



DEALING WITH MISSING DATA IN R

Performing and tracking imputation

Nicholas Tierney
Statistician



Lesson overview

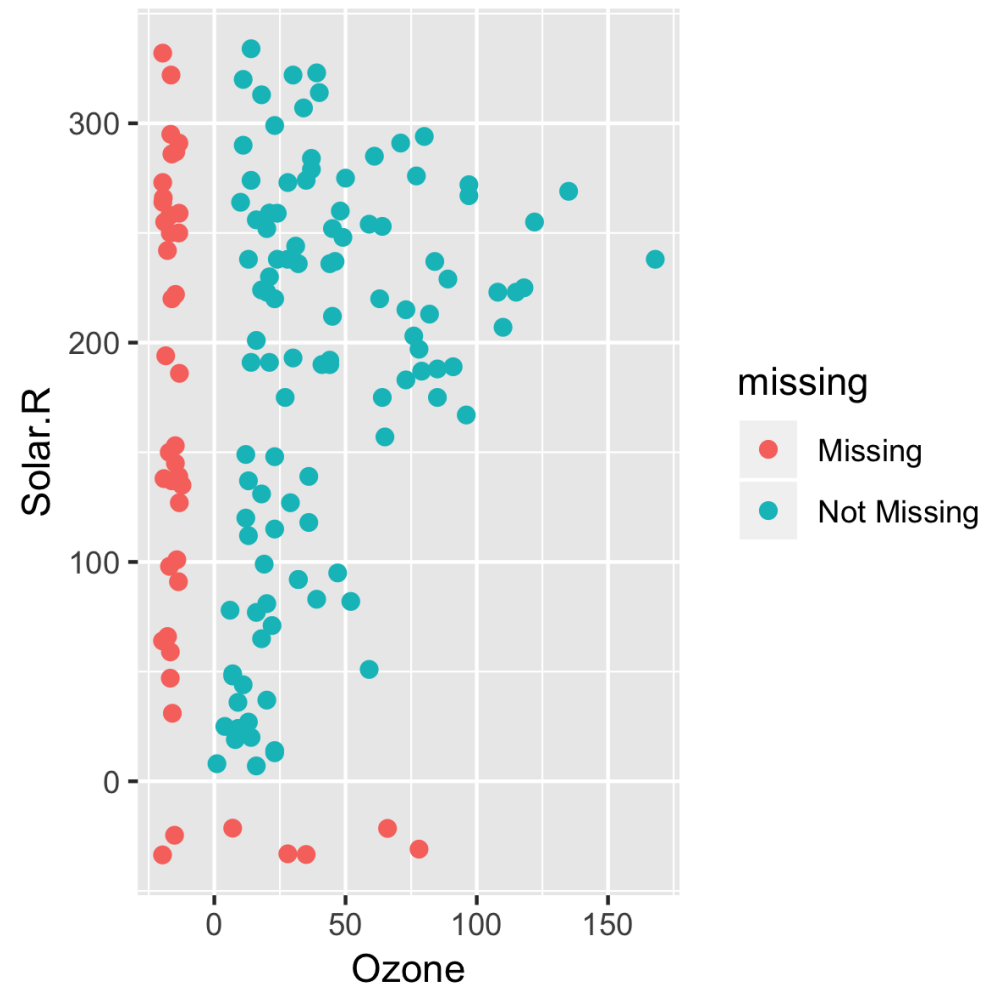
Using imputations to understand data structure

Visualising + exploring imputed values

- Imputing data to explore missingness
- Track missing values
- Visualise imputed values against data



Using imputations to understand data structure



```
> impute_below(c(5, 6, 7, NA, 9, 10))  
[1] 5.00000 6.00000 7.00000  
[4] 4.40271 9.00000 10.00000
```



impute_below

- `impute_below_if()`:

```
impute_below_if(data, is.numeric)
```

- `impute_below_at()`:

```
impute_below_at(data, vars(var1, var2))
```

- `impute_below_all()`:

```
impute_below_all(data)
```

Tracking missing values

```
> df
# A tibble: 6 x 1
  var1
<dbl>
1     5
2     6
3     7
4    NA
5     9
6    10
```

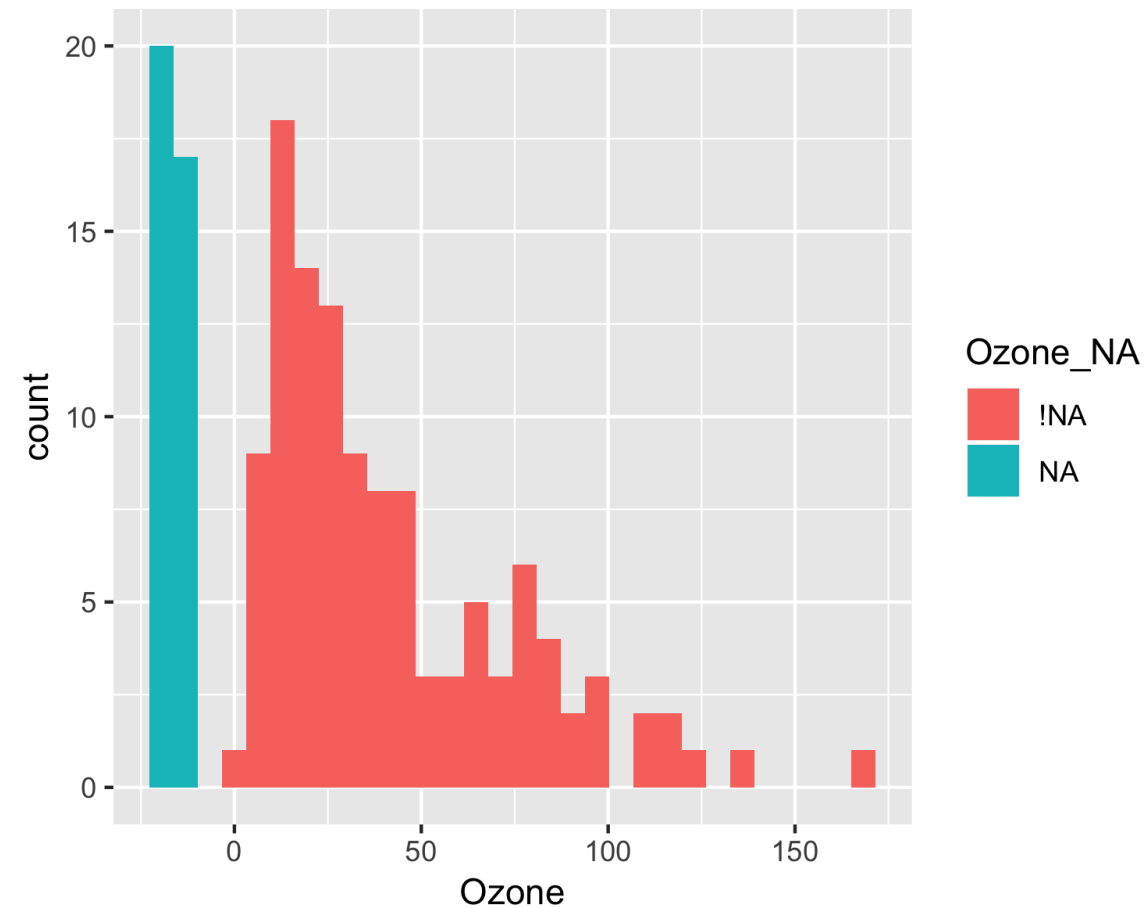
```
> impute_below_all(df)
# A tibble: 6 x 1
  var1
<dbl>
1     5
2     6
3     7
4  4.40
5     9
6    10
```

```
> bind_shadow(df)
# A tibble: 6 x 2
  var1 var1_NA
<dbl> <fct>
1     5    !NA
2     6    !NA
3     7    !NA
4    NA     NA
5     9    !NA
6    10    !NA
```

```
> bind_shadow(df) %>%
  impute_below_all()
# A tibble: 6 x 2
  var1 var1_NA
<dbl> <fct>
1     5    !NA
2     6    !NA
3     7    !NA
4  4.40    NA
5     9    !NA
6    10    !NA
```

