CSSS508, Week 5

Importing, Exporting, and Cleaning Data

Chuck Lanfear

May 1, 2019

Updated: Mar 28, 2019

UW CS&SS

Today's Theme:

"Data Custodian Work"

Issues around getting data *in* and *out* of R and making it analytically *ready*:

- Working directories and projects
- Importing and exporting data: `readr` and `haven`
- Cleaning and reshaping data: `tidyr`
- Dates and times: `lubridate`
- Controlling factor variables: `forcats`

Directories

Working Directory

You may recall that the **working directory** is where R will look for and save things by default.

You can find out what it is using the function `getwd()`.

On my computer when I knitted these slides, it happened to be:

```
getwd()
```

```
## [1] "C:/Users/cclan/OneDrive/GitHub/CSSS508/Lectures/Week5"
```

Changing Your Working Directory

You can use `setwd(dir = "C:/path/to/new/working/directory")` to change the working directory.

Working Directory Suggestions:

- `.Rmd` files use their current directory as a working directory: Just put everything you need in there!
- For larger projects, instead of setting a working directory, it is usually better to use RStudio projects to manage working directories.
- Windows users: If you copy a path from Explorer, make sure to change back slashes (`\`) to forward slashes (`/`) for the filepaths
- If you need to set a working, put `setwd()` at the start of your file so that someone using another computer knows they need to modify it

Projects in RStudio

A better way to deal with working directories: RStudio's **project** feature in the top-right dropdown. This has lots of advantages:

- Sets your working directory to be the project directory.
- Remembers objects in your workspace, command history, etc. next time you re-open that project.
- Reduces risk of intermingling different work using the same variable names (e.g. `n`) by using separate RStudio instances for each project.
- Easy to integrate with version control systems (e.g. `git`)
    - I usually make each RStudio project its own GitHub repository.

If you're interested in advanced project management, ask me after class or check out my presentation on reproducible research with rrtools.

Relative Paths

Once you've set the working directoryâ€”or you're in an RStudio projectâ€”you can refer to folders and files within the working directory using relative paths.

```
library(ggplot2)
a_plot <- ggplot(data = cars, aes(x = speed, y = dist)) +
    geom_point()
ggsave("graphics/cars_plot.png", plot = a_plot)
```

The above would save an image called "cars_plot.png" inside an existing folder called "graphics" within my working directory.

Relative paths are nice, because all locations of loaded and saved files can be changed just by altering the working directory.

Relative paths also allow others to download your files or entire project and use them on their computer without modifying all the paths!

Importing and Exporting Data

Special Data Access Packages

If you are working with a popular data source, try Googling to see if it has a devoted R package on *CRAN* or *Github* (use `devtools::install_github("user/repository")` for these). Examples:

- `WDI`: World Development Indicators (World Bank)
- `WHO`: World Health Organization API
- `tidycensus`: Census and American Community Survey [1]
- `quantmod`: financial data from Yahoo, FRED, Google

[1] We'll use this in our lecture on geographical data!

Delimited Text Files

Besides a package, the easiest way to work with external data is for it to be stored in a *delimited* text file, e.g. comma-separated values (**.csv**) or tab-separated values (**.tsv**). Here is `csv` data:

```
"Subject","Depression","Sex","Week","HamD","Imipramine"
101,"Non-endogenous","Male",0,26,NA
101,"Non-endogenous","Male",1,22,NA
101,"Non-endogenous","Male",2,18,4.04305
101,"Non-endogenous","Male",3,7,3.93183
101,"Non-endogenous","Male",4,4,4.33073
101,"Non-endogenous","Male",5,3,4.36945
103,"Non-endogenous","Female",0,33,NA
103,"Non-endogenous","Female",1,24,NA
103,"Non-endogenous","Female",2,15,2.77259
```

`readr`

R has a variety of built-in functions for importing data stored in text files, like `read.table()` and `read.csv()`. I recommend using the versions in the `readr` package instead: `read_csv()`, `read_tsv()`, and `read_delim()`:

`readr` function features:

- Faster!
- Better defaults (e.g. doesn't automatically convert character data to factors)
- A *little* smarter about dates and times
- Handy function `problems()` you can run if there are errors
- Loading bars for large files

```
library(readr)
```

`readr` Importing Example

Let's import some data about song ranks on the Billboard Hot 100 in 2000:

```
billboard_2000_raw <- read_csv(file = "https://raw.githubusercontent.com/hadley/tidyr/master
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   artist = col_character(),
##   track = col_character(),
##   time = col_time(format = ""),
##   date.entered = col_date(format = ""),
##   wk66 = col_logical(),
##   wk67 = col_logical(),
##   wk68 = col_logical(),
##   wk69 = col_logical(),
##   wk70 = col_logical(),
##   wk71 = col_logical(),
##   wk72 = col_logical(),
##   wk73 = col_logical(),
##   wk74 = col_logical(),
##   wk75 = col_logical(),
##   wk76 = col_logical()
## )

## See spec(...) for full column specifications.
```

Did It Load?

Look at the data types for the last few columns:

```
str(billboard_2000_raw[, 65:ncol(billboard_2000_raw)])
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    317 obs. of  17 variables:
##  $ wk60: num  NA NA NA NA NA NA NA NA NA NA ...
##  $ wk61: num  NA NA NA NA NA NA NA NA NA NA ...
##  $ wk62: num  NA NA NA NA NA NA NA NA NA NA ...
##  $ wk63: num  NA NA NA NA NA NA NA NA NA NA ...
##  $ wk64: num  NA NA NA NA NA NA NA NA NA NA ...
##  $ wk65: num  NA NA NA NA NA NA NA NA NA NA ...
##  $ wk66: logi  NA NA NA NA NA NA ...
##  $ wk67: logi  NA NA NA NA NA NA ...
##  $ wk68: logi  NA NA NA NA NA NA ...
##  $ wk69: logi  NA NA NA NA NA NA ...
##  $ wk70: logi  NA NA NA NA NA NA ...
##  $ wk71: logi  NA NA NA NA NA NA ...
##  $ wk72: logi  NA NA NA NA NA NA ...
##  $ wk73: logi  NA NA NA NA NA NA ...
##  $ wk74: logi  NA NA NA NA NA NA ...
##  $ wk75: logi  NA NA NA NA NA NA ...
##  $ wk76: logi  NA NA NA NA NA NA ...
```

What Went Wrong?

`readr` uses the values in the first 1000 rows to guess the type of the column (integer, logical, numeric, character). There are not many songs in the data that charted for 60+ weeksâ€”and none in the first 1000 that charted for 66+ weeks!

Since it encountered no values, `readr` assumed the `wk66`-`wk76` columns were *character* to be sure nothing would be lost. Use the `col_types` argument to fix this:

```
# paste is a string concatenation function
# i = integer, c = character, D = date
# rep("i", 76) does the 76 weeks of integer ranks
bb_types <- paste(c("icccD", rep("i", 76)), collapse="")

billboard_2000_raw <-
  read_csv(file = "https://raw.githubusercontent.com/hadley/tidyr/mas
           col_types = bb_types)
```

Alternate Solutions

You could also deal with this by adjusting the maximum rows used by `readr` to guess column types:

```
read_csv(file, guess_max=5000) # Default is 1000
```

Or you could use `read.csv()` in the `foreign` package. This is a base R alternative that is slower and a bit dumber.

