Introduction to missing data

DEALING WITH MISSING DATA IN R

Nicholas Tierney
Statistician

Introduction

> The best thing to do with missing data is to not have any

> **--Gertrude Mary Cox**

- Working with real-world data = working with missing data
- Missing data can have unexpected effects on your analysis
- Bad imputation can lead to poor estimates and decisions.

What will you learn

- What missing values are
- How to find missing data
- How to wrangle and tidy missing data
- Explore why is data missing
- Impute missing values

Assumed knowledge

- **Basic** to **intermediate** experience with R
- Experience creating plots using **ggplot2**
- Experience **using dplyr** to manipulate and rearrange data
- Experience **fitting linear models** in R

What are missing values?

> Missing values are values that should have been recorded but were not.

`NA` = **N**ot **A**vailable.

How do I check if I have missing values?

```r
x <- c(1, NA, 3, NA, NA, 5)
any_na(x)
```

```
TRUE
```

```r
are_na(x)
```

```
FALSE  TRUE FALSE  TRUE  TRUE FALSE
```

```r
n_miss(x)
```

```
3
```

```r
prop_miss(x)
```

```
0.5
```

Working with missing data

`NA` **+ anything =** `NA`

```
heights
```

```
Sophie    Dan   Fred
   165    177     NA
```

```
sum(heights)
```

```
NA
```

Missing data gotchas

`NaN` : Not a Number.

```
any_na(NaN)
```

```
TRUE
```

```
any_na(NULL)
```

```
FALSE
```

```
any_na(Inf)
```

```
FALSE
```

Missing data gotchas (2)

```
NA | TRUE
```

```
TRUE
```

```
NA | FALSE
```

```
NA
```

```
NA + NaN
```

```
NA
```

```
NaN + NA
```

```
NaN
```

Let's practice!

DEALING WITH MISSING DATA IN R

How to summarise missing values

DEALING WITH MISSING DATA IN R

Nicholas Tierney
Statistician

Introduction to missingness summaries

Basic summaries of missingness:

- `n_miss`
- `n_complete`

Dataframe summaries of missingness:

- `miss_var_summary`
- `miss_case_summary`

These functions work with `group_by`

Missing data summaries: Variables

```
miss_var_summary(airquality)
```

```
# A tibble: 6 x 3
  variable n_miss pct_miss
  <chr>     <int>    <dbl>
1 Ozone        37    24.2
2 Solar.R       7     4.58
3 Wind          0     0
4 Temp          0     0
5 Month         0     0
6 Day           0     0
```

Missing data summaries: Cases

```
miss_case_summary(airquality)
```

```
# A tibble: 153 x 3
    case n_miss pct_miss
   <int>  <int>    <dbl>
 1     5      2     33.3
 2    27      2     33.3
 3     6      1     16.7
 4    10      1     16.7
 5    11      1     16.7
 6    25      1     16.7
 7    26      1     16.7
 8    32      1     16.7
 9    33      1     16.7
10    34      1     16.7
# ... with 143 more rows
```

Missing data tabulations

```
miss_var_table(airquality)
```

```
# A tibble: 3 x 3
  n_miss_in_var n_vars pct_var
          <int>  <int>    <dbl>
1             0      4     66.7
2             7      1     16.7
3            37      1     16.7
```

```
miss_case_table(airquality)
```

```
# A tibble: 3 x 3
  n_miss_in_case n_cases pct_case
           <int>   <int>    <dbl>
1              0     111    72.5
2              1      40    26.1
3              2       2     1.31
```

Missing data summaries: Spans of missing data

```r
miss_var_span(pedestrian, var = hourly_counts, span_every = 4000)
```

```
# A tibble: 10 x 5
   span_counter n_miss n_complete prop_miss prop_complete
          <int>  <int>      <dbl>     <dbl>         <dbl>
 1            1      0       4000   0                 1
 2            2      1       3999   0.00025           1.000
 3            3    121       3879   0.0302            0.970
 4            4    503       3497   0.126             0.874
 5            5    745       3255   0.186             0.814
 6            6      0       4000   0                 1
 7            7      1       3999   0.00025           1.000
 8            8      0       4000   0                 1
 9            9    745       3255   0.186             0.814
10           10    432       3568   0.108             0.892
```

Missing data summaries: Runs of missing data

```
miss_var_run(pedestrian, hourly_counts)
```

```
# A tibble: 35 x 2
   run_length is_na
        <int> <chr>
 1       6628 complete
 2          1 missing
 3       5250 complete
 4        624 missing
 5       3652 complete
 6          1 missing
 7       1290 complete
 8        744 missing
 9       7420 complete
10          1 missing
# ... with 25 more rows
```

Using summaries with group_by

```
airquality %>%
  group_by(Month) %>%
  miss_var_summary()
```

```
# A tibble: 25 x 4
   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
 7     6 Solar.R       0      0
# ... with 18 more rows
```

Let's practice!

