Key concepts in data viz and `ggplot`

Author: Nicholas G Reich

Types of data graphics

Using graphics to explore data

- The most valuable graphics are often the simple ones you make for yourself.
- Exploratory graphics can introduce you to a dataset.
- Key goal: understand the variation.
- What do you want to know about these data?

```
data(airquality)
head(airquality)

##   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
## 6    28      NA 14.9   66     5   6
```

Exploratory summaries: airquality data

Understanding what the rows and columns are

```
nrow(airquality)

## [1] 153

str(airquality)

## 'data.frame': 153 obs. of  6 variables:
##  $ Ozone  : int  41 36 12 18 NA 28 23 19 8 NA ...
##  $ Solar.R: int  190 118 149 313 NA NA 299 99 19 194 ...
##  $ Wind   : num  7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...
##  $ Temp   : int  67 72 74 62 56 66 65 59 61 69 ...
##  $ Month  : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ Day    : int  1 2 3 4 5 6 7 8 9 10 ...
```

Exploratory summaries: airquality data

Tabulating different values of the data using

```
stem(airquality$Ozone)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##    0 | 1467778999
##    1 | 01112233334444666688889
##    2 | 00001111233333334478889
##    3 | 001222455667799
##    4 | 01444556789
##    5 | 0299
##    6 | 134456
##    7 | 13367889
##    8 | 024559
##    9 | 1677
##   10 | 8
##   11 | 058
##   12 | 2
##   13 | 5
##   14 |
##   15 |
##   16 | 8
```

Exploratory summaries: airquality data

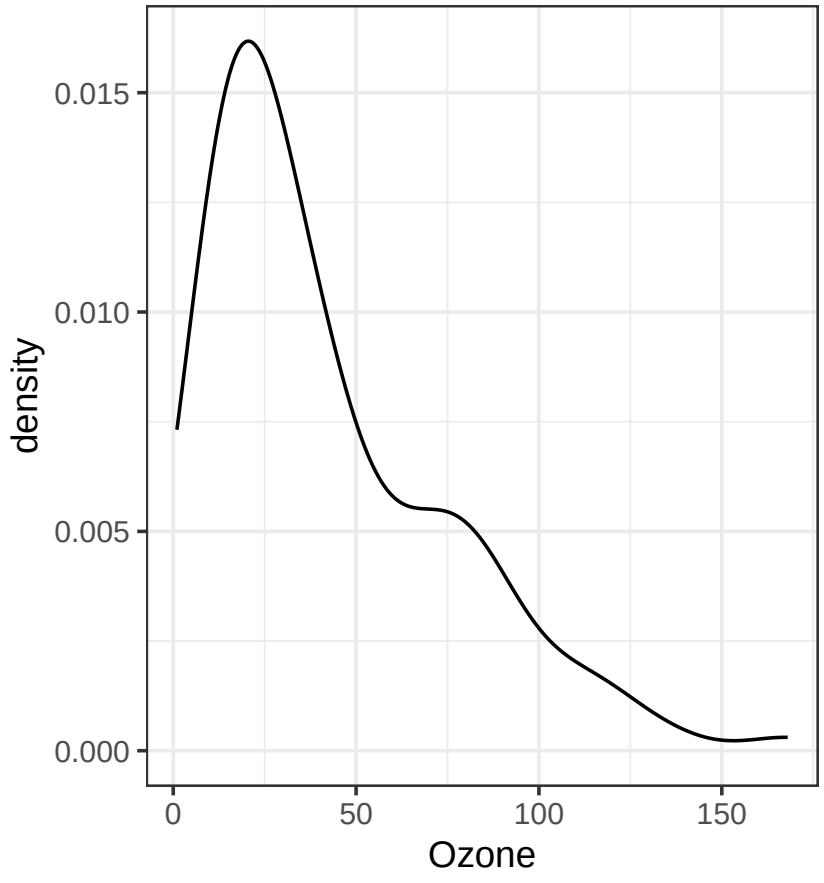
Tabulating different values of the data

```
table(airquality$Month)

##
##  5  6  7  8  9
## 31 30 31 31 30

with(airquality, table(Month, Day))

##       Day
## Month 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 2
##     5 1 1 1 1 1 1 1 1 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##     6 1 1 1 1 1 1 1 1 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##     7 1 1 1 1 1 1 1 1 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##     8 1 1 1 1 1 1 1 1 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##     9 1 1 1 1 1 1 1 1 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##       Day
## Month 28 29 30 31
##     5  1  1  1  1
##     6  1  1  1  0
##     7  1  1  1  1
##     8  1  1  1  1
##     9  1  1  1  0
```