With `read.csv()` you will need to specify if the data have column names (`header=TRUE`) and you'll want to use `stringsAsFactors=FALSE` to prevent character columns from becoming factors. `read_csv()` does this for us!

Excel Files

The simplest thing to do with Excel files (`.xls` or `.xlsx`) is open them up, export to CSV, then import in Râ€”and compare carefully to make sure everything worked!

For Excel files that might get updated and you want the changes to flow to your analysis, I recommend using an R package such as `readxl` or `openxlsx`. For Google Docs Spreadsheets, there's the `googlesheets` package.

You won't keep text formatting, color, comments, or merged cells so if these mean something in your data (*bad!*), you'll need to get creative.

Writing Delimited Files

Getting data out of R into a delimited file is very similar to getting it into R:

```
write_csv(billboard_2000_raw, path = "billboard_data.csv")
```

This saved the data we pulled off the web in a file called `billboard_data.csv` in my working directory.

Saving in R Formats

Exporting to a `.csv` drops R metadata, such as whether a variable is a character or factor. You can save objects (data frames, lists, etc.) in R formats to preserve this.

- `.Rds` format:
  - Used for single objects, doesn't save original the object name
  - Save: `write_rds(old_object_name, "path.Rds")`
  - Load: `new_object_name <- read_rds("path.Rds")`
- `.Rdata` or `.Rda` format:
  - Used for saving multiple files where the original object names are preserved
  - Save: `save(object1, object2, ... , file = "path.Rdata")`
  - Load: `load("path.Rdata")` *without assignment operator*

I pretty much always just save as `.Rdata`.

dput()

For asking for help, it is useful to prepare a snippet of your data with `dput()`:

```
dput(head(cars, 8))
```

```
## structure(list(speed = c(4, 4, 7, 7, 8, 9, 10, 10), dist = c(2,
## 10, 4, 22, 16, 10, 18, 26)), row.names = c(NA, 8L), class = "data.frame")
```

The output of `dput()` can be copied and assigned to an object in R:

```
temp <- structure(list(speed = c(4, 4, 7, 7, 8, 9, 10, 10),
                       dist = c(2, 10, 4, 22, 16, 10, 18, 26)),
                       .Names = c("speed", "dist"),
                       row.names = c(NA, 8L), class = "data.frame")
```

Reading in Data from Other Software

Working with **Stata** or **SPSS** users? You can use a package to bring in their saved data files:

- `haven` for Stata, SPSS, and SAS.
  - Part of the `tidyverse` family
- `foreign` for Stata, SPSS, Minitab
  - Part of base R

For less common formats, Google it. I've yet to encounter a data format without an R package to handle it (or at least a clever hack).

If you encounter a mysterious file extension (e.g. `.dat`), try opening it with a good text editor first (e.g. Atom or Sublime); there's a good chance it is actually raw text with a delimiter or fixed format that R can handle!

Cleaning Data

Initial Spot Checks

First things to check after loading new data:

- Did the last rows/columns from the original file make it in?
  - May need to use different package or manually specify range
- Are the column names in good shape?
  - Modify a `col_names=` argument or fix with `rename()`
- Are there "decorative" blank rows or columns to remove?
  - `filter()` or `select()` out those rows/columns
- How are missing values represented: `NA`, `" "` (blank), `.` (period), `999`?
  - Use `mutate()` with `ifelse()` to fix these (perhaps *en masse* with looping)
- Are there character data (e.g. ZIP codes with leading zeroes) being incorrectly represented as numeric or vice versa?
  - Modify `col_types=` argument, or use `mutate()` and `as.numeric()`

Slightly Messy Data

Program	Female	Male
Evans School	10	6
Arts & Sciences	5	6
Public Health	2	3
Other	5	1

- What is an observation?
  - A group of students from a program of a given gender
- What are the variables?
  - Program, gender
- What are the values?
  - Program: Evans School, Arts & Sciences, Public Health, Other
  - Gender: Female, Male -- **in the column headings, not its own column!**
  - Count: **spread over two columns!**

Tidy Version

Program	Gender	Count
Evans School	Female	10
Evans School	Male	6
Arts & Sciences	Female	5
Arts & Sciences	Male	6
Public Health	Female	2
Public Health	Male	3
Other	Female	5
Other	Male	1

Each variable is a column.

Each observation is a row.

Ready to throw into `ggplot()`!

Billboard is Just Ugly-Messy

year	artist	track	time	date.entered	wk1	wk2	wk3	wk4	wk5
2000	2 Pac	Baby Don't Cry (Keep...	4:22	2000-02-26	87	82	72	77	87
2000	2Ge+her	The Hardest Part Of ...	3:15	2000-09-02	91	87	92	NA	NA
2000	3 Doors Down	Kryptonite	3:53	2000-04-08	81	70	68	67	66
2000	3 Doors Down	Loser	4:24	2000-10-21	76	76	72	69	67
2000	504 Boyz	Wobble Wobble	3:35	2000-04-15	57	34	25	17	17
2000	98^0	Give Me Just One Nig...	3:24	2000-08-19	51	39	34	26	26
2000	A*Teens	Dancing Queen	3:44	2000-07-08	97	97	96	95	100
2000	Aaliyah	I Don't Wanna	4:15	2000-01-29	84	62	51	41	38
2000	Aaliyah	Try Again	4:03	2000-03-18	59	53	38	28	21
2000	Adams, Yolanda	Open My Heart	5:30	2000-08-26	76	76	74	69	68
2000	Adkins, Trace	More	3:05	2000-04-29	84	84	75	73	73
2000	Aguilera, Christina	Come On Over Baby (A...	3:38	2000-08-05	57	47	45	29	23

Week columns continue up to `wk76`!

Billboard

- What are the **observations** in the data?
    - Week since entering the Billboard Hot 100 per song
- What are the **variables** in the data?
    - Year, artist, track, song length, date entered Hot 100, week since first entered Hot 100 (**spread over many columns**), rank during week (**spread over many columns**)
- What are the **values** in the data?
    - e.g. 2000; 3 Doors Down; Kryptonite; 3 minutes 53 seconds; April 8, 2000; Week 3 (**stuck in column headings**); rank 68 (**spread over many columns**)

Tidy Data

Tidy data (aka "long data") are such that:

1. The values for a single observation are in their own row.
2. The values for a single variable are in their own column.
3. The observations are all of the same nature.

Why do we want tidy data?

