**Data Source 1: YCOM (Yale Climate Opinion Map) Data**

1. This data, which indicates the result of climate opinion surveys at a county-level, was provided through the Yale Program on Climate Communication, which studies climate opinion and messaging. In each given year, they conduct a survey on a variety of climate-related topics, which is reflected in this map. After conducting a survey of over 25,000 households, they use a statistical model (specifically, “multilevel regression with post stratification”) to get the final output. The results of this statistical model are cross-validated and compared to results of other independent, local-level surveys conducted in various states, and third-party survey data.
2. **Data integrity**
   1. **\*\*Structure\*\*:** This data is stored as a CSV. There is a separate CSV for each year, where each row represents a given geographic area (i.e. a given county) and each column is the percentage of respondents who answered in a given way. The question wording and meaning of each response is included in the meta-data.
   2. **\*\*Granularity\*\*:** Each record represents a given geographic area, and the percent of respondents who answered in a certain way. Some records capture granularity at a state-level, but we isolate records which capture granularity at a county-level. While there isn’t explicit aggregation, it’s important to note that results were based on a combination of survey responses and modeling off of those responses—so there could be some aggregation in how certain responses were used to model for other areas. We don’t need to do any additional aggregation.
   3. **\*\*Scope\*\*:** The dataset contains a row for every county in the 50 sets, and there is data for 2014, 2016, 2018, and 2020.
   4. **\*\*Temporality\*\*:** Each year indicates the year in which the survey was conducted. We don’t have any funky timestamps because each year is given as its own csv. We don’t know the exact date that each survey was conducted, but that shouldn't’ have much of an impact on the data or modeling.
   5. **\*\*Faithfulness\*\*:** We believe data is trustworthy first because of the reputation of the Yale Center on Climate Communication, whose results have been published in academic papers. Additionally, they have detailed the extensive steps taken to ensure their data is valid (i.e. cross-validation, comparison with surveys).
3. We use this data for **target variables**. In each year, we isolate 3 specific response columns (“are you worried about climate change”, “do you support CO2 limits on power plants” and “do you think climate change will harm people in the US”). These represent questions that were asked in all surveys across time, and represents what we are trying to predict.

**Data Source 2: MIT Election Lab Data (Presidential Elections)**

1. We found county-level presidential election data from MIT’s Election Data Science Lab. They collected the data using publicly-available data on election results in each county from 2000-2020, and compiled it into a single CSV for every county in the US.
2. **Data integrity**
   1. **\*\*Structure\*\*:** Data is stored in a single CSV, which contains columns for the county, year, candidate, party, number of votes for a given candidate, and total votes in the county. It does not include an explicit column for the proportion of votes going to a given candidate, which we must add.
   2. **\*\*Granularity\*\*:** Data is aggregated at the county level. Each row represents a vote for a given candidate in a given county in a given year for a presidential election (i.e. for the 2020 election, there is a row for Joe Biden in County X, Donald Trump in County Y, and other 3rd party candidates in County X). Since county-level election returns are already available, the data was not manipulated or aggregated. We do not have to aggregate additionally, because we want to use county-level data. All records are captured at the same level of granularity (county-level).
   3. **\*\*Scope\*\*:** This covers the entire 50 states from 2000-2020 in presidential elections only. The data is completed: there are no missing values or presidential election years in which data is not available.
   4. **\*\*Temporality\*\*:** Each year represents the year of the presidential election. Time is only represented in years, and since presidential elections happen every 4 years, it would be easy to tell if records were entered incorrect((which they weren’t), because we’d see a year like 2017. We don’t have to worry about funky timestamps, because all we care about is the year, which are all entered in the same format
   5. **\*\*Faithfulness\*\*:** Since presidential election data is publicly available and released, we don’t have concerns about falsified data. Additionally, the reputation of MIT university research gives us additional confidence in the validity of this data. There are some candidates for whom the data says receive 98%+ of the vote in a county. This seemed exceptionally large, but we examined the same county in previous elections and found that these tend to be smaller counties which leaned heavily towards one party— so while we thought these values were exceptionally large, we did additional fact checking to make sure the data did not have errors.
3. This data is used as a feature: for a given county, we’ll use presidential election data (i.e. proportion for a democratic candidate) as a feature to predict opinions on climate change.

**Data Source 3: Dave Leip’s Atlas of US Election (Midterm Election Data)**

1. This data source provides county-level midterm election data (i.e. for House races). While some sources provide county-level data for some states, this provides county-level House election results across the United States. According to the metadata, these are compiled using primary sources, such as reports from election agencies (Board of Election or Secretary of State). For
2. Data integrity
   1. **\*\*Structure\*\*:** For each election year (2014, 2016, etc.), there is a CSV that includes each county in a state and the proportion of votes in that county going to the republican and democrat candidate. The data is cleanly organized into rows. The data includes other columns besides proportions that are not relevant to the project.
   2. **\*\*Granularity\*\*:** Each record represents a given county in a given election year. Almost all records capture granularity at the county-level, however the data also includes summaries for each state (i.e. aggregation at a state-level for proportion of votes going to each party). We do not need to aggregate further beyond this.
   3. **\*\*Scope\*\*:** There is available for elections going back to 1992. We are using data from 2014-2020. While it covers almost every county, we do not have data for
   4. **\*\*Temporality\*\*:** Each year represents the year an election happened. Each CSV is for a specific year, so we don’t have specific date/time fields in the data, only in the original CSV download.
   5. **\*\*Faithfulness\*\*:** Since this data is compiled from primary sources, we have confidence in its values. This atlas is also offered as a library resource at numerous colleges/universities, which give us additional confidence in its trustworthiness.
3. This data is used as a feature: for a given county, we’ll use midterm election data (i.e. proportion for a democratic candidate) as a feature to predict opinions on climate change.

**Data Source 4: Extreme Heat Data**

1. This dataset of extreme heat events by county was found at the CDC’s EPH (Environmental Public Health) tracking database. It provides a database of environmental measures. This specific dataset depicts extreme heat data, defined as 2 or more consecutive days where the daily maximum temperatures exceeded the county’s 90th percentile. Temperature data is primarily taken from sensors, although some is modeled using the North American Land Data Assimilation System. The percentile values (i.e. for the 90th percentile) are taken from 1979-2016 data. County-level estimates of temperature are taken by processing model data, since there isn’t a single sensor point that encompasses an entire county. Importantly, these metrics are only taken in summer months (May-Sept), since extreme heat likely won’t occur outside of those months (but if it did, it wasn’t included in this data)
2. Data integrity
   1. **\*\*Structure\*\*:** Data is given in a CSV. Each row represents a county, and column include a year and the value of the metric (i.e. # of extreme heat events).
   2. **\*\*Granularity\