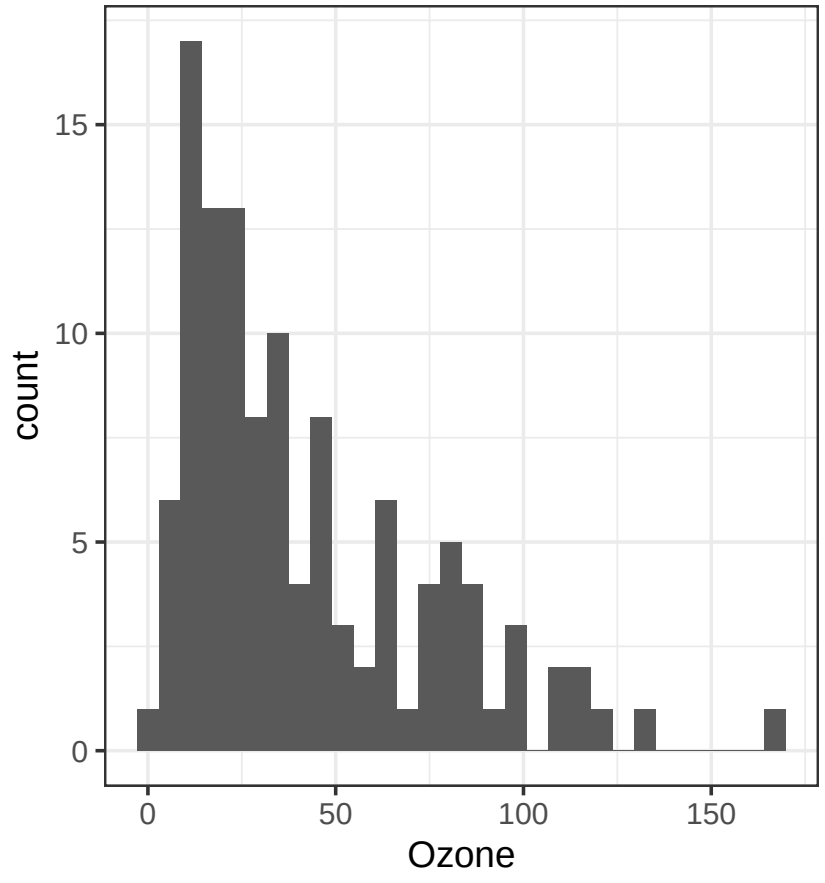
Univariate graphics: airquality data

Univariate graphics show you the distribution of or the variation in the observations of a single variable.

A density plot



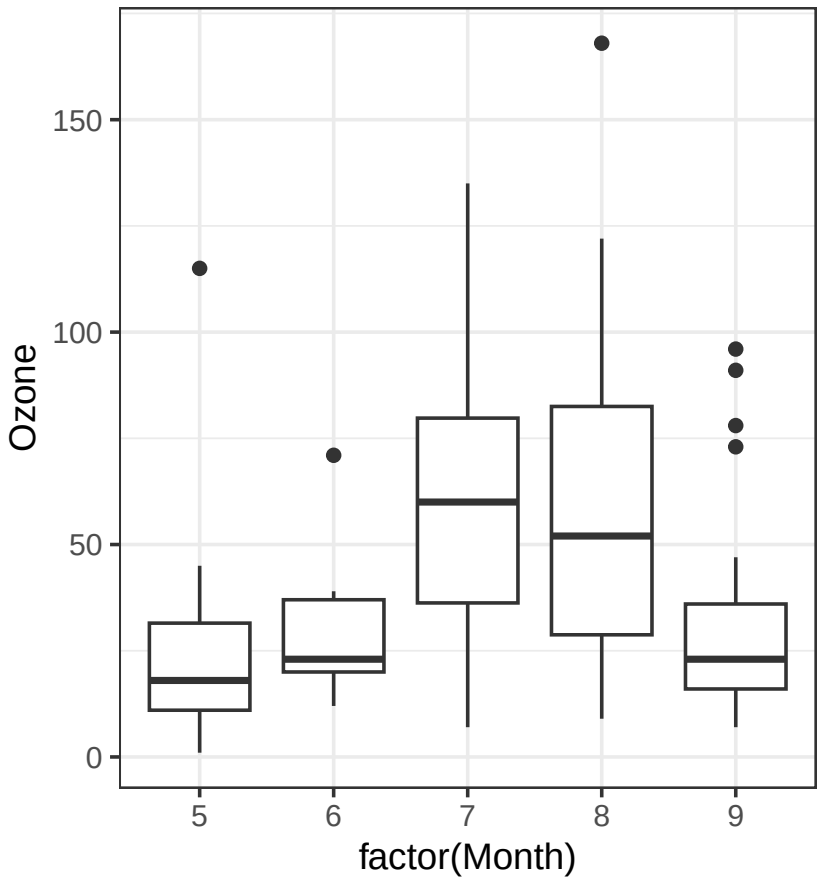
A histogram



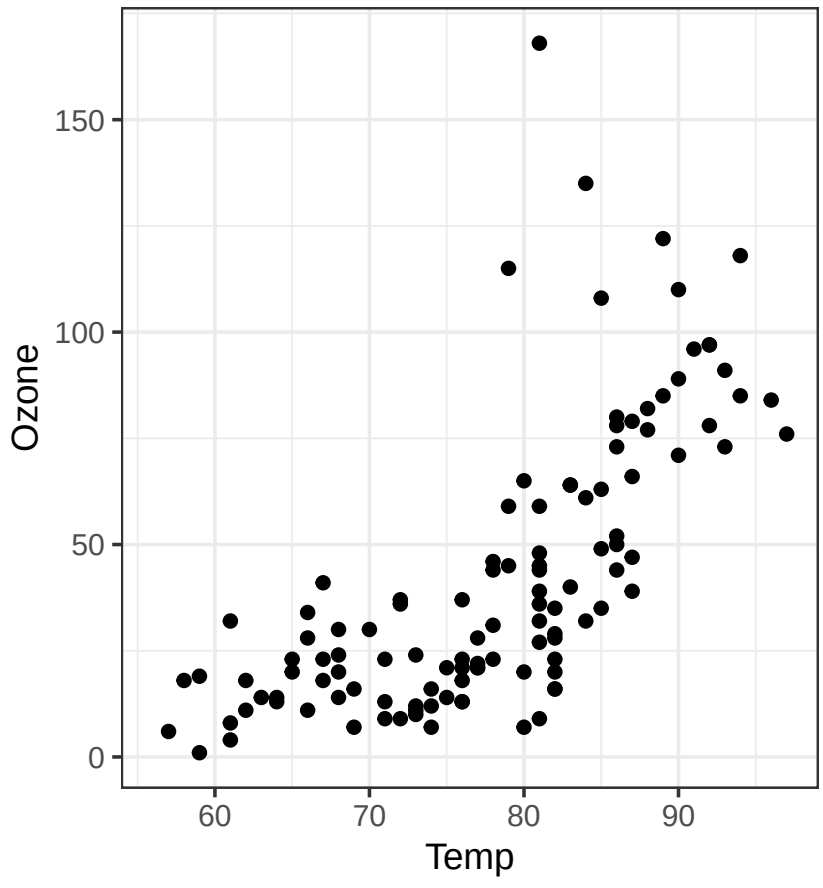
Bivariate graphics: airquality data

Bivariate graphics show you the relationship between two variables in your dataset.

