Introduction to the tidyverse

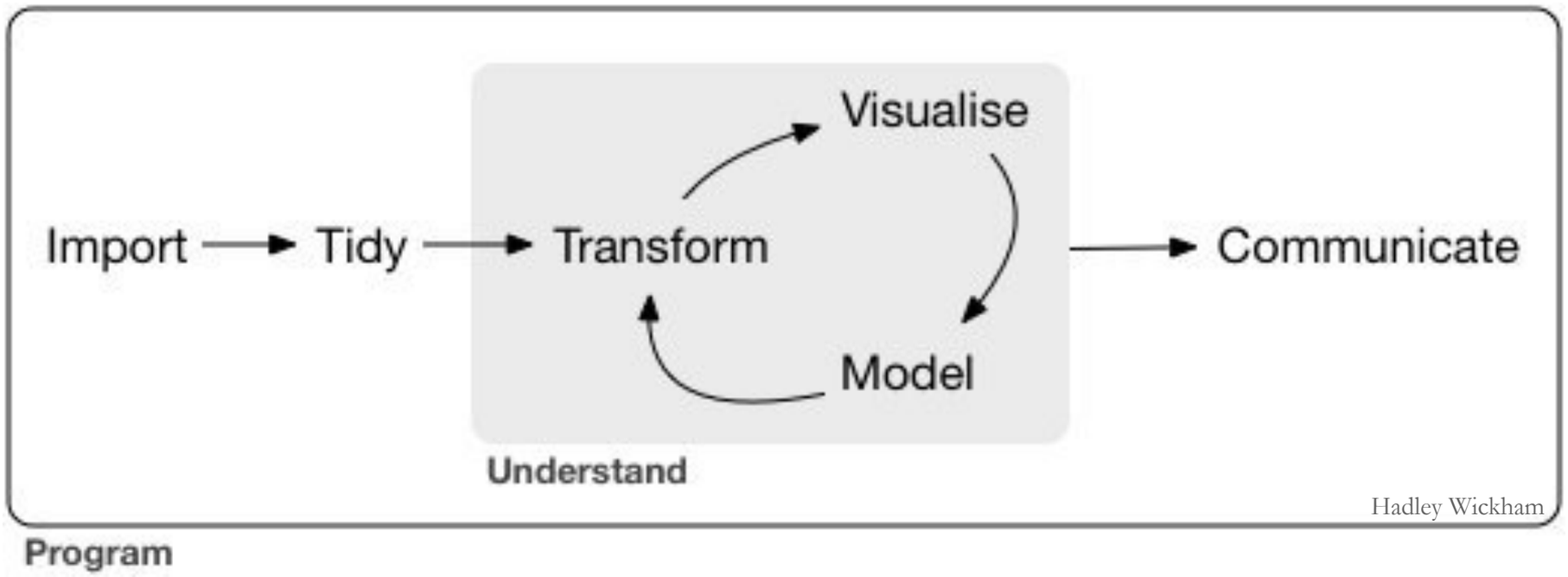
Gargi Datta

thedattadoctor

gargi-datta

www.gargidatta.com

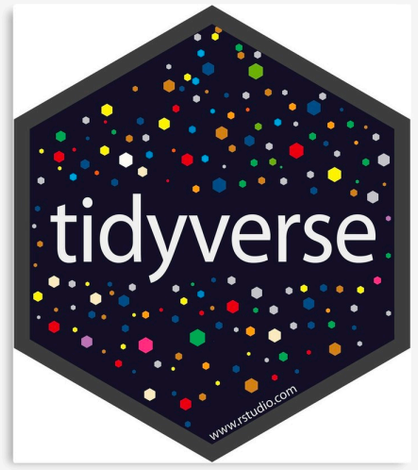
The Machine Learning Lifecycle



Tidy Data:

1. **Variable make up the columns**
2. **Observations make up the rows**
3. **Values go into cells**

The “tidyverse”



- Collection of R-packages designed for data science
  - Share a common data representation and API design

Tidy API principles

- Reuse existing data structures
- Compose simple functions with the pipe
- Embrace functional programming
- Design for humans

The “tidyverse”

Import	Tidy	Transform	Visualize	Model	Communicate
readr, haven, readxl	tidyr	dplyr, purrr	ggplot2	broom, tidymodels	knitr, rmarkdown, Shiny

%>%, tibble, forcats, stringr

Installing the tidyverse

- Have R-studio and R installed.
- install.packages("tidyverse")
- library(tidyverse)

magrittr

- install.packages("magrittr")
- library(magrittr)

%>%

```
foo_foo <- little_bunny()

bop_on(
  scoop_up(
    hop_through(foo_foo, forest),
    field_mouse
  ),
  head
)

# vs
foo_foo %>%
  hop_through(forest) %>%
  scoop_up(field_mouse) %>%
  bop_on(head)
```

Hadley Wickham

Datasets

- data()
- data("airquality")

```
> head(airquality, 5)
  Ozone Solar.R Wind Temp Month Day
1    41     190  7.4   67     5   1
2    36     118  8.0   72     5   2
3    12     149 12.6   74     5   3
4    18     313 11.5   62     5   4
5    NA      NA 14.3   56     5   5
> nrow(airquality)
[1] 153
> dim(airquality)
[1] 153   6
> class(airquality)
[1] "data.frame"
```

Summarize the data

- %>%

```
> airquality %>% summary()
     Ozone           Solar.R         Wind             Temp           Month            Day
 Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00   Min.   :5.000   Min.   : 1.0
 1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00   1st Qu.:6.000   1st Qu.: 8.0
 Median : 31.50   Median :205.0   Median : 9.700   Median :79.00   Median :7.000   Median :16.0
 Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88   Mean   :6.993   Mean   :15.8
 3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00   3rd Qu.:8.000   3rd Qu.:23.0
 Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00   Max.   :9.000   Max.   :31.0
 NA's   :37       NA's   :7
```

airquality %>% summary()

Noun

Verb

Summarize the data

- %>%

```
> airquality %>% summary()
     Ozone           Solar.R          Wind             Temp           Month            Day
 Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00   Min.   :5.000   Min.   : 1.0
 1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00   1st Qu.:6.000   1st Qu.: 8.0
 Median : 31.50   Median :205.0   Median : 9.700   Median :79.00   Median :7.000   Median :16.0
 Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88   Mean   :6.993   Mean   :15.8
 3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00   3rd Qu.:8.000   3rd Qu.:23.0
 Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00   Max.   :9.000   Max.   :31.0
 NA's   :37       NA's   :7
```

