Analysis of Mail-in Ballots Rejection Rate in the State of New Jersey

Khamanna Iskandarova, Aidan Hughes, and Ariel Pérez
(aka the Non-QAnons)

Background and Objectives

- Population: NJ mail-in ballots in the 2020 election:
  - First all-mail election in NJ
  - Good opportunity to study VBM as a voting method

- Identifying if mail-in-ballot rejection rates impacted by: party affiliation, race, language, income, and/or educational attainment

- These insights will provide actionable insights to Action Together New Jersey (ATNJ) to target their voter education efforts, craft & lobby for electoral reforms, and understand the potential benefits/drawbacks of widespread VBM

Data Sources and Aggregation

- Ideally, we would have demographic information for each voter/ballot cast, allowing us to build models at the individual ballot level

- However, despite highly personal contact information being available for each voter, there is no demographic data connected to each voter

- As a result, analyzing demographic trends in voting behavior (without putting a poll in the field) requires the aggregation of voting data into regional statistics that can be connected to population data for those regions

Data Sources and Aggregation

2019 American Community Survey Estimates (https://data.census.gov/)

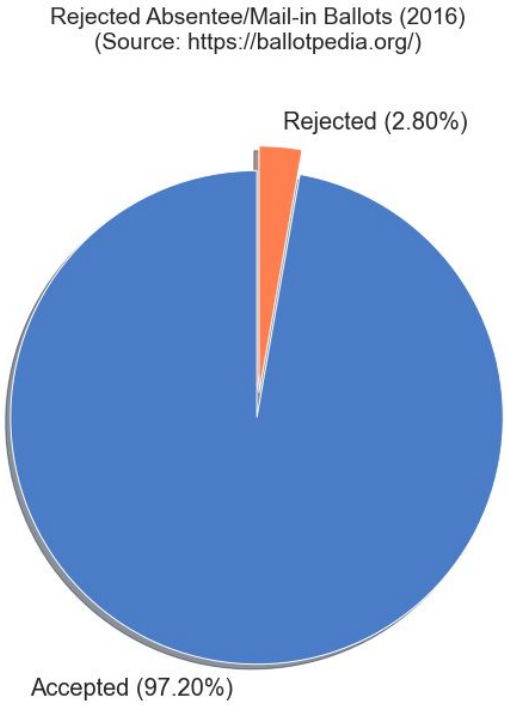
- Income features by county (median income, % of pop. in income brackets, % of population below the poverty line)
- Language (Language spoken at home, % of pop. speaks English “very well”)
- Race
- Educational Attainment
- Age Brackets

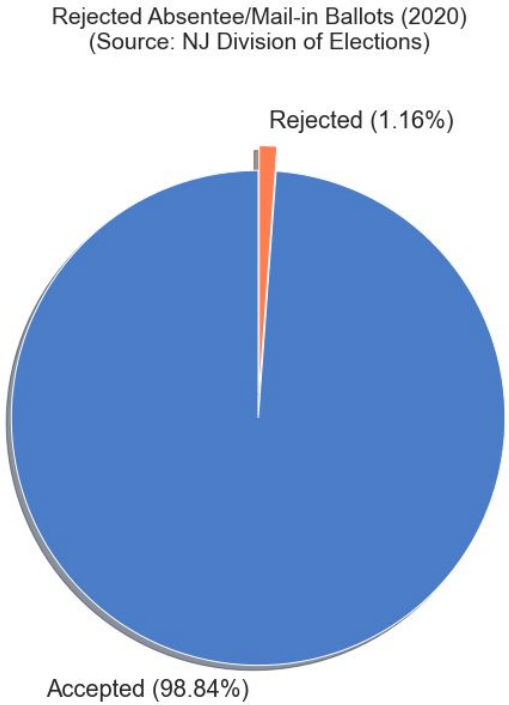
Dataset 2: Mail-in Ballots Information (NJ Division of Elections)

- Ballot / voter information
- Raw dataset = 6047966 rows (voters), 44 columns
- Obtained dataset from an Open Public Records Act (OPRA) request to the NJ Division of Elections. The dataset/report is dated Dec 13 2020

	County	ballots_cast	ballots_rejected	percent_rejected	Percent_Ballots_Cast_Dem	Percent_Ballots_Cast_Repub	Percent_Ballots_Cast_Unaff	Percent_Ba
0	Atlantic	135484	910	0.671666	38.171297	31.508518	29.313424	
1	Bergen	476733	5986	1.255629	40.954790	23.590563	34.724049	
2	Burlington	253339	1911	0.754325	42.231555	27.698459	29.278556	
3	Camden	246440	703	0.285262	52.869258	17.708570	28.477926	
4	Cape May	56565	367	0.648811	26.306020	45.342526	27.642535	
5	Cumberland	60047	761	1.267341	39.633953	27.545090	31.488667	
6	Essex	307537	2184	0.710158	57.282213	11.854834	30.258148	
7	Gloucester	170985	100	0.058485	42.053981	26.179489	30.837793	
8	Hudson	224243	5605	2.499521	61.905165	10.421284	26.567607	
9	Hunterdon	84247	618	0.733557	28.574311	41.551628	29.239023	
10	Mercer	169955	1314	0.773146	47.604954	16.915654	34.468536	
11	Middlesex	355993	6837	1.920543	46.086580	17.443882	35.498170	
12	Monmouth	377148	5403	1.432594	30.591439	30.918101	37.574374	
13	Morris	293123	1122	0.382774	30.459909	36.895433	31.884908	
14	Ocean	338676	3727	1.100462	24.230533	39.618101	35.125016	
15	Passaic	210776	1797	0.852564	42.123392	23.975690	32.785042	
16	Salem	33022	572	1.732179	33.671492	32.015020	33.280843	
17	Somerset	183143	2369	1.293525	37.081406	27.613941	34.557695	
18	Sussex	87304	992	1.136260	24.167278	45.194951	29.428205	
19	Union	243523	6284	2.580454	51.267437	17.382342	30.497735	
20	Warren	60053	970	1.615240	27.092735	42.783874	29.157577	

