Hockey Analysis in R:
Public and Private Perspectives

Namita Nandakumar
@nnstats

Things I Have Recently Been:

- Philly sports fan
- Wharton undergrad
- avid public hockey analyst

Things I Am Now:

- Quantitative Analyst 📊 @ the Philadelphia Eagles 🏈
- #rstats + #tidyverse enthusiast
- lazy public hockey analyst on occasion

My R Credentials:

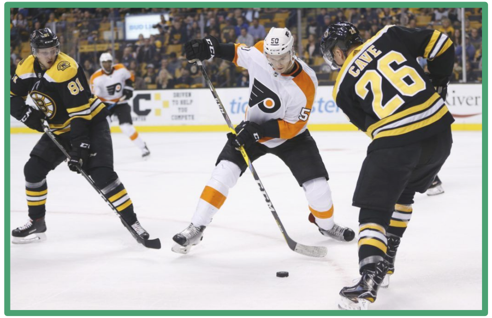

Namita @nnstats · 9 Dec 2018

sometimes Python evangelists get mad at me and it's like please neither of us have time for this, you need to get back to troubleshooting your package installation issues

56 382 2.7K

What is (ice) hockey?

- Violent soccer on ice.
- Very fast-paced and exciting.
- Always feels like a goal can be scored at any time.



How do I feel when I watch hockey?

- In general:
  Nervous.
- If my favorite team is losing:
  There's no way the other team blows this lead.
- If my favorite team is winning:
  *My team is **certainly** going to blow this lead.*

But I am often wrong about things.

- So let's dig into some NHL play-by-play data and try to find the truth about blowing leads and win probability.
- MoneyPuck is kind enough to post an updated .csv of shot data for every season, including 2018-19, which we can access directly in R:

```
library(tidyverse)
library(janitor)

temp <- tempfile()
download.file('http://peter-tanner.com/moneypuck/downloads/shots_2018.zip', temp)
pbp <- read_csv(unz(temp, 'shots_2018.csv'))
unlink(temp)
```

It’s time to put the #tidyverse to work.

```
pbp_clean <- pbp %>%
  clean_names() %>%
  # keep goals only
  filter(event == 'GOAL') %>%
  select(game_id, period, time, is_home_team, home_team_won, home_team_won,
         home_team = home_team_code, home_goals = home_team_goals,
         away_team = away_team_code, away_goals = away_team_goals) %>%
  # add goal that was just scored to running tally
  mutate(home_goals = (is_home_team == 1) + home_goals,
         away_goals = (is_home_team == 0) + away_goals) %>%
  group_by(game_id) %>%
  # no OT games
  filter(max(period) <= 3) %>%
  ungroup() %>%
  select(-c(is_home_team, period))
```

game_id	time	home_team_won	home_team	home_goals	away_team	away_goals
20002	24	1	WSH	1	BOS	0
20002	107	1	WSH	2	BOS	0
20002	1457	1	WSH	3	BOS	0
20002	1573	1	WSH	4	BOS	0
20002	1648	1	WSH	5	BOS	0
20002	2145	1	WSH	6	BOS	0
20002	3052	1	WSH	7	BOS	0

```
pbp_clean %>% View()
```

tidyr::expand() is the MVP of this code.