- tidyr::drop_na()

```
> airquality %>% drop_na() %>% summary()
     Ozone           Solar.R          Wind             Temp           Month            Day
 Min.   :  1.0   Min.   :  7.0   Min.   : 2.30   Min.   :57.00   Min.   :5.000   Min.   : 1.00
 1st Qu.: 18.0   1st Qu.:113.5   1st Qu.: 7.40   1st Qu.:71.00   1st Qu.:6.000   1st Qu.: 9.00
 Median : 31.0   Median :207.0   Median : 9.70   Median :79.00   Median :7.000   Median :16.00
 Mean   : 42.1   Mean   :184.8   Mean   : 9.94   Mean   :77.79   Mean   :7.216   Mean   :15.95
 3rd Qu.: 62.0   3rd Qu.:255.5   3rd Qu.:11.50   3rd Qu.:84.50   3rd Qu.:9.000   3rd Qu.:22.50
 Max.   :168.0   Max.   :334.0   Max.   :20.70   Max.   :97.00   Max.   :9.000   Max.   :31.00
```

Exercise 1:
Replace NA's in data using tidyr

- Replace NA's in the column Ozone with 0 and summarize
- tidyr::replace_na(): The replace argument is a list specified as:
  - list(column_name = replace_value) [eg. list(Ozone = 0)]
  - To access help, do **?replace_na()**

```
> airquality %>% replace_na(list(Ozone=0)) %>% summary()
     Ozone           Solar.R           Wind             Temp           Month            Day       
 Min.   :  0.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00   Min.   :5.000   Min.   : 1.0  
 1st Qu.:  4.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00   1st Qu.:6.000   1st Qu.: 8.0  
 Median : 21.00   Median :205.0   Median : 9.700   Median :79.00   Median :7.000   Median :16.0  
 Mean   : 31.94   Mean   :185.9   Mean   : 9.958   Mean   :77.88   Mean   :6.993   Mean   :15.8  
 3rd Qu.: 46.00   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00   3rd Qu.:8.000   3rd Qu.:23.0  
 Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00   Max.   :9.000   Max.   :31.0  
                  NA's   :7
```

tibbles

- as_tibble()

```
> aqt ← as_tibble(airquality)
> aqt
# A tibble: 153 x 6
   Ozone Solar.R  Wind  Temp Month   Day
   <int>   <int> <dbl> <int> <int> <int>
 1    41     190   7.4    67     5     1
 2    36     118   8      72     5     2
 3    12     149  12.6    74     5     3
 4    18     313  11.5    62     5     4
 5    NA      NA  14.3    56     5     5
 6    28      NA  14.9    66     5     6
 7    23     299   8.6    65     5     7
 8    19      99  13.8    59     5     8
 9     8      19  20.1    61     5     9
10    NA     194   8.6    69     5    10
# … with 143 more rows
```

Subsetting in tibbles

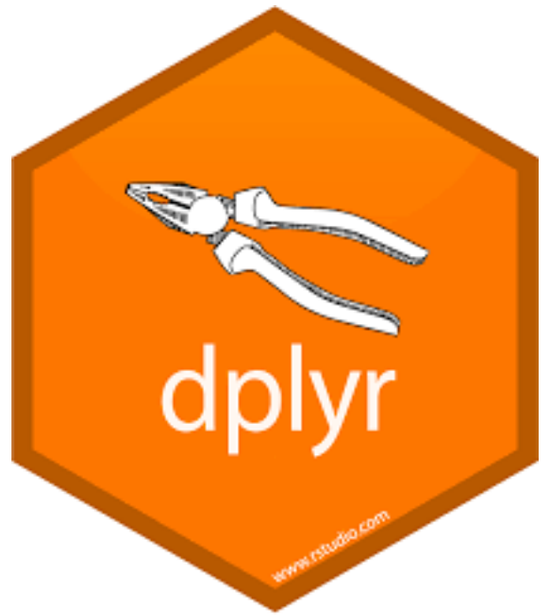
```
> airquality$yea
NULL
> aqt$yea
NULL
Warning message:
Unknown or uninitialised column: 'yea'.
```

Try aqt$wind

Wrangling data

- dplyr!!
  - mutate()
  - select() and pull()
  - filter()
  - arrange()
  - summarise()
  - group_by()

Verbs



dplyr

- filter()

```
> aqt %>% filter(Month == 6)
# A tibble: 30 x 6
   Ozone Solar.R  Wind  Temp Month   Day
   <int>   <int> <dbl> <int> <int> <int>
 1    NA     286   8.6    78     6     1
 2    NA     287   9.7    74     6     2
 3    NA     242  16.1    67     6     3
 4    NA     186   9.2    84     6     4
 5    NA     220   8.6    85     6     5
 6    NA     264  14.3    79     6     6
 7    29     127   9.7    82     6     7
 8    NA     273   6.9    87     6     8
 9    71     291  13.8    90     6     9
10    39     323  11.5    87     6    10
# … with 20 more rows
```

dplyr

- select() and pull()

```
> aqt %>% filter(Month == 6) %>% select(Ozone)
# A tibble: 30 x 1
   Ozone
   <int>
 1    NA
 2    NA
 3    NA
 4    NA
 5    NA
 6    NA
 7    29
 8    NA
 9    71
10    39
# … with 20 more rows
> aqt %>% filter(Month == 6) %>% pull(Ozone)
 [1] NA NA NA NA NA NA 29 NA 71 39 NA NA 23 NA NA 21 37 20 12 13 NA NA NA NA NA NA NA NA NA NA
```

Exercise 2:
Pull down a variable and look at the summary

- No filters, pull Ozone
  - See the summary of the results
- Next, fil
  - See s

```
> aqt %>% filter(Day == 1) %>%
+     pull(Ozone)
[1]   41   NA 135   39   96
```

```
> aqt %>%
+   pull(Ozone) %>%
+   summary()
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   1.00   18.00   31.50   42.13   63.25  168.00      37
>
> aqt %>% filter(Day == 1) %>%
+   pull(Ozone) %>%
+   summary()
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  39.00   40.50   68.50   77.75  105.75  135.00       1
```

