EXPLORATORY DATA ANALYSIS

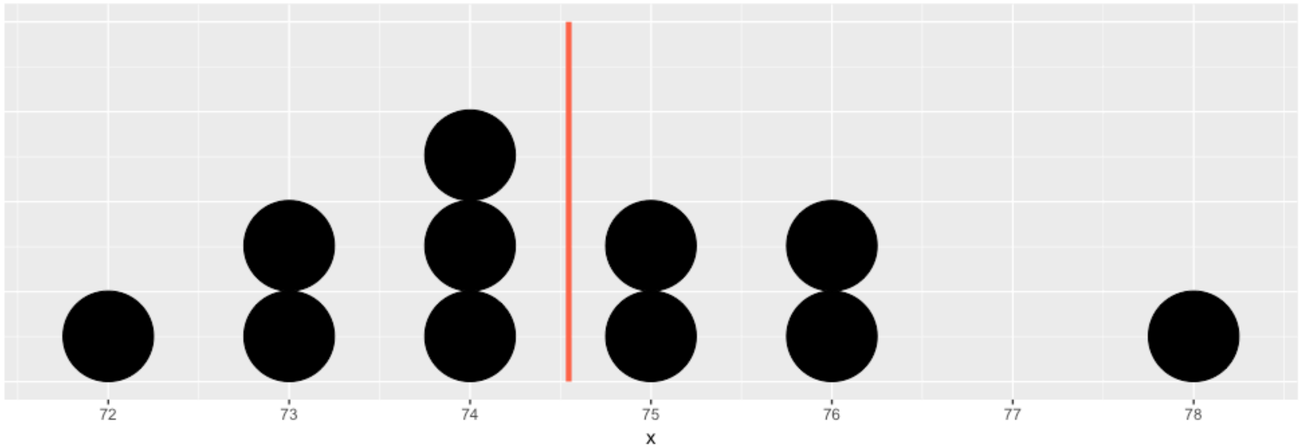
Measures of center

County demographics

```
> life
# A tibble: 3,142 x 4
      state           county expectancy income
      <chr>            <chr>      <dbl>  <int>
1   Alabama  Autauga County     76.060  37773
2   Alabama  Baldwin County     77.630  40121
3   Alabama  Barbour County     74.675  31443
4   Alabama     Bibb County     74.155  29075
5   Alabama   Blount County     75.880  31663
6   Alabama  Bullock County     71.790  25929
7   Alabama   Butler County     73.730  33518
8   Alabama  Calhoun County     73.300  33418
9   Alabama Chambers County     73.245  31282
10 Alabama Cherokee County     74.650  32645
# ... with 3,132 more rows
```

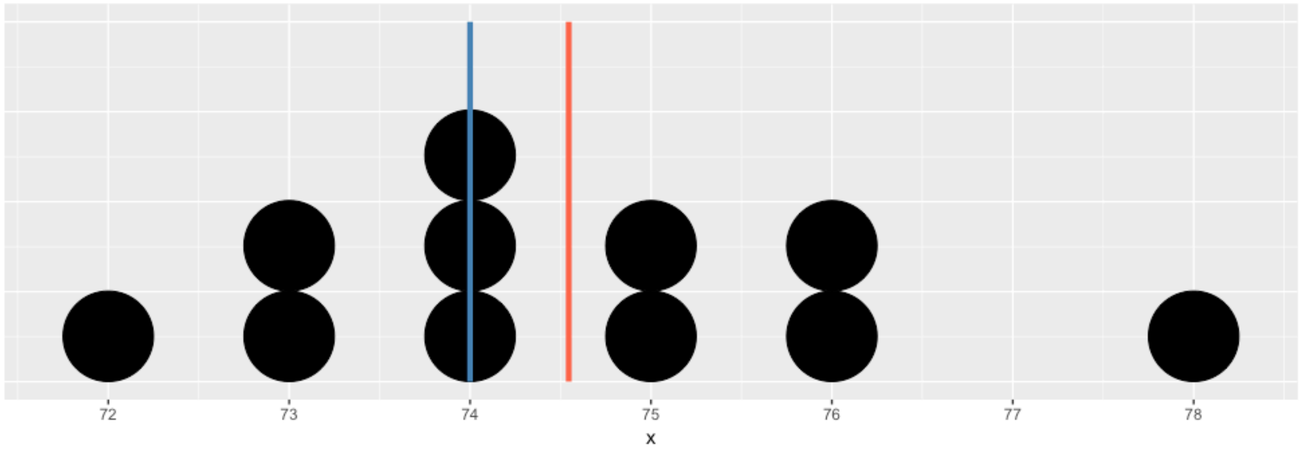
Center: mean

```
> x <- head(round(life$expectancy), 11)
> x
 [1] 76 78 75 74 76 72 74 73 73 75 74
> sum(x)/11
[1] 74.54545
> mean(x)
[1] 74.54545
```



