



DEALING WITH MISSING DATA IN R

# Searching for and replacing missing values

Nicholas Tierney  
Statistician



# What we are going to cover

- How to look for hidden missing values
- Replacing missing value labels with `NA`
- Checking your assumptions on missingness

# Searching for and replacing missing values

- Ideal = `NA`
- Missing values can be coded incorrectly: e.g. "missing", "Not Available", "N/A"
- Assuming that missing values are coded as `NA`. This is a mistake.



# Understanding Chaos

score	grade	place
3	N/A	-99
-99	E	97
4	missing	95
-99	na	92
7	n/a	-98
10		missing
12	.	88
16		.
9	N/a	86



# Searching for missing values

```
miss_scan_count()
```

```
chaos %>%  
  miss_scan_count(search = list("N/A"))
```

```
# A tibble: 3 x 2  
  Variable      n  
  <chr>    <int>  
1 score          0  
2 grade          1  
3 place          0
```



# Searching for missing values

```
chaos %>%  
  miss_scan_count(search = list("N/A",  
                                "N/a"))
```

```
# A tibble: 3 x 2  
  Variable      n  
  <chr>    <int>  
1 score        0  
2 grade        2  
3 place        0
```



# Replacing missing values

```
chaos %>%  
  replace_with_na(replace = list(grade = c("N/A", "N/a")))
```

```
# A tibble: 9 x 3  
  score grade    place  
  <dbl> <chr>   <chr>  
1      3 NA      -99  
2    -99 E       97  
3      4 missing 95  
4    -99 na      92  
5      7 n/a     -98  
6     10 " "     missing  
7     12 .      88  
8     16 ""      .  
9      9 NA      86
```



# "scoped variants" of `replace_with_na`

- `replace_with_na` can be repetitive:
  - Use it across many different variables and values
  - Complex cases, replacing values less than -1, only affect character columns.
- `replace_with_na_all()` All variables.
- `replace_with_na_at()` A subset of selected variables.
- `replace_with_na_if()` A subset of variables that fulfill some condition ( numeric, character).



# Using scoped variants of `replace_with_na`

```
chaos %>%  
  replace_with_na_all(condition = ~.x == -99)
```

```
# A tibble: 9 x 3  
  score grade  place  
  <dbl> <chr>  <chr>  
1     3 N/A    NA  
2    NA E      97  
3     4 missing 95  
4    NA na     92  
5     7 n/a    -98  
6    10 " "    missing  
7    12 .      88  
8    16 ""     .  
9     9 N/a    86
```



# Using scoped variants of `replace_with_na`

```
chaos %>%  
  replace_with_na_all(condition = ~.x %in% c("N/A", "missing", "na"))
```

```
# A tibble: 9 x 3  
  score grade place  
  <dbl> <chr> <chr>  
1      3 NA    -99  
2    -99 E      97  
3      4 NA     95  
4    -99 NA     92  
5      7 n/a    -98  
6     10 " "     NA  
7     12 .      88  
8     16 ""      .  
9      9 N/a     86
```



## DEALING WITH MISSING DATA IN R

**Let's practice!**



DEALING WITH MISSING DATA IN R

# Missing, missing data

Nicholas Tierney  
Statistician



# Another perspective on missing data

name	time	value
robin	morning	358
robin	afternoon	534
robin	evening	100
sam	morning	139
sam	afternoon	177
blair	morning	963
blair	afternoon	962
blair	evening	929

name	afternoon	evening	morning
blair	962	929	963
robin	534	100	358
sam	177	NA	139



# Explicit and Implicit missing values

- **explicitly:** They are missing with `NA`
- **implicitly:** Not shown in the data, but implied

# Making implicit missings explicit

```
tetris %>%  
  tidyr::complete(name, time)
```

```
# A tibble: 9 x 3  
  name    time    value  
  <fct> <fct>    <dbl>  
1 blair  afternoon  962  
2 blair  evening    929  
3 blair  morning    963  
4 robin  afternoon  534  
5 robin  evening    100  
6 robin  morning    358  
7 sam    afternoon  177  
8 sam    evening    NA  
9 sam    morning    139
```



