Missing Data Workflows: The Shadow matrix and Nabular data

DEALING WITH MISSING DATA IN R

Nicholas Tierney

Statistician

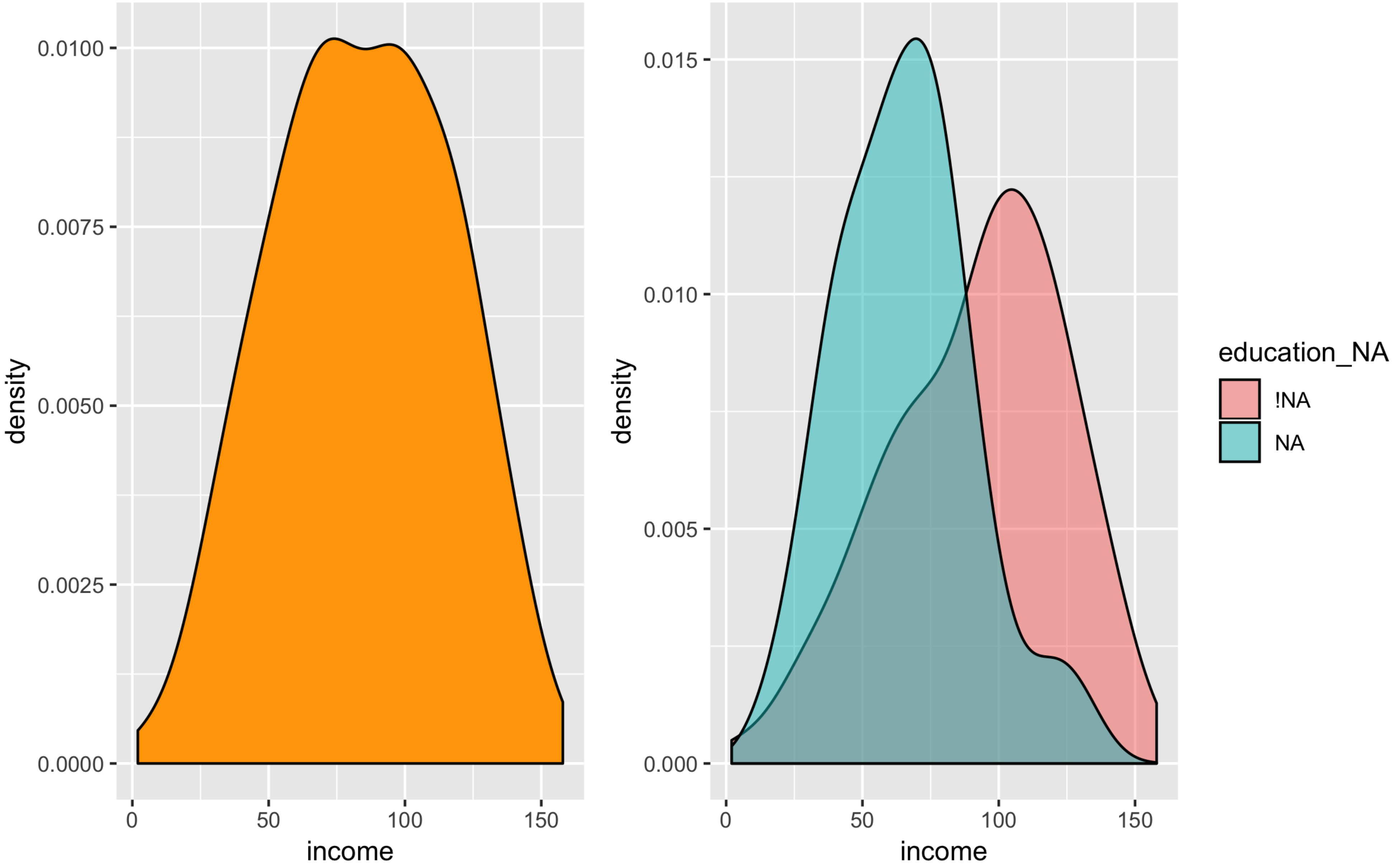
An example

Census data containing:

- Income
- Education

income	education
48.69087	NA
40.93218	NA
52.69245	high_school
31.33808	NA
89.35671	university
103.87278	university

What we are going to cover



The shadow matrix

name	height	age
Sophie	174	NA
NA	185	26
Dan	NA	42

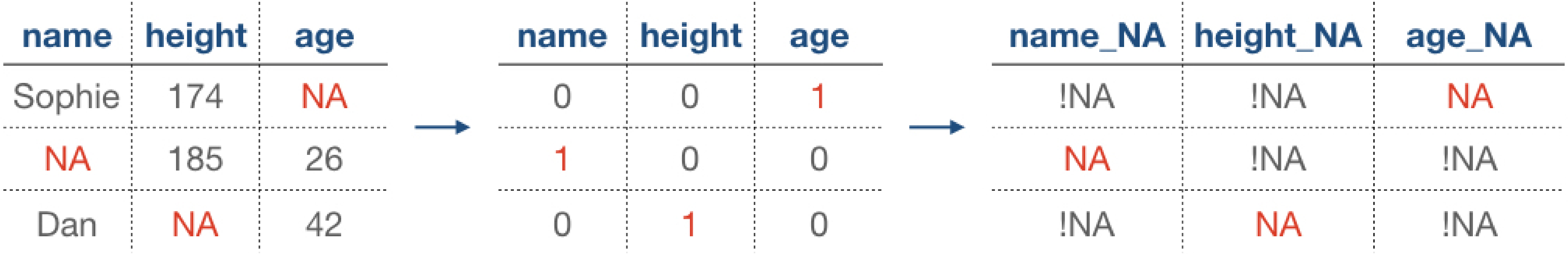
→

name	height	age
0	0	1
1	0	0
0	1	0

→

name_NA	height_NA	age_NA
!NA	!NA	NA
NA	!NA	!NA
!NA	NA	!NA

The shadow matrix



name	height	age
Sophie	174	NA
NA	185	26
Dan	NA	42

name	height	age
0	0	1
1	0	0
0	1	0

name_NA	height_NA	age_NA
!NA	!NA	NA
NA	!NA	!NA
!NA	NA	!NA

Two main features

1. Coordinated names
2. Clear values

Creating nabular data

income	education	income_NA	education_NA
48.69087	NA	!NA	NA
40.93218	NA	!NA	NA
52.69245	high_school	!NA	!NA
31.33808	NA	!NA	NA
89.35671	university	!NA	!NA
103.87278	university	!NA	!NA

