EXPLORATORY DATA ANALYSIS

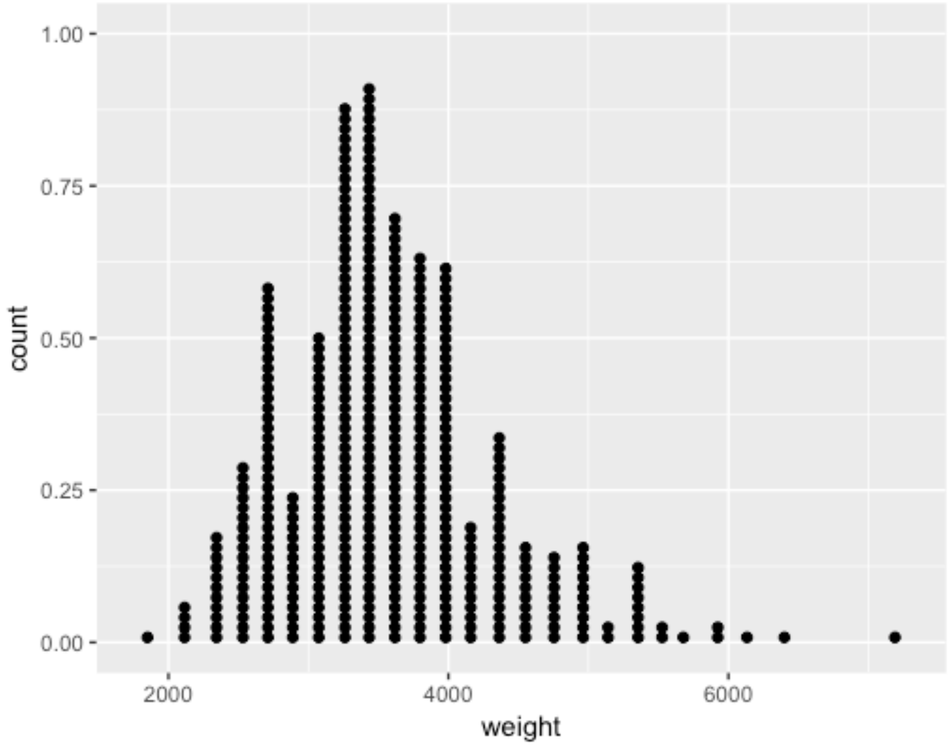
Exploring numerical data

Cars dataset

```
> str(cars)
  Classes ‘tbl_df’, ‘tbl’ and 'data.frame':  428 obs. of  19 variables:
  $ name       : chr  "Chevrolet Aveo 4dr" "Chevrolet Aveo LS 4dr hatch" ...
  $ sports_car : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
  $ suv        : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
  $ wagon      : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
  $ minivan    : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
  $ pickup     : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
  $ all_wheel  : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
  $ rear_wheel : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
  $ msrp       : int  11690 12585 14610 14810 16385 13670 15040 13270 ...
  $ dealer_cost: int  10965 11802 13697 13884 15357 12849 14086 12482 ...
  $ eng_size   : num  1.6 1.6 2.2 2.2 2.2 2 2 2 2 2 ...
  $ ncyl       : int  4 4 4 4 4 4 4 4 4 4 ...
  $ horsepwr   : int  103 103 140 140 140 132 132 130 110 130 ...
  $ city_mpg   : int  28 28 26 26 26 29 29 26 27 26 ...
  $ hwy_mpg    : int  34 34 37 37 37 36 36 33 36 33 ...
  $ weight     : int  2370 2348 2617 2676 2617 2581 2626 2612 2606 ...
  $ wheel_base : int  98 98 104 104 104 105 105 103 103 103 ...
  $ length     : int  167 153 183 183 183 174 174 168 168 168 ...
  $ width      : int  66 66 69 68 69 67 67 67 67 67 ...
```

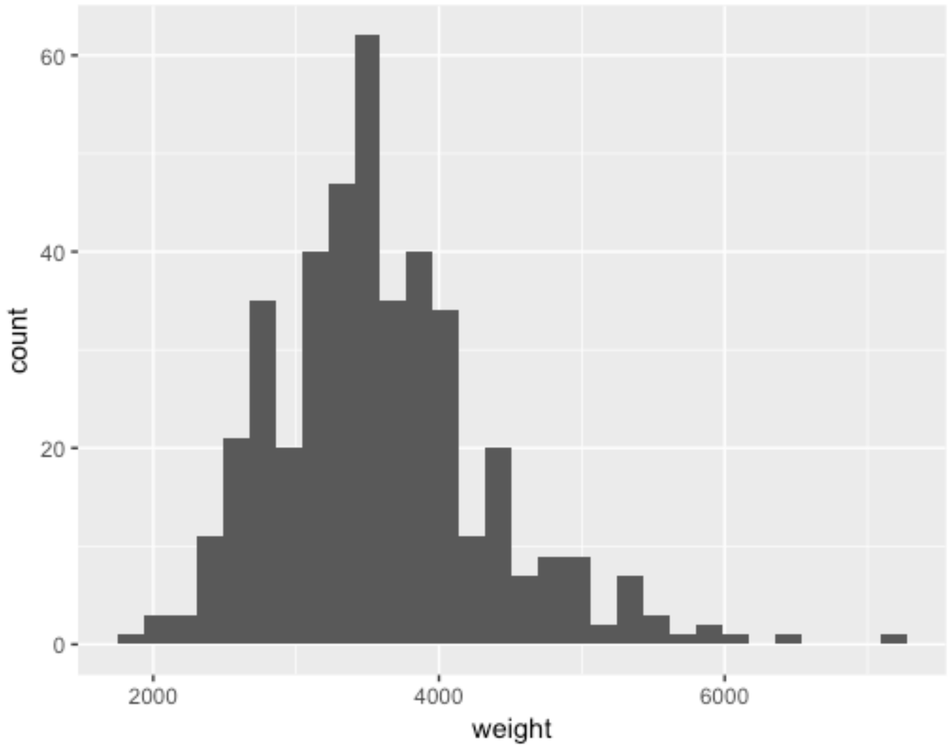
Dotplot

```
> ggplot(data, aes(x = weight)) +
    geom_dotplot(dotsize = 0.4)
```



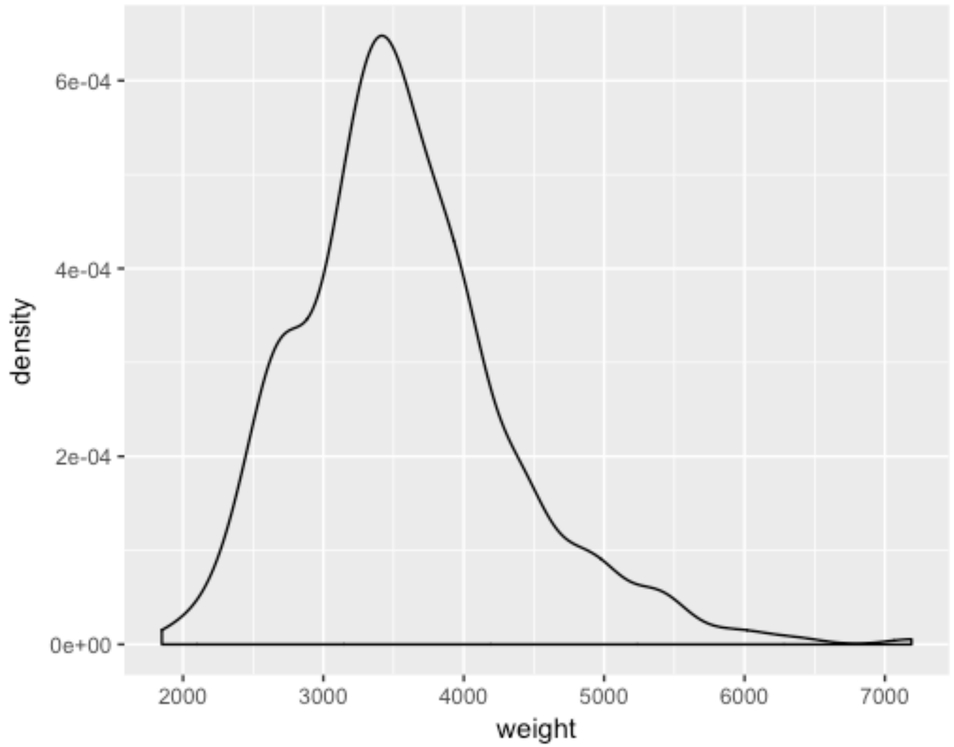
Histogram

```
> ggplot(data, aes(x = weight)) +
    geom_histogram()
```