How do we visualize missing values?

DEALING WITH MISSING DATA IN R

Nicholas Tierney
Statistician

Introduction to missing data visualizations in naniar

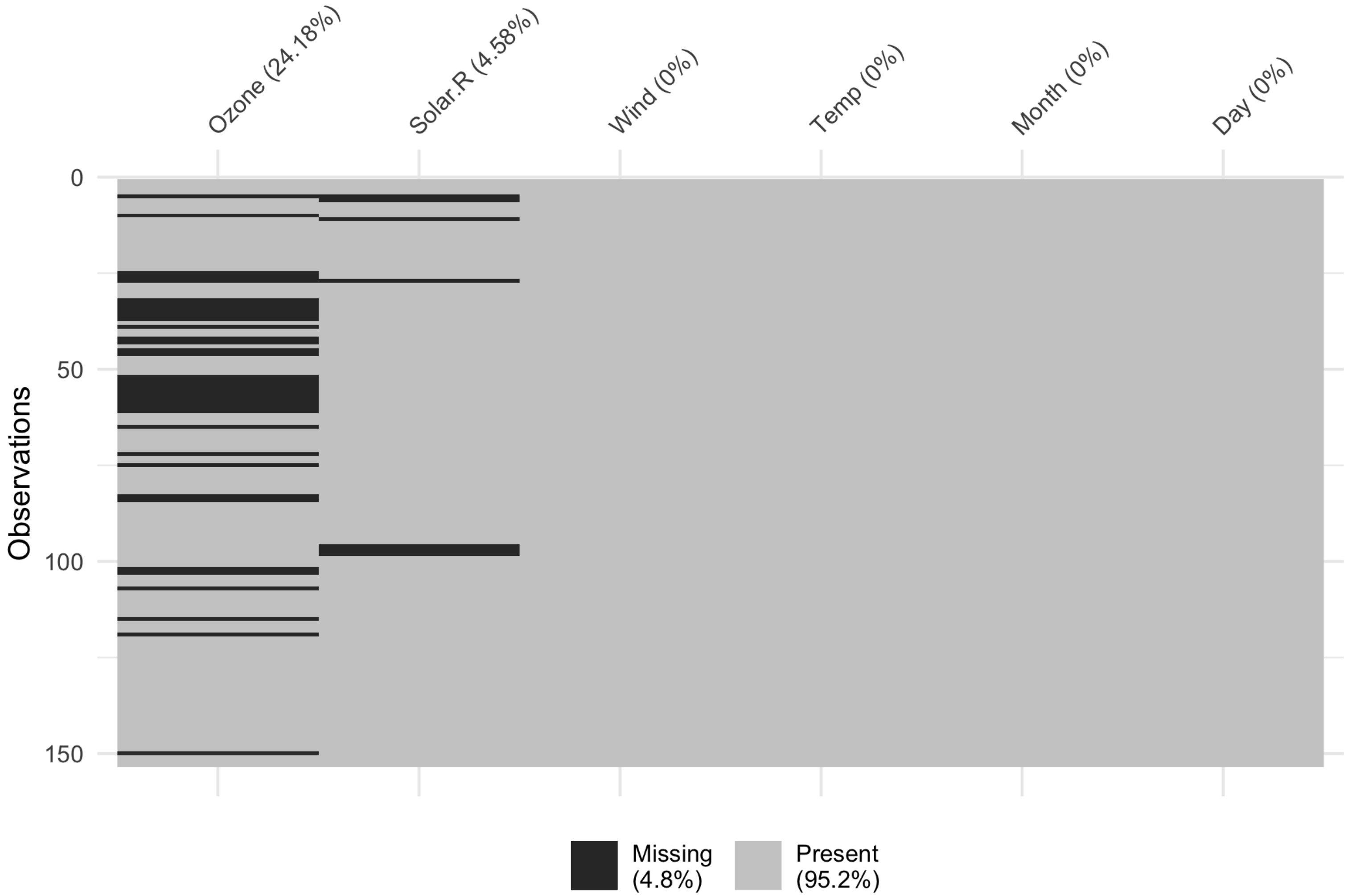
- Visualization can quickly capture an idea or thought.
- `naniar` provides a friendly family of missing data visualization functions.
- Each visualization corresponds to a data summary.
- Visualizations help you operate closer to the speed of thought.

Lesson overview

- How to get a bird's eye view of the data
- How to look at missings in the variables and cases
- How to generate visualizations for missing spans and across groups in the data.

