Concerns about Dr. Frank's Election Analysis

efg

2022-08-24

Concerns about Dr. Frank's Election Analysis

Outline

- "Data leakage" in predictive analytics results in overly good fits.
- Normalized turnout curves changed little from 2006-2020.
- Normalized turnout for a specific age varies slightly by state, county, precinct, but is highly correlated.
- A 6^{th} degree polynomial may not be the "best" fit. Isn't needed.
- Correlation is a weak way to compare predictive models.
- Turnout varies by gender and political party.
- Inflated voter rolls are not new.

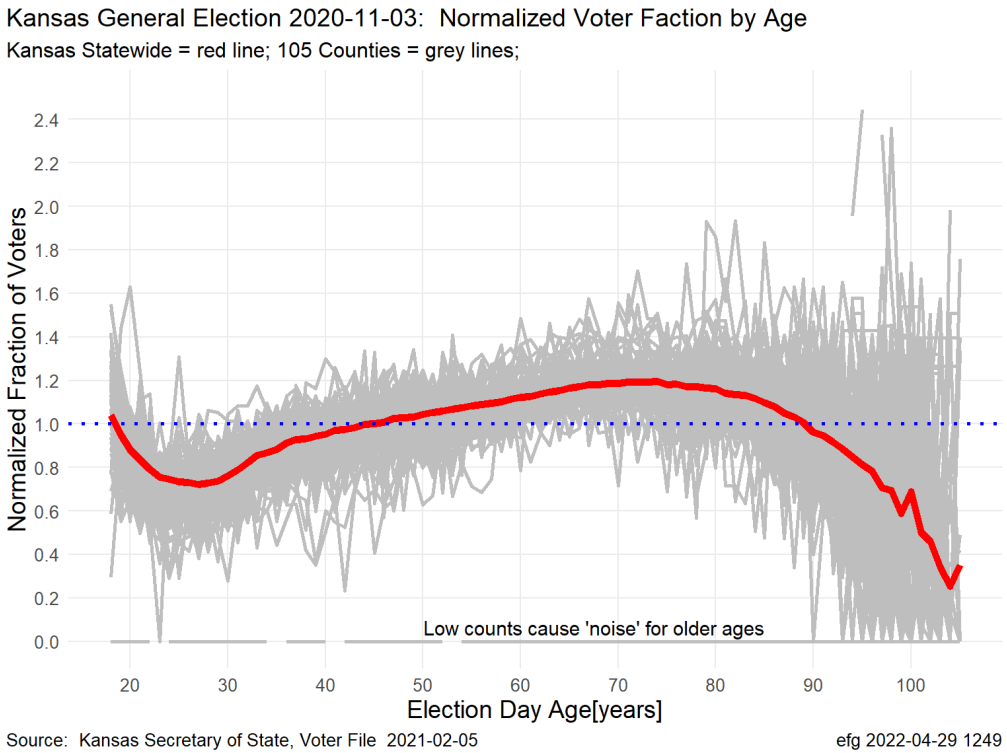
“Data leakage” in predictive analytics results in overly good fits

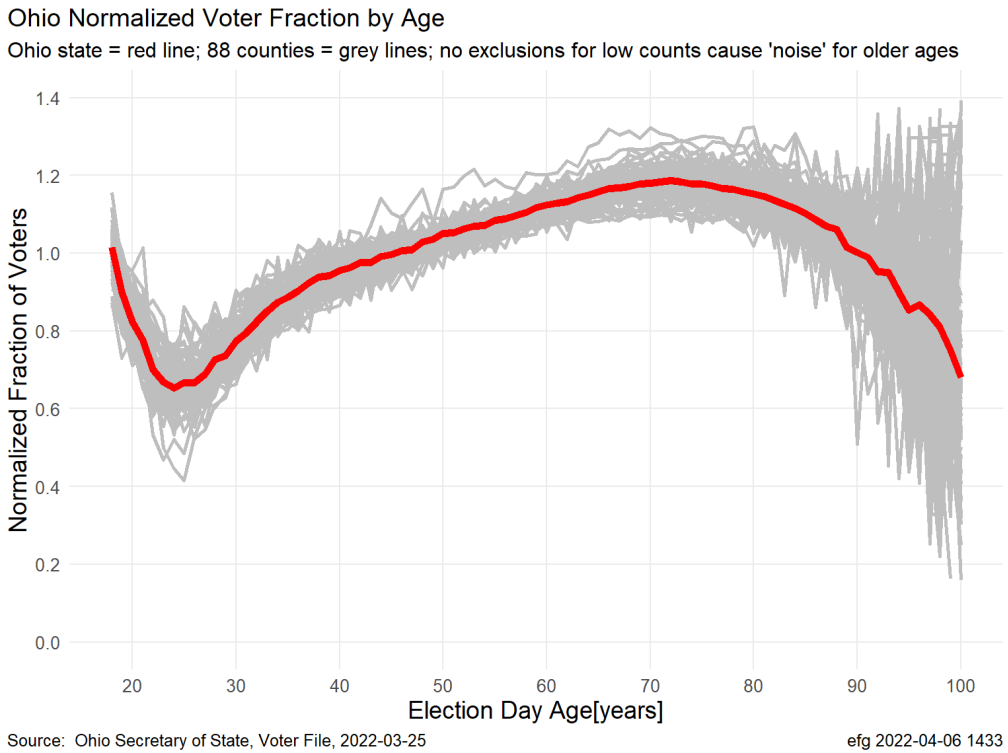
- One way “data leakage” occurs is when the development of a predictive model uses data being predicted to create the model.
- I may be mistaken, but Dr. Frank’s voter predictions start with a state “key” or a county “key,” which is a normalized voter turnout curve. All these keys are roughly interchangeable.

- *How are “predictions” for the 2022 elections possible without having the voter registration numbers and actual numbers of voted by age that are not known until after the election?*

“Data leakage” in predictive analytics results in overly good fits

- A statewide “key” or a county “key” is a normalized turnout curve.
- The state key and counties keys are so correlated that predictions can be made with any of them, i.e., the state key or any county key.





“Data leakage” in predictive analytics results in overly good fits

But how is this “key” (normalized turnout curve) computed?

For each age interval, 18 to 105 years:

$$\textbf{Normalized Fraction Voted[age]} = \frac{\frac{Nov\ 2020\ voters\ [age]}{Registered\ voters[age]}}{Overall\ Voted\ Fraction} = \text{Turnout Key[age] fitted with polynomial}$$

where

$$\textbf{Overall Voted Fraction = Overall Turnout} = \frac{\sum_{age} Nov\ 2020\ voters}{\sum_{age} Registered\ voters}$$

Predicted Ballots[age] = Overall Voted Fraction * Turnout Key[age] * Registered[age]

Using *Nov 2020 voters* in the key computation is DATA LEAKAGE when used for predictions.

“Data leakage” in predictive analytics results in overly good fits

After computing the “key” (normalized turnout curve) using the actual counts of *Nov 2020 voters*, the prediction of ballots cast is based on the original numbers of registered voters.

This is predictive analytics “data leakage” and explains the overly good prediction fits that have been observed everywhere.

The “key” contains information about the relative turnout by age for the specific election being predicted, which is why the predictions are so close.

Using the 6^{th} degree polynomial adds some “fuzz” to the computations.

Normalized turnout curves changed little from 2006-2020

- Raw Turnout
- Percentage Turnout
- Normalized Turnout for State
- Normalized Turnout for Each County

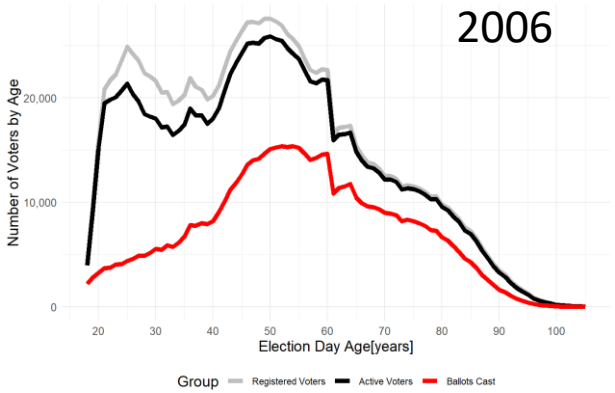
Raw Turnout

Kansas November General Elections: Voter Counts by Age

Grey Line = Registered Voters; Black Line = "Active" Voters; Red Line = Ballots Cast

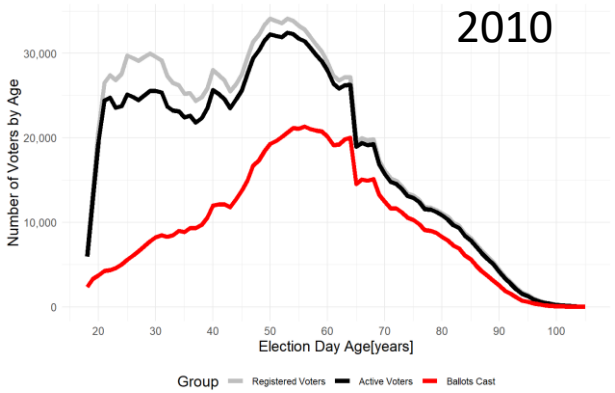
Gubernatorial Election Years

Kansas General Election 2006-11-07: Voters by Age - Kansas Statewide
1,306,339 Registered Voters; 1,202,426 Active Voters; 625,300 Ballots Cast



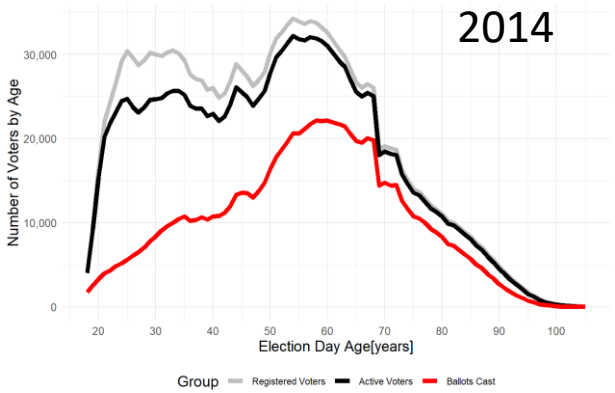
Source: Kansas Secretary of State, Voter File 2006-12-27 efg 2022-04-29 1236

Kansas General Election 2010-11-08: Voters by Age - Kansas Statewide
1,648,089 Registered Voters; 1,530,590 Active Voters; 841,250 Ballots Cast



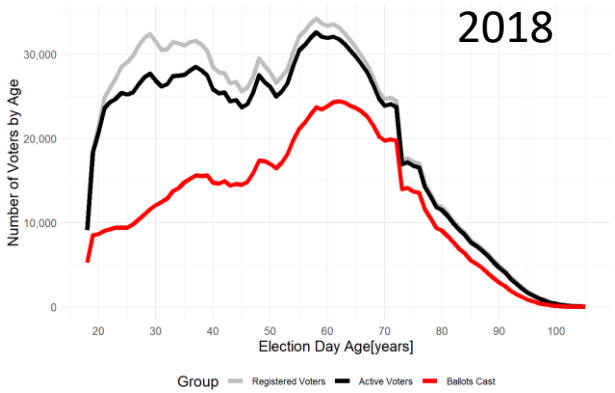
Source: Kansas Secretary of State, Voter File 2011-05-12 efg 2022-04-29 1240

Kansas General Election 2014-11-04: Voters by Age - Kansas Statewide
1,703,501 Registered Voters; 1,551,142 Active Voters; 874,627 Ballots Cast



