Data cleaning and preparation

All data cleaning was done in Excel and Pandas.

Outcome Variables

First I needed to find a file with the outcome variable. I found one on github. I cannot use the county name as the identifier column since there are multiple counties with the same name. Thus I used county fips codes instead. There was no column that indicated directly which party had won the election so I had to make one instead (just check whether the diff was > 0). In this project, Republican = 1 and Democrat = 0.

The data has fips codes 11002-11008 for Washington D.C which is not correct as they are not voting jurisdictions so I removed them.

Alaska does not handle elections by counties like the other states do. The fips codes are thus not consistent with the data in the other datasets and there is no readily available information on the results in each county in Alaska for 2024. I thus decided to assume a Republican outcome (based on Alaska's historical voting outcome) for each county knowing there is bound to be some error. Alaska only accounts for ~1.3% of the total observations.

The following 4 datasets were taken from:
County-level Data Sets: Download Data | Economic Research Service

Poverty Estimate

For Poverty estimates dataset there were columns for the country and the states which I had to remove. There were many attributes for each observation with 90% confidence intervals, them being:
(a) Total in poverty (POVALL_2023)
(b) Poverty **rate** (PCTPOVALL_2023)
(d) Children in poverty age 0-17 (POV017_2023)
(e) Child poverty **rate** age 0-17 (PCTPOV017 _2023)
(f) School-age children (5–17) in poverty (POV517_2023)
(g) Median Household income
(h) Poverty **rate** for children under 5

I decided to use only Poverty rate, Child poverty rate age 0-17, and Median household income as the other attributes were either only confidence intervals or highly correlated with each other. Only the count of people would have not been enough for the dataset because the population is not distributed evenly so it is good that the rate was included in the data as well.

Additionally, the attributes were distributed amongst the rows instead of the columns. I had to restructure the data file to accommodate for that.

Education Level

The education file was laid out in similar fashion to the poverty file (indeed they were both obtained from the same website) and thus similar transformations were performed. Only the most recent data pertaining to 2019-2023 was kept and only the percentage of the population relating to level of education was kept (Less than High School, High School, Some College, College+).

Unemployment and median household income

The unemployment and median household income was laid out in similar fashion to the poverty file. Only the unemployment rate, median household income, and median household income as a percentage of median household income of the state was kept.

Population Estimate

The population estimate was laid out in similar fashion to the poverty file. The features that were used in the final model were the population estimate, death rate, rate of net migration, and two metrics defined by the dataset makers that roughly determine the economic development of the county (Rural Urban Continuum Code, and Economic Typology).

Annual County Resident Population Estimates by Age, Sex, Race

County Population by Characteristics: 2020-2023

The dataset only had count values for the age groups so I decided to change it to the rate of population at a certain age group as I felt that was more valuable. The same goes for male and female count, race count, and race combined with gender count.

Additional Data Sources

Some more data points that could help the prediction are Religion, “urban-ness” of the county (housing density, population density, area of the county), polling, and betting odds data leading up to the election.