# 06 데이터 시각화(1)

## 학습 내용

- ggplot2 패키지 알아보기
- 산점도 그래보기
- qplot과 ggplot의 차이 알아보기
- 그래프 이미지, pdf로 저장해 보기

## 6-1 그래프 or 차트 or 다이어그램

- 일련의 데이터에 대한 추세를 보여준다.
- 데이터를 그래프로 표현하면 **추세와 경향성**이 드러나 특징을 쉽게 이해 가능하다.
- 그래프를 만드는 과정중에 **새로운 패턴이 발견**되기도 한다.
- R이 제공하는 기본 함수가 있고, 현재 많이 사용되고 있는 그래프 패키지는 ggplot2이다.

## ggplot2

- ggplot2는 그래프를 만들 때 가장 많이 사용하는 패키지.
- 짧은 문법을 통해 그래프를 만들 수 있다.
- ggplot2는 레이어 구조로 되어 있음.

| 단계 | 설명 |
|---|---|
| 1단계 | 배경 설정(축) |
| 2단계 | 그래프 추가(점, 막대, 선) |
| 3단계 | 설정 추가(축 범위, 색, 표식) |

## 6-2 산점도(Scatterplot)

- 두 수치형 변수의 관계를 표시한다.
- 각 점은 관측치를 나타낸다.

In [1]:

```
library(ggplot2)
```

In [6]:

```
head(mpg,8)
```

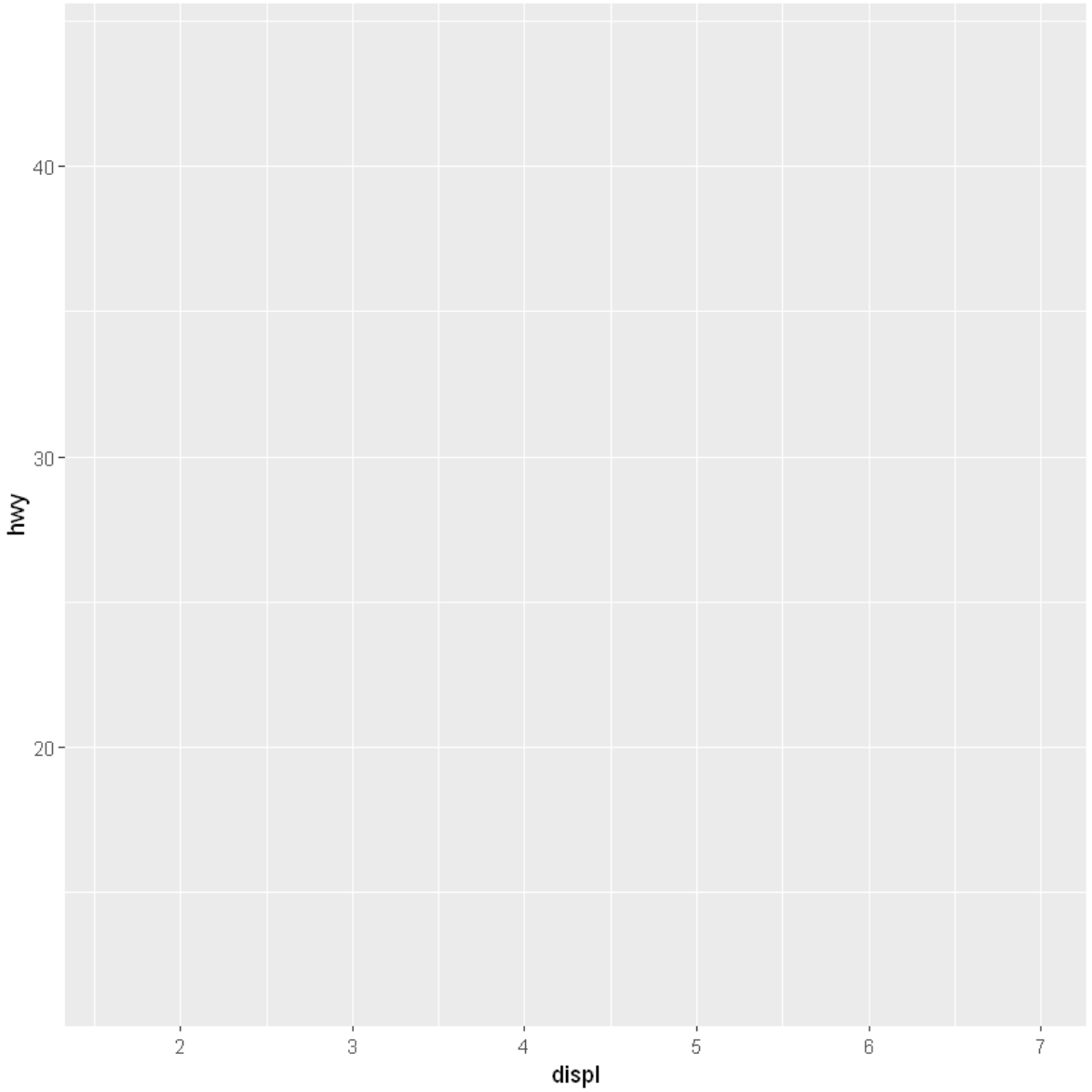
| manufacturer | model | displ | year | cyl | trans | drv | cty | hwy | fl | class |
|---|---|---|---|---|---|---|---|---|---|---|
| audi | a4 | 1.8 | 1999 | 4 | auto(l5) | f | 18 | 29 | p | compact |
| audi | a4 | 1.8 | 1999 | 4 | manual(m5) | f | 21 | 29 | p | compact |
| audi | a4 | 2.0 | 2008 | 4 | manual(m6) | f | 20 | 31 | p | compact |
| audi | a4 | 2.0 | 2008 | 4 | auto(av) | f | 21 | 30 | p | compact |
| audi | a4 | 2.8 | 1999 | 6 | auto(l5) | f | 16 | 26 | p | compact |
| audi | a4 | 2.8 | 1999 | 6 | manual(m5) | f | 18 | 26 | p | compact |
| audi | a4 | 3.1 | 2008 | 6 | auto(av) | f | 18 | 27 | p | compact |
| audi | a4 quattro | 1.8 | 1999 | 4 | manual(m5) | 4 | 18 | 26 | p | compact |