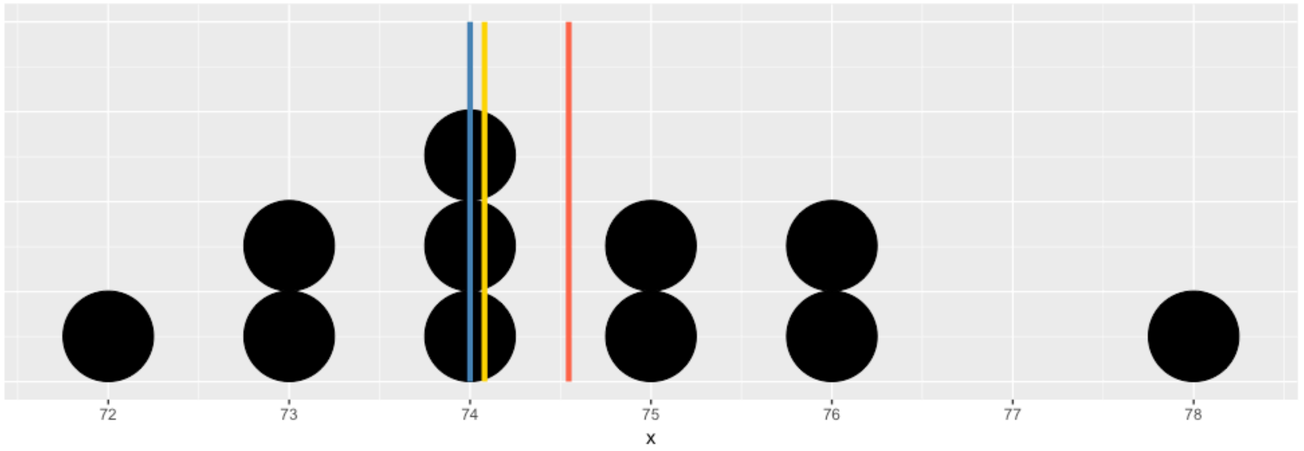
Center: mean, median

```
> x
 [1] 76 78 75 74 76 72 74 73 73 75 74
> sort(x)
 [1] 72 73 73 74 74 74 75 75 76 76 78
> median(x)
[1] 74
```



Center: mean, median, mode

```
> x
 [1] 76 78 75 74 76 72 74 73 73 75 74
> table(x)
x
72 73 74 75 76 78
 1  2  3  2  2  1
```



Groupwise means

```
> life <- life %>%
    mutate(west_coast = state %in% c("California", "Oregon", "Washington"))

> life %>%
    group_by(west_coast) %>%
    summarize(mean(expectancy),
              median(expectancy))
# A tibble: 2 x 3
  west_coast mean(expectancy) median(expectancy)
       <lgl>            <dbl>              <dbl>
1      FALSE         77.12750              77.31
2       TRUE         78.90545              78.65
```