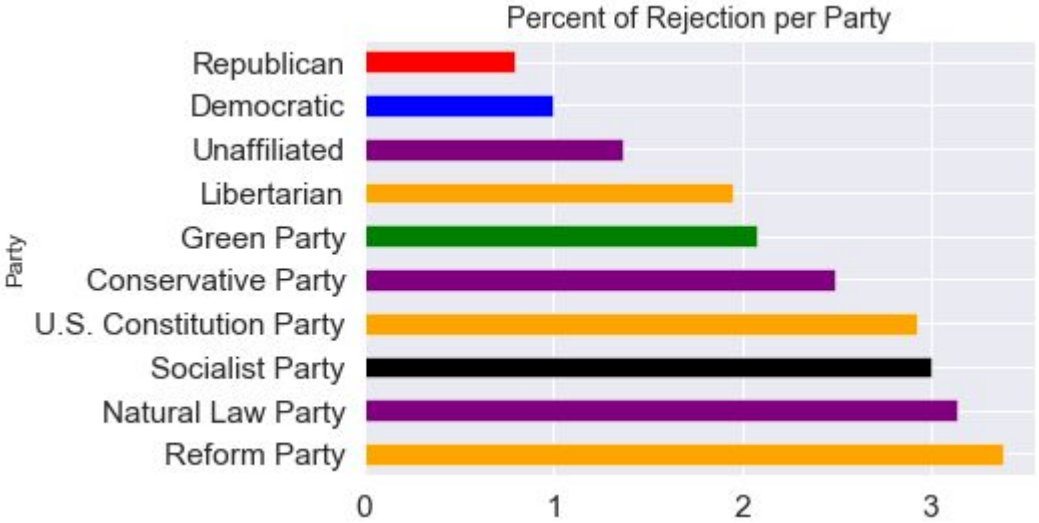
Rejected Absentee/Mail-in Ballots

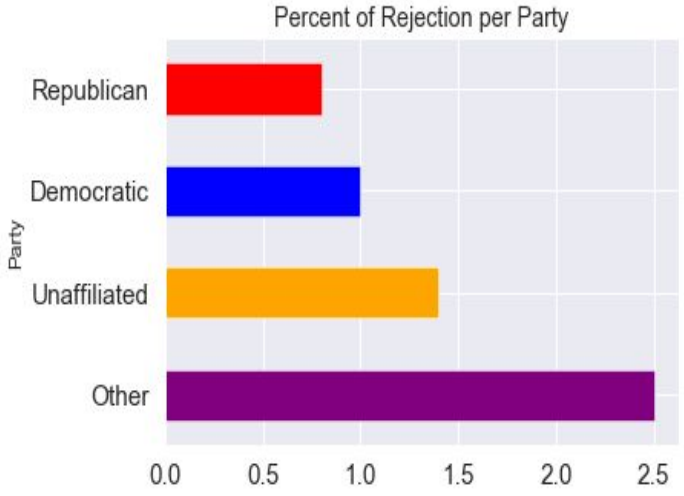




* Wines, Michael. "November Surprise: Fewer Ballots Rejected by Election Officials." *The New York Times*, 2 November 2020, https://www.nytimes.com/2020/11/02/us/election-ballots-rejections.html

Percentage Rejected per Party





Statistical Test - Chi-Square Analysis

- *We conducted a statistical test and were able to show that the ballot rejection rates between parties are statistical significant.*

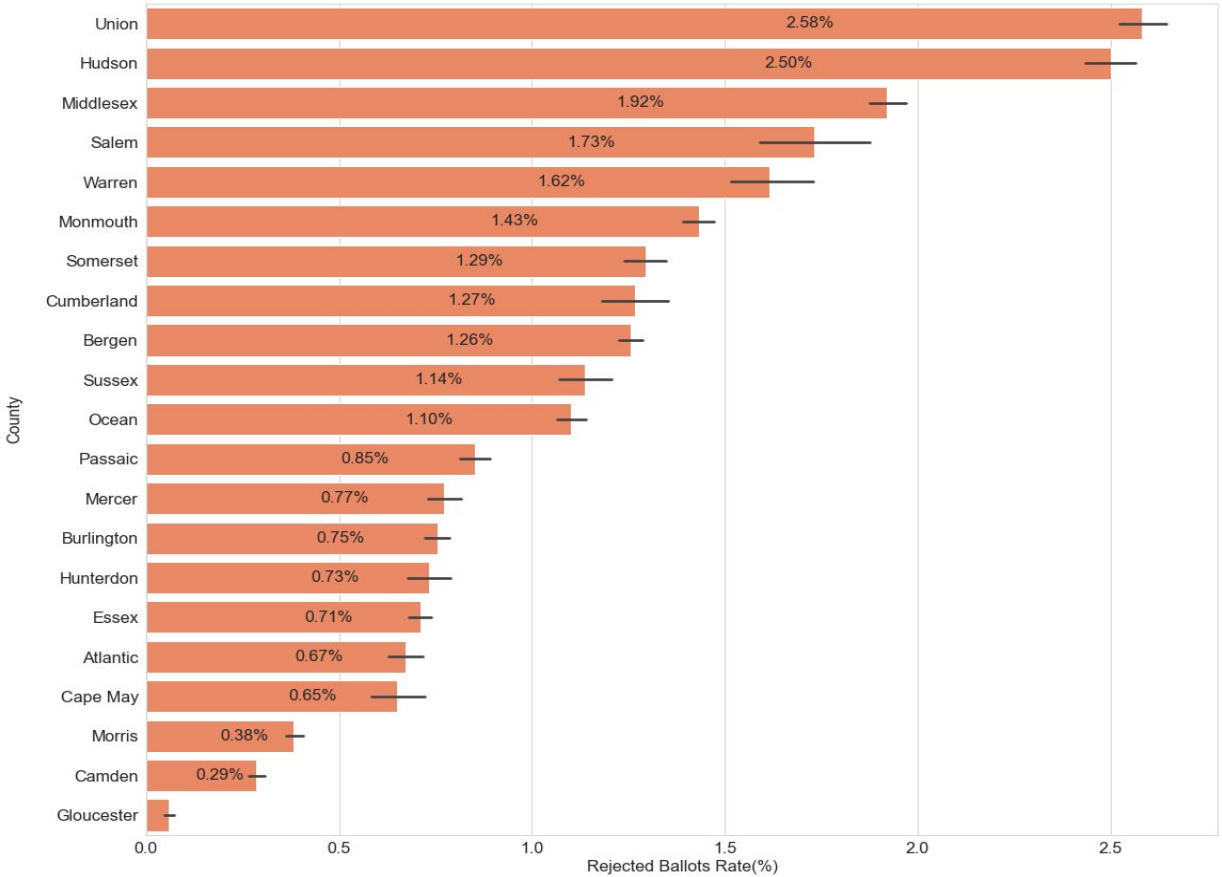
ballot_vtr_party	Conservative Party	Democratic	Green Party	Libertarian	Natural Law Party	Reform Party	Republican	Socialist Party	U.S. Constitution Party	Unaffiliated
ballot_status										
Accepted	8142	1777234	5594	10199	2743	825	1109512	3032	7364	1393159
Rejected	224	18376	133	224	95	32	10178	103	240	20927

P-value < 0.05
Party affiliation does affect the ballot rejection rate - the ballots rejection rate for each party are statistically significant.

