
# Leveraging Data Science to promote transparency in elections: The case of Peru

Giancarlo Marchesi

Final Presentation - Data Science Bootcamp

# Problem Statement

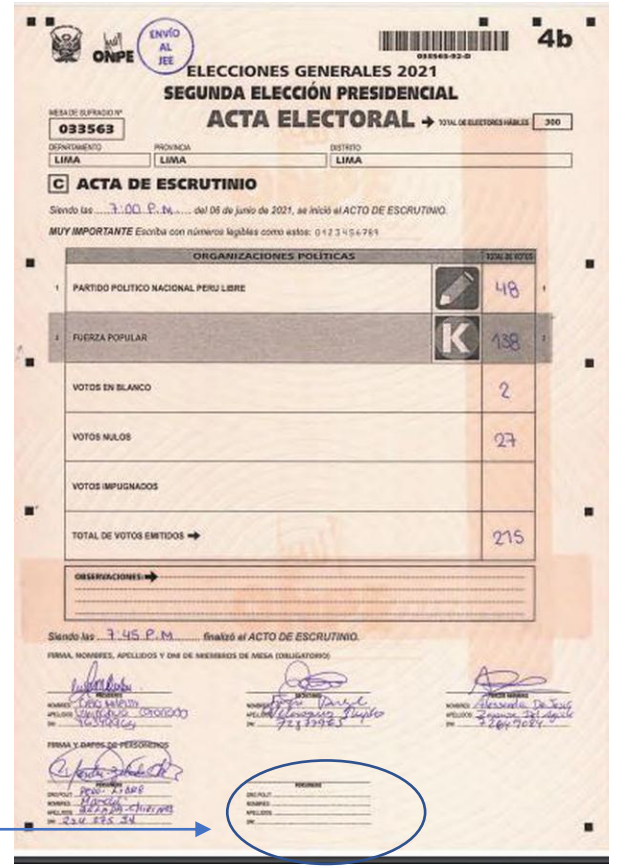
- Investigate if the controversy around the results of the run-off presidential election in Peru has merits
- Will the outcome of the election change if votes from *alleged controversial* polling stations were discarded?

# Background

- Run-off election on June 6th
- Castillo won by 44,240 votes
- Fujimori has claimed foul play

***Why?***

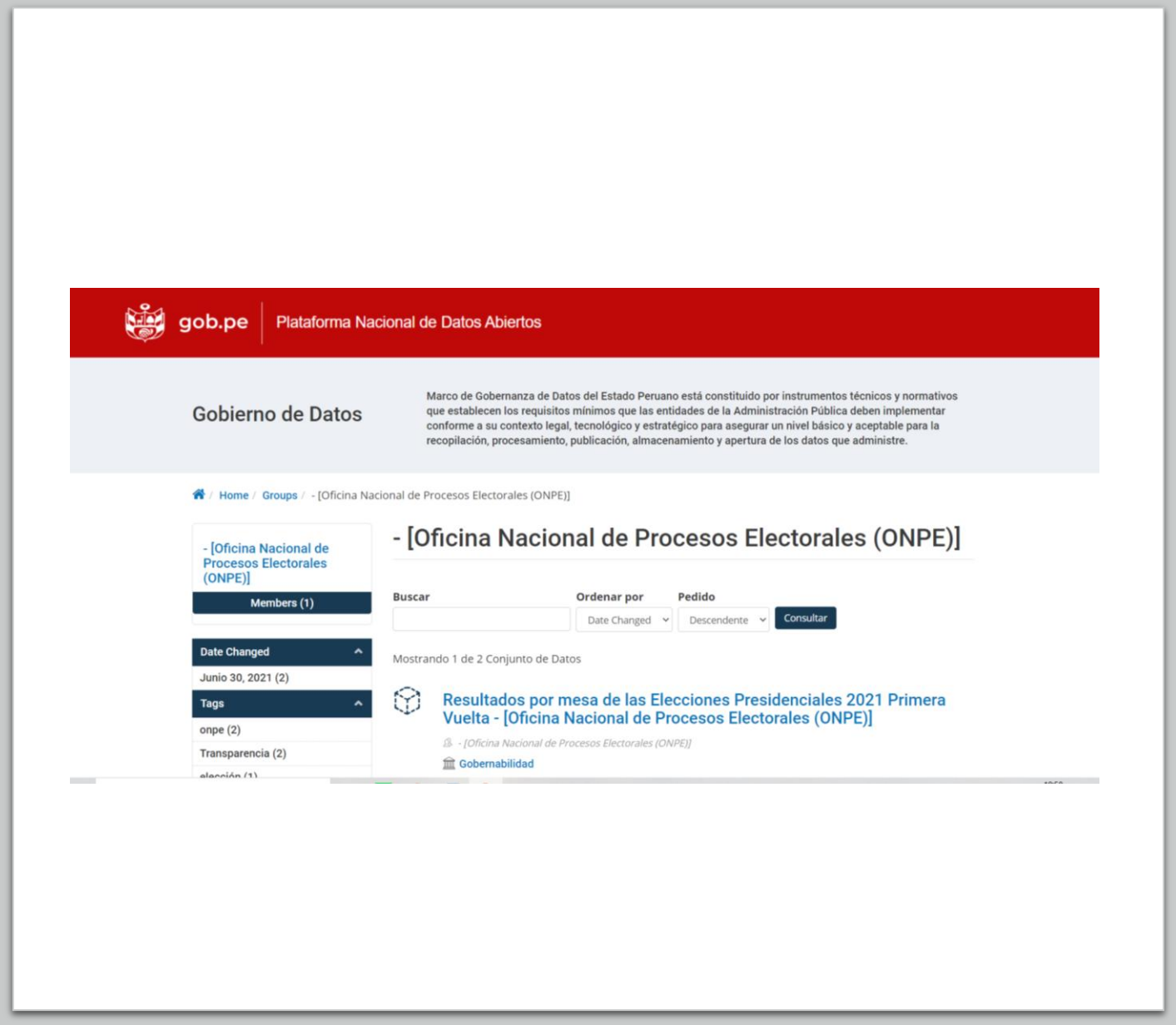
- The results are returned in ***handwritten summaries*** for tabulation
- Paper ballots are discarded after they are tallied when polls close.
- Summaries are signed by the poll workers and 2 representatives from parties.