Visualise imputed values against data values using histograms

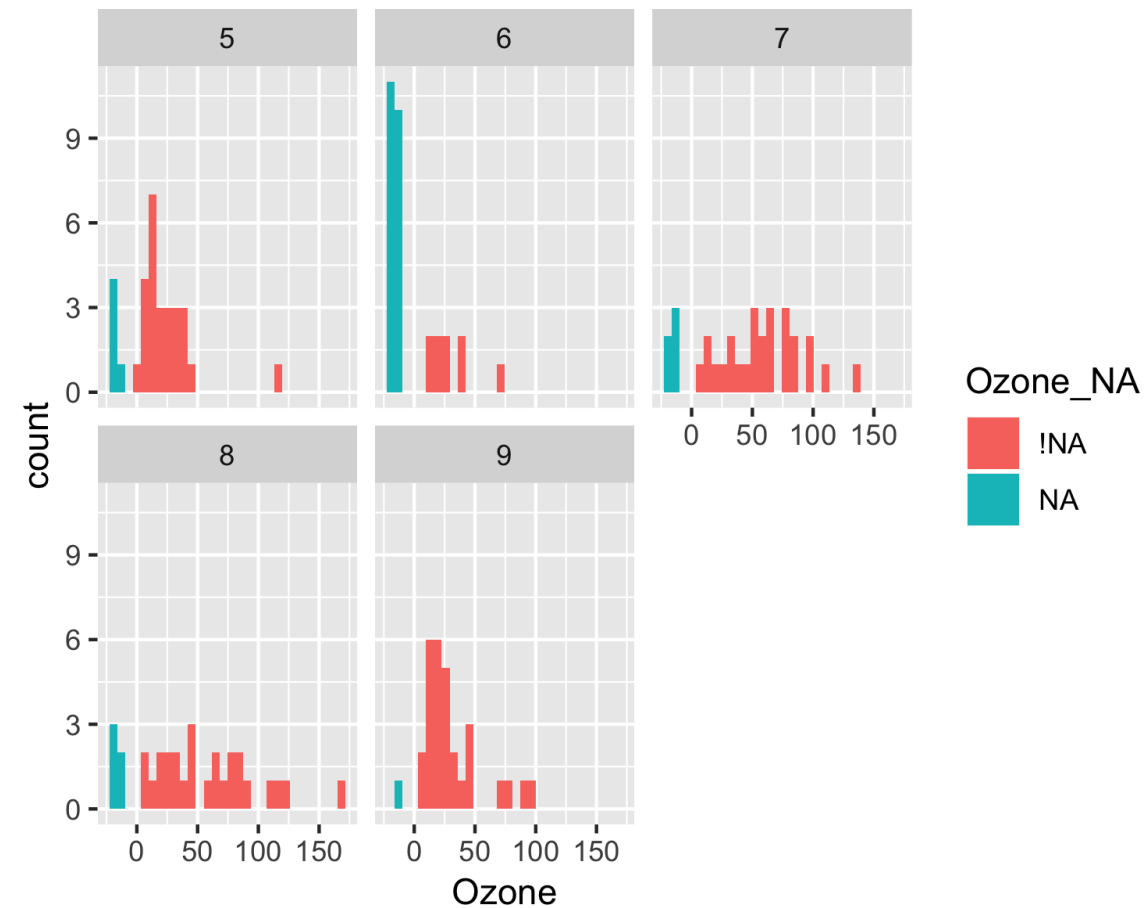
```
> aq_imp <- airquality %>%  
  bind_shadow() %>%  
  impute_below_all()  
  
> ggplot(aq_imp,  
  aes(x = Ozone,  
      fill = Ozone_NA)) +  
  geom_histogram()
```





Visualize imputed values against data values using facets

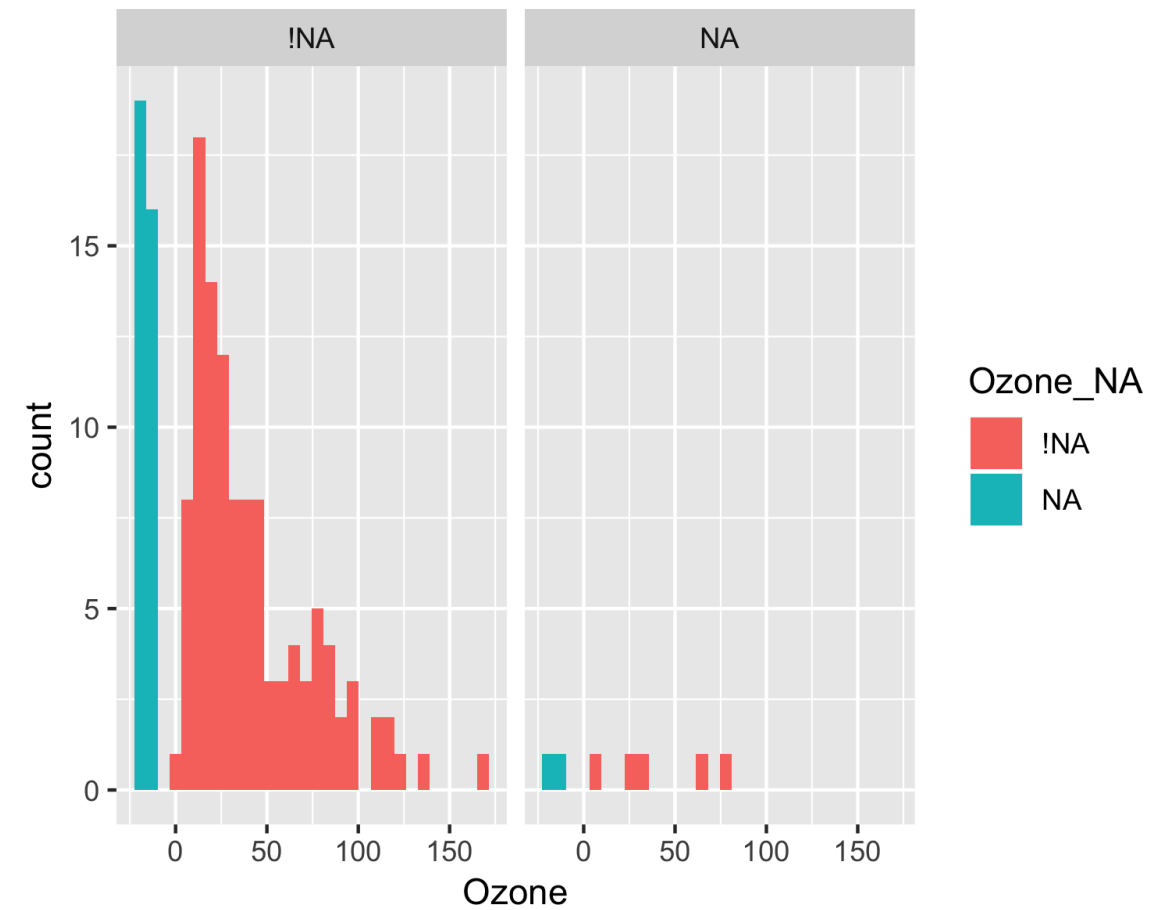
```
ggplot(aq_imp,
       aes(x = Ozone,
           fill = Ozone_NA)) +
  geom_histogram() +
  facet_wrap(~Month)
```