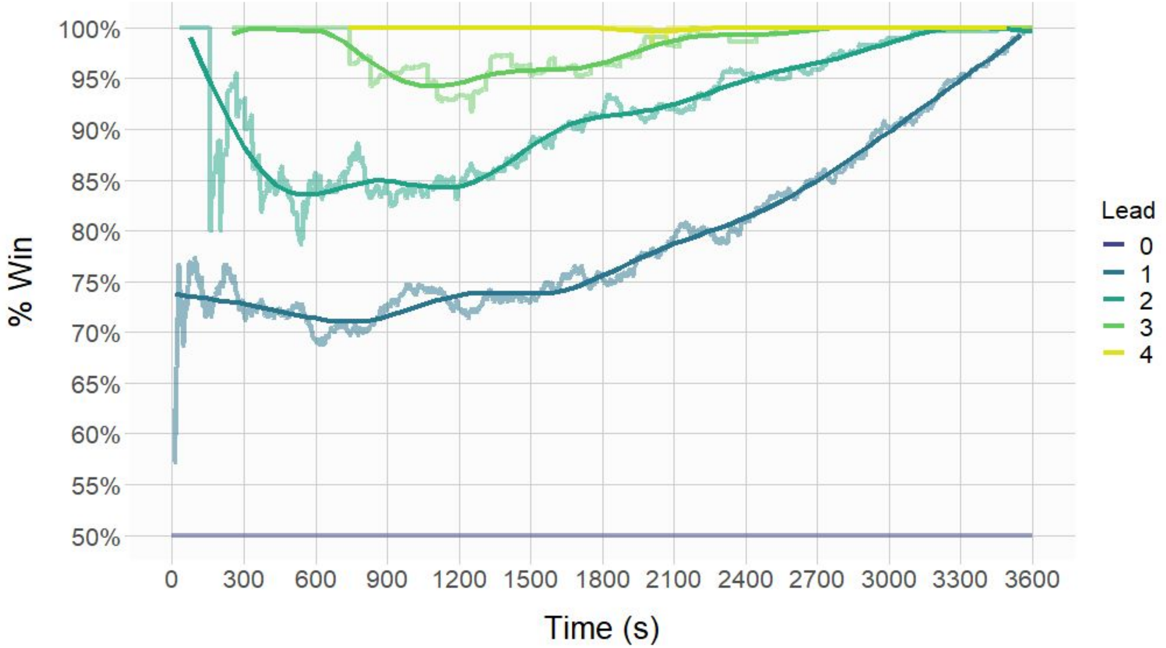
```
pbp_full <- pbp_clean %>%
  group_by(game_id) %>%
  # row for every second of every game
  expand(time = 0:3600) %>%
  left_join(pbp_clean, by=c('game_id', 'time')) %>%
  # fill in score info for times between goals
  fill(home_team_won, home_team, home_goals,
       away_team, away_goals, .direction='down') %>%
  fill(home_team_won, home_team, away_team, .direction='up') %>%
  # every game starts 0-0
  replace_na(list(home_goals = 0, away_goals = 0)) %>%
  # calculate magnitude of lead and indicator for whether leading team won
  mutate(lead = abs(home_goals - away_goals),
         lead_won = ifelse(home_goals > away_goals,
                           1*(home_team_won == 1), 1*(home_team_won == 0)),
         lead_won = replace(lead_won, lead == 0, 0.5)) %>%
  ungroup()
```

game_id	time	home_team_won	home_team	home_goals	away_team	away_goals	lead	lead_won
20002	19	1	WSH	0	BOS	0	0	0.5
20002	20	1	WSH	0	BOS	0	0	0.5
20002	21	1	WSH	0	BOS	0	0	0.5
20002	22	1	WSH	0	BOS	0	0	0.5
20002	23	1	WSH	0	BOS	0	0	0.5
20002	24	1	WSH	1	BOS	0	1	1.0
20002	25	1	WSH	1	BOS	0	1	1.0
20002	26	1	WSH	1	BOS	0	1	1.0
20002	27	1	WSH	1	BOS	0	1	1.0
20002	28	1	WSH	1	BOS	0	1	1.0
20002	29	1	WSH	1	BOS	0	1	1.0

```
pbp_full %>% slice(20:30) %>% View()
```

Empirical probabilities.

