Models and Markets

Kiernan Nicholls

December 4, 2018

Why predict elections?

- Resource allocation
- Strategy adjustment
- Quantitative journalism

How to Predict Elections

1. Opinion polling
2. Polling aggregation
3. Forecast modeling
4. Prediction markets

Opion Polling

e.g., Washington Post/ABC

In 1824 *The Harrisburg Pennsylvanian* had Jackson over Adams, 335 to 169.

- Sample Size
- Methodology
- Partisanship

Polling Aggrigation

e.g., RealClearPolitics

- 21st century invention
- Average out all polls
- Minimize errors and reduce bias
- Possibly weighted

Forecasting Models

Montel carlo simulations = probability distribution

1. Define a domain of possible inputs
2. Generate inputs randomly from a probability distribution over the domain
3. Perform a deterministic computation on the inputs
4. Aggregate the results

- Draw share of vote, compared
- 20,000 interations
- Law of large numbers

About FiveThirtyEight

- Founded in 2008, sold to NYT then ABC
- Least inaccurate in 2016

  Someone could look like a genius simply by doing some fairly basic research into what really has predictive power in a political campaign

FiveThirtyEight Forecast

> *It takes lots of polls, performs various types of adjustments to them, and then blends them with other kinds of empirically useful indicators... Then it accounts for the uncertainty in the forecast and simulates the election thousands of times.*

1. **Polling**: District-by-district polling, adjusted for house effects and other factors.
2. **CANTOR**: Infers results for districts with little or no polling from comparable districts with polling.
3. **Fundamentals**: District partisanship, past performance, generic ballot, fundraising, experience, scandals

Trained off elections since 1998. Only miscalled 3.3% of past races.