West coast counties

Without `group_by()`

```
> life %>%
    slice(240:247) %>%
    summarize(mean(expectancy))
# A tibble: 1 x 1
  mean(expectancy)
             <dbl>
1          79.2775
```

state	county	expectancy	income	west_coast
California	Tuolumne	79.6	41770	TRUE
California	Ventura	81.1	54155	TRUE
California	Yolo	80.0	49063	TRUE
California	Yuba	76.3	37535	TRUE
Colorado	Adams	80.1	36962	FALSE
Colorado	Alamosa	77.4	34088	FALSE
Colorado	Arapahoe	80.3	52545	FALSE
Colorado	Archuleta	79.1	40307	FALSE

With group_by()

```
> life %>%
    slice(240:247) %>%
    group_by(west_coast) %>%
    summarize(mean(expectancy))
# A tibble: 2 x 2
  west_coast mean(expectancy)
       <lgl>            <dbl>
1      FALSE         79.26125
2       TRUE         79.29375
```

state	county	expectancy	income	west_coast
California	Tuolumne	79.6	41770	TRUE
California	Ventura	81.1	54155	TRUE
California	Yolo	80.0	49063	TRUE
California	Yuba	76.3	37535	TRUE
Colorado	Adams	80.1	36962	FALSE
Colorado	Alamosa	77.4	34088	FALSE
Colorado	Arapahoe	80.3	52545	FALSE
Colorado	Archuleta	79.1	40307	FALSE

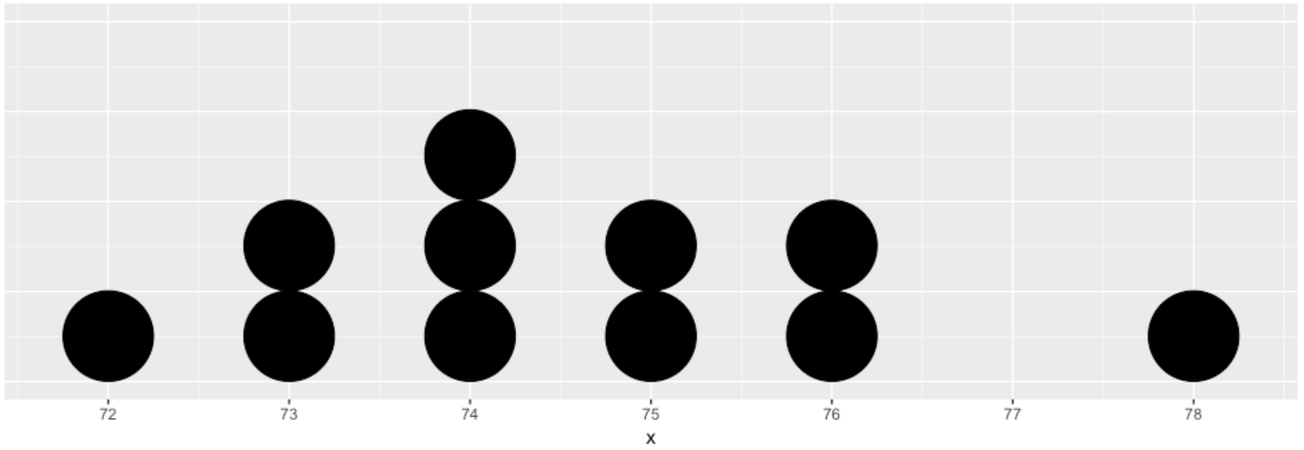
EXPLORATORY DATA ANALYSIS

Let’s practice!