## 1단계 배경 및 축설정

- data = 데이터 셋
- aes(x=col1, y=col2) : x축과 y축에 사용할 변수를 지정
- displ : 배기량, hwy : 고속도로 연비

In [4]:

```
ggplot(data = mpg, aes(x=displ, y=hwy))
```

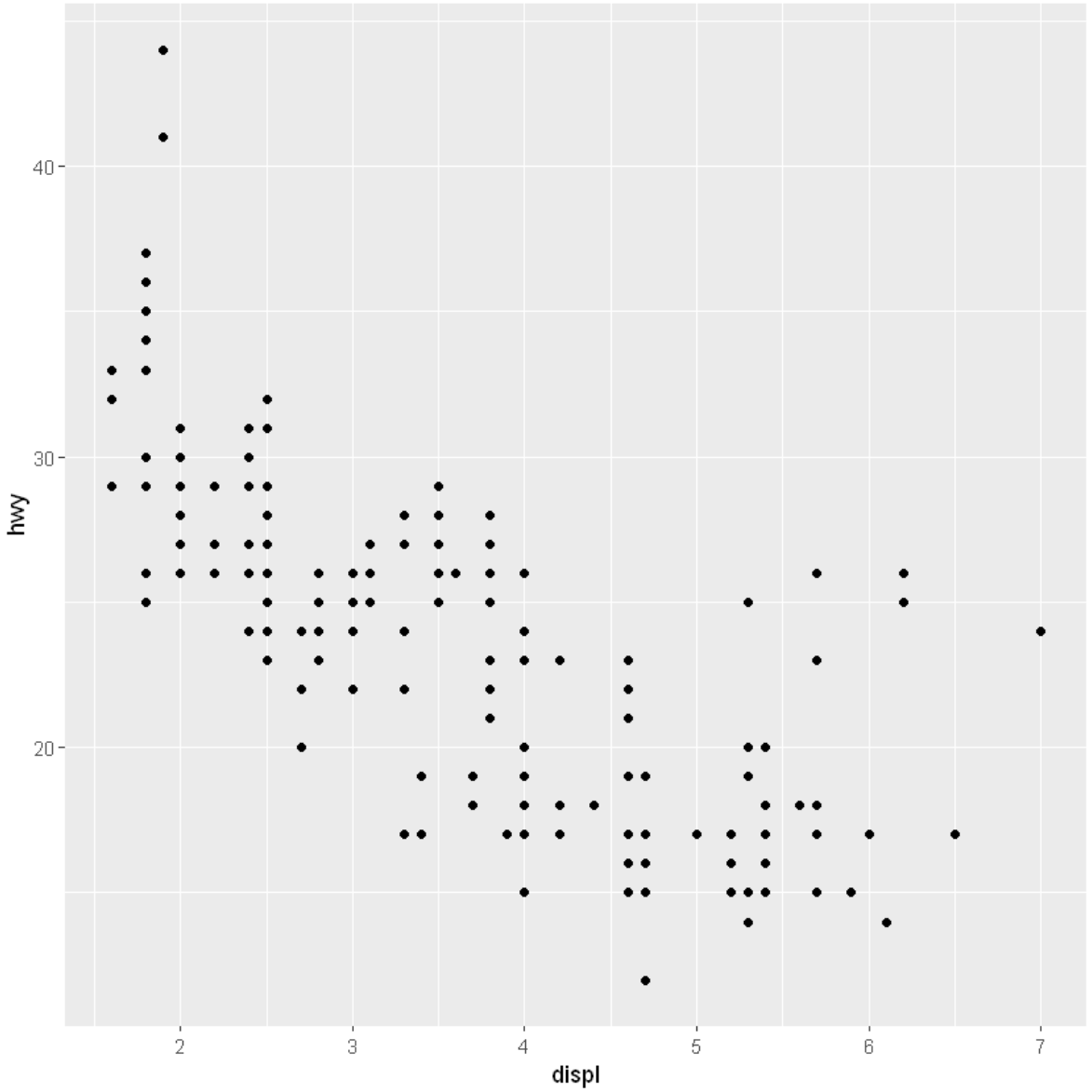



## 2단계 그래프 추가

- 배경에 산점도를 추가한다. + geom_point()
- '+' 기호를 사용하여 패키지 함수를 연결한다.
- 산점도 위에 표시된 점은 각각의 데이터(자동차 모델을 의미)

