

Tidy Principles for Missing Data Exploration and Analysis using naniar

Nicholas Tierney

Why care about missing data?

Dealing with missing data in real world analysis is usually unavoidable and almost always frustrating.

naniar uses principles of **tidy missing data**, and provides **tidyverse-friendly** tools to clean up, summarise, and assess missing values, so you can easily:

- **Identify** missing values
- **Explore** why data is missing
- **Investigate** imputation

Tidy Data and Data Shadow

Tidy Data

Variables in columns
Observations in rows

A	B
2018	NA
NA	"Sam"

Data Shadow

Variables end in `_NA`
Values are Missing (NA) or Not Missing (!NA)

A_NA	B_NA
!NA	NA
NA	!NA

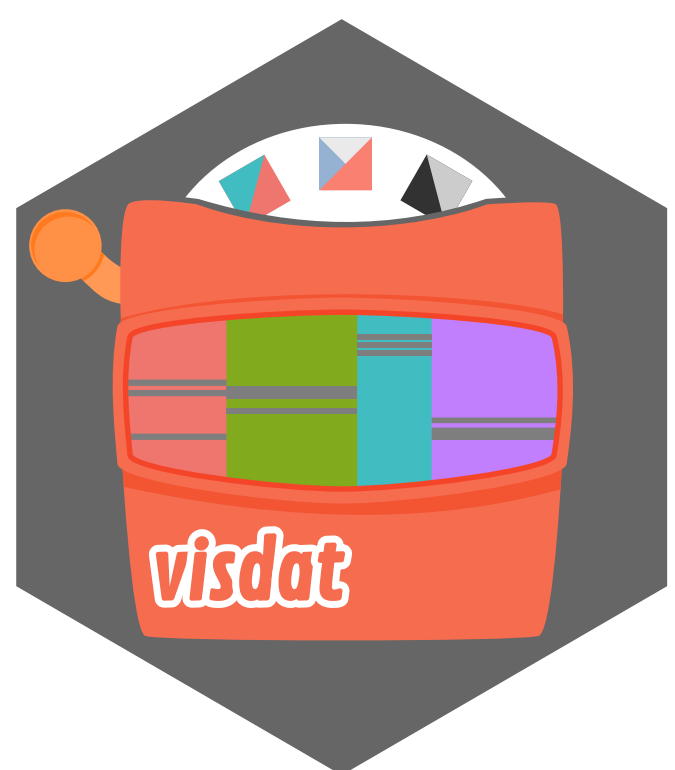
Tidy Missing Data

Bind data with shadow using `bind_shadow()`

A	B	A_NA	B_NA
2018	NA	!NA	NA
NA	"Sam"	NA	!NA



naniar
