



Political Data Science Interview Challenge

Ruidi Zhao
08/01/2021



Project Outline

Data Collection

Data Preprocessing

Exploratory Data Analysis

Modeling

Summary and Future Work



Data Collection

For the City of Los Altos:

- Voting Data (Precinct Level)
- Census Data (Block Level)
 - SEX BY AGE
 - SEX BY AGE (WHITE ONLY)



Data Preprocessing

- Join the data by transforming block level census data to precinct level census data
 - Not 100% accurate mapping >> use ratio to get estimated census data for that precinct, ratio = registered voters/Above 18 years population.
- Clean the data frame
 - Strip out the white space before column names
 - Change the data type from string to integer/float for future calculations
- Merge the census data and voting data into one data frame

| Precinct | CBG | Total | Male | Male Under 5 years | Male 5 to 9 years | Male 10 to 14 years | Male 15 to 17 years | Male 18 and 19 years | Male 20 years | ... | F to | | |
|----------|------|---|-------------|--------------------|-------------------|---------------------|---------------------|----------------------|---------------|------|------|----|--|
| 0 | 2301 | [5105.1, 5105.2, 5105.3] | 4541.000000 | 2145.000000 | 74.000000 | 193.000000 | 187.000000 | 98.000000 | 55.0 | 9.0 | ... | 15 | |
| 1 | 2305 | [5103.1, 5103.2, 5104.1, 5104.2, 5104.3] | 6956.000000 | 3494.000000 | 203.000000 | 254.000000 | 294.000000 | 201.000000 | 34.0 | 30.0 | ... | 20 | |
| 2 | 2309 | [5102.3half, 5103.3] | 1858.500000 | 932.000000 | 28.500000 | 82.000000 | 107.000000 | 79.500000 | 13.5 | 0.0 | ... | 6 | |
| 3 | 2314 | [5102.1, 5102.2, 5102.3half] | 3760.500000 | 1802.000000 | 90.500000 | 134.000000 | 149.000000 | 79.500000 | 16.5 | 52.0 | ... | 11 | |
| 4 | 2317 | [5100.02.1, 5100.02.2, 5100.02.3] | 3616.000000 | 1581.000000 | 47.000000 | 84.000000 | 144.000000 | 107.000000 | 6.0 | 0.0 | ... | 12 | |
| 5 | 2330 | [5100.01.1, 5100.01.2, 5100.01.3, 5100.01.4, 5... | 6564.000000 | 3083.000000 | 104.000000 | 156.000000 | 245.000000 | 183.000000 | 14.0 | 23.0 | ... | 21 | |
| 6 | 2338 | [5078.05.2ratio_1] | 223.041779 | 111.633423 | 16.204852 | 4.501348 | 15.754717 | 0.000000 | 0.0 | 0.0 | ... | | |
| 7 | 2351 | [5101.1, 5101.2, 5101.3] | 2906.000000 | 1493.000000 | 90.000000 | 181.000000 | 133.000000 | 47.000000 | 16.0 | 14.0 | ... | 9 | |
| 8 | 2353 | [5077.03.1ratio_2] | 64.743396 | 26.719497 | 0.757233 | 0.757233 | 4.597484 | 0.540881 | 0.0 | 0.0 | ... | | |

9 rows × 100 columns



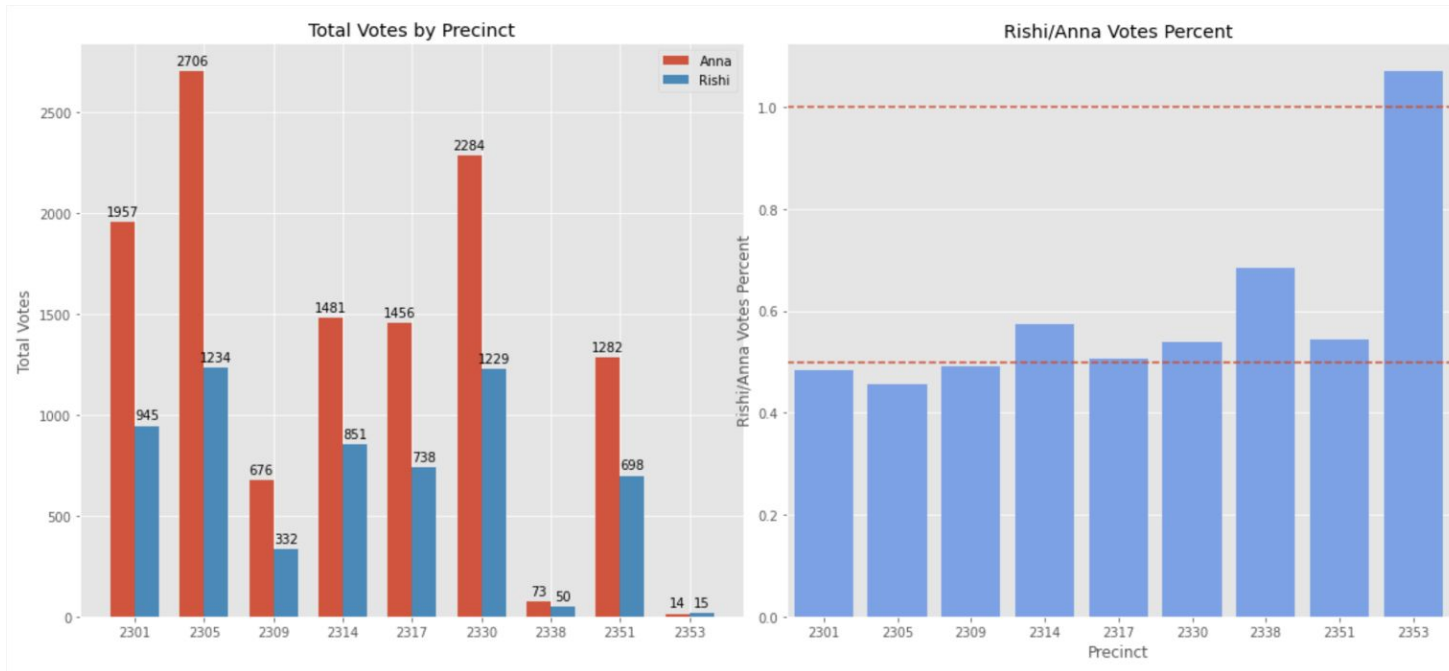
Exploratory Data Analysis

1. Voting Data
 - 1.1 Total Votes
 - 1.2 Mail Votes Percent
2. Census Data
 - 2.1 Population Number
 - 2.2 Gender
 - 2.3 Age Distribution
 - 2.4 Race (white)