EXPLORATORY DATA ANALYSIS

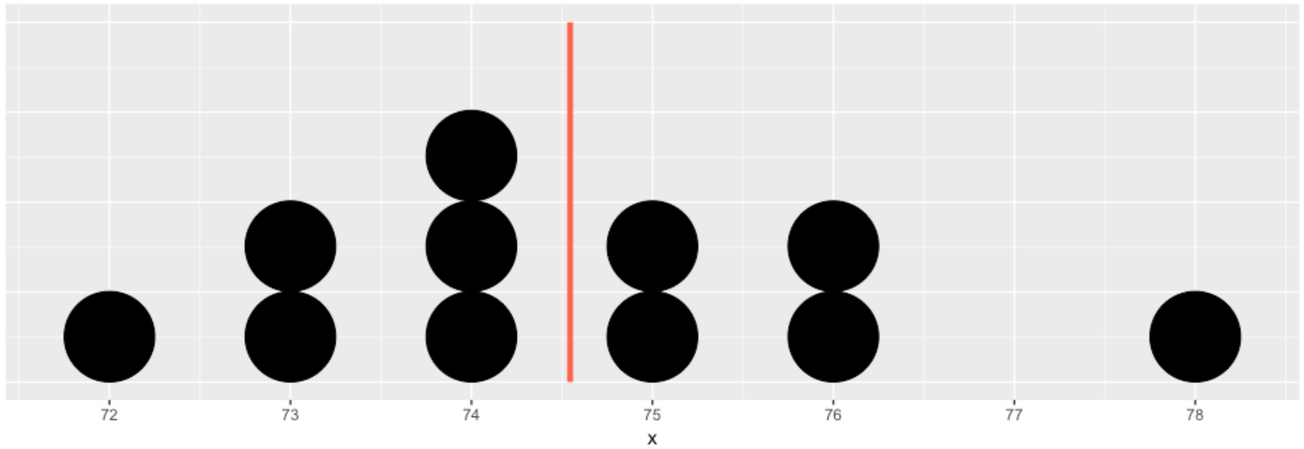
Measures of variability

```
> x
 [1] 76 78 75 74 76 72 74 73 73 75 74
```



```
> x
 [1] 76 78 75 74 76 72 74 73 73 75 74
```

```
> x - mean(x)
 [1]  1.4545  3.4545  0.4545 -0.5455  1.4545 -2.5455
 [7] -0.5455 -1.5455 -1.5455  0.4545 -0.5455
> sum(x - mean(x))
[1] -1.421085e-14
> sum((x - mean(x))^2)
[1] 28.72727
> n <- 11
> sum((x - mean(x))^2)/n
[1] 2.61157
> sum((x - mean(x))^2)/(n - 1)
[1] 2.872727
> var(x)
[1] 2.872727
```



```
> x
 [1] 76 78 75 74 76 72 74 73 73 75 74
```

```
> sd(x) # Standard deviation
[1] 1.694912  years

> var(x) # Variance
[1] 2.872727  years squared

> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  72.00   73.50   74.00   74.55   75.50   78.00

> IQR(x) # Interquartile range
[1] 2

> diff(range(x)) # Range
[1] 6
```

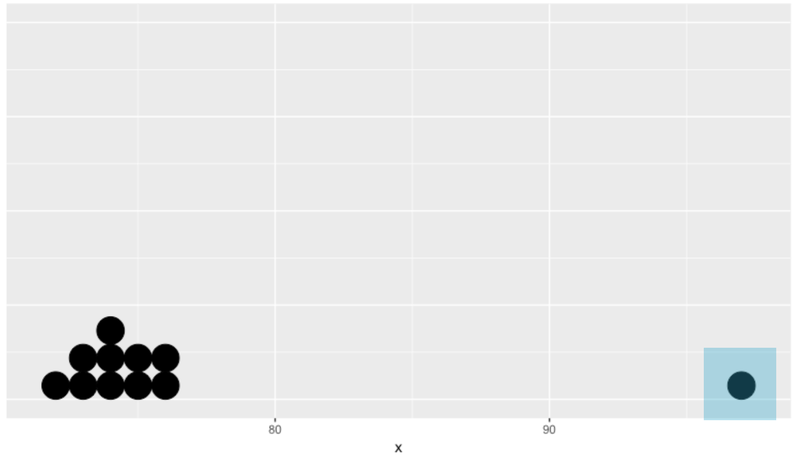
```
> x
 [1] 76 78 75 74 76 72 74 73 73 75 74
> x_new
 [1] 76 97 75 74 76 72 74 73 73 75 74
```

```
> sd(x_new) # Was 1.69
[1] 6.987001

> var(x_new) # Was 2.87
[1] 48.81818

> diff(range(x_new)) # Was 6
[1] 25

> IQR(x_new) # Doesn't change
[1] 2
```