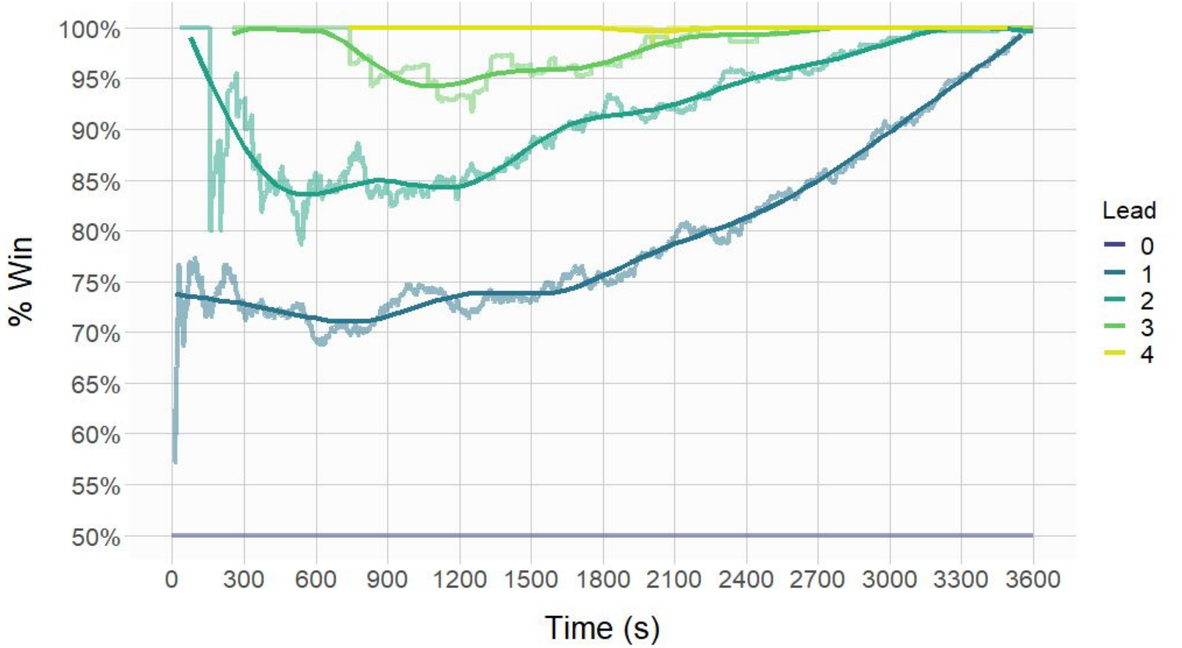
```
pbp_full %>%
  mutate(lead = replace(lead, lead > 4, 4)) %>%
  group_by(lead, time) %>%
  summarize(empirical_prob = mean(lead_won)) %>%
  ggplot(...
```



Empirical probabilities.

- Interesting but kind of unsatisfying.
- There should probably be monotonic + smooth relationships between time left and P(win), holding lead constant.
- Not sure if standard smoothing procedures can fix this.

```
pbp_full %>%
  mutate(lead = replace(lead, lead > 4, 4)) %>%
  group_by(lead, time) %>%
  summarize(empirical_prob = mean(lead_won)) %>%
  ggplot(...
```



One Weird Hack to Ensure Monotonicity: xgboost.