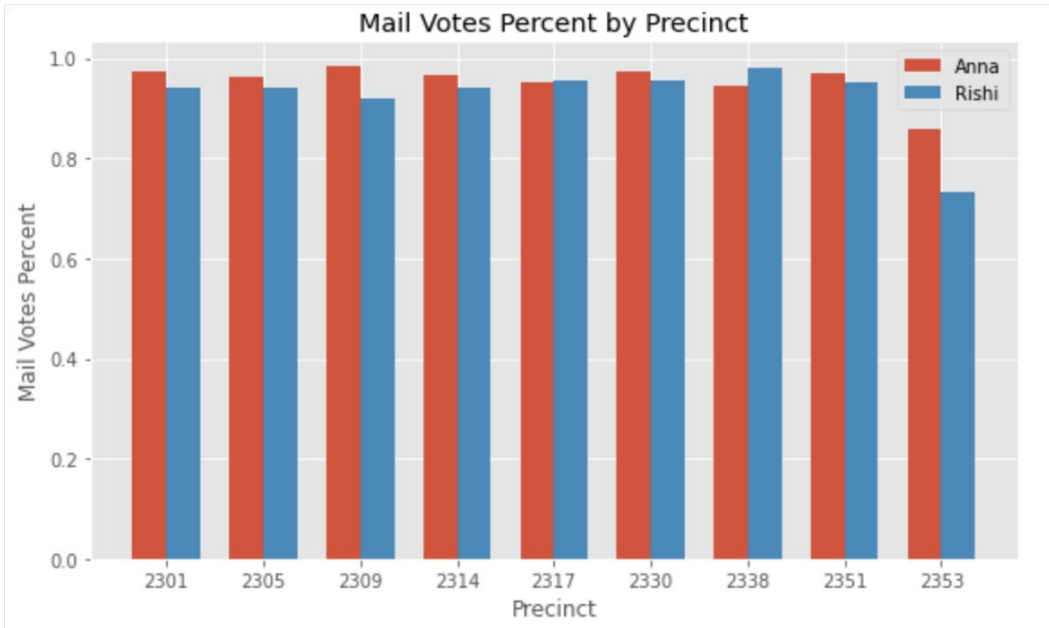
1. Voting Data

1.1 Total Votes by Precinct



- Anna won in every precinct except 2353
- 2314, 2330, 2338, 2351, has percent > 0.5

1.2 Mail Votes Percent by Precinct



- Anna's Mail Votes Percent is larger in every precinct except 2338
- Democrats are more likely to mail their votes

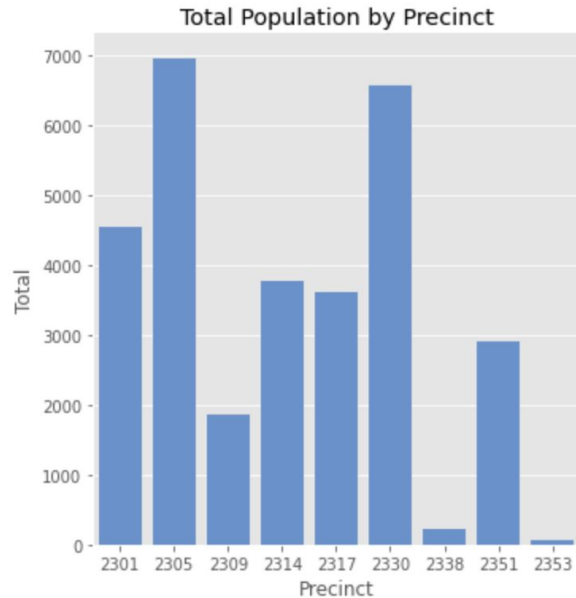


2. Census Data



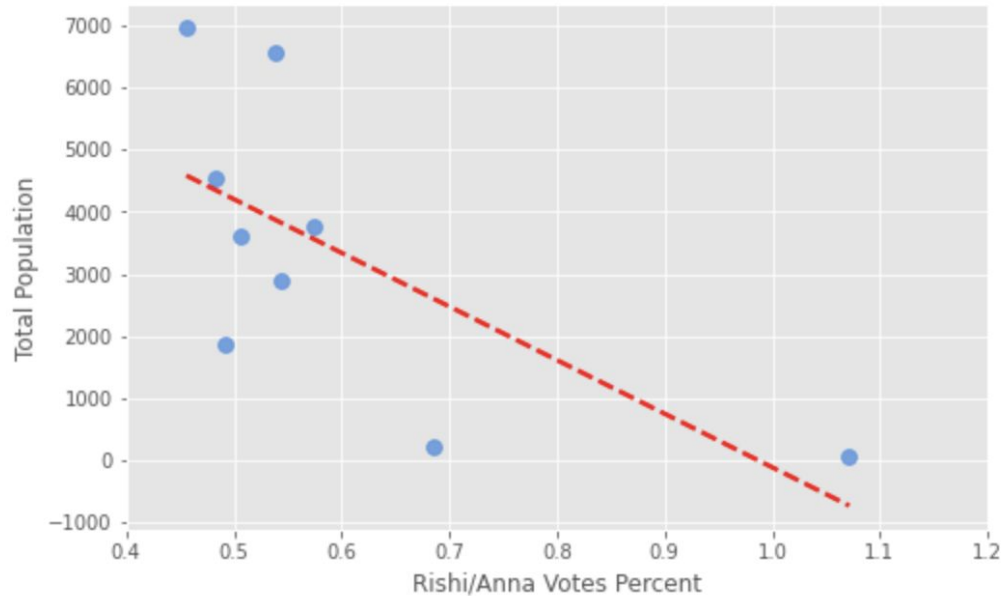
2.1 Census Data (Population Number)

Total Population by Precinct



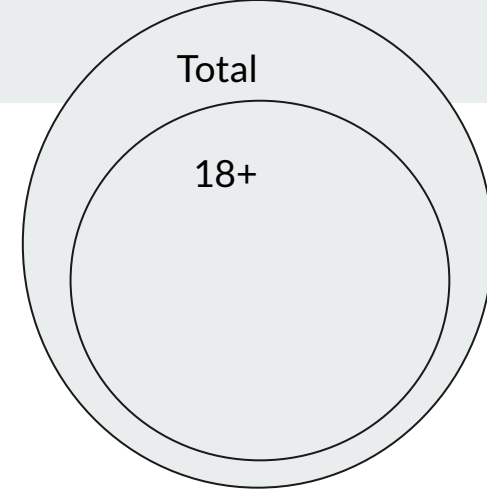
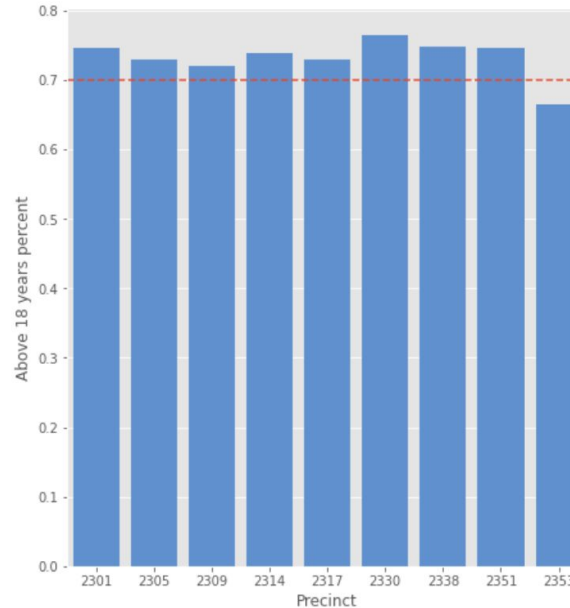
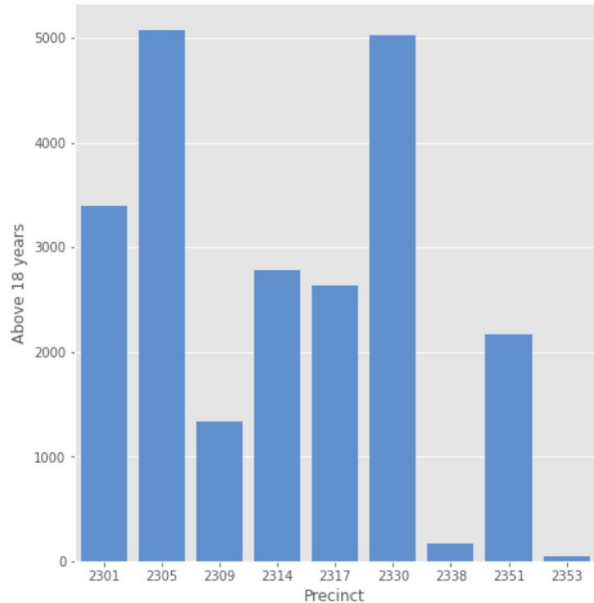
- 2305, 2330 have the largest population
- 2338, 2353 have the smallest population