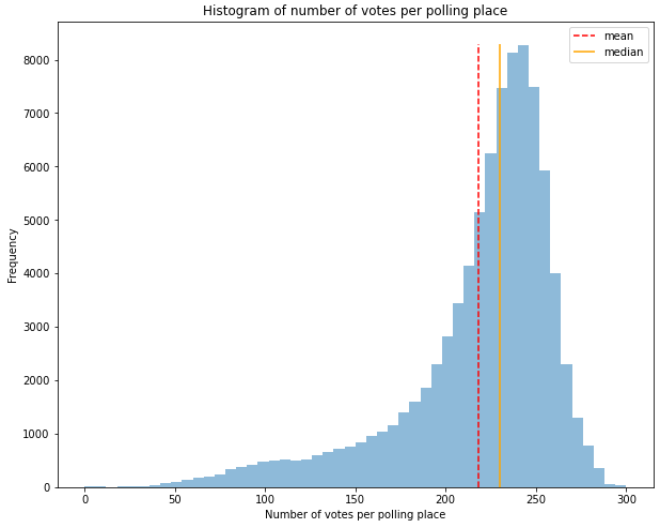

Lower than expected vote count in regions with few representatives

# The Data

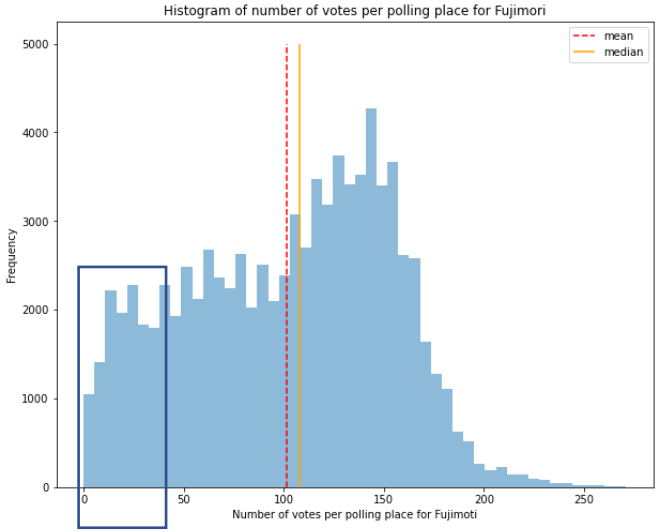
- Open data from the Peruvian Government
- Official ONPE vote count at the poll level for first round and runoff election.
- 86488 polling stations for both rounds
  - Consistency checks:
    - Only polling stations with valid data for both rounds
    - Eligible voters equal in both rounds
  - Feature engineering: delta variables