Using nabular data to perform summaries

```
bind_shadow(airquality)
```

```
# A tibble: 153 x 12
   Ozone Solar.R  Wind  Temp Month   Day Ozone_NA Solar.R_NA Wind_NA Temp_NA
   <int>   <int> <dbl> <int> <int> <int> <fct>    <fct>      <fct>   <fct>
 1    41     190   7.4    67     5     1 !NA      !NA        !NA     !NA
 2    36     118   8      72     5     2 !NA      !NA        !NA     !NA
 3    12     149  12.6    74     5     3 !NA      !NA        !NA     !NA
 4    18     313  11.5    62     5     4 !NA      !NA        !NA     !NA
 5    NA      NA  14.3    56     5     5 NA       NA         !NA     !NA
 6    28      NA  14.9    66     5     6 !NA      NA         !NA     !NA
 7    23     299   8.6    65     5     7 !NA      !NA        !NA     !NA
 8    19      99  13.8    59     5     8 !NA      !NA        !NA     !NA
 9     8      19  20.1    61     5     9 !NA      !NA        !NA     !NA
10    NA     194   8.6    69     5    10 NA       !NA        !NA     !NA
# ... with 143 more rows, and 2 more variables: Month_NA <fct>, Day_NA <fct>
```

Using nabular data to perform summaries

```
airquality %>%
  bind_shadow() %>%
  group_by(Ozone_NA) %>%
  summarize(mean = mean(Wind))
```

Ozone_NA	mean
!NA	9.862069
NA	10.256757

Let's practice!

Exploring conditional missings with ggplot

DEALING WITH MISSING DATA IN R

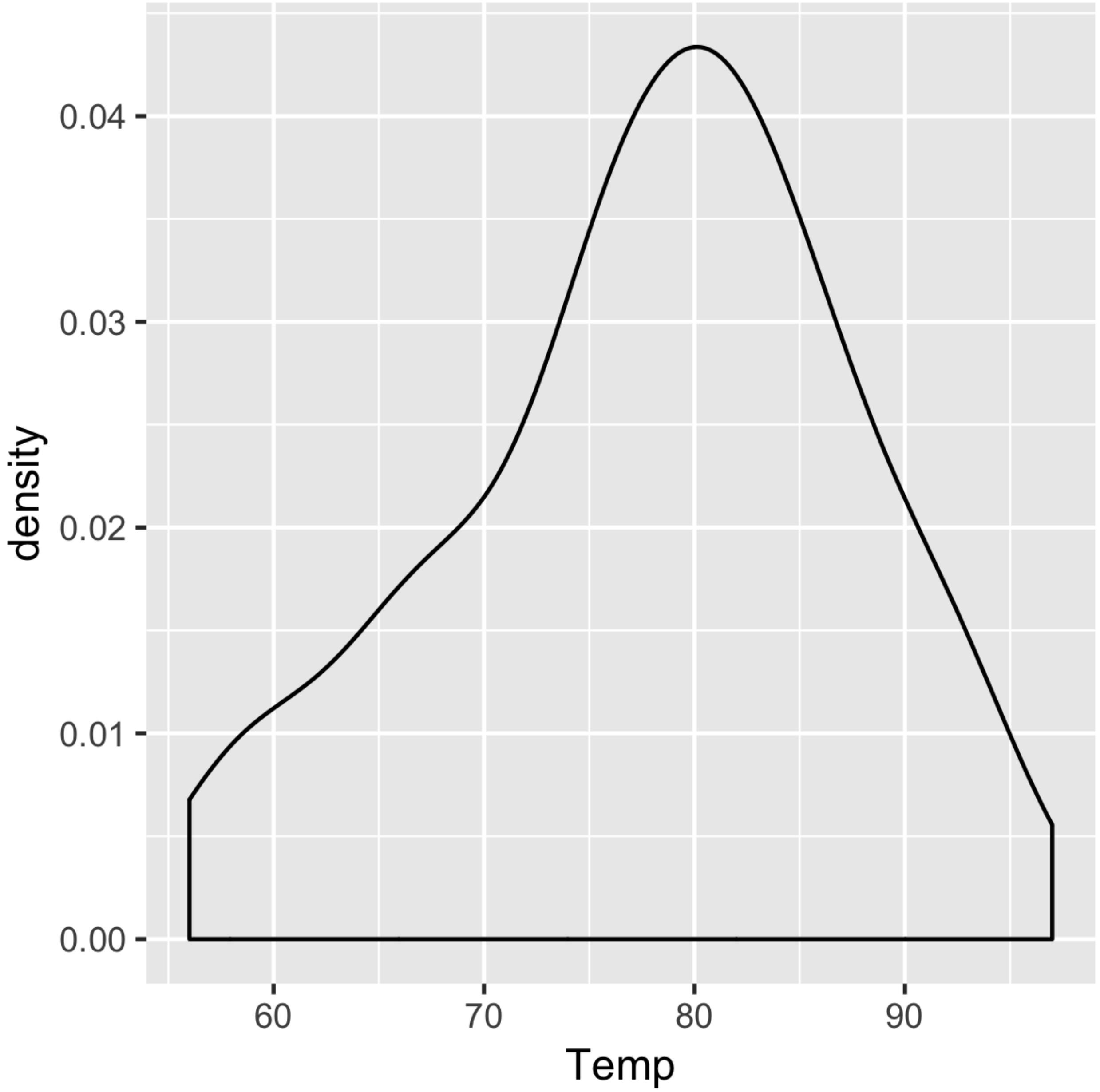
Nicholas Tierney
Statistician

What we are going to cover

- How to use nabular data to explore how values change according to other values going missing
- Explore visualizations:
  - densities
  - box plots
  - different methods of splitting the visualization