Total Population and Votes Ratio



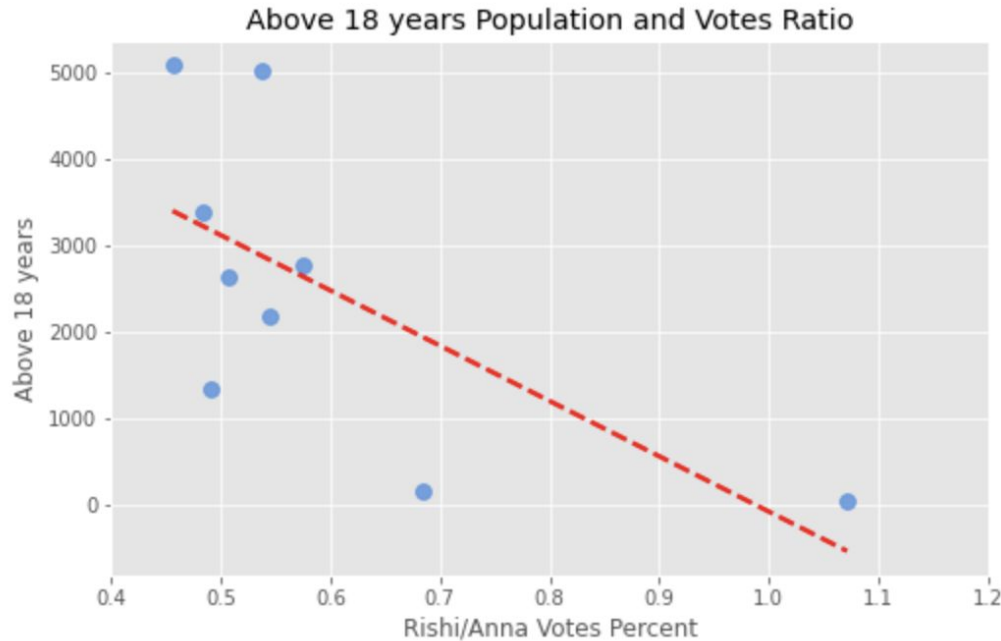
- Negative Slope

Above 18 years Population by Precinct



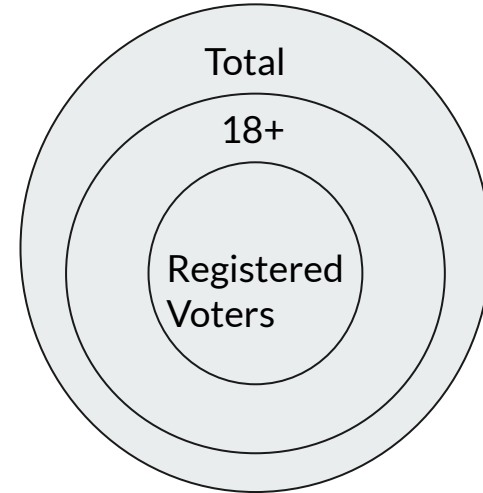
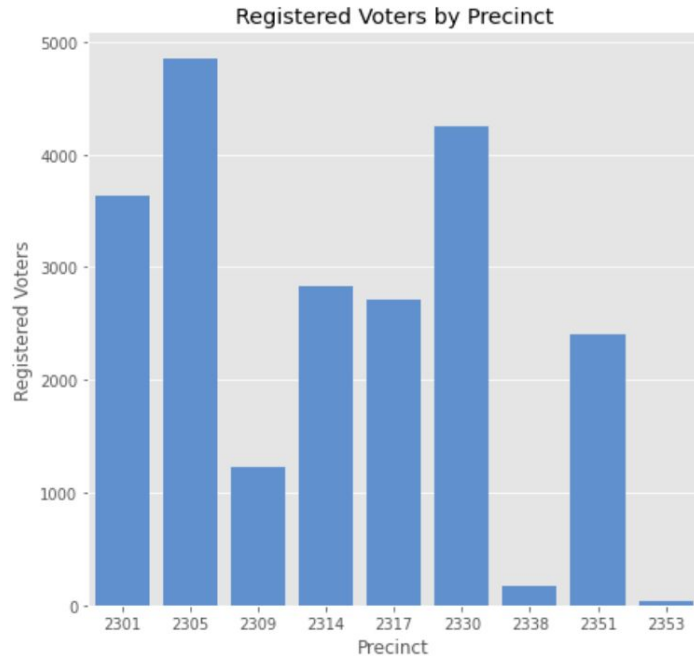
- Similar Pattern with Total Population by Precinct; the relative difference between 2305 and 2330 is smaller here because 2330 has the highest percent
- All the percents are between 0.7 and 0.8 except 2353