Source: Kansas Secretary of State, Voter File 2015-01-13 efg 2022-04-29 1243

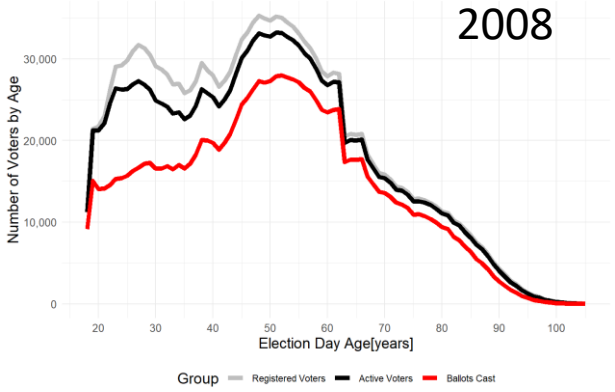
Kansas General Election 2018-11-06: Voters by Age - Kansas Statewide
1,802,129 Registered Voters; 1,673,990 Active Voters; 1,059,977 Ballots Cast



Source: Kansas Secretary of State, Voter File 2019-01-15 efg 2022-04-29 1247

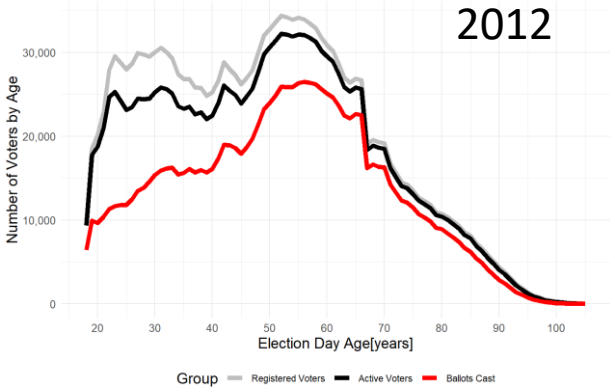
Presidential Election Years

Kansas General Election 2008-11-04: Voters by Age - Kansas Statewide
1,693,207 Registered Voters; 1,574,690 Active Voters; 1,233,513 Ballots Cast



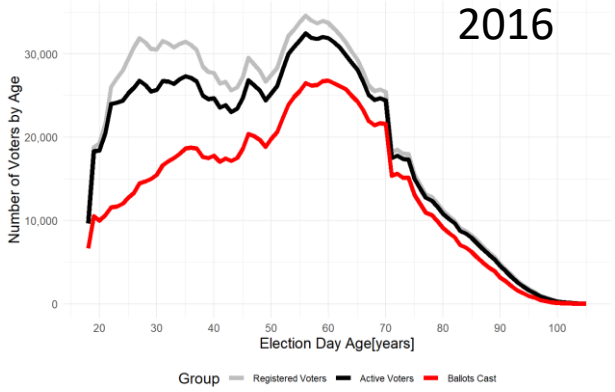
Source: Kansas Secretary of State, Voter File 2009-01-22 efg 2022-04-29 1238

Kansas General Election 2012-11-06: Voters by Age - Kansas Statewide
1,693,016 Registered Voters; 1,549,448 Active Voters; 1,151,198 Ballots Cast



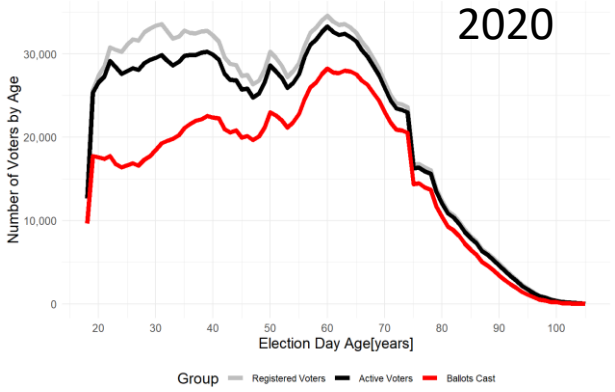
Source: Kansas Secretary of State, Voter File 2013-07-15 efg 2022-04-29 1241

Kansas General Election 2016-11-08: Voters by Age - Kansas Statewide
1,772,685 Registered Voters; 1,623,709 Active Voters; 1,207,269 Ballots Cast



Source: Kansas Secretary of State, Voter File 2017-02-13 efg 2022-04-29 1245

Kansas General Election 2020-11-03: Voters by Age - Kansas Statewide
1,897,481 Registered Voters; 1,787,645 Active Voters; 1,381,516 Ballots Cast



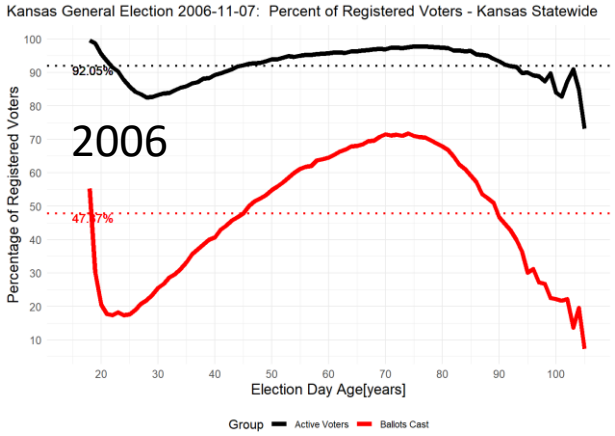
Source: Kansas Secretary of State, Voter File 2021-02-05 efg 2022-04-29 1249

Percentage Turnout

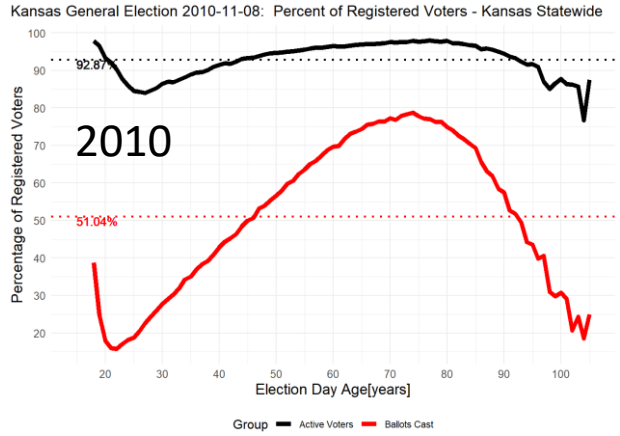
Kansas November General Elections: Percentages by Age

Black Line = “Active” Voters; Red Line = Ballots Cast

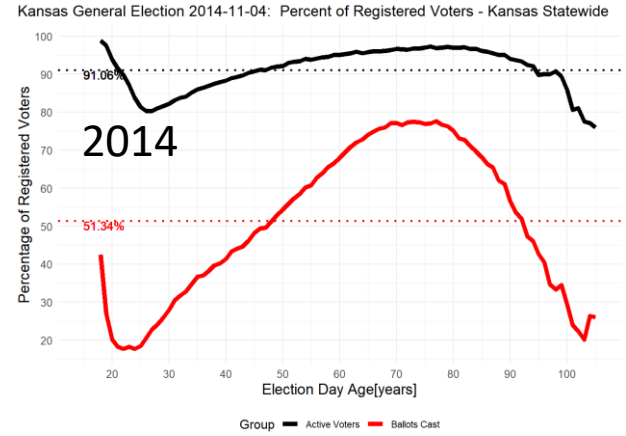
Gubernatorial Election Years



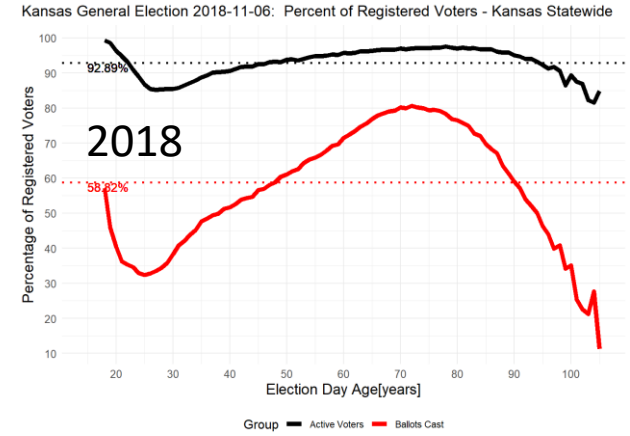
Source: Kansas Secretary of State, Voter File 2006-12-27 efg 2022-04-29 1236



Source: Kansas Secretary of State, Voter File 2011-05-12 efg 2022-04-29 1240

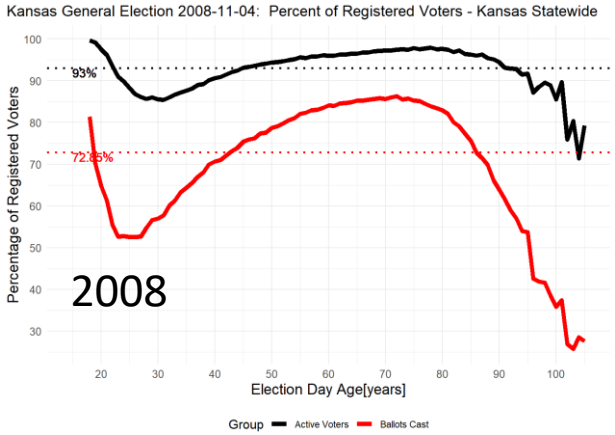


Source: Kansas Secretary of State, Voter File 2015-01-13 efg 2022-04-29 1243

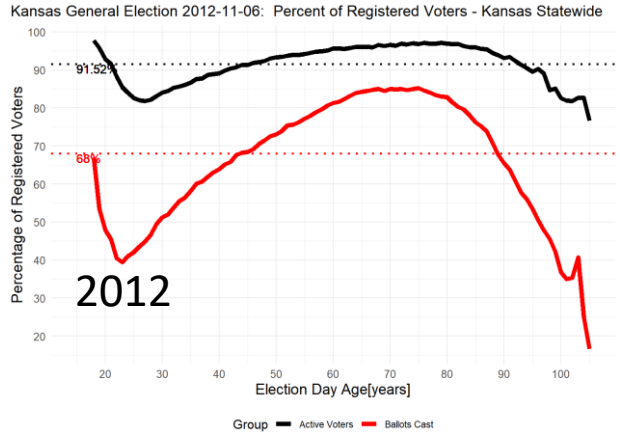


Source: Kansas Secretary of State, Voter File 2019-01-15 efg 2022-04-29 1247

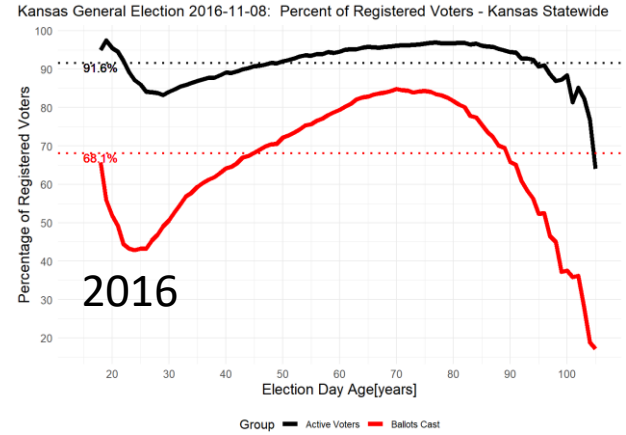
Presidential Election Years



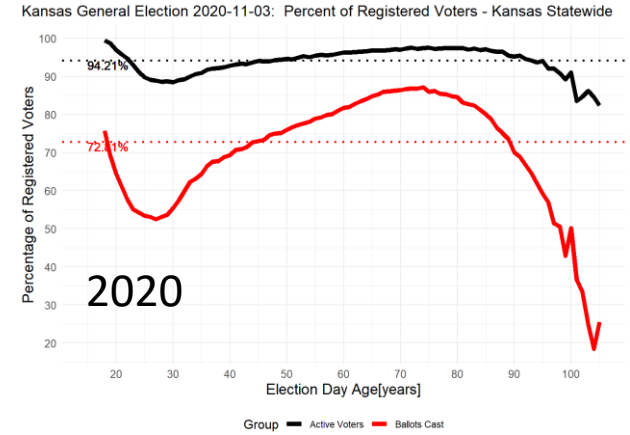
Source: Kansas Secretary of State, Voter File 2009-01-22 efg 2022-04-29 1238