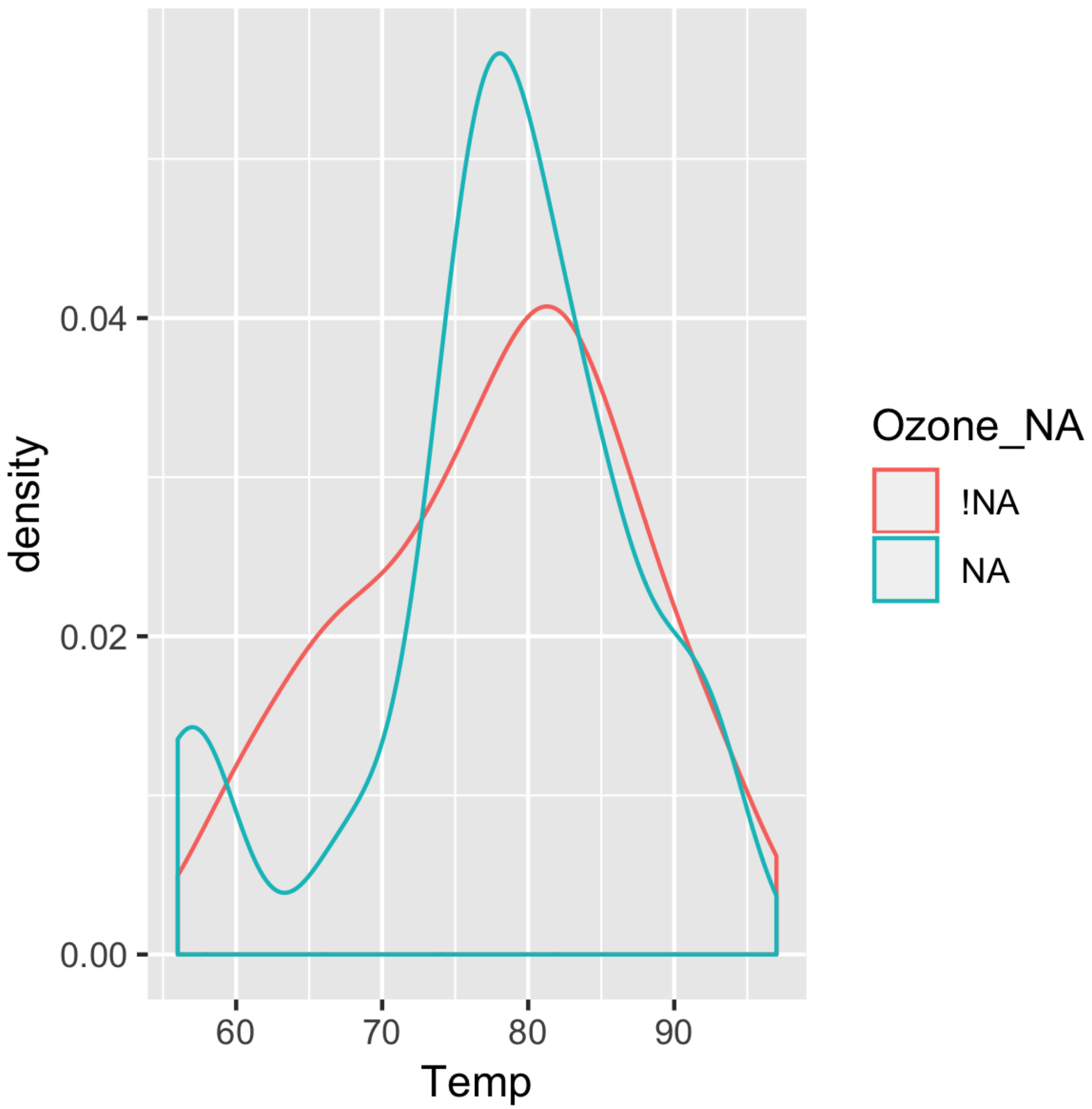
Visualizing missings using densities

```
ggplot(airquality,
       aes(x = Temp)) +
  geom_density()
```



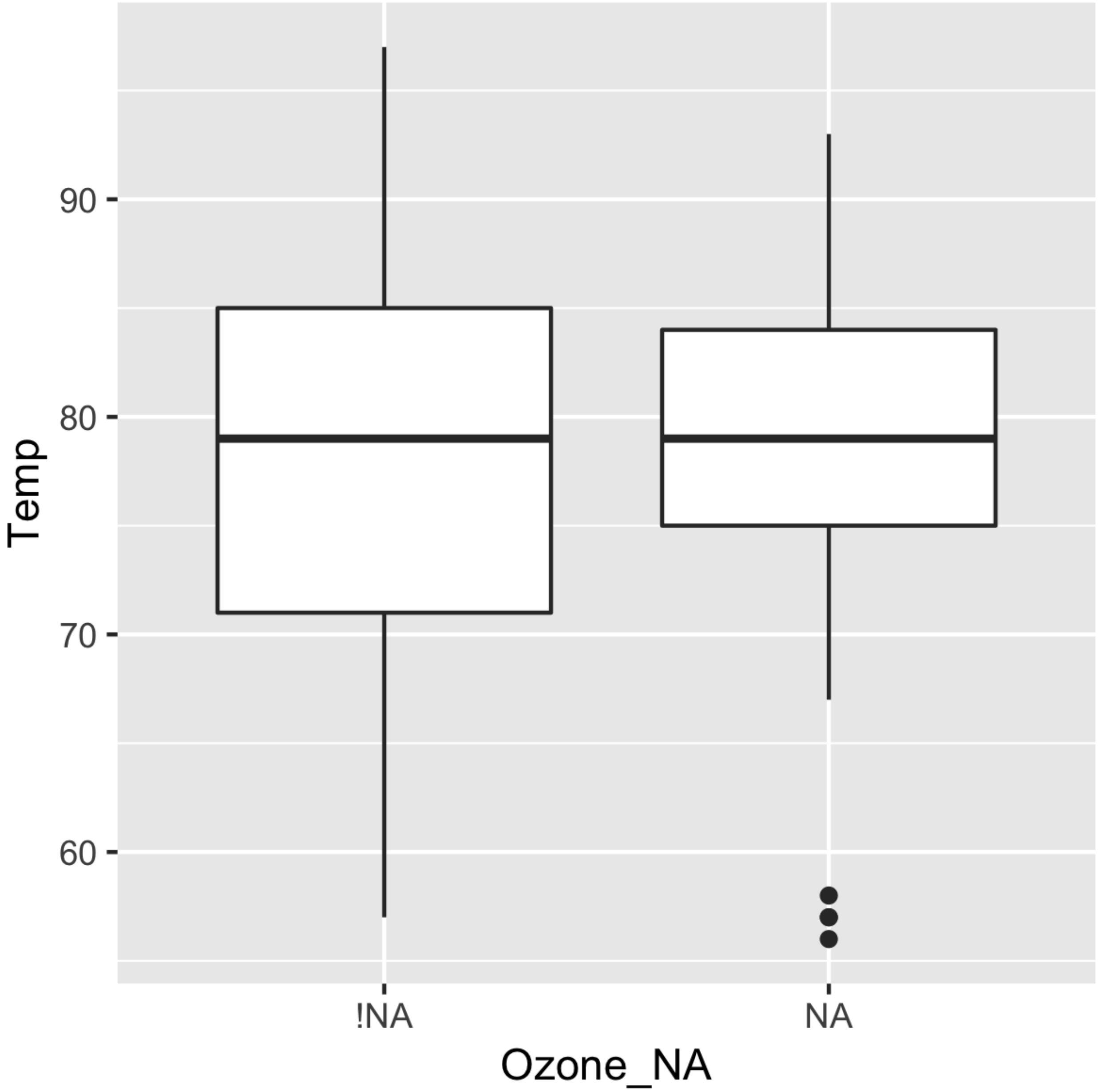
Visualizing missings using densities

```
airquality %>%
  bind_shadow() %>%
  ggplot(aes(x = Temp,
             color = Ozone_NA)) +
  geom_density()
```



