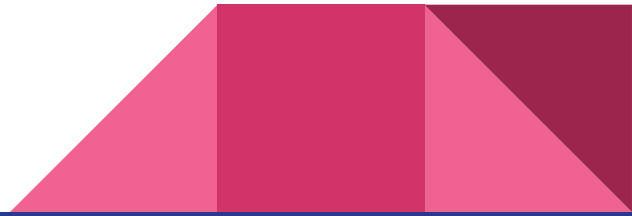
Predicting Age from Political Views

Pooja Barai, Emily Nomura, Megan Sindhi

Introduction

- Campaigns want to find demographics who will vote for them
- Increasing separation of views between generations (Gen Z vs. Boomers)
- Determining associations between political views and age could be helpful for campaigns
- Can we predict an individual’s age from their political views?



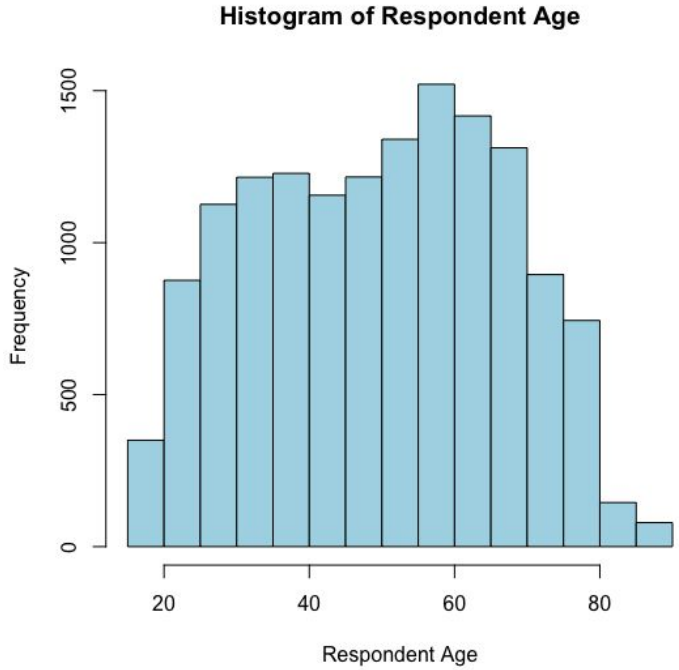
Data

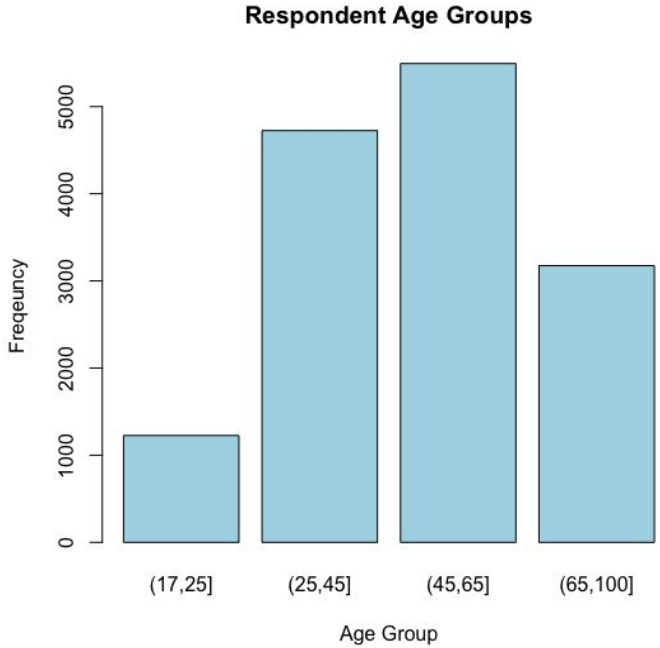
- American National Election Studies (ANES) Time Series Cumulative Data
  - Contained data from 1948-2020
  - Series of questions about a person's political views and plans to vote in presidential election
  - Participants differed from year to year: subjects all independent
  - Data limitations: 4000-8000 subjects per year may not represent full US population view
- Selected a subset of data from 2012-2020
  - Questions changed depending on relevant issues (ex: COVID in 2020, Civil Rights in 1964)
  - To get a similar group of questions, picked 3 consecutive years