Source: Kansas Secretary of State, Voter File 2013-07-15 efg 2022-04-29 1241



Source: Kansas Secretary of State, Voter File 2017-02-13 efg 2022-04-29 1245



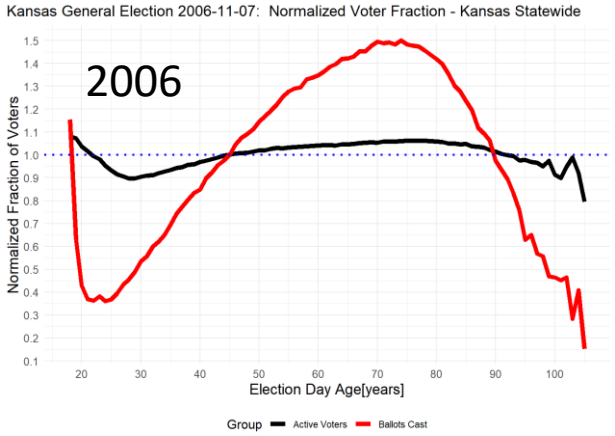
Source: Kansas Secretary of State, Voter File 2021-02-05 efg 2022-04-29 1249

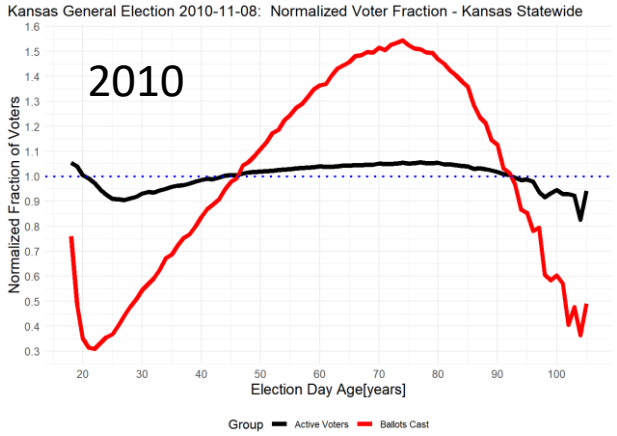
Normalized Turnout for State

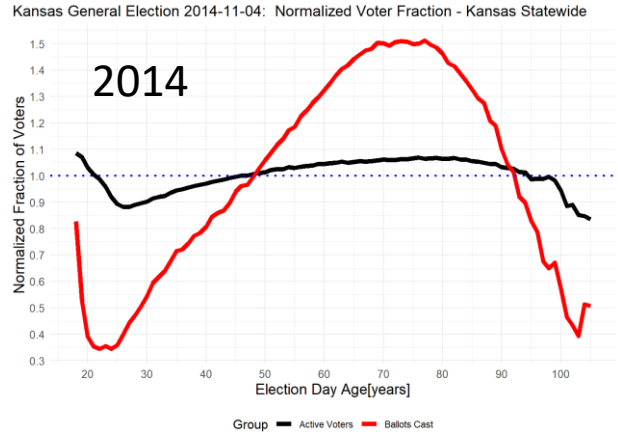
Kansas November General Elections: Normalized by Age

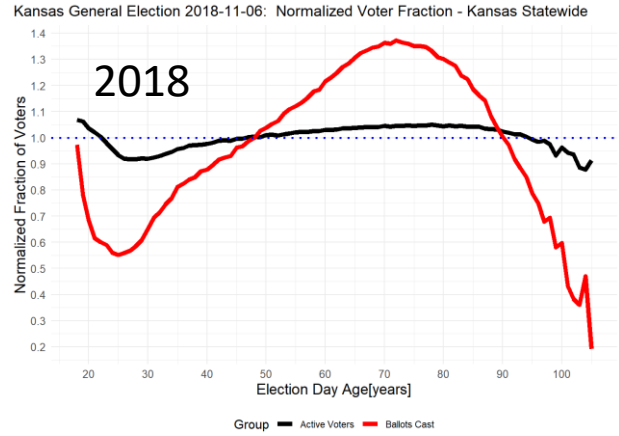
Black Line = “Active” Voters; Red Line = Ballots Cast, **Statewide Turnout**

Gubernatorial Election Years

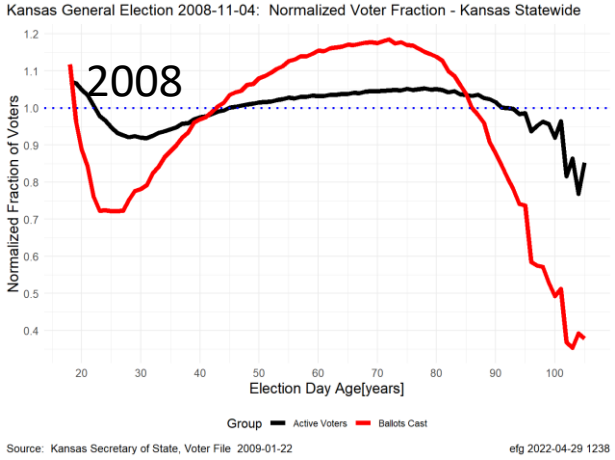


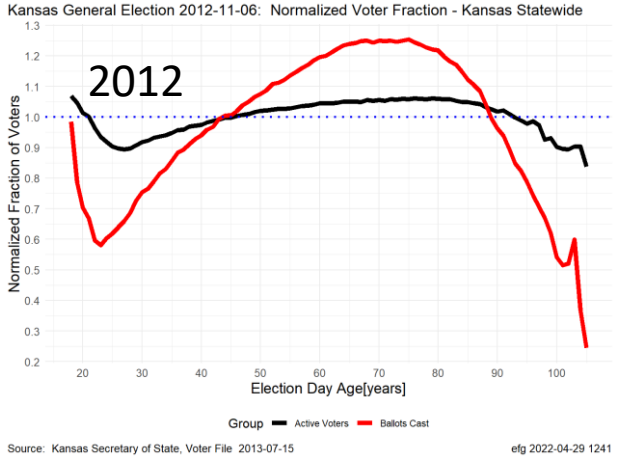


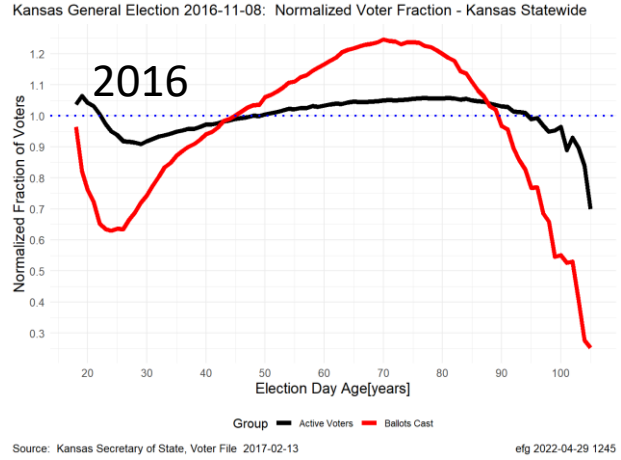


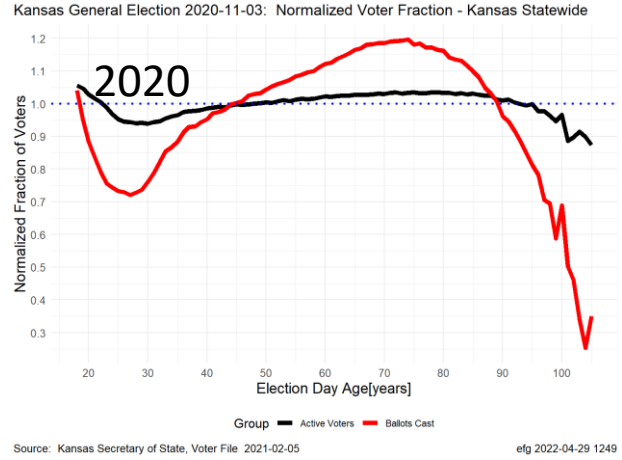


Presidential Election Years







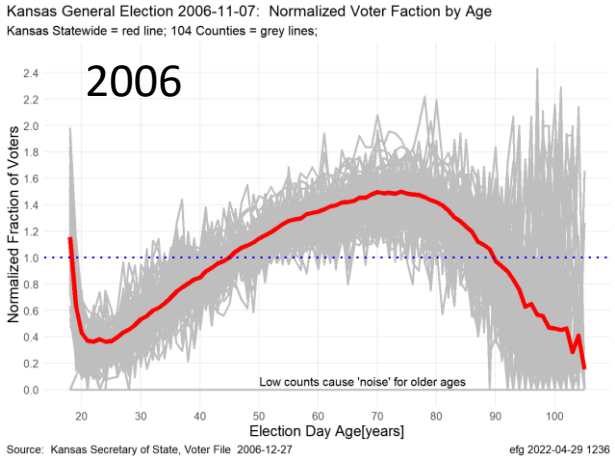


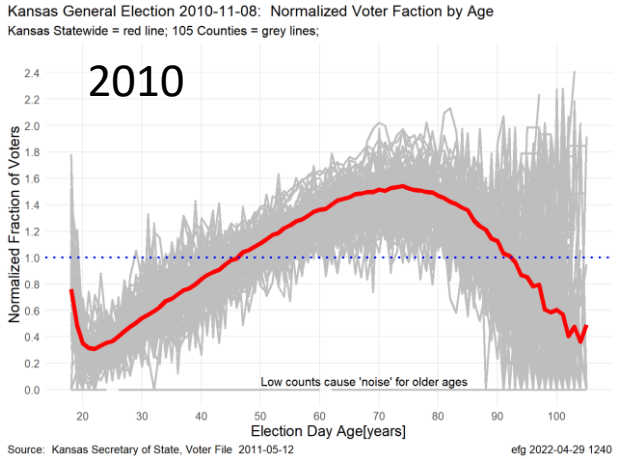
Normalized Turnout for State and Counties

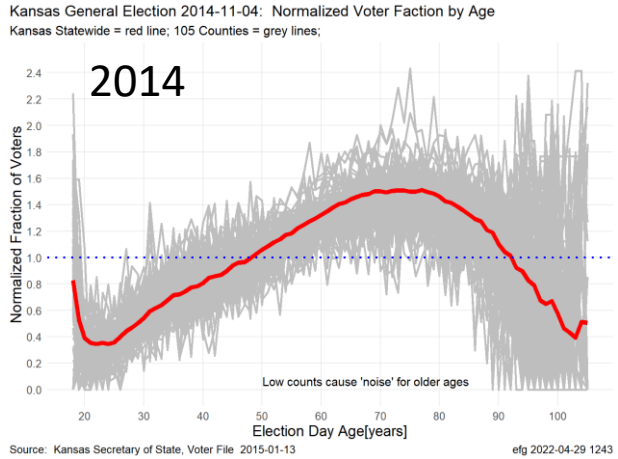
Kansas November General Elections: Normalized Voter Turnout by Age