# Handling explicitly missing values

name	time	value
robin	morning	936
NA	afternoon	635
NA	evening	438
sam	morning	208
NA	afternoon	92
NA	evening	79
blair	morning	969
NA	afternoon	918
NA	evening	954

name	time	value
robin	morning	936
robin	afternoon	635
robin	evening	438
sam	morning	208
sam	afternoon	92
sam	evening	79
blair	morning	969
blair	afternoon	918
blair	evening	954



# Handling explicitly missing values

name	time	value
robin	morning	936
NA	afternoon	635
NA	evening	438
sam	morning	208
NA	afternoon	92
NA	evening	79
blair	morning	969
NA	afternoon	918
NA	evening	954

```
tetris %>%  
  tidyr::fill(name)
```

```
# A tibble: 9 x 3  
  name    time    value  
  <chr> <chr>    <dbl>  
1 robin morning    936  
2 robin afternoon  635  
3 robin evening   438  
4 sam    morning    208  
5 sam    afternoon   92  
6 sam    evening    79  
7 blair  morning    969  
8 blair  afternoon  918  
9 blair  evening   954
```

# A Warning

```
tetris %>%  
  tidyr::fill(name)
```

```
# A tibble: 9 x 3  
  name    time    value  
  <chr> <chr>    <dbl>  
1 robin morning    936  
2 robin afternoon  635  
3 robin evening   438  
4 sam    morning    208  
5 sam    afternoon   92  
6 sam    evening    79  
7 blair  morning    969  
8 blair  afternoon  918  
9 blair  evening   954
```



## DEALING WITH MISSING DATA IN R

**Let's practice!**



DEALING WITH MISSING DATA IN R

# Missing Data dependence

Nicholas Tierney  
Statistician



# Outline

- **MCAR** Missing Completely at Random
- **MAR** Missing At Random
- **MNAR** Missing Not At Random



# MCAR: What is it?

**Missingness has no association with any data you have observed, or not observed.**

test	vacation
NA	TRUE
11.533340	FALSE
10.126115	TRUE
NA	FALSE
NA	TRUE
8.551881	FALSE
NA	FALSE
NA	TRUE
10.608264	TRUE
8.611877	TRUE



# MCAR: What are the implications

## Implications

- Imputation is advisable
- Deleting observations may reduce sample size, limiting inference, but will not bias
- You should be imputing data



# MAR: What is it?

**Missingness depends on data observed, but not data observed**

Implications:

- Impute
- Deleting observations not ideal, may lead to bias

test	vacation	depression
NA	TRUE	87.93109
11.533340	FALSE	40.02708
10.126115	TRUE	48.62883
NA	FALSE	88.21743
NA	TRUE	90.29282
8.551881	FALSE	44.77343
NA	FALSE	89.48865
NA	TRUE	89.99209
10.608264	TRUE	45.56832
8.611877	TRUE	42.41686



# MNAR: What is it?

**Missingness of the response is related to an unobserved value relevant to the assessment of interest.**

Implications:

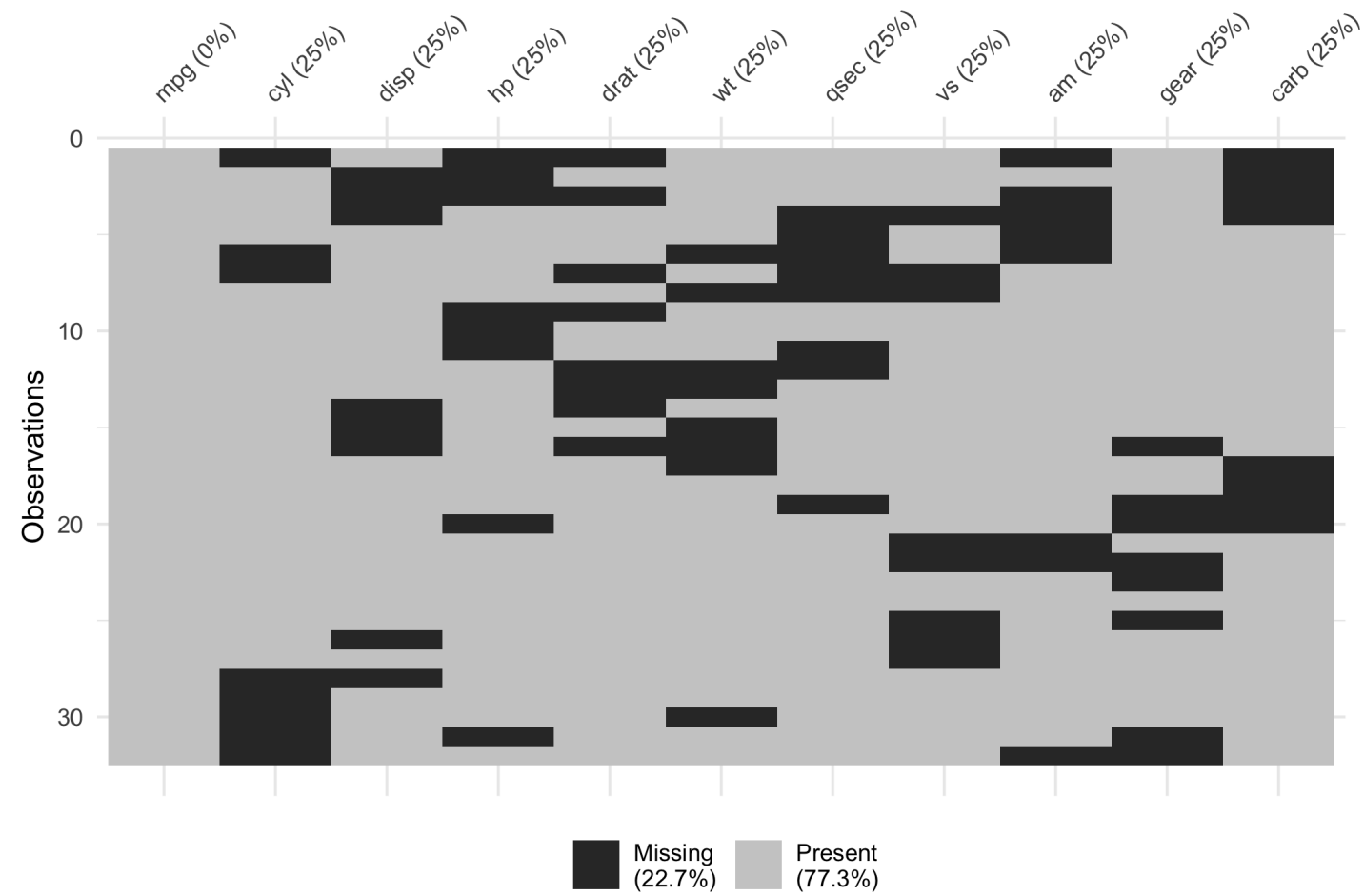
- Data will be biased from deletion and imputation
- Inference can be limited, proceed with caution.

test	vacation	depression
NA	TRUE	NA
11.533340	FALSE	11.533340
10.126115	TRUE	10.126115
NA	FALSE	NA
NA	TRUE	NA
8.551881	FALSE	8.551881
NA	FALSE	NA
NA	TRUE	NA
10.608264	TRUE	10.608264
8.611877	TRUE	8.611877



