

```
Midwest <- as.data.frame(ggplot2::midwest)
```

```
# Q1. popadults(해당 지역의 성인 인구), poptotal(전체 인구) midwest 데이터에  
'전체 인구 대비 미성년 인구 백분율' 변수 추가!!
```

```
Midwest <- Midwest %>% mutate(underage =  
(poptotal-popadults)/poptotal*100)
```

```
# Q2. 미성년 인구 백분율이 가장 높은 상위 5개 county(지역)의 미성년 인구 백분  
율 출력
```

```
Midwest %>% arrange(desc(underage)) %>% select(county, underage) %>%  
head(5)
```

```
# Q3. 분류표의 기준에 따라 미성년 비율 등급 변수 추가, 각 등급에 몇 개의 지역이  
있는가?
```

```
Midwest <- Midwest %>% mutate(grade= ifelse(underage >= 40, "Large",
```

```
ifelse(underage >= 30, "Middle", "Small"))))  
table(Midwest$grade)
```

Q4. popasian(해당 지역의 아시아인 인구), '전체 인구 대비 아시아인 인구 백분율' 변수 추가, 하위 10개 지역의 state(주), county(지역명), 아시아인 인구 백분율을 출력.

```
Midwest %>% mutate(asian = (popasian/poptotal)*100) %>% arrange(asian)  
%>% select(state, county, asian) %>% head(10)
```