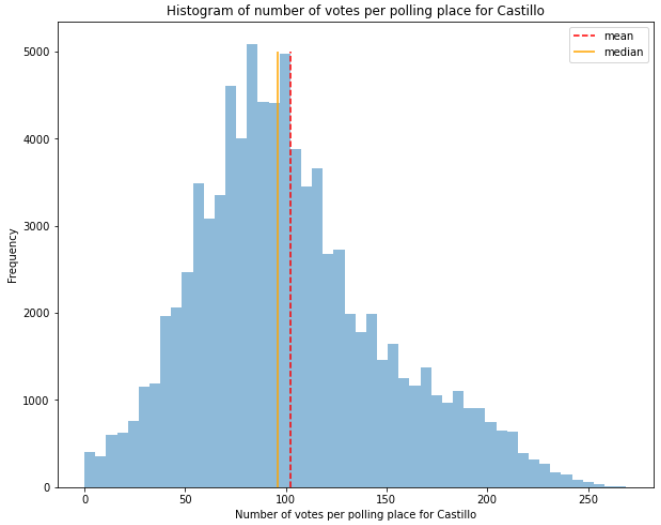
Sum of both candidates




Fujimori




Castillo




# Exploratory Data Analysis: vote count per candidate

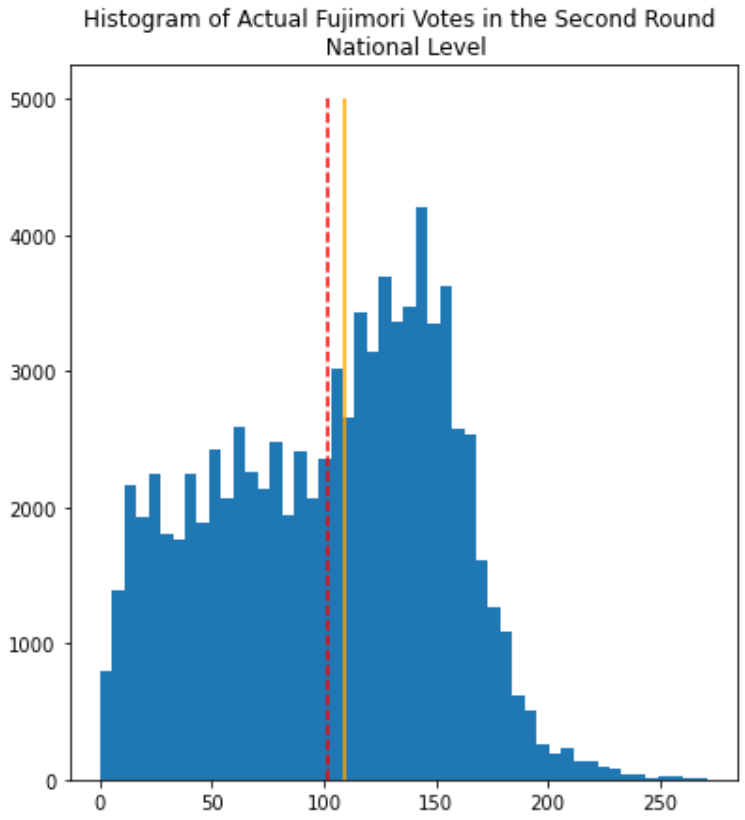
# A linear regression model does a pretty good job at predicting the results of the second round

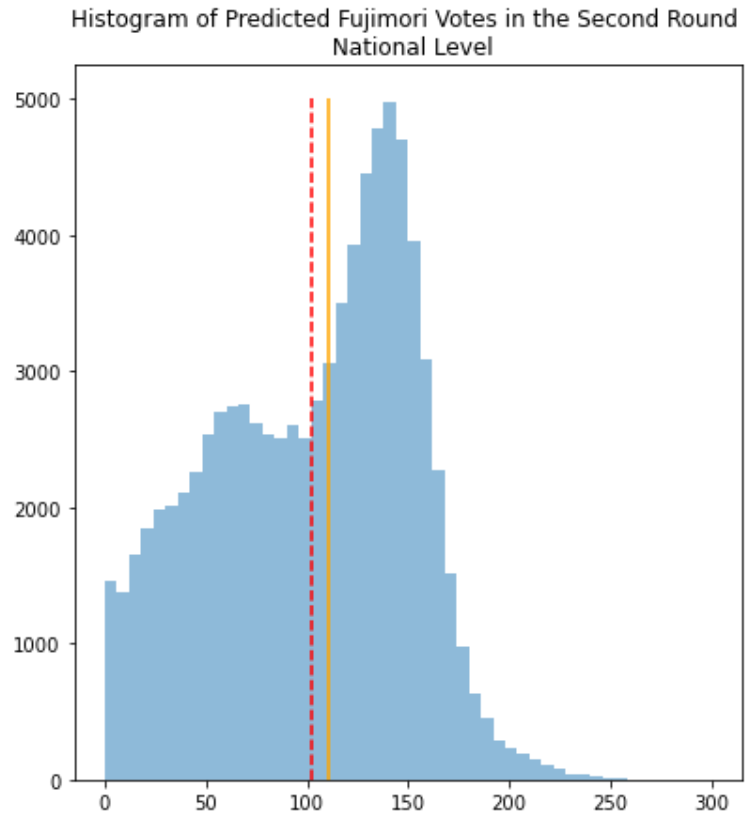
Regressors:

- Votes of 18 candidates in the first round
- Change in turnout between rounds

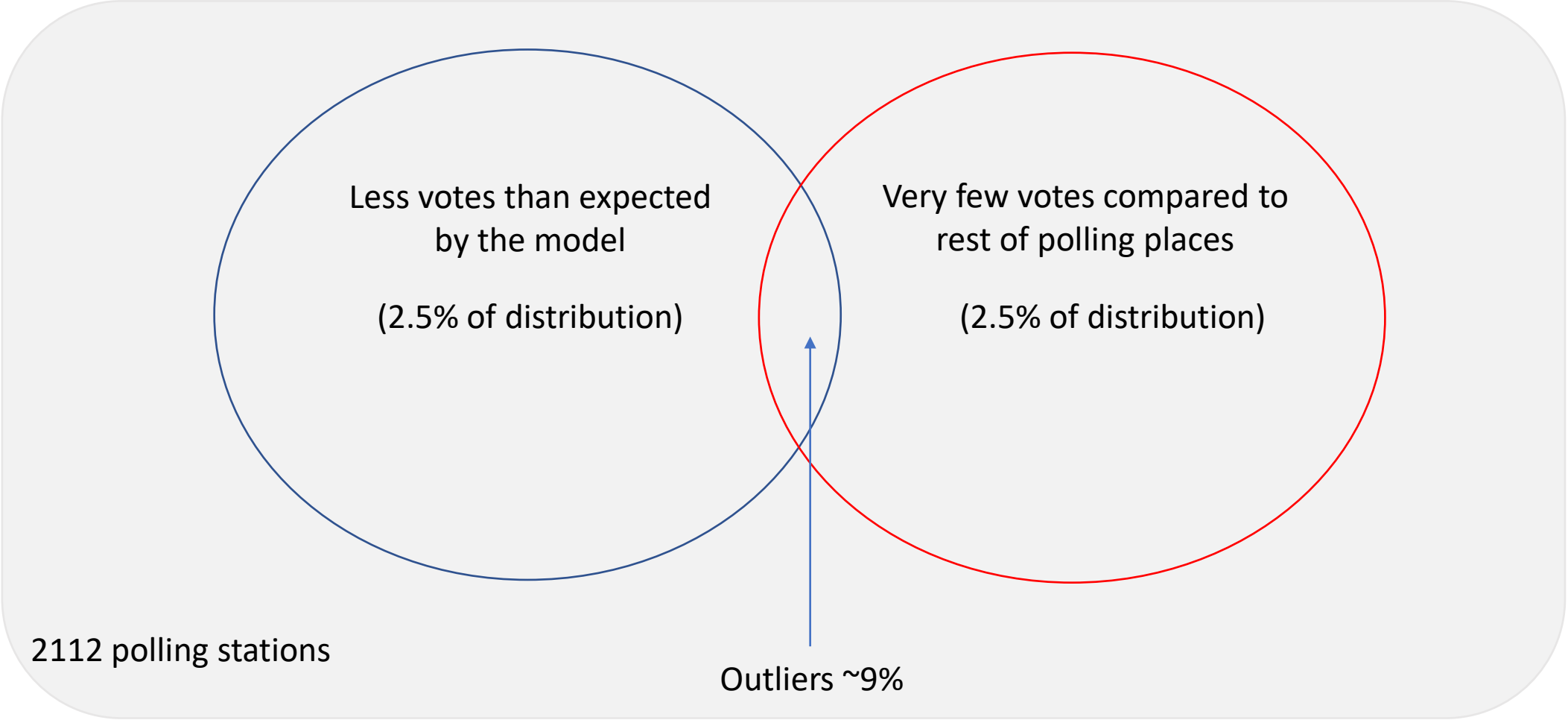
Model for Castillo: R2=0.90

R2= 0.92

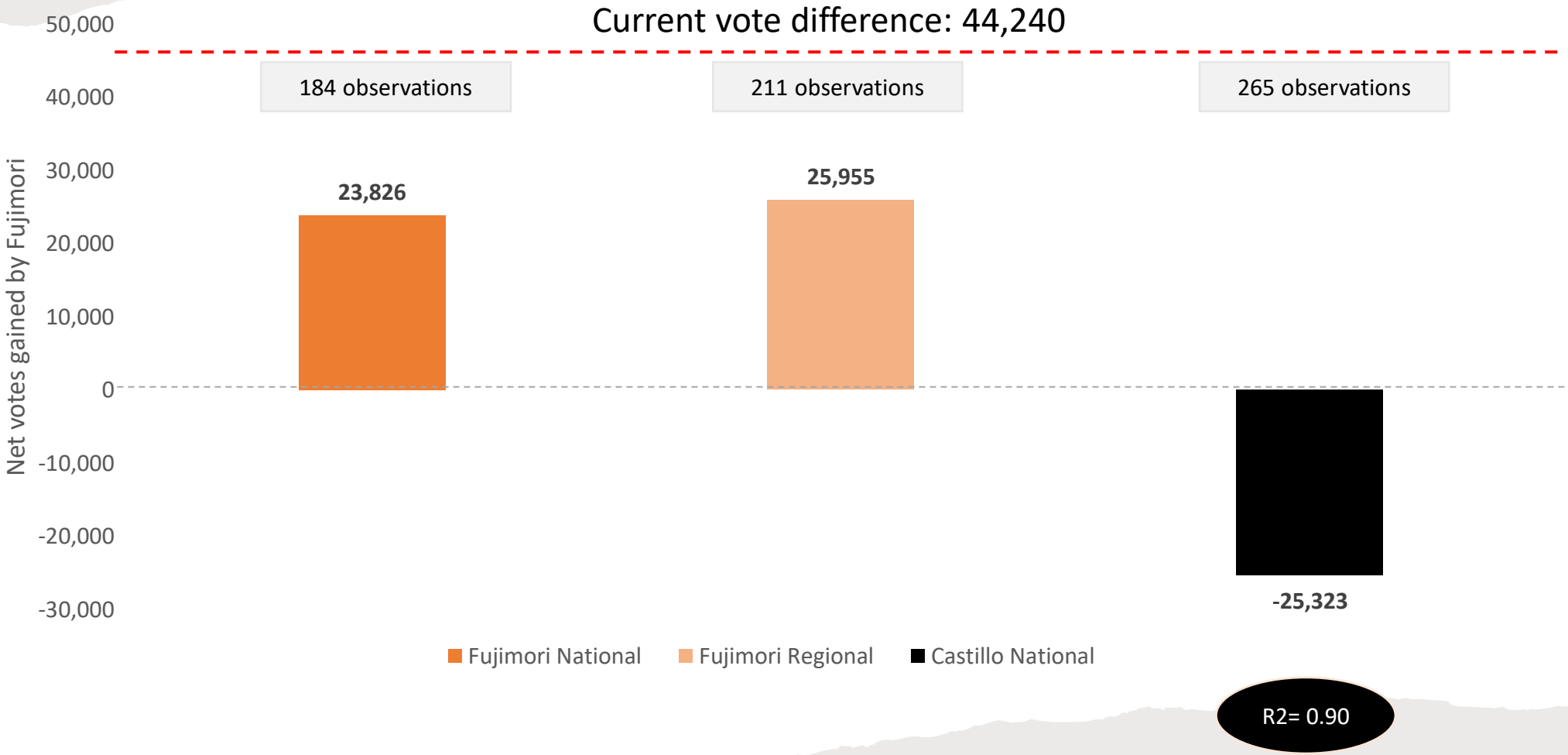