A boxplot



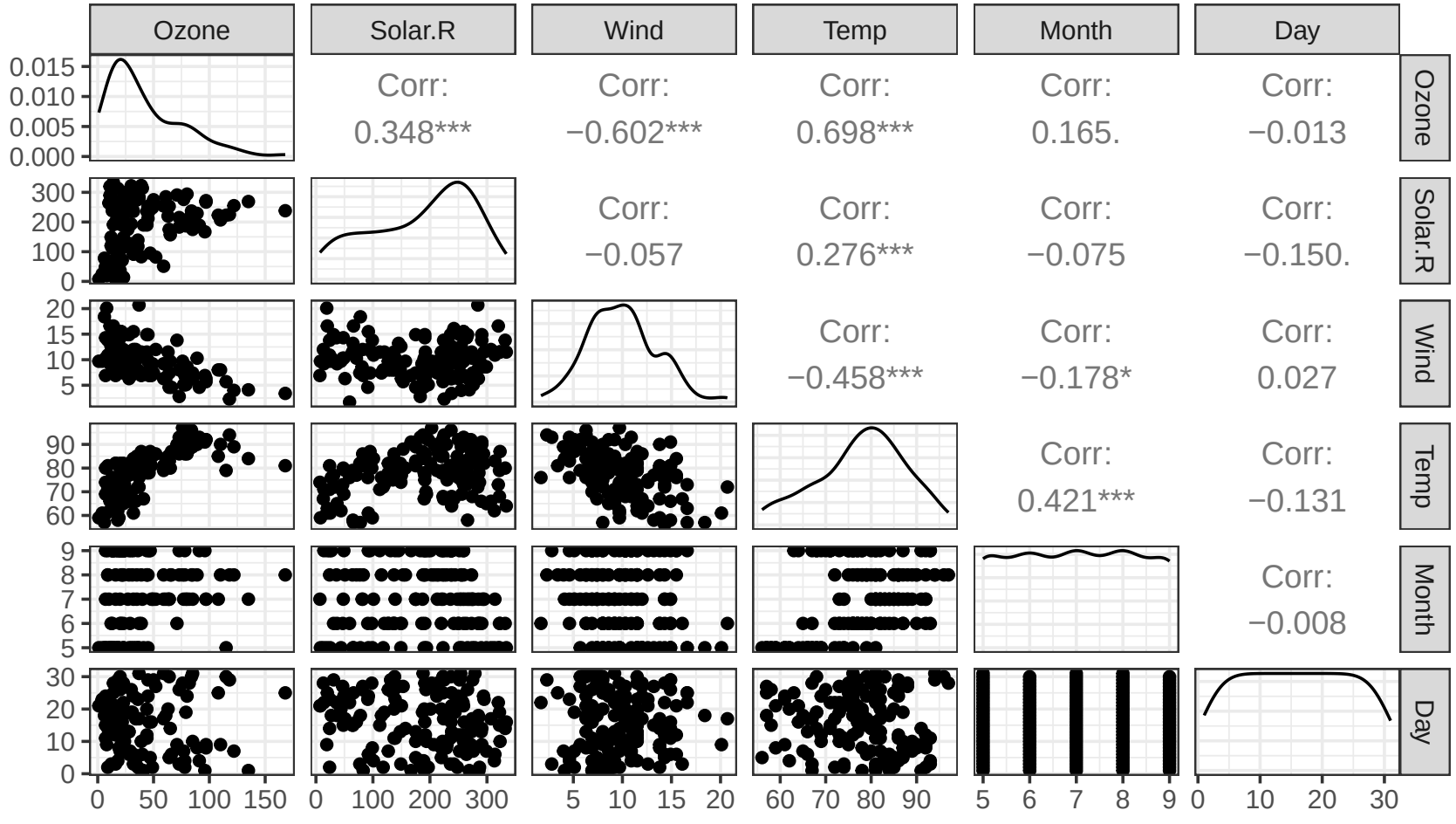
A scatterplot



Bivariate graphics: pairs plots!