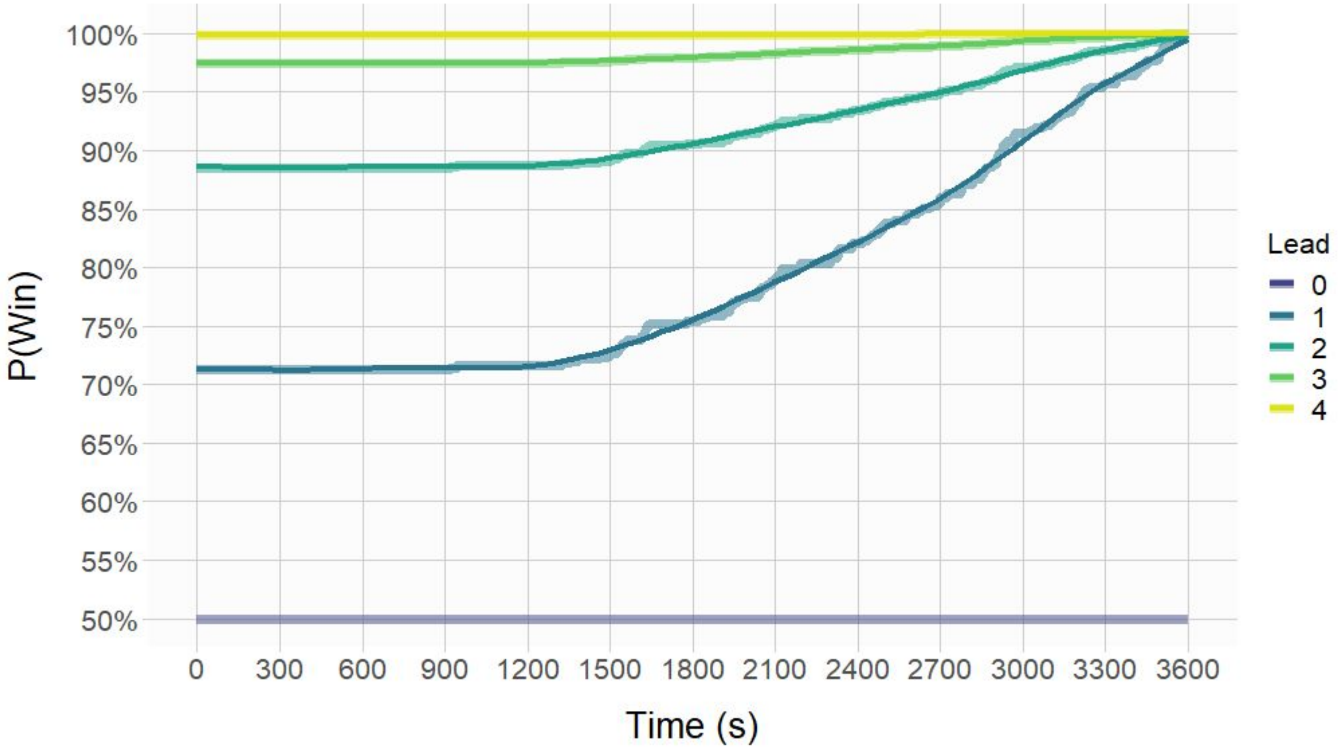
- Get P(win) of 2018-19 regular season games based on lead, time.
- We don't mind getting as close as possible to the empirical percentages, subject to a couple constraints.

```
param <- list(objective = 'binary:logistic',
              eval_metric = 'logloss',
              max_depth = 1,
              eta = 0.01,
              monotone_constraints = c(1, 1))
```

would normally overfit to hell and back

ensures that a larger lead and/or being later in the game is always better

Let's take a look at our "predictions."