Why are these different?

dplyr

- mutate()

What's this?!!!

```
> aqt %<>% mutate(TempC = (Temp - 32) * 5/9)
> aqt
# A tibble: 153 x 7
   Ozone Solar.R  Wind  Temp Month   Day TempC
   <int>   <int> <dbl> <int> <int> <int> <dbl>
 1    41     190   7.4    67     5     1  19.4
 2    36     118   8      72     5     2  22.2
 3    12     149  12.6    74     5     3  23.3
 4    18     313  11.5    62     5     4  16.7
 5    NA      NA  14.3    56     5     5  13.3
 6    28      NA  14.9    66     5     6  18.9
 7    23     299   8.6    65     5     7  18.3
 8    19      99  13.8    59     5     8  15
 9     8      19  20.1    61     5     9  16.1
10    NA     194   8.6    69     5    10  20.6
# … with 143 more rows
```

Exercise 3:
Create a new variable

- Create a variable that is Ozone/Wind, using mutate()

```
> aqt %<>% mutate(OzoneByWind = Ozone/Wind)
> aqt
# A tibble: 153 x 8
   Ozone Solar.R  Wind  Temp Month   Day TempC OzoneByWind
   <int>   <int> <dbl> <int> <int> <int> <dbl>       <dbl>
 1    41     190   7.4    67     5     1  19.4        5.54
 2    36     118   8      72     5     2  22.2        4.5
 3    12     149  12.6    74     5     3  23.3        0.952
 4    18     313  11.5    62     5     4  16.7        1.57
 5    NA      NA  14.3    56     5     5  13.3       NA
 6    28      NA  14.9    66     5     6  18.9        1.88
 7    23     299   8.6    65     5     7  18.3        2.67
 8    19      99  13.8    59     5     8  15          1.38
 9     8      19  20.1    61     5     9  16.1        0.398
10    NA     194   8.6    69     5    10  20.6       NA
# … with 143 more rows
```

dplyr

- arrange()

```
> aqt %>% drop_na() %>% arrange(Ozone)
# A tibble: 111 x 7
   Ozone Solar.R  Wind  Temp Month   Day TempC
   <int>   <int> <dbl> <int> <int> <int> <dbl>
 1     1       8   9.7    59     5    21  15
 2     4      25   9.7    61     5    23  16.1
 3     6      78  18.4    57     5    18  13.9
 4     7      48  14.3    80     7    15  26.7
 5     7      49  10.3    69     9    24  20.6
 6     8      19  20.1    61     5     9  16.1
 7     9      24  13.8    81     8     2  27.2
 8     9      36  14.3    72     8    22  22.2
 9     9      24  10.9    71     9    14  21.7
10    10     264  14.3    73     7    12  22.8
# … with 101 more rows
```

```
> aqt %>% drop_na() %>% arrange(Ozone) %>% tail()
# A tibble: 6 x 7
  Ozone Solar.R  Wind  Temp Month   Day TempC
  <int>   <int> <dbl> <int> <int> <int> <dbl>
1   110     207   8      90     8     9  32.2
2   115     223   5.7    79     5    30  26.1
3   118     225   2.3    94     8    29  34.4
4   122     255   4      89     8     7  31.7
5   135     269   4.1    84     7     1  28.9
6   168     238   3.4    81     8    25  27.2
```

dplyr

- summarise() and group_by()

```
> aqt %>% summarise(mean_temp = mean(Temp))
# A tibble: 1 x 1
  mean_temp
      <dbl>
1      77.9
```

```
> aqt %>%
+   group_by(Month) %>%
+   summarise(mean_temp = mean(Temp))
# A tibble: 5 x 2
  Month mean_temp
  <int>     <dbl>
1     5      65.5
2     6      79.1
3     7      83.9
4     8      84.0
5     9      76.9
```

Exercise 4: Summarise() and group_by()

- Get median solar radiation by month
  - Remember to have na.rm =T in your median() function

```
> aqt %>% group_by(Month) %>%
+   summarise(median_rad = median(Solar.R, na.rm = T))
# A tibble: 5 x 2
  Month median_rad
  <int>      <dbl>
1     5        194
2     6        188.
3     7        253
4     8        198.
5     9        192
```