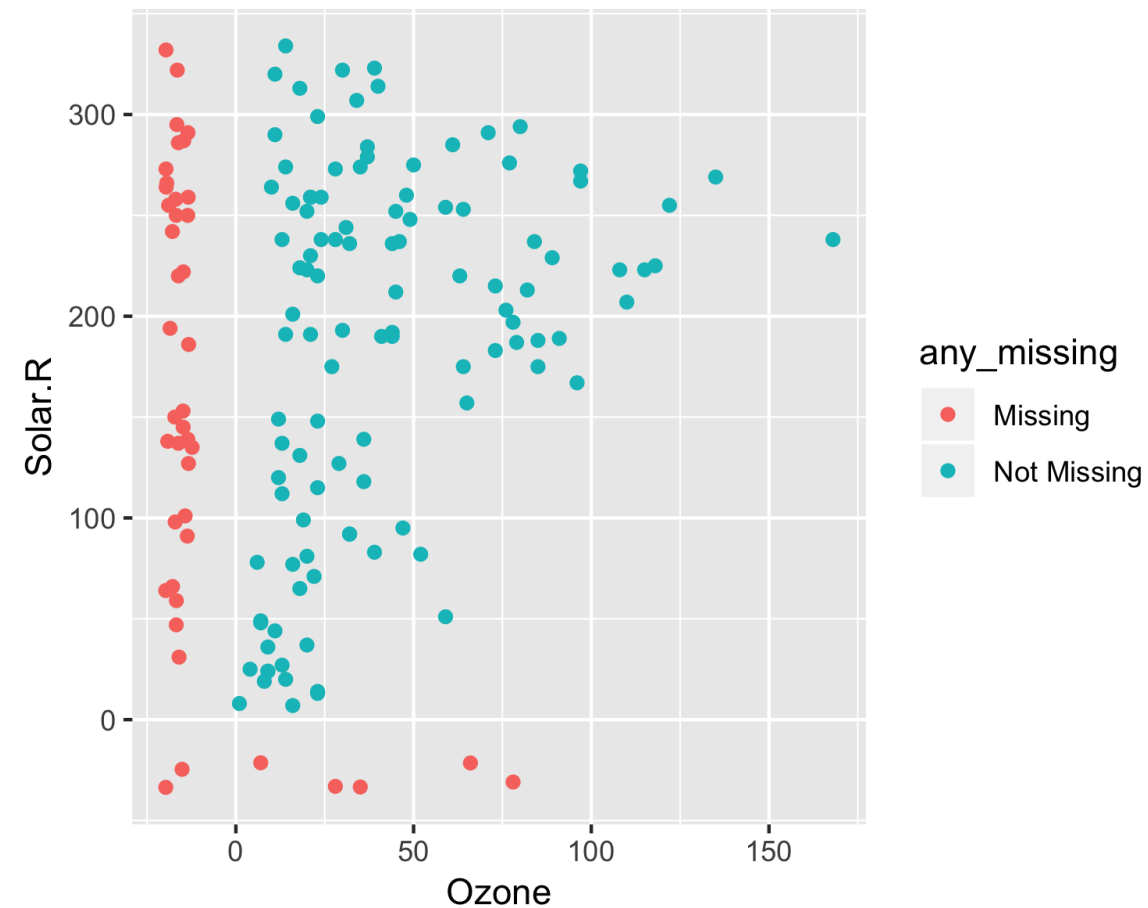
Visualize imputed values using facets

```
ggplot(aq_imp,
       aes(x = Ozone,
           fill = Ozone_NA)) +
  geom_histogram() +
  facet_wrap(~Solar.R_NA)
```



Visualize imputed values against data values using scatterplots

```
aq_imp <- airquality %>%  
  bind_shadow() %>%  
  add_label_missings() %>%  
  impute_below_all()  
  
ggplot(aq_imp,  
       aes(x = Ozone,  
           y = Solar.R,  
           colour = any_missing)) +  
  geom_point()
```





DEALING WITH MISSING DATA IN R

Let's practice!



DEALING WITH MISSING DATA IN R

What makes a good imputation

Nicholas Tierney
Statistician



Lesson overview

- Understand good and bad imputations
- Evaluate missing values:
 - Mean, Scale, Spread
- Using visualisations
 - Boxplots
 - Scatterplots
 - Histograms
 - Many variables



Understanding the good by understanding the bad

```
#> # A tibble: 6 x 1  
#>       x  
#>   <dbl>  
#> 1     1  
#> 2     4  
#> 3     9  
#> 4    16  
#> 5    NA  
#> 6    36
```

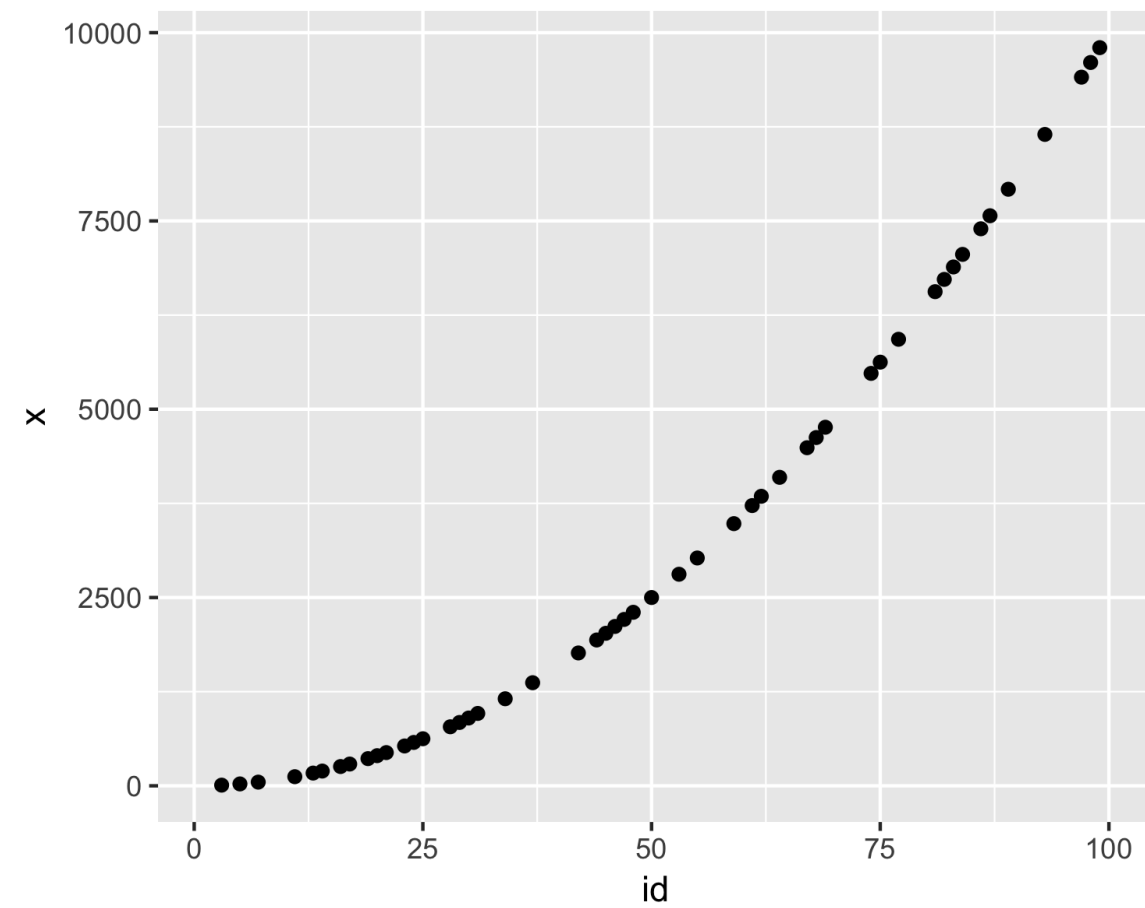
```
> mean(df$x, na.rm = TRUE)  
[1] 13.2
```

```
#> # A tibble: 6 x 1  
#>       x  
#>   <dbl>  
#> 1     1  
#> 2     4  
#> 3     9  
#> 4    16  
#> 5  13.2  
#> 6    36
```