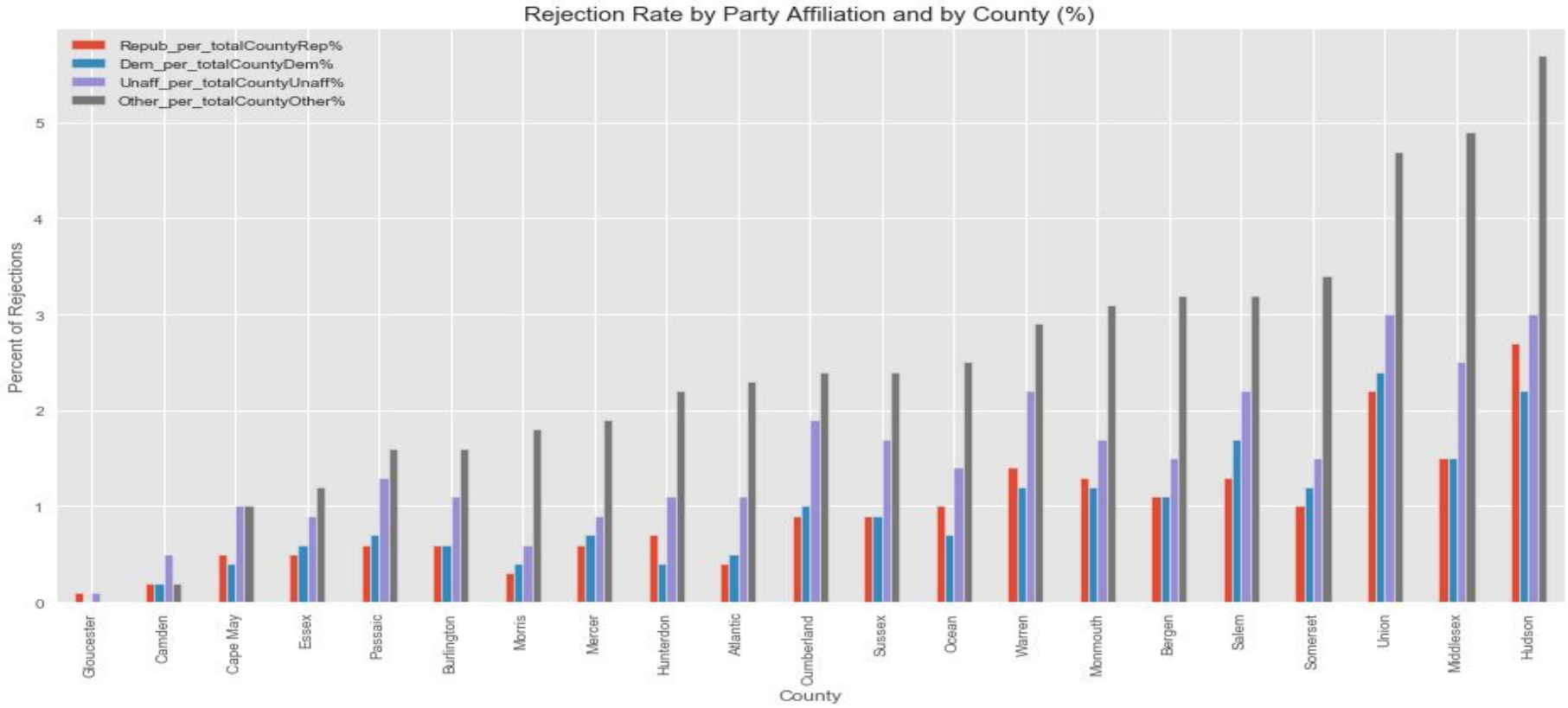
- *We also compared the Republican and Democratic parties rejection rates, and even so they seem to be close in the graph, the statistical test showed that the rejection rates are statistical significant.*

P-value = 5.140129872709784e-22 < 0.05
We reject Null Hypothesis - the variables are statistically significant

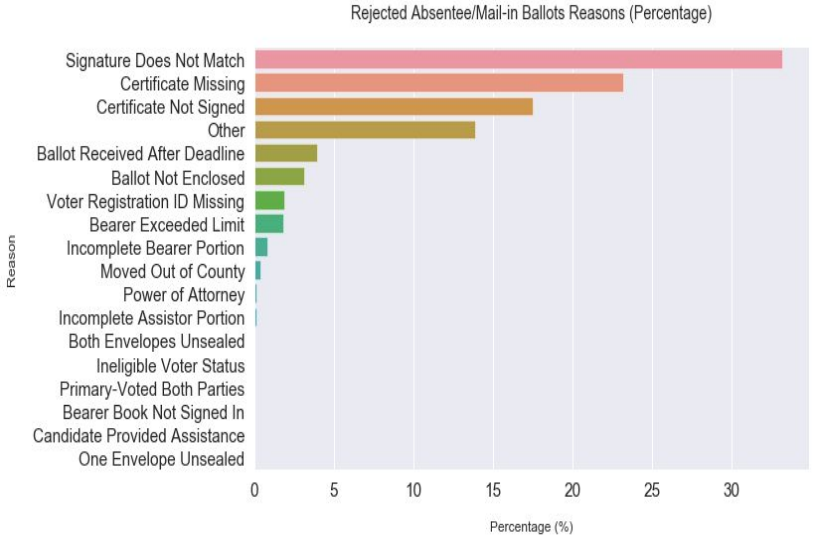
Rejected Absentee/Mail-in Ballots by County