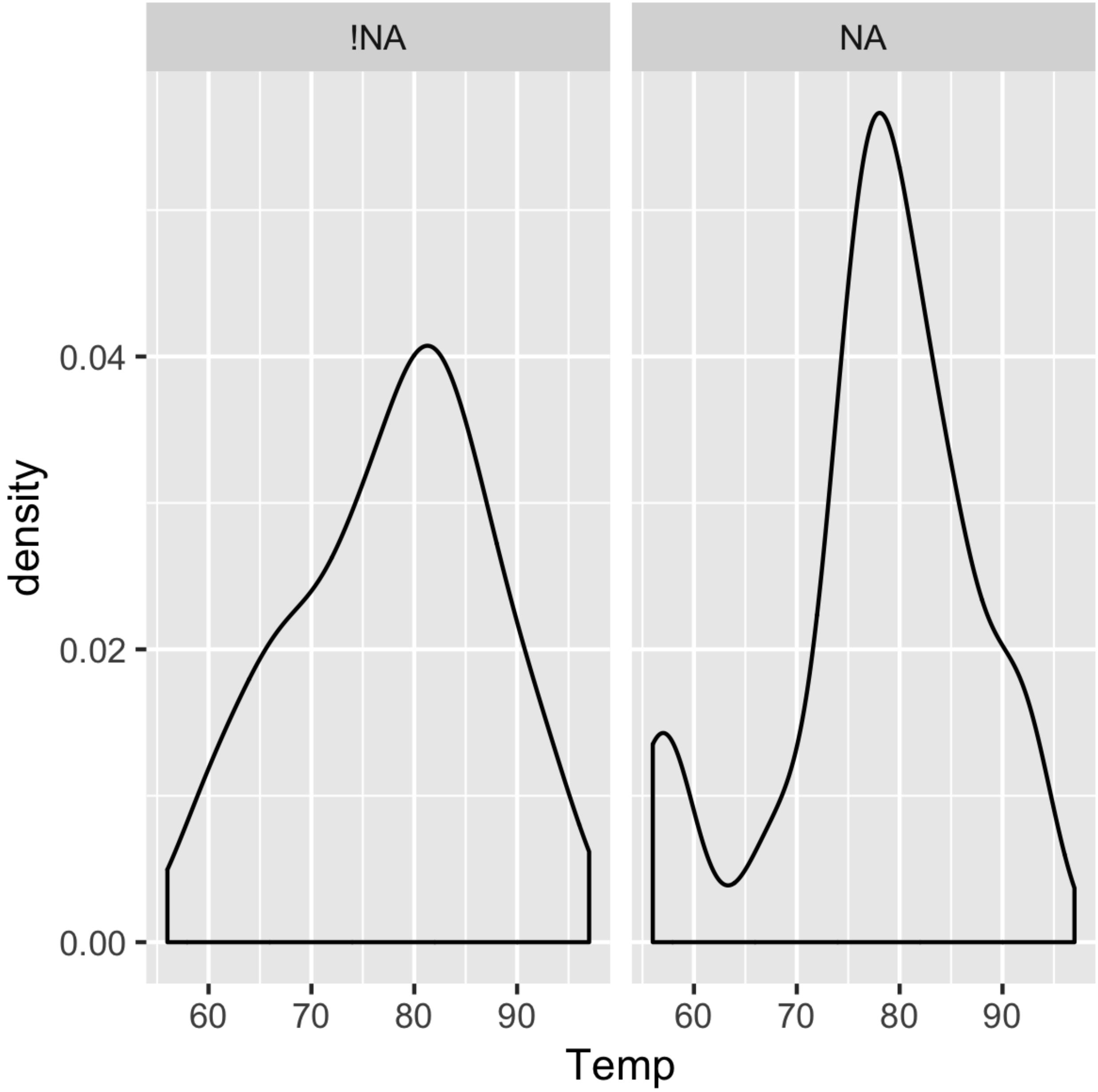
Visualizing missings using box plots

```
airquality %>%
  bind_shadow() %>%
  ggplot(aes(x = Ozone_NA,
             y = Temp)) +
  geom_boxplot()
```