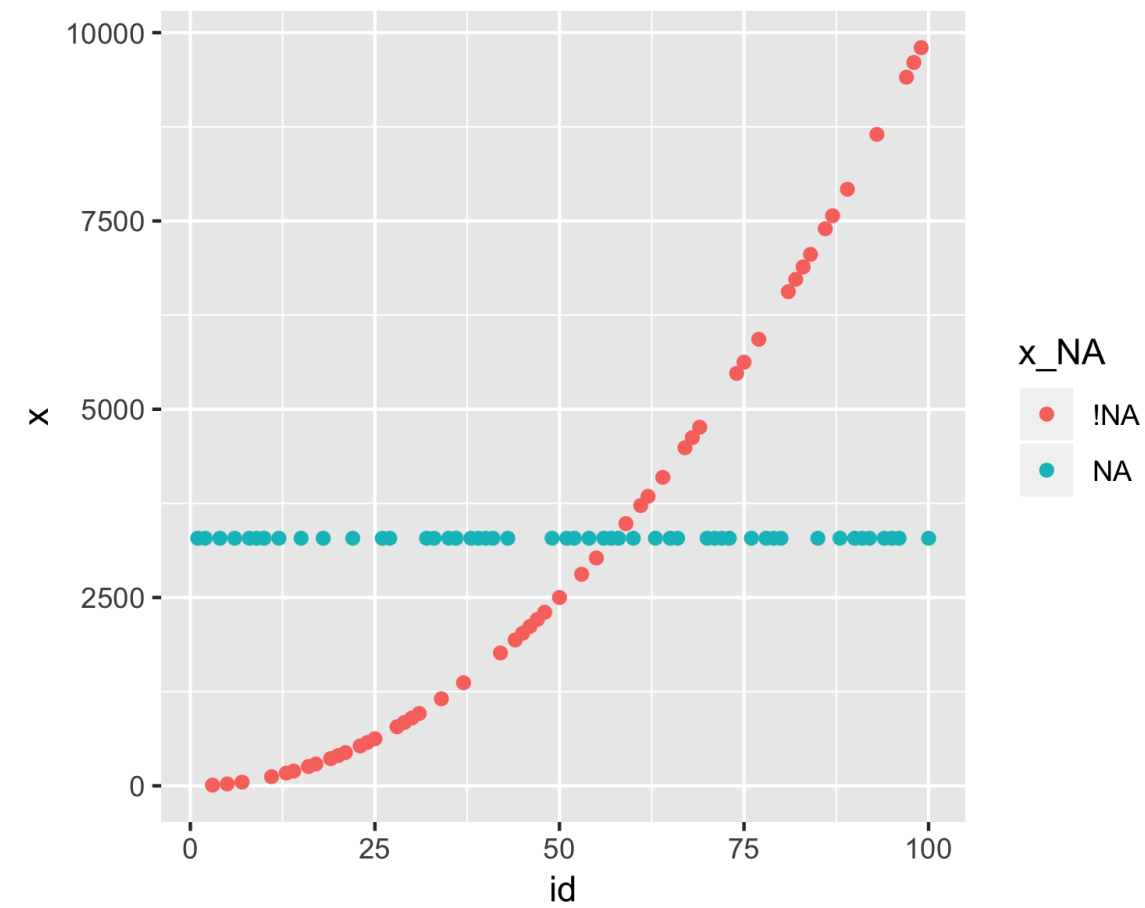


Demonstrating mean imputation

Data with missing values



Data with mean imputations





Explore bad imputations: The mean

- `impute_mean(data$variable)`
- `impute_mean_if(data, is.numeric)`
- `impute_mean_at(data, vars(variable1, variable2))`
- `impute_mean_all(data)`

Tracking missing values

```
aq_impute_mean <- airquality %>%  
  bind_shadow(only_miss = TRUE) %>%  
  impute_mean_all() %>%  
  add_label_shadow()
```

```
aq_impute_mean
```

```
# A tibble: 153 x 9  
  Ozone Solar.R Wind Temp Month Day Ozone_NA Solar.R_NA any_missing  
  <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl> <fct>    <fct>      <chr>  
1    41     190   7.4   67     5     1 !NA      !NA      Not Missing  
2    36     118    8    72     5     2 !NA      !NA      Not Missing  
3    12     149  12.6   74     5     3 !NA      !NA      Not Missing  
4    18     313  11.5   62     5     4 !NA      !NA      Not Missing  
5   42.1    186.  14.3   56     5     5 NA       NA       Missing  
6    28     186.  14.9   66     5     6 !NA      NA       Missing
```




Exploring imputations using a boxplot

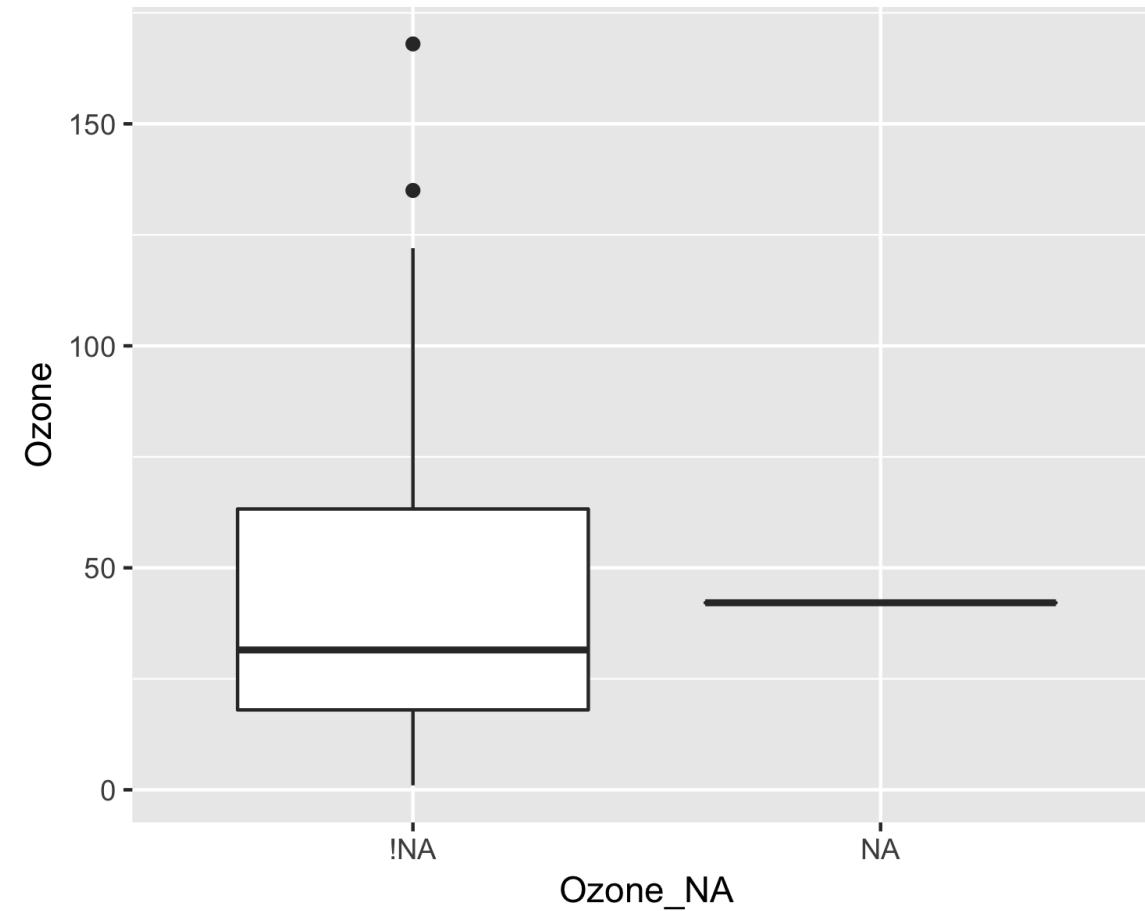
When evaluating imputations, explore changes / similarities in

- **The mean/median** (boxplot)
- The spread
- The scale



Visualizing imputations using the boxplot

```
ggplot(aq_impute_mean,  
      aes(x = Ozone_NA,  
          y = Ozone)) +  
  geom_boxplot()
```