Model Uncertainty

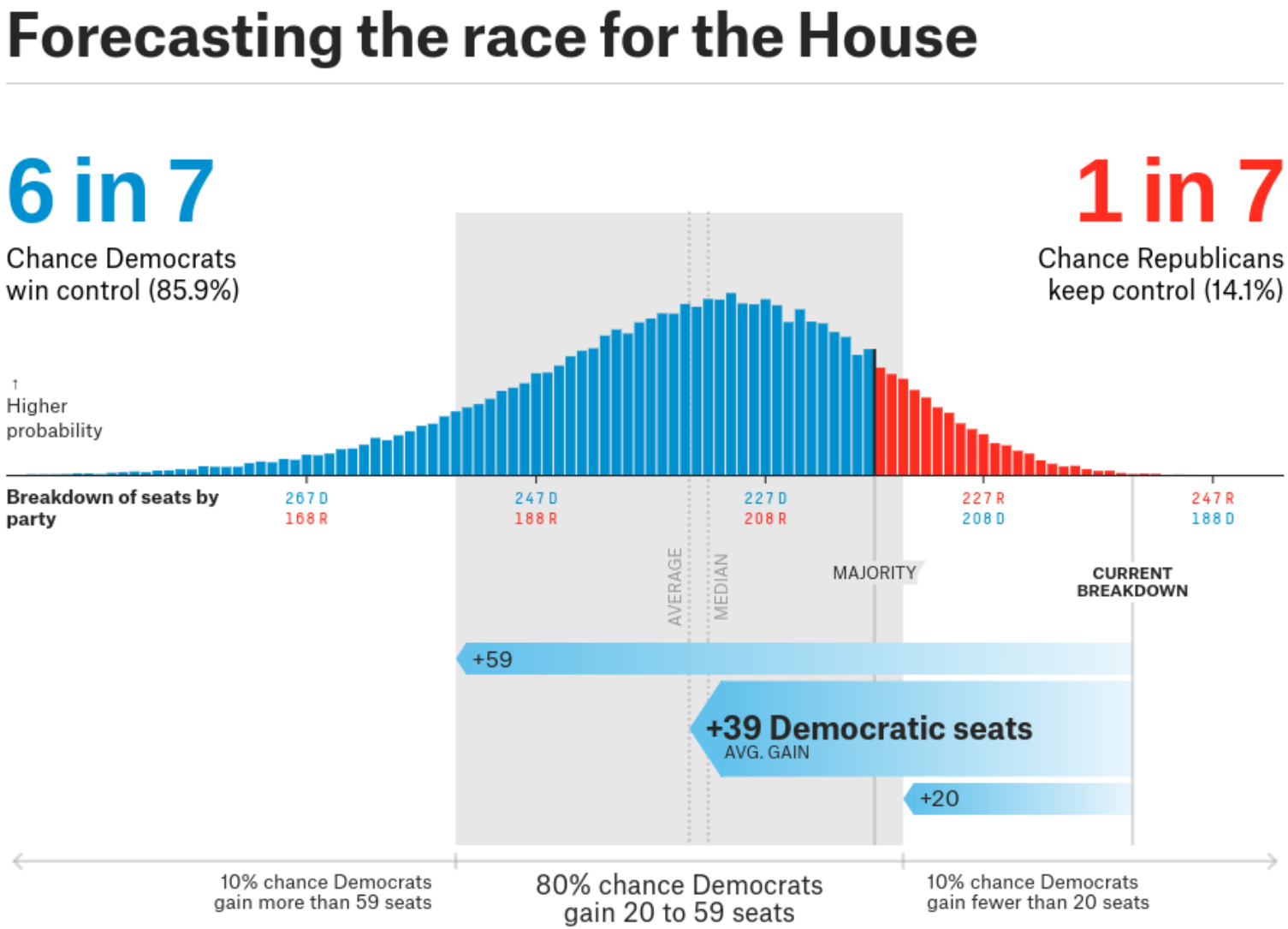


Figure 1: model_histogram

Model Data

```
## # A tibble: 89,918 x 6
##    date       chamber code  party voteshare  prob
##    <date>     <chr>   <chr> <chr>      <dbl> <dbl>
##  1 2018-08-01 senate  AZ-99 D          0.511 0.738
##  2 2018-08-01 senate  AZ-99 R          0.461 0.262
##  3 2018-08-01 senate  CA-99 D          0.636 0.999
##  4 2018-08-01 senate  CA-99 D          0.364 0.001
##  5 2018-08-01 senate  CT-99 D          0.641 0.999
##  6 2018-08-01 senate  CT-99 R          0.324 0.001
##  7 2018-08-01 senate  DE-99 D          0.607 0.989
##  8 2018-08-01 senate  DE-99 R          0.367 0.011
##  9 2018-08-01 senate  FL-99 D          0.511 0.616
## 10 2018-08-01 senate  FL-99 R          0.489 0.384
## # ... with 89,908 more rows
```

Prediction Markets

In 1503 traders bet on Papal successor. Iowa Election Market founded in 1988.

- Exchange-traded markets
- Binary options
- Contract price $=$ probability
- Crowd-sourcing
- Efficient market hypothesis
- Price equilibrium
- Risk aversion

PredictIt

PredictIt is a unique and exciting real money site that tests your knowledge of political events by letting you trade shares on everything from the outcome of an election to a Supreme Court decision to major world events... PredictIt is run by Victoria University of Wellington, New Zealand, a not-for-profit university, for educational purposes

PredictIt Contracts

- Real money
- Elections, Justice, Administration, World
- Futures contracts
- Two buyers
- Executes at time or condition
- Either $1 or $0
- Sell at any time

PredictIt Markets

- Will Donald Trump be president at year-end 2018?
- Will the federal government be shut down on February 9?
- Will Ted Cruz be re-elected to the U.S. Senate in Texas in 2018?
- Will Facebook's Mark Zuckerberg run for president in 2020?
- How many tweets will @realDonaldTrump post from noon Oct. 10 to noon Oct. 17?