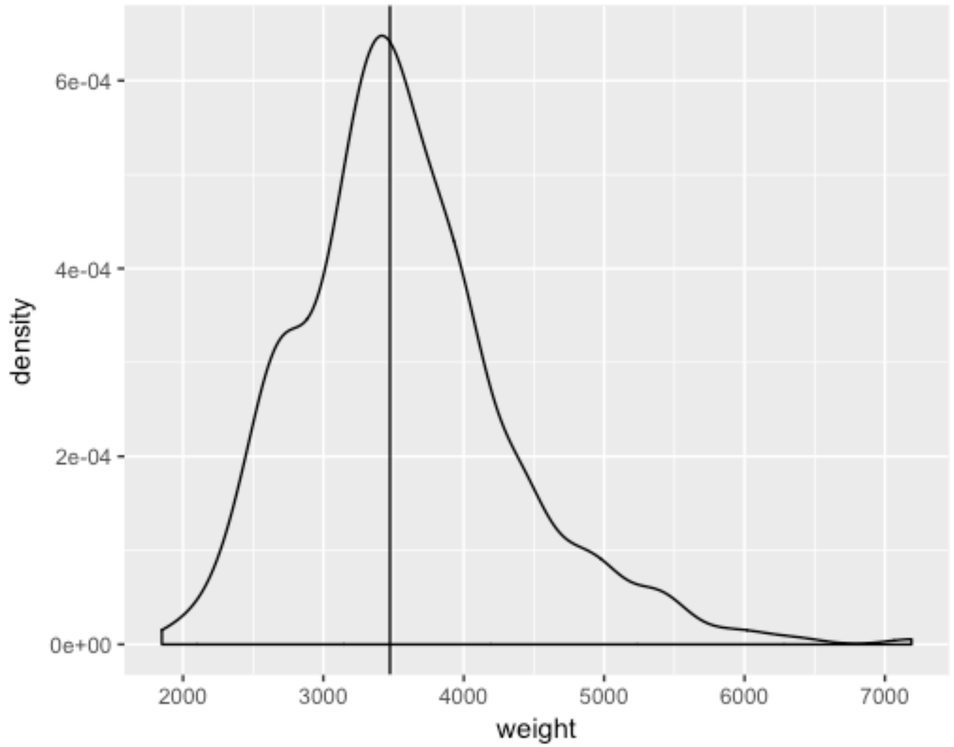
Density plot

```
> ggplot(data, aes(x = weight)) +
    geom_density()
```



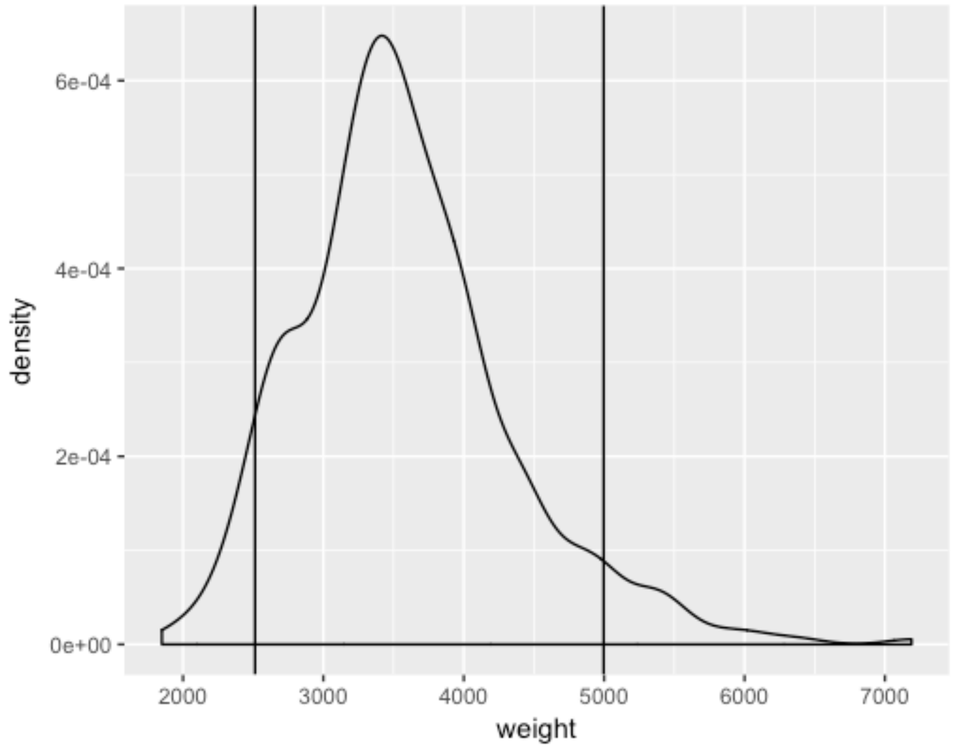
Density plot

```
> ggplot(data, aes(x = weight)) +
    geom_density()
```



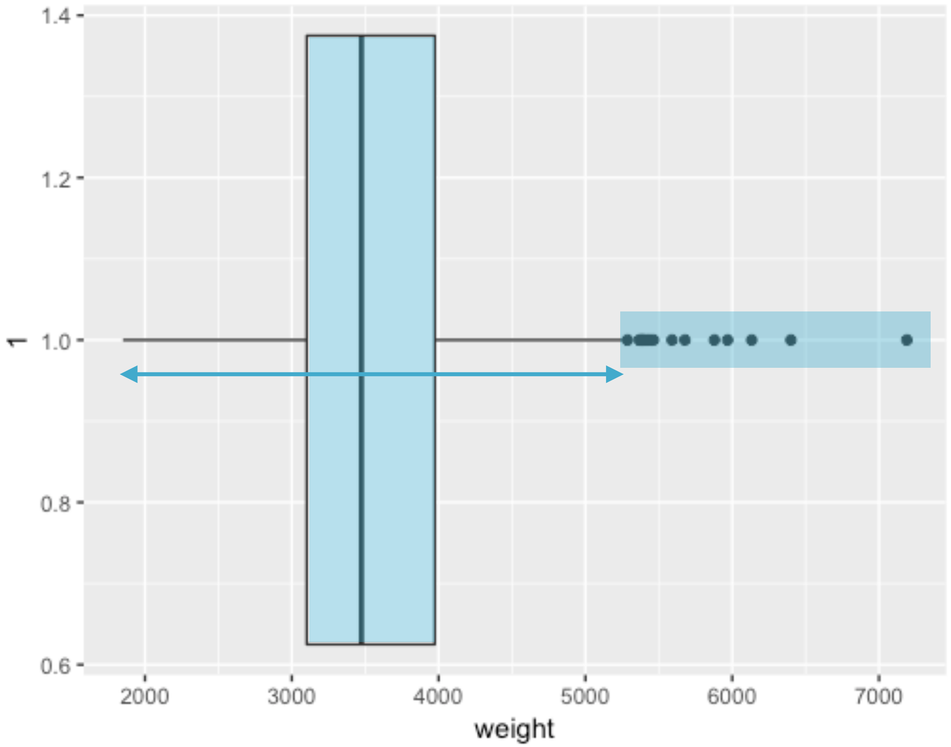
Density plot

```
> ggplot(data, aes(x = weight)) +
    geom_density()
```



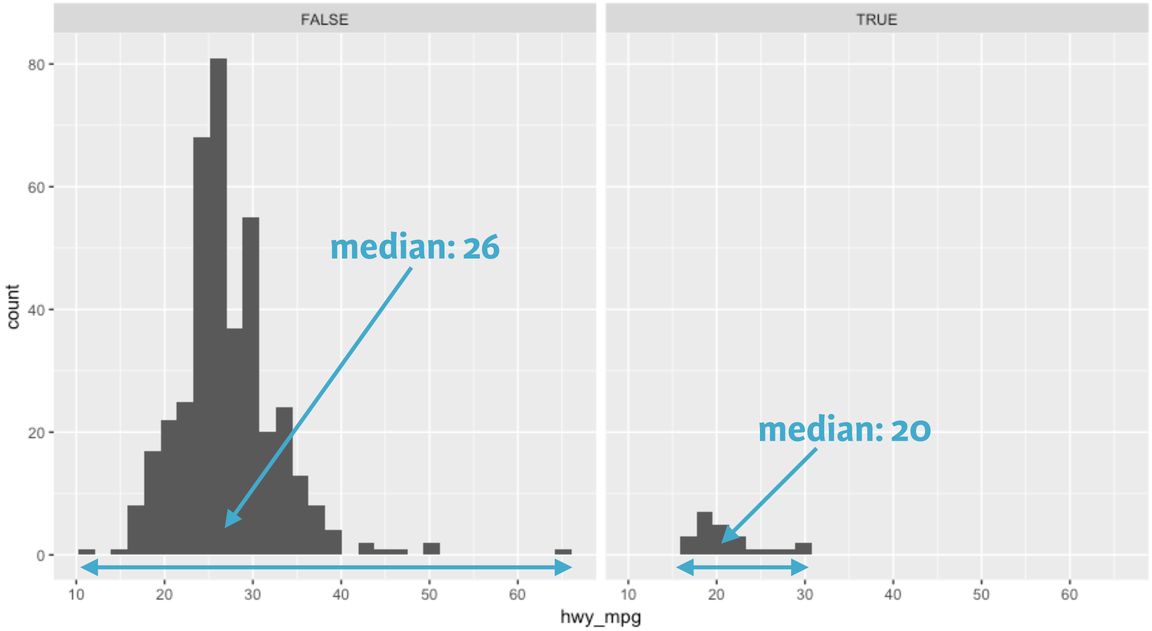
Boxplot

```
> ggplot(data, aes(x = 1, y = weight)) +
    geom_boxplot() +
    coord_flip()
```