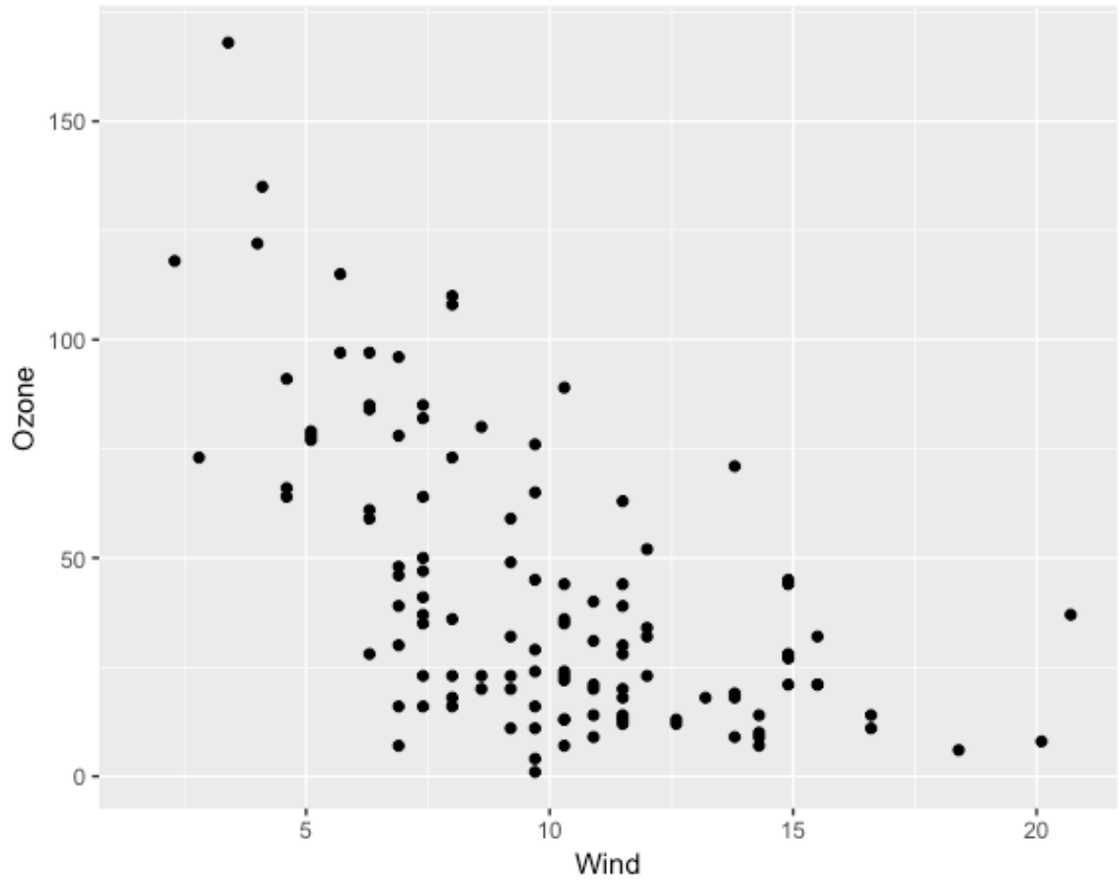
Visualization

- ggplot2
  - Graphics language for creating elegant and complex plots

geom_point()

- Is Wind correlated with Ozone?

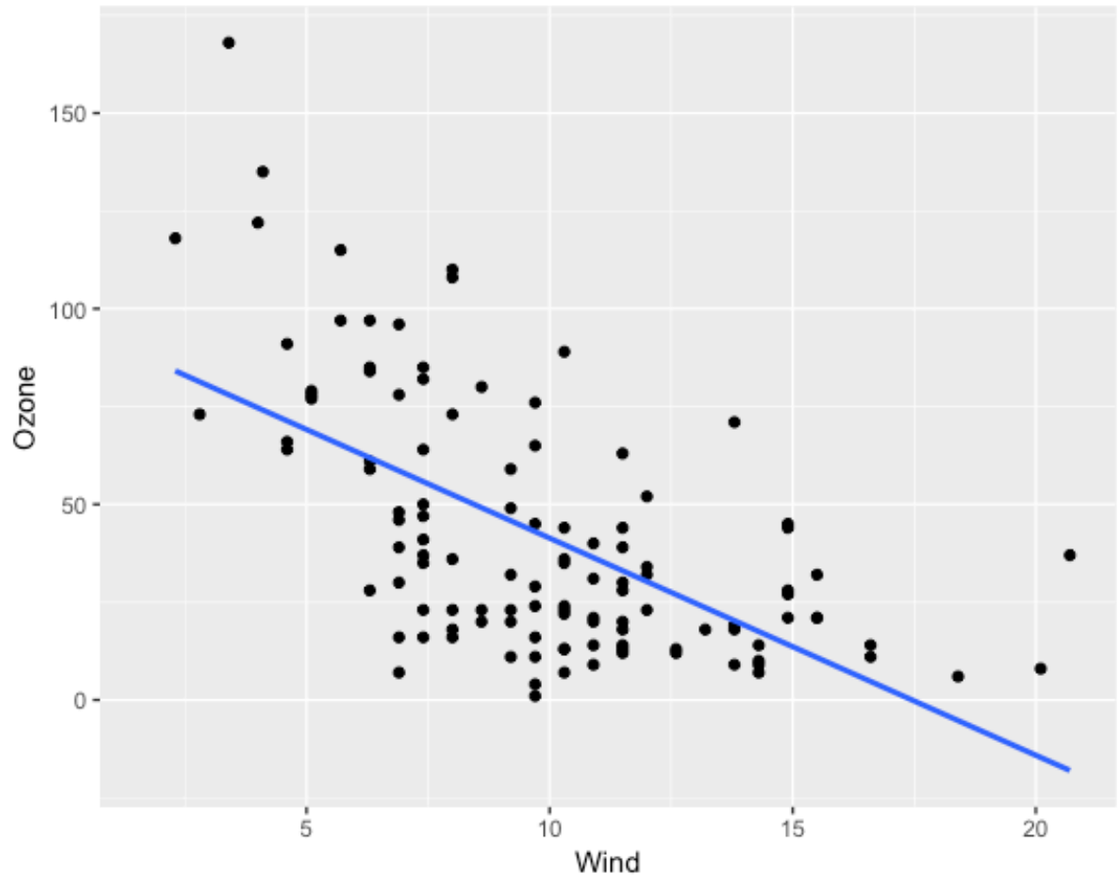
```
> aqt %>% ggplot(aes(x=Wind, y=Ozone)) +
+   geom_point()
Warning message:
Removed 37 rows containing missing values (geom_point).
```



geom_point()

- Is wind correlated with Ozone?