In [5]:

```
ggplot(data=mpg, aes(x=displ, y=hwy)) + geom_point()
```



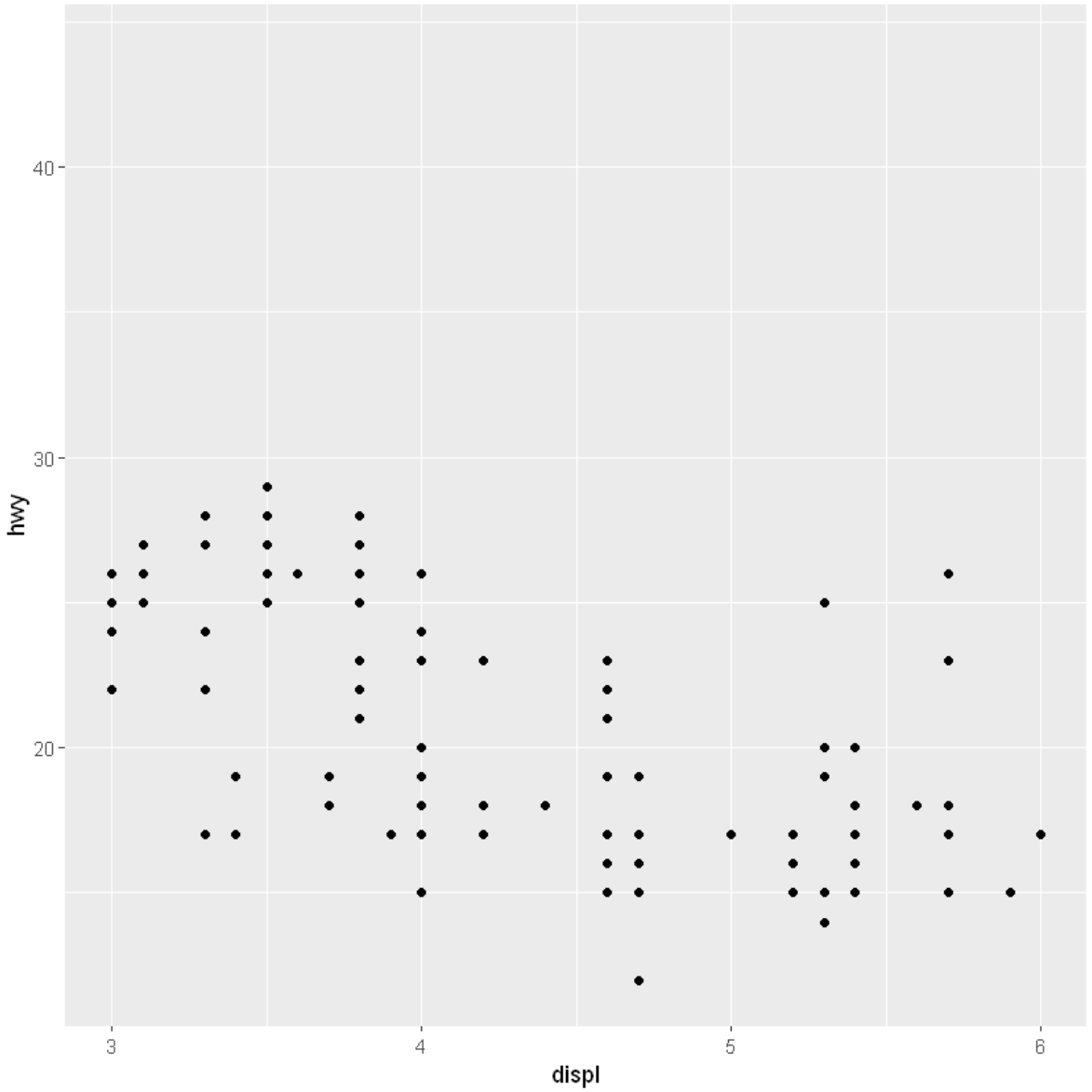

## 3단계 축 범위를 조정하는 설정 추가

- '+' 기호를 이용하여 그래프 설정 변경 코드 추가
- 축범위 설정 : xlim(), ylim()
- warning message는 105행의 데이터가 보이지 않는다는 알림.