Faceted histogram

```
> ggplot(cars, aes(x = hwy_mpg)) +
    geom_histogram() +
    facet_wrap(~pickup)
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
Warning message:
Removed 14 rows containing non-finite values (stat_bin).
```



EXPLORATORY DATA ANALYSIS

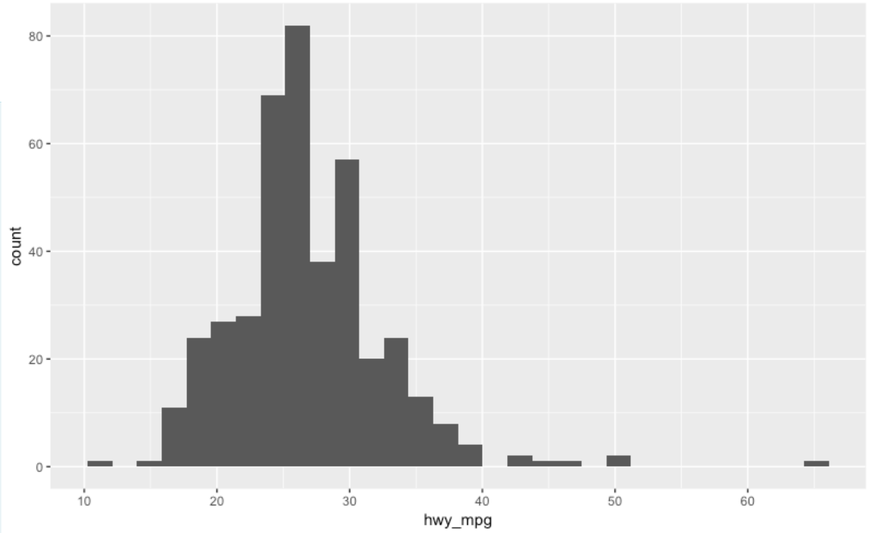
Let’s practice!

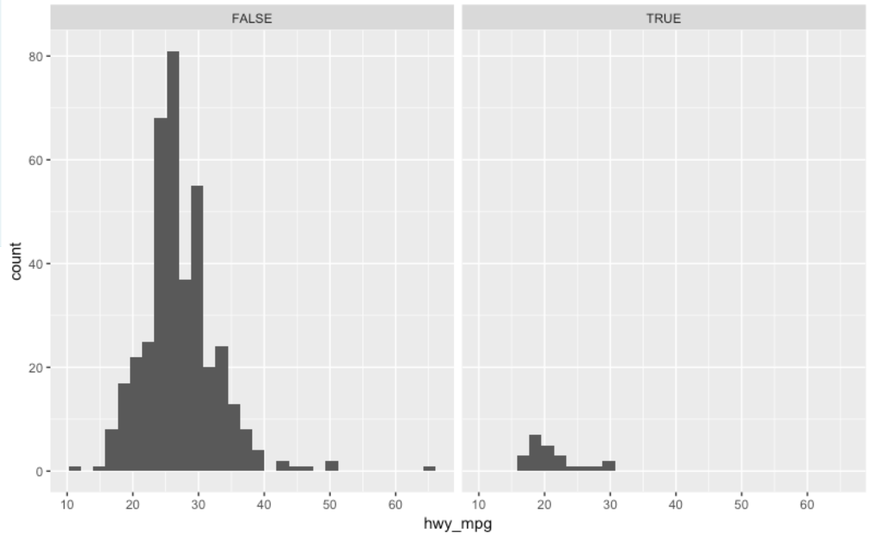
EXPLORATORY DATA ANALYSIS

Distribution of one variable

Marginal vs. conditional

```
> ggplot(cars, aes(x = hwy_mpg)) +
    geom_histogram()
`stat_bin()` using `bins = 30`. Pick better value with
Warning message:
Removed 14 rows containing non-finite values (stat_bin)

> ggplot(cars, aes(x = hwy_mpg)) +
    geom_histogram() +
    facet_wrap(~pickup)
`stat_bin()` using `bins = 30`. Pick better value with
Warning message:
Removed 14 rows containing non-finite values (stat_bin)
```





Building a data pipeline

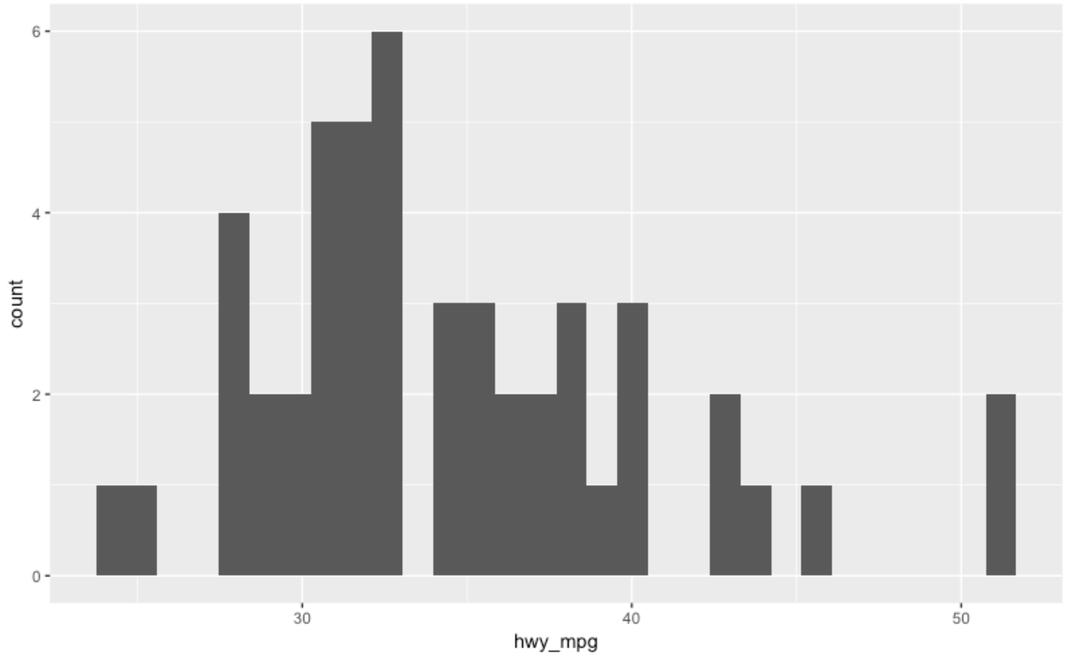
```
cars2 <- cars %>%
  filter(eng_size < 2.0)

ggplot(cars2, aes(x = hwy_mpg)) +
  geom_histogram()
```

Building a data pipeline

```
cars %>%
  filter(eng_size < 2.0) %>%
  ggplot(aes(x = hwy_mpg)) +
  geom_histogram()
```