Data Cleaning and Preprocessing

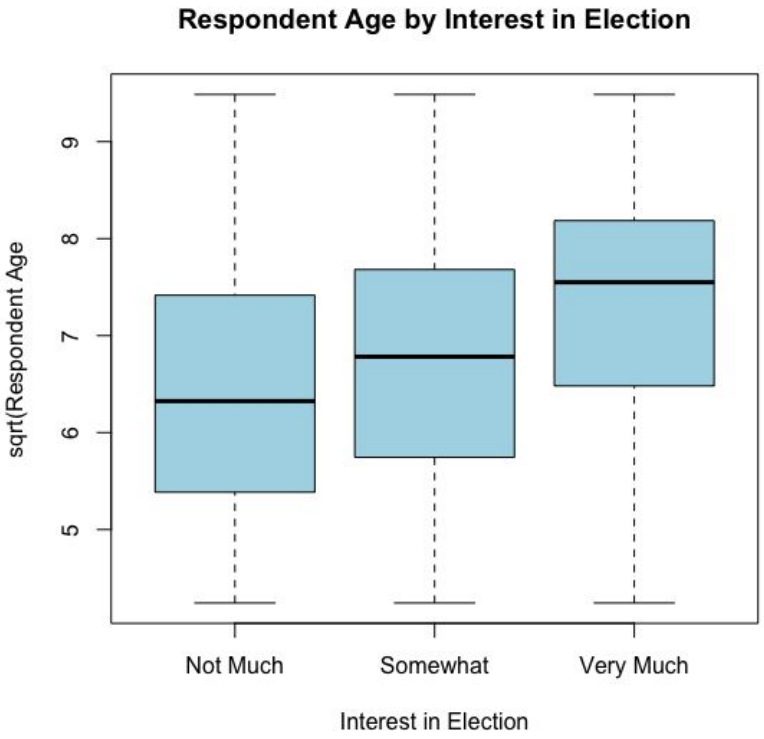
- Years considered: 2012, 2016, 2020
- Outcome variable: age
- Kept 59/1029 variables, all categorical factors
- Removed demographic variables (gender, race, etc.)
- Coerced continuous thermometer variables into 2-3 categories
- Replaced ‘missing’ survey values with NAs
  - Omit observations in columns with <1000 NAs
  - Create new ‘missing’ category for categorical variables with a large portion of missing data
- Converted all variables except for age to factors