- Easier to understand many rows than many columns
- Required for plotting in `ggplot2`
- Required for many types of statistical procedures (e.g. hierarchical or mixed effects models)
- Fewer confusing variable names
- Fewer issues with missing values and "imbalanced" repeated measures data

tidyr

The `tidyr` package provides functions to tidy up data, similar to `reshape` in Stata or `varstocases` in SPSS. Key functions:

- `gather()`: takes a set of columns and rotates them down to make two new columns (which you can name yourself):
    - A `key` that stores the original column names
    - A `value` with the values in those original columns
- `spread()`: inverts `gather()` by taking two columns and rotating them up into multiple columns
- `separate()`: pulls apart one column into multiple columns (common with `gather`ed data where values had been embedded in column names)
    - `extract_numeric()` does a simple version of this for the common case when you just want grab the number part
- `unite()`: inverts `separate()` by gluing together multiple columns into one character column (less common)

`gather()`

Let's use `gather()` to get the week and rank variables out of their current layout into two columns (big increase in rows, big drop in columns):

```
library(dplyr)
library(tidyr)
billboard_2000 <- billboard_2000_raw %>%
    gather(key = week, value = rank, starts_with("wk"))
dim(billboard_2000)
```

```
## [1] 24092     7
```

`starts_with()` and other helper functions from `dplyr::select()` work here too.

We could instead use: `gather(key = week, value = rank, wk1:wk76)` to pull out these contiguous columns.

`gather`ed Weeks

```
head(billboard_2000)
```

```
## # A tibble: 6 x 7
##    year artist       track                   time  date.entered week   rank
##   <int> <chr>        <chr>                   <chr> <date>       <chr> <int>
## 1  2000 2 Pac        Baby Don't Cry (Keep... 4:22  2000-02-26   wk1      87
## 2  2000 2Ge+her      The Hardest Part Of ... 3:15  2000-09-02   wk1      91
## 3  2000 3 Doors Down Kryptonite              3:53  2000-04-08   wk1      81
## 4  2000 3 Doors Down Loser                   4:24  2000-10-21   wk1      76
## 5  2000 504 Boyz     Wobble Wobble           3:35  2000-04-15   wk1      57
## 6  2000 98^0         Give Me Just One Nig... 3:24  2000-08-19   wk1      51
```

Now we have a single week column!

Gathering Better?

```
summary(billboard_2000$rank)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##       1      26      51      51      76     100   18785 
```

This is an improvement, but we don't want to keep the 18785 rows with missing ranks (i.e. observations for weeks since entering the Hot 100 that the song was no longer on the Hot 100).

Gathering Better: `na.rm`

The argument `na.rm = TRUE` to `gather()` will remove rows with missing ranks.

```
billboard_2000 <- billboard_2000_raw %>%
    gather(key = week, value = rank, starts_with("wk"),
           na.rm = TRUE)
summary(billboard_2000$rank)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0    26.0    51.0    51.1    76.0   100.0
```

`separate()`

The track length column isn't analytically friendly. Let's convert it to a number rather than the character (minutes:seconds) format:

```
billboard_2000 <- billboard_2000 %>%
    separate(time, into = c("minutes", "seconds"),
             sep = ":", convert = TRUE) %>%
    mutate(length = minutes + seconds / 60) %>%
    select(-minutes, -seconds)
summary(billboard_2000$length)
```

```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     2.60    3.67    3.93    4.03    4.28    7.83
```

`sep = :` tells `separate()` to split the column into two where it finds a colon (`:`).

Then we add `seconds / 60` to `minutes` to produce a numeric `length` in minutes.

`parse_number()`

`tidyr` provides a convenience function to grab just the numeric information from a column that mixes text and numbers:

```
billboard_2000 <- billboard_2000 %>%
    mutate(week = parse_number(week))
summary(billboard_2000$week)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0     5.0    10.0    11.5    16.0    65.0
```

For more sophisticated conversion or pattern checking, you'll need to use string parsing (to be covered in week 8).

`spread()` Motivation

`spread()` is the opposite of `gather()`, which you use if you have data for the same observation taking up multiple rows.

Example of data that we probably want to spread (unless we want to plot each statistic in its own facet):

Group	Statistic	Value
A	Mean	1.28
A	Median	1.0
A	SD	0.72
B	Mean	2.81
B	Median	2
B	SD	1.33

A common cue to use `spread()` is having measurements of different quantities in the same column.

Before `spread()`

```
(too_long_data <- data.frame(Group = c(rep("A", 3), rep("B", 3)),
                             Statistic = rep(c("Mean", "Median", "SD"), 2),
                             Value = c(1.28, 1.0, 0.72, 2.81, 2, 1.33)))
```

```
##   Group Statistic Value
## 1     A      Mean  1.28
## 2     A    Median  1.00
## 3     A        SD  0.72
## 4     B      Mean  2.81
## 5     B    Median  2.00
## 6     B        SD  1.33
```

After `spread()`

```
(just_right_data <- too_long_data %>%
    spread(key = Statistic, value = Value))
```

```
##   Group Mean Median   SD
## 1     A 1.28      1 0.72
## 2     B 2.81      2 1.33
```

Charts of 2000: Data Prep

Let's look at songs that hit #1 at some point and look how they got there versus songs that did not:

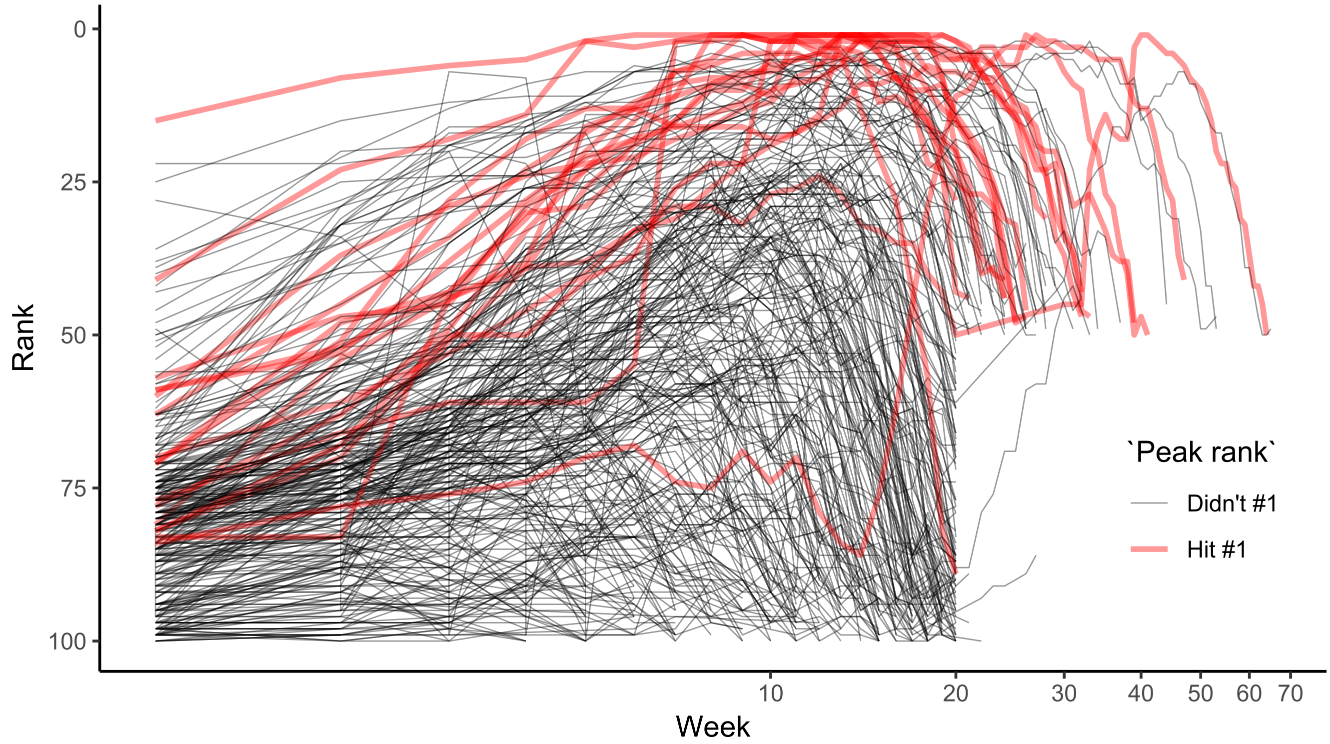
```
# find best rank for each song
best_rank <- billboard_2000 %>%
    group_by(artist, track) %>%
    summarize(min_rank = min(rank),
              weeks_at_1 = sum(rank == 1)) %>%
    mutate(`Peak rank` = ifelse(min_rank == 1,
                                "Hit #1",
                                "Didn't #1"))

# merge onto original data
billboard_2000 <- billboard_2000 %>%
    left_join(best_rank, by = c("artist", "track"))
```