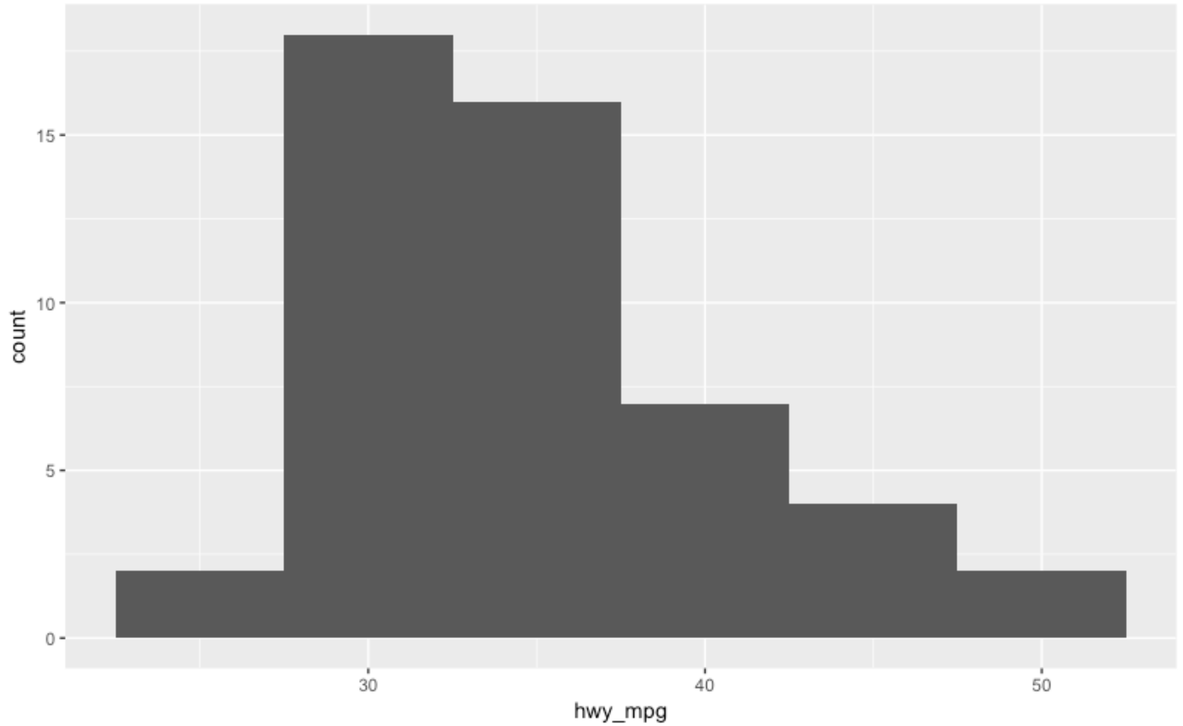
Filtered and faceted histogram

```
> cars %>%
    filter(eng_size < 2.0) %>%
    ggplot(aes(x = hwy_mpg)) +
    geom_histogram()
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



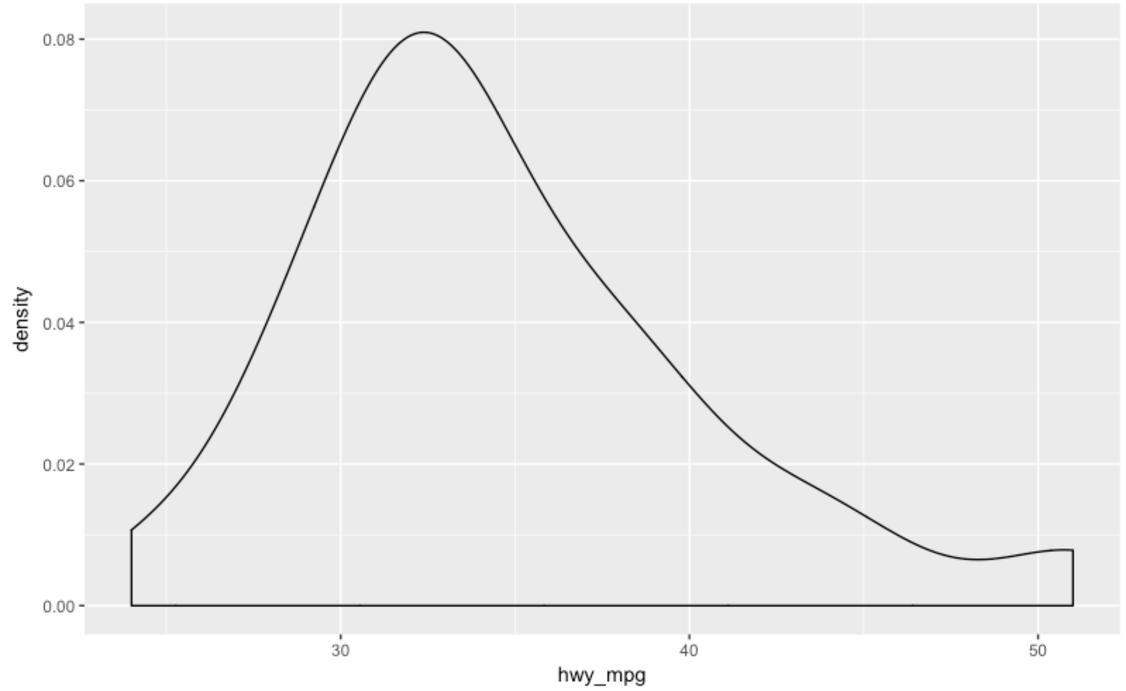
Wide bin width

```
> cars %>%
    filter(eng_size < 2.0) %>%
    ggplot(aes(x = hwy_mpg)) +
    geom_histogram(binwidth = 5)
```



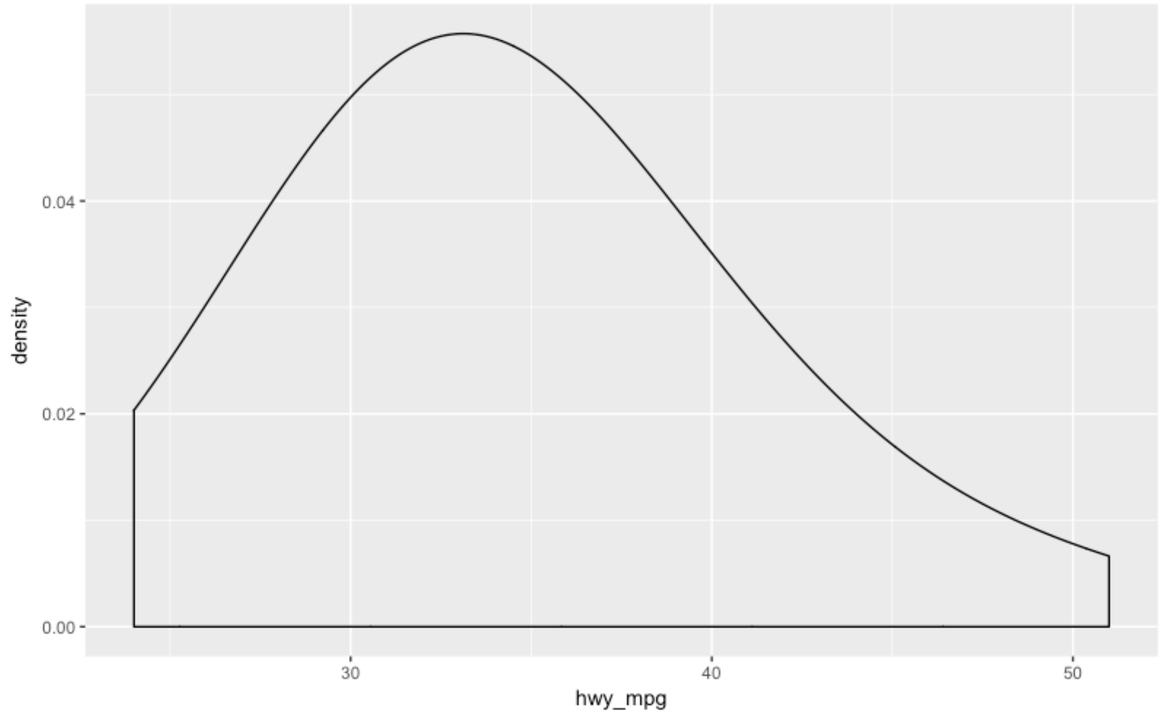
Density plot

```
> cars %>%
    filter(eng_size < 2.0) %>%
    ggplot(aes(x = hwy_mpg)) +
    geom_density()
```



Wide bandwidth