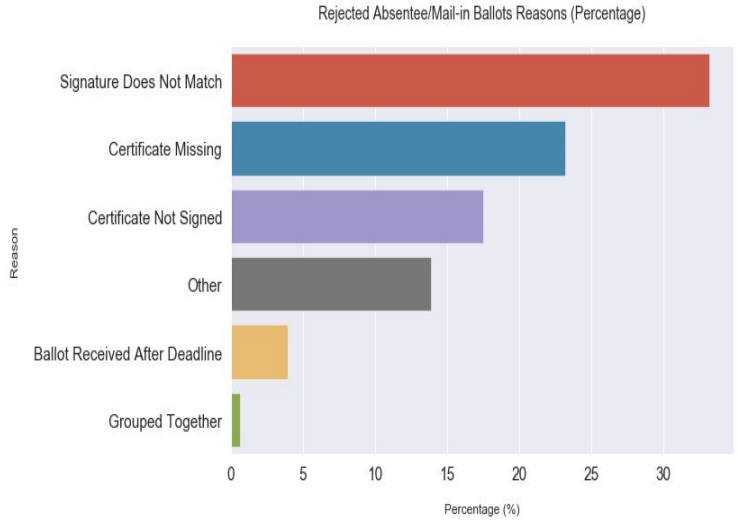


Rejection Rate by Party Affiliation and County

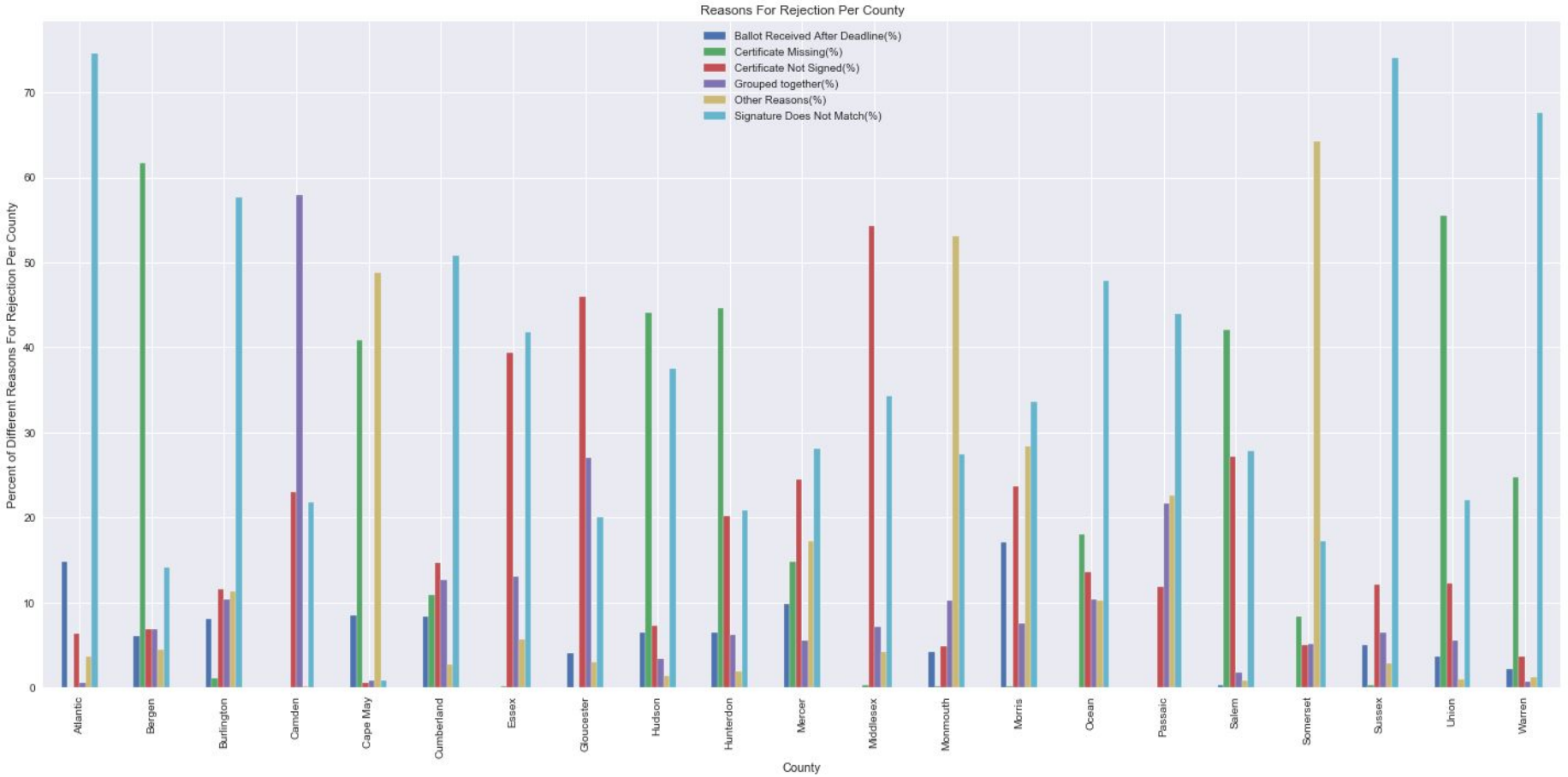


Percentage of Rejected Reasons by State





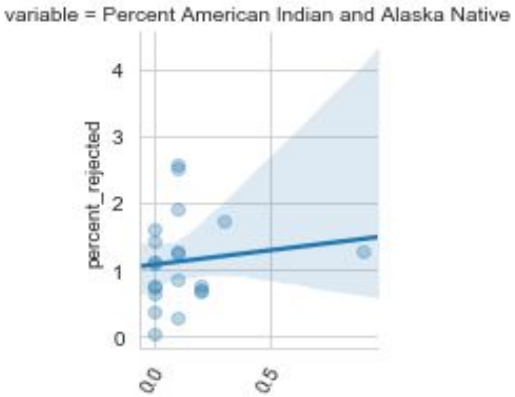
Reasons for Rejection Per County (percentage)

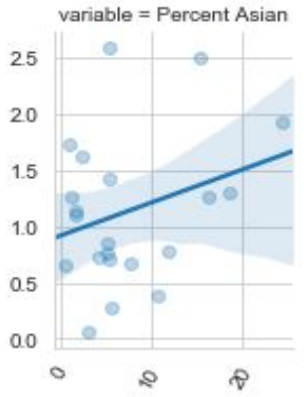


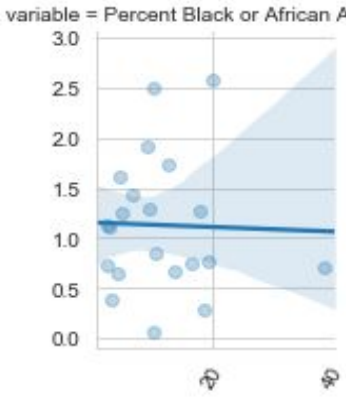
Correlation of Demographic Features

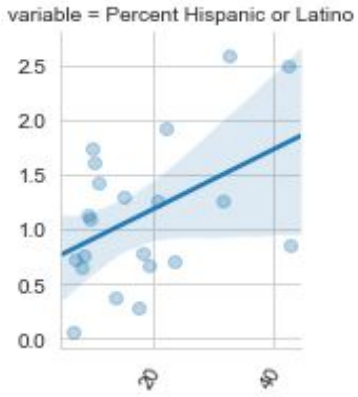
- We found that there was meaningful correlation between ballot rejection rates and certain individual features related to race....
  - Percent “Other” Race: 0.457867
  - Percent Hispanic or Latino: 0.455597
  - Percent Asian: 0.288592
  - Percent White: -0.371413

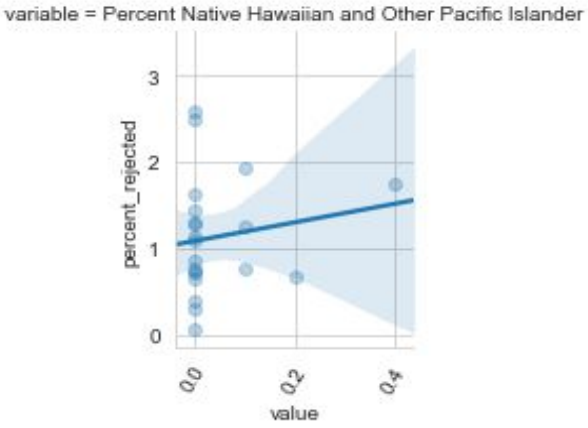
Rejected Absentee/Mail-in Ballots vs Race