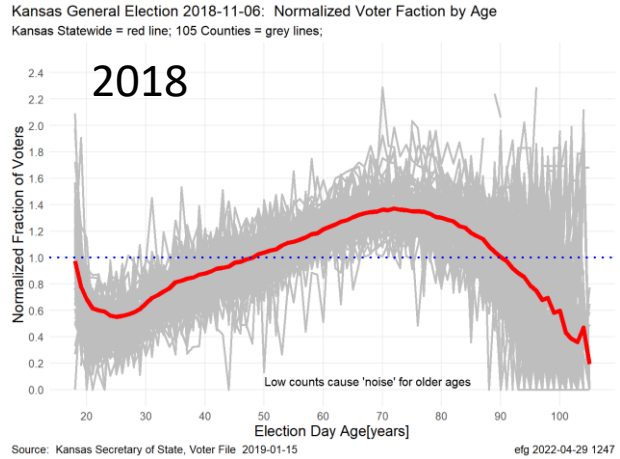
Red Line = Statewide Turnout; Grey Lines = **Turnout for Each of 105 Counties**

Gubernatorial Election Years

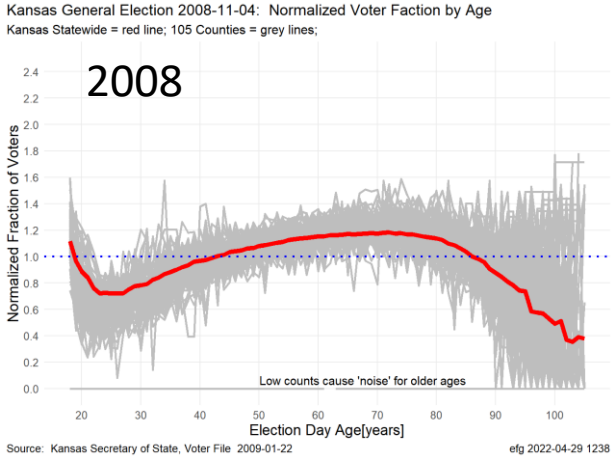


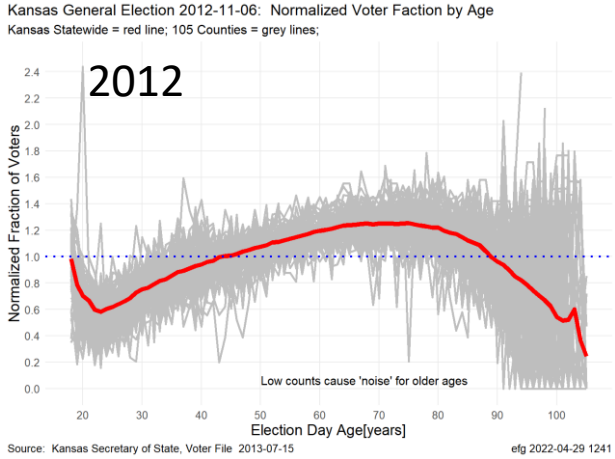


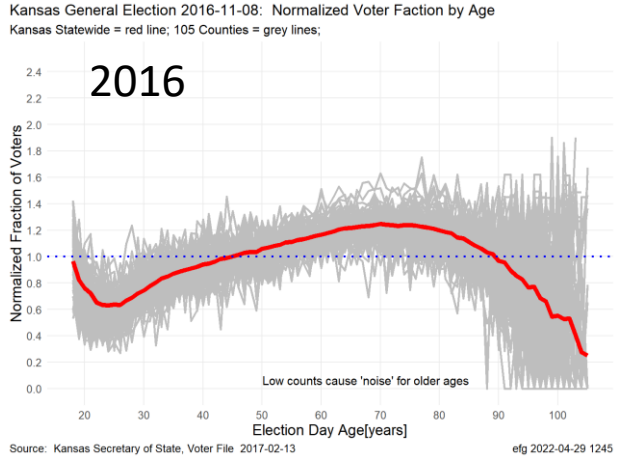


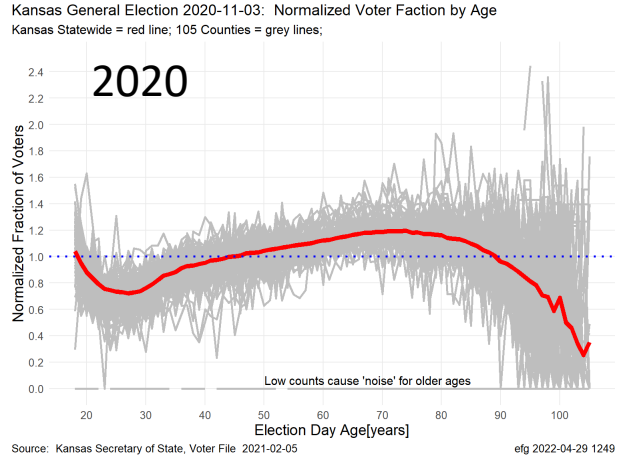


Presidential Election Years









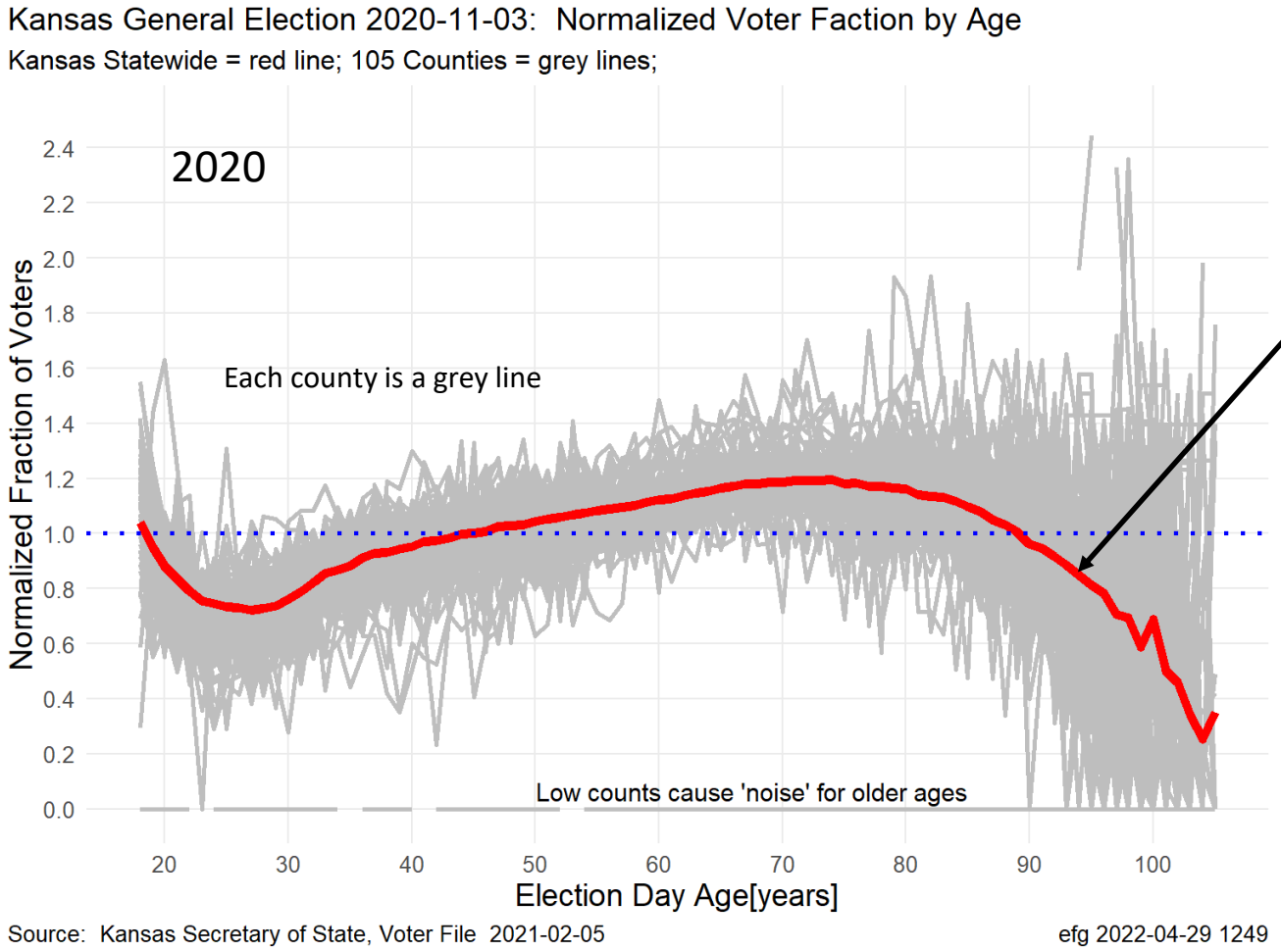
Normalized turnout for a specific age varies slightly by state, county, precinct,
but is highly correlated

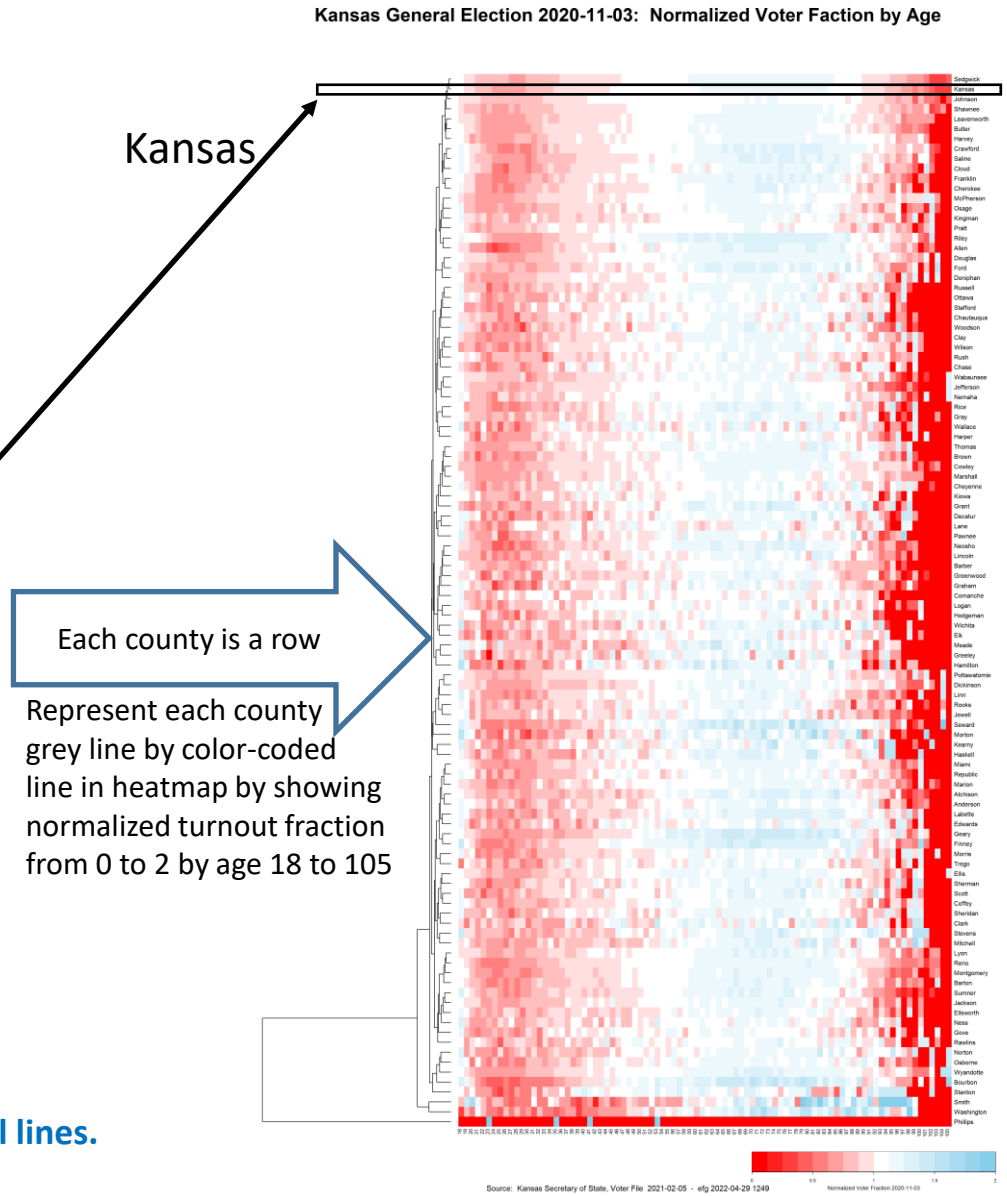
The numbers show turnout is not constant by age within a state when viewed by county or precinct.