Note that because the "highest" rank is *numerically lowest* (1), we are summarizing with `min()`.

Charts of 2000: `ggplot2`

```r
library(ggplot2)
billboard_trajectories <-
  ggplot(data = billboard_2000,
         aes(x = week, y = rank, group = track,
             color = `Peak rank`)
         ) +
  geom_line(aes(size = `Peak rank`), alpha = 0.4) +
    # rescale time: early weeks more important
  scale_x_log10(breaks = seq(0, 70, 10)) +
  scale_y_reverse() + # want rank 1 on top, not bottom
  theme_classic() +
  xlab("Week") + ylab("Rank") +
  scale_color_manual(values = c("black", "red")) +
  scale_size_manual(values = c(0.25, 1)) +
  theme(legend.position = c(0.90, 0.25),
        legend.background = element_rect(fill="transparent"))
```

Charts of 2000: Beauty!



Observation: There appears to be censoring around week 20 for songs falling out of the top 50 that I'd want to follow up on.

Which Were #1 the Most Weeks?

```
billboard_2000 %>%
    select(artist, track, weeks_at_1) %>%
    distinct(artist, track, weeks_at_1) %>%
    arrange(desc(weeks_at_1)) %>%
    head(7)
```

```
## # A tibble: 7 x 3
##   artist             track                   weeks_at_1
##   <chr>              <chr>                        <int>
## 1 Destiny's Child    Independent Women Pa...         11
## 2 Santana            Maria, Maria                    10
## 3 Aguilera, Christina Come On Over Baby (A...         4
## 4 Madonna            Music                            4
## 5 Savage Garden      I Knew I Loved You               4
## 6 Destiny's Child    Say My Name                      3
## 7 Iglesias, Enrique  Be With You                      3
```

Dates and Times

Getting Usable Dates

We have the date the songs first charted, but not the dates for later weeks. We can calculate these now that the data are tidy:

```
billboard_2000 <- billboard_2000 %>%
    mutate(date = date.entered + (week - 1) * 7)
billboard_2000 %>% arrange(artist, track, week) %>%
    select(artist, date.entered, week, date, rank) %>% head(4)
```

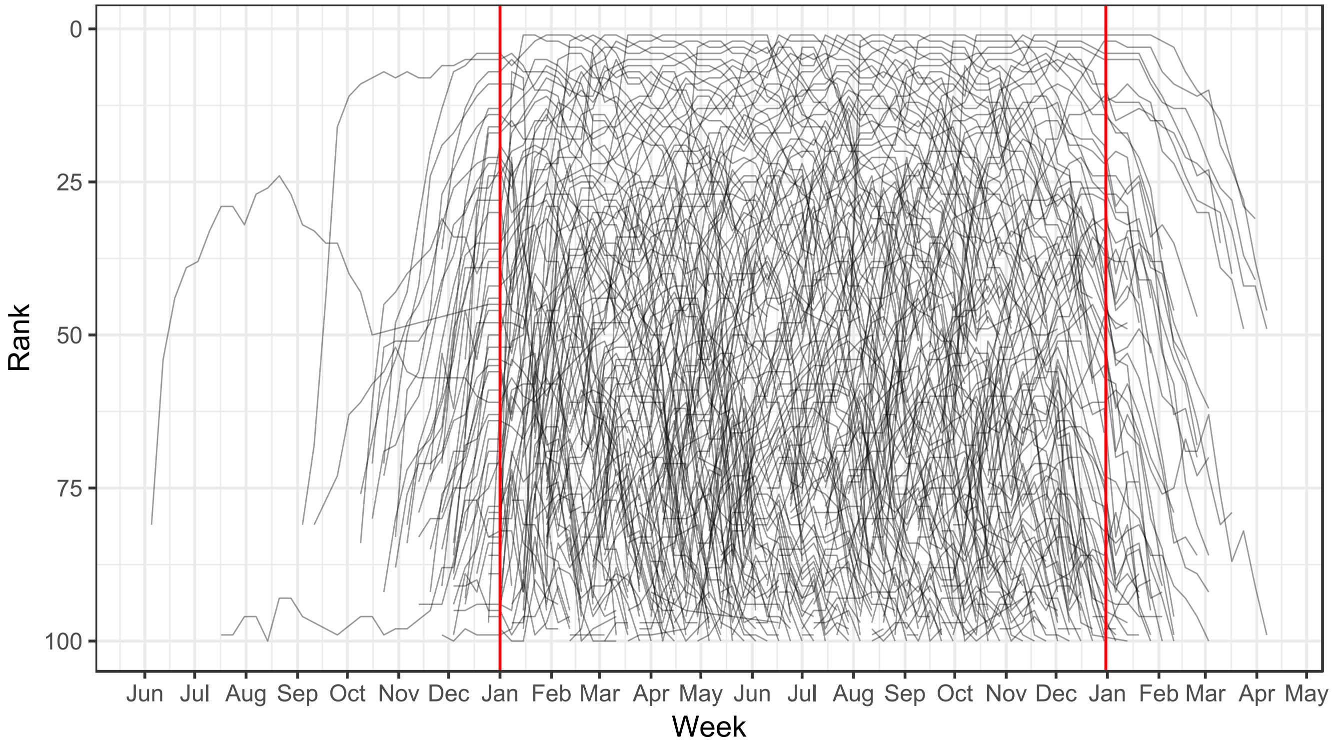
```
## # A tibble: 4 x 5
##   artist date.entered  week date        rank
##   <chr>  <date>       <dbl> <date>     <int>
## 1 2 Pac  2000-02-26       1 2000-02-26    87
## 2 2 Pac  2000-02-26       2 2000-03-04    82
## 3 2 Pac  2000-02-26       3 2000-03-11    72
## 4 2 Pac  2000-02-26       4 2000-03-18    77
```

This works because `date` objects are in units of daysâ€”we just add 7 days per week to the start date.