```
> cars %>%
    filter(eng_size < 2.0) %>%
    ggplot(aes(x = hwy_mpg)) +
    geom_density(bw = 5)
```

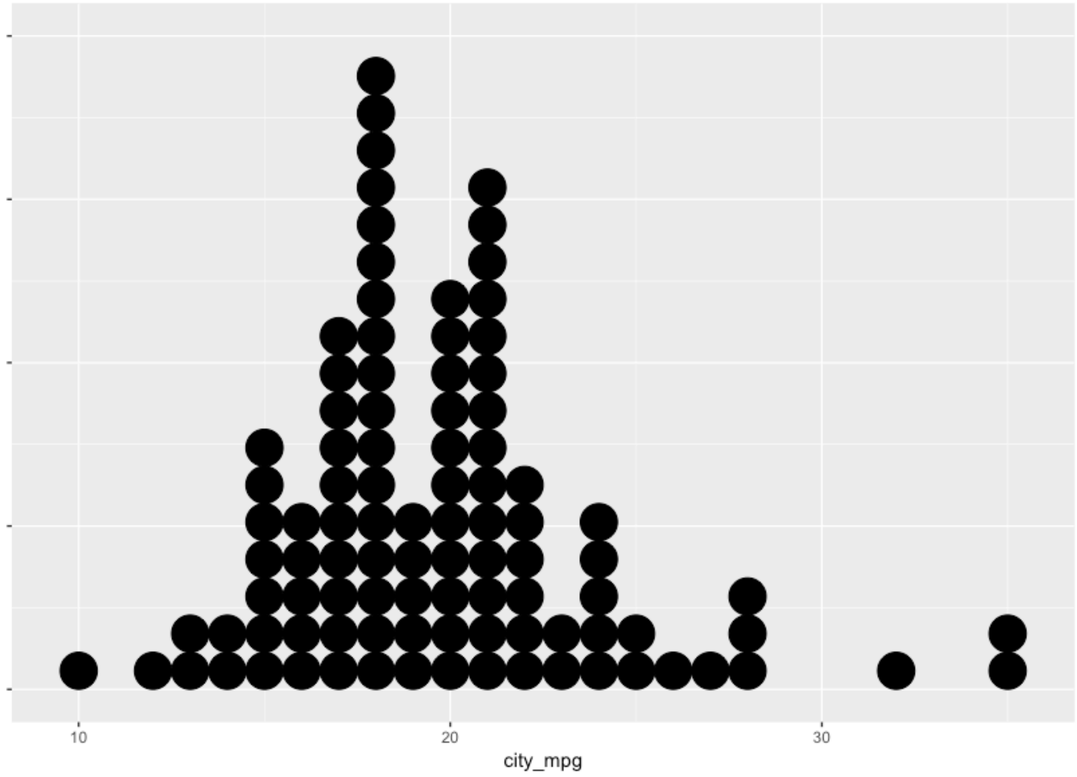


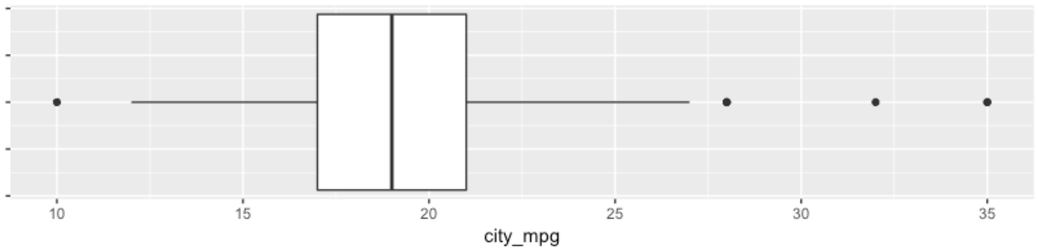
EXPLORATORY DATA ANALYSIS

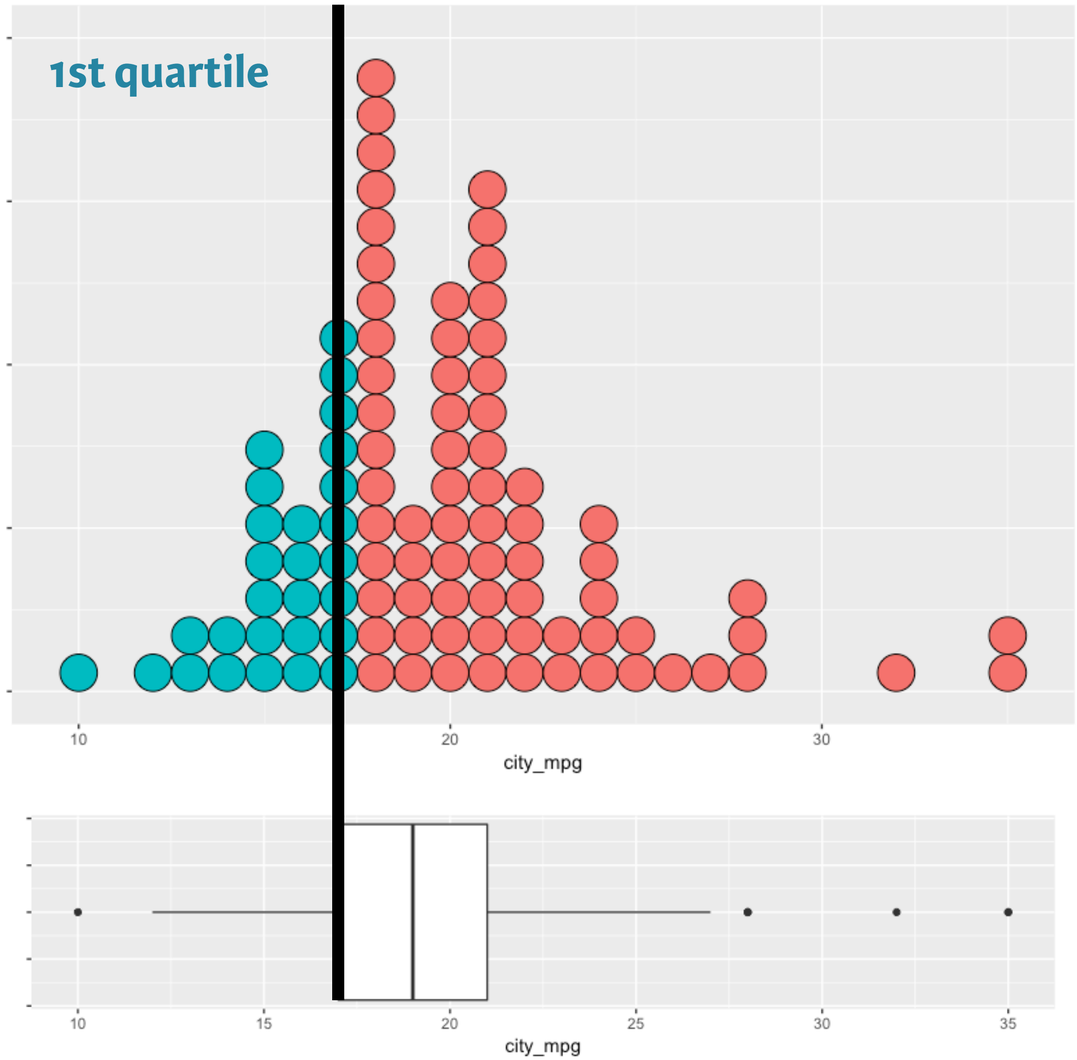
Let’s practice!

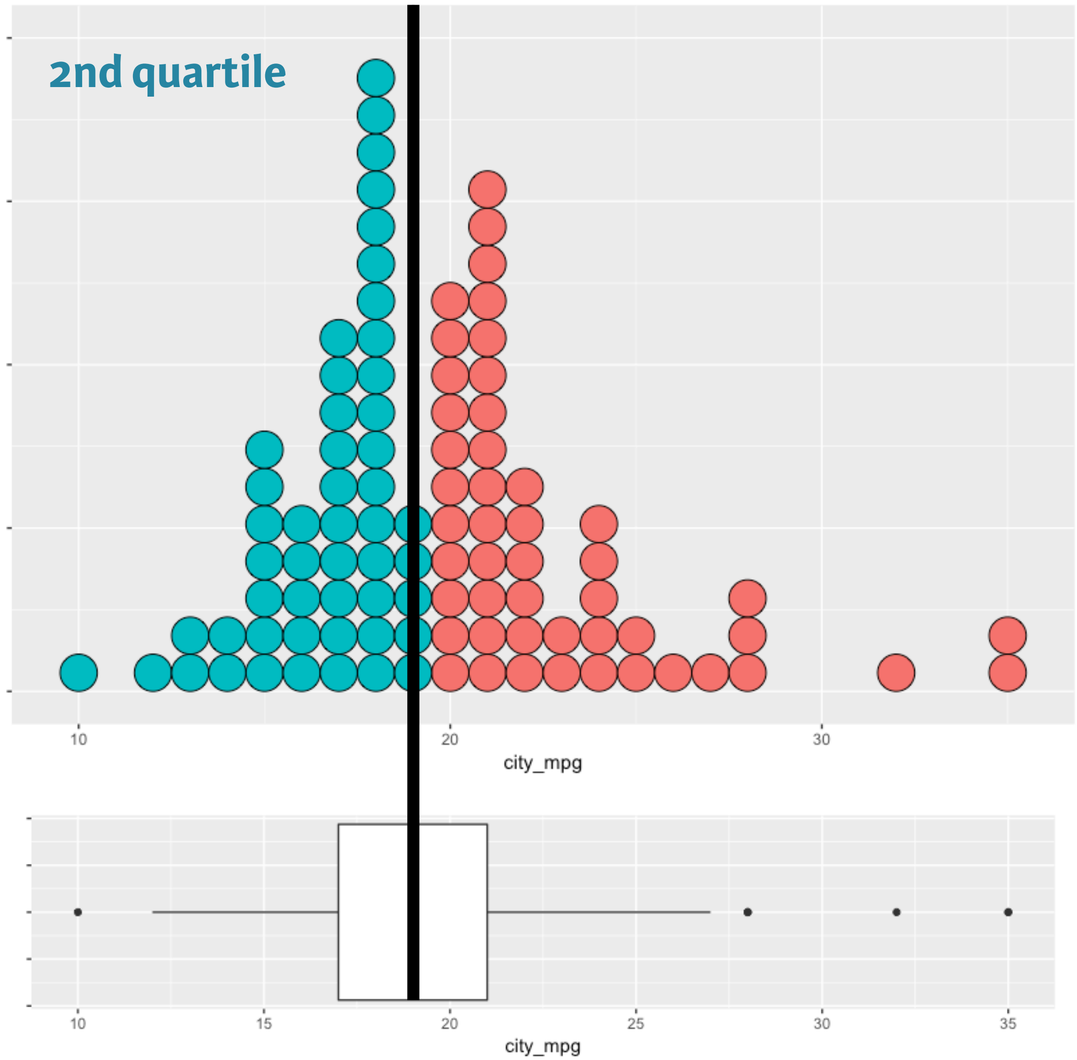
EXPLORATORY DATA ANALYSIS

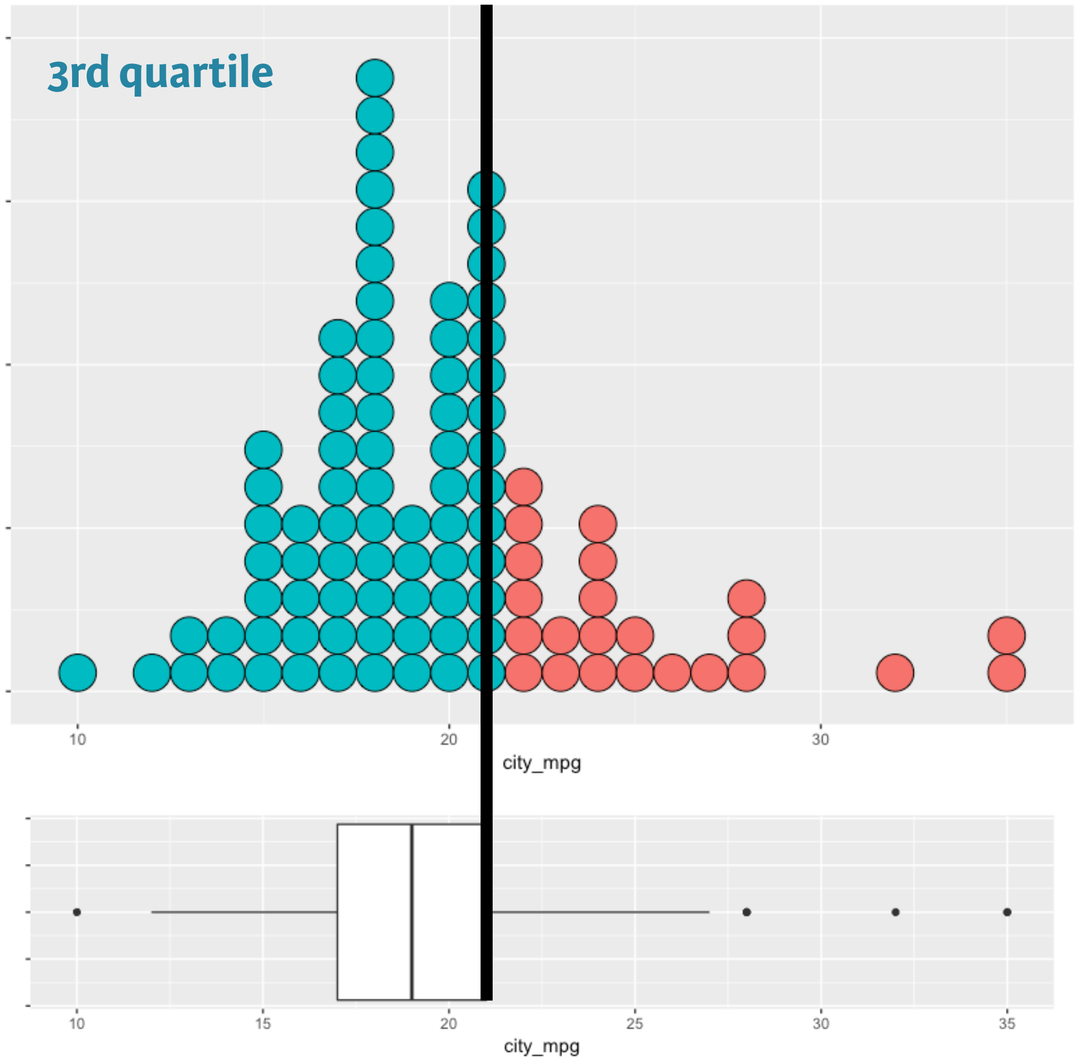
Box plots

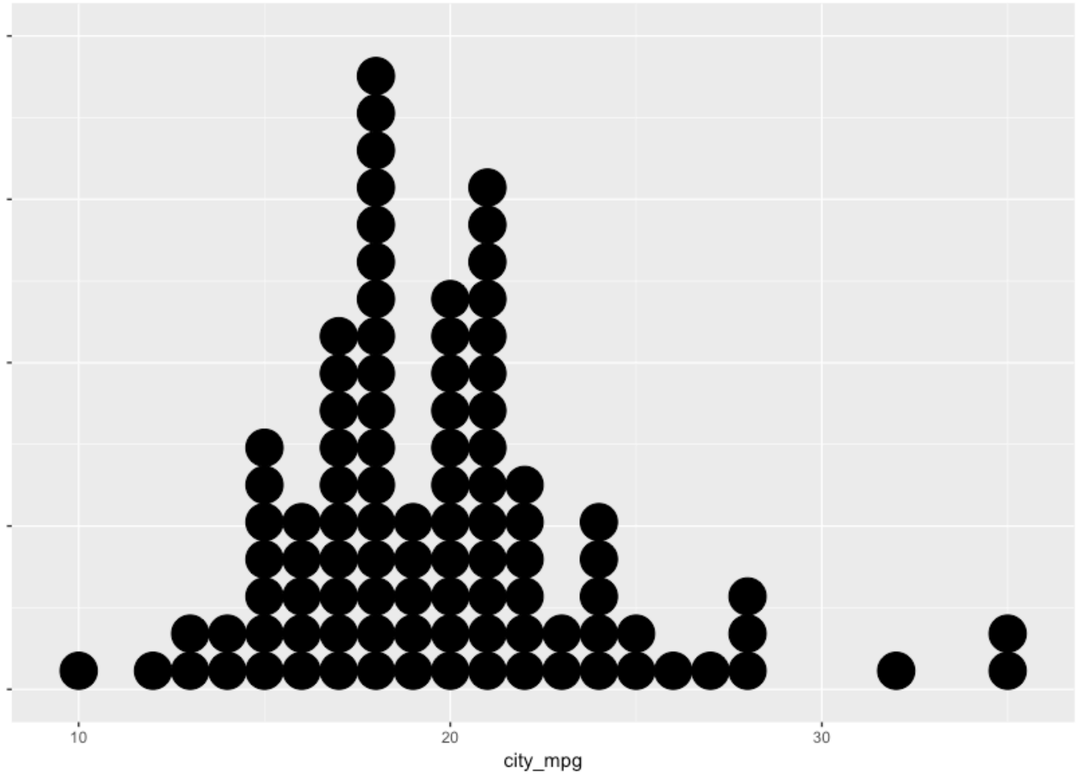

10
20
30