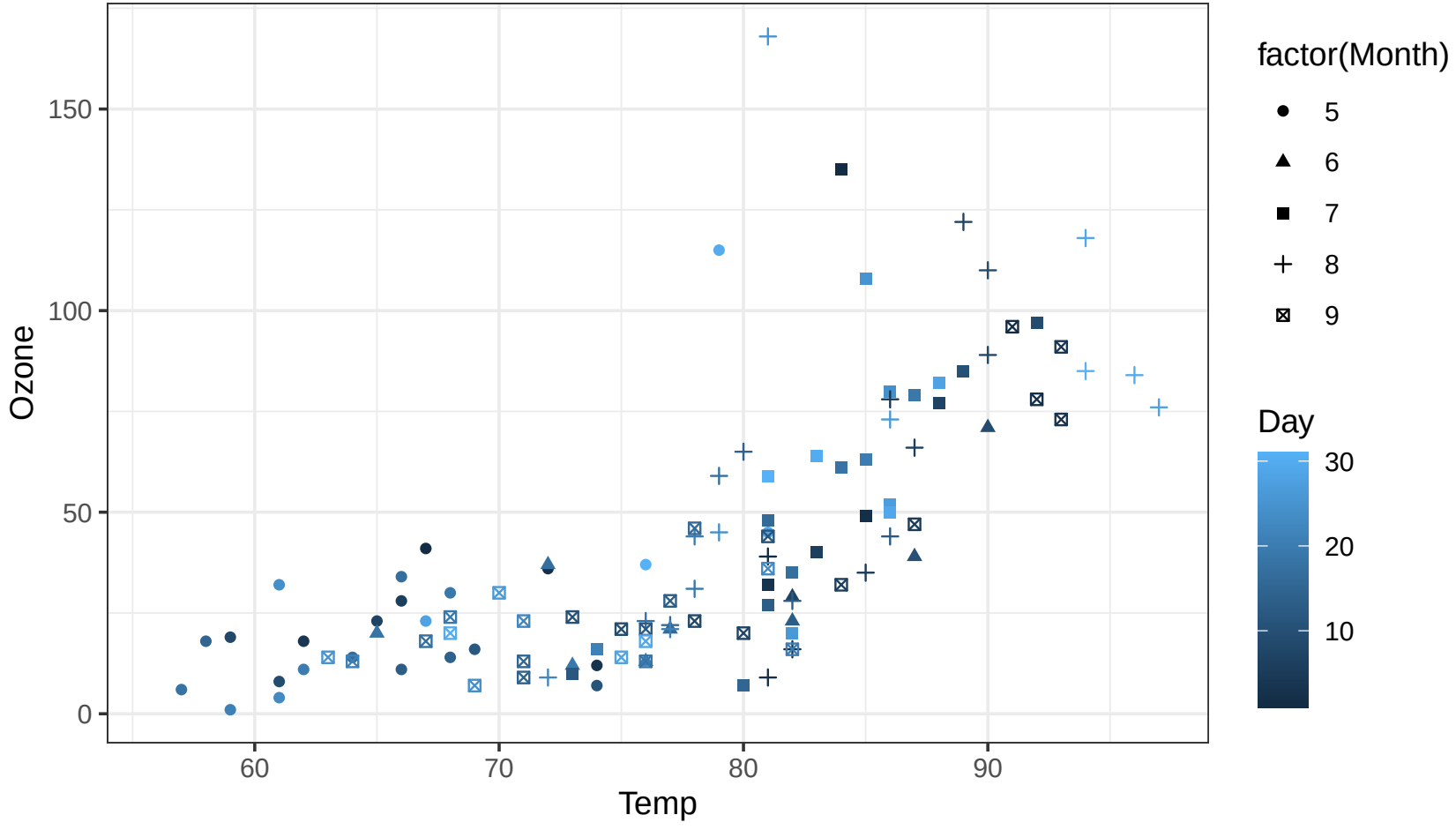
Pairs plots are nice, but can take some time to render (especially for big datasets).

```
GGally::ggpairs(airquality)
```



Multivariate graphics

Multivariate graphics show you the relationship between multiple variables in your dataset. Rather than using some fancy "3D" plot, it is often better to use other features like color or shapes or facets to show a third or fourth variable.



Using ggplot

Choices for R graphics

You have three central choices for making graphics in R:

- "Base R graphics"
- `ggplot2`
- lattice

I use `ggplot` because:

1. it is integrated with the `tidyverse`
2. it is actively developed/maintained
3. there are a ton of extensions (see more later)

Understanding the “grammar” of ggplot2

The grammar ...

- layers (a ‘geom’, a ‘stat’, an ‘annotation’)
- aesthetics (‘aes‘)
- scales
- facets
- data
- ... and more here: http://ggplot2.tidyverse.org/reference/

What is a layer?

Layers define the basic structure of the elements on the plot

- Geoms: point, line, tile, boxplot, ribbon, ...
- Stats: histogram, smooth, density, ...
- Annotation: hline, vline, text, ...

For more info check out the documentation:
http://ggplot2.tidyverse.org/reference/.

What are "aesthetics"?

Aesthetics define a mapping between **tidy data** and the information required to create a specific graphic[1]

length	width	depth	trt
2	3	4	a
1	2	1	a
4	5	15	b
9	10	80	b

x	y	colour
2	3	a
1	2	a
4	5	b
9	10	b

[1] Figure credits: Hadley Wickham

geom_point

Each geom has a different set of aesthetics.
What information do we need to draw a scatterplot?
Or, asked another way, what aesthetics do we need for `geom_point`?