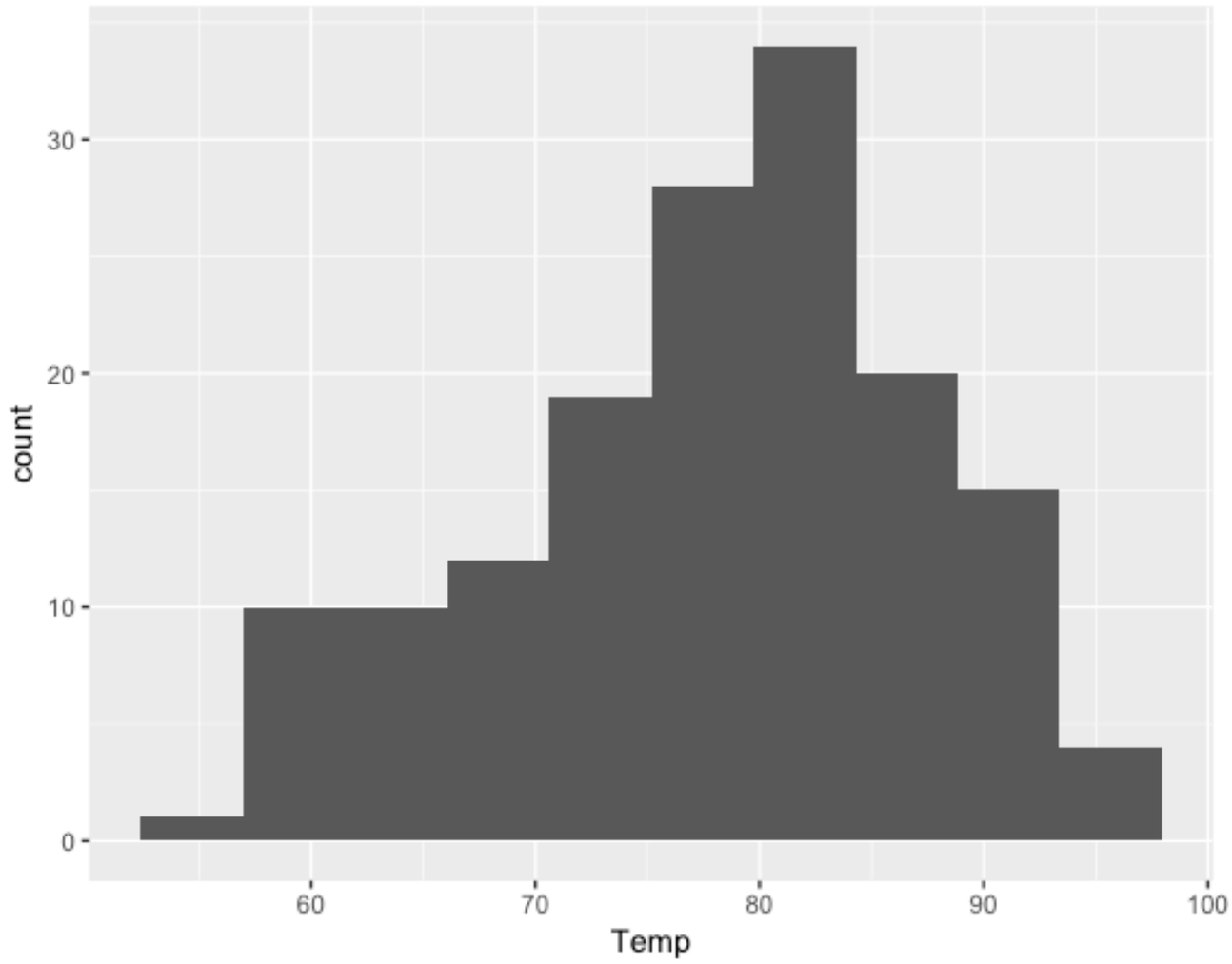
```
> aqt %>% ggplot(aes(x=Wind, y=Ozone)) +
+   geom_point() +
+   stat_smooth(method = 'lm', se = F)
Warning messages:
1: Removed 37 rows containing non-finite values (stat_smooth).
2: Removed 37 rows containing missing values (geom_point).
```



geom_histogram()

- Histogram of temperature

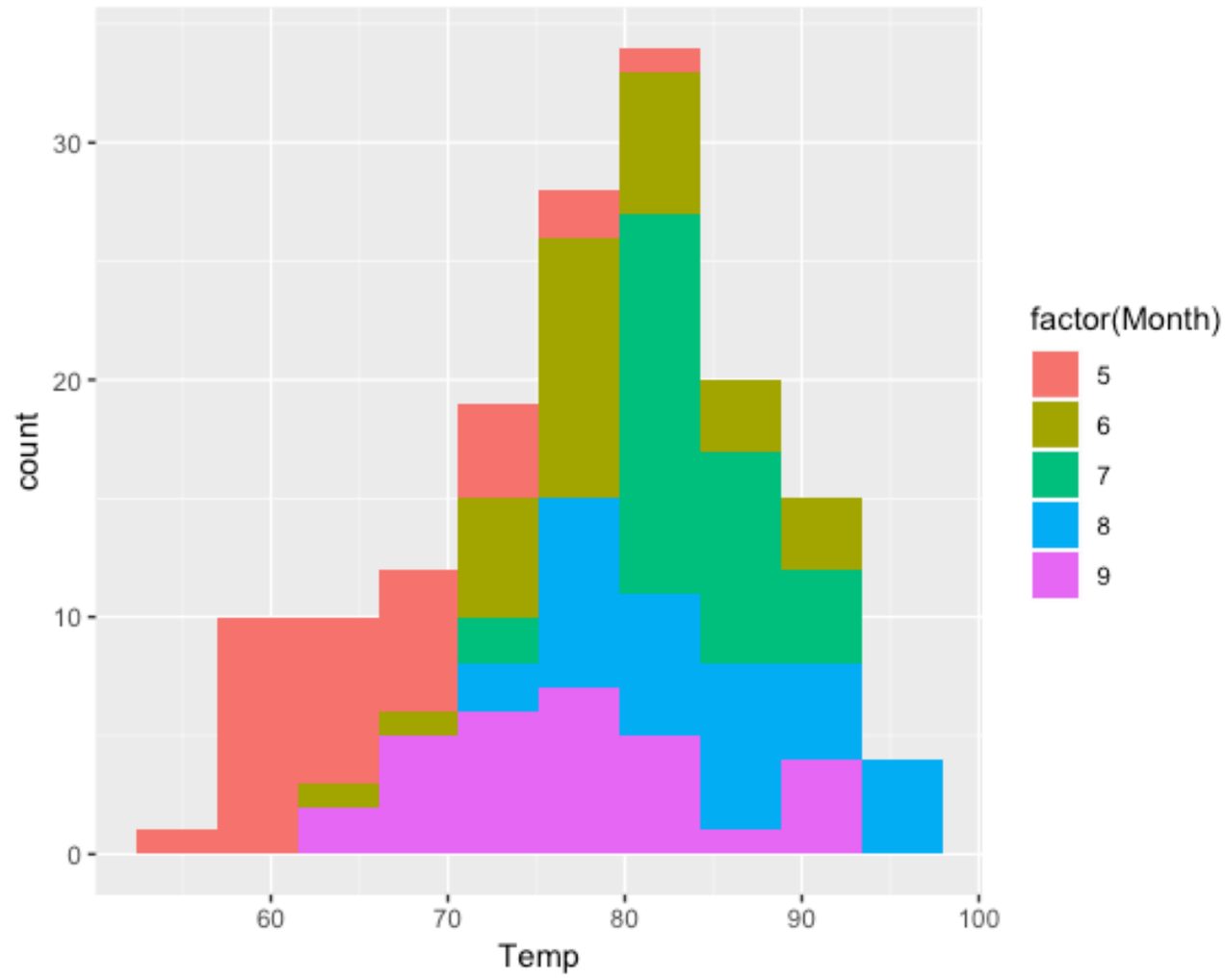
```
> aqt %>% ggplot(aes(x=Temp)) +
+   geom_histogram(bins = 10)
```



geom_histogram()