# Example: MCAR

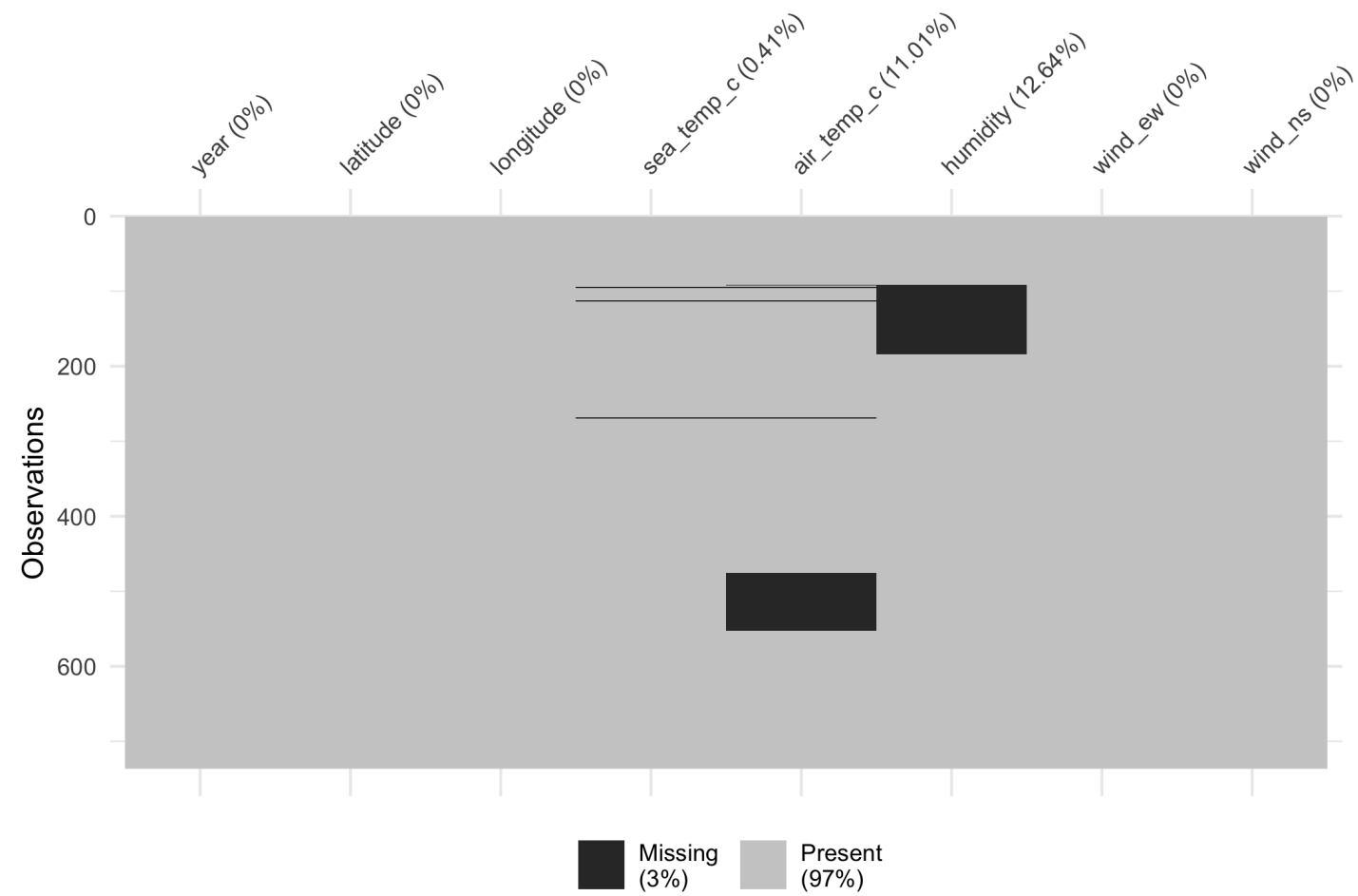
```
vis_miss(mt_cars, cluster = TRUE)
```