In [9]:

```
ggplot(data=mpg, aes(x=displ, y=hwy)) + geom_point() + xlim(3,6)
```

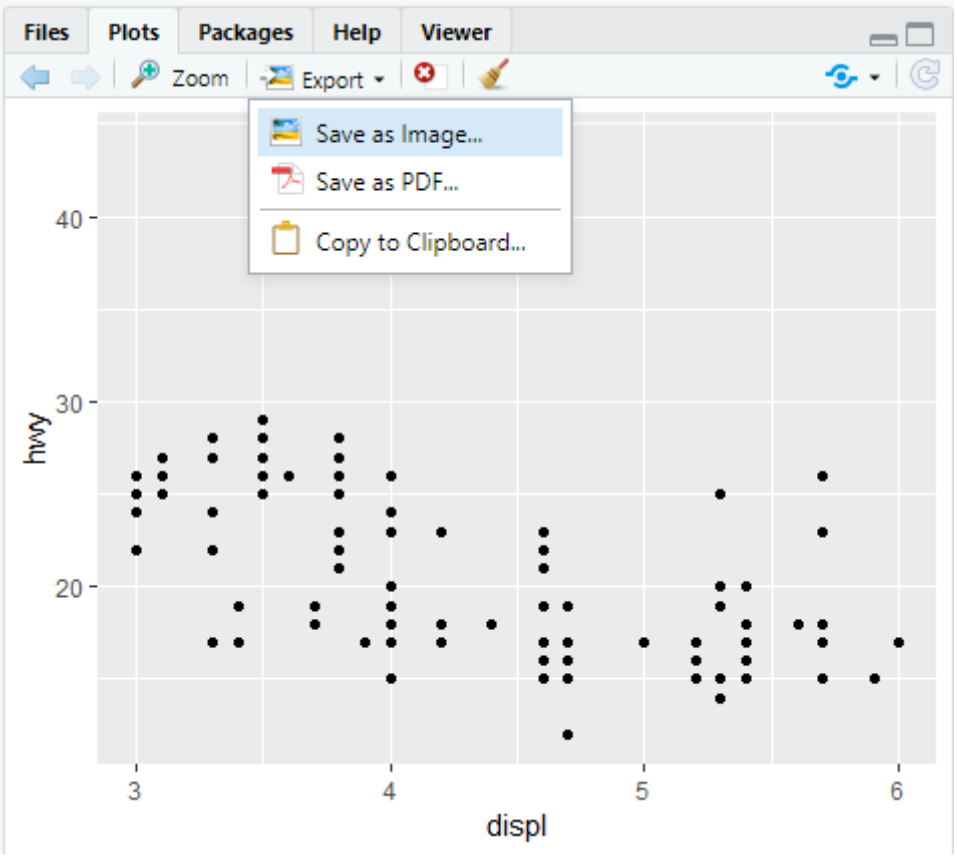
```
Warning message:
"Removed 105 rows containing missing values (geom_point)."
```