- Histogram of temperature

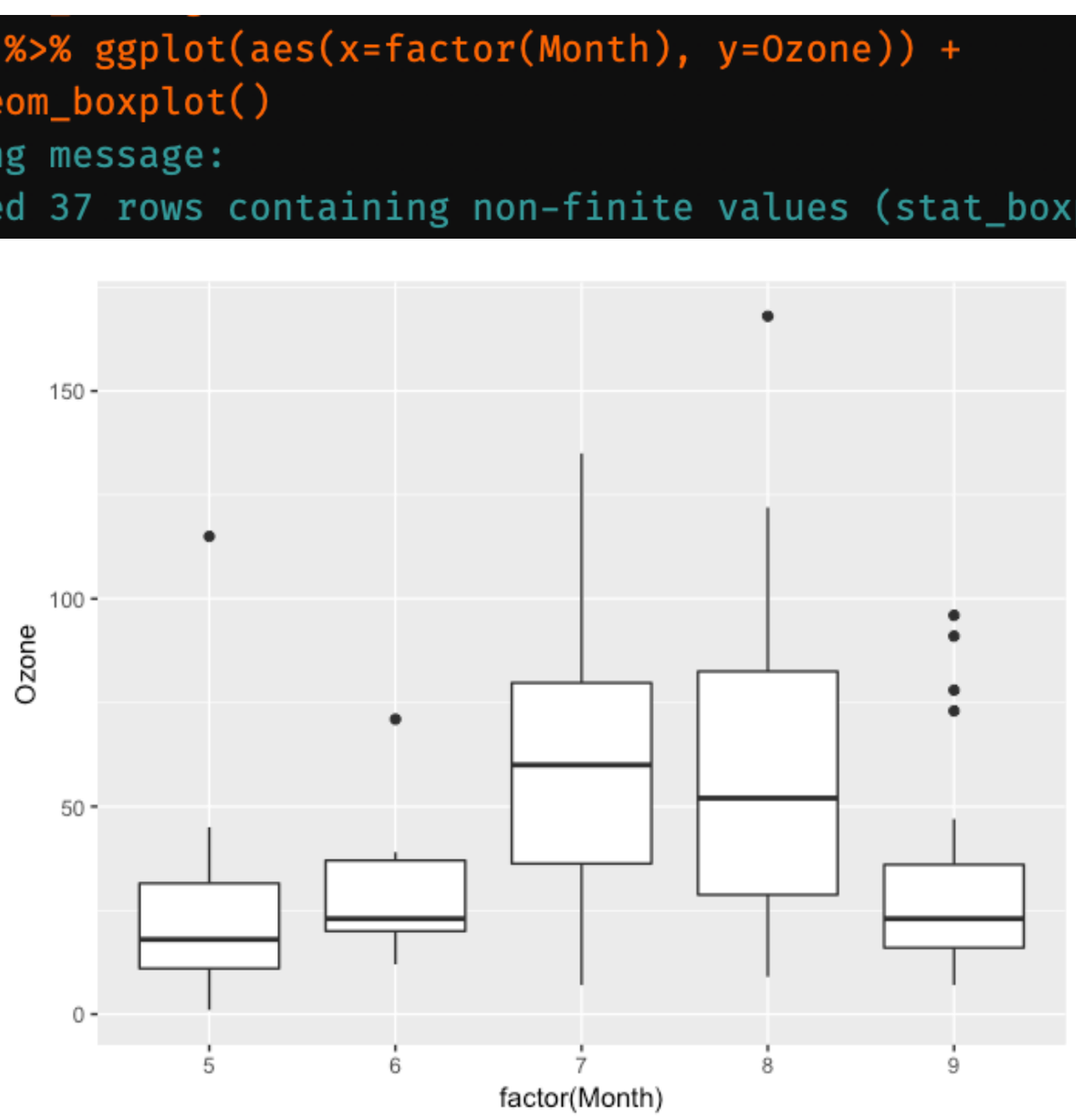
```
> aqt %>% ggplot(aes(x=Temp, fill = factor(Month))) +
+   geom_histogram(bins = 10)
```



geom_boxplot()