# What constitutes an outlier?



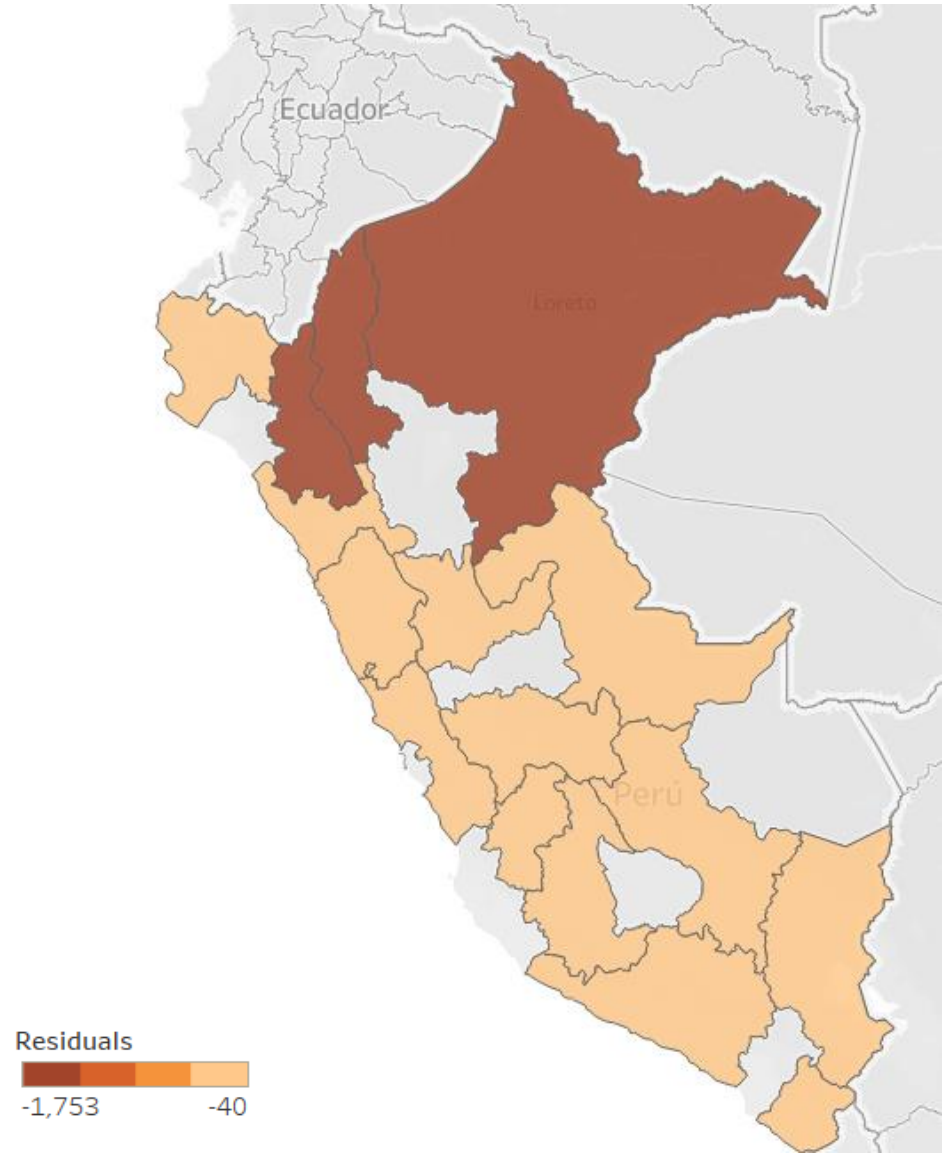
# The results of the election don’t change, even if outliers are completely discarded