PredicIt Data

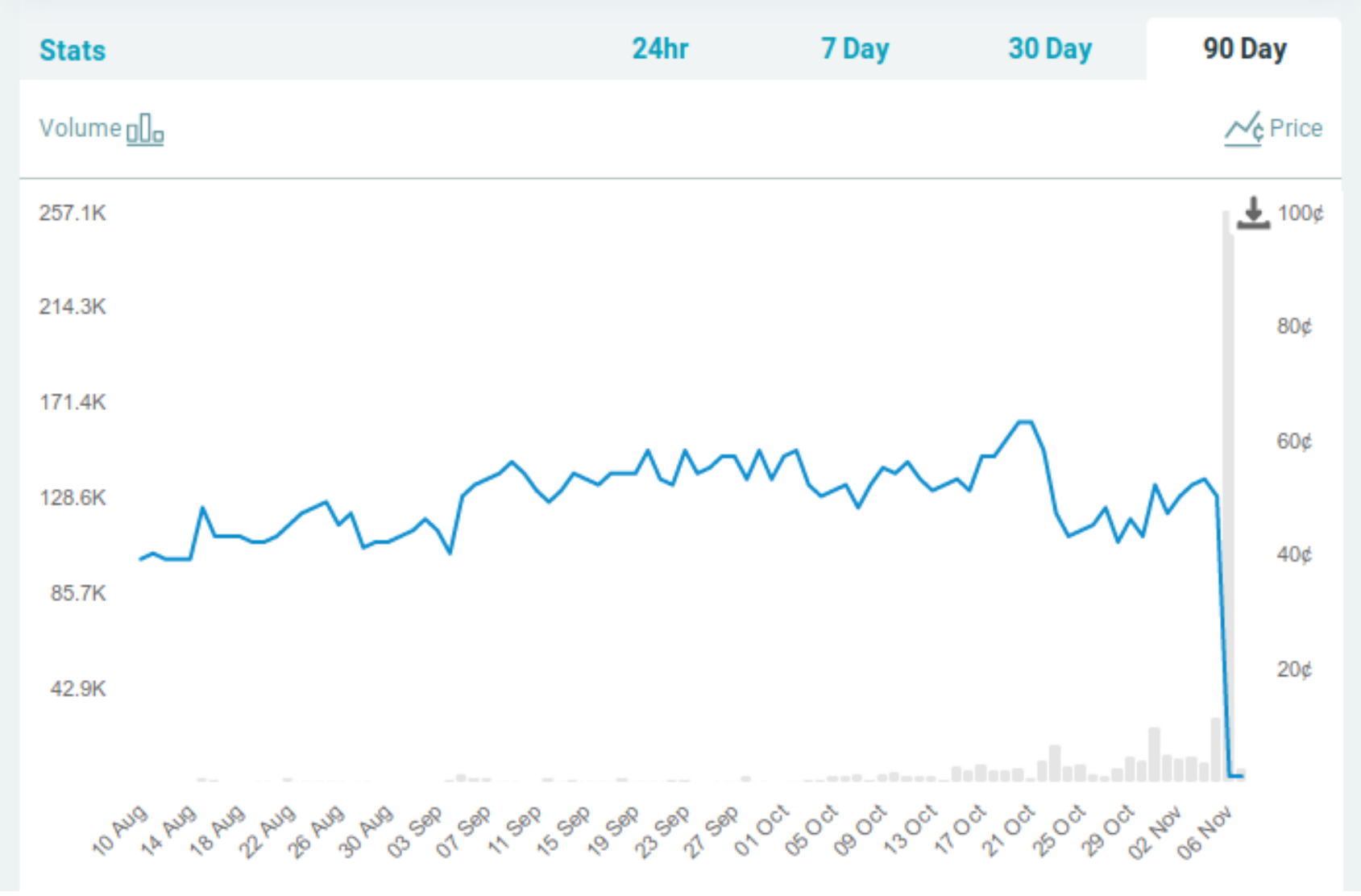


Figure 2: Donnelly Chart

PredictIt Data Collection

1. Get all active relevant market names from API
2. Grab chart data from all above markets
3. Merge chart data with API names
4. Turn market names into district codes and party affiliation

Scraped Market Data

```
## # A tibble: 24,556 x 5
##    date       mid   cid   price volume
##    <date>     <chr> <chr> <dbl>  <dbl>
##  1 2018-08-10 2918  5264   0.95     56
##  2 2018-08-11 2918  5264   0.95     50
##  3 2018-08-12 2918  5264   0.89    100
##  4 2018-08-13 2918  5264   0.9      40
##  5 2018-08-14 2918  5264   0.91     61
##  6 2018-08-15 2918  5264   0.91     85
##  7 2018-08-16 2918  5264   0.91     59
##  8 2018-08-17 2918  5264   0.91      0
##  9 2018-08-18 2918  5264   0.91      0
## 10 2018-08-19 2918  5264   0.95     50
## # ... with 24,546 more rows
```

Market API Names

- Which party will win GA-07?
- Which party will win AK at-large?
- Will Brian Fitzpatrick be re-elected?
- Which party will win MS Senate special?
- Will Pelosi be re-elected?
- Will a Dem candidate win the 2018 House of Reps race in WA's 3rd district?

Formatting Names

```
if_else(str_detect(market_history$code, "re-elected"),
        word(market_history$code, 3),
if_else(str_detect(market_history$code, "at-large"),
        paste(word(market_history$code, 5), "01", sep = "-"
if_else(str_detect(market_history$code, "special"),
        paste(word(market_history$code, 5), "98", sep = "-"
if_else(str_detect(market_history$code, "Senate"),
        paste(word(market_history$code, 5), "99", sep = "-"
if_else(str_detect(market_history$code, "re-elected"),
        word(market_history$code, 3),
if_else(str_detect(market_history$code, "Which party"),
        word(market_history$code, 5), "ERROR"))))))
```

Market Data Combination

```
## # A tibble: 24,466 x 7
##    date         mid   cid price volume code  party
##    <date>     <dbl> <dbl> <dbl>  <dbl> <chr> <chr>
##  1 2018-08-10  2918  5264  0.95     56 MA-99 D
##  2 2018-08-11  2918  5264  0.95     50 MA-99 D
##  3 2018-08-12  2918  5264  0.89    100 MA-99 D
##  4 2018-08-13  2918  5264  0.9      40 MA-99 D
##  5 2018-08-14  2918  5264  0.91     61 MA-99 D
##  6 2018-08-15  2918  5264  0.91     85 MA-99 D
##  7 2018-08-16  2918  5264  0.91     59 MA-99 D
##  8 2018-08-17  2918  5264  0.91      0 MA-99 D
##  9 2018-08-18  2918  5264  0.91      0 MA-99 D
## 10 2018-08-19  2918  5264  0.95     50 MA-99 D
## # ... with 24,456 more rows
```