- Ozone distributions by month

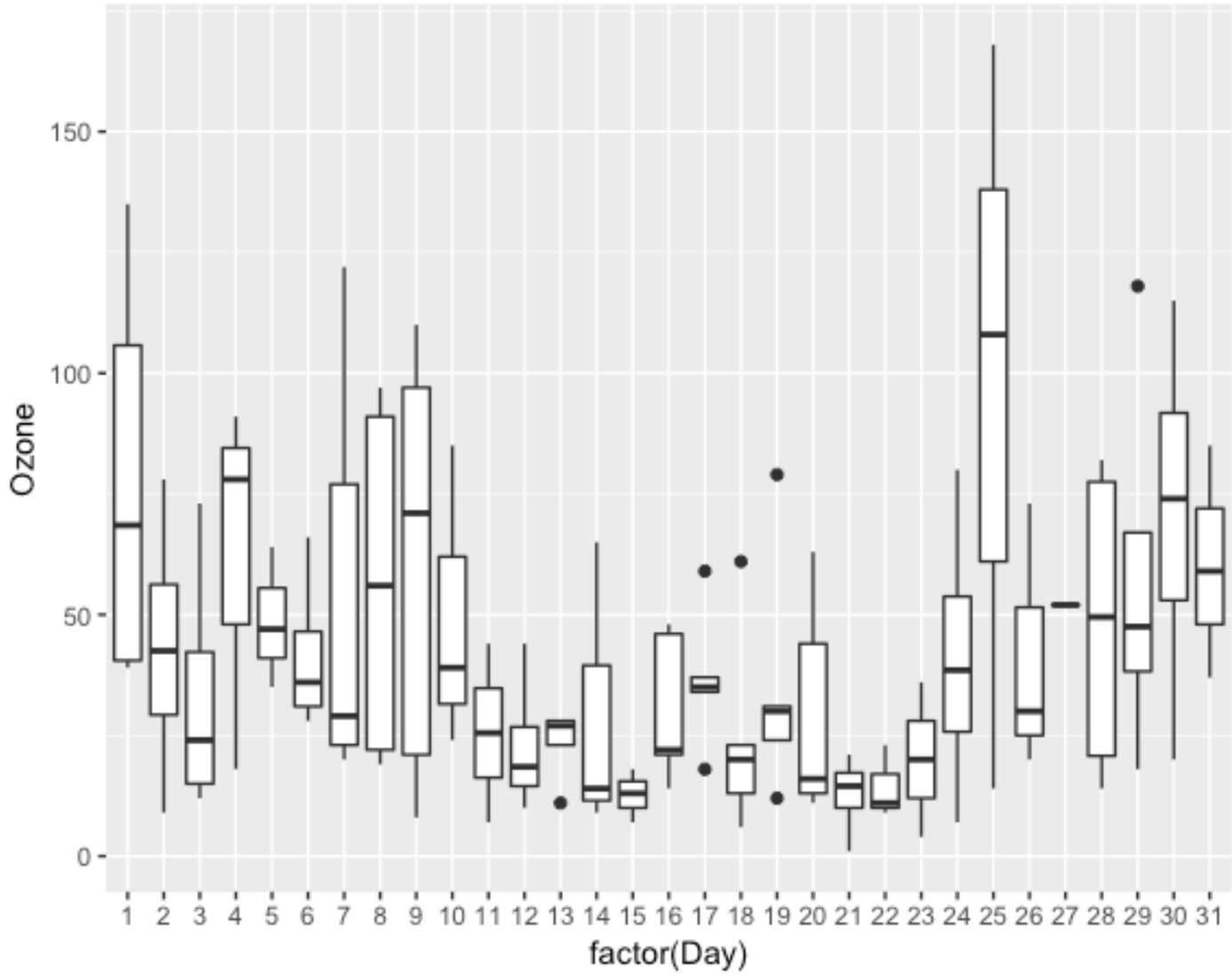
```
> aqt %>% ggplot(aes(x=factor(Month), y=Ozone)) +
+   geom_boxplot()
Warning message:
Removed 37 rows containing non-finite values (stat_boxplot).
```



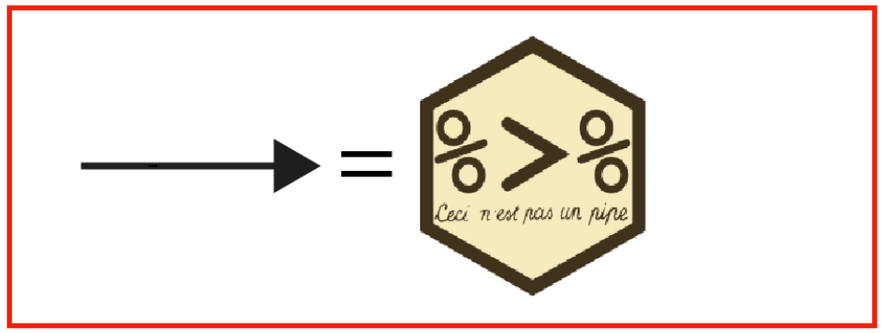
Exercise 5:
How do Ozone distributions vary by day?

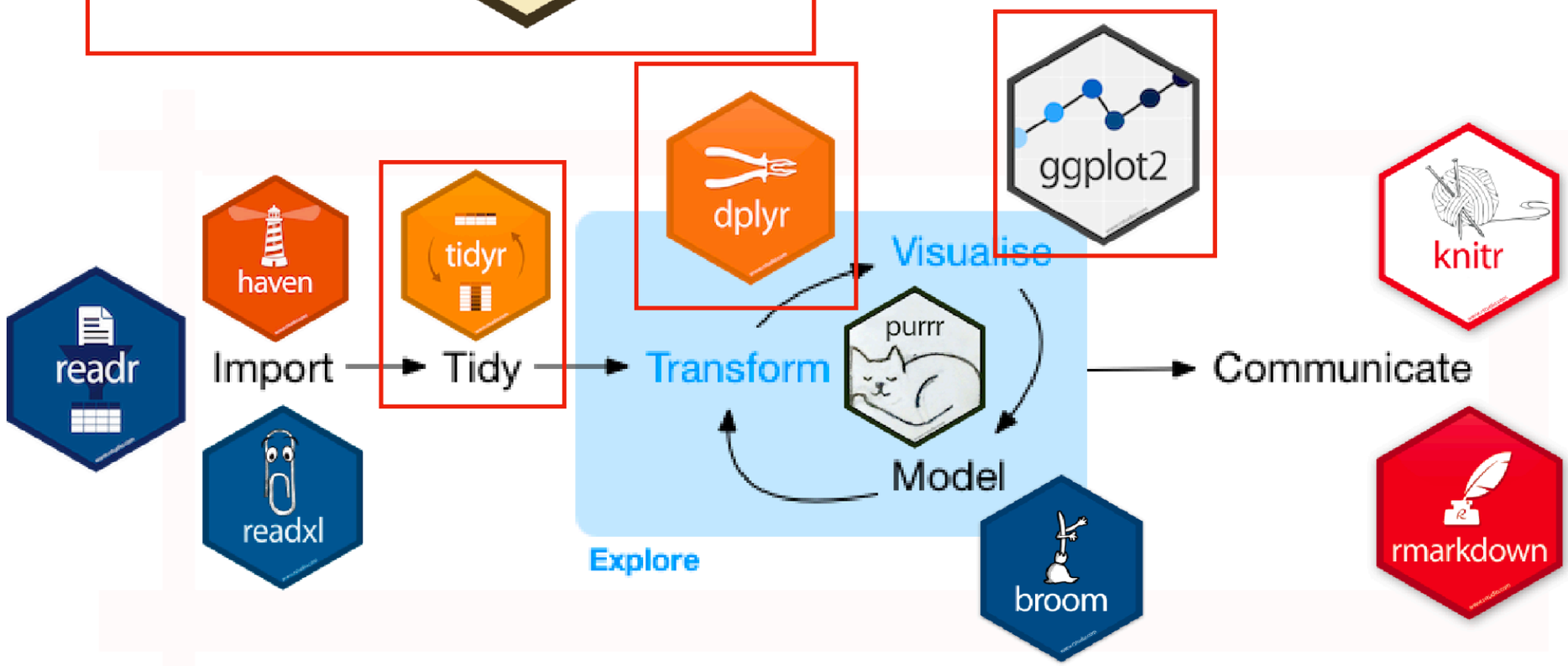
- What plot would you use?

```
> aqt %>% ggplot(aes(x=factor(Day), y=Ozone)) + geom_boxplot()
Warning message:
Removed 37 rows containing non-finite values (stat_boxplot).
```