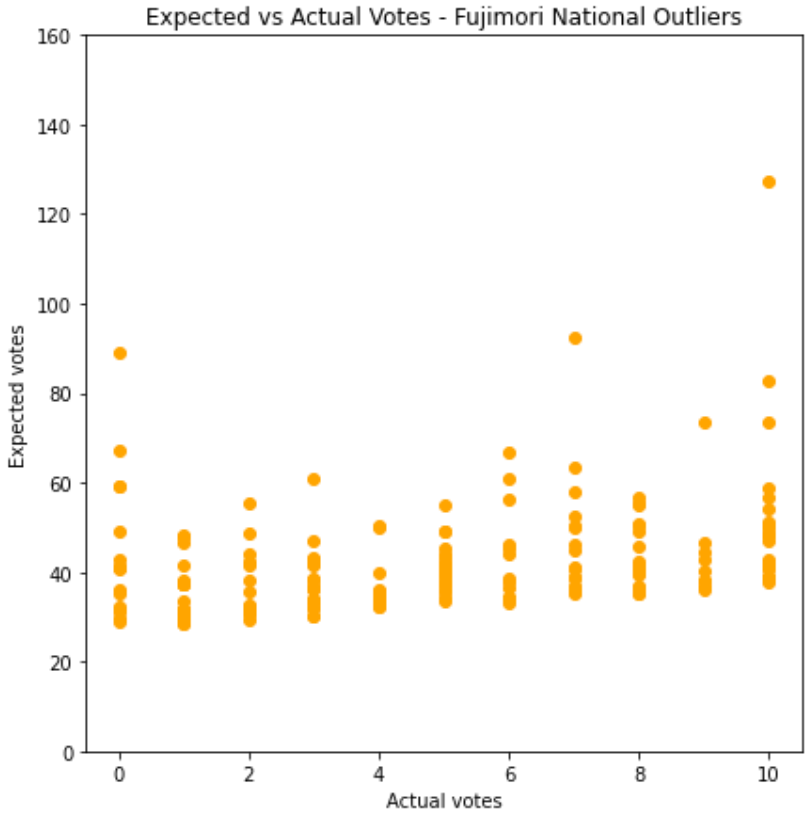
Residuals (less than expected votes) :
National model: 6910 votes
Regional model: 6995 votes

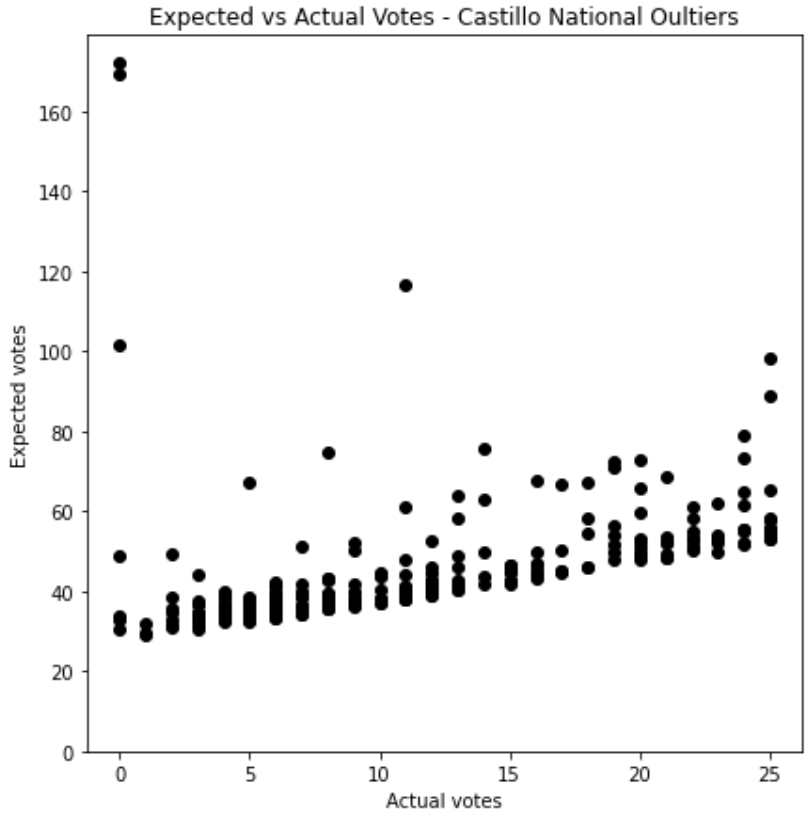
# No marked geographical pattern for Fujimori outliers at national level