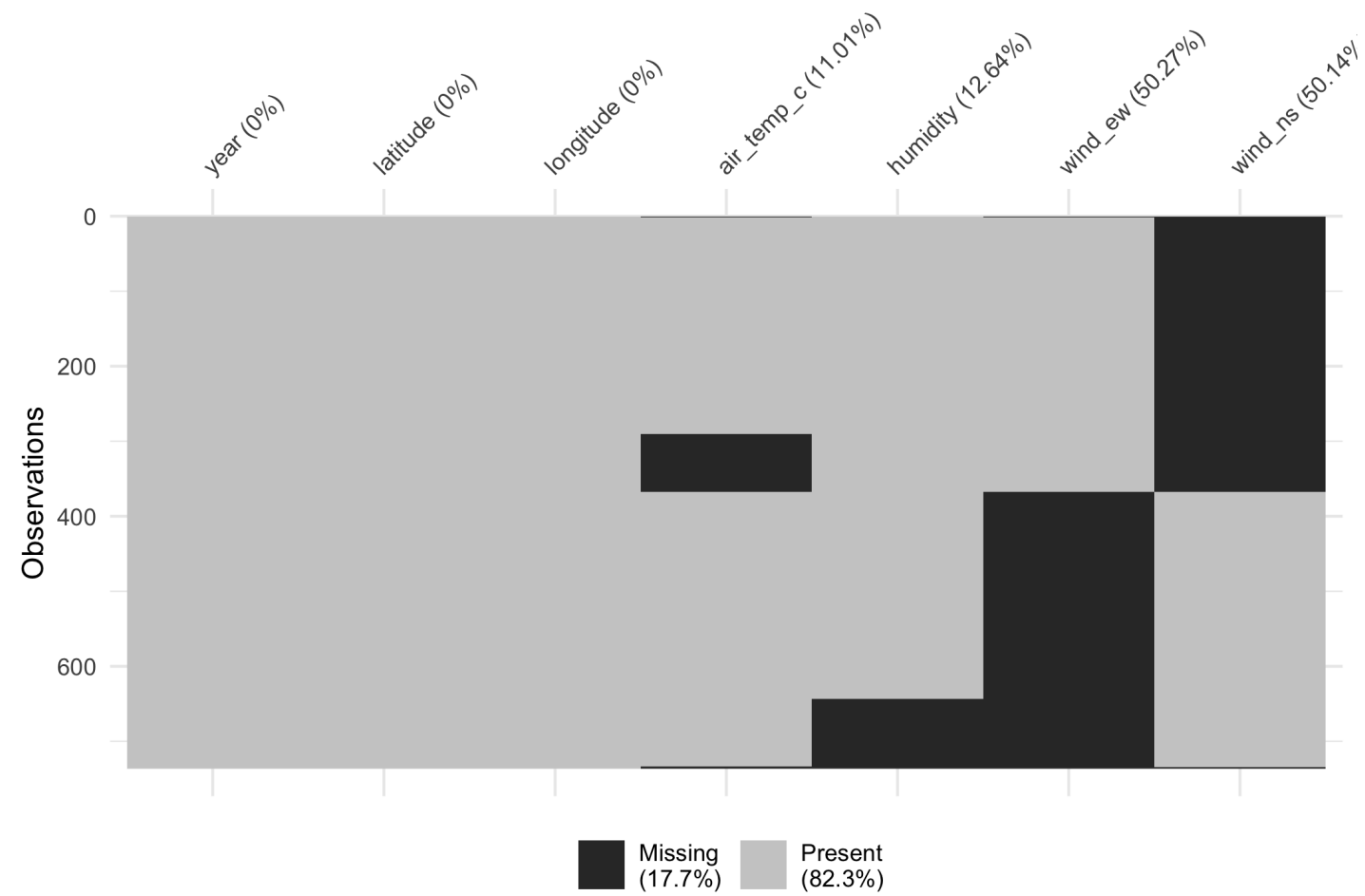
# Example: MAR

```
oceanbuoys %>% arrange(year) %>% vis_miss()
```



# Example: MNAR

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





## DEALING WITH MISSING DATA IN R

**Let's practice!**