# Fun with ggplot2

## Visualizing data in R

Download the section 5 .Rmd handout to `STAT240/lecture/sect05-more-ggplot`.

Download the file `lake-mendota-winters-2023.csv` to `STAT240/data`

Material in this section is covered by Chapter 6 on the notes website.

Let's use the Lake Mendota data to explore other plot types.

- `geom_smooth` is a useful addition to a scatterplot.

There are also geoms for plotting a single variable.

- `geom_hist`
- `geom_density`
- `geom_boxplot`

`geom_smooth` shows the overall tend in a time series scatterplot.

- Can optionally show *confidence intervals*
- Several different methods for calculation

So far, all of the plots are motivated by the relationship between year and duration.

Now, let's study the duration variable on its own.

Histograms, density plots, and boxplots are useful tools for a single numeric variable.

`geom_hist` divides the data into bins and draws bars based on the number of observations. Specify:

- `binwidth` for how wide the bins should be
- `bins` for the number of bins
- `center` for the center of a bin
- `boundary` for a specific breakpoint

Use only one of (`binwidth`, `bins`) and only one of (`center`, `boundary`).

`geom_density` builds a density plot. It is similar to a histogram, but has a smooth curve instead of discrete bins.

- Good to emphasize "general trend"
- Related to integration

Consider layering both a density and histogram plot.

`geom_boxplot` creates a "box-and-whisker" plot. This visualizes the **quartiles**.

- Shows minimum, 25th, 50th, 75th percentiles, and maximum
- The box shows the middle 50% of the data
- Outliers are drawn as dots

The box width is the **interquartile range** (IQR).

- The "threshold" for outliers is 1.5× IQR
- Anything that is 1.5 "box lengths" away is a dot

Note: the lines only go out to data that exists.

Consider making a categorical variable for century.

- Add `fill = century` to color-code the one-variable plots
- What happens if we use `col = century` instead?
- Make a change to the density plot to make the overlapping plots more readable.

Lines are a useful way to annotate different types of numeric plots.

- Use `geom_vline` or `geom_hline`
- Can add multiple lines

`geom_text` can do variable mapping but is also useful for text annotations.

Histograms, density plots, and boxplots are tools to visualize a single numeric variable.

A bar graph is used to visualize a single categorical variable.

We draw bars (similar to a histogram) based on the number of occurrences in each category.

We see that the x-axis is organized alphabetically.

The bar plot creates a y-variable, "count", for us. We can also provide the height manually with `geom_col`.

This geom takes a categorical and numeric variable.

- Requires more manual calculation
- More flexible, not just counts

We can edit the axes of our plots to be more useful and informative.

- Use `scale_x` and `scale_y` to specify the axis
- Can be `continuous` or `discrete` depending on the data type
- Helpful arguments: `breaks`, `labels`, `limits`, `trans`

The most fun part of graphing is choosing a color scheme.

- Useful (colorblind friendly) built-in scales in `viridis`
- Can make your own custom scale with `manual`
- Specify `d` or `c` for discrete and continuous color schemes

Here are the viridis options.
Here is a list of predefined R colors.

Use the `labs` addition to customize labels.

- Title, subtitle, and caption
- Can change labels for any mapping present in your graph
- Can make labels blank as well