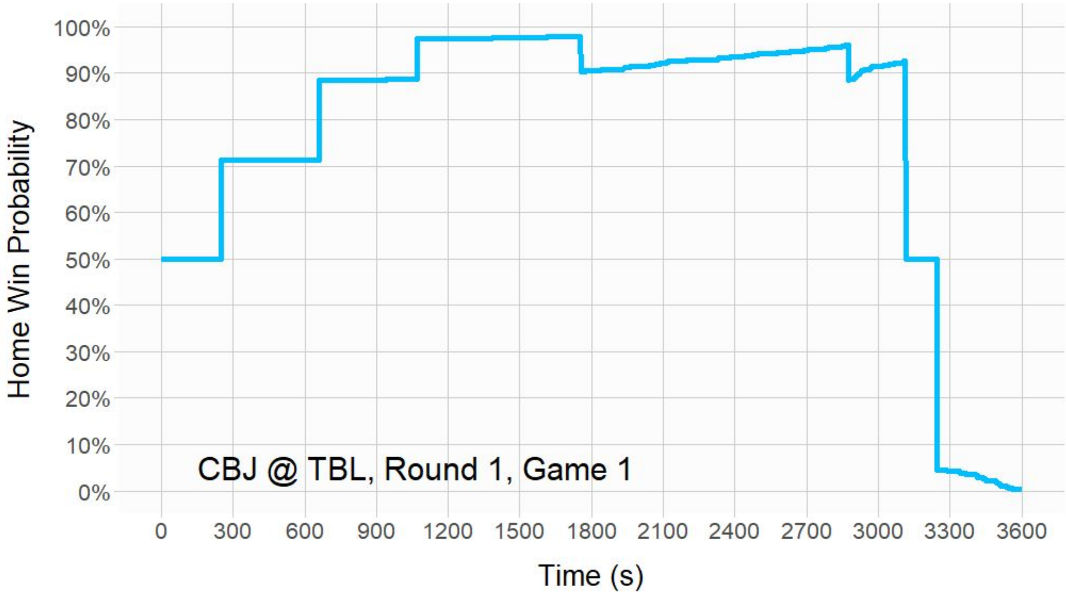


Now, let’s find the most exciting 2019 playoff game.

- One idea: identify the largest spread between maximum and minimum team win probability throughout any playoff game (that ended in regulation).

Now, let's find the most exciting 2019 playoff game.

- One idea: identify the largest spread between maximum and minimum team win probability throughout any playoff game (that ended in regulation).



What’s the point?

- Honestly, there isn’t one, aside from satisfying my curiosity.
- It’s not really useful for teams, but not everything has to be!
- Still, part of my perspective is exploring what teams find actionable.
- So, let’s switch gears.

Rapid Fire NHL Draft Facts

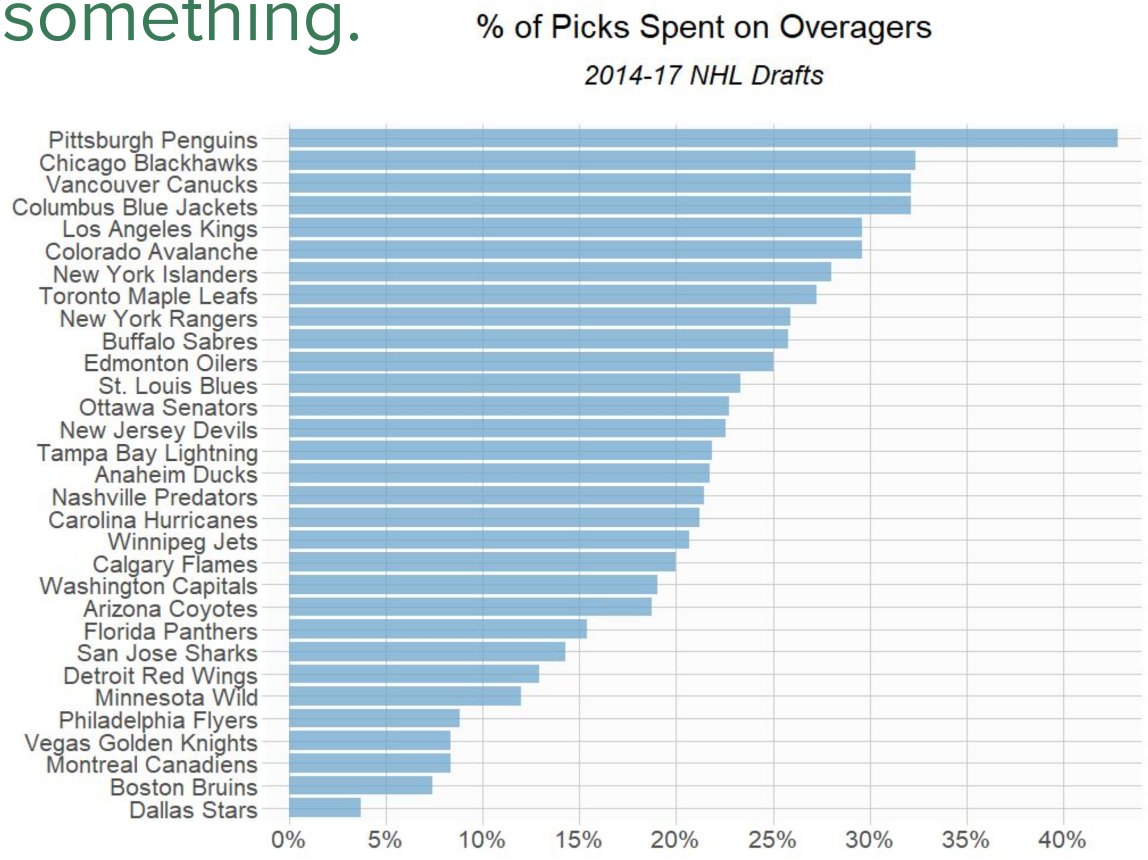
- Every year, there is a 7-round entry draft for NHL teams.
- Each team is initially allocated one pick per round.
- If a prospect turns 18 by September 15th of the year of the draft, he is eligible to be drafted.
- Most drafted players are in their first year of eligibility. Some (~20%) are in their second (or third, or fourth...).
- Understanding opposing team draft tendencies can be useful when trying to make draft decisions.