city_mpg

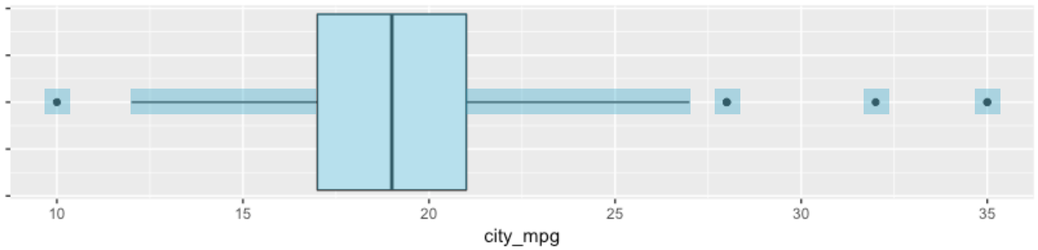

10
15
20
25
30
35
city_mpg


1st quartile
10
20
30
city_mpg
10
15
20
25
30
35
city_mpg


2nd quartile
10
20
30
city_mpg
10
15
20
25
30
35
city_mpg

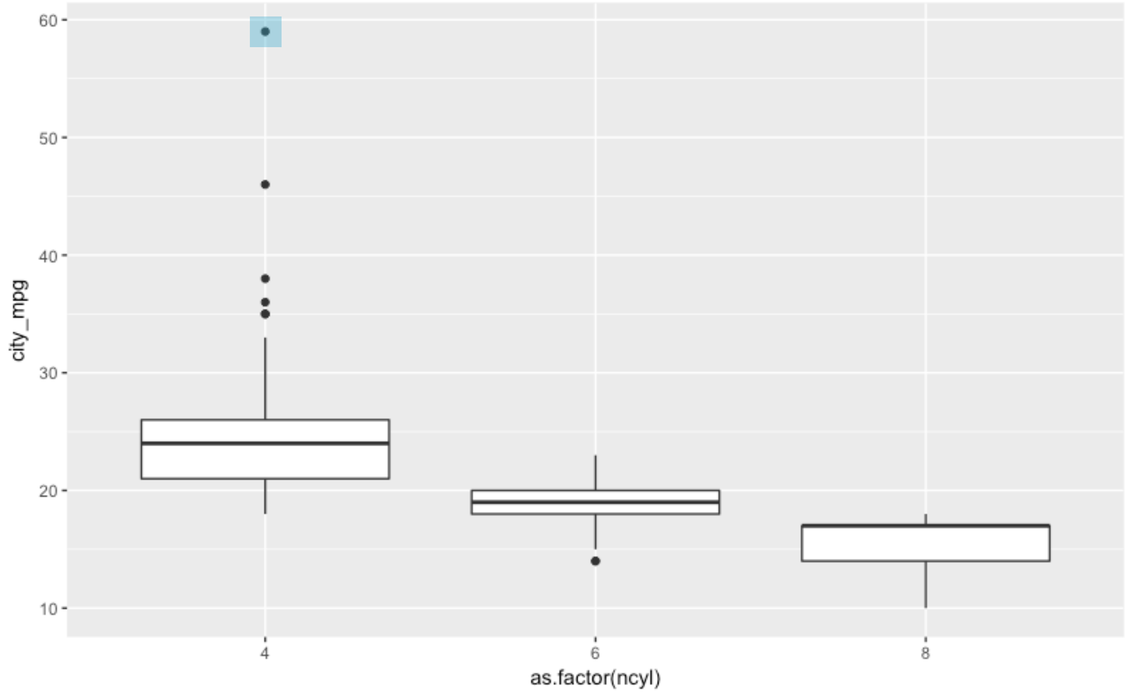

3rd quartile
10
20
30
city_mpg
10
15
20
25
30
35
city_mpg

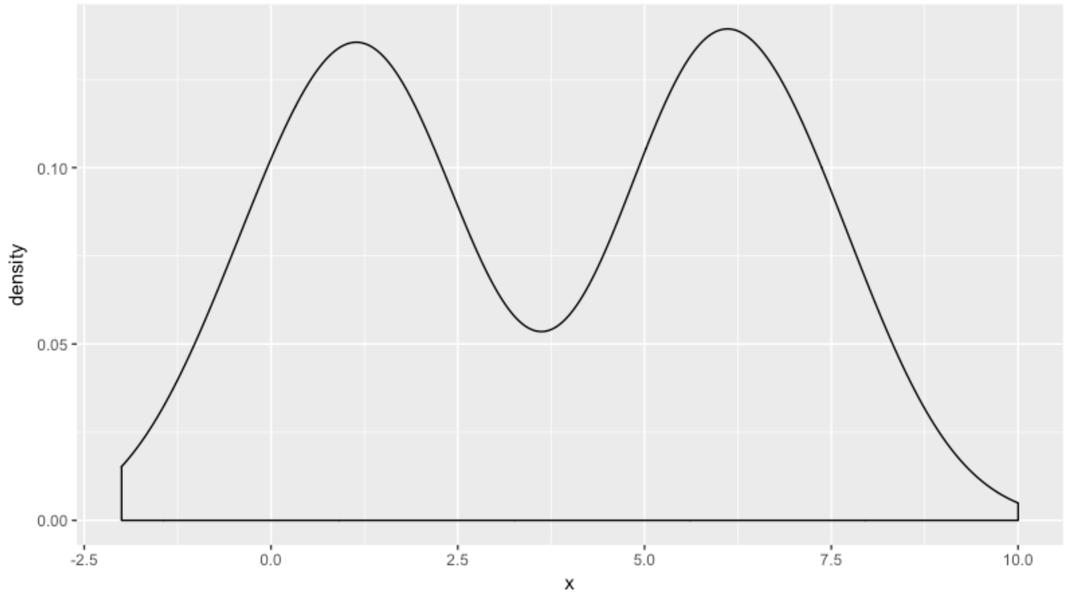

10
20
30
city_mpg

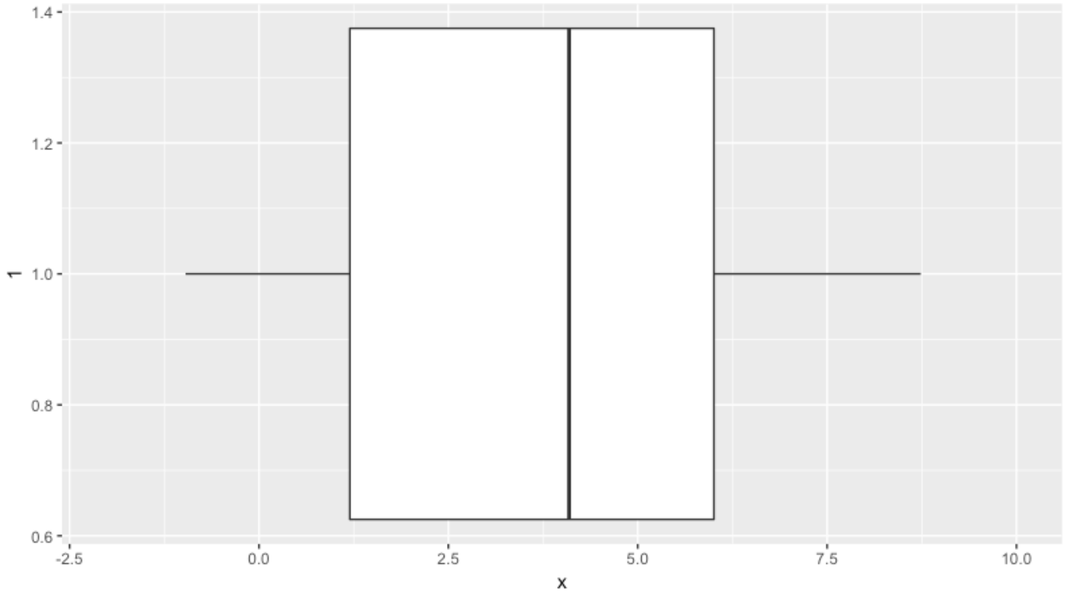

10
15
20
25
30
35
city_mpg

Side-by-side box plots