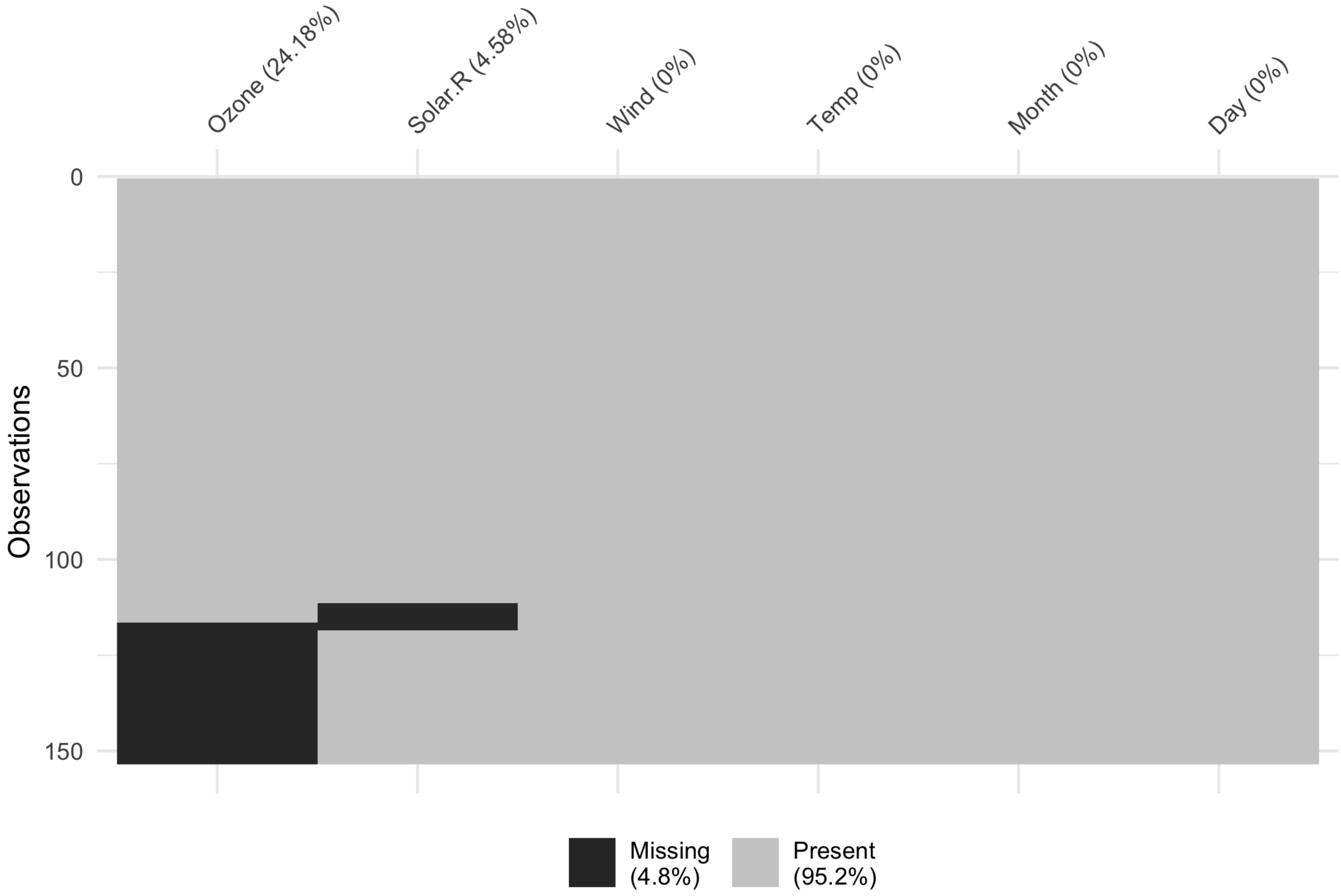
Get a bird's eye view of the missing data

```
vis_miss(airquality)
```