Overall	Player	Age
62	Joonas Korpisalo	18
63	Jujhar Khaira	18
64	Tim Bozon	18
65	Adam Pelech	18
66	Jimmy Vesey	19
67	Mackenzie MacEachern	18
68	John Draeger	18
69	Daniel Altshuller	18
70	Scott Kosmachuk	18
71	Tanner Richard	19
72	Troy Bourke	18
73	Justin Kea	18
74	Esa Lindell	18
75	Jon Gillies	18
76	Chris Driedger	18
77	Chandler Stephenson	18
78	Shayne Gostisbehere	19

2012 draft snapshot via Hockey Reference

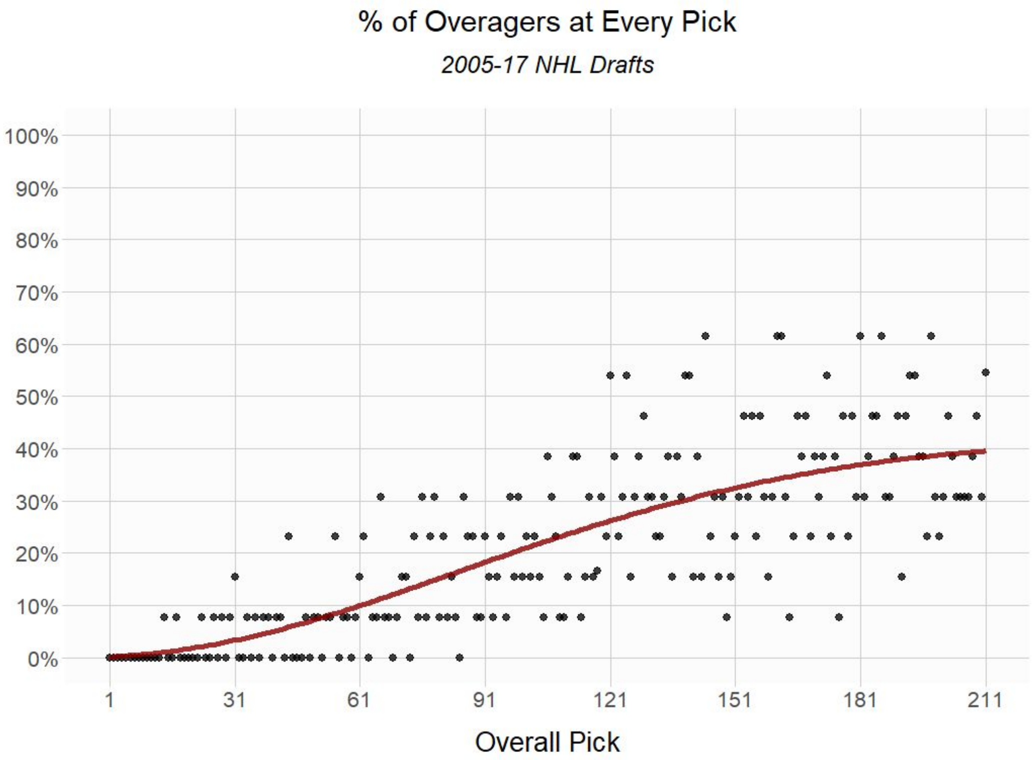
Last year, I noticed something.