```
> ggplot(common_cyl, aes(x = as.factor(ncyl), y = city_mpg)) +
    geom_boxplot()
Warning message:
Removed 11 rows containing non-finite values (stat_boxplot).
```




density
0.10
0.05
0.00
-2.5
0.0
2.5
5.0
7.5
10.0
x


1.4
1.2
1.0
0.8
0.6
1
-2.5
0.0
2.5
5.0
7.5
10.0
x

EXPLORATORY DATA ANALYSIS

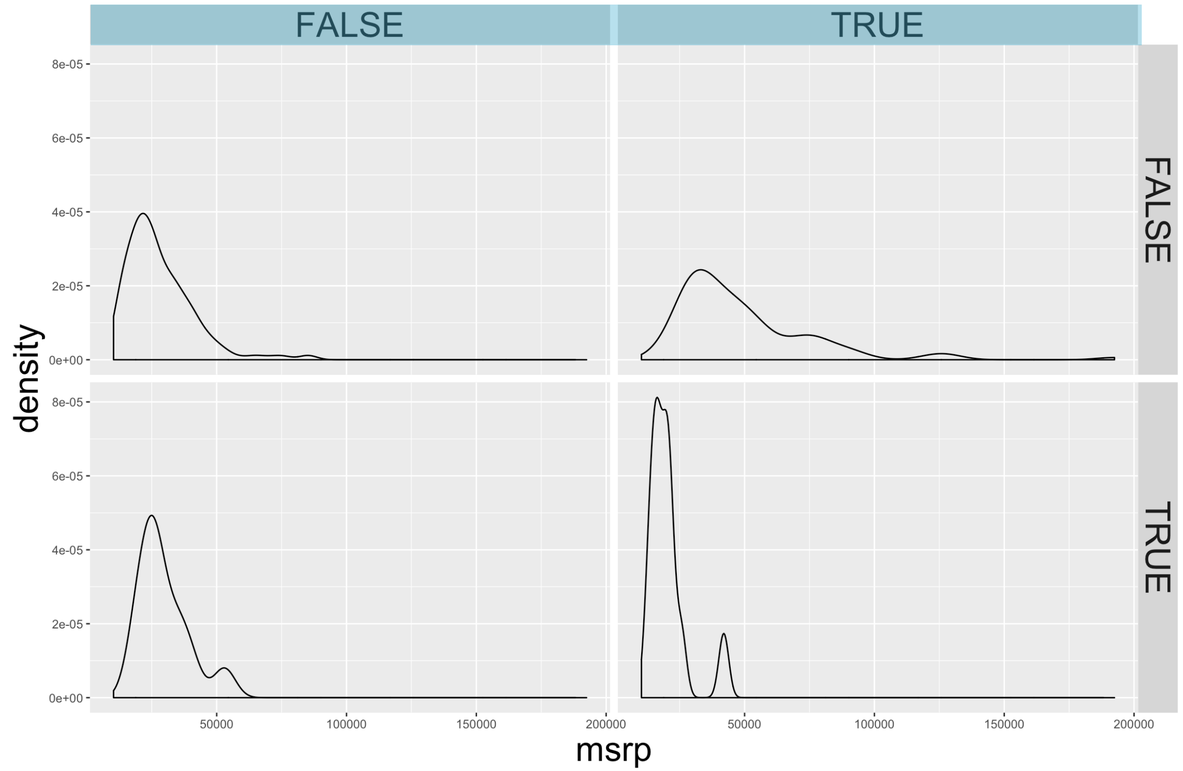
Let’s practice!

EXPLORATORY DATA ANALYSIS

Visualization in higher dimensions

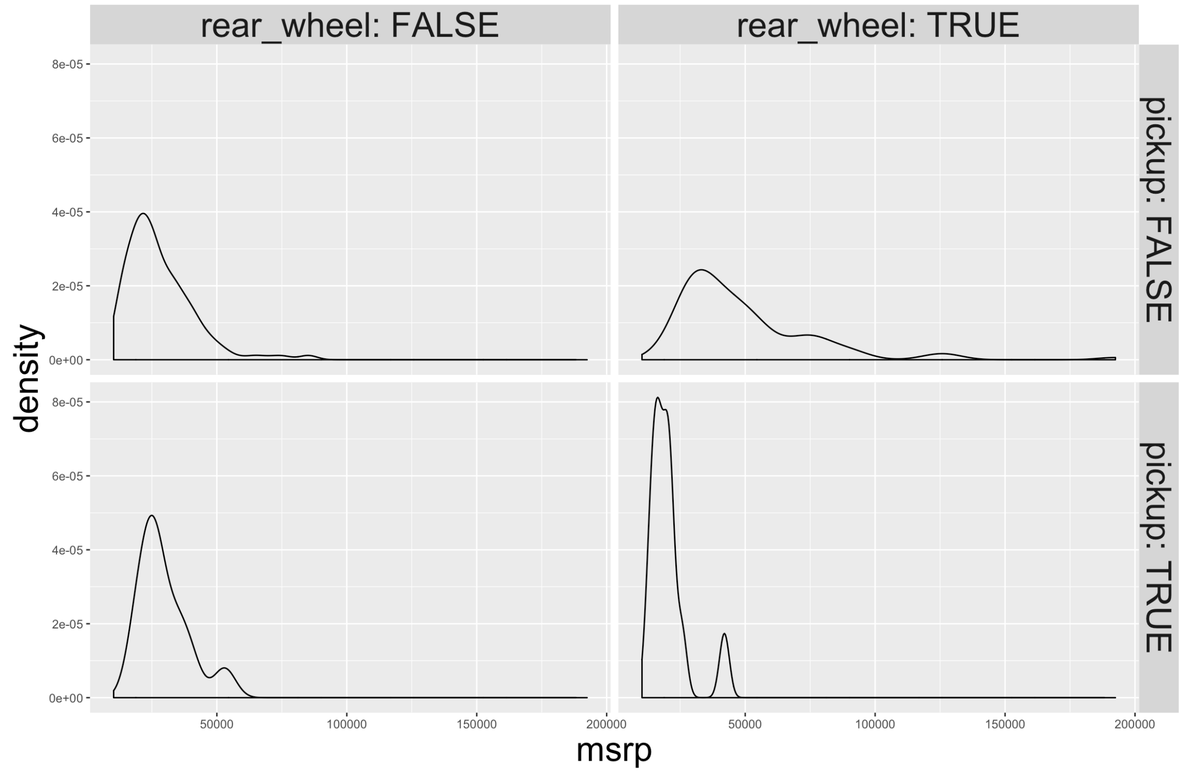
Plots for 3 variables