Above 18 years Population and Votes Ratio

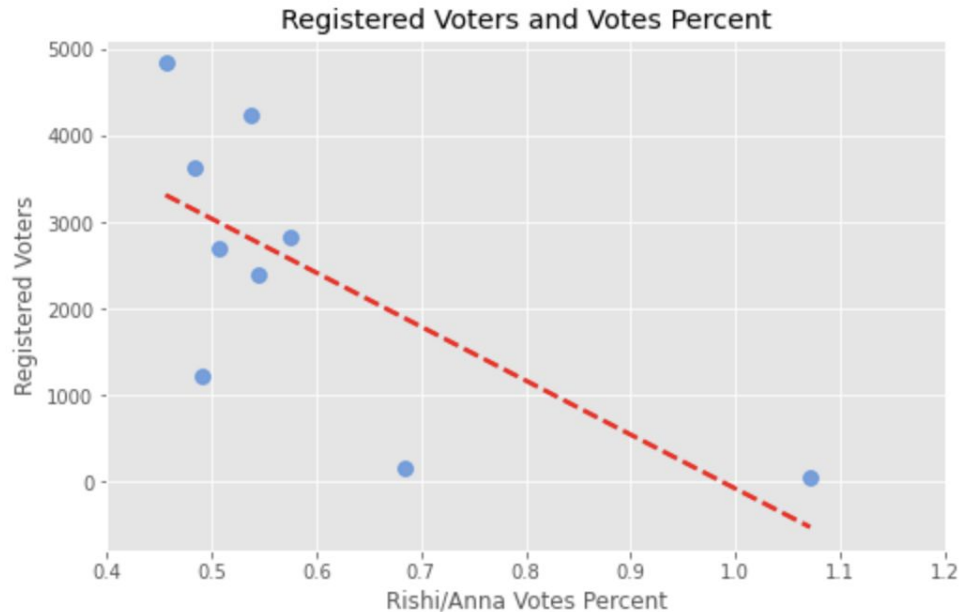


- Negative Slope
- Similar to Total Population

Registered Voters Population by Precinct



Registered Voters and Votes Ratio



- Negative Slope
- Similar Pattern Again



Total Population, 18+, and Registered Voters

| | Total | Above 18 years | Registered Voters |
|-------------------|----------|----------------|-------------------|
| Total | 1.000000 | 0.999376 | 0.989851 |
| Above 18 years | 0.999376 | 1.000000 | 0.987532 |
| Registered Voters | 0.989851 | 0.987532 | 1.000000 |

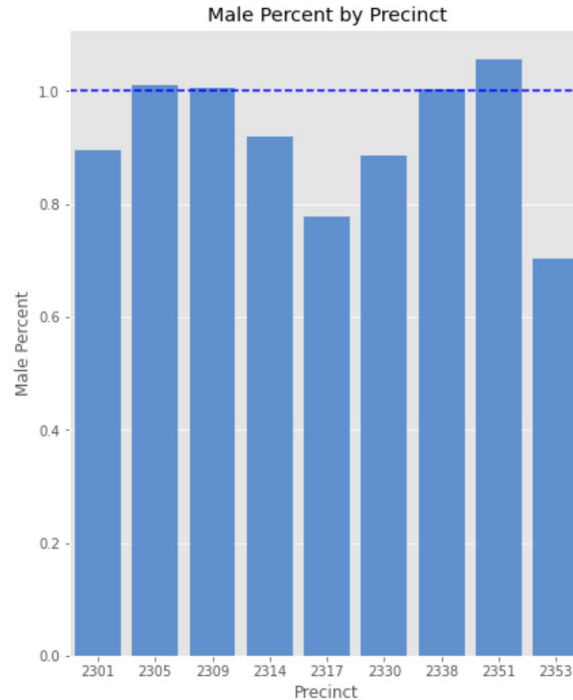
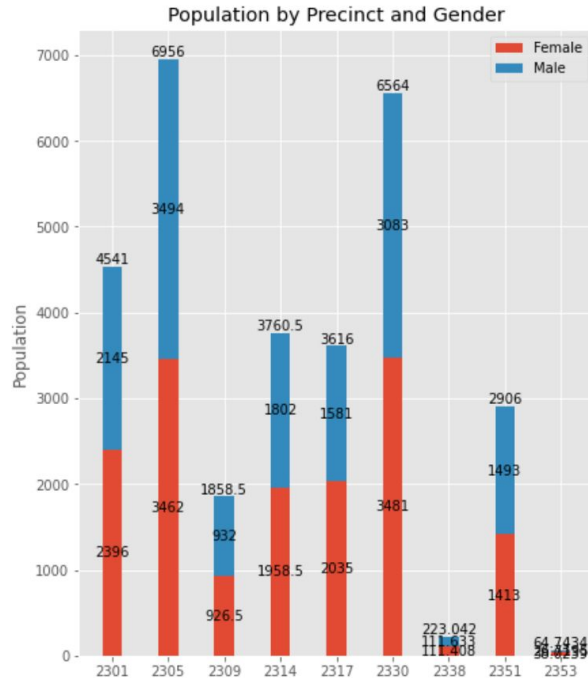
Only use Registered Voters as feature



2.2 Census Data (gender)

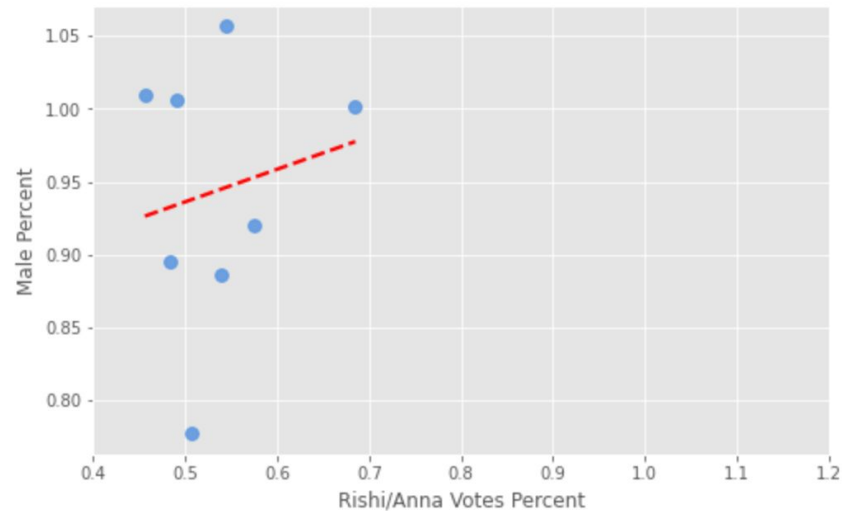
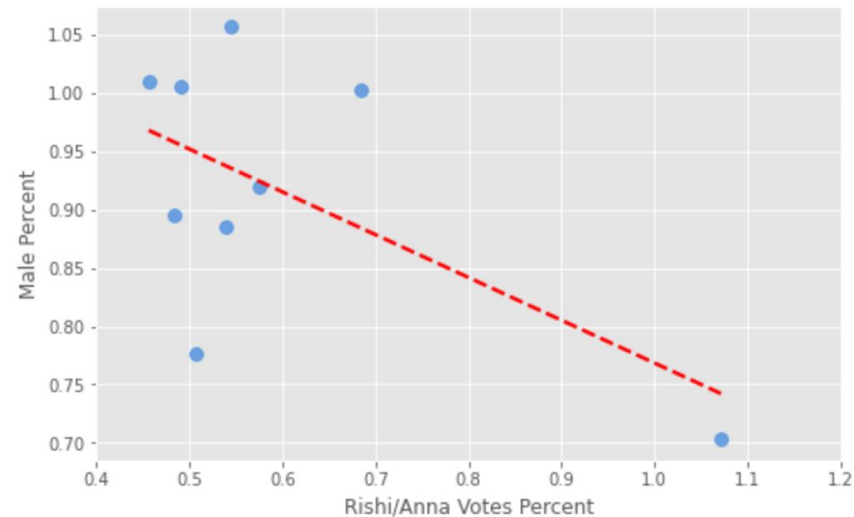
- Total Population by Gender, Male Population/Female Population Percent
- Above 18 years Population by Gender, Male 18+ Population/ Female 18+ Population Percent