Preparing to Plot Over Calendar Time

```r
plot_by_day <-
  ggplot(billboard_2000, aes(x = date, y = rank, group = track)) +
  geom_line(size = 0.25, alpha = 0.4) +
  # just show the month abbreviation label (%b)
  scale_x_date(date_breaks = "1 month", date_labels = "%b") +
  scale_y_reverse() + theme_bw() +
  # add lines for start and end of year:
  # input as dates, then make numeric for plotting
  geom_vline(xintercept = as.numeric(as.Date("2000-01-01", "%Y-%m-%d")),
             col = "red") +
  geom_vline(xintercept = as.numeric(as.Date("2000-12-31", "%Y-%m-%d")),
             col = "red") +
  xlab("Week") + ylab("Rank")
```

Calendar Time Plot!



We see some of the entry dates are before 2000---presumably songs still charting during 2000 that came out earlier.

Dates and Times

To practice working with finer-grained temporal information, let's look at one day of Seattle Police response data obtained from data.seattle.gov:

```
spd_raw <- read_csv("https://raw.githubusercontent.com/clanfear/CSSS508/master/Seattle_Police_Department_911_Incide
```

```
## Parsed with column specification:
## cols(
##   `CAD CDW ID` = col_double(),
##   `CAD Event Number` = col_double(),
##   `General Offense Number` = col_double(),
##   `Event Clearance Code` = col_character(),
##   `Event Clearance Description` = col_character(),
##   `Event Clearance SubGroup` = col_character(),
##   `Event Clearance Group` = col_character(),
##   `Event Clearance Date` = col_character(),
##   `Hundred Block Location` = col_character(),
##   `District/Sector` = col_character(),
##   `Zone/Beat` = col_character(),
##   `Census Tract` = col_double(),
##   Longitude = col_double(),
##   Latitude = col_double(),
##   `Incident Location` = col_character(),
##   `Initial Type Description` = col_character(),
##   `Initial Type Subgroup` = col_character(),
##   `Initial Type Group` = col_character(),
##   `At Scene Time` = col_character()
## )
```

SPD Data

```r
glimpse(spd_raw)
```

```
## Observations: 706
## Variables: 19
## $ `CAD CDW ID`                   <dbl> 1701856, 1701857, 1701853, 17018...
## $ `CAD Event Number`             <dbl> 1.6e+10, 1.6e+10, 1.6e+10, 1.6e+...
## $ `General Offense Number`       <dbl> 2.02e+09, 2.02e+09, 2.02e+09, 2....
## $ `Event Clearance Code`         <chr> "063", "064", "161", "245", "202...
## $ `Event Clearance Description`  <chr> "THEFT - CAR PROWL", "SHOPLIFT",...
## $ `Event Clearance SubGroup`     <chr> "CAR PROWL", "THEFT", "TRESPASS"...
## $ `Event Clearance Group`        <chr> "CAR PROWL", "SHOPLIFTING", "TRE...
## $ `Event Clearance Date`         <chr> "03/25/2016 11:58:30 PM", "03/25...
## $ `Hundred Block Location`       <chr> "S KING ST / 8 AV S", "92XX BLOC...
## $ `District/Sector`              <chr> "K", "S", "D", "M", "M", "B", "B...
## $ `Zone/Beat`                    <chr> "K3", "S3", "D2", "M1", "M3", "B...
## $ `Census Tract`                 <dbl> 9100, 11801, 7200, 8002, 8100, 5...
## $ Longitude                      <dbl> -122, -122, -122, -122, -122, -1...
## $ Latitude                       <dbl> 47.6, 47.5, 47.6, 47.6, 47.6, 47...
## $ `Incident Location`            <chr> "(47.598347, -122.32245)", "(47....
## $ `Initial Type Description`     <chr> "THEFT (DOES NOT INCLUDE SHOPLIF...
## $ `Initial Type Subgroup`        <chr> "OTHER PROPERTY", "SHOPLIFTING",...
## $ `Initial Type Group`           <chr> "THEFT", "THEFT", "TRESPASS", "C...
## $ `At Scene Time`                <chr> "03/25/2016 10:25:51 PM", "03/25...
```

lubridate

```
str(spd_raw$`Event Clearance Date`)
```

```
##  chr [1:706] "03/25/2016 11:58:30 PM" "03/25/2016 11:57:22 PM" ...
```

We want this to be in a date/time format ("POSIXct"), not character. We will work with dates using the `lubridate` package.

```
# install.packages("lubridate")
library(lubridate)
spd <- spd_raw %>%
  mutate(`Event Clearance Date` =
           mdy_hms(`Event Clearance Date`,
                   tz = "America/Los_Angeles"))
str(spd$`Event Clearance Date`)
```

```
##  POSIXct[1:706], format: "2016-03-25 23:58:30" "2016-03-25 23:57:22" ...
```

`mdy_hms()` automatically processes datetimes in month-day-year, hour-minute-second format. It figures out separators for you!

An Aside on Time

Time data are a bit weird.

R uses two primary formats for storing data on times and dates:

- `POSIXct`: Numeric vector of seconds since the beginning of 1970.
- `POSIXlt`: Named list of vectors containing lots of date/time information.

We usually work with `POSIXct`.

`lubridate` gives us many convenience functions for dealing with date/time data.

It is often easiest to just convert time to standard numeric values and work with it that way, however, particularly if it will be used as a variable in a statistical model.