What did we learn?





data.frames

factors

strings

Bonus Exercise: The iris dataset

- data("iris")
- What unique species of flowers are represented?
- Mean Sepal length and width by Species?
- Plots
    - Visualize sepal length distributions by species (what plot would you use?)
    - Visualize Sepal length by sepal width (what plot?_
        - Color by species
        - Add linear fits!
- Whatever else you'd like!

Bonus: purrr

- Functional programming toolkit
    - To work on functions and vectors

Bonus: purrr

- map()

```
> l ← list('one' = 1, 'two' = 2, 'three' = 3)
> lroot ← list()
> for(i in names(l)) {
+     lroot[[i]] ← sqrt(l[[i]])
+ }
> lroot
$one
[1] 1

$two
[1] 1.414214

$three
[1] 1.732051
```

```
> lroot1 ← l %>% map(sqrt)
> lroot1
$one
[1] 1

$two
[1] 1.414214

$three
[1] 1.732051
```

Bonus: purrr

- pluck()

```
> aqt %>% filter(Month == 9) %>%
+   mutate(Feeling = ifelse(Temp < 70, 'Cool', 'Warm'))
# A tibble: 30 x 8
   Ozone Solar.R  Wind  Temp Month   Day TempC Feeling
   <int>   <int> <dbl> <int> <int> <int> <dbl> <chr>
 1    96     167   6.9    91     9     1  32.8 Warm
 2    78     197   5.1    92     9     2  33.3 Warm
 3    73     183   2.8    93     9     3  33.9 Warm
 4    91     189   4.6    93     9     4  33.9 Warm
 5    47      95   7.4    87     9     5  30.6 Warm
 6    32      92  15.5    84     9     6  28.9 Warm
 7    20     252  10.9    80     9     7  26.7 Warm
 8    23     220  10.3    78     9     8  25.6 Warm
 9    21     230  10.9    75     9     9  23.9 Warm
10    24     259   9.7    73     9    10  22.8 Warm
# … with 20 more rows
> aqt %>% filter(Month == 9) %>%
+   mutate(Feeling = ifelse(Temp < 70, 'Cool', 'Warm')) %>%
+   pluck(8)
 [1] "Warm" "Warm" "Warm" "Warm" "Warm" "Warm" "Warm" "Warm" "Warm" "Warm" "Warm" "Warm" "Warm" "Warm" "Warm" "Warm" "Cool" "Warm" "Cool"
[20] "Warm" "Cool" "Warm" "Warm" "Cool" "Cool" "Warm" "Warm" "Warm" "Warm" "Cool"
```

Bonus: broom

- install.library("broom")
- library(broom)
- tidy()

```
> library(broom)
> fit ← lm(Ozone ~ Wind, aqt)
> fit %>% summary()

Call:
lm(formula = Ozone ~ Wind, data = aqt)

Residuals:
    Min      1Q  Median      3Q     Max
-51.572 -18.854  -4.868  15.234  90.000

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  96.8729     7.2387   13.38  < 2e-16 ***
Wind         -5.5509     0.6904   -8.04 9.27e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.47 on 114 degrees of freedom
  (37 observations deleted due to missingness)
Multiple R-squared:  0.3619,    Adjusted R-squared:  0.3563
F-statistic: 64.64 on 1 and 114 DF,  p-value: 9.272e-13

> tidy(fit)
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)    96.9      7.24      13.4  3.99e-25
2 Wind           -5.55     0.690     -8.04 9.27e-13
```

Bonus: broom

- glance()

```
> glance(fit)
# A tibble: 1 x 11
  r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC deviance df.residual
      <dbl>         <dbl> <dbl>     <dbl>    <dbl> <int>  <dbl> <dbl> <dbl>    <dbl>       <int>
1     0.362         0.356  26.5      64.6 9.27e-13     2  -544. 1093. 1101.   79859.         114
```

Bonus: broom

- augment()

```
> augment(fit)
# A tibble: 116 x 10
   .rownames Ozone  Wind .fitted .se.fit .resid    .hat .sigma    .cooksd .std.resid
   <chr>     <int> <dbl>   <dbl>   <dbl>  <dbl>   <dbl>  <dbl>      <dbl>      <dbl>
 1 1            41   7.4    55.8    2.99 -14.8  0.0127    26.5 0.00204       -0.563
 2 2            36   8      52.5    2.77 -16.5  0.0110    26.5 0.00217       -0.626
 3 3            12  12.6    26.9    3.10 -14.9  0.0137    26.5 0.00224       -0.568
 4 4            18  11.5    33.0    2.71 -15.0  0.0104    26.5 0.00172       -0.571
 5 6            28  14.9    14.2    4.26  13.8  0.0259    26.6 0.00373        0.530
 6 7            23   8.6    49.1    2.61 -26.1  0.00970   26.5 0.00482       -0.992
 7 8            19  13.8    20.3    3.66  -1.27 0.0192    26.6 0.0000229     -0.0485
 8 9             8  20.1   -14.7    7.48  22.7  0.0799    26.5 0.0347         0.894
 9 11            7   6.9    58.6    3.20 -51.6  0.0146    26.1 0.0285        -1.96
10 12           16   9.7    43.0    2.46 -27.0  0.00864   26.5 0.00458       -1.03
# … with 106 more rows
```

Other resources

https://www.tidyverse.org/learn/

https://bookdown.org/Tazinho/Tidyverse-Cookbook/

https://jennybc.github.io/purrr-tutorial/

http://varianceexplained.org/r/teach-tidyverse/

https://monashbioinformaticsplatform.github.io/r-more/topics/tidyverse.html

R for data science, https://r4ds.had.co.nz/

Thank you!

thedattadoctor

gargi-datta

www.gargidatta.com