Total Population by Gender

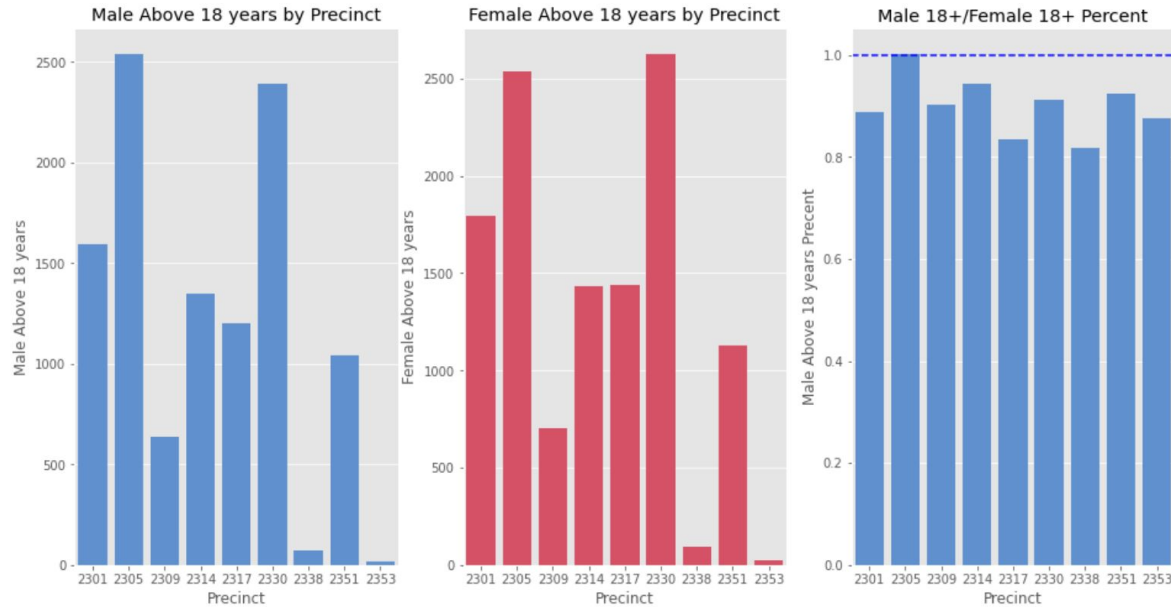


- 2301, 2314, 2317, 2330, 2353 more female
- 2305, 2309, 2338 almost balanced
- 2351 more male

Male Population Percent and Votes Ratio

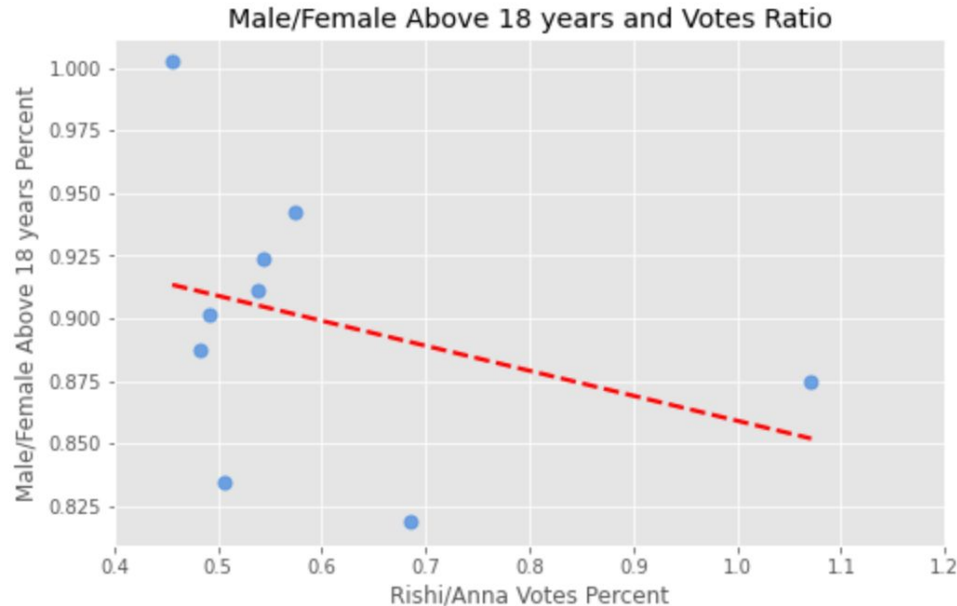


Above 18 years Population by Gender



- More female than male except 2305
- 2305 is almost balanced, with percent 1.002

Above 18 years Male/Female Percent and Votes Ratio





Male Percent and Male/Female Above 18 + Percent

| | Male Percent | Male/Female Above 18 years Present |
|------------------------------------|--------------|------------------------------------|
| Male Percent | 1.000000 | 0.394637 |
| Male/Female Above 18 years Present | 0.394637 | 1.000000 |

Both will be used as features



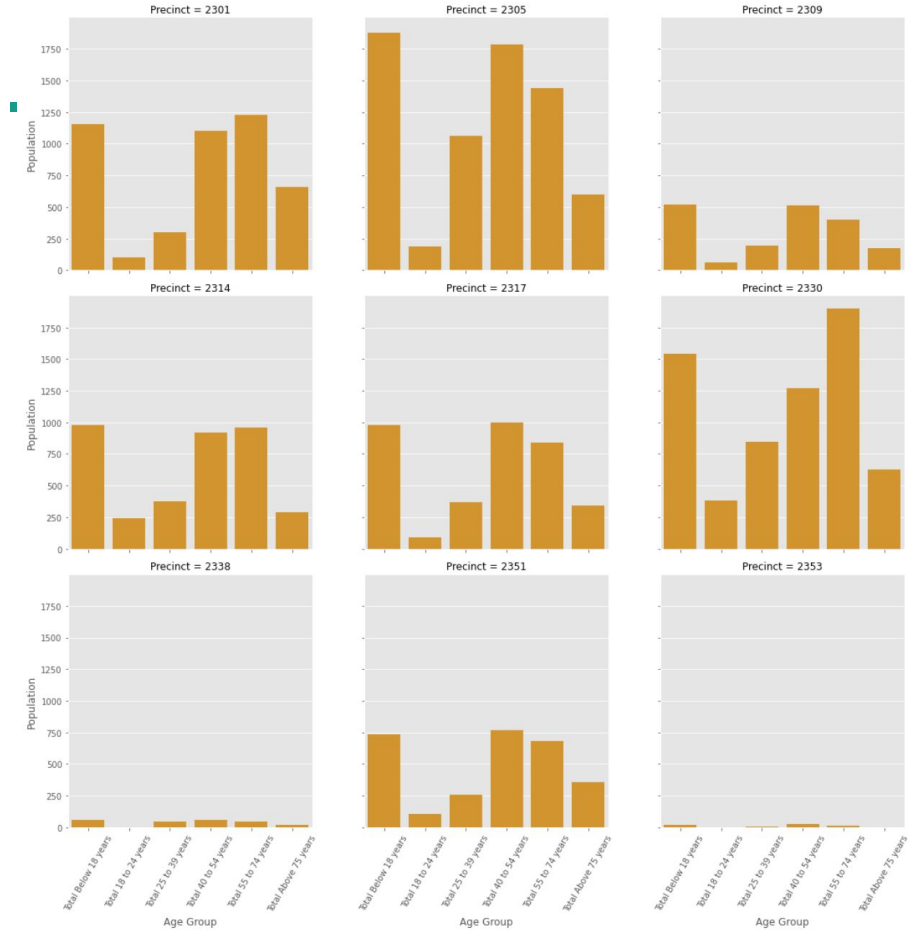
Census Data (Age Distribution)