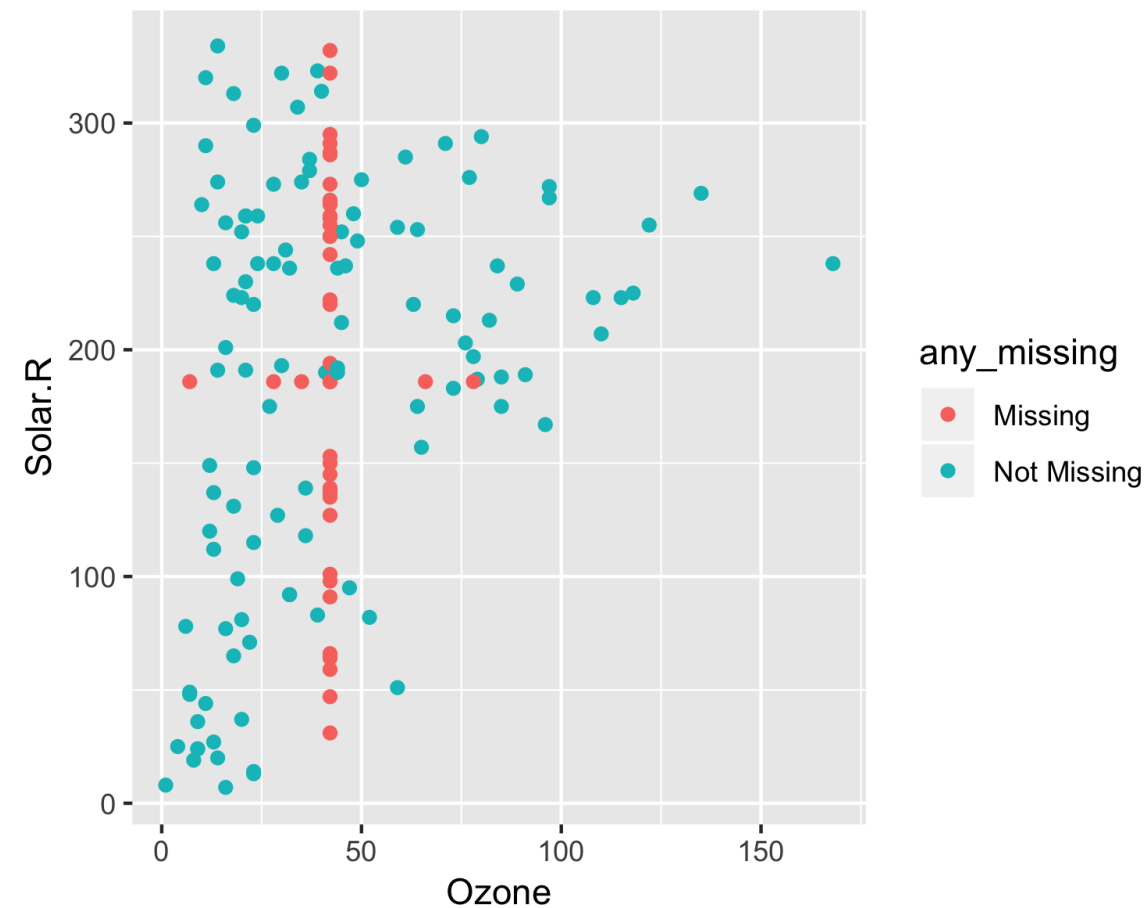


Explore bad imputations using a scatterplot

When evaluating imputations, explore changes/similarities in

- **The spread (scatterplot)**

```
ggplot(aq_impute_mean,  
  aes(x = Ozone,  
      y = Solar.R,  
      colour = any_missing)) +  
  geom_point()
```



Exploring imputations for many variables

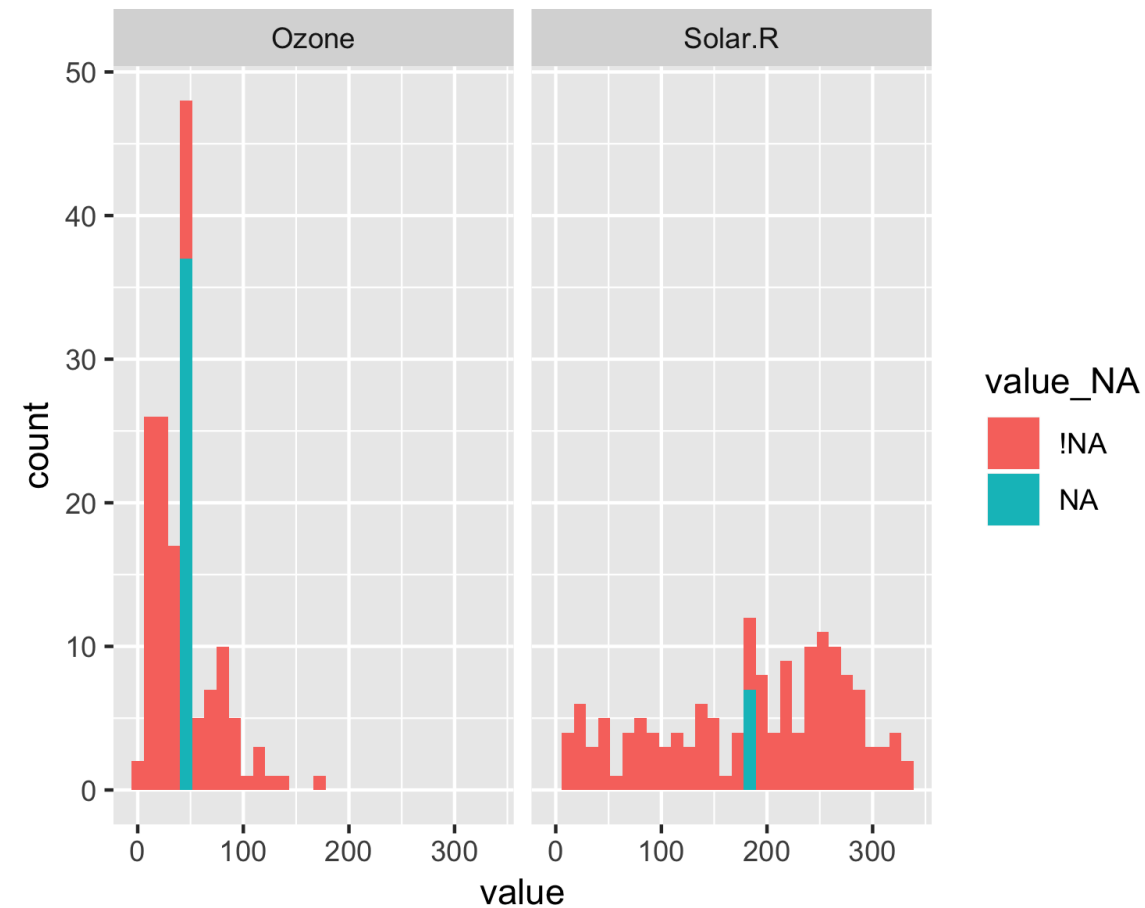
```
aq_imp <- airquality %>%  
  bind_shadow() %>%  
  impute_mean_all()  
  
aq_imp_long <- shadow_long(aq_imp,  
                           Ozone,  
                           Solar.R)  
  
aq_imp_long
```

```
# A tibble: 306 x 4  
  variable value variable_NA value_NA  
  <chr>     <dbl> <chr>      <chr>  
1 Ozone      41 Ozone_NA   !NA  
2 Ozone      36 Ozone_NA   !NA  
3 Ozone      12 Ozone_NA   !NA  
4 Ozone      18 Ozone_NA   !NA  
5 Ozone     42.1 Ozone_NA   NA  
6 Ozone      28 Ozone_NA   !NA  
7 Ozone      23 Ozone_NA   !NA  
8 Ozone      19 Ozone_NA   !NA  
9 Ozone       8 Ozone_NA   !NA  
10 Ozone     42.1 Ozone_NA   NA  
# ... with 296 more rows
```



Exploring imputations for many variables

```
ggplot(aq_imp_long,  
       aes(x = value,  
           fill = value_NA)) +  
  geom_histogram() +  
  facet_wrap(~variable)
```





DEALING WITH MISSING DATA IN R

Let's Practice!



DEALING WITH MISSING DATA IN R

Practicing imputing with different models

Nicholas Tierney
Statistician



Lesson Overview

- Imputation using the `simputation` package
- Use linear model to impute values with `impute_lm`
- Assess new imputations
- Build many imputation models
- Compare imputations across different models and variables

How imputing using a linear model works

```
> df
# A tibble: 5 x 3
      y      x1      x2
  <dbl> <dbl> <dbl>
1  2.67  2.43  3.27
2  3.87  3.55  1.45
3  NA    2.90  1.49
4  5.21  2.72  1.84
5  NA    4.29  1.15
```

```
df %>%
  bind_shadow(only_miss = TRUE) %>%
  add_label_shadow() %>%
  impute_lm(y ~ x1 + x2)
```

```
# A tibble: 5 x 7
      y      x1      x2  y_NA any_missing
  <dbl> <dbl> <dbl> <fct> <chr>
1  2.67  2.43  3.27 !NA    Not Missing
2  3.87  3.55  1.45 !NA    Not Missing
3  5.54  2.90  1.49 NA     Missing
4  5.21  2.72  1.84 !NA    Not Missing
5  2.56  4.29  1.15 NA     Missing
```

Using impute_lm

```
aq_imp_lm <- airquality %>%  
  bind_shadow() %>%  
  add_label_shadow() %>%  
  impute_lm(Solar.R ~ Wind + Temp + Month) %>%  
  impute_lm(Ozone ~ Wind + Temp + Month)
```