Get a bird's eye view of the missing data

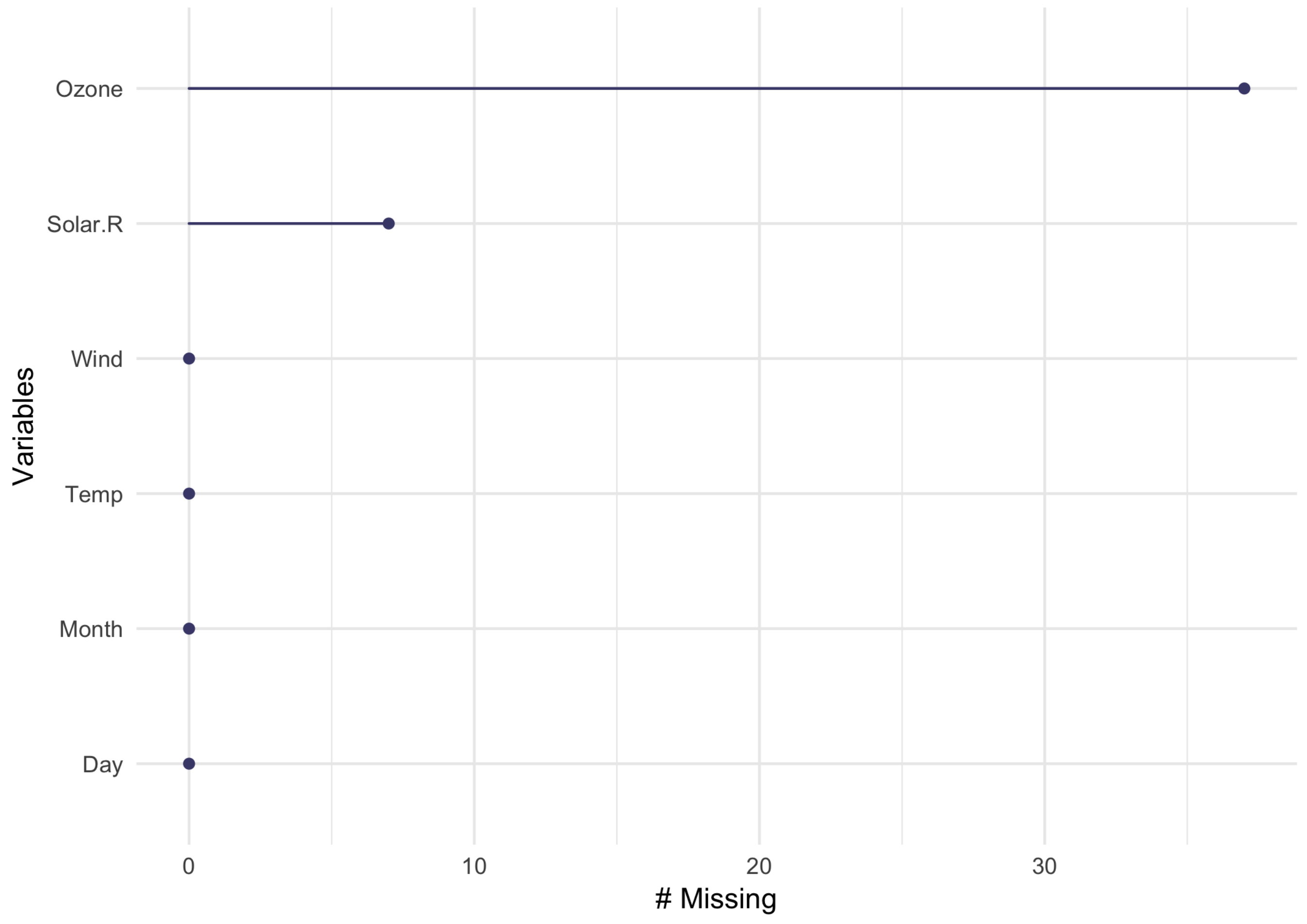
```
vis_miss(airquality, cluster = TRUE)
```