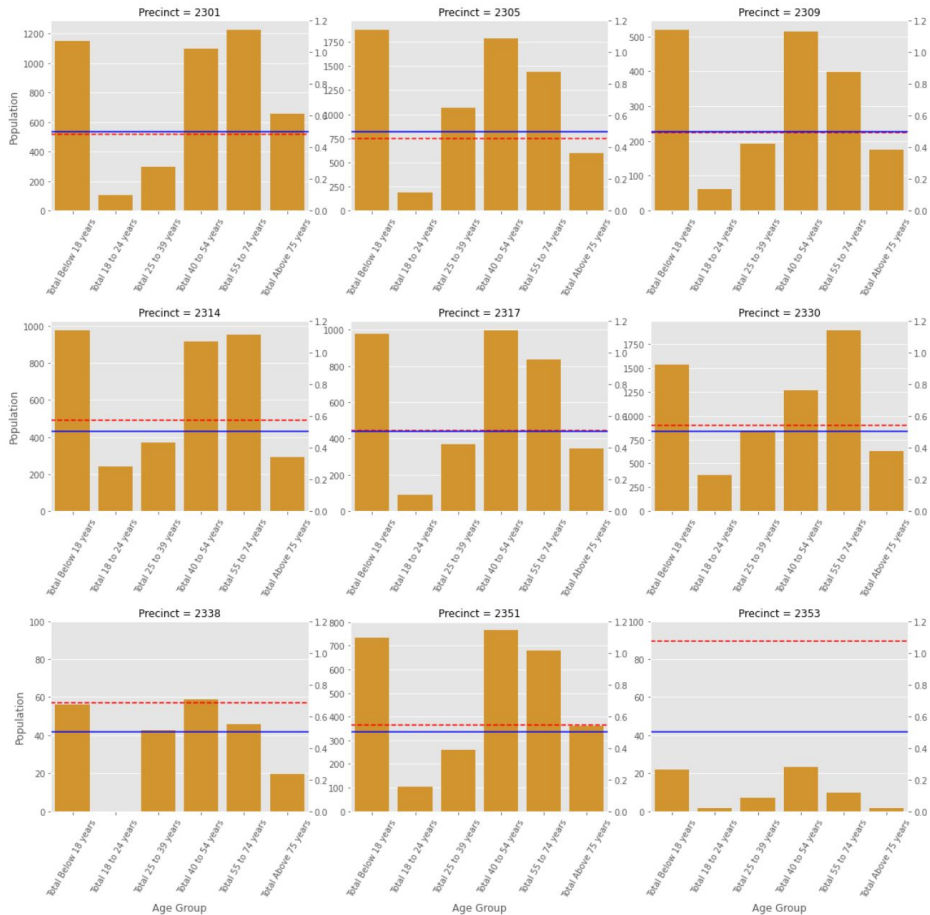
- Combine original age groups into larger groups

- Below 18
- 18 - 24
- 25 - 39
- 40 - 54
- 55 - 74
- Above 75

| | Born | Ages |
|--------------------|-------------|---------|
| Gen Z | 1997 – 2012 | 9 – 24 |
| Millennials | 1981 – 1996 | 25 – 40 |
| Gen X | 1965 – 1980 | 41 – 56 |
| Boomers II | 1955 – 1964 | 57 – 66 |
| Boomers I | 1946 – 1954 | 67 – 75 |
| Post War | 1928 – 1945 | 76 – 93 |
| WW II | 1922 – 1927 | 94 – 99 |

Total Age Distribution by Precinct





Total Age Distribution by Precinct

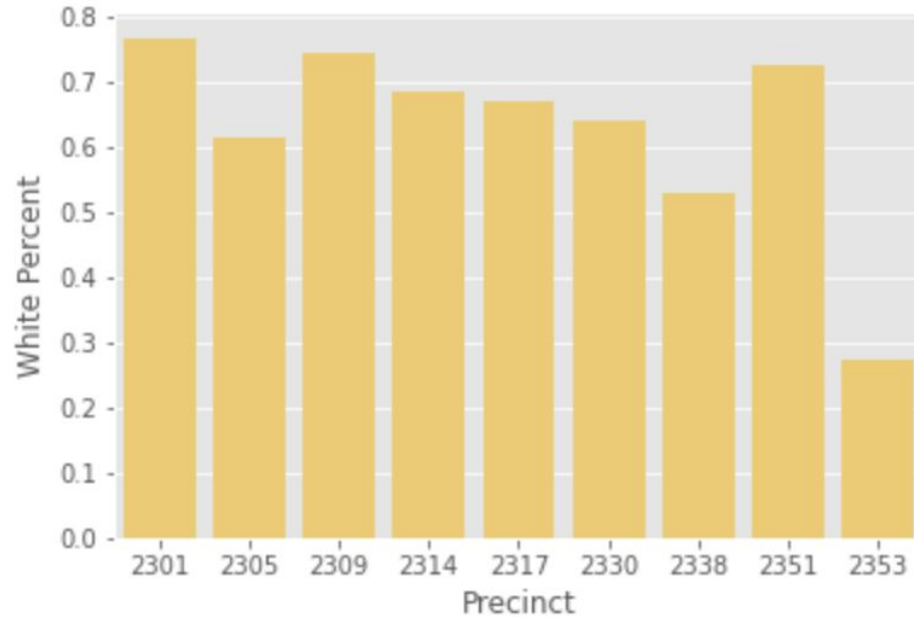
Blue Line: 0.5
Red Line: Rishi/Anna Votes Ratio

- (Blue > Red) 2305: young population
- (Red > Blue) 2314, 2330, 2351: middle age population to elderly
- (Red >> Blue) 2338 and 2353: both have small population



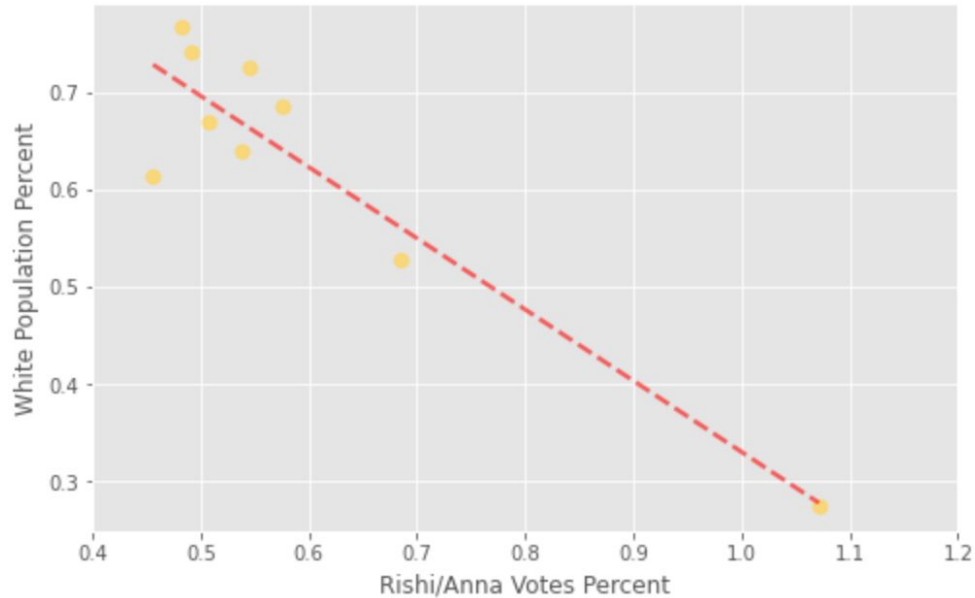
Census Date (Race)