EXPLORATORY DATA ANALYSIS

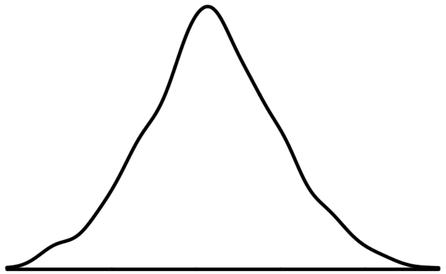
Let’s practice!

EXPLORATORY DATA ANALYSIS

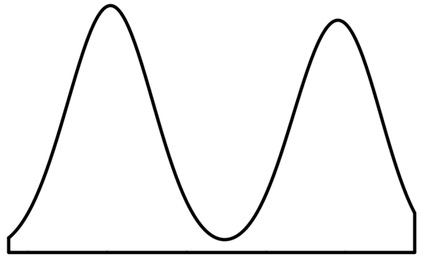
Shape and transformations

Modality

Unimodal

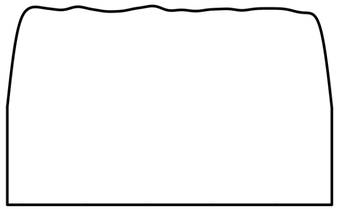


Bimodal



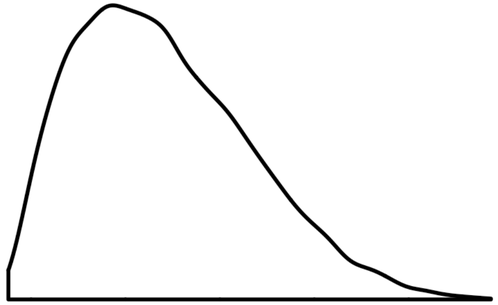
Multimodal

Uniform

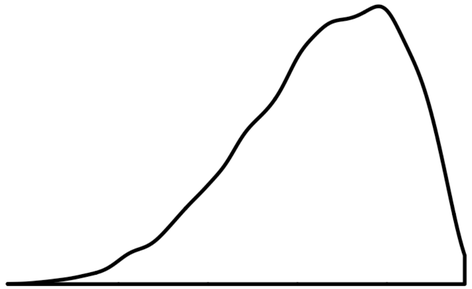


Skew

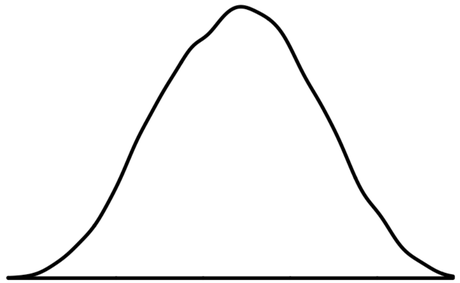