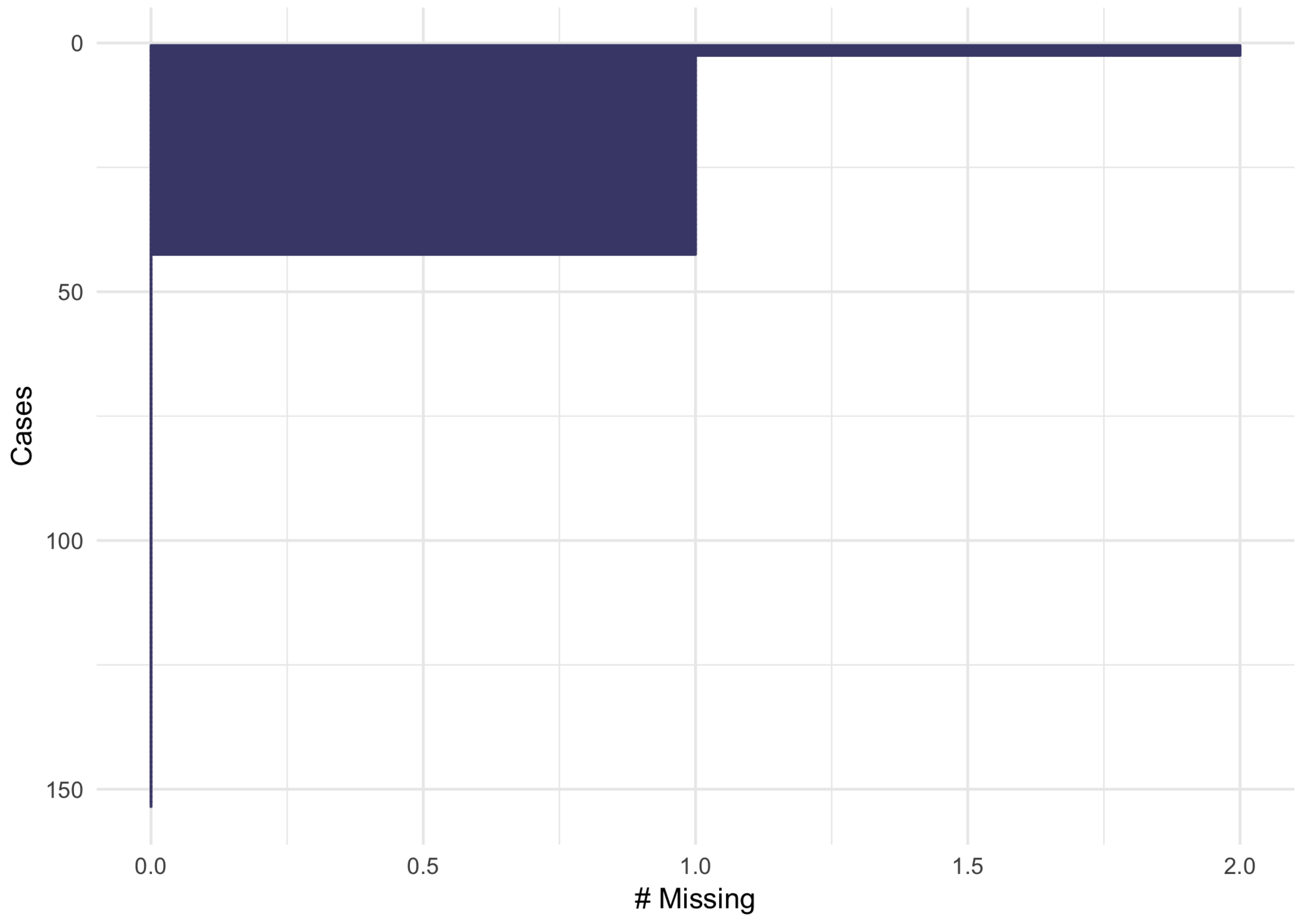


Look at missings in variables and cases

```
gg_miss_var(airquality)
```

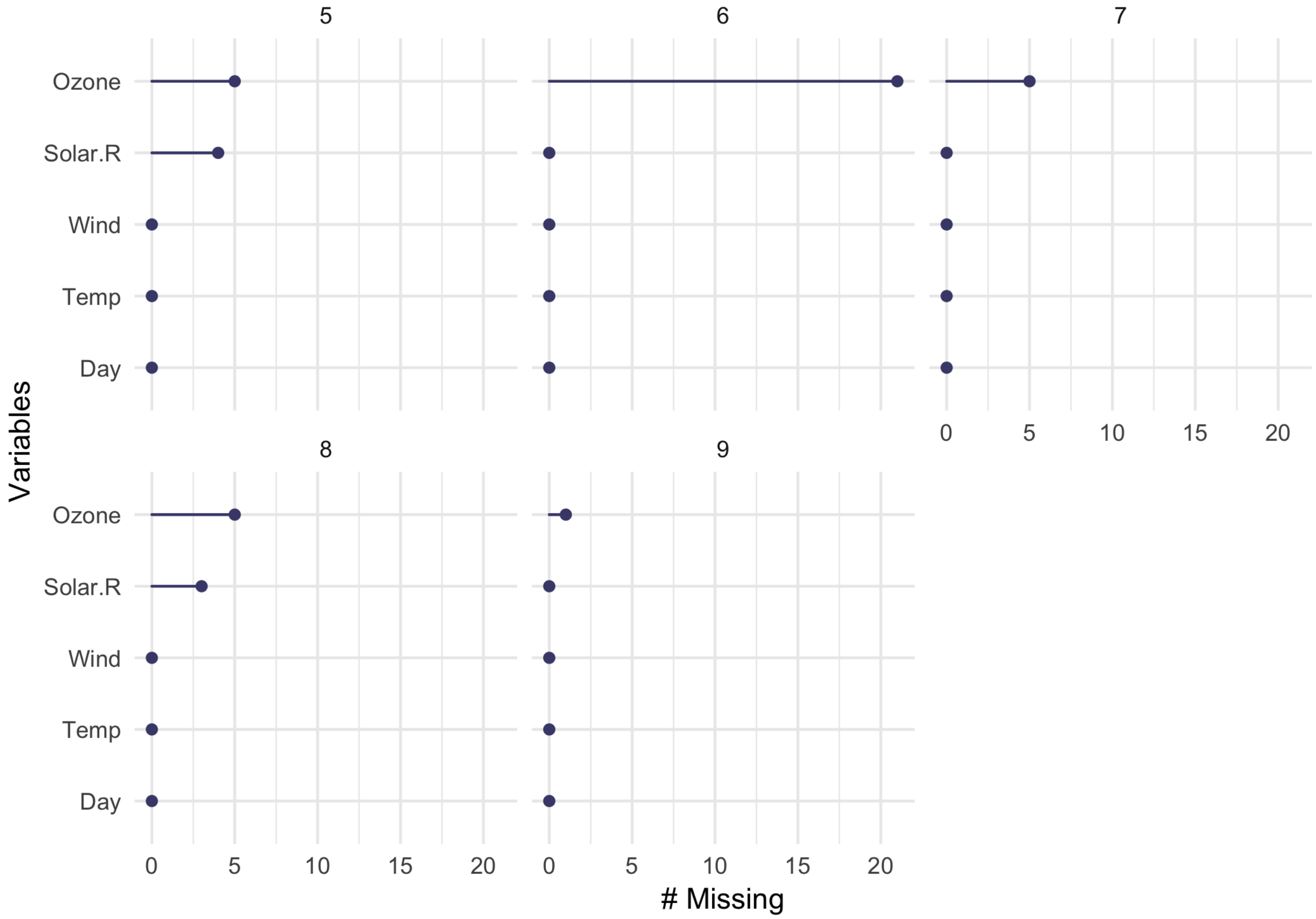
```