Useful Date/Time Functions

```
demo_dts <- spd$`Event Clearance Date`[1:2]
(date_only <- as.Date(demo_dts, tz = "America/Los_Angeles"))
```

```
## [1] "2016-03-25" "2016-03-25"
```

```
(day_of_week_only <- weekdays(demo_dts))
```

```
## [1] "Friday" "Friday"
```

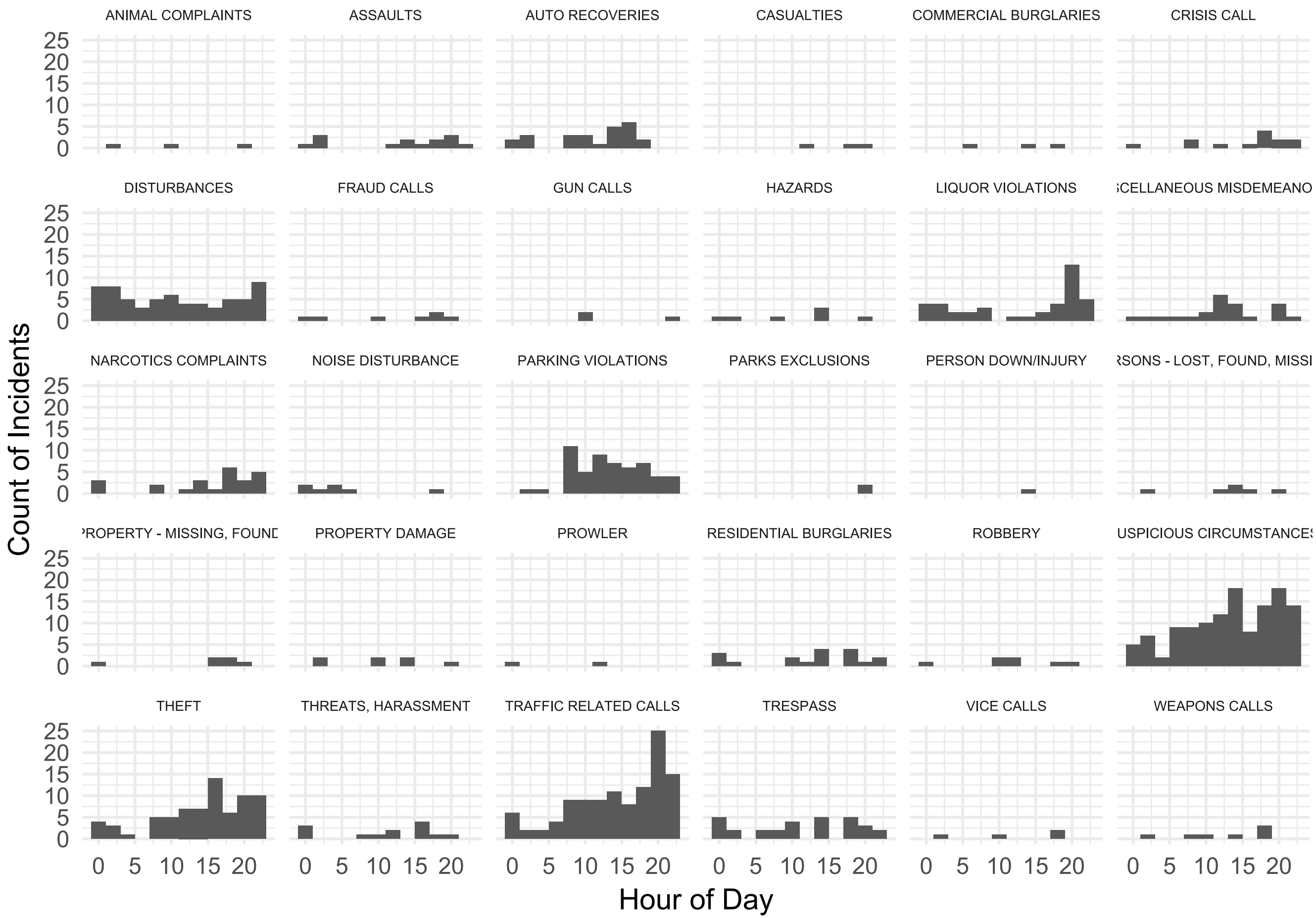
```
(one_hour_later <- demo_dts + dhours(1))
```

```
## [1] "2016-03-26 00:58:30 PDT" "2016-03-26 00:57:22 PDT"
```

What Time of Day were Incidents Cleared?

```
spd_times <- spd %>%
    select(`Initial Type Group`, `Event Clearance Date`) %>%
    mutate(hour = hour(`Event Clearance Date`))

time_spd_plot <- ggplot(spd_times, aes(x = hour)) +
    geom_histogram(binwidth = 2) +
    facet_wrap( ~ `Initial Type Group`) +
    theme_minimal() +
    theme(strip.text.x = element_text(size = rel(0.6))) +
    ylab("Count of Incidents") + xlab("Hour of Day")
```

SPD Event Clearances, March 25



Managing Factor Variables

Factor Variables

Factors are such a common (and fussy) vector type in R that we need to get to know them a little better when preparing data:

- The order of factor levels controls the order of categories in tables, on axes, in legends, and in facets in `ggplot2`.
    - Often we want to plot in interpretable/aesthetically pleasing order, e.g. from highest to lowest valuesâ€”not **"Alabama first"**.
- The lowest level of a factor is treated as a reference for regression, and the other levels get their own coefficients.
    - Reference levels are by default alphabetical, which doesn't necessarily coincide with the easiest to understand baseline category.

`forcats`

The `tidyverse` family of packages includes the package `forcats` (an anagram of "factors") that is "for cat(egorical)s".

This package supersedes the functionality of the base factor functions with somewhat more logical and uniform syntax.

To find more, look at the `forcats` manual.

Character to Factor

```r
# install.packages("forcats")
library(forcats)
str(spd_times$`Initial Type Group`)
```

```
##  chr [1:706] "THEFT" "THEFT" "TRESPASS" "CRISIS CALL" ...
```

```r
spd_times$`Initial Type Group` <-
  parse_factor(spd_times$`Initial Type Group`, levels=NULL)
str(spd_times$`Initial Type Group`)
```

```
##  Factor w/ 30 levels "THEFT","TRESPASS",..: 1 1 2 3 4 5 6 7 8 5 ...
```