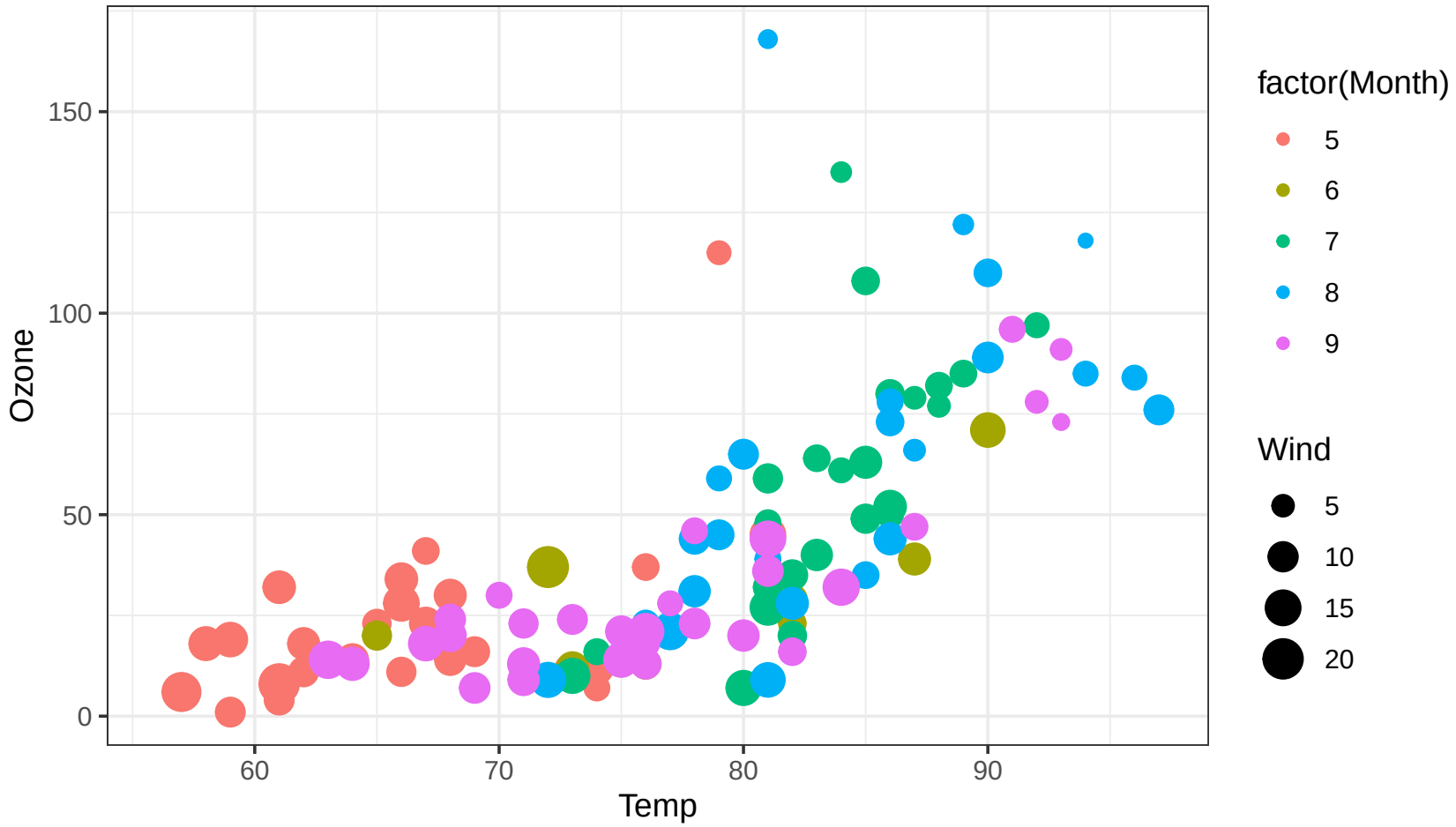
geom_point

Each geom has a different set of aesthetics.
What information do we need to draw a scatterplot?
Or, asked another way, what aesthetics do we need for `geom_point`?

- x (required)
- y (required)
- alpha
- color
- fill
- shape
- size

geom_point

```
library(ggplot2)
theme_set(theme_bw())
ggplot(airquality) +
  geom_point(aes(x=Temp, y=Ozone, color=factor(Month), size=Wind))
```



geom_line

What information do we need to draw a line of connected points?
Or, asked another way, what aesthetics do we need for `geom_line`?