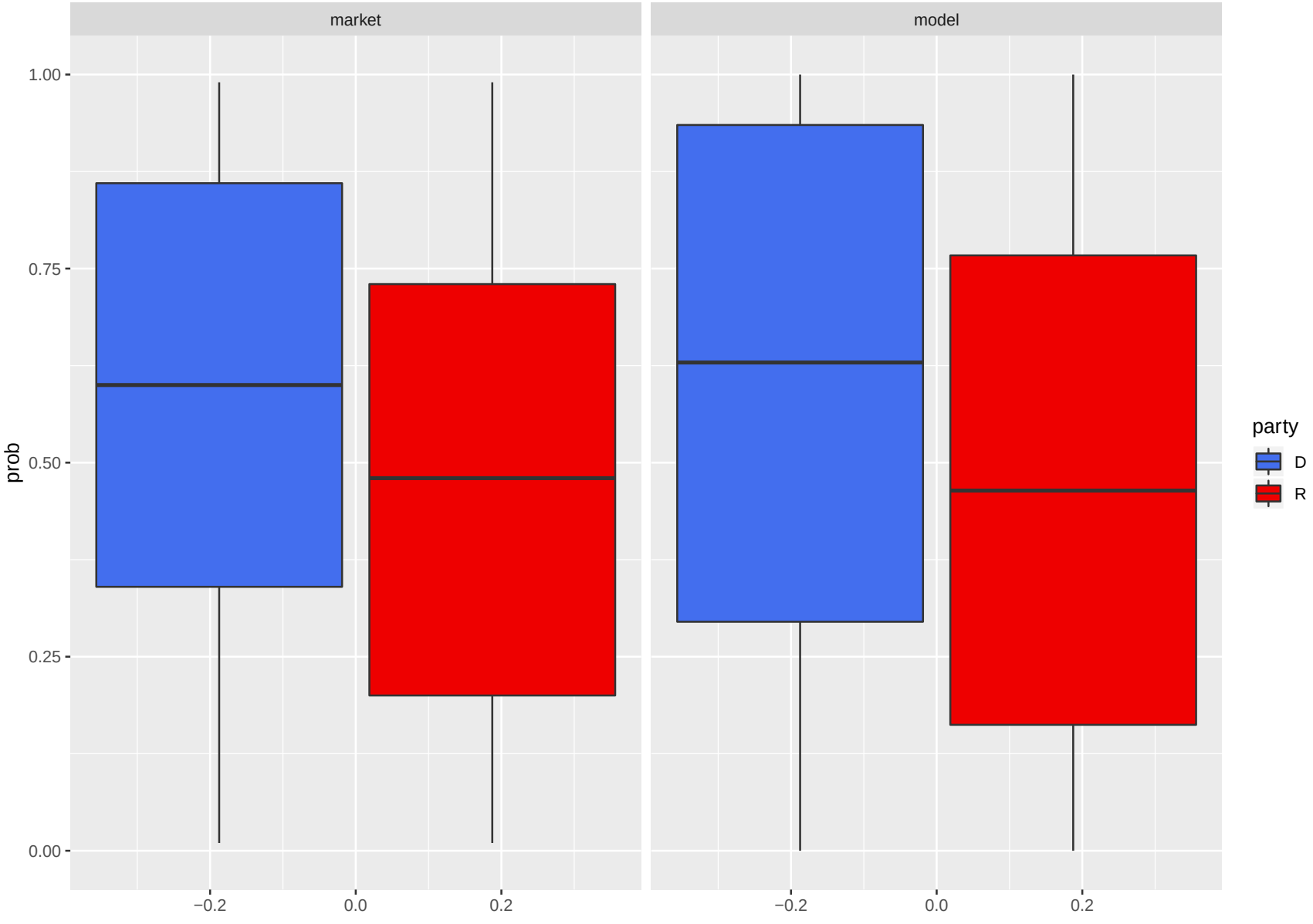
Joining Markets and Models

```
## # A tibble: 24,555 x 6
##    date       code  party  prob price volume
##    <date>     <chr> <chr> <dbl> <dbl>  <dbl>
##  1 2018-08-10 MA-99 D     0.999  0.95     56
##  2 2018-08-10 TX-99 R     0.742  0.7    1303
##  3 2018-08-10 VT-99 D     1      0.95    542
##  4 2018-08-10 WV-99 D     0.859  0.75    533
##  5 2018-08-10 IN-99 D     0.864  0.39     12
##  6 2018-08-10 CA-12 D     1      0.9      51
##  7 2018-08-10 ND-99 D     0.594  0.42     81
##  8 2018-08-10 MO-99 D     0.733  0.47    333
##  9 2018-08-10 WI-99 D     0.977  0.83      0
## 10 2018-08-10 MI-99 D     0.985  0.79    390
## # ... with 24,545 more rows
```