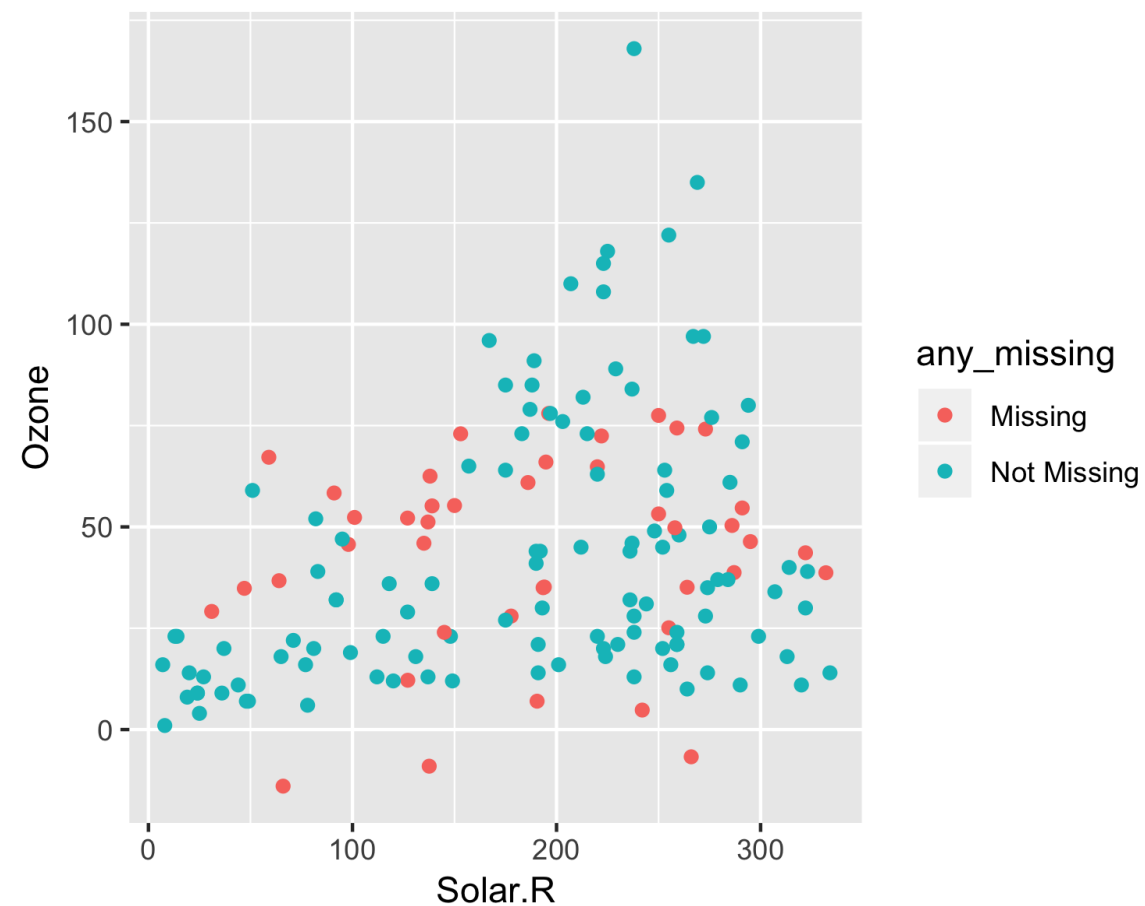
```
aq_imp_lm
```

```
# A tibble: 153 x 13  
  Ozone Solar.R Wind Temp Month Day Ozone_NA Solar.R_NA  
*   <dbl>   <dbl> <dbl> <int> <int> <int> <fct>    <fct>  
1    41     190   7.4    67     5     1 !NA      !NA  
2    36     118    8     72     5     2 !NA      !NA  
3    12     149  12.6    74     5     3 !NA      !NA  
4    18     313  11.5    62     5     4 !NA      !NA  
5   -9.04    138.  14.3    56     5     5 NA       NA  
6    28     178.  14.9    66     5     6 !NA      NA  
# ... with 147 more rows, and 5 more variables: Wind_NA <fct>,  
# Temp_NA <fct>, Month_NA <fct>, Day_NA <fct>,  
# any_missing <chr>
```

Tracking missing values

```
aq_imp_lm <-  
airquality %>%  
  bind_shadow() %>%  
  add_label_missings() %>%  
  impute_lm(Solar.R ~ Wind + Temp +  
            Month) %>%  
  impute_lm(Ozone ~ Wind + Temp +  
            Month)
```

```
ggplot(aq_imp_lm,  
       aes(x = Solar.R,  
           y = Ozone,  
           colour = any_missing)) +  
  geom_point()
```



Evaluating imputations: Evaluating and comparing imputations

```
aq_imp_small <- airquality %>%  
  bind_shadow() %>%  
  impute_lm(Ozone ~ Wind + Temp) %>%  
  impute_lm(Solar.R ~ Wind + Temp) %>%  
  add_label_shadow()  
  
aq_imp_large <- airquality %>%  
  bind_shadow() %>%  
  impute_lm(Ozone ~ Wind + Temp + Month + Day) %>%  
  impute_lm(Solar.R ~ Wind + Temp + Month + Day) %>%  
  add_label_shadow()
```



Evaluating imputations: Binding and visualising many models

```
bound_models <-  
bind_rows(small = aq_imp_small,  
          large = aq_imp_large,  
          .id = "imp_model")
```

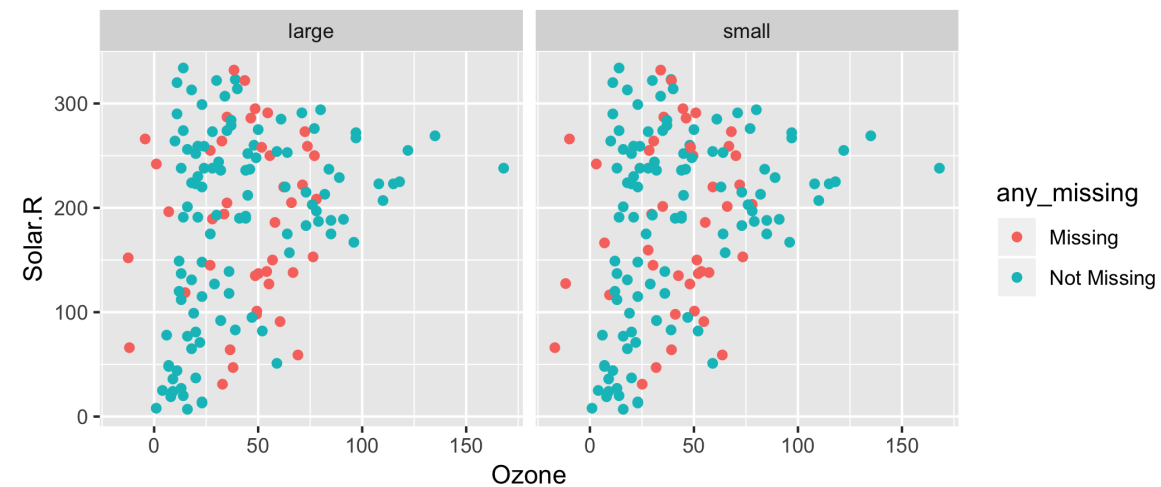
```
bound_models
```

```
  imp_model  Ozone  Solar.R Wind Temp Month Day  
1:    small 41.00000 190.0000  7.4   67     5   1  
2:    small 36.00000 118.0000  8.0   72     5   2  
3:    small 12.00000 149.0000 12.6   74     5   3  
4:    small 18.00000 313.0000 11.5   62     5   4  
5:    small -11.67673 127.4317 14.3   56     5   5  
---  
302:   large 30.00000 193.0000  6.9   70     9  26  
303:   large 26.92183 145.0000 13.2   77     9  27  
304:   large 14.00000 191.0000 14.3   75     9  28  
305:   large 18.00000 131.0000  8.0   76     9  29  
306:   large 20.00000 223.0000 11.5   68     9  30
```



Evaluating imputations: exploring many imputations

```
ggplot(bound_models,  
       aes(x = Ozone,  
           y = Solar.R,  
           colour = any_missing)) +  
  geom_point() +  
  facet_wrap(~imp_model)
```





Explore imputations in multiple variables and models

```
bound_models_gather <- bound_models %>%  
  select(Ozone, Solar.R,  
         any_missing, imp_model) %>%  
  gather(key = "variable", value = "value",  
         -any_missing, -imp_model)
```