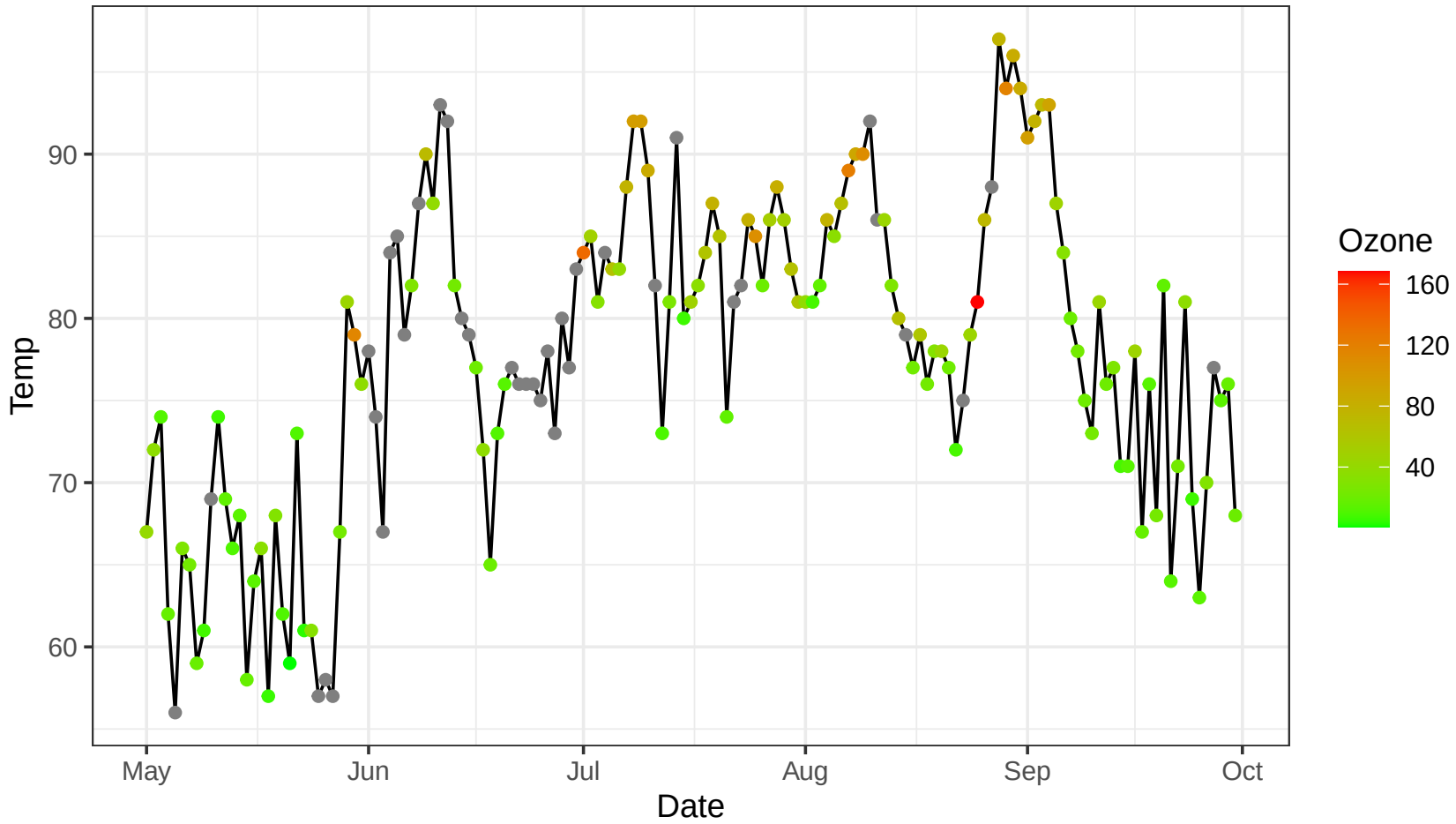
geom_line

What information do we need to draw a line of connected points?
Or, asked another way, what aesthetics do we need for geom_line?

- x (required)
- y (required)
- alpha
- color
- linetype
- size

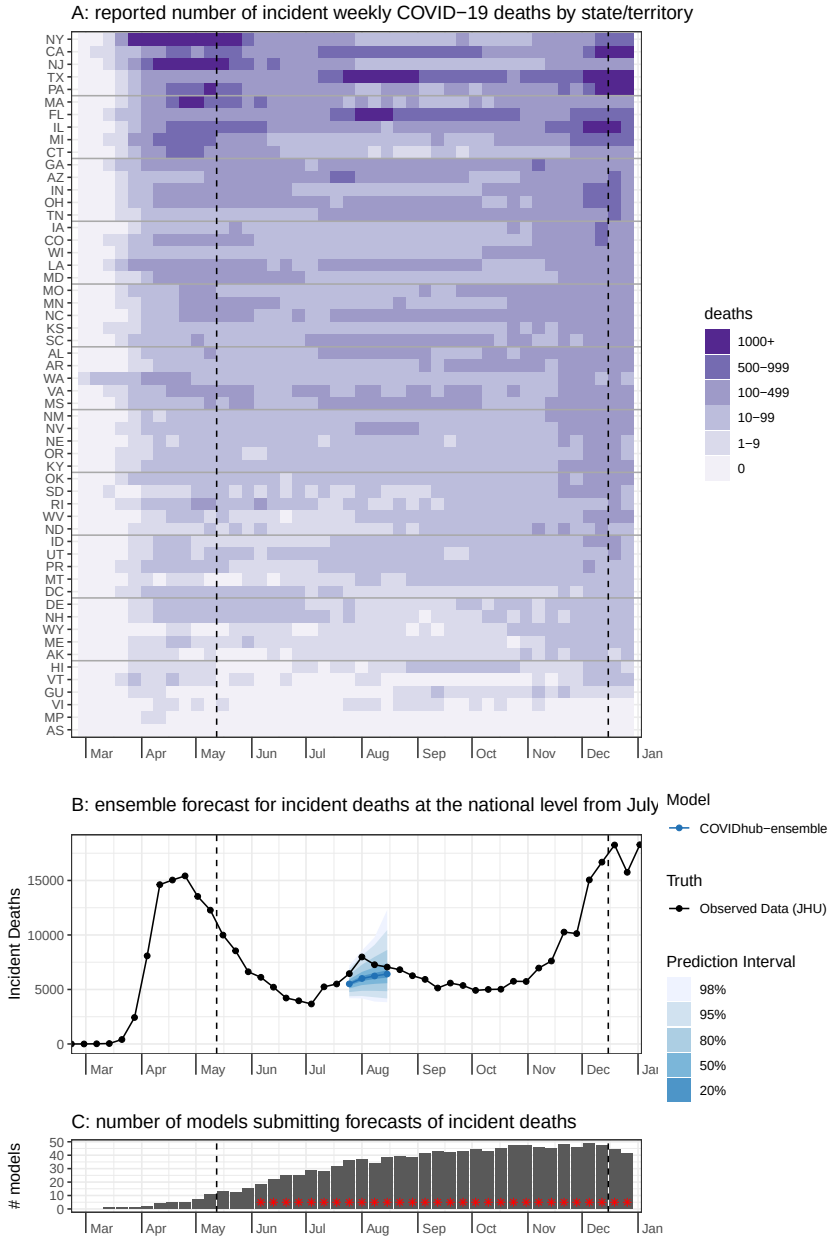
geom_line

```
airquality$Date <- lubridate::ymd(paste(1973, airquality$Month, airquality$Day))
ggplot(airquality, aes(x=Date, y=Temp)) +
  geom_line() + geom_point(aes(color=Ozone)) +
  scale_color_gradient(low="green", high="red")
```