White Population Percent by Precinct



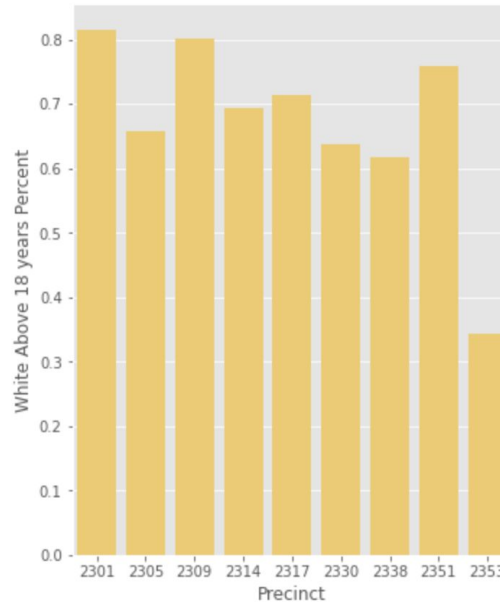
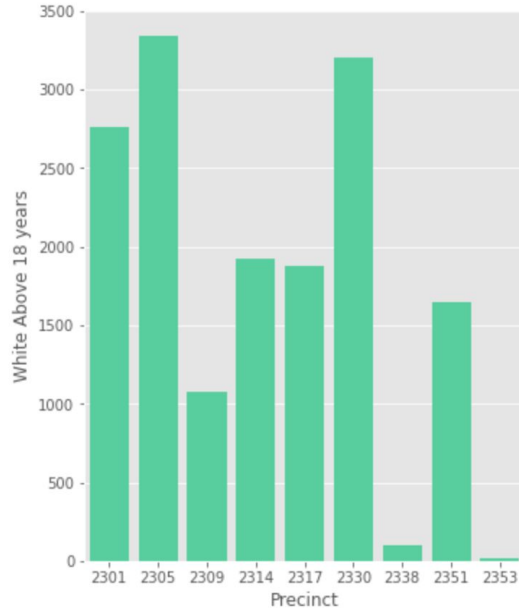
- White population is the majority

White Population Percent and Votes Ratio



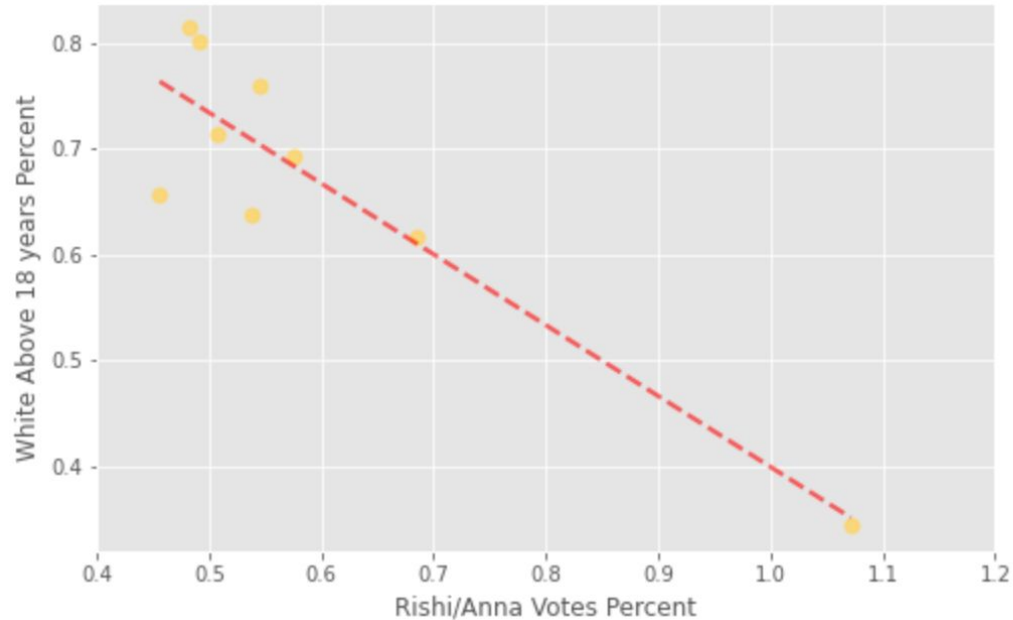
- Clear negative trend

White Above 18 years Population by Precinct



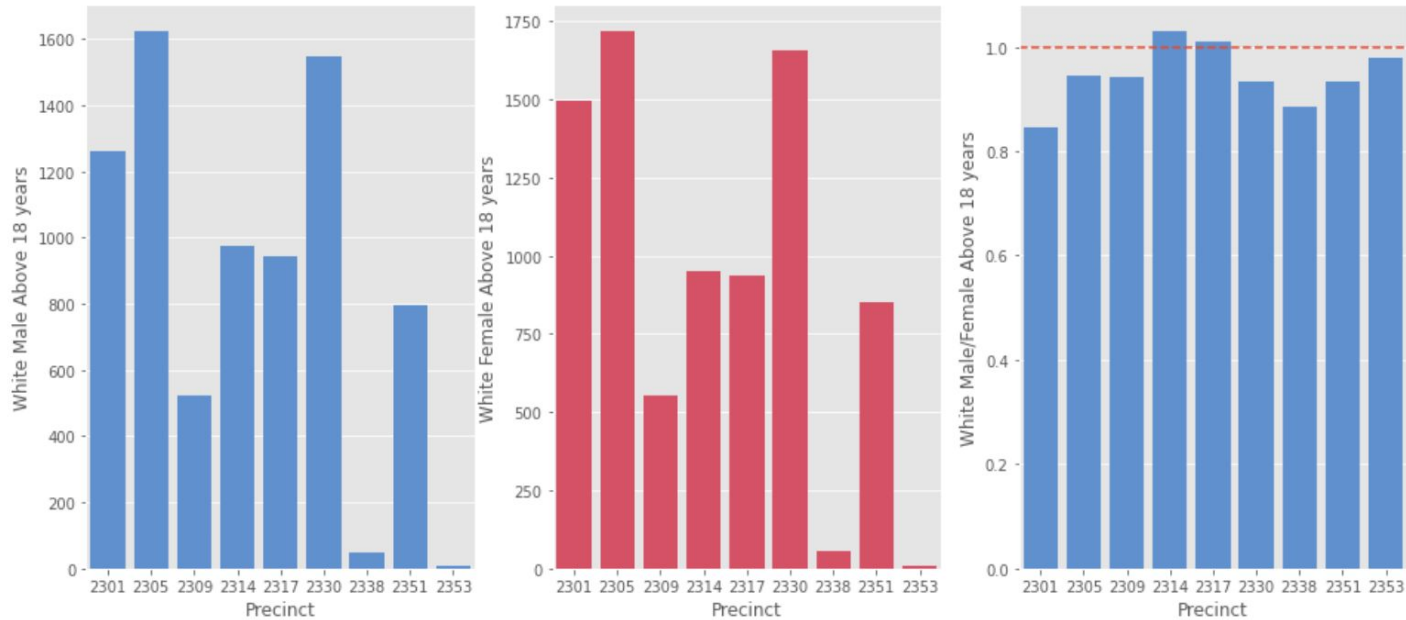
White Above 18 years
Percent = White Above 18
years/Total Above 18 years