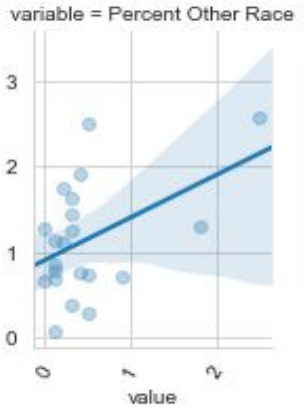


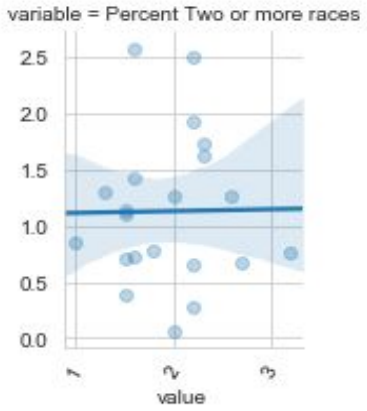


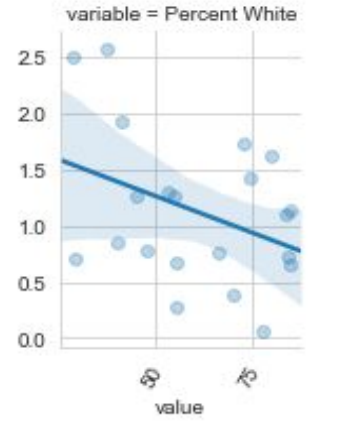








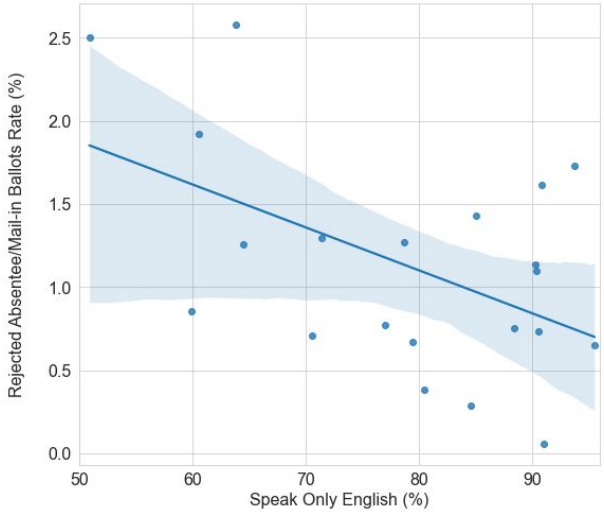


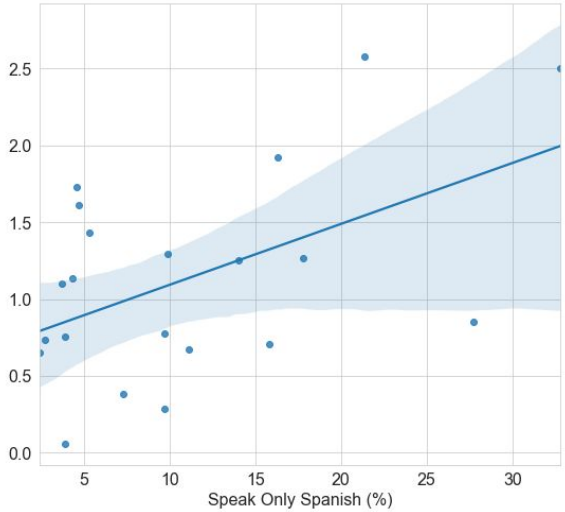


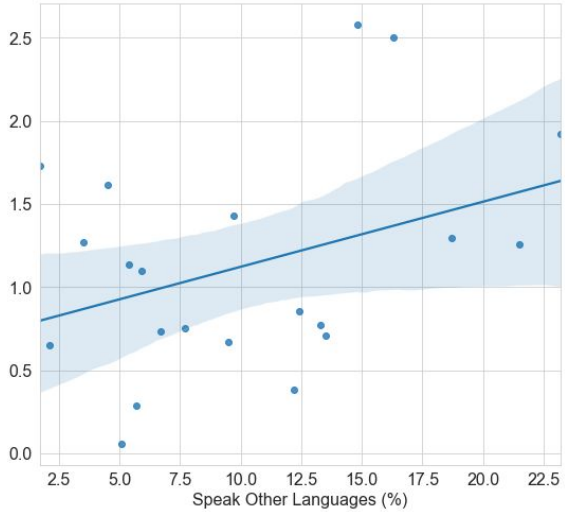
Rejected Absentee/Mail-in Ballots vs Language Spoken at Home

- And language....
  - Language Spoken at Home = Spanish: 0.504456
  - Percent of Population Speaks English Well / Not Very Well: +/- 0.465768
  - Language Spoken at Home = “Other”: 0.366002
  - Language Spoken at Home = English: - 0.503830

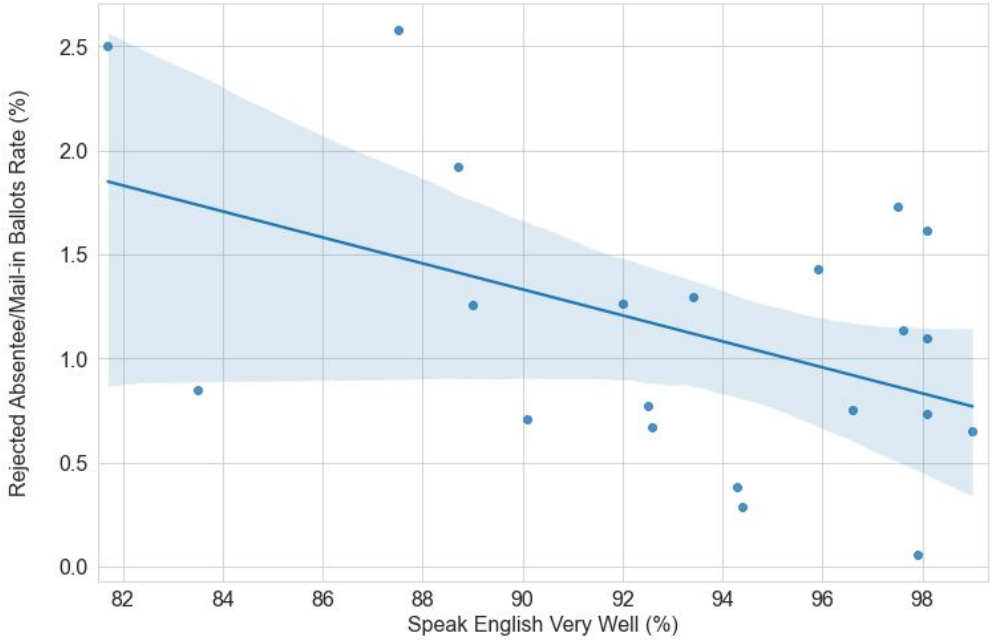
Rejected Absentee/Mail-in Ballots vs Language Spoken at Home