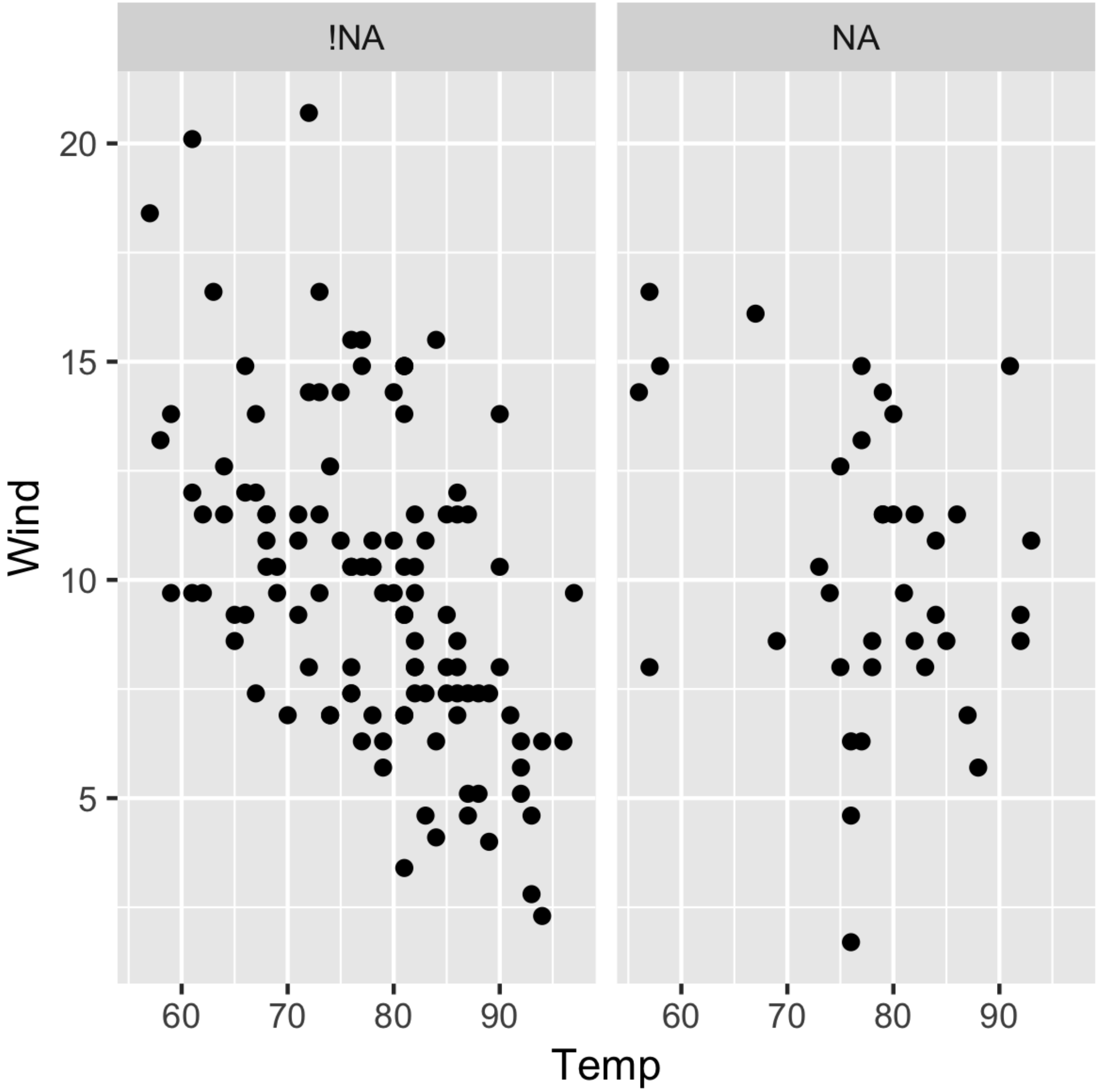
Visualizing missings using facets

```
airquality %>%
  bind_shadow() %>%
  ggplot(aes(x = Temp)) +
  geom_density() +
  facet_wrap(~Ozone_NA)
```



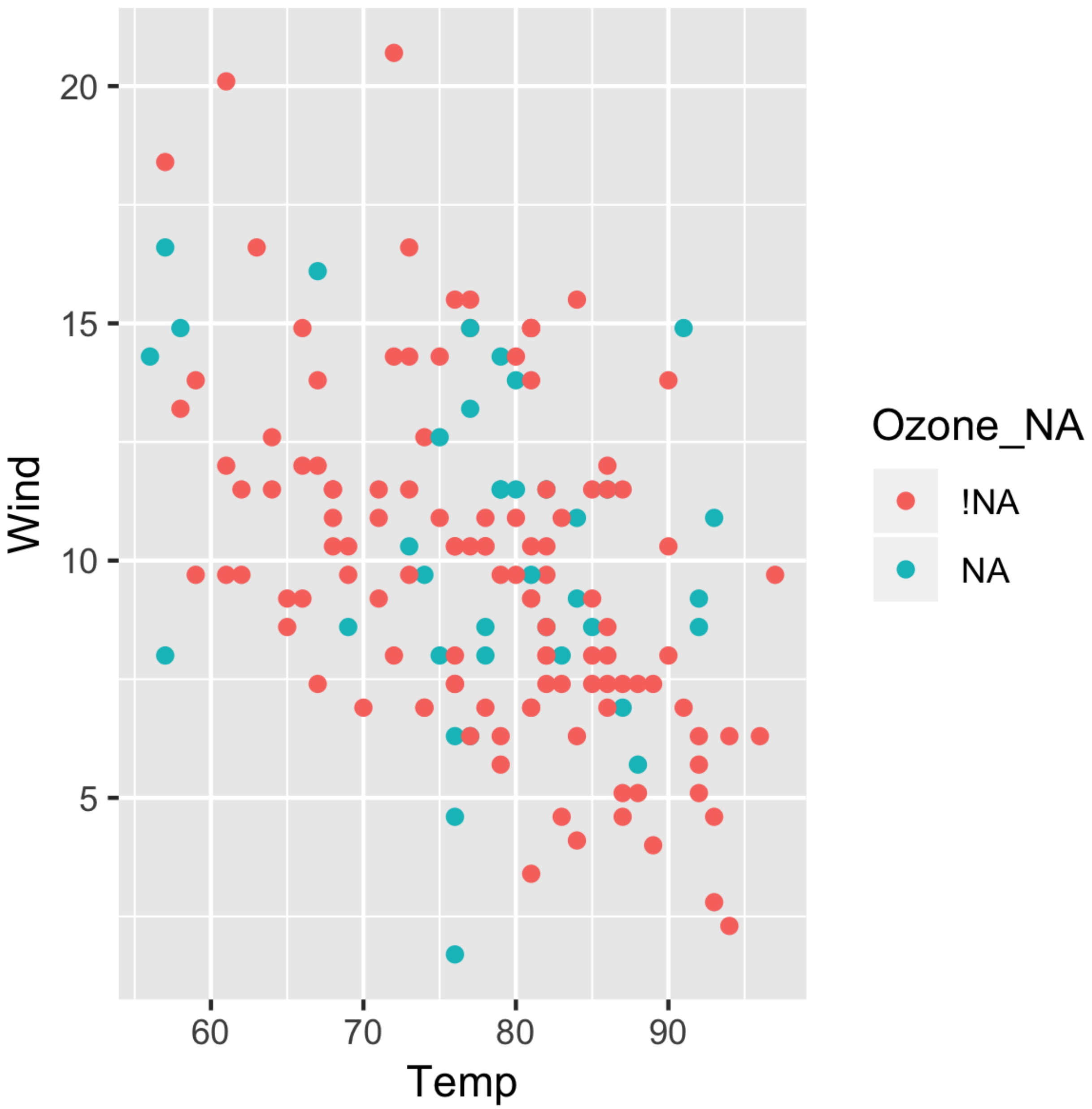
Visualizing missings using facets

```
airquality %>%
  bind_shadow() %>%
  ggplot(aes(x = Temp,
             y = Wind)) +
  geom_point() +
  facet_wrap(~ Ozone_NA)
```