`bind_shadow()`
`geom_miss_point()`

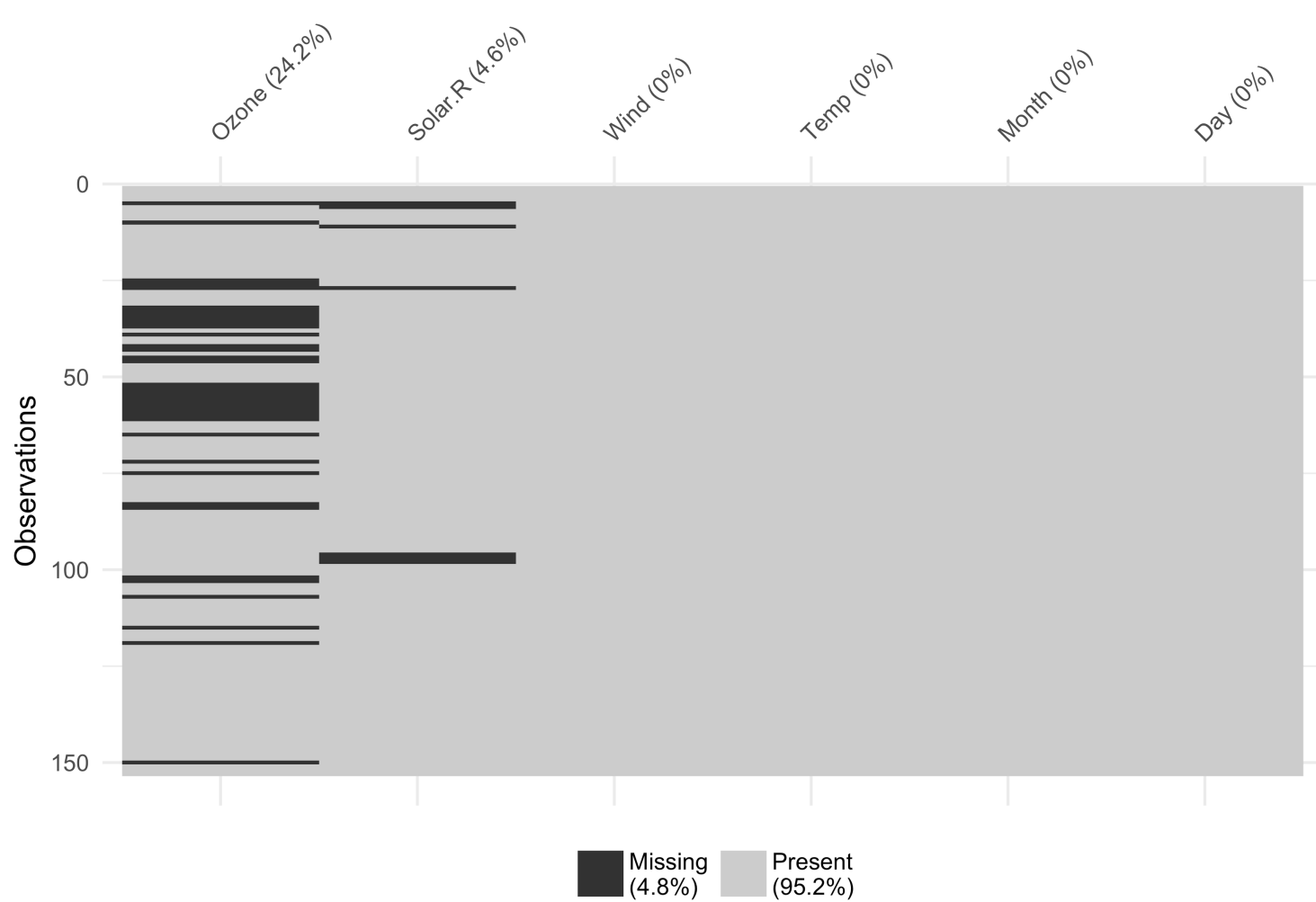


visdat
`vis_dat()`
`vis_miss()`

1. Identify missing values

Get a sense of the **missingness** in the whole data frame

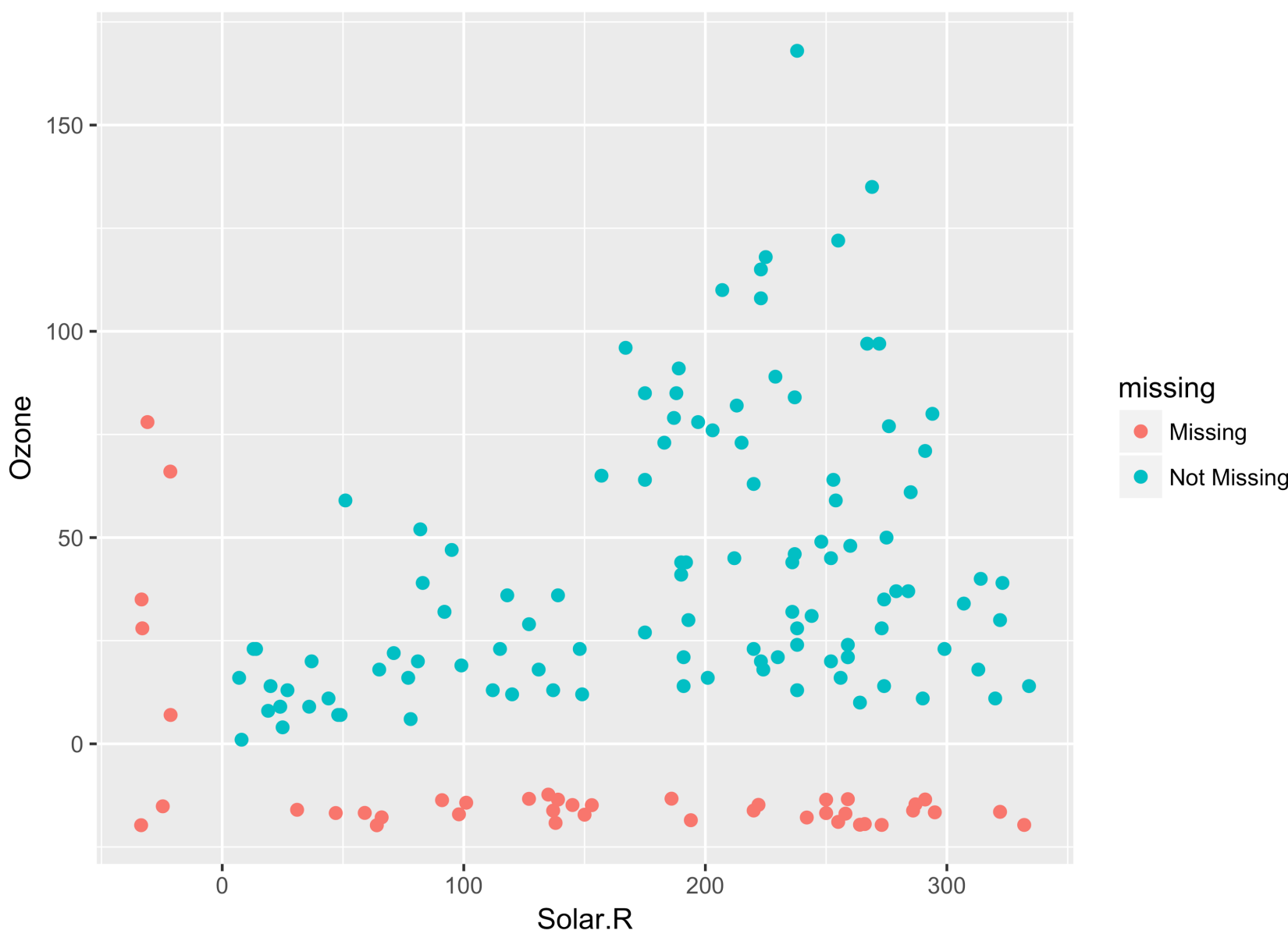
`vis_miss(airquality)`



2. Explore why data is missing

`geom_miss_point()`
Reveals missing values in ggplot2:

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



Summarise missing data:

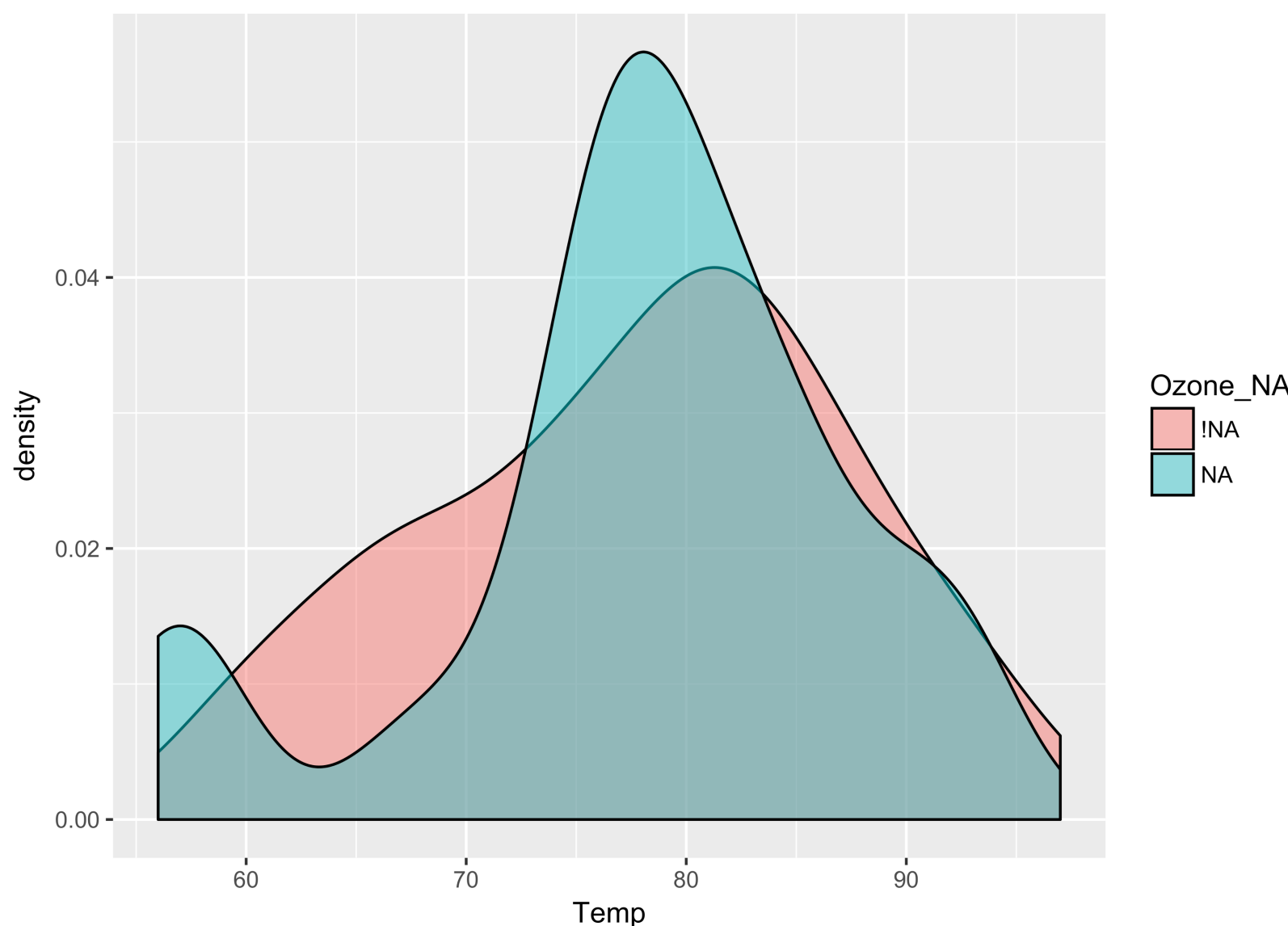
Combine with `dplyr::group_by()`

`miss_var_summary(airquality)`

```
airquality %>%  
  group_by(Month) %>%  
  miss_var_summary()  
  
# A tibble: 25 x 5  
  Month variable n_miss pct_miss  
  <int> <chr>      <int>    <dbl>  
1     5   Ozone         5     16.1  
2     5  Solar.R         4     12.9  
3     5   Wind          0         0  
4     5   Temp          0         0  
5     5   Day           0         0  
6     6   Ozone        21     70.0  
7     6  Solar.R         0         0
```

Refer to `_NA` variables to plot variables according to being missing:

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



3. Investigate imputation

Highlight imputed values by referring to whether they were previously missing using `_NA`

```
airquality %>%  
  bind_shadow() %>%  
  simulation::impute_lm(Ozone ~ Temp + Wind) %>%  
  ggplot(aes(x = Temp,  
            y = Ozone,  
            colour = Ozone_NA)) +  
  geom_point()
```