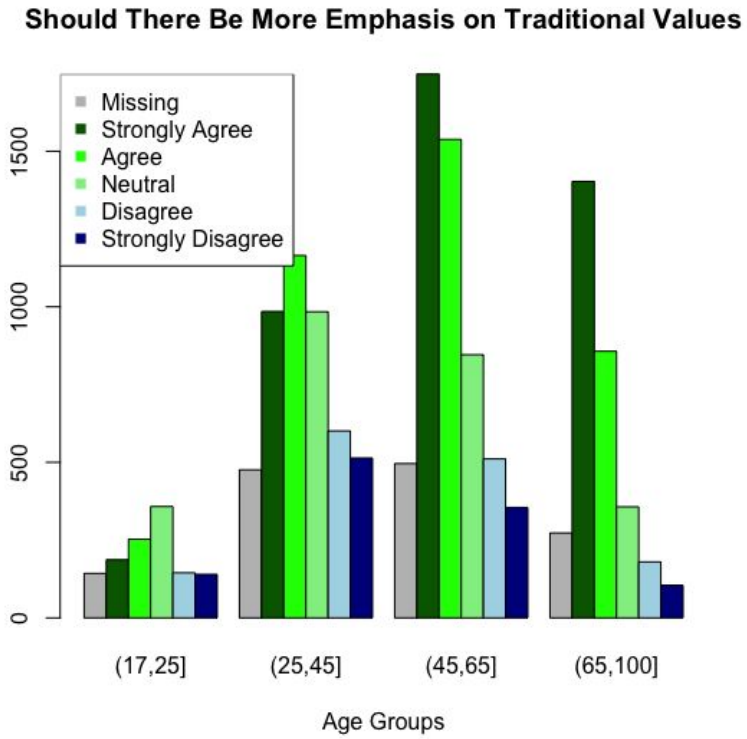
EDA: Response Variable



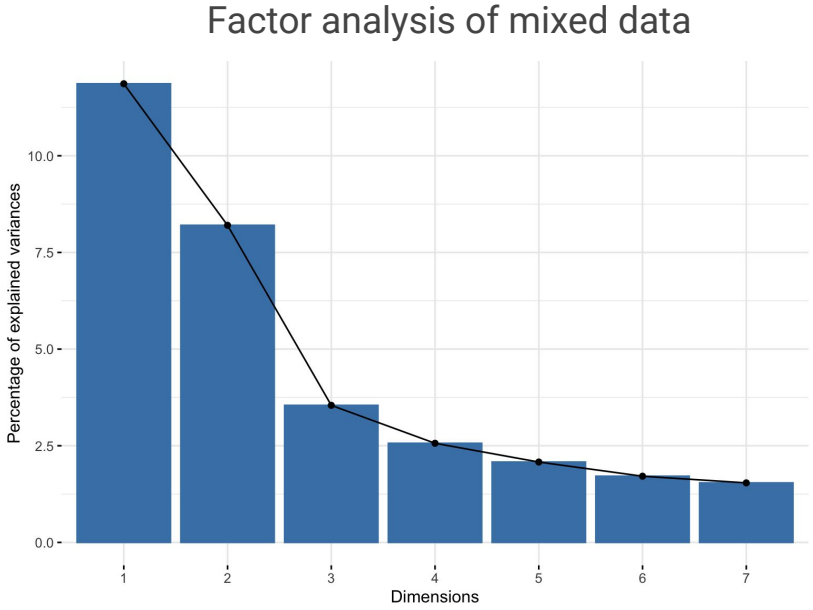


EDA: Plots of Interest





Dimension Reduction



Factor Analysis- 10 Factors

	Factor1	Factor2	Factor3	Factor4	Factor5
SS loadings	9.885	9.421	2.791	2.485	1.777
Proportion Var	0.139	0.133	0.039	0.035	0.025
Cumulative Var	0.139	0.272	0.311	0.346	0.371

	Factor6	Factor7	Factor8	Factor9	Factor10
SS loadings	1.740	1.417	1.219	0.839	0.679
Proportion Var	0.025	0.020	0.017	0.012	0.010
Cumulative Var	0.396	0.416	0.433	0.445	0.454

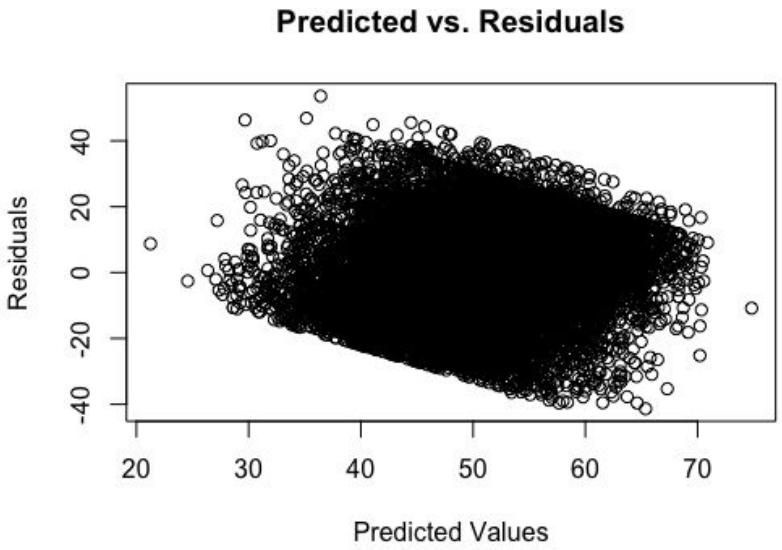
Linear Regression Model

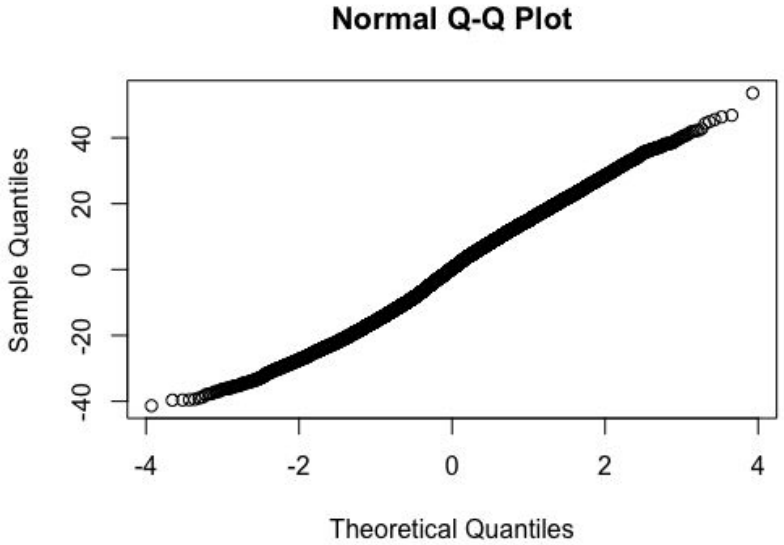
- Used a Lasso Linear Regression model for variable selection
- Year was kept as a predictor variable
  - Could be differences as political landscape has shifted considerably from 2012-2020
- 80% train 20% test
- 10 fold cross validation to select best lambda
  - Eval Metric: RMSE
- Tried several interaction terms
  - Mainly interactions between party and specific issues