- State and Counties
- Johnson County and Precincts

Kansas November General Elections: Normalized Voter Turnout by Age

Red Line = Statewide Turnout; Grey Lines = Turnout for Each of 105 Counties



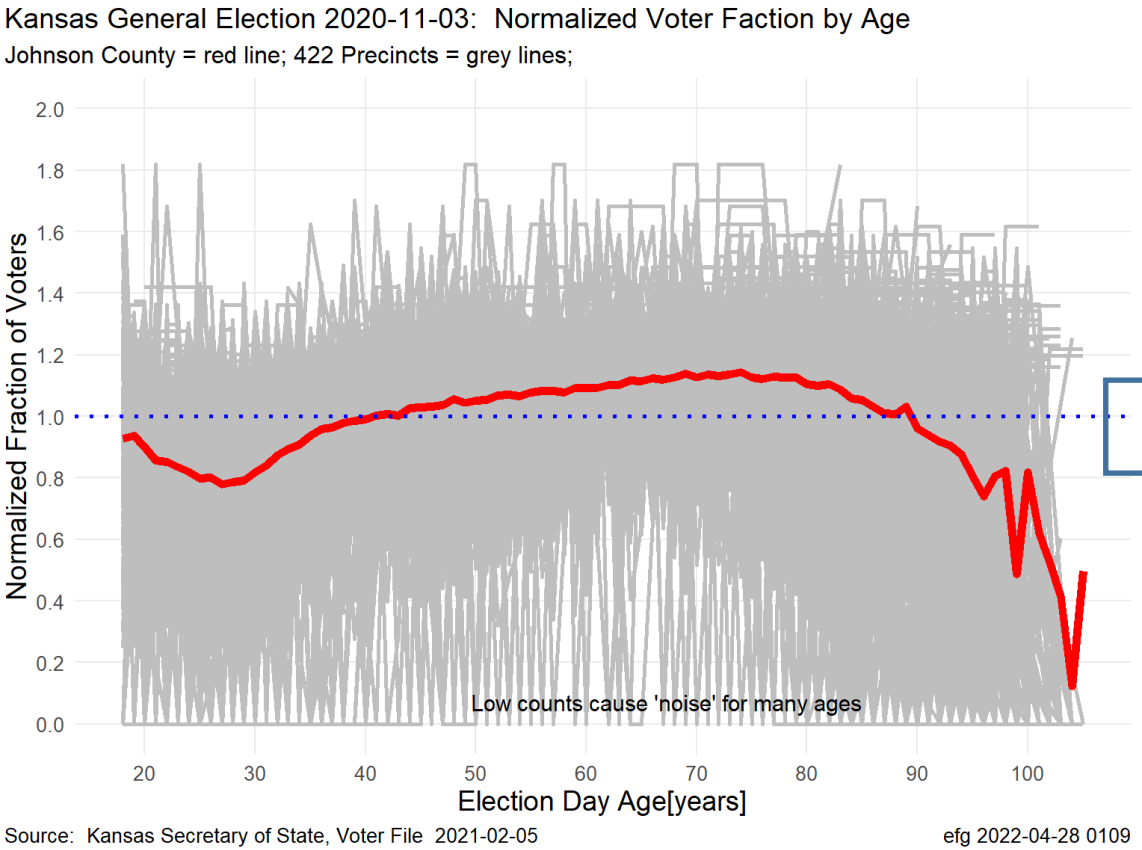


If normalized turnout were constant for a given age, the heatmap would be all vertical lines.

Normalized turnout is only "constant" when evaluated using same polynomial curve fit.

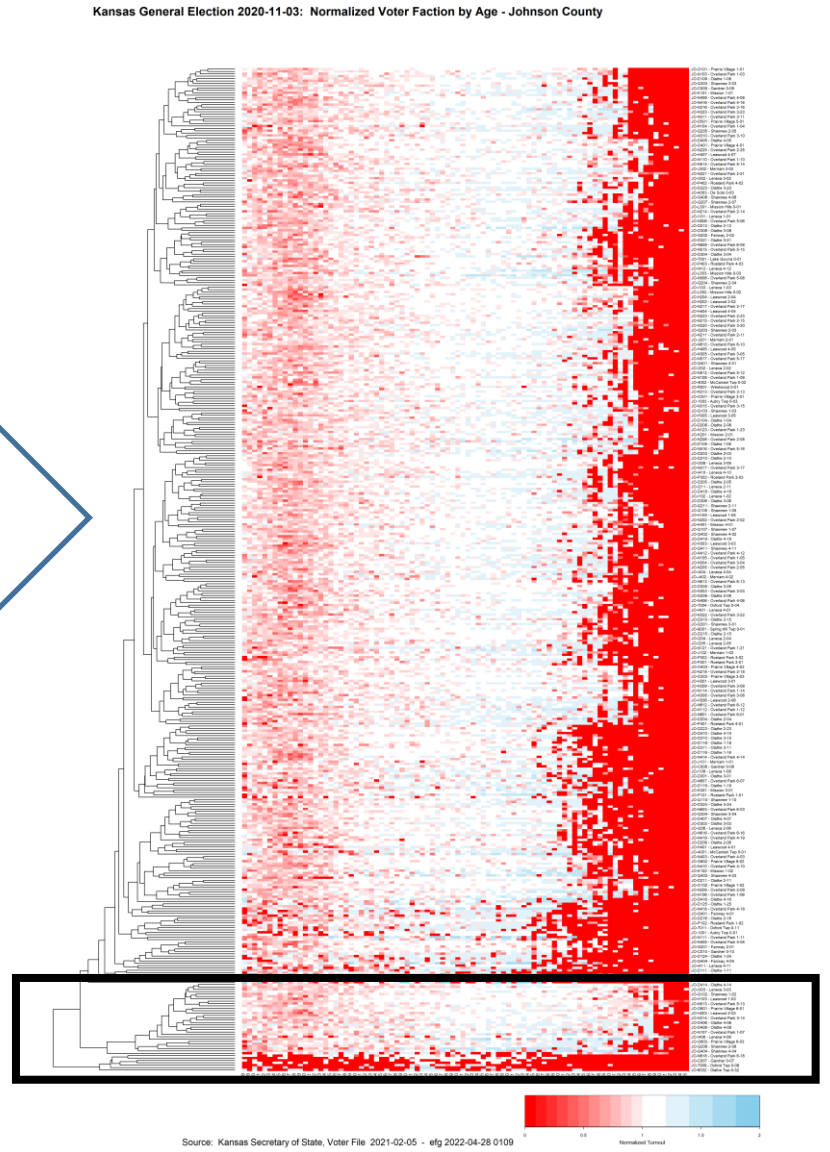
Johnson County Kansas November General Elections: Normalized Voter Turnout by Age

Red Line = County Turnout; Grey Lines = Turnout for Each of 463 Counties



Excluding counties with fewer than 25 registered voters

Each precinct is a row



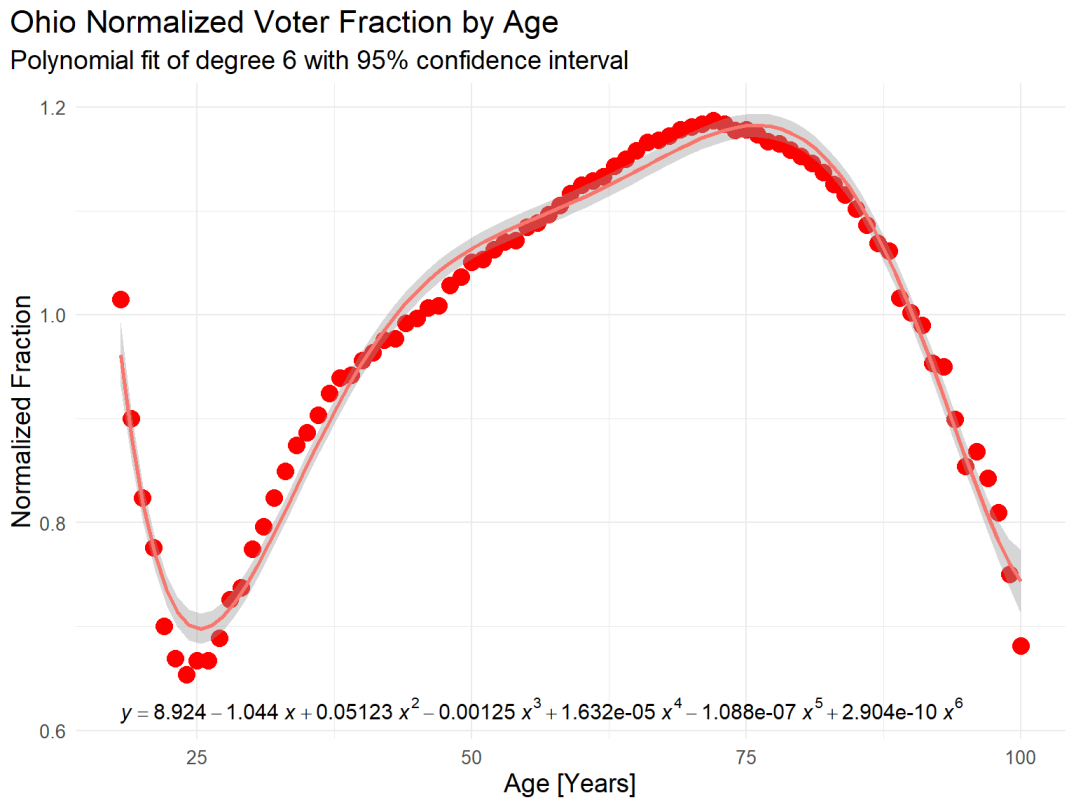
Clade of Older Voting Precincts

I live in one.

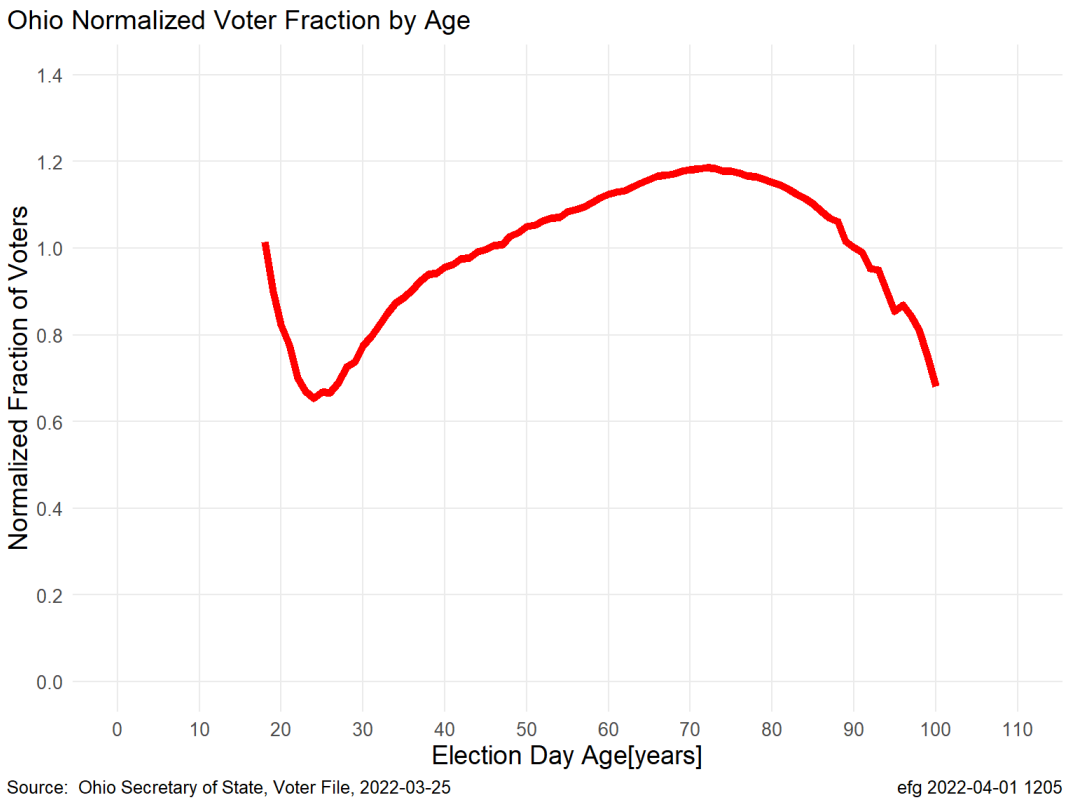
A 6th degree polynomial may not be the “best” fit. Isn’t needed.