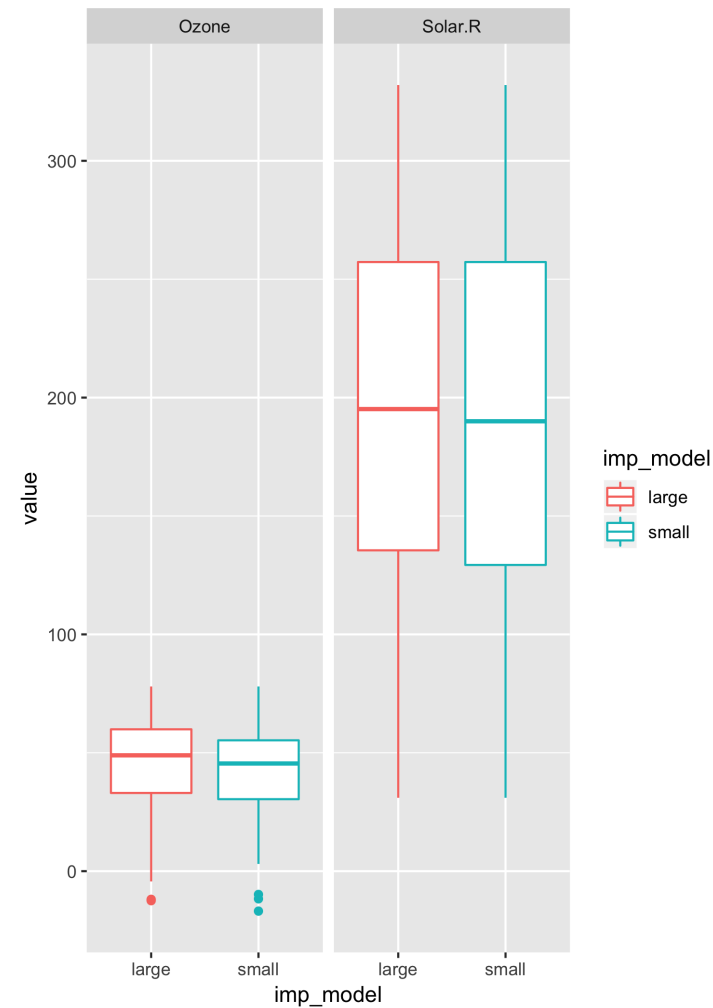
```
bound_models_gather
```

```
   any_missing imp_model variable    value  
1: Not Missing    small    Ozone  41.00000  
2: Not Missing    small    Ozone  36.00000  
3: Not Missing    small    Ozone  12.00000  
4: Not Missing    small    Ozone  18.00000  
5:      Missing    small    Ozone -11.67673  
---  
608: Not Missing   large  Solar.R 193.00000  
609:      Missing   large  Solar.R 145.00000  
610: Not Missing   large  Solar.R 191.00000  
611: Not Missing   large  Solar.R 131.00000  
612: Not Missing   large  Solar.R 223.00000
```



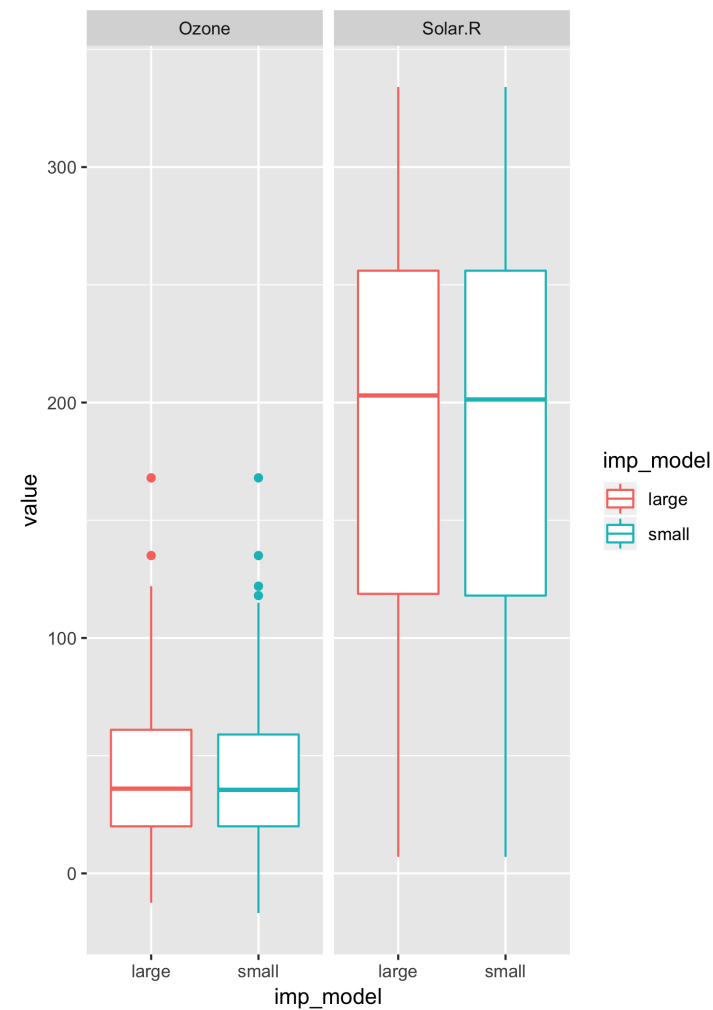
Explore imputations in multiple variables and models

```
ggplot(bound_models_gather,  
       aes(x = imp_model,  
           y = value)) +  
  geom_boxplot() +  
  facet_wrap(~key)
```



Explore imputations in multiple variables and models

```
bound_models_gather %>%  
  filter(any_missing == "Missing") %>%  
  ggplot(aes(x = imp_model,  
            y = value)) +  
  geom_boxplot() +  
  facet_wrap(~key)
```





DEALING WITH MISSING DATA IN R

Let's practice!



DEALING WITH MISSING DATA IN R

Assessing inference from imputed data in a modelling context

Nicholas Tierney
Statistician



Exploring parameters of one model

```
lm(Temp ~ Ozone + Solar.R + Wind + Month + day, data = airquality)
```

1. Complete case analysis
2. Imputation using the imputed data from the last lesson

Combining the datasets together

```
#1. Complete cases
aq_cc <- airquality %>%
  na.omit() %>%
  bind_shadow() %>%
  add_label_shadow()
```

```
#2. Imputation using the imputed data from the last lesson
aq_imp_lm <- bind_shadow(airquality) %>%
  add_label_shadow() %>%
  impute_lm(Ozone ~ Temp + Wind + Month + Day) %>%
  impute_lm(Solar.R ~ Temp + Wind + Month + Day)
```

```
# 3. Bind the models together
bound_models <- bind_rows(cc = aq_cc,
                          imp_lm = aq_imp_lm,
                          .id = "imp_model")
```

Combining the datasets together

```
bound_models
```

imp_model	Ozone	Solar.R	Wind	Temp	Month	Day	Ozone_NA	Solar.R_NA	any_missing
cc	41	190	7.4	67	5	1	!NA	!NA	Not Missing
cc	36	118	8.0	72	5	2	!NA	!NA	Not Missing
cc	12	149	12.6	74	5	3	!NA	!NA	Not Missing
cc	18	313	11.5	62	5	4	!NA	!NA	Not Missing
cc	23	299	8.6	65	5	7	!NA	!NA	Not Missing
imp_lm	30	193	6.9	70	9	26	!NA	!NA	Not Missing
imp_lm	NA	145	13.2	77	9	27	NA	!NA	Missing
imp_lm	14	191	14.3	75	9	28	!NA	!NA	Not Missing
imp_lm	18	131	8.0	76	9	29	!NA	!NA	Not Missing
imp_lm	20	223	11.5	68	9	30	!NA	!NA	Not Missing

Exploring the models

```
model_summary <- bound_models %>%  
  group_by(imp_model) %>%  
  nest() %>%  
  mutate(mod = map(data,  
                    ~lm(Temp ~ Ozone + Solar.R + Wind + Temp + Days + Month  
                        data = .)),  
         res = map(mod, residuals),  
         pred = map(mod, predict),  
         tidy = map(mod, broom::tidy))
```

```
model_summary
```

```
# A tibble: 2 x 6  
  imp_model data          mod      res      pred      tidy  
  <chr>      <list>          <list> <list>    <list>    <list>  
1 cc        <tibble [111 x 13]> <S3: lm> <dbl [111]> <dbl [111]> <tibble [3 x 5]  
2 imp_lm    <tibble [153 x 13]> <S3: lm> <dbl [153]> <dbl [153]> <tibble [3 x 5]
```