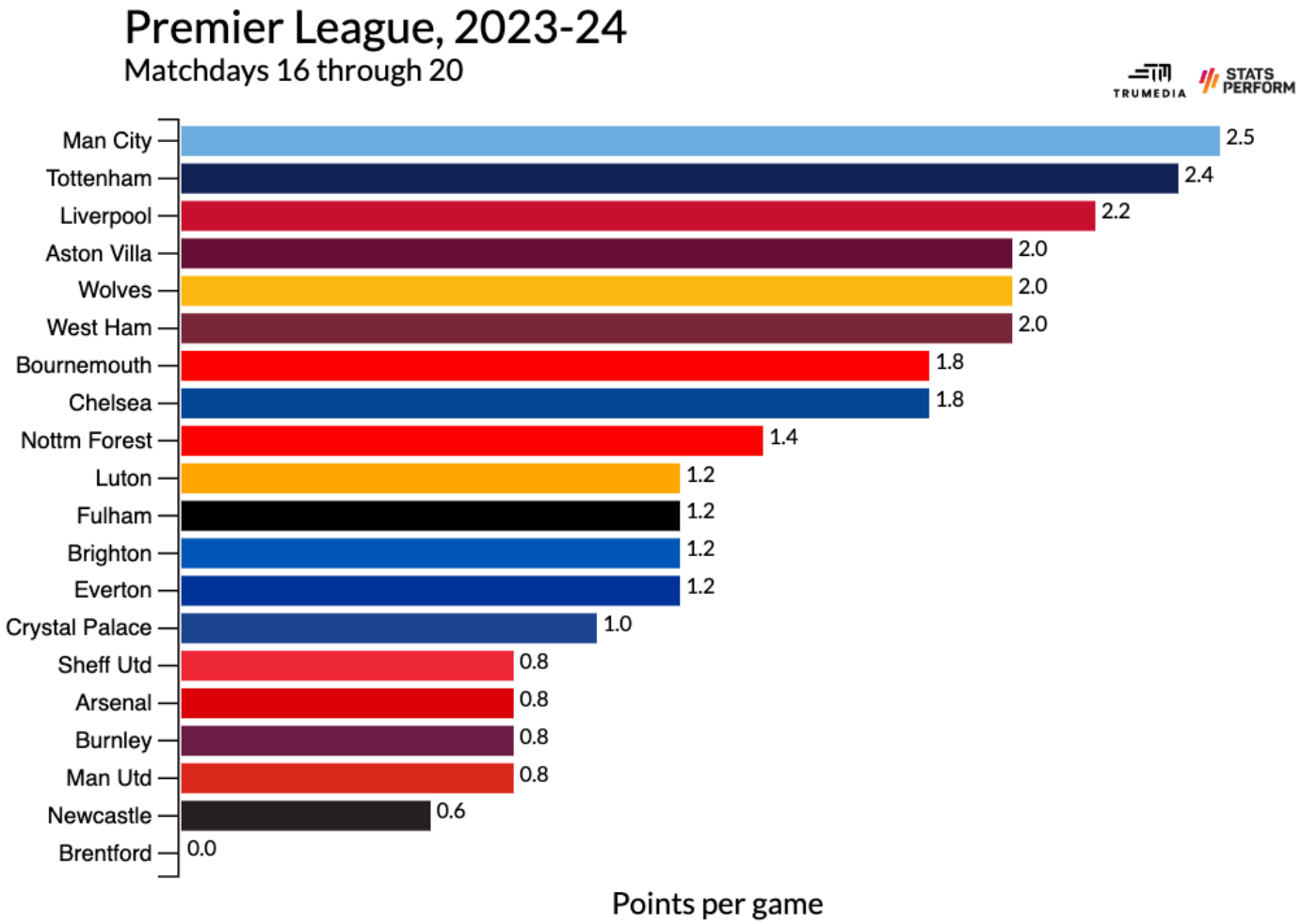
*Remove* labels with `NULL`.

Themes change the overall appearance of the background of your plot.

- Default is `theme_gray`
- Some nice ones are `theme_minimal` and `theme_classic`

We can also specify the font size and family.

Here is the list of ggplot themes.




Recreate this graphic using the partial dataframe in the .Rmd.

It can be difficult to view many overlapping plots. **Faceting** splits each plot onto its own panel.

Facet based on one variable with `facet_wrap`, or two variables with `facet_grid`.

- Need to specify `vars()`.

Let's explore different ways to facet the duration data. Note that R will always try to fill in every spot of `facet_grid`.

Let's create a new column for whether a year is a leap year.

How can we use faceting to explore trends in duration across both century and leap year?

Consider the `facet_grid` graph we just made.

- The bottom right panel shows the durations among (leap years/non-leap years) in the (19th/20th/21st) century.
- We don't expect there to be a difference in duration between non-leap years and leap years. So, each (row/column) has roughly the same center across its panels.
- We expect there to be a difference in duration across centuries. So, each (row/column) has different centers across its panels.

The `economics` dataset in `ggplot2` describes the US economy over time.

Make *any kind* of ggplot you want to. What questions might be interesting to answer? How can we visualize them?

- Note that dates can appear on a numeric axis
- Feel free to search for other techniques

Post your plot under Discussions in Canvas!