Ohio Turnout “Key”

Key: 6th Degree Polynomial Fit to Data (**7 numbers**)

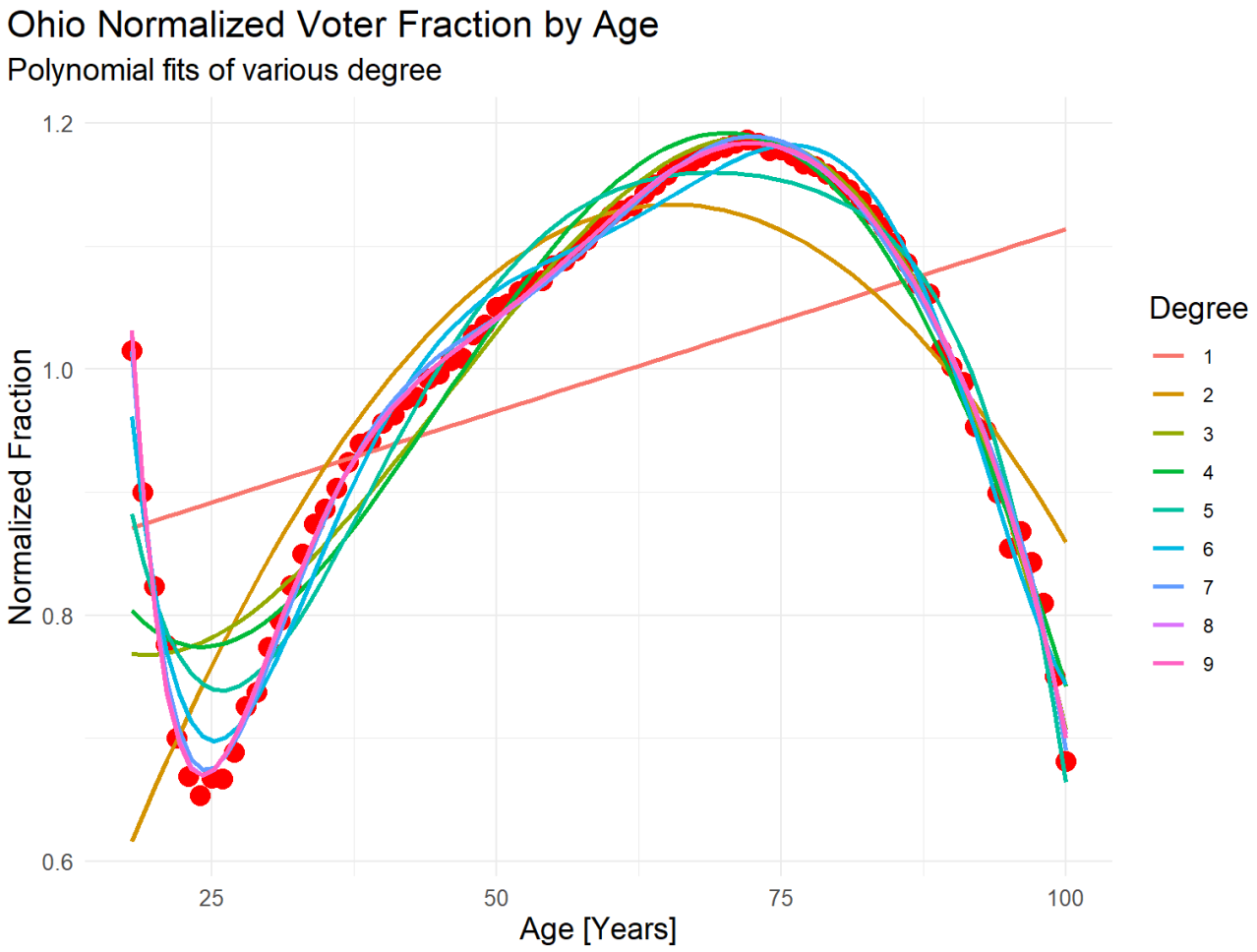


Key: Based on Yearly Data (**83 numbers**)



Scaling is different between plots but both are based on the same data

A 6th degree polynomial may not be the “best” fit. Isn’t needed.

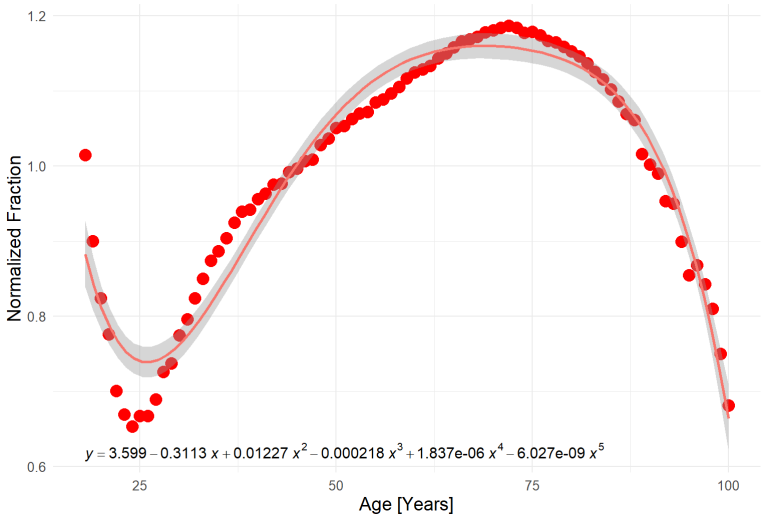


- Here age range limited to [18, 100].
- Fit Ohio Normalized Turnout curve (red dots) to polynomials of various degrees
- Higher degree provides curvature/ “wiggliness”, but too high can lead to overfitting.
- Akaike Information Criterion (AIC) indicates highest degree over range 1 to 9 was the “best” model.
- R^2 approaches 1 as degree increases.
- There is nothing “remarkable” about these curve fits.
- Curve fits offer few insight about data but provide good numerical interpolation.

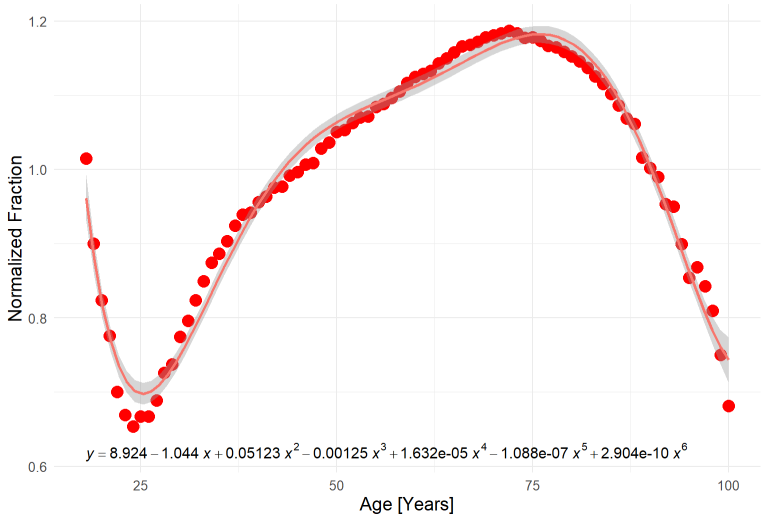
RStudio Notebook: Ohio-Key-Polynomials.html, Section 5.1 (AIC), Section 5.2 (R^2), polyfitsMany-1.png

A 6th degree polynomial may not be the “best” fit. Isn’t needed.

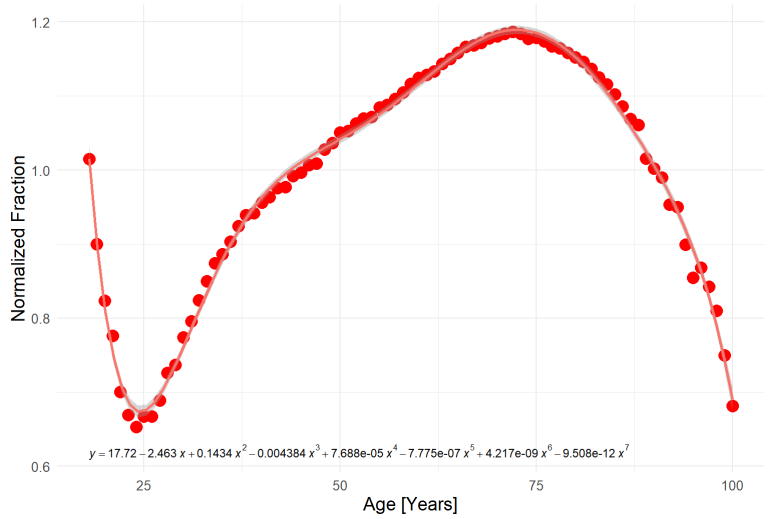
Degree 5



Degree 6



Degree 7



In polynomial equations above, x = Age, y = Normalized Fraction

Largest residual over range: 0.132 (5^{th}), 0.062 (6^{th}), 0.038 (7^{th})

But, there is *no need for polynomial fit if original normalized turnout curve is used directly!*

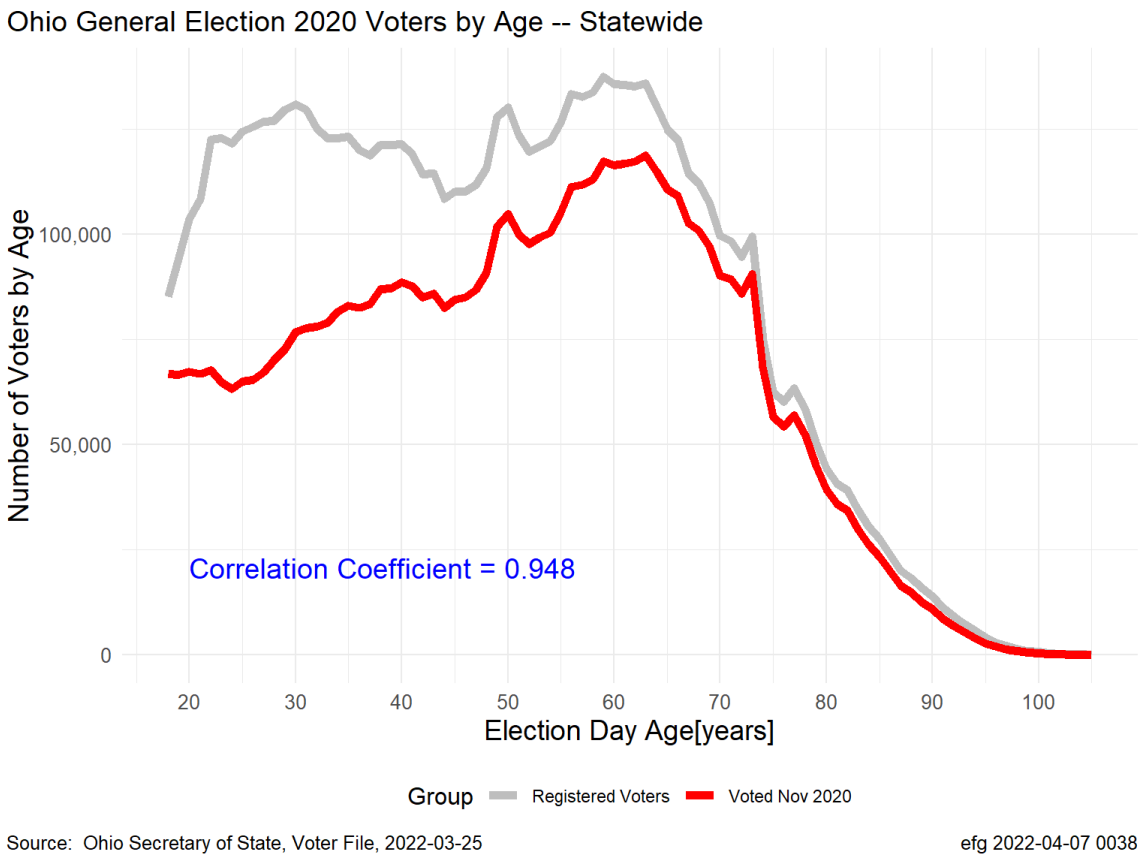
RStudio Notebook: Ohio-Key-Polynomials.html, Section 6, polyfit5-1.png, polyfit6-1.png, polyfit7-1.png