Residuals are not more than 1800 in any state

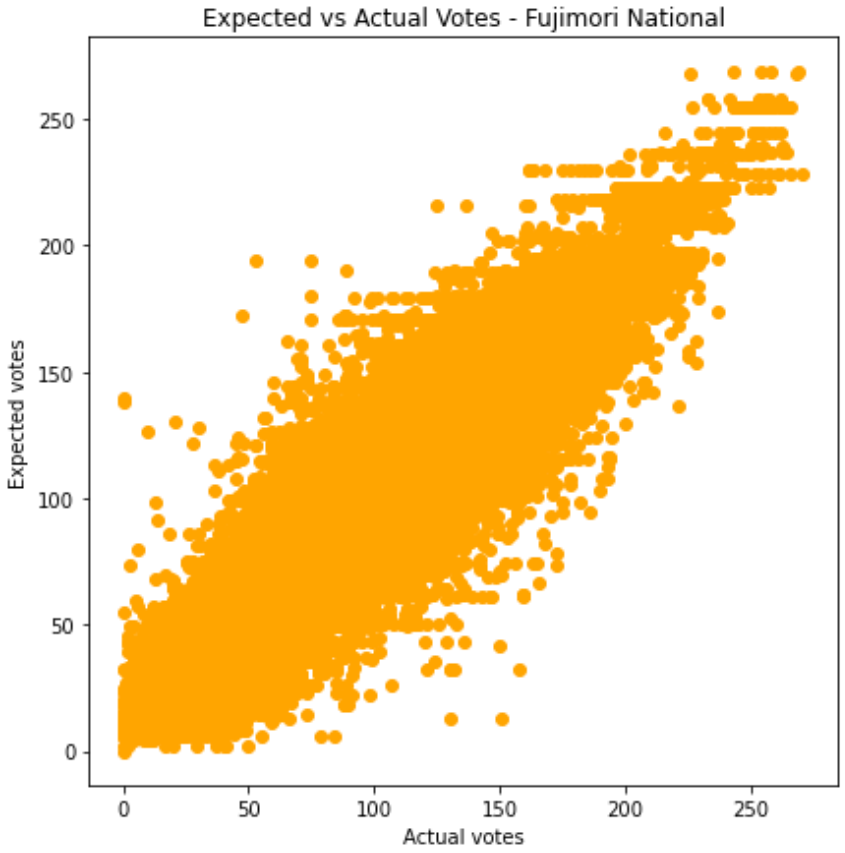


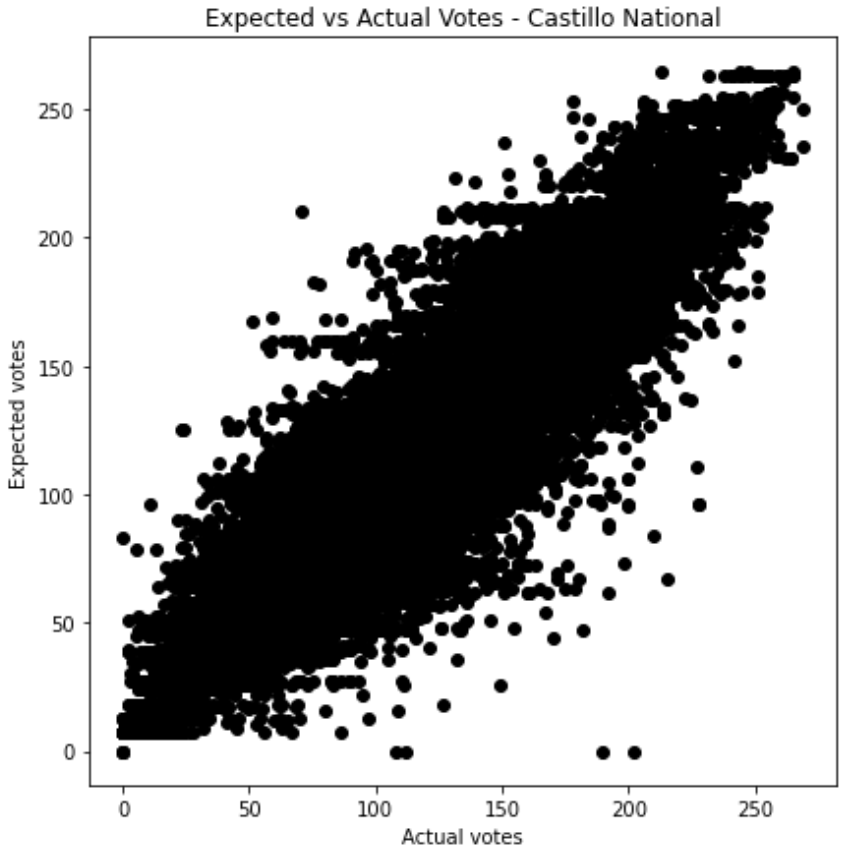
# Outliers for both candidates look similar