`Right-skewed`



`Left-skewed`

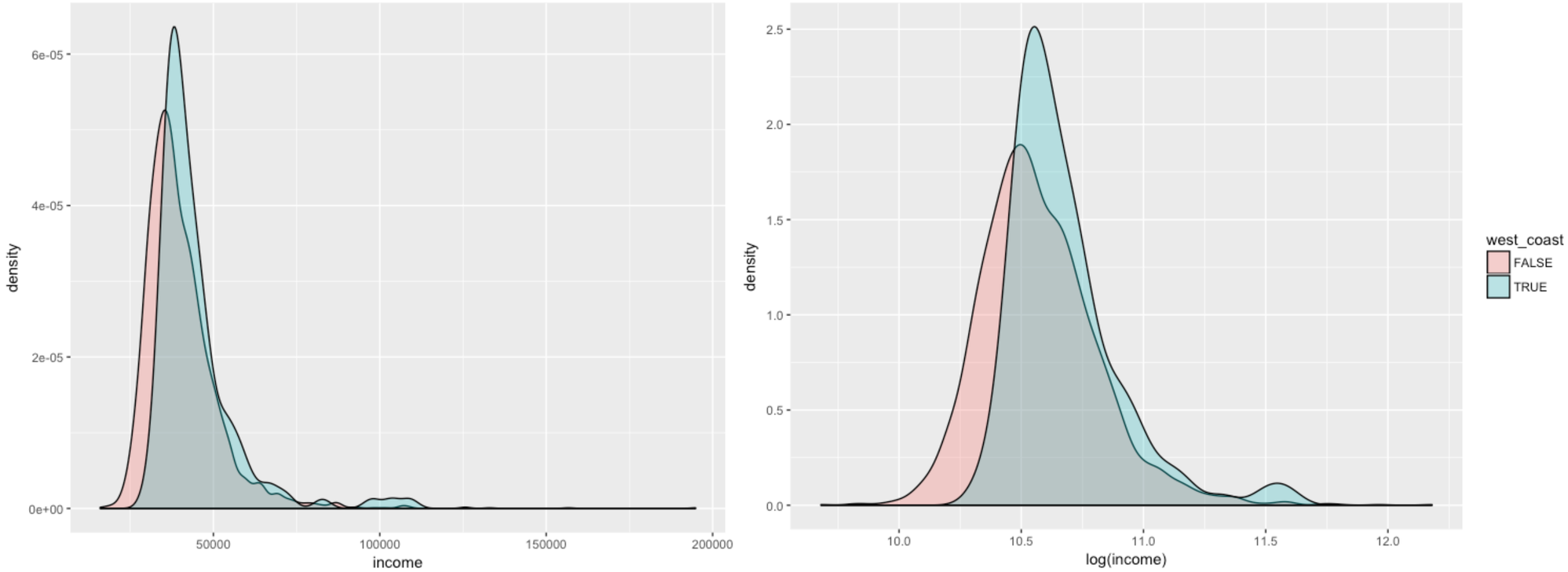


`Symmetric`



Shape of income

```
> ggplot(life, aes(x = income, fill = west_coast)) +
    geom_density(alpha = .3)
> ggplot(life, aes(x = log(income), fill = west_coast)) +
    geom_density(alpha = .3)
```



EXPLORATORY DATA ANALYSIS

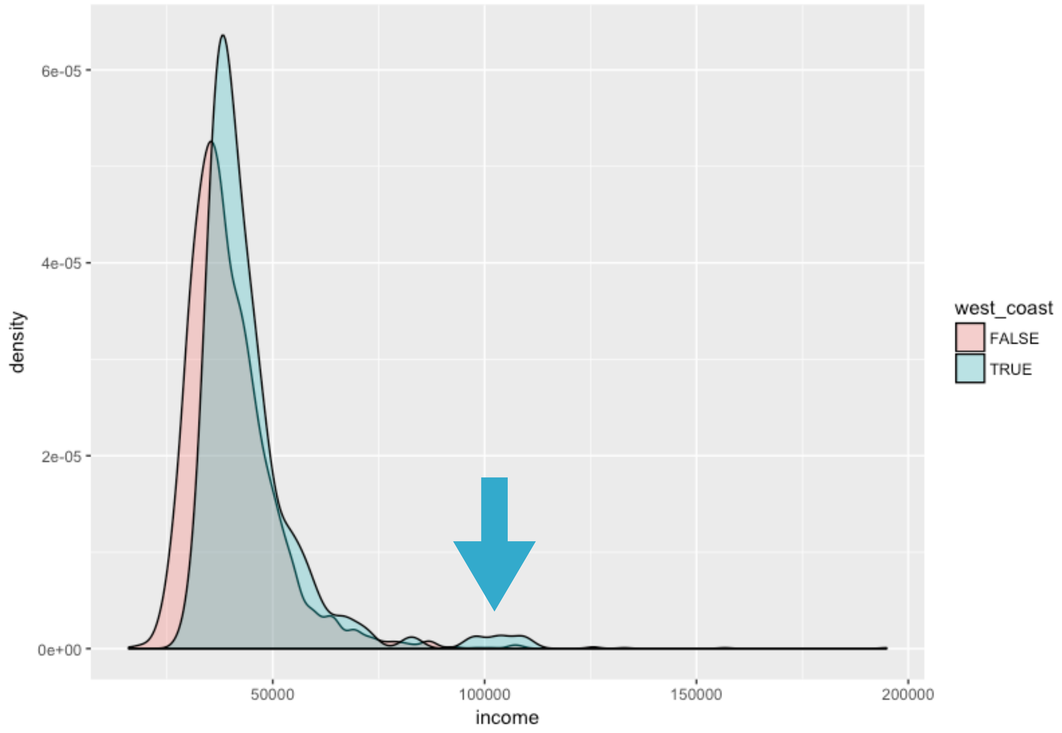
Let’s practice!