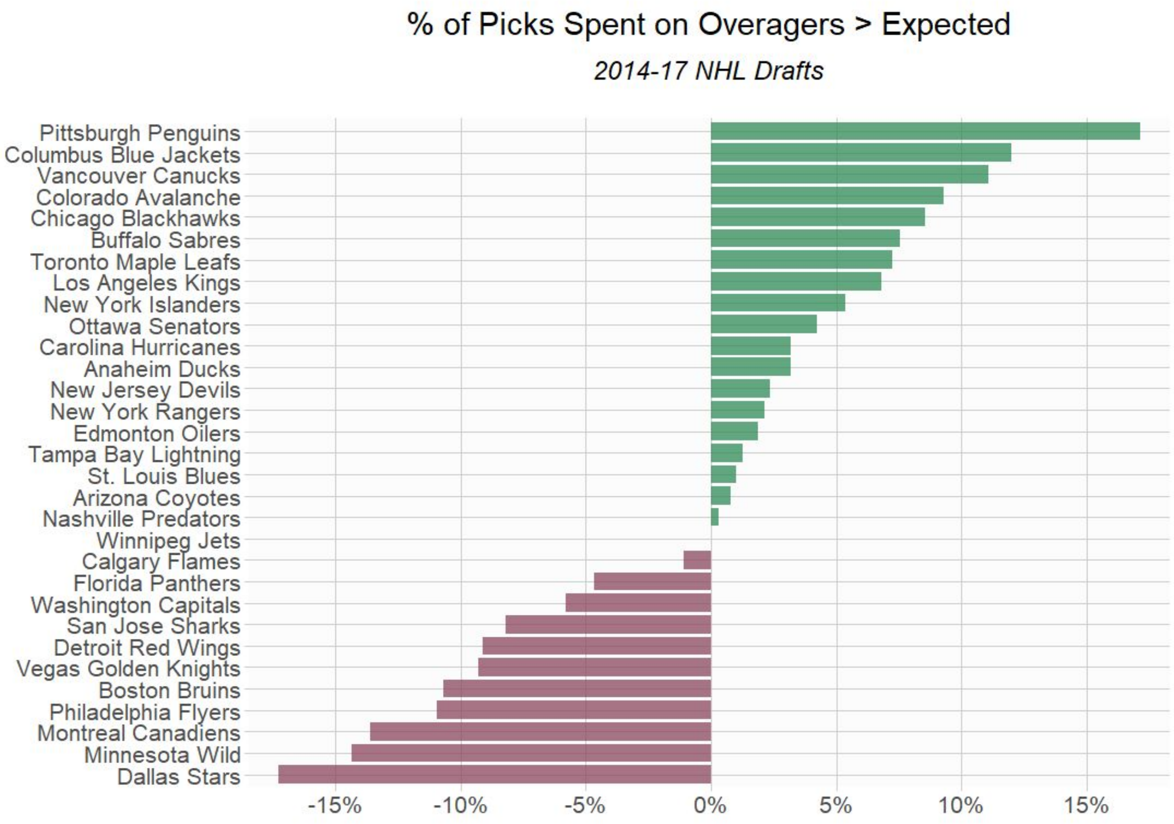
- But first, I used *rvest* to scrape Hockey Reference draft data.
- It appeared that the Penguins had been selecting a lot of overage (19+) prospects recently.