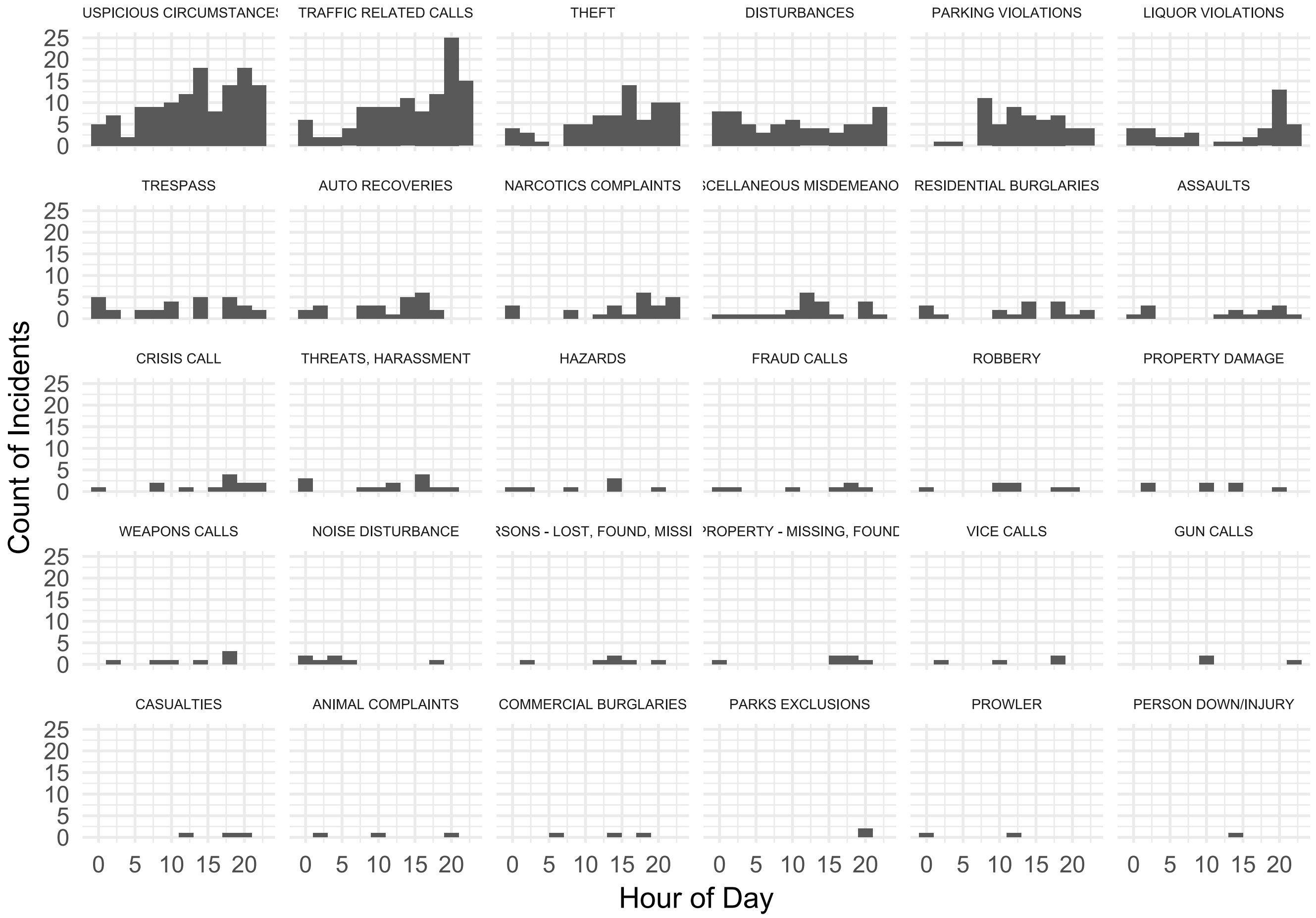
```r
head(as.numeric(spd_times$`Initial Type Group`))
```

```
## [1] 1 1 2 3 4 5
```

Releveling by Frequency

```
spd_vol <- spd_times %>%
  group_by(`Initial Type Group`) %>%
  summarize(n_events = n()) %>%
  arrange(desc(n_events))

# set levels using order from sorted frequency table
spd_times_2 <- spd_times %>%
  mutate(`Initial Type Group` =
         factor(`Initial Type Group`,
                levels = spd_vol$`Initial Type Group`))

# replot
time_spd_plot_2 <- ggplot(spd_times_2, aes(x = hour)) +
  geom_histogram(binwidth = 2) +
  facet_wrap( ~ `Initial Type Group`) +
  theme_minimal() +
  theme(strip.text.x = element_text(size = rel(0.6))) +
  ylab("Count of Incidents") + xlab("Hour of Day")
```

Better Ordered Plot



Other Ways to Reorder

Another way to reorder a factor is through the `fct_reorder()` function:

```
fct_reorder(factor_vector,
        quantity_to_order_by,
        function_to_apply_to_quantities_by_factor)
```

This is especially useful for making legends go from highest to lowest value visually using `max()` as your function, or making axis labels go from lowest to highest value using `mean()`.

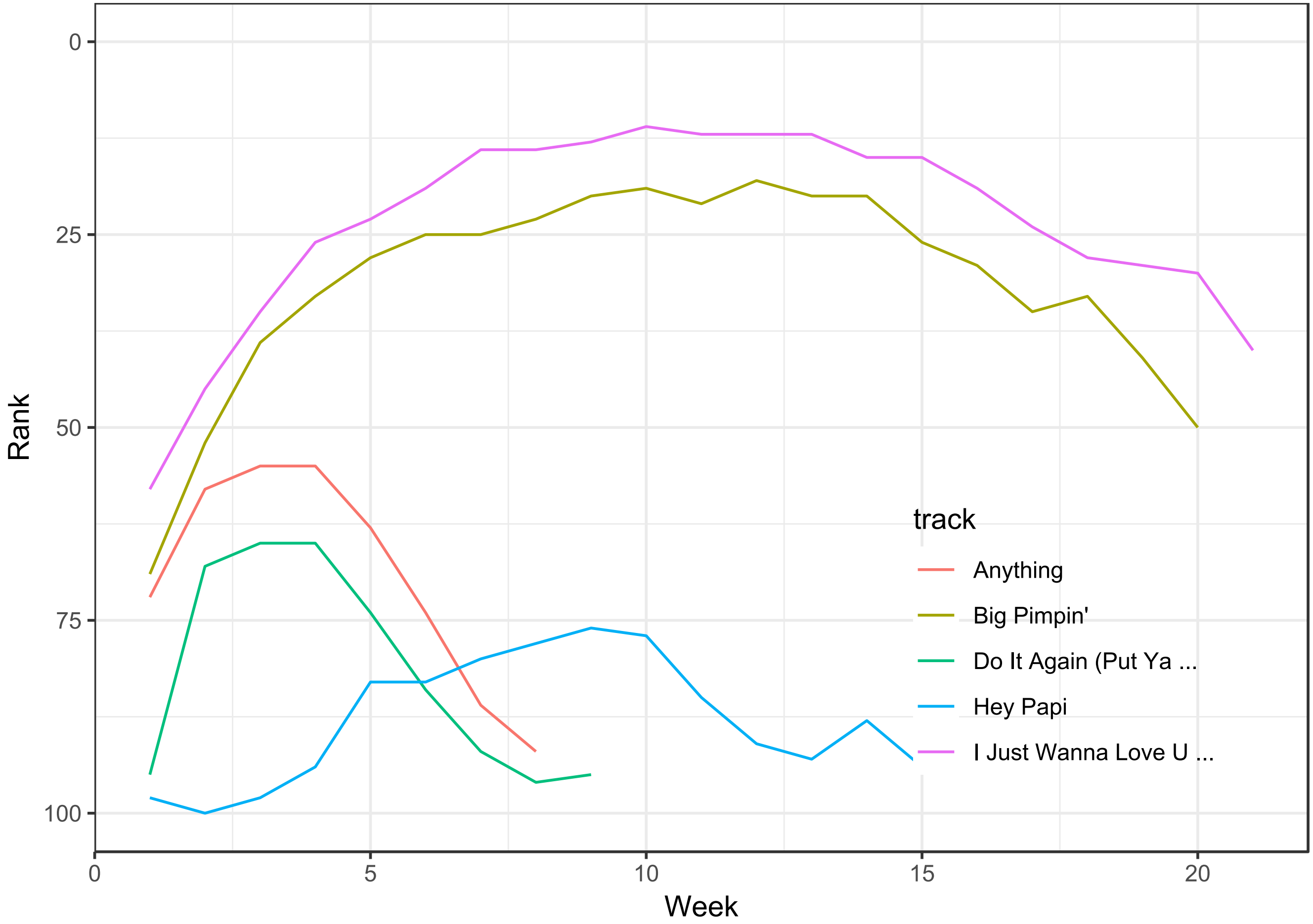
Use `fct_relevel()` and use the `ref=` argument to change the reference category