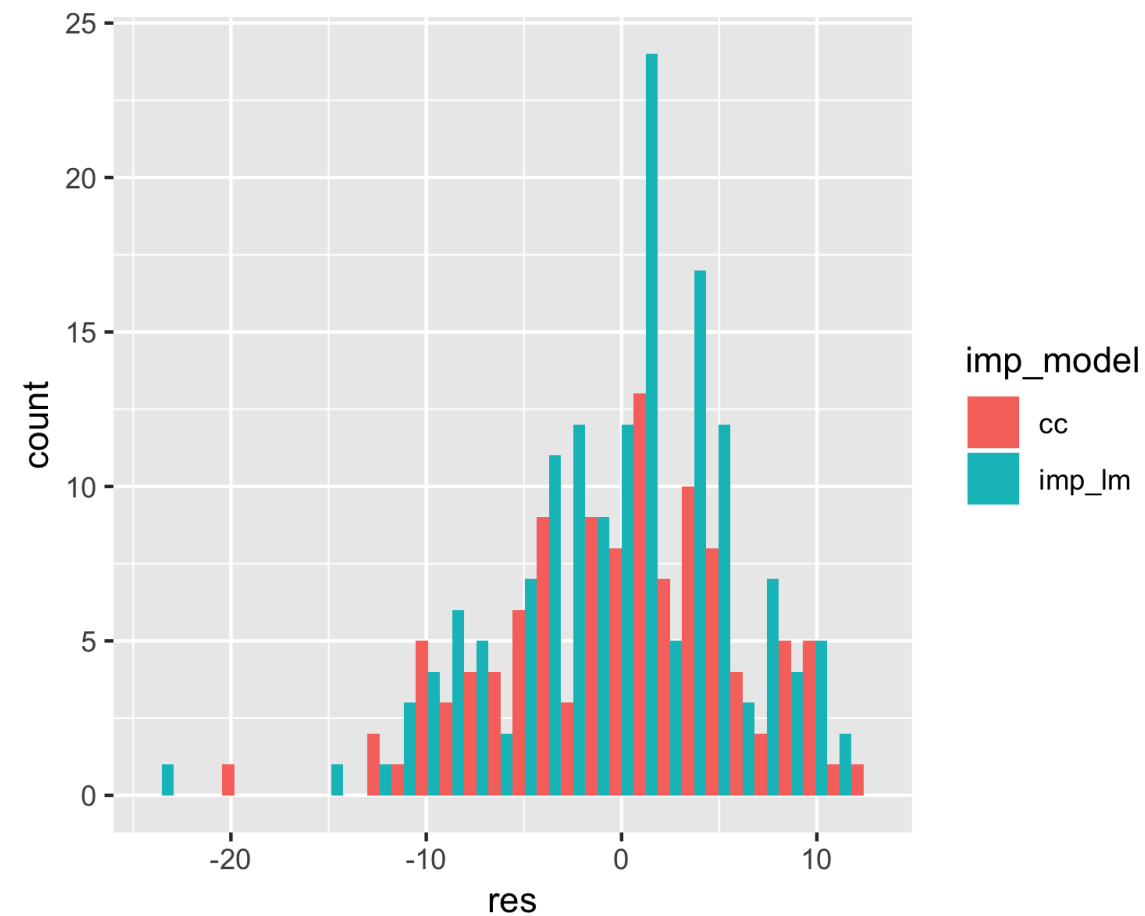
Exploring coefficients of multiple models

```
model_summary %>%  
  select(imp_model,  
         tidy) %>%  
  unnest()
```

```
# A tibble: 6 x 6  
  imp_model term      estimate std.error statistic  p.value  
  <chr>      <chr>      <dbl>    <dbl>    <dbl>    <dbl>  
1 cc        (Intercept) 68.5      1.53     44.8 1.31e-71  
2 cc        Ozone        0.194    0.0210     9.26 2.22e-15  
3 cc        Solar.R      0.00604   0.00766     0.789 4.32e- 1  
4 imp_lm    (Intercept) 67.2      1.30     51.5 2.68e-97  
5 imp_lm    Ozone        0.215    0.0180    12.0 1.40e-23  
6 imp_lm    Solar.R      0.00787   0.00630     1.25 2.13e- 1
```

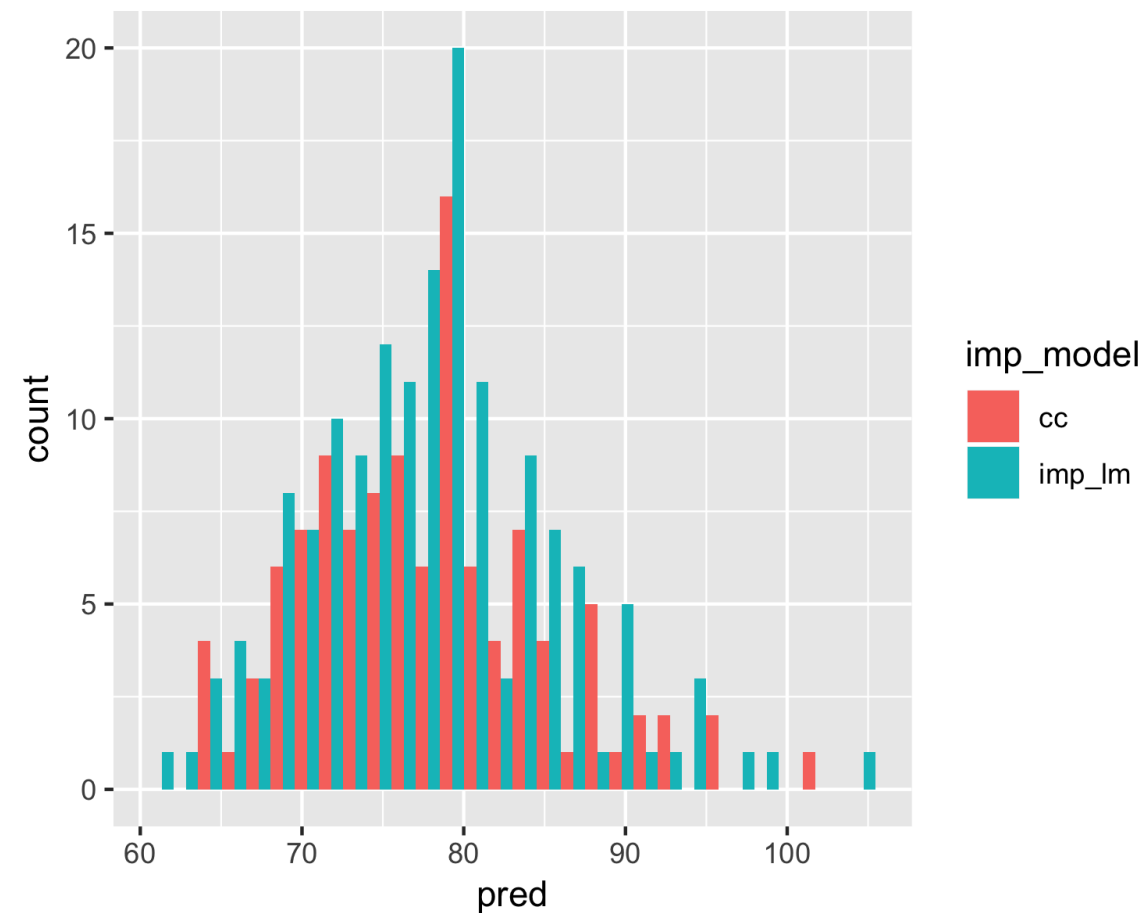

Exploring residuals of multiple models

```
model_summary %>%  
  select(imp_model,  
         res) %>%  
  unnest() %>%  
  ggplot(aes(x = res,  
            fill = imp_model)) +  
  geom_histogram(position = "dodge")
```



Exploring predictions of multiple models

```
model_summary %>%  
  select(imp_model,  
         pred) %>%  
  unnest() %>%  
  ggplot(aes(x = pred,  
             fill = imp_model)) +  
  geom_histogram(position = "dodge")
```





DEALING WITH MISSING DATA IN R

Let's practice!