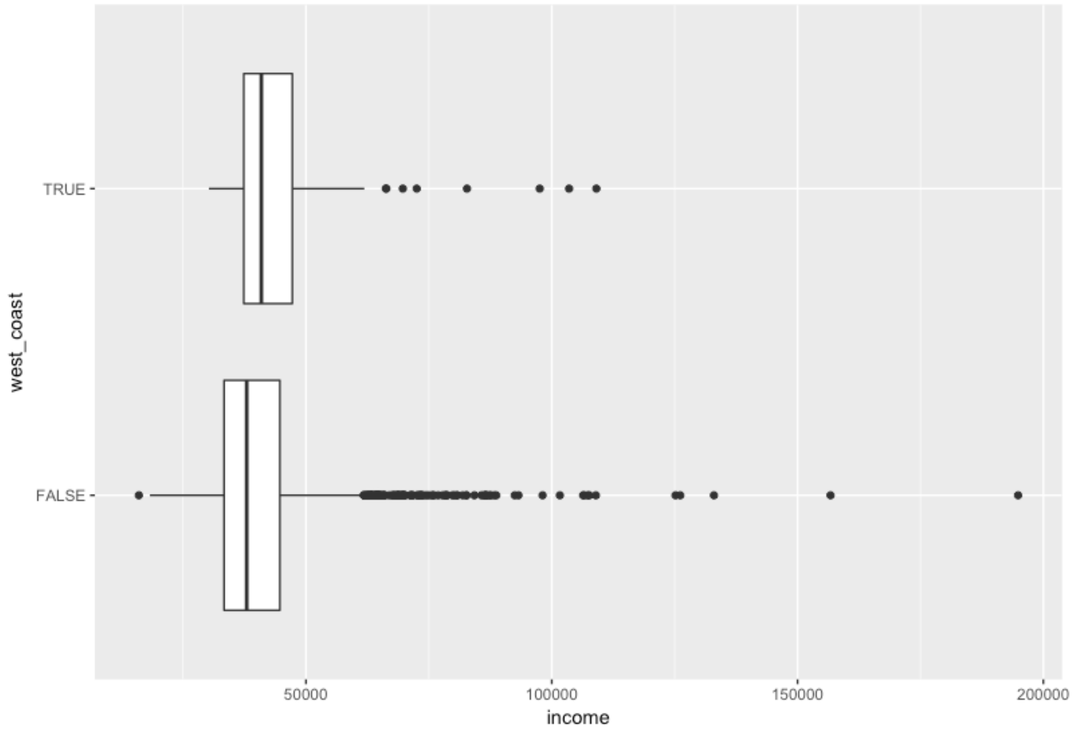
EXPLORATORY DATA ANALYSIS

Outliers

Characteristics of a distribution

- Center
- Variability
- Shape
- Outliers


6e-05
4e-05
2e-05
0e+00
density
50000
100000
150000
200000
income
west_coast
FALSE
TRUE