## (직접해보기 01) ylim 범위를 지정해 보자.

## 그래프를 이미지 파일로 저장해 보기

- [Export] 버튼 선택 - [Save as Image] 선택 : 이미지 파일로 선택
- [Export] 버튼 선택 - [Save as PDF] 선택 : PDF 포맷으로 선택
- [Export] 버튼 선택 - [Copy to Clipboard] 선택 : 그래프를 메모리에 저장하는 기능




## qplot() vs ggplot()

- qplot()은 기능은 상대적으로 적지만 문법이 간단하여 주로 전처리 단계에서 빠르게 데이터 확인 용도로 사용
- ggplot() 분석 결과를 보고하기 위해 그래프를 만들 때는 ggplot()을 사용

## (실습2-1)

- mpg 데이터 셋의 cty(도시 연비)와 hwy(고속도로 연비)간의 관계를 알아보기 위해 산점도를 그려보자.
- midwest(데이터 셋)의 전체인구와 아시아 인구간의 어떤 관계가 있는지 알아보자.
  - poptotal(전체인구), popwhite(백인 인구)
  - 전체인구는 50만명 이하, 백인 인구는 8만명 이하인 지역만 설정.

In [10]:

```
head(midwest)
```

| PID | county | state | area | poptotal | popdensity | popwhite | popblack | popamerindian | popa |
|---|---|---|---|---|---|---|---|---|---|
| 561 | ADAMS | IL | 0.052 | 66090 | 1270.9615 | 63917 | 1702 | 98 | |
| 562 | ALEXANDER | IL | 0.014 | 10626 | 759.0000 | 7054 | 3496 | 19 | |
| 563 | BOND | IL | 0.022 | 14991 | 681.4091 | 14477 | 429 | 35 | |
| 564 | BOONE | IL | 0.017 | 30806 | 1812.1176 | 29344 | 127 | 46 | |
| 565 | BROWN | IL | 0.018 | 5836 | 324.2222 | 5264 | 547 | 14 | |
| 566 | BUREAU | IL | 0.050 | 35688 | 713.7600 | 35157 | 50 | 65 | |

In [11]:

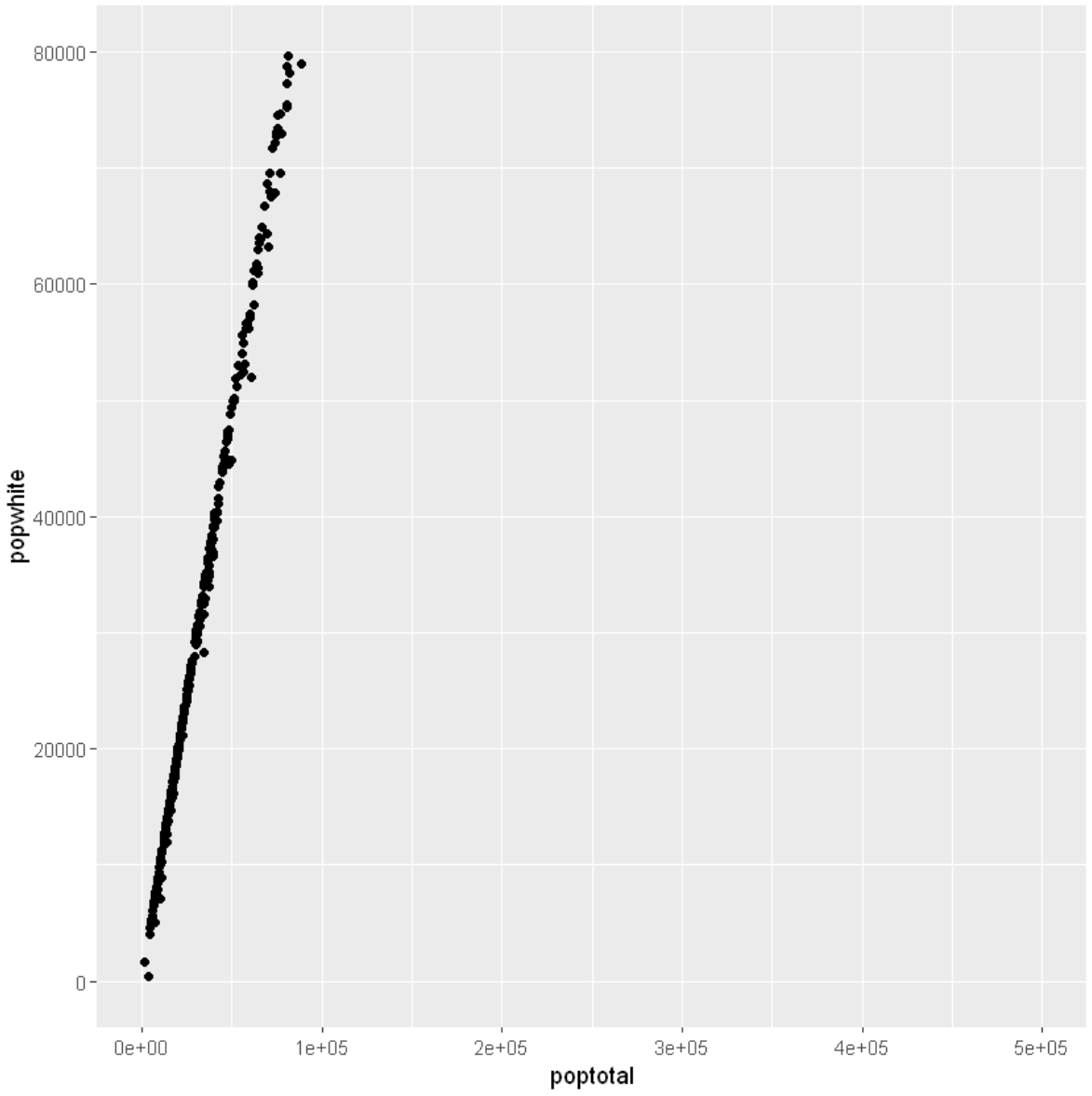
```
summary(midwest)
```

```
      PID          county             state                area        
 Min.   : 561   Length:437         Length:437         Min.   :0.00500  
 1st Qu.: 670   Class :character   Class :character   1st Qu.:0.02400  
 Median :1221   Mode  :character   Mode  :character   Median :0.03000  
 Mean   :1437                                         Mean   :0.03317  
 3rd Qu.:2059                                         3rd Qu.:0.03800  
 Max.   :3052                                         Max.   :0.11000  
    poptotal         popdensity          popwhite          popblack      
 Min.   :   1701   Min.   :   85.05   Min.   :   416   Min.   :     0  
 1st Qu.:  18840   1st Qu.:  622.41   1st Qu.:  18630   1st Qu.:    29  
 Median :  35324   Median : 1156.21   Median :  34471   Median :   201  
 Mean   :  96130   Mean   : 3097.74   Mean   :  81840   Mean   :  11024  
 3rd Qu.:  75651   3rd Qu.: 2330.00   3rd Qu.:  72968   3rd Qu.:   1291  
 Max.   :5105067   Max.   :88018.40   Max.   :3204947   Max.   :1317147  
 popamerindian        popasian         popother        percwhite     
 Min.   :    4.0   Min.   :     0   Min.   :     0   Min.   :10.69  
 1st Qu.:   44.0   1st Qu.:    35   1st Qu.:    20   1st Qu.:94.89  
 Median :   94.0   Median :   102   Median :    66   Median :98.03  
 Mean   :  343.1   Mean   :  1310   Mean   :  1613   Mean   :95.56  
 3rd Qu.:  288.0   3rd Qu.:   401   3rd Qu.:   345   3rd Qu.:99.07  
 Max.   :10289.0   Max.   :188565   Max.   :384119   Max.   :99.82  
   percblack        percamerindan        percasian        percother      
 Min.   : 0.0000   Min.   : 0.05623   Min.   :0.0000   Min.   :0.00000  
 1st Qu.: 0.1157   1st Qu.: 0.15793   1st Qu.:0.1737   1st Qu.:0.09102  
 Median : 0.5390   Median : 0.21502   Median :0.2972   Median :0.17844  
 Mean   : 2.6763   Mean   : 0.79894   Mean   :0.4872   Mean   :0.47906  