White Above 18 years and Votes Ratio

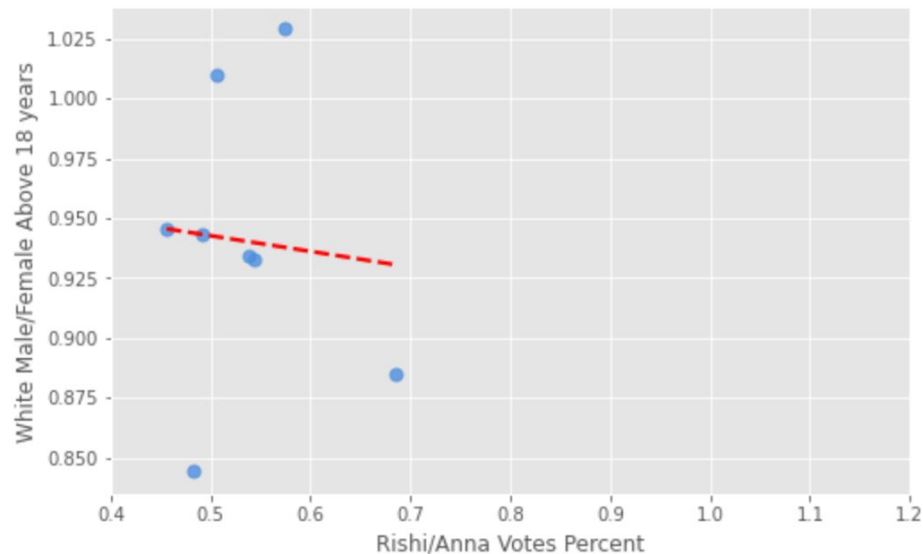
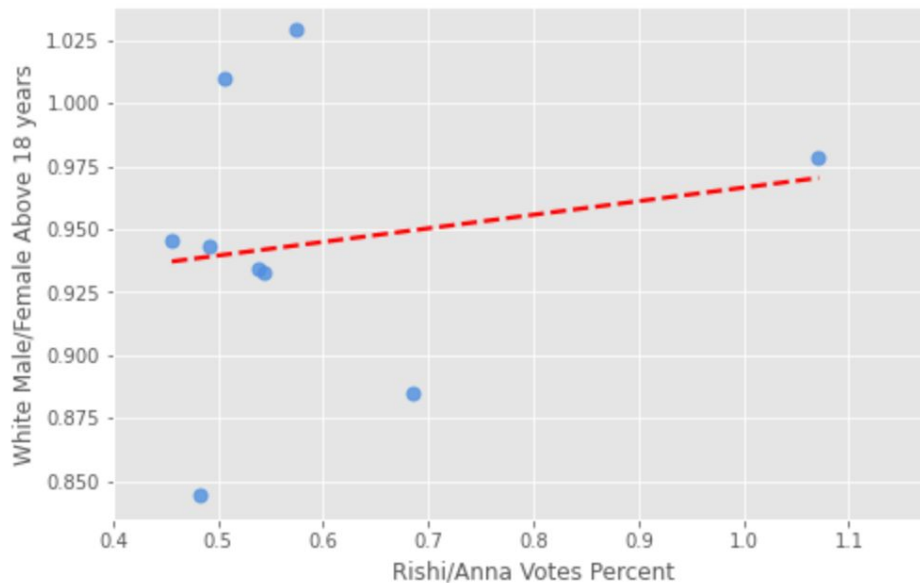


- Clear negative trend

White Above 18 years by Gender (Race & Gender Mixed Factor)



White Male/Female Above 18 years and Votes Ratio





Modeling

Features:

- Registered Voters
- Male Percent
- Male/Female Above 18 years Percent
- Age Distribution
- White Above 18 years Percent
- White Male/Female Above 18 years Percent

Target: Rishi/Anna Votes Ratio

Train/Test: Leave One Out

Metric: MSE

| | Model Name | Model Parameters | Average MSE |
|---|---------------------------|---|-------------|
| 4 | LinearRegression | {'copy_X': True, 'fit_intercept': True, 'n_job... | 580.722334 |
| 6 | Ridge | {'alpha': 1.0, 'copy_X': True, 'fit_intercept'... | 457.453342 |
| 5 | Lasso | {'alpha': 1.0, 'copy_X': True, 'fit_intercept'... | 0.223858 |
| 8 | XGBRegressor | {'objective': 'reg:squarederror', 'base_score'... | 0.202878 |
| 7 | DecisionTreeRegressor | {'ccp_alpha': 0.0, 'criterion': 'mse', 'max_de... | 0.176652 |
| 3 | GradientBoostingRegressor | {'alpha': 0.9, 'ccp_alpha': 0.0, 'criterion': ... | 0.095575 |
| 2 | ExtraTreesRegressor | {'bootstrap': False, 'ccp_alpha': 0.0, 'criter... | 0.066500 |
| 0 | RandomForestRegressor | {'bootstrap': True, 'ccp_alpha': 0.0, 'criteri... | 0.065337 |
| 1 | BaggingRegressor | {'base_estimator': None, 'bootstrap': True, 'b... | 0.027177 |



Feature importances generated by Random Forest Regressor

| | Feature | Importance |
|----|------------------------------------|------------|
| 7 | Total 40 to 54 years | 0.163732 |
| 3 | White Above 18 years Percent | 0.123969 |
| 0 | Registered Voters | 0.107240 |
| 8 | Total 55 to 74 years | 0.098009 |
| 4 | Total Below 18 years | 0.097287 |
| 9 | Total Above 75 years | 0.092717 |
| 1 | Male Percent | 0.077991 |
| 6 | Total 25 to 39 years | 0.071833 |
| 2 | Male/Female Above 18 years Percent | 0.069045 |
| 5 | Total 18 to 24 years | 0.061915 |
| 10 | White Male/Female Above 18 years | 0.036262 |



Summary and Future Work

- Anna:
 - Pros: overall significantly more votes(63.2%), represents district 18 since 2013
 - Opportunities: Young Voters, White Voters, Democrats
- Rishi:
 - Cons: overall significantly less votes(36.8%)
 - Opportunities: Middle-Age Voters, Non-White Voters, Non-Democrats
- Future Work:
 - Obtain data of the remaining cities; current dataset is too small so it is hard to find true trend and test the model
 - Better mapping of the block and precinct (e.g. , find a better estimation of registered voters/18+ population), so that the census data is more accurate
 - Gather census data of other topics, such as religions, education, and income



Thank you for listening!