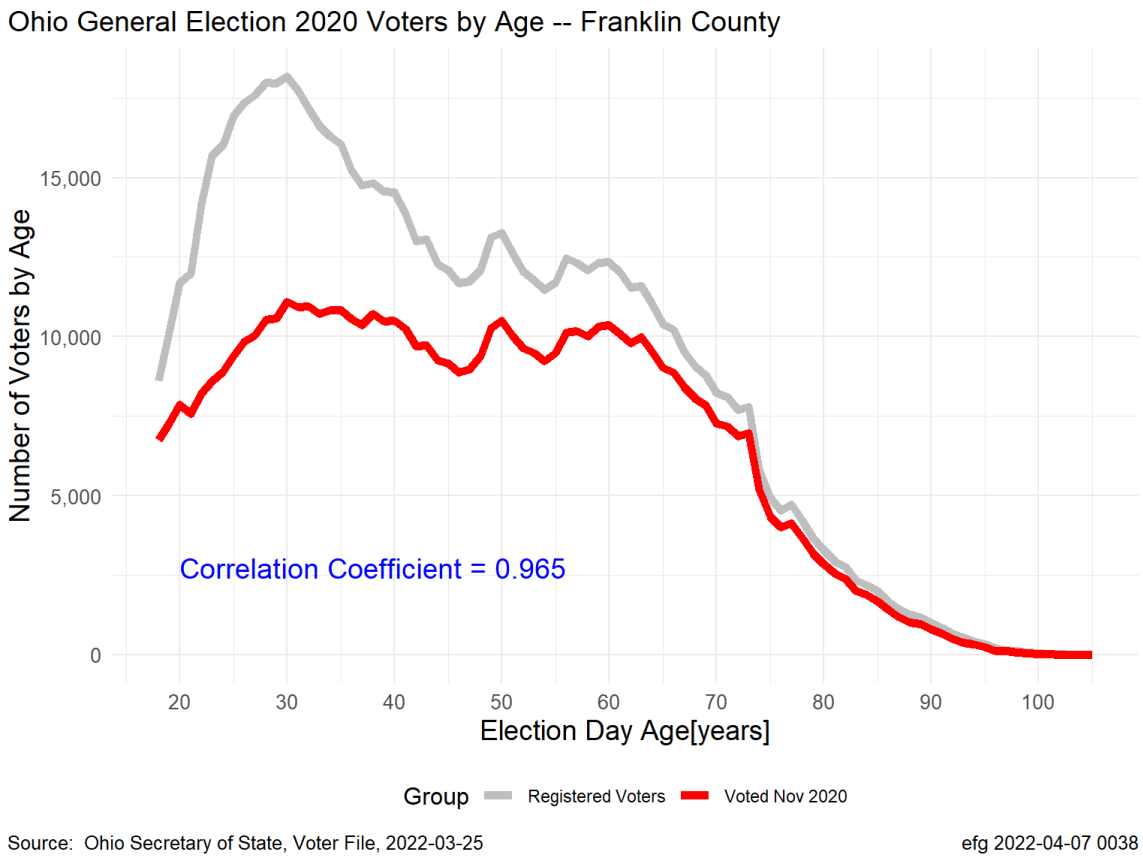
Correlation is a weak way to compare predictive models

Let’s use correlation to measure similarity between number of registered voters and the number casting ballots by age.

- Examples: Ohio Statewide, Franklin County

Correlation is not a good comparison metric for predictions

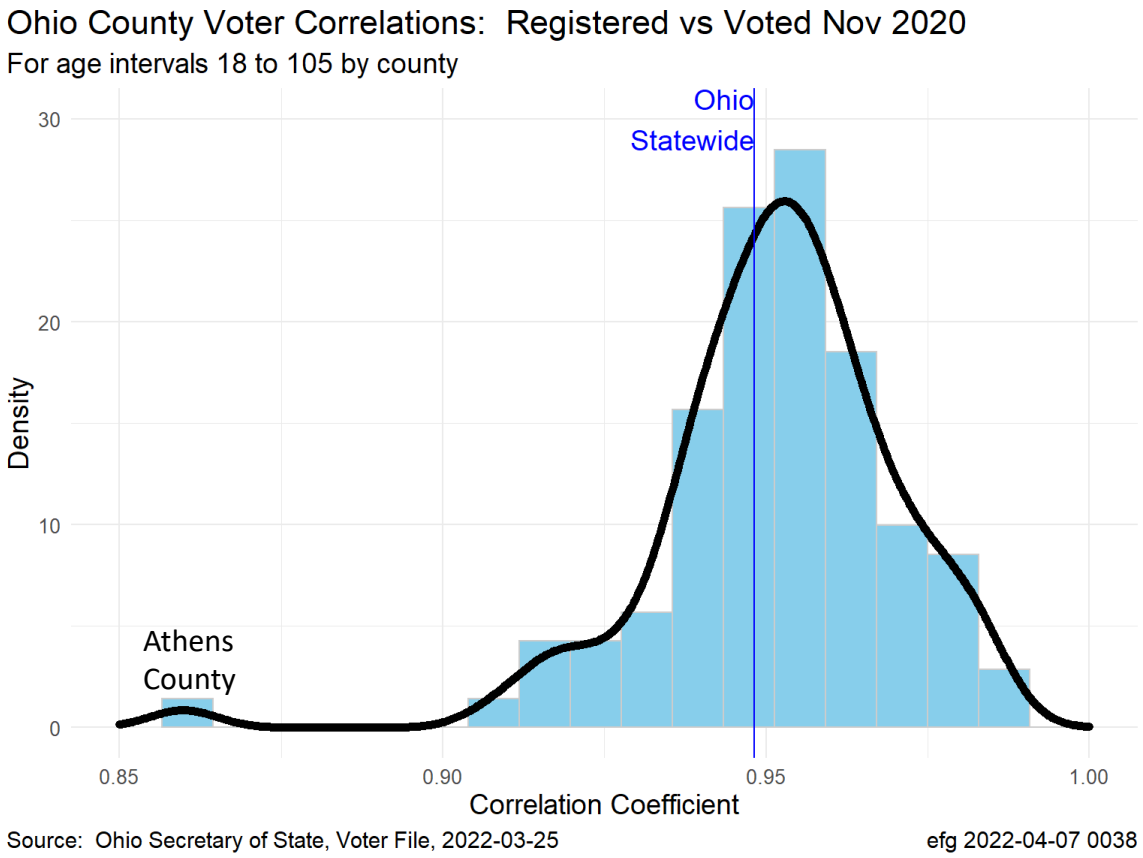


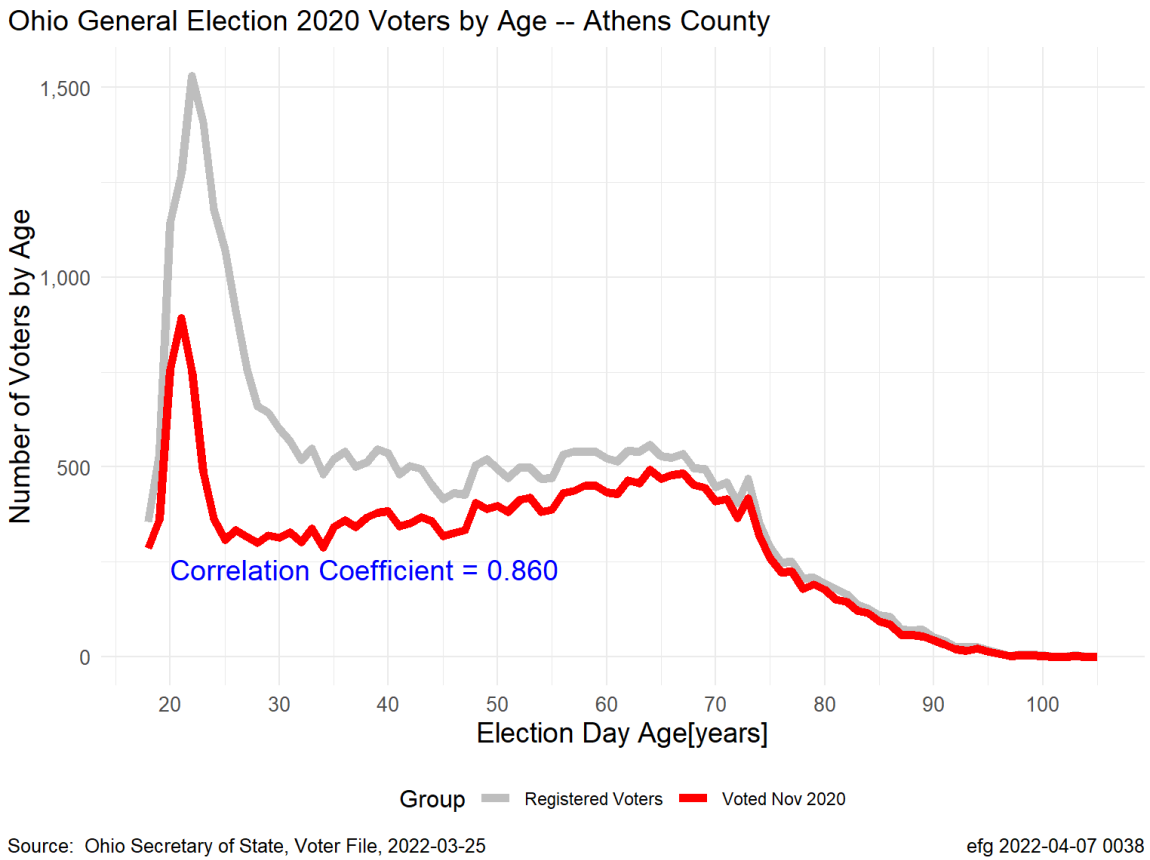


Most population and voter turnout curves are highly correlated

Ohio-Correlations.html, Statewide-1.png, FranklinCounty-1.png

Correlation is not a good comparison metric for predictions





In most counties the correlation between the number of registered voters and the number voting in Nov 2020 over age intervals 18 to 100 was between 0.90 and 1.00 with a an overall state value of about 0.95.

Athens County was the outlier in the density plot.