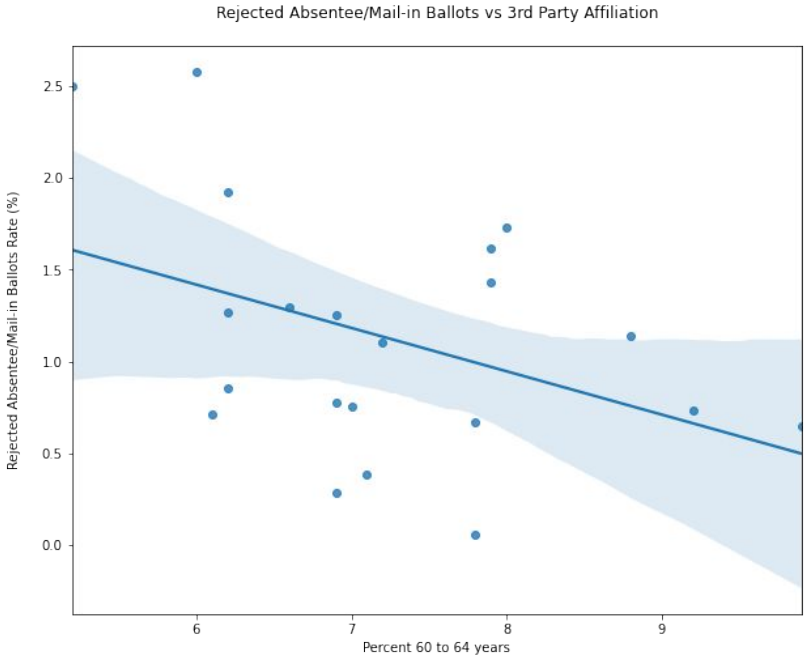


Rejected Absentee/Mail-in Ballots vs English Speaking Population



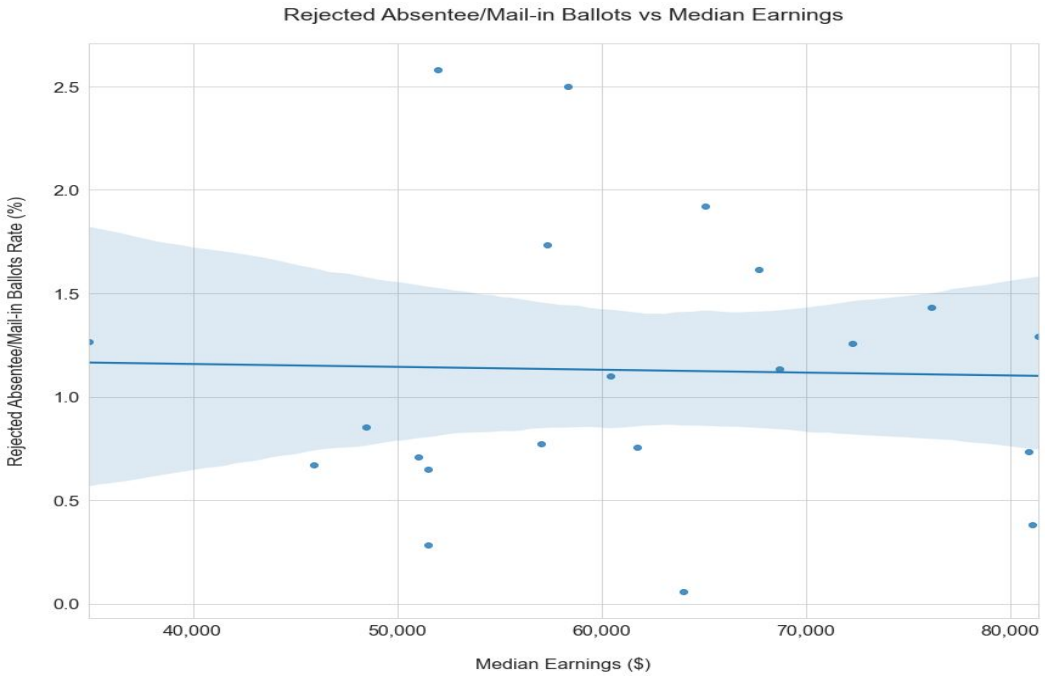
Rejected Absentee/Mail-in Ballots vs English Speaking Population

- And meaningful correlation with a county's age breakdown
  - Percent 35 to 44: 0.433374
  - Percent 25 to 34: 0.406391
  - Percent 60 to 64: -0.407653



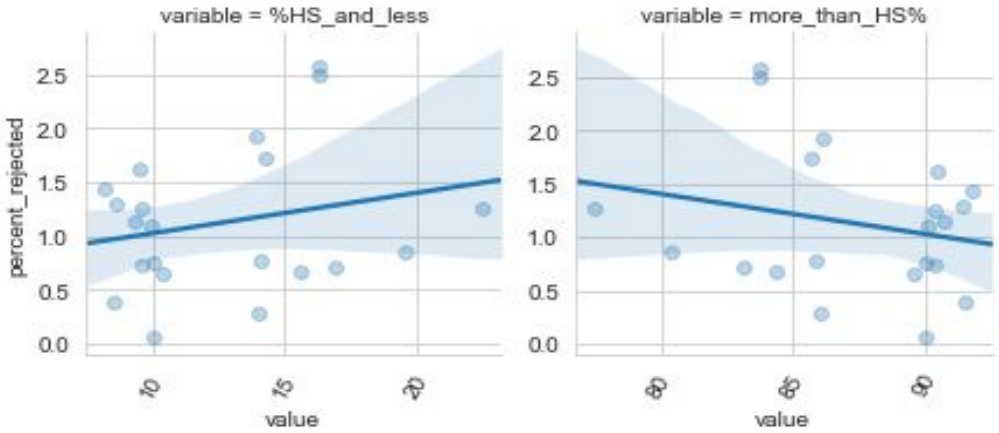
Rejected Absentee/Mail-in Ballots vs Median Earnings

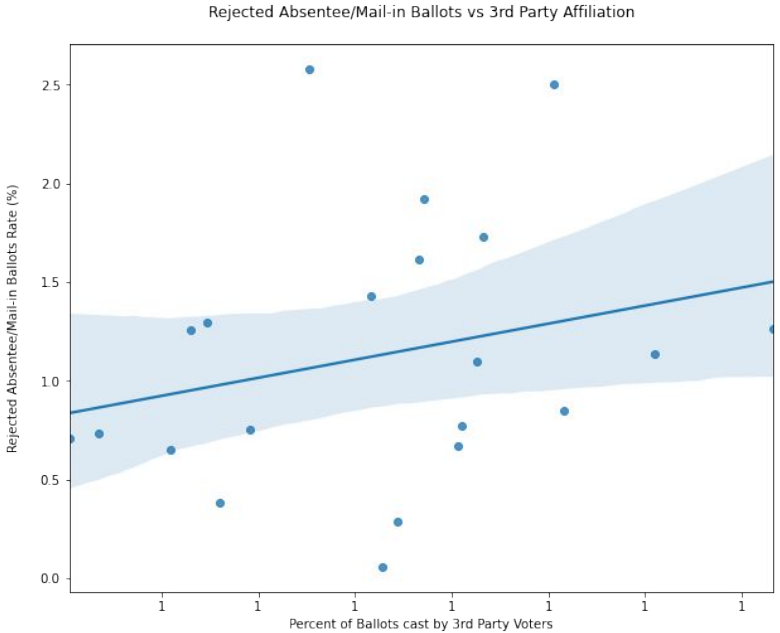
But limited correlation associated with income-related features, despite large amounts of variation across counties…