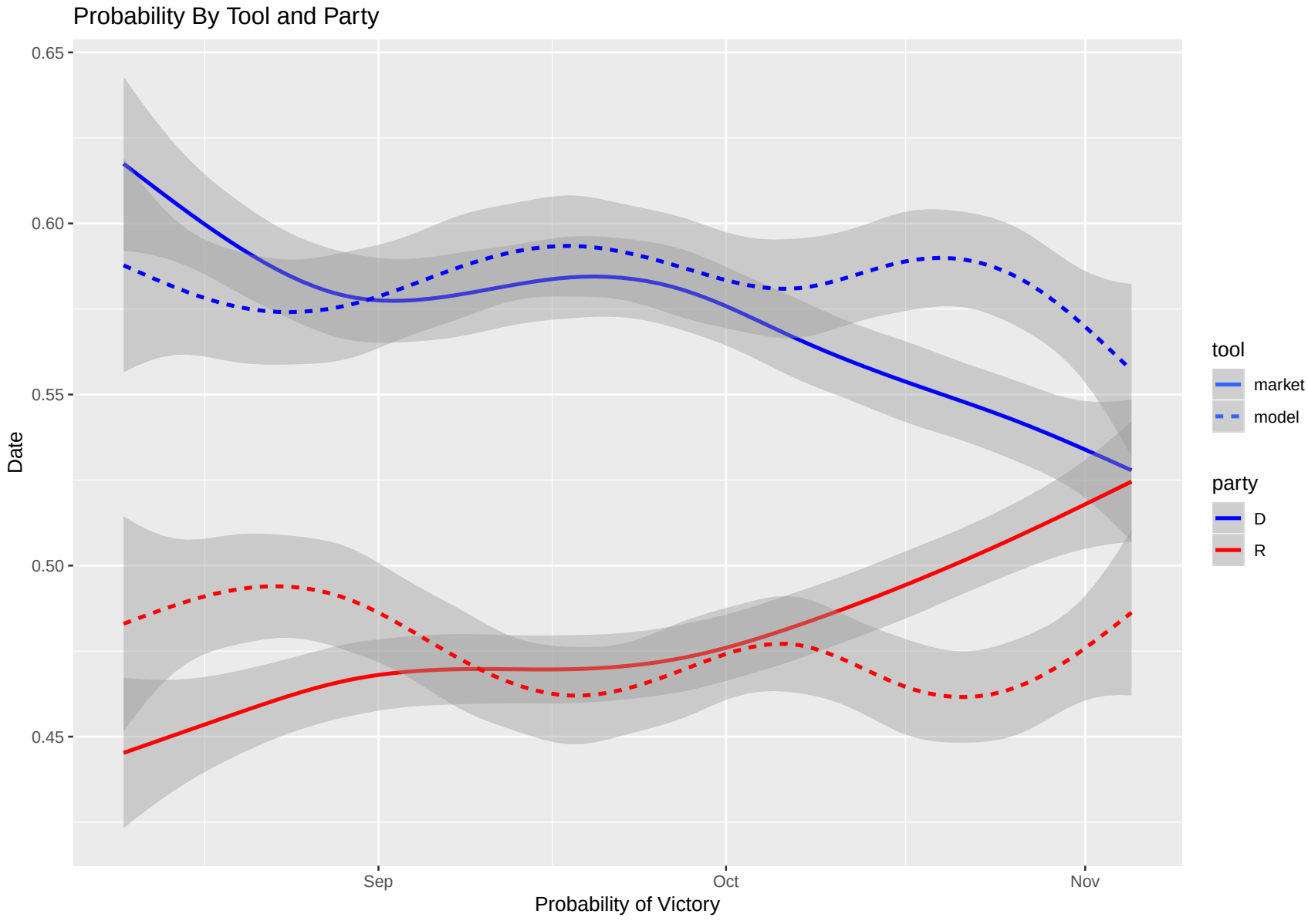
Tidy Data

```
## # A tibble: 46,138 x 5
##    date       code  party tool    prob
##    <date>     <chr> <chr> <chr>  <dbl>
##  1 2018-08-10 AZ-99 R     model  0.272
##  2 2018-08-10 AZ-99 R     market 0.02
##  3 2018-08-10 CA-12 D     model  1
##  4 2018-08-10 CA-12 D     market 0.9
##  5 2018-08-10 CA-22 R     model  0.96
##  6 2018-08-10 CA-22 R     market 0.65
##  7 2018-08-10 CA-49 R     model  0.197
##  8 2018-08-10 CA-49 R     market 0.03
##  9 2018-08-10 CA-99 D     model  0.999
## 10 2018-08-10 CA-99 D     model  0.001
## # ... with 46,128 more rows
```