Results

- Optimal lambda: 0.247
- Model kept 66 categories from 44 different variables
  - Several thermometer variables that captured sentiment on groups like big business, white people, democrats/conservatives
  - Variables related to if person approved of job government was doing
  - Variables related to views on health insurance, public school funding, traditional values, tax money
- Interaction terms didn’t improve the model
- Model performance was low
  - Test RMSE was 14.57
  - R2 was 0.23

Checking Model Assumptions



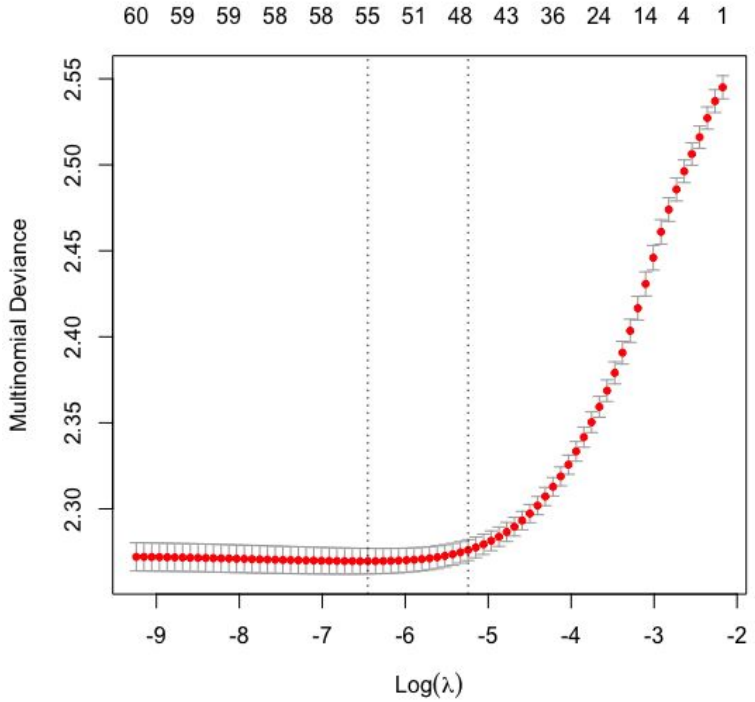


Results: Interpretations

- Intercept: Average expected age was 42
- Party was not included in final model
- Compared to being an independent, being weakly partisan meant a person was 1.05 years younger, and being strongly partisan meant a person was 1.48 years older
- Traditional Values: ‘This country would have many fewer problems if there were more emphasis on traditional family ties.’
    - Agreeing or somewhat agreeing added years to age
    - Being neutral or disagreeing subtracted years from age
    - Base case is not answering the question

An Alternative Approach: Multinomial Regression

- Turned age into a categorical variable
  - Young adults:17-25
  - Adults: 26-45
  - Middle Aged: 46-65
  - Senior: 66+
- Predicted which class each person would belong to
- Used lasso regression for variable selection: optima lambda = 0.00529



Multinomial Results

- Performance not great but better than random
  - AUCs above base: (0.53-0.81)
  - Overall accuracy: 48.2%
  - Class-wise accuracy:

(17,25]	(25,45]	(45,65]	(65,100]
0.4%	58.7%	60.7%	54.3%

- Interpretations
  - similar variables to lasso model eg. trad_values
  - Therm_white1: -0.41
  - Interest_in_elec3: 0.49
  - Donate_campaign1: 0.64

Confusion Matrix

True/ Pred	(17,25]	(25,45]	(45,65]	(65,100]
(17,25]	1	1	0	0
(25,45]	157	538	316	78
(45,65]	80	347	681	377
(65,100]	5	30	124	189

Conclusions and Next Steps

- Creating categories and using logistic regression may be a better approach
  - Continue exploring multinomial regression
  - Other ML techniques may be better (random forest, knn, etc.)
- Additionally, could attempt other types of models
  - Proportional odds cumulative logit- for ordinal data
    - Predicts odds of falling into or below a category
- May not be possible to accurately predict age with just political views
  - There may be too much variation to model accurately

Questions?

Group Member Contributions

Pooja: Factor Analysis, EDA, Multinomial Regression

Emily: Data Cleaning and Prep, PCA

Megan: Linear Regression, EDA