At least partially because they had late picks.

- We can confirm this by estimating the likelihood of selecting an overager based on pick #.
- I used *rstanarm::stan_glm()* to fit a logistic regression with weakly informative normal priors on the coefficients.

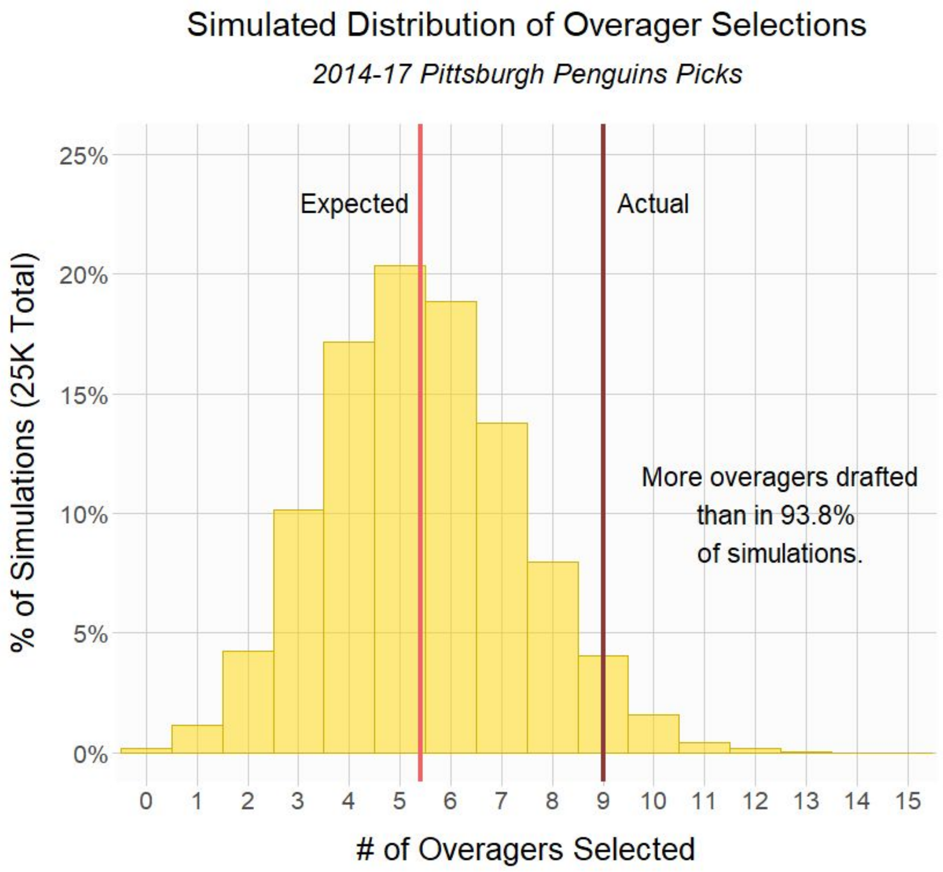




But even after adjusting for picks, the tendency seems pretty real.

And after simulating the Penguins' drafts repeatedly, I wasn't as worried about small samples.

- *rstanarm::posterior_predict()* makes this very easy to do.



My Impact™

Namita
@nnstats

so...on Tuesday I tweeted “take an overager right before a Penguins pick out of spite” and the Leafs like...actually did that

11:50 AM - 23 Jun 2018

Thanks for listening!

- I'll tweet out slides/code/data @nnstats on Twitter.
- Shoutout to MoneyPuck and Hockey Reference for making hockey data readily and painlessly available to the public.
- Some great sources for advanced hockey stats (a lecture for another time) include corsica.hockey, Evolving Hockey, HockeyViz, and Natural Stat Trick.