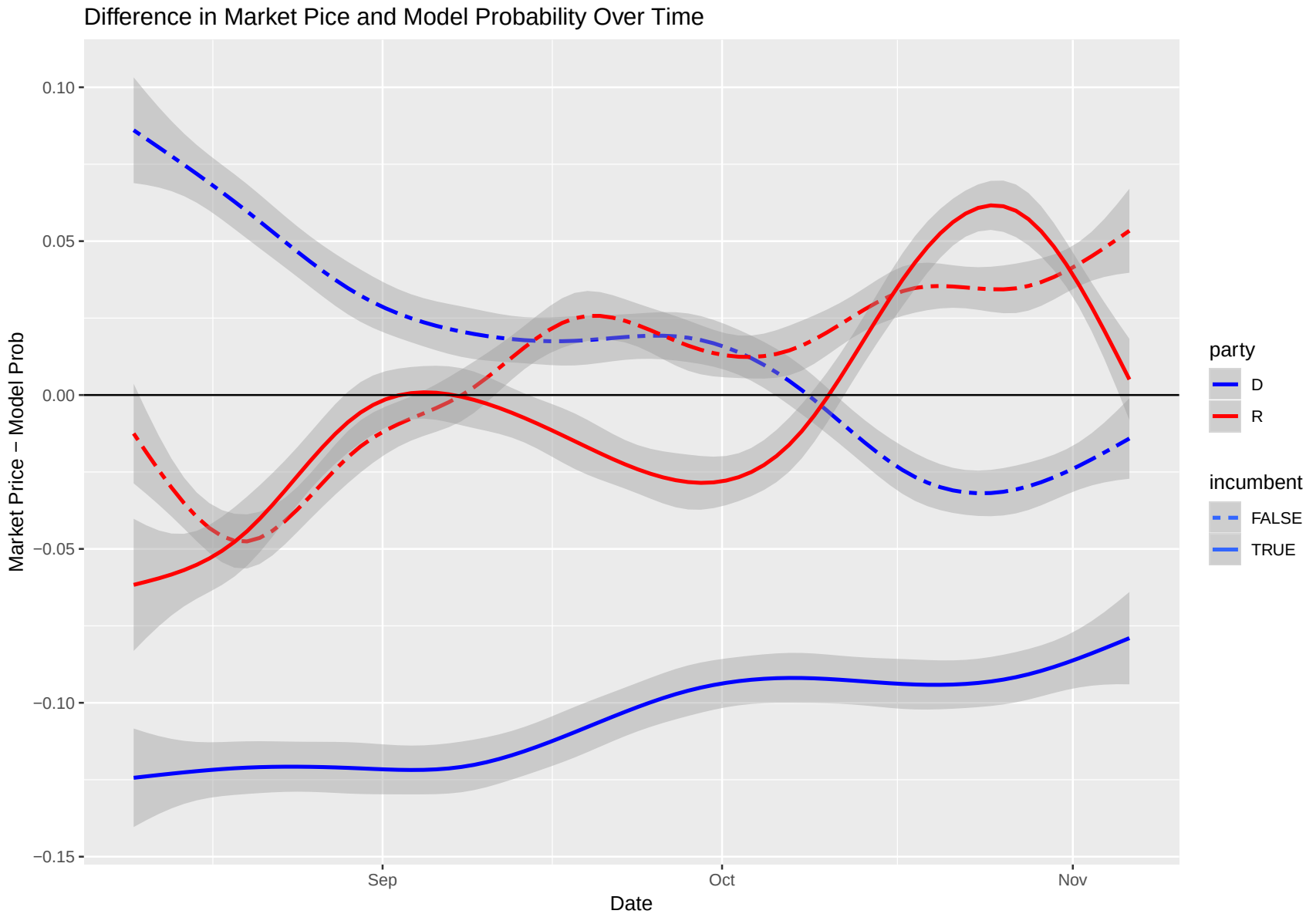
Probability Boxplots



Probability by Tool



Difference in Tools



AP Election Results

Safe Dem

	D	R	% report.
AL-7	Unc.	*	
AZ-3	63.4%	36.6%	99.4%
AZ-7	85.7%	*	100%
AZ-9	60.9%	39.1%	100%
CA-2	76.8%	23.2%	100%
CA-3	56.9%	43.1%	100%
CA-5	78.4%	*	100%
CA-6	80.9%	*	100%
CA-9	55.9%	44.1%	100%
CA-11	74.0%	26.0%	100%
CA-12	86.8%	13.2%	100%
CA-13	88.4%	*	100%
CA-14	79.2%	20.8%	100%
CA-15	73.0%	27.0%	100%
CA-17	75.3%	24.7%	100%
CA-18	74.5%	25.5%	100%
CA-19	73.7%	26.3%	100%
CA-20	81.2%	*	100%
CA-24	58.0%	42.0%	100%
CA-26	61.8%	38.2%	100%
CA-27	79.3%	*	100%
CA-28	78.3%	21.7%	100%
CA-29	80.6%	19.4%	100%
CA-30	73.4%	26.6%	100%
CA-31	58.2%	41.8%	100%
CA-32	68.7%	31.3%	100%
CA-33	70.0%	30.0%	100%
CA-34	72.6%	*	100%
CA-35	69.1%	30.9%	100%
CA-36	58.0%	42.0%	100%

Likely Dem

	D	R	% report.
CA-7	54.8%	45.2%	100%
CA-16	56.8%	43.2%	100%
FL-7	57.7%	42.3%	100%
MN-7	52.1%	47.9%	100%
NH-1	53.5%	45.0%	100%
NJ-5	55.2%	43.5%	100%
PA-8	54.6%	45.4%	99.2%

Vulnerable Dem

	D	R	% report.
AZ-1	53.8%	46.2%	100%
MN-1	49.8%	50.2%	100%
MN-8	45.2%	50.7%	100%
NV-3	51.9%	42.8%	100%
NV-4	51.9%	43.7%	100%
PA-14	42.0%	58.0%	100%

Vulnerable GOP

	D	R	% report.
AK	46.7%	53.3%	100%
AZ-2	54.4%	45.6%	99.0%
CA-10	51.7%	48.3%	100%
CA-25	54.2%	45.8%	100%
CA-39	51.5%	48.5%	100%
CA-45	51.9%	48.1%	100%
CA-48	53.5%	46.5%	100%
CA-49	56.1%	43.9%	100%
CA-50	48.2%	51.8%	100%
CO-6	54.1%	42.9%	99.2%
FL-6	43.7%	56.3%	100%
FL-15	47.0%	53.0%	100%
FL-16	45.4%	54.6%	100%
FL-18	45.7%	54.3%	100%
FL-25	39.5%	60.5%	100%
FL-26	50.9%	49.1%	100%
FL-27	51.8%	45.8%	100%
GA-6	50.5%	49.5%	100%
GA-7	49.9%	50.1%	100%
IA-1	50.9%	46.0%	100%
IA-3	49.0%	47.5%	100%
IA-4	47.0%	50.4%	100%
IL-6	52.8%	47.2%	99.5%
IL-12	45.2%	51.8%	100%
IL-13	49.5%	50.5%	100%
IL-14	51.9%	48.1%	100%
KS-2	46.4%	48.1%	100%
KS-3	53.3%	44.2%	100%
KY-6	47.8%	51.0%	100%
ME-2	45.5%	46.2%	95.5%

Likely GOP

	D	R	% report.
AR-2	45.8%	52.1%	100%
AZ-6	44.7%	55.3%	100%
AZ-8	44.4%	55.6%	100%
CA-1	45.0%	55.0%	100%
CA-4	45.7%	54.3%	100%
CA-21	50.2%	49.8%	100%
CA-22	46.6%	53.4%	100%
CO-3	43.4%	51.7%	98.6%
IN-2	45.2%	54.8%	95.1%
MI-1	43.7%	56.3%	100%
MI-3	43.2%	54.4%	100%
MI-7	46.2%	53.8%	100%
NC-8	44.6%	55.4%	100%
NY-1	46.4%	52.5%	100%
NY-2	46.7%	53.3%	100%
NY-21	41.8%	56.7%	100%
NY-23	45.0%	55.0%	100%
OH-10	41.9%	56.2%	100%
OH-14	44.6%	55.4%	100%
OK-5	50.7%	49.3%	100%
TX-2	45.5%	52.9%	100%
TX-6	45.4%	53.1%	100%
TX-10	46.9%	50.9%	100%
TX-21	47.5%	50.3%	100%
TX-24	47.5%	50.7%	100%
TX-31	47.6%	50.6%	100%
WI-6	44.5%	55.5%	100%
WV-2	42.9%	54.0%	100%