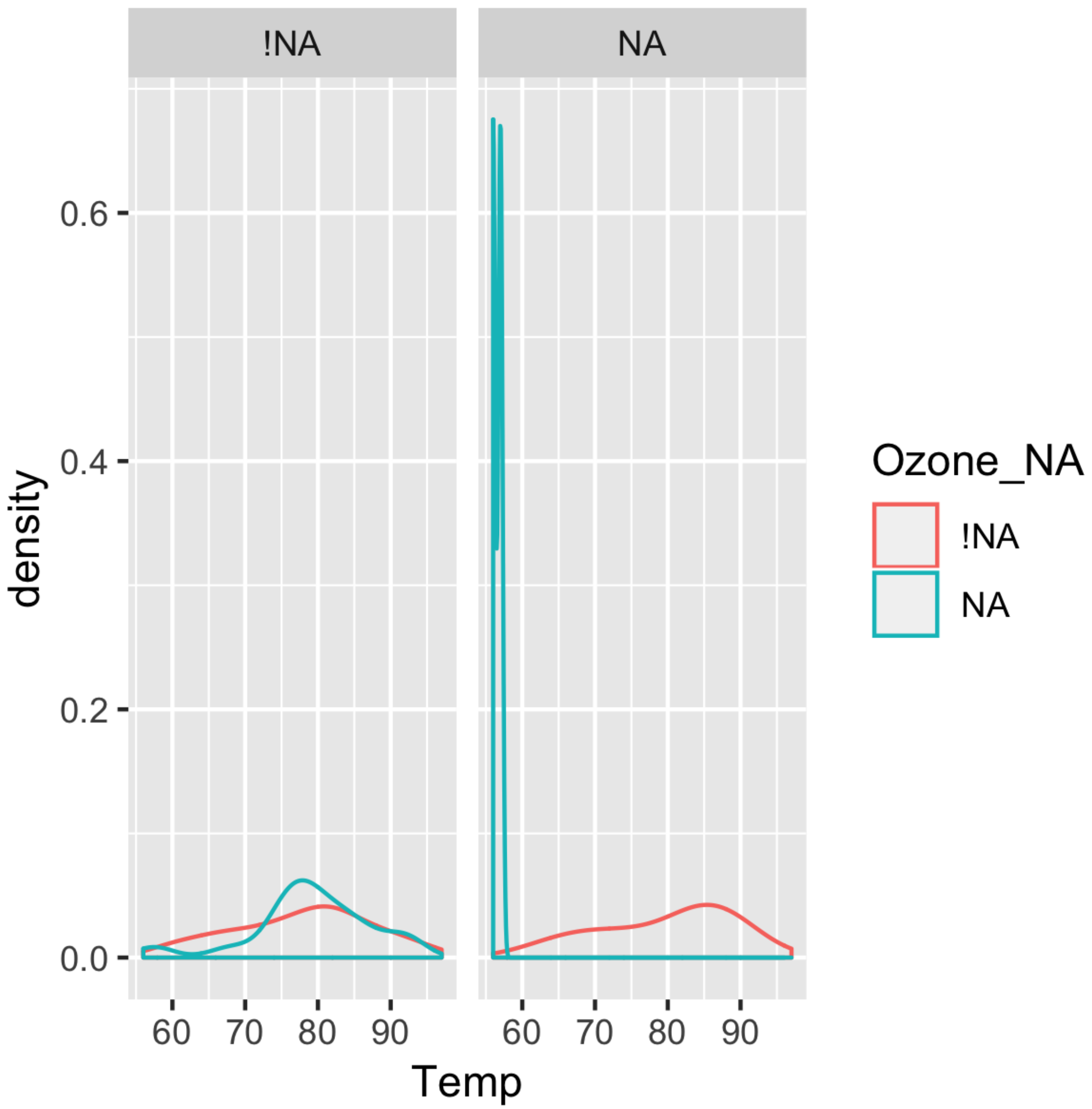
Visualizing missings using color

```r
airquality %>%
  bind_shadow() %>%
  ggplot(aes(x = Temp,
             y = Wind,
             color = Ozone_NA)) +
  geom_point()
```



Adding layers of missingness

```
airquality %>%
  bind_shadow() %>%
  ggplot(aes(x = Temp,
             color = Ozone_NA)) +
  geom_density()  +
  facet_wrap(~ Solar.R_NA)
```



Let's practice!