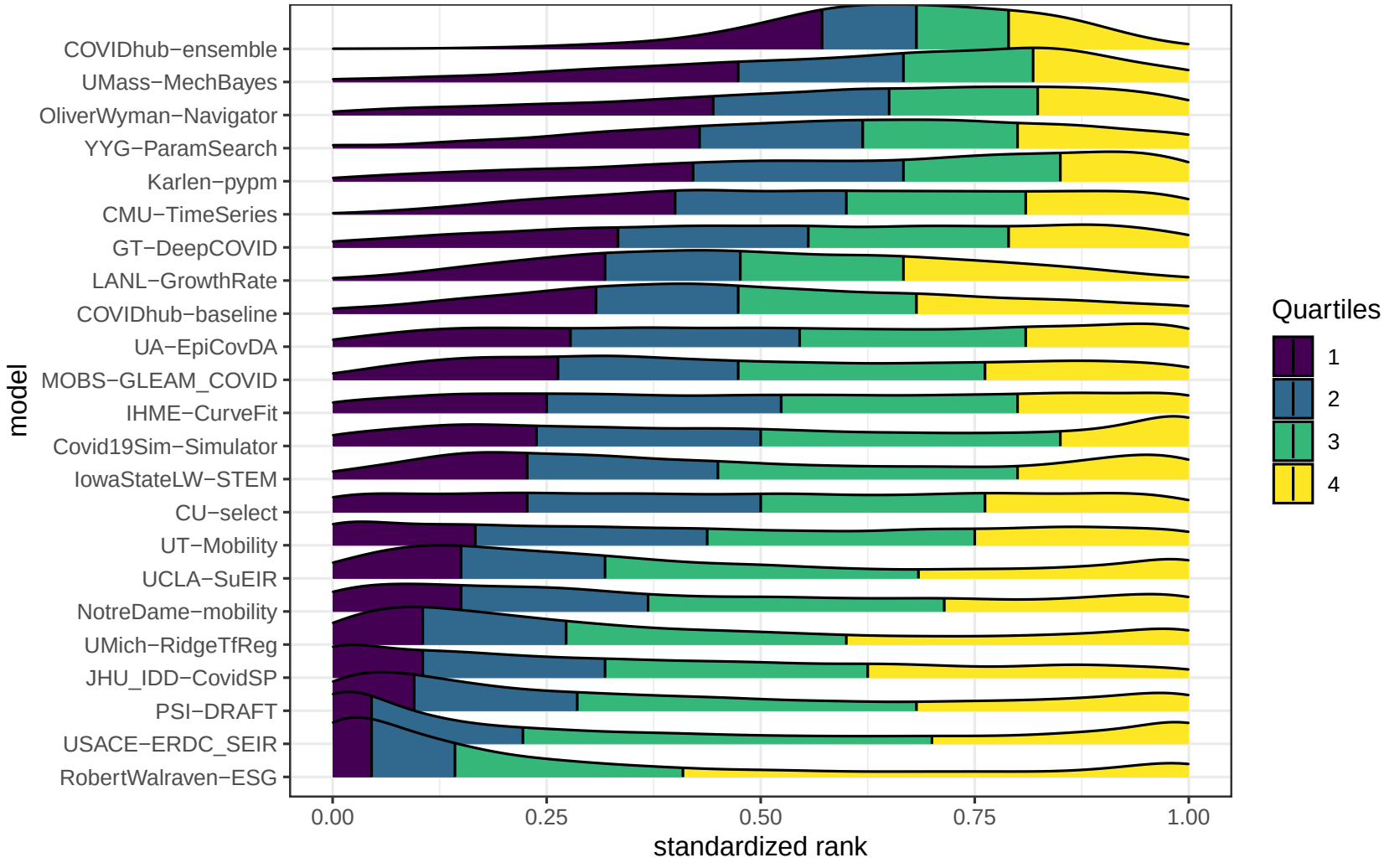


`ggplot` extensions that I used in a recent paper

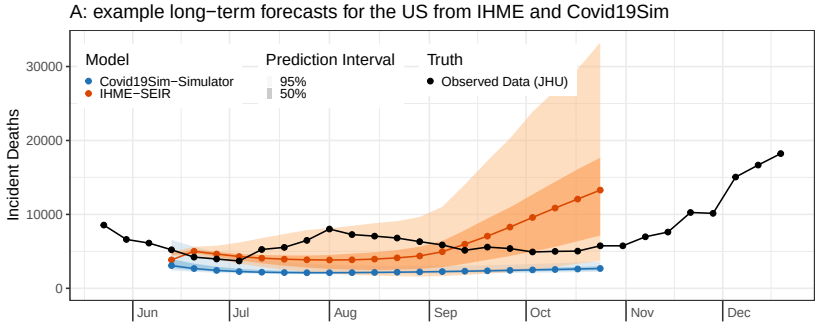
gridExtra or cowplot for multi-plot alignment