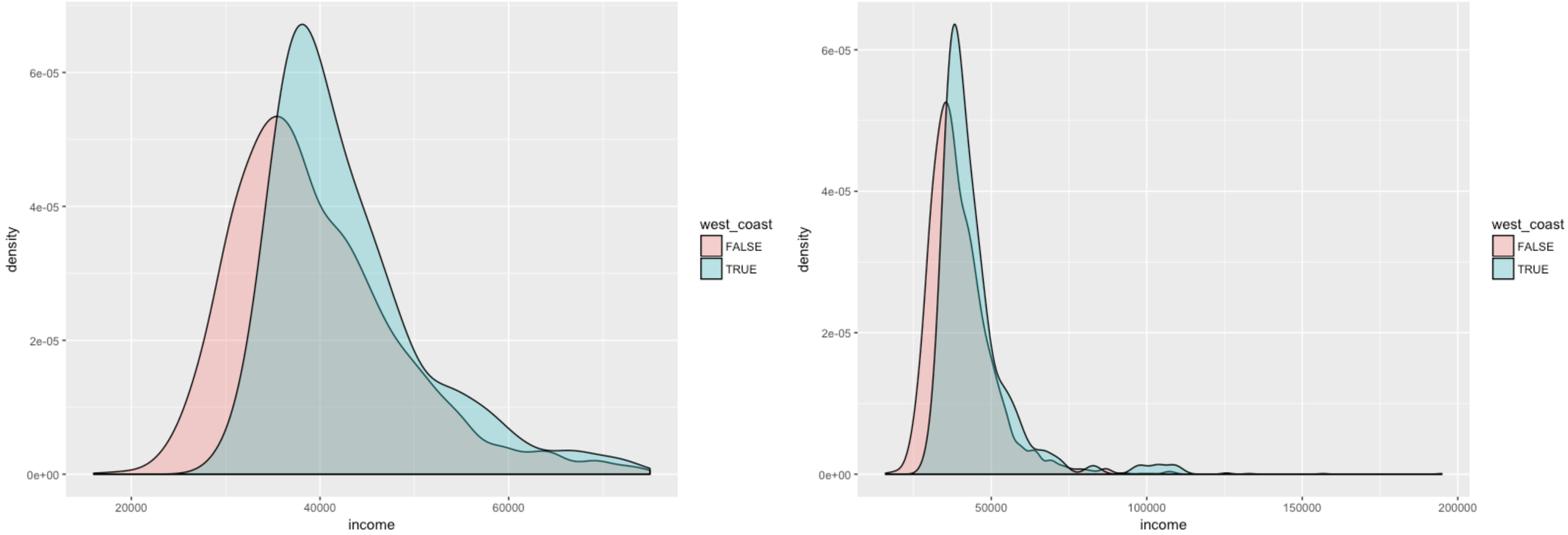

TRUE
FALSE
west_coast
50000
100000
150000
200000
income

Indicating outliers

```
> life <- life %>%
    mutate(is_outlier = income > 75000)
> life %>%
    filter(is_outlier) %>%
    arrange(desc(income))
# A tibble: 45 x 6
           state              county expectancy income west_coast is_outlier
           <chr>               <chr>      <dbl>  <int>       <lgl>      <lgl>
1        Wyoming        Teton County     82.110 194861       FALSE       TRUE
2       New York    New York County     81.675 156708       FALSE       TRUE
3          Texas Shackelford County     75.400 132989       FALSE       TRUE
4       Colorado      Pitkin County     82.990 126137       FALSE       TRUE
5       Nebraska     Wheeler County     79.180 125171       FALSE       TRUE
6     California       Marin County     83.230 109076        TRUE       TRUE
7       Nebraska     Kearney County     79.630 108975       FALSE       TRUE
8          Texas    McMullen County     77.320 107627       FALSE       TRUE
9  Massachusetts   Nantucket County     80.325 107341       FALSE       TRUE
10         Texas     Midland County     77.830 106588       FALSE       TRUE
# ... with 35 more rows
```

Plotting without outliers

```
> life %>%
    filter(!is_outlier) %>%
    ggplot(aes(x = income, fill = west_coast)) +
    geom_density(alpha = .3)
```



EXPLORATORY DATA ANALYSIS

Let’s practice!