DEALING WITH MISSING DATA IN R

Visualizing missingness across two variables

DEALING WITH MISSING DATA IN R

Nicholas Tierney
Instructor

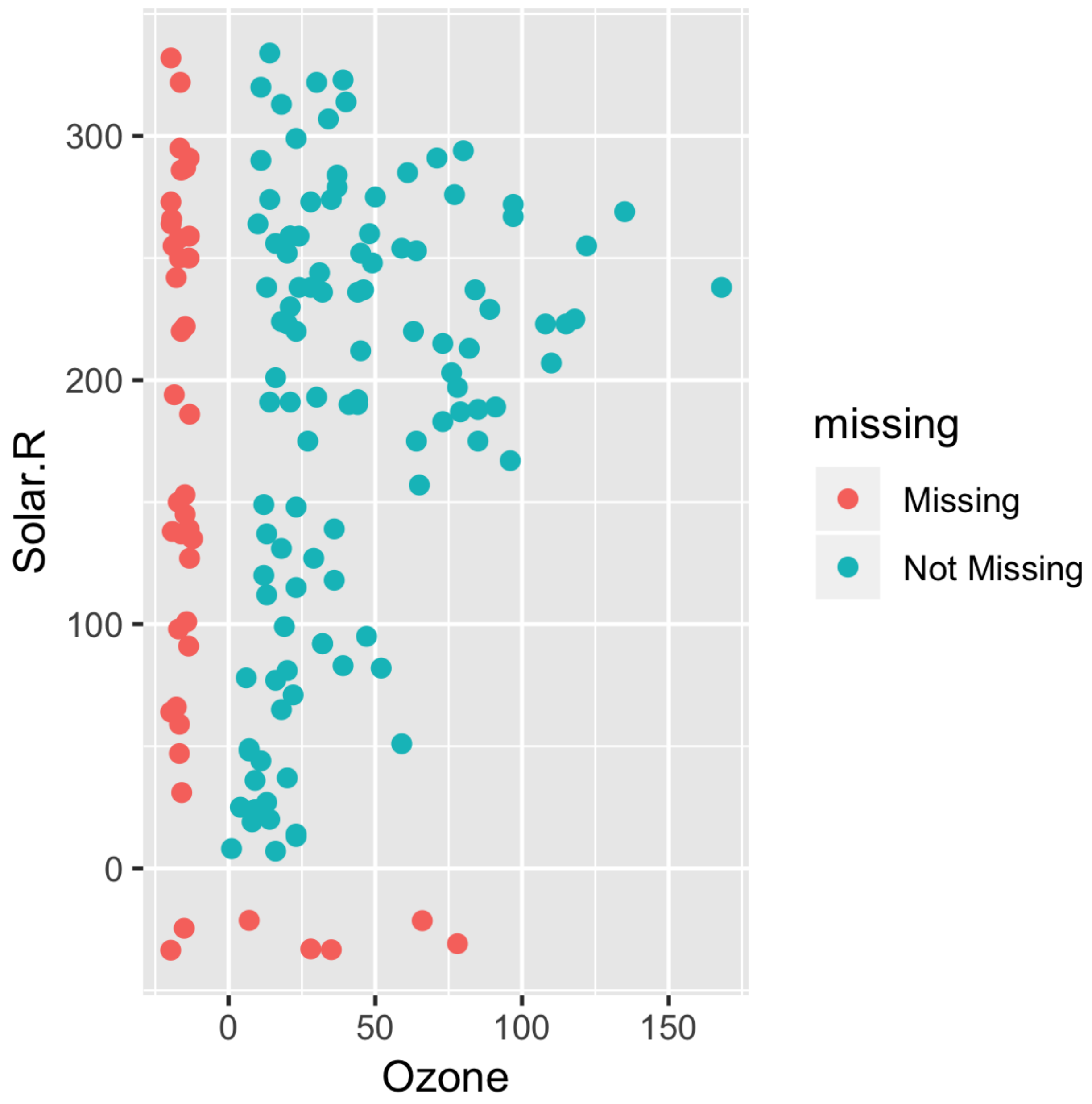
The problem of visualizing missing data in two dimensions

```r
ggplot(airquality,
       aes(x = Ozone,
           y = Solar.R)) +
  geom_point()
```

```
Warning message:
Removed 42 rows containing
missing values (geom_point).
```

Introduction to geom_miss_point()

```
ggplot(airquality,
       aes(x = Ozone,
           y = Solar.R)) +
  geom_miss_point()
```