Safe GOP

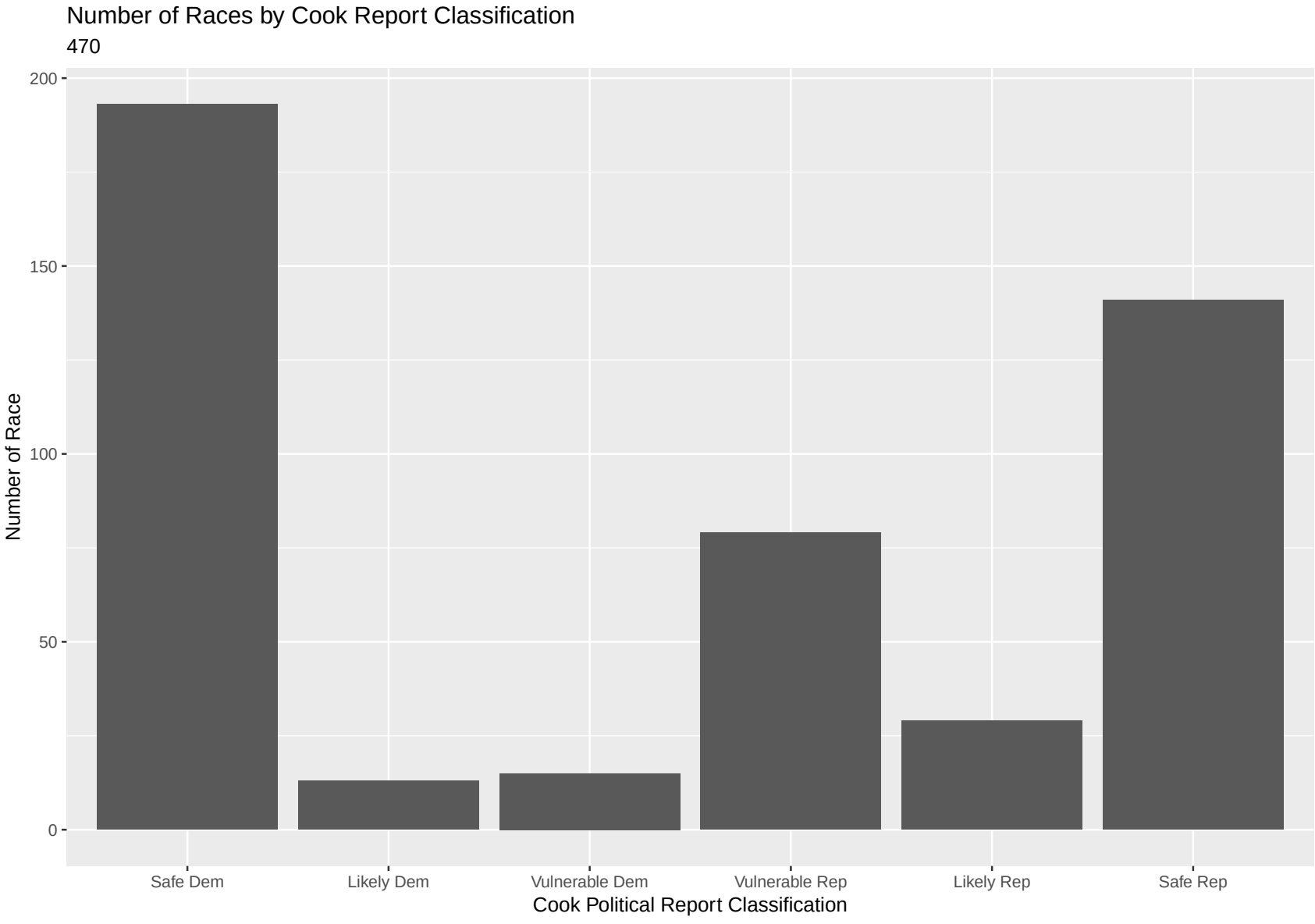
	D	R	% report.
AL-1	36.8%	63.2%	100%
AL-2	38.5%	61.5%	100%
AL-3	36.2%	63.8%	100%
AL-4	20.1%	79.9%	100%
AL-5	38.9%	61.1%	100%
AL-6	30.8%	69.2%	100%
AR-1	28.7%	69.0%	100%
AR-3	32.5%	64.9%	97.0%
AR-4	31.3%	66.8%	98.0%
AZ-4	30.5%	68.2%	100%
AZ-5	40.6%	59.4%	100%
CA-8	*	60.1%	100%
CA-23	36.1%	63.9%	100%
CA-42	42.7%	57.3%	100%
CO-4	39.1%	60.9%	97.8%
CO-5	38.3%	58.3%	87.2%
FL-1	32.9%	67.1%	100%
FL-2	32.6%	67.4%	100%
FL-3	42.4%	57.6%	100%
FL-4	32.4%	65.2%	100%
FL-8	39.5%	60.5%	100%
FL-11	34.8%	65.2%	100%
FL-12	39.7%	58.1%	100%
FL-17	37.7%	62.3%	100%
FL-19	37.7%	62.3%	100%
GA-1	42.3%	57.7%	100%
GA-3	34.5%	65.5%	100%
GA-8	*	Unc.	
GA-9	20.5%	79.5%	100%
GA-10	37.1%	62.9%	100%