Rejected Absentee/Mail-in Ballots vs Education

… And somewhat weak correlations associated with educational attainment and affiliation with 3rd Parties





Race + Language + Party-based Regression

- A hybrid of three types of demographic/voter features provided the best performance:
  - Percent of Ballots cast by 3rd party voters
  - Percent of pop. Black or African American
  - Percent of pop. Native Hawaiian and Other Pacific Islander
  - Percent of pop. Other Race
  - Percent of households speaking Spanish
- Explained ~62% of variation in mail-in-ballot rejection rates across NJ's counties
- Some features with high correlation couldn't be included in the model (especially age-based features), due to high multicollinearity.

	variables	VIF
0	Percent_Ballots_Cast_Other	4.010286
1	Percent Black or African American	3.347087
2	Percent Native Hawaiian and Other Pacific Isla...	1.304543
3	Percent Other Race	1.862095
4	percent_spanish	3.732140

```
RSS: 3.39
R^2: 0.61945
```

```
Percent_Ballots_Cast_Other                          1.026140
Percent Black or African American                  -0.027437
Percent Native Hawaiian and Other Pacific Islander  2.036379
Percent Other Race                                  0.681579
percent_spanish                                     0.031350
```

Feature-specific models

- A model exclusively containing racial features still achieved a relatively strong R^2 score (~0.54)

```
['Percent Hispanic or Latino',
 'Percent Black or African American',
 'Percent American Indian and Alaska Native',
 'Percent Asian',
 'Percent Native Hawaiian and Other Pacific Islander',
 'Percent Other Race',
 'Percent Two or more races']
```

- However, other feature-specific models (age, income, language, etc) did not perform well and/or suffered from strong multicollinearity

Classification Model - Counties, Parties, Demographics

- Mapped Demographics Data on Voter Ballot Mail-in dataset
- Scanned the dataset for multicollinearity, eliminated “Other Race” and “Percent Hispanic|Latino”
- Attributed 0 to Accepted Ballots and 1 to Rejected

	ballot_county	ballot_vtr_party	ballot_status	received_rejReason	County ballots_cast	County ballots_rejected	County percent_english	County percent_spanish	County percent_other	County percent_english_very_well	County percent_english_not_very_well	County median_earnings	County HS_and_less	County more_than_HS
0	Atlantic	Unaffiliated	Accepted	0	135484	910	79.4	11.1	9.5	92.6	7.4	45935	32464	175711
1	Atlantic	Democratic	Accepted	0	135484	910	79.4	11.1	9.5	92.6	7.4	45935	32464	175711
2	Atlantic	Republican	Accepted	0	135484	910	79.4	11.1	9.5	92.6	7.4	45935	32464	175711
3	Atlantic	Democratic	Accepted	0	135484	910	79.4	11.1	9.5	92.6	7.4	45935	32464	175711
4	Atlantic	Democratic	Accepted	0	135484	910	79.4	11.1	9.5	92.6	7.4	45935	32464	175711
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4368331	Warren	Unaffiliated	Accepted	0	60053	970	90.8	4.7	4.5	98.1	1.9	67667	8063	76852
4368332	Warren	Democratic	Accepted	0	60053	970	90.8	4.7	4.5	98.1	1.9	67667	8063	76852
4368333	Warren	Unaffiliated	Accepted	0	60053	970	90.8	4.7	4.5	98.1	1.9	67667	8063	76852
4368334	Warren	Unaffiliated	Accepted	0	60053	970	90.8	4.7	4.5	98.1	1.9	67667	8063	76852
4368335	Warren	Unaffiliated	Accepted	0	60053	970	90.8	4.7	4.5	98.1	1.9	67667	8063	76852

4368336 rows × 39 columns

Classification Model - continued

- Set “class_weight” to “balanced”
- Tried GridSearchCV
- Tried different sets of variables i.e Parties + Demographics
  Counties + Demographics
  Counties + Parties + Demographics (excluding multicollinear columns)

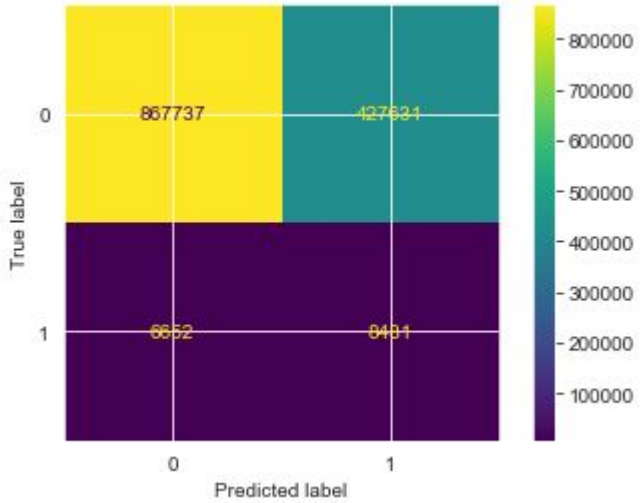
Classification Model - Results for Counties + Parties + Demographics (37 features excluding multicollinear columns and poverty with age data)

- No overfitting:
  Best Score for Train Set
  0.630

  Best Score for Test Set
  0.633

Confusion Matrix

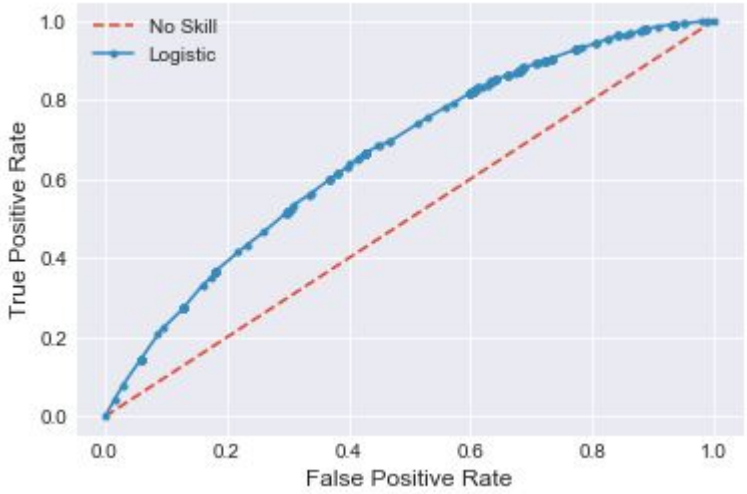
867737	427631
6652	8481



AUC curve

No Skill: ROC AUC=0.500
Logistic: ROC AUC=0.664

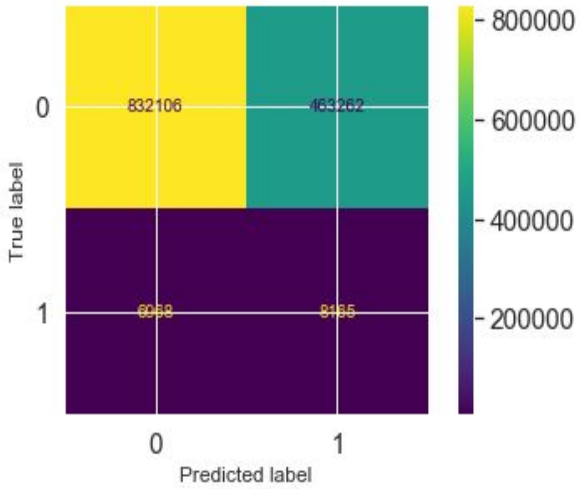
In [271]:



Classification Model after running Lasso for Features Importance (61 features down to 20)

['county_Monmouth',
 'county_Union',
 'party_Democratic',
 'party_Republican',
 'Percent $50,000 to $74,999',
 'Percent Other Race',
 'Percent 45 to 54 years',
 'Percent $15,000 to $24,999',
 'Percent $10,000 to $14,999',
 'Percent 25 to 34 years',
 'Percent $100,000 to $149,999',
 'percent_english',
 'Percent Hispanic or Latino',
 'Percent Black or African American',
 'Percent Two or more races',
 'Percent Less than $10,000',
 'Percent Asian',
 'HS_and_less']

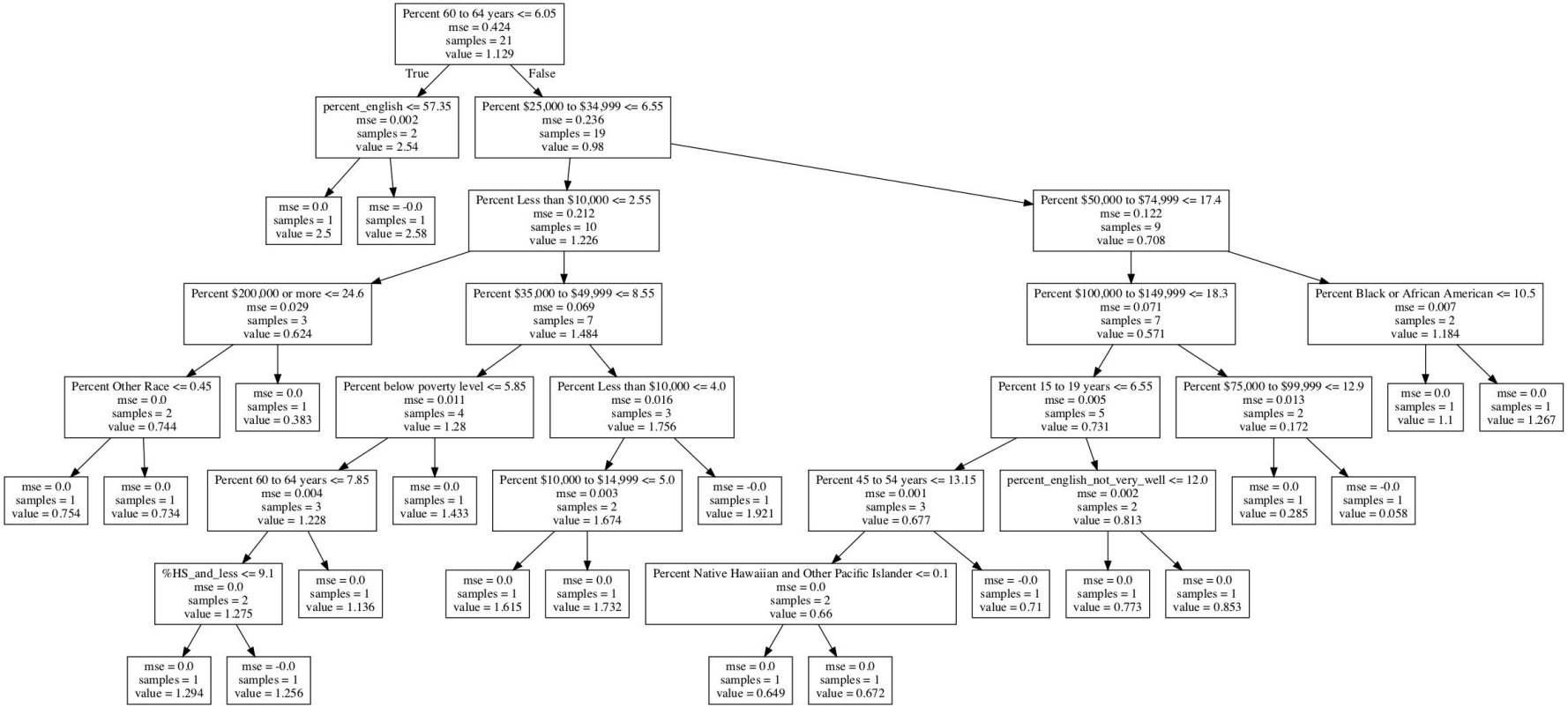
832106	463262
6968	8165



No Skill: ROC AUC=0.500
Logistic: ROC AUC=0.601

Scores for train and test set respectively: 0.65 and .63

Decision Tree



Decision Tree

- The Decision Tree uses five features to determine the rejection rate: age, earnings, race, language, and education.
- The most important feature is related to age, population between 60 and 64 years.

Features	Importance
Percent 60 to 64 years	0.496244
Percent Less than $10,000	0.179204
Percent 25,000to34,999	0.142926
Percent 50,000to74,999	0.065585
Percent 100,000to149,999	0.050246
Percent 35,000to49,999	0.043747
Percent $200,000 or more	0.009774
Percent below poverty level	0.003512
Percent 75,000to99,999	0.002890
Percent 15 to 19 years	0.002494
Percent Black or African American	0.001565
Percent 10,000to14,999	0.000769
percent_english	0.000368
percent_english_not_very_well	0.000354
Percent 45 to 54 years	0.000187
%HS_and_less	0.000081
Percent Native Hawaiian and Other Pacific Islander	0.000029
Percent Other Race	0.000024

Takewaways / Insights

- A substantial amount of variation in ballot rejection rates in counties across NJ can be explained by a relatively small number of demographic features (especially race, language, and age)

- However, without demographic data at the individual/voter level, the predictive power of demographic features is limited

- This analysis provides important and actionable insights regarding the ways that different voting methods may have on different groups

Takeaways/Insights

- For example, even though many counties with high proportions of Spanish-speaking voters already send out bilingual ballots/instructions, clearly language is still a barrier

- Same with age -- perhaps suggests an association w/ how many times a voter has cast a ballot previously and their familiarity with the voting process

- Future additions: deeper analysis of factors impacting reasons for rejection; new/different approach to voter-level analysis