```
> ggplot(cars, aes(x = msrp)) +
    geom_density() +
    facet_grid(pickup ~ rear_wheel)
```



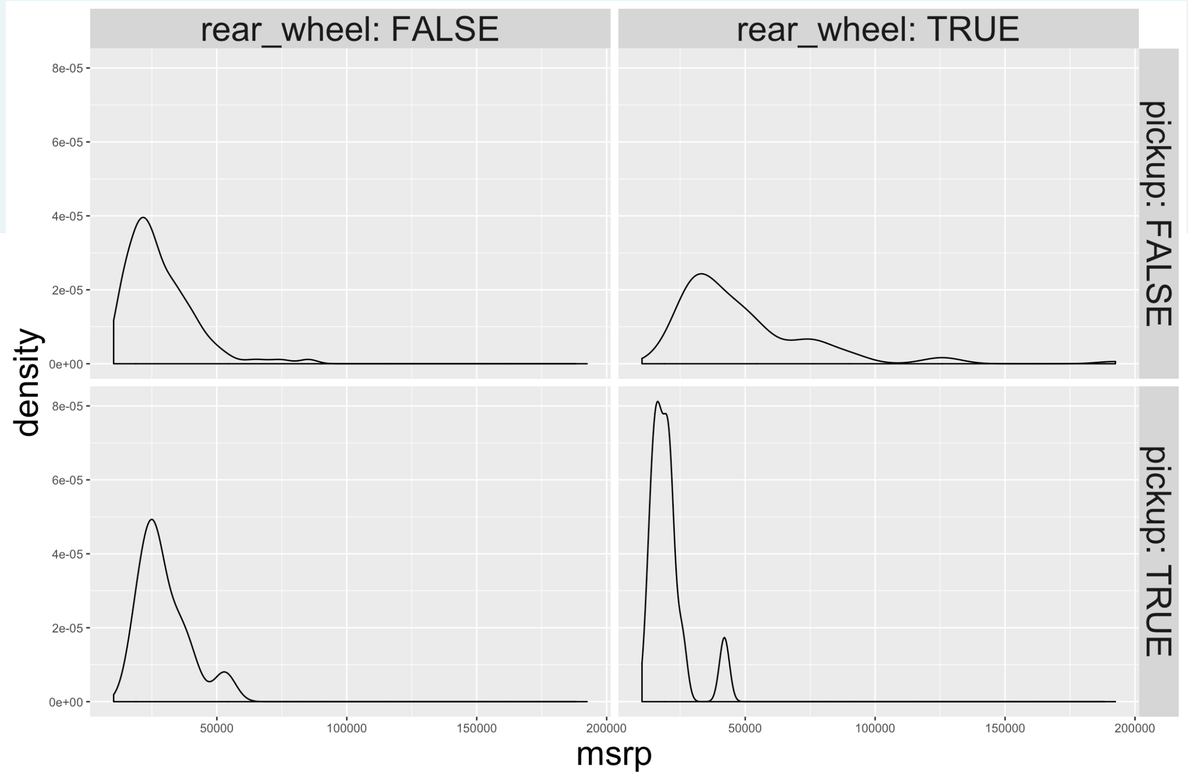
Plots for 3 variables

```
> ggplot(cars, aes(x = msrp)) +
    geom_density() +
    facet_grid(pickup ~ rear_wheel, labeller = label_both)
```



Plots for 3 variables

```
> ggplot(cars, aes(x = msrp)) +
    geom_density() +
    facet_grid(pickup ~ rear_wheel, labeller = label_both)
> table(cars$rear_wheel, cars$pickup)

       FALSE TRUE
  FALSE   306   12
  TRUE     98   12
```



Higher dimensional plots

- Shape
- Size
- Color
- Pattern
- Movement
- x-coordinate
- y-coordinate

EXPLORATORY DATA ANALYSIS

Let’s practice!