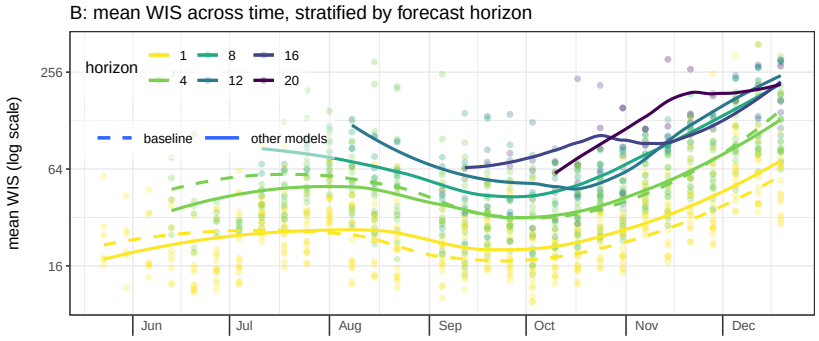


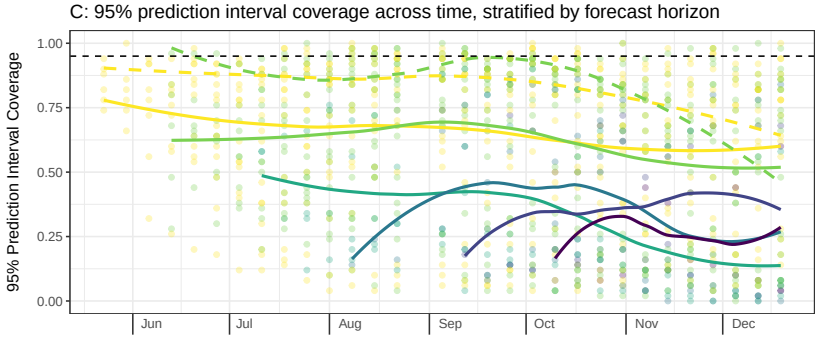
ggrides for ridgeplots

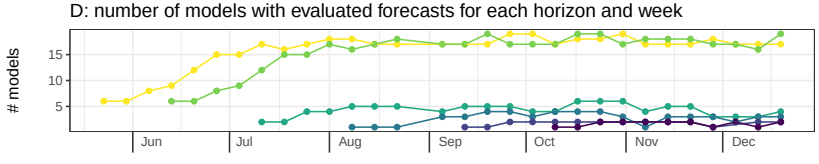


RColorBrewer and viridis for colors









Note-catcher

A figure from "Cities, traffic, and CO2: A multidecadal assessment of trends, drivers, and scaling relationships", Gately et al, PNAS, 2015. Original paper on Moodle.

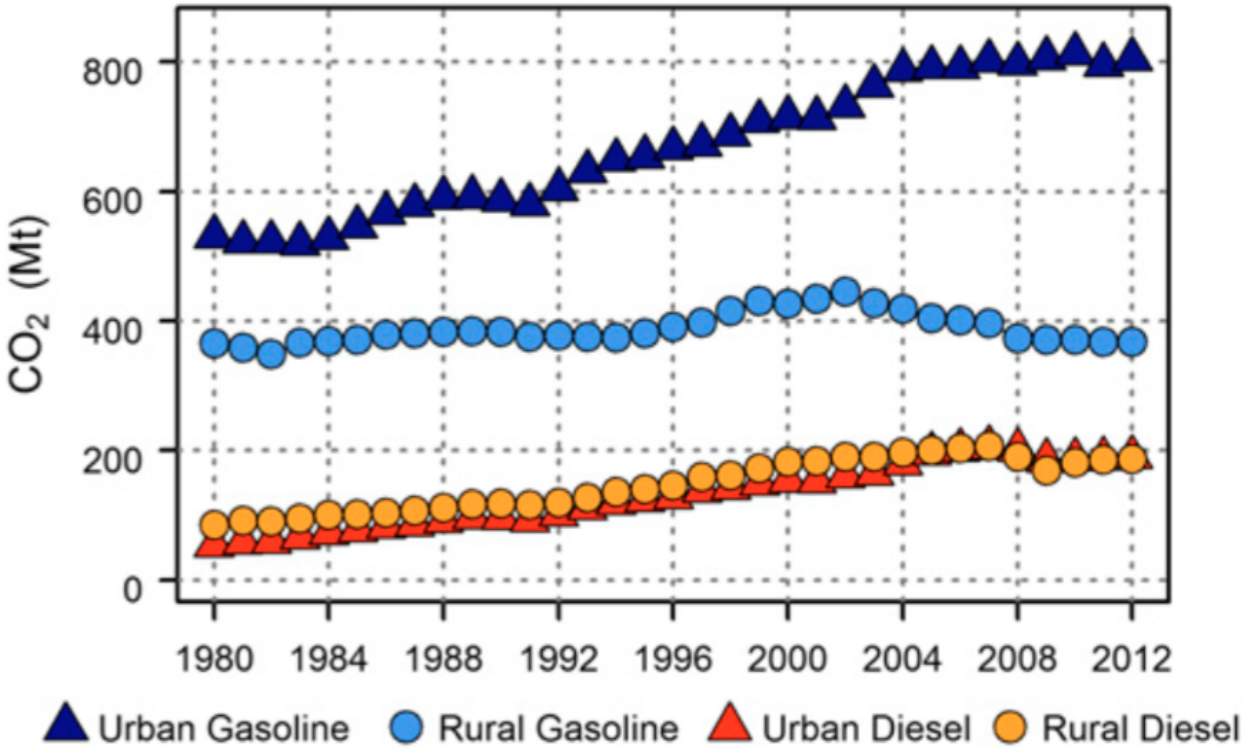


Fig. 2. Time series of US on-road CO_2 emissions. Urban roads accounted for 80% of total emissions growth since 1980. Rural road emissions have been declining since 2002.

Note-catcher

We have made the data from the CO2 emissions figure available on Canvas. As a group, you will be asked to complete the following tasks:

1. Recreate the figure as close as possible to the original.
2. Improve the figure. Make some changes that you think make the figure more clear.
3. Post your final figures on the Note-catcher document.

The class will vote on which figure is (1) closest to the original and (2) the best improvement. Extra credit on a future homework assignment will be awarded to all members at the table(s) that win the votes.

NOTE

Everyone should be doing some coding here, and having a version of the graphic working on their laptop! Make sure it's not just 1-2 people typing the code and having it work for them.