Figure 3: Election Tables

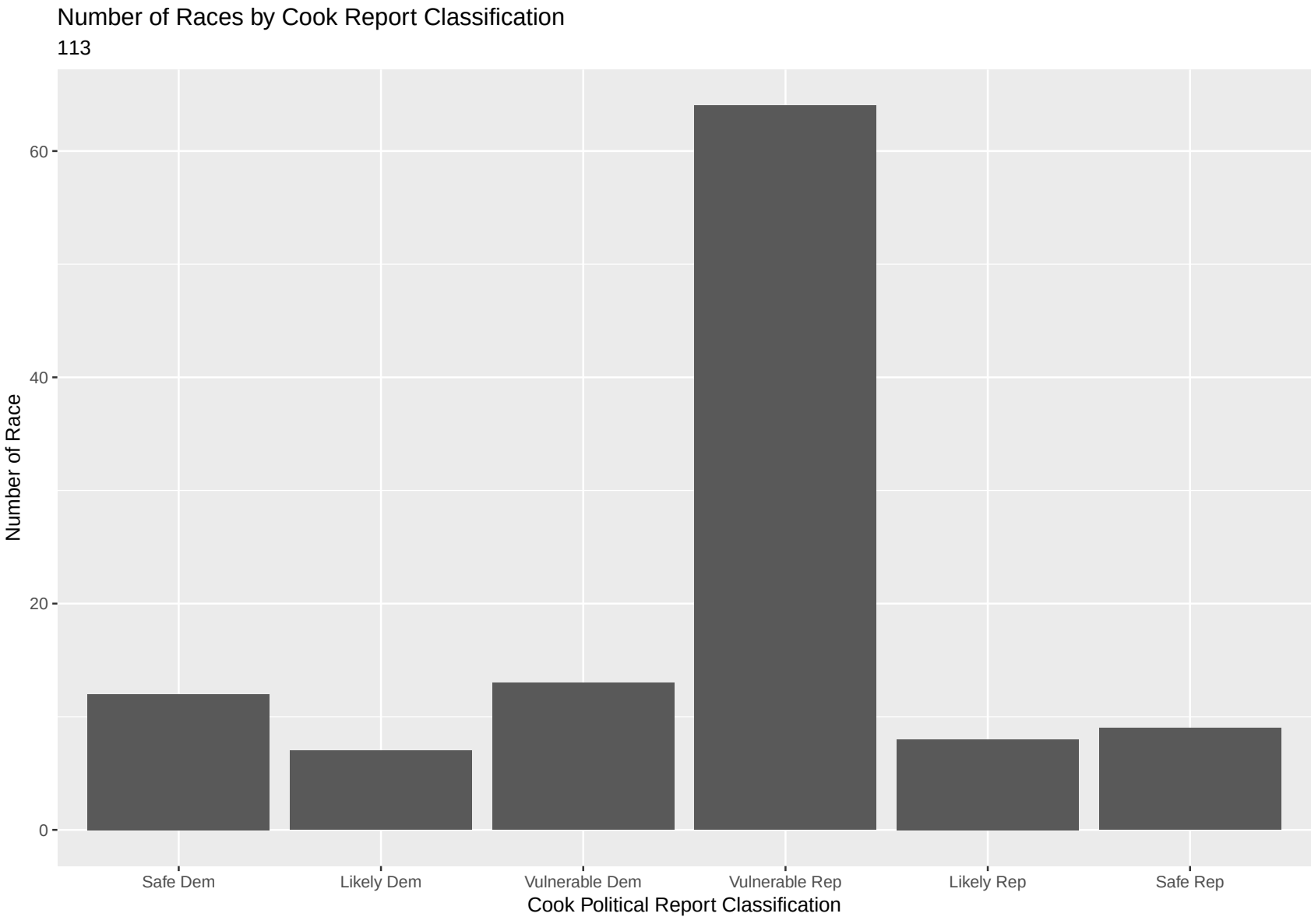
Scraping Results

```
## # A tibble: 470 x 5
##    code    dem   rep class  winner
##    <chr> <dbl> <dbl> <fct>  <chr>
##  1 AK-01 0.46  0.54  vul R  R
##  2 AL-01 0.367 0.633 safe R R
##  3 AL-02 0.385 0.615 safe R R
##  4 AL-03 0.362 0.638 safe R R
##  5 AL-04 0.201 0.799 safe R R
##  6 AL-05 0.389 0.611 safe R R
##  7 AL-06 0.307 0.693 safe R R
##  8 AL-07 1     0     safe D D
##  9 AR-01 0.287 0.69  safe R R
## 10 AR-02 0.458 0.521 lkly R R
## # ... with 460 more rows
```

Cook Race Classifications



Cook Race Classifications



Post-Election Results

1. Any given time, $>50\%$ is a predicted winner
2. For each day, ask if guess matches winner
3. Average across all races
4. Plot over time

Accuracy Over Time