- Good when fitting regressions where you don't care about the overall ordering, just which level is the reference

Reorder Example: Jay-Z

```r
jayz <- billboard_2000 %>%
  filter(artist == "Jay-Z") %>%
  mutate(track = factor(track))

jayz_bad_legend <-
  ggplot(jayz, aes(x = week, y = rank,
                   group = track, color = track)) +
  geom_line() +
  theme_bw() +
  scale_y_reverse(limits = c(100, 0)) +
  theme(legend.position = c(0.80, 0.25),
        legend.background = element_rect(fill="transparent")) +
  xlab("Week") + ylab("Rank")
```

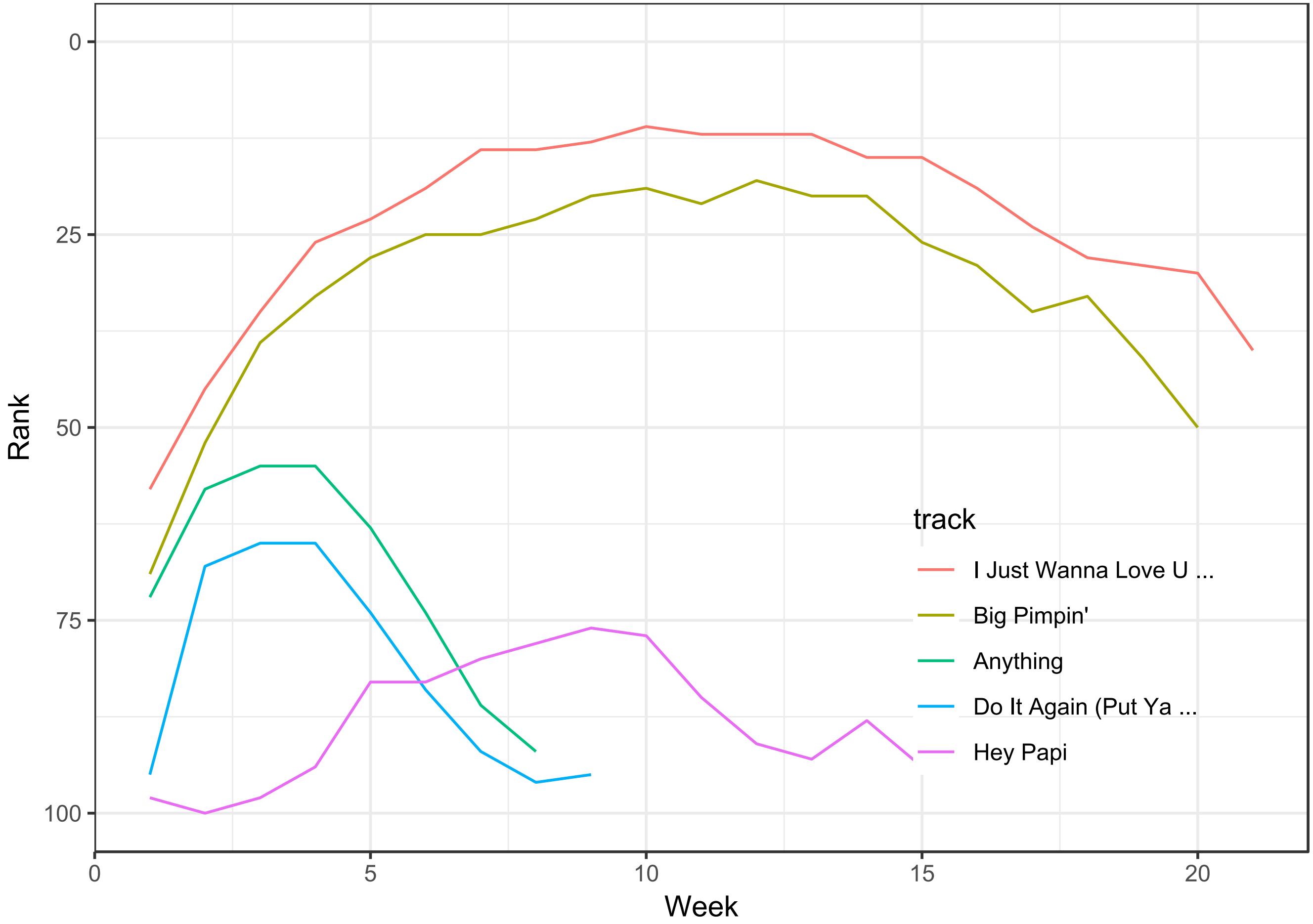
Jay-Z with Bad Legend Order



Better Ordering for Jay-Z

```
jayz <- jayz %>% mutate(track = fct_reorder(track, rank, min))

jayz_good_legend <-
  ggplot(jayz, aes(x = week, y = rank,
                   group = track, color = track)) +
  geom_line() +
  theme_bw() +
  scale_y_reverse(limits = c(100, 0)) +
  theme(legend.position = c(0.80, 0.25),
        legend.background = element_rect(fill="transparent")) +
  xlab("Week") + ylab("Rank")
```

Jay-Z with Good Legend Order



Dropping Unused Levels

After subsetting you can end up with fewer *realized* levels than before, but old levels remain linked and can cause problems for regressions. Drop unused levels from variables or your *entire data frame* using `droplevels()`.

```
jayz_biggest <- jayz %>%
  filter(track %in% c("I Just Wanna Love U ...", "Big Pimpin'"))
levels(jayz_biggest$track)
```

```
## [1] "I Just Wanna Love U ..." "Big Pimpin'"
## [3] "Anything"                "Do It Again (Put Ya ..."
## [5] "Hey Papi"
```

```
jayz_biggest <- jayz_biggest %>% droplevels(.)
levels(jayz_biggest$track)
```

```
## [1] "I Just Wanna Love U ..." "Big Pimpin'"
```

Homework

Vote tallies in King County from the 2016 general election are in a 60 MB comma-delimited text file downloaded from King County. They can be found on the course website.

The data have no documentation (aside from what I provide), so show your detective work to answer questions about the data and clean them up in the R Markdown template on the course website. Use *âŒ˜-Click* on Mac or *Right-Click* on Windows to download the .Rmd to the folder you plan to work from, then open it in RStudio.

This homework is two parts to be completed in each of the next two weeks. It can be daunting, so do not wait until Monday to start. I recommend reading instructions closely, working with others, and using the mailing list.

PART 1 DUE: 11:59 PM, May 1st

PART 2 DUE: 11:59 PM, May 8th