Accuracy Better than Correlation for Assessing Predictions

Ohio-Analysis-5-Predictions-vs-Acutal-Votes.ipynb

Prediction = overall turnout * registered * keyvalue

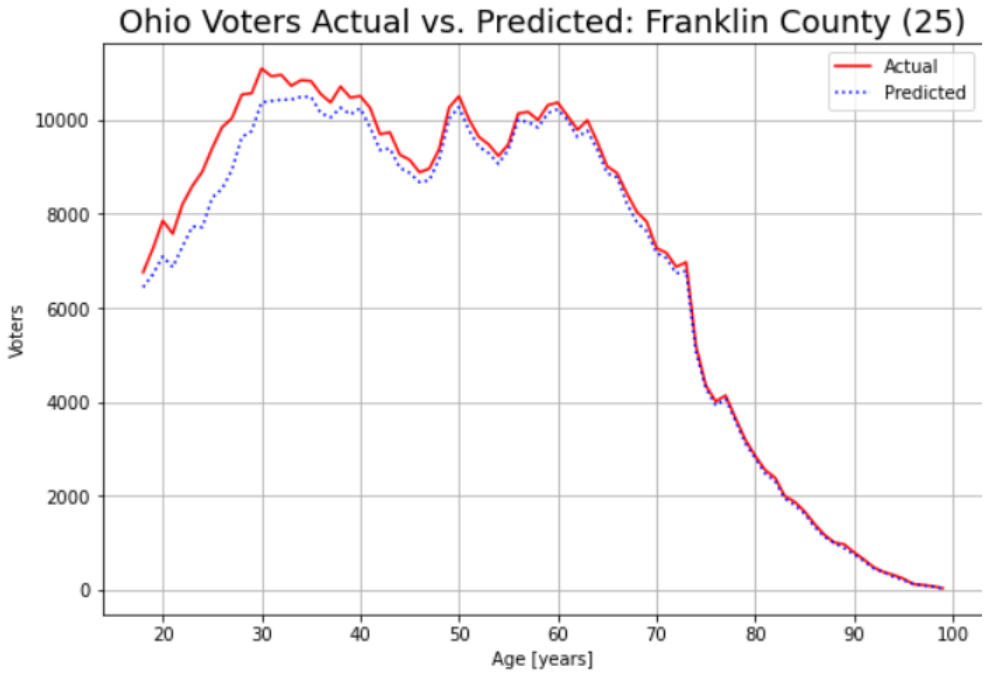
Ohio Key

```
> 0.73707*8607*1.01483
[1] 6438.042
```

Franklin County

age	voted	registered	keyvalue	prediction	Delta	absDelta
18	6,756	8,607	1.01483	6438.0	-318	318
19	7,264	10,124	0.89988	6714.9	-549	549
20	7,857	11,689	0.82371	7096.7	-760	760
21	7,579	12,001	0.77607	6864.8	-714	714
22	8,207	14,157	0.70037	7308.1	-899	899
. . .						
100	27	48	0.68139	24.1	-3	3
101	12	20	0.60789	9.0	-3	3
102	6	15	0.54378	6.0	0	0
103	3	8	0.59521	3.5	1	1
104	3	5	0.37173	1.4	-2	2
105	1	1	0.33391	0.2	-1	1
Total	**574,067**	**778,854**	**1.00000**	**574067.0**	**0**	**22,915**
overall turnout		**0.73707**				
% Error						**3.99%**

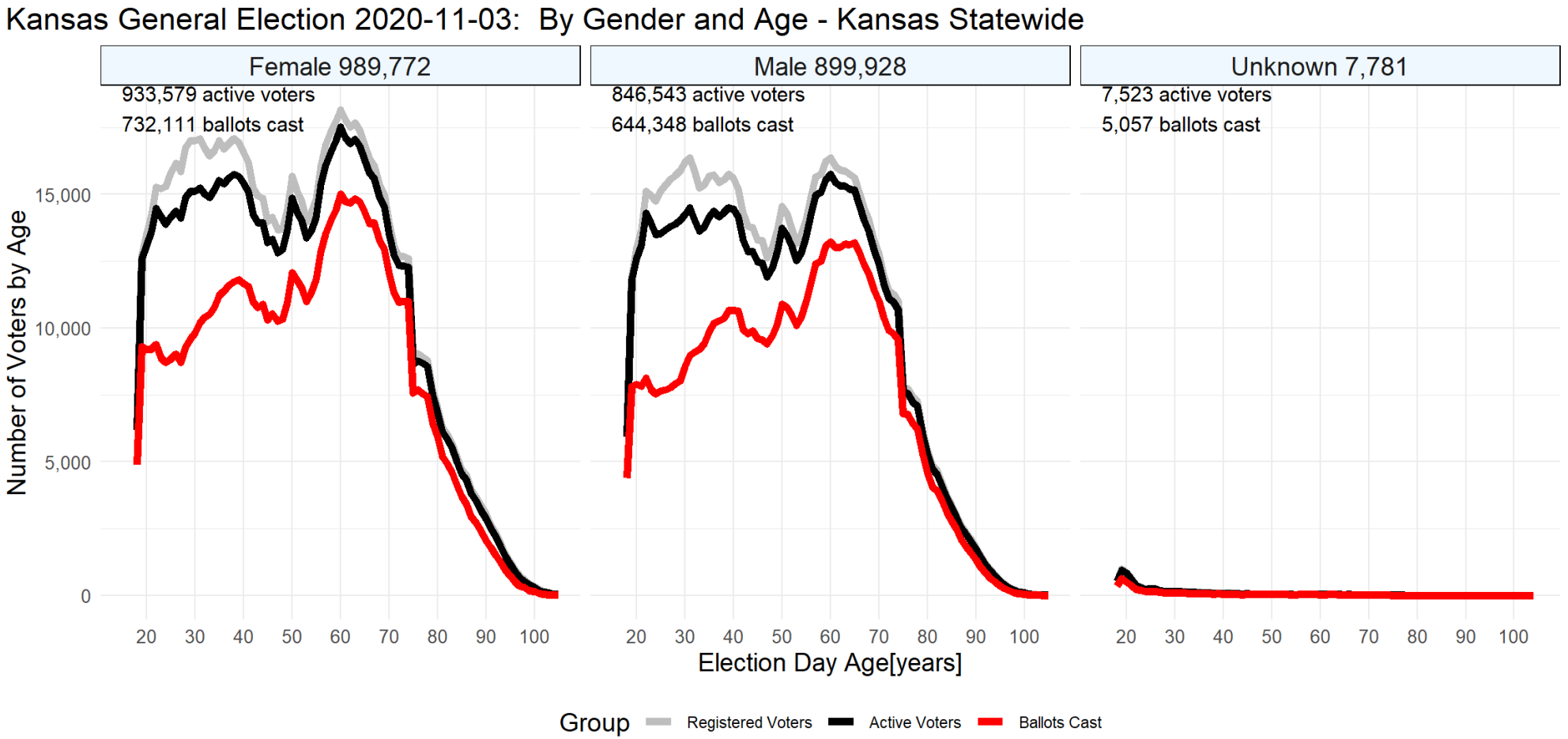
Accuracy Metric



Adapted from Lee's script: predict.py
Ohio-Analysis-2-Single-County-Franklin.html
Ohio-Analysis-5-Predictions-vs-Acutal-Votes-Franklin.html, **Compare-Franklin-25.xlsx**

Turnout varies by gender

(but party not available in all states)

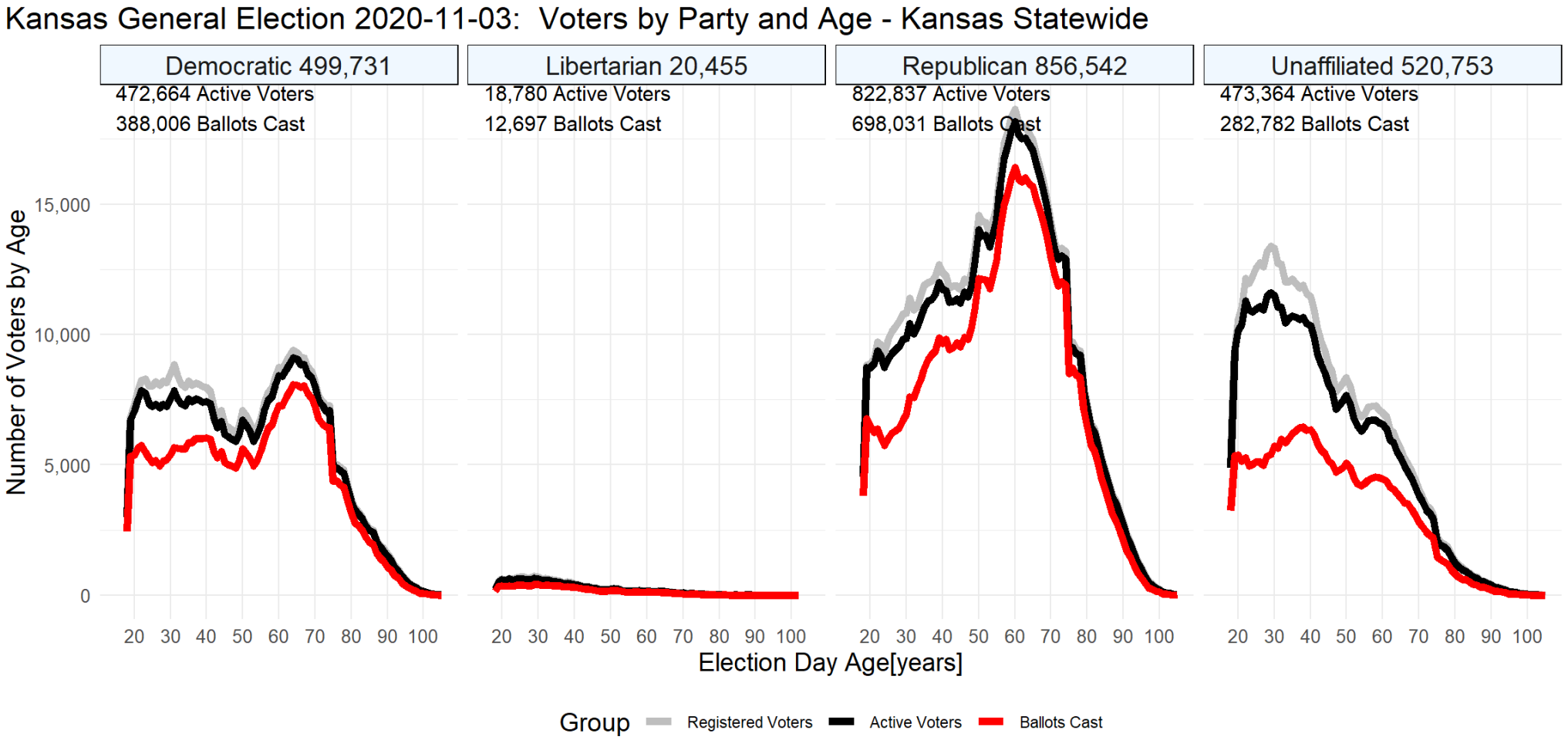


Source: Kansas Secretary of State, Voter File 2021-02-05

efg 2022-04-29 1249

Turnout varies by political party

(but party not available in all states)



Source: Kansas Secretary of State, Voter File 2021-02-05

efg 2022-04-29 1249

Inflated voter rolls are not new but may be worse

WATCHDOGLABS.org

Michigan lost 55,000 people but gained 500,000 voters between 2000 and 2010 census

efg October 24, 2012 at 2:47 pm

By Earl F Glynn | Franklin Center

Oct. 24, 2012
93 counties in 17 states with voter registration ≥ 100%
132 counties in 17 states with registration ≥ 95%

Oct. 16, 2020
359 counties in 29 states with voter registration > 100%

Write a story about it and some counties fix the problem!

The cost to monitor all the states is prohibitive, but is affordable for many states, e.g., OH, MI, NJ, DC, FL.