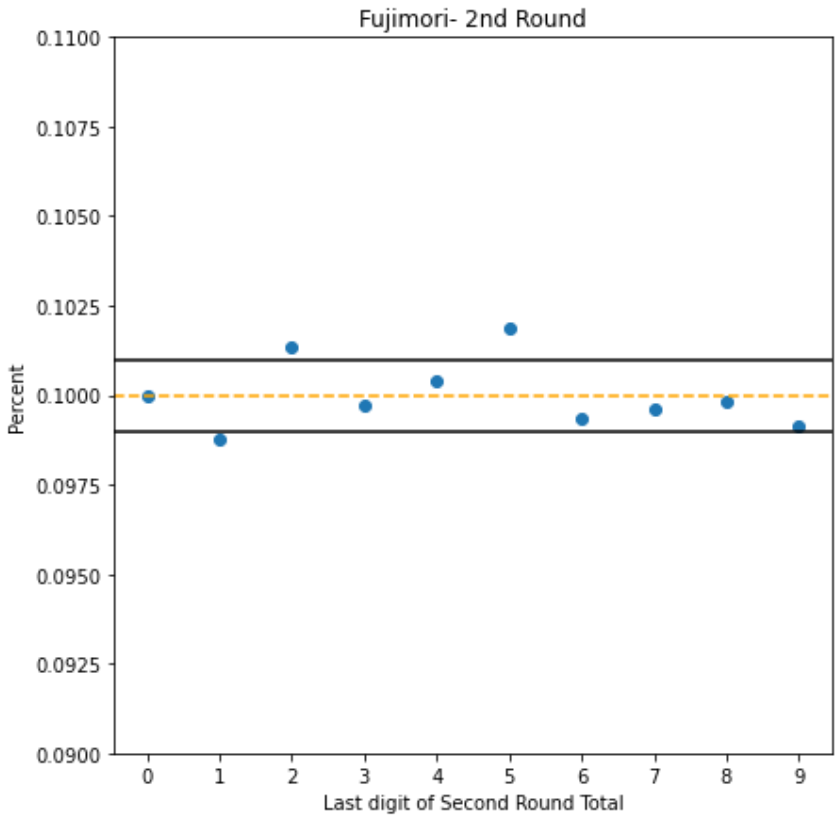



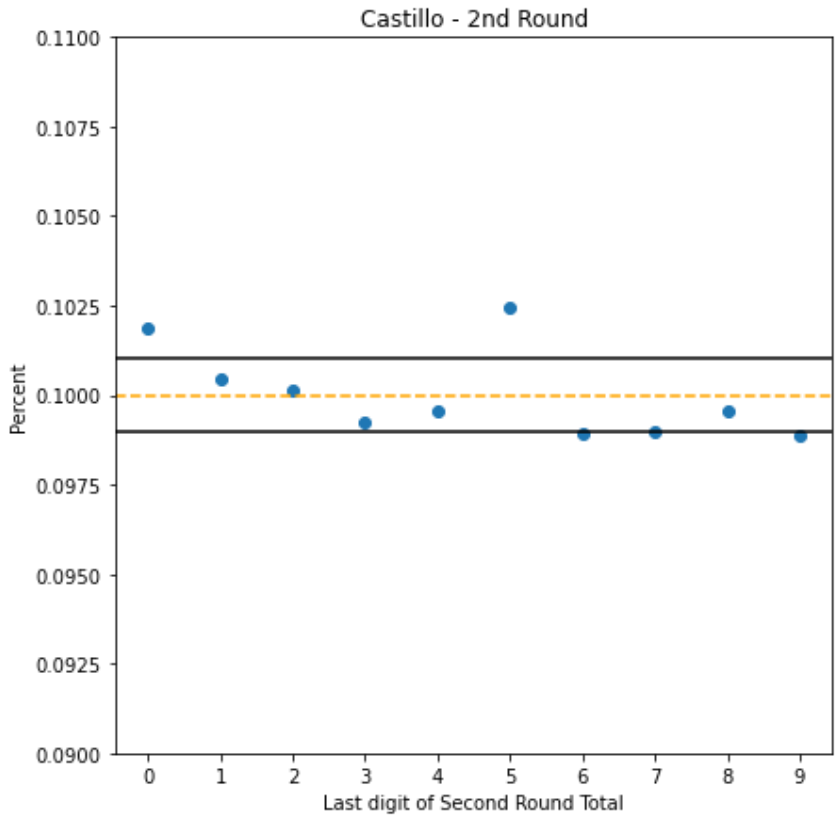
# Decision Tree Regressor has the same message, with a higher R2



# Benford’s law is an extra check. Both look similar



# Future direction

- Validate 2016 election with this methodology
- Apply a Bayesian Naïve Classifiers to this dataset
- Explore academic research on additional methods