gg_miss_case(airquality)
```





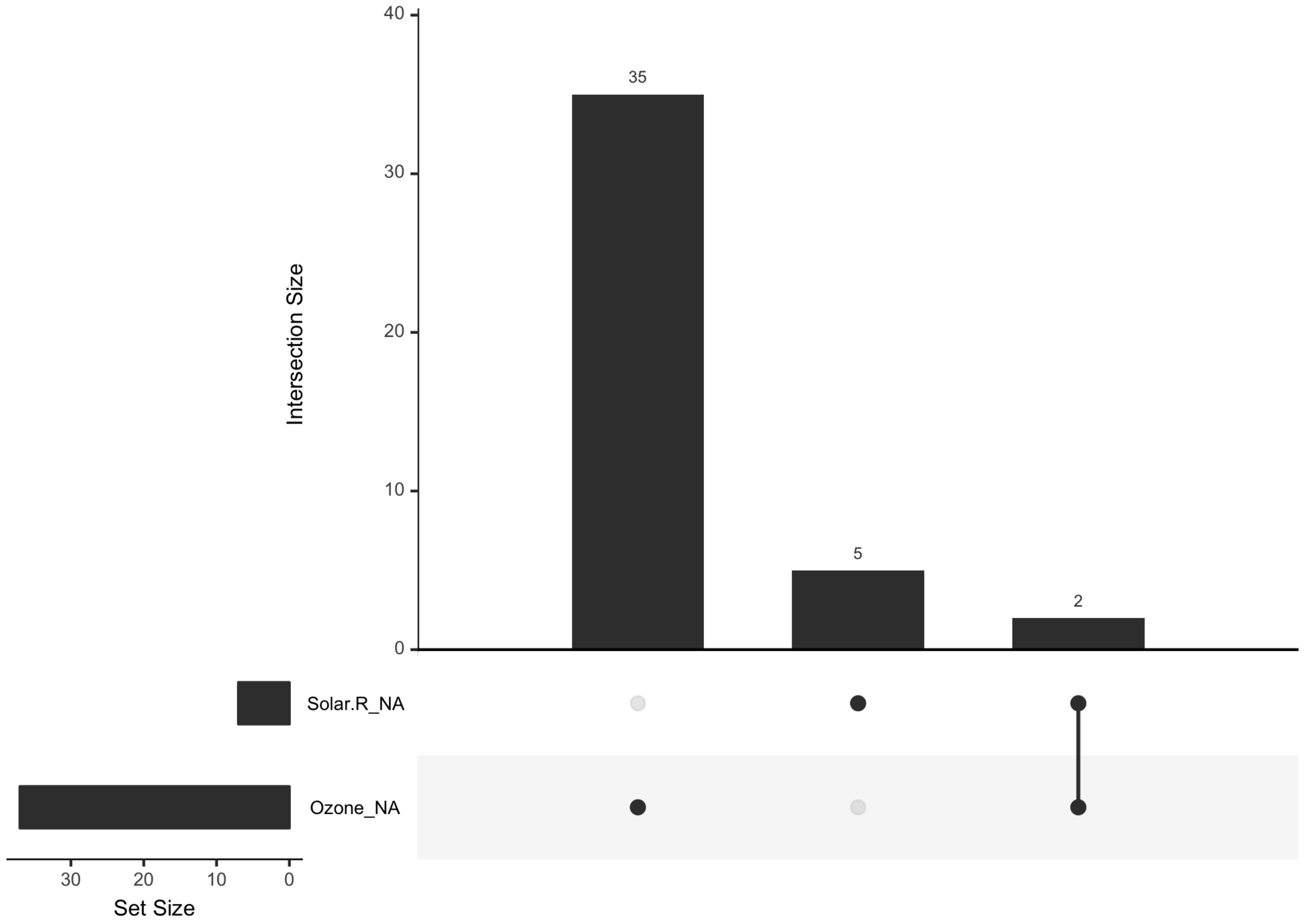
Look at missings in variables and cases

```
gg_miss_var(airquality, facet = Month)
```