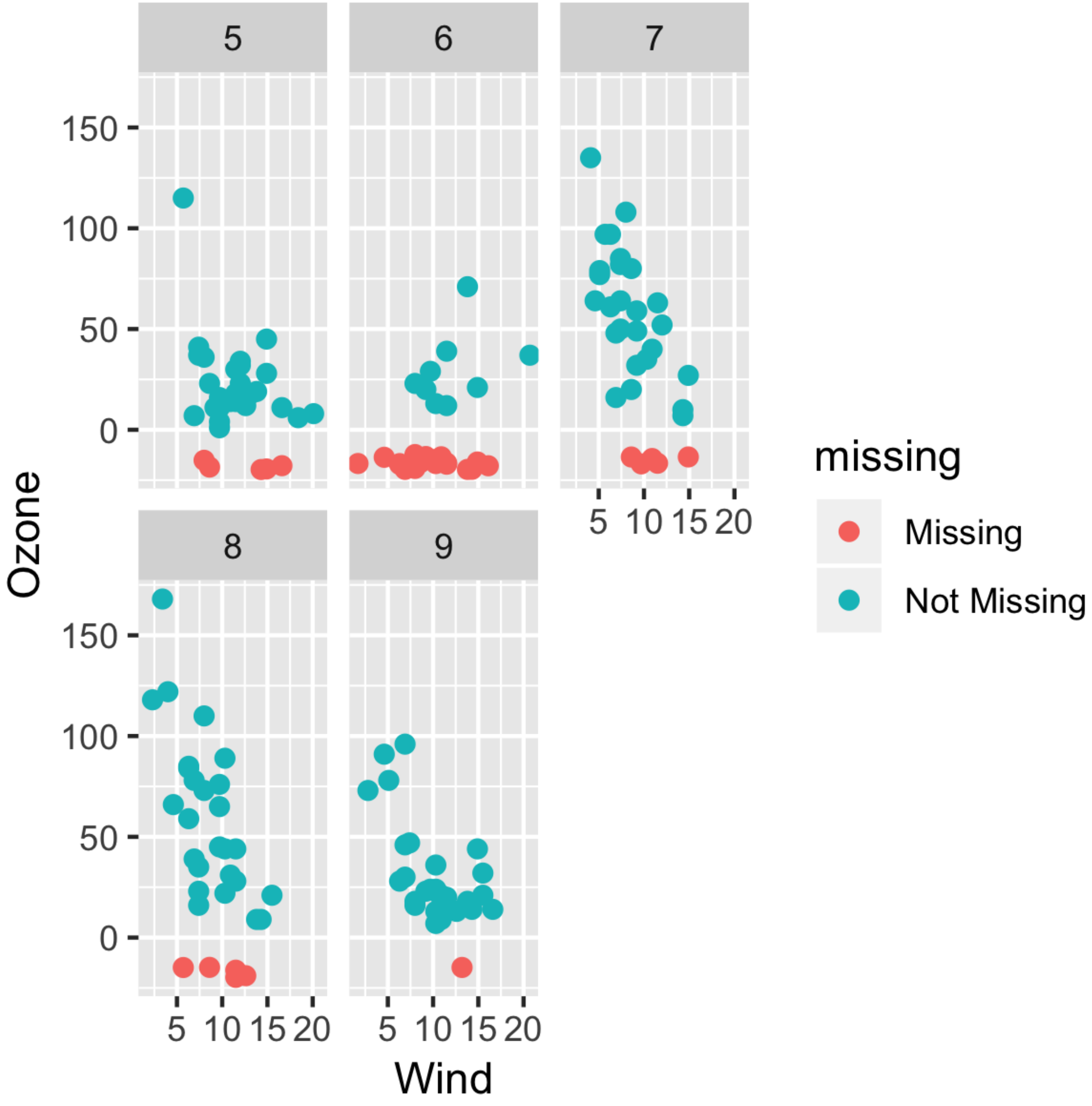


Aside: How geom_miss_point() works

Ozone	Ozone_shift	Ozone_NA
41	41.00000	!NA
36	36.00000	!NA
12	12.00000	!NA
18	18.00000	!NA
NA	-19.72321	NA
28	28.00000	!NA

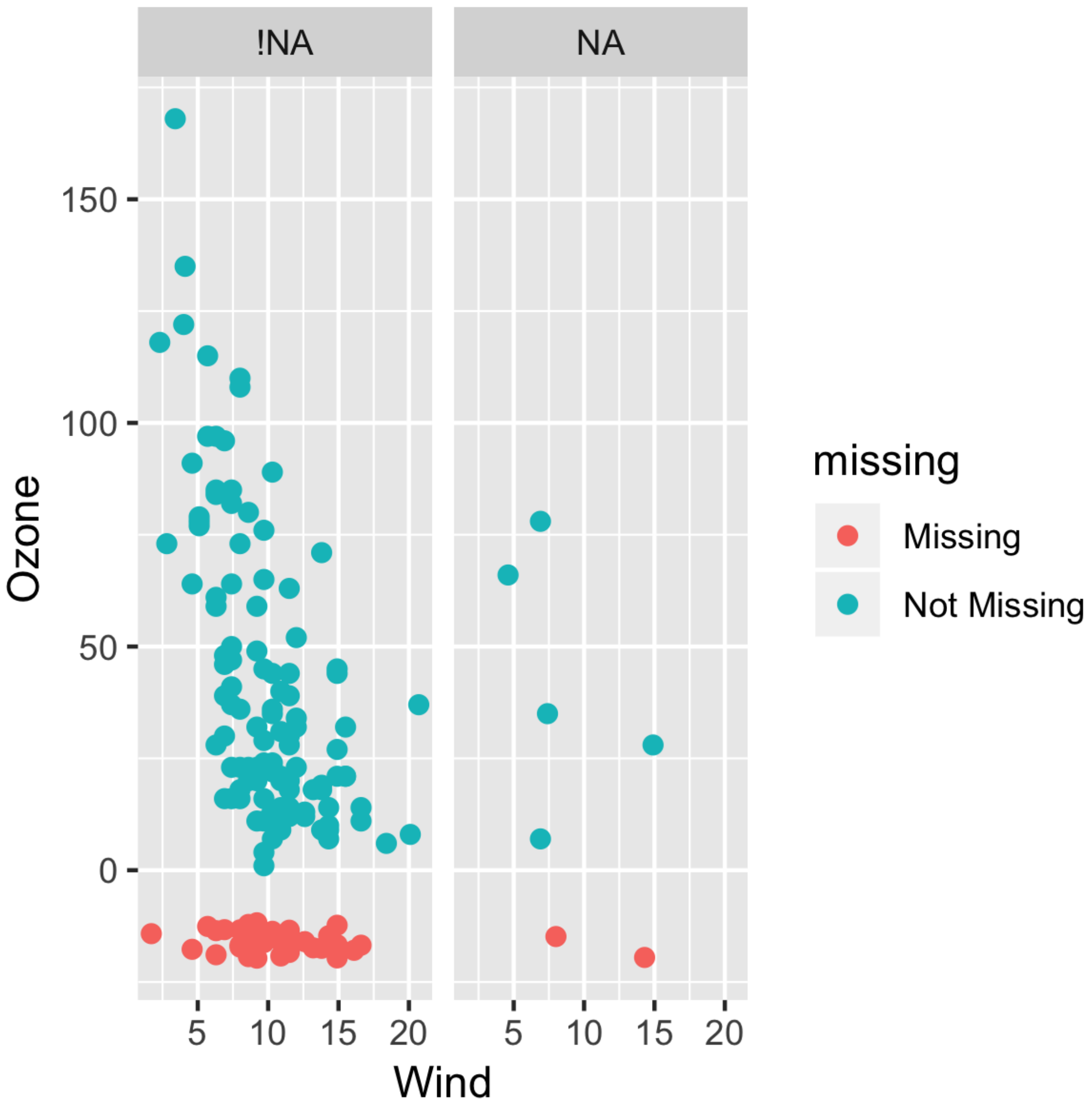
Exploring missingness using facets

```
ggplot(airquality,
       aes(x = Wind,
           y = Ozone)) +
  geom_miss_point() +
  facet_wrap(~ Month)
```



Exploring missingness using facets

```
airquality %>%
  bind_shadow() %>%
ggplot(aes(x = Wind,
           y = Ozone)) +
  geom_miss_point() +
  facet_wrap(~ Solar.R_NA)
```



Let's practice!

DEALING WITH MISSING DATA IN R