 3rd Qu.: 2.6014   3rd Qu.: 0.38362   3rd Qu.:0.5212   3rd Qu.:0.48050  
 Max.   :40.2100   Max.   :89.17738   Max.   :5.0705   Max.   :7.52427  
   popadults          perchsd         percollege        percprof      
 Min.   :   1287   Min.   :46.91   Min.   : 7.336   Min.   : 0.5203  
 1st Qu.:  12271   1st Qu.:71.33   1st Qu.:14.114   1st Qu.: 2.9980  
 Median :  22188   Median :74.25   Median :16.798   Median : 3.8142  
 Mean   :  60973   Mean   :73.97   Mean   :18.273   Mean   : 4.4473  
 3rd Qu.:  47541   3rd Qu.:77.20   3rd Qu.:20.550   3rd Qu.: 4.9493  
 Max.   :3291995   Max.   :88.90   Max.   :48.079   Max.   :20.7913  
 poppovertyknown   percpovertyknown percbelowpoverty percchildbelowpovert
 Min.   :   1696   Min.   :80.90    Min.   : 2.180   Min.   : 1.919      
 1st Qu.:  18364   1st Qu.:96.89    1st Qu.: 9.199   1st Qu.:11.624      
 Median :  33788   Median :98.17    Median :11.822   Median :15.270      
 Mean   :  93642   Mean   :97.11    Mean   :12.511   Mean   :16.447      
 3rd Qu.:  72840   3rd Qu.:98.60    3rd Qu.:15.133   3rd Qu.:20.352      
 Max.   :5023523   Max.   :99.86    Max.   :48.691   Max.   :64.308      
 percadultpoverty percelderlypoverty    inmetro         category        
 Min.   : 1.938   Min.   : 3.547     Min.   :0.0000   Length:437        
 1st Qu.: 7.668   1st Qu.: 8.912     1st Qu.:0.0000   Class :character  
 Median :10.008   Median :10.869     Median :0.0000   Mode  :character  
 Mean   :10.919   Mean   :11.389     Mean   :0.3432                     
 3rd Qu.:13.182   3rd Qu.:13.412     3rd Qu.:1.0000                     
 Max.   :43.312   Max.   :31.162     Max.   :1.0000                     
```

In [13]:

```
ggplot(data=midwest, aes(x=poptotal, y=popwhite)) +
    geom_point() +
    xlim(0, 500000) +
    ylim(0, 80000)
```

```
Warning message:
"Removed 98 rows containing missing values (geom_point)."
```




## REF

- The R Graph Gallery : https://www.r-graph-gallery.com/ (https://www.r-graph-gallery.com/)