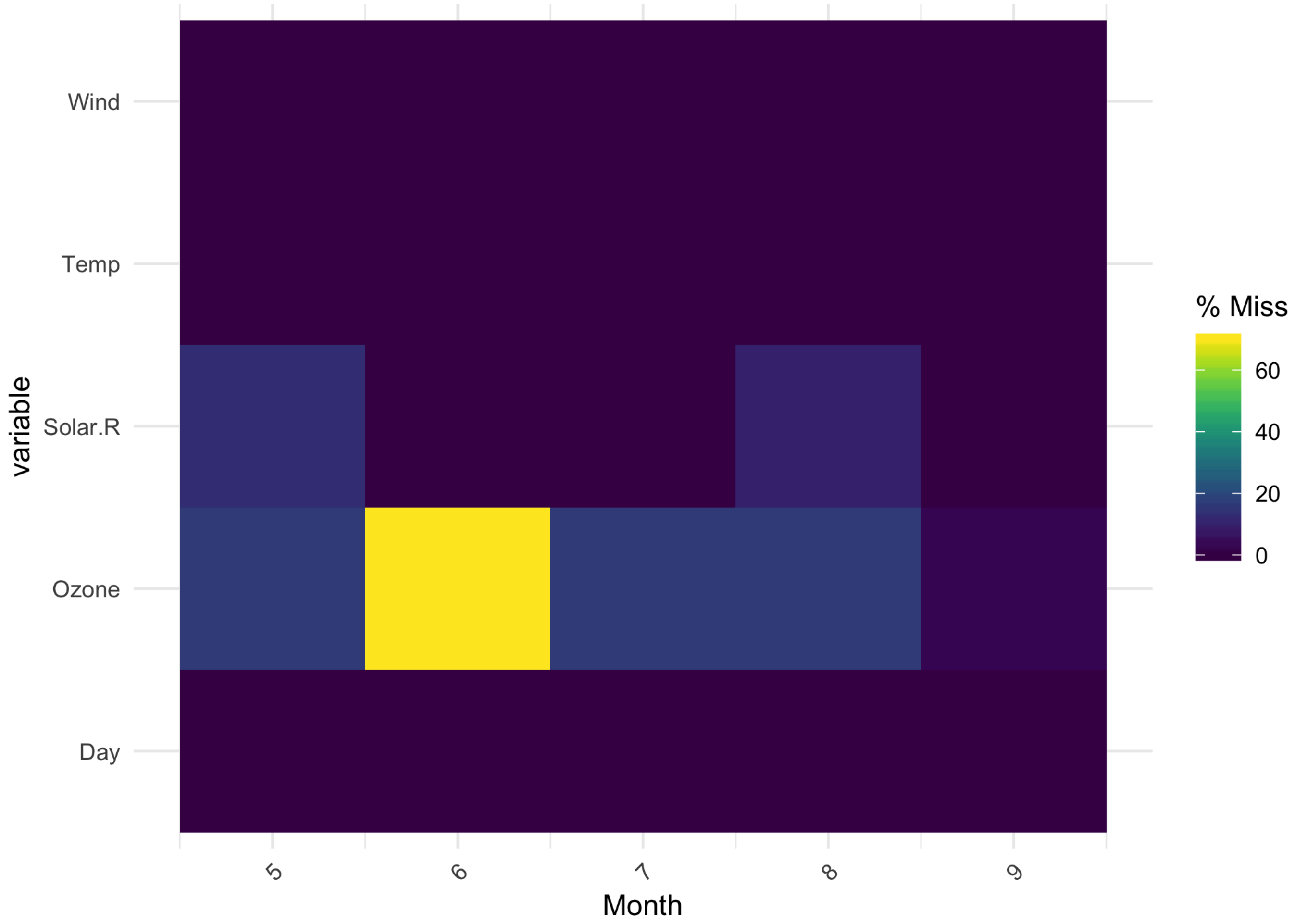
Visualizing missingness patterns

```
gg_miss_upset(airquality)
```



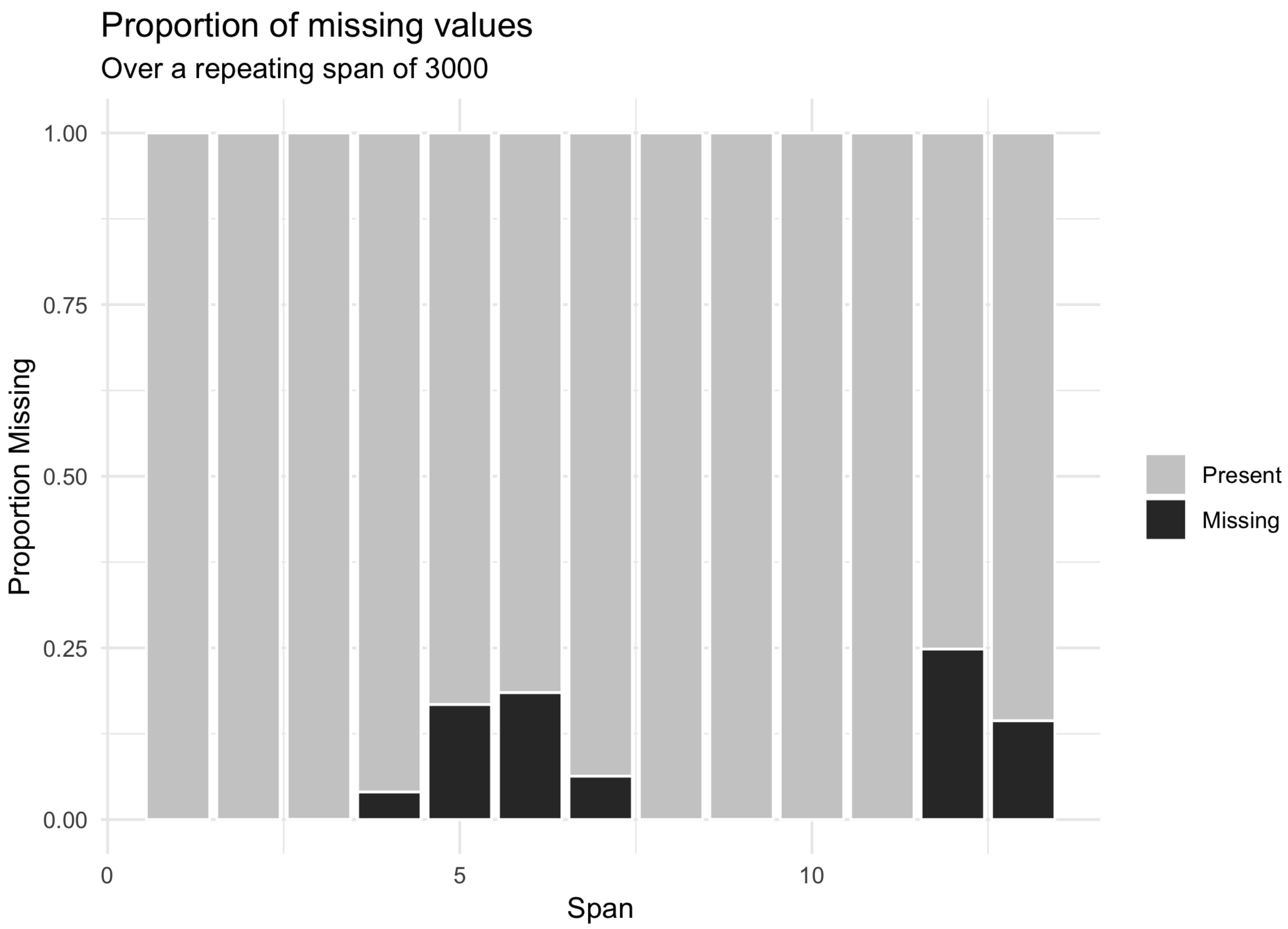
Visualizing factors of missingness

```
gg_miss_fct(x = airquality, fct = Month)
```



Visualizing spans of missingness

```
gg_miss_span(pedestrian, hourly_counts, span_every = 3000)
```



Let's practice!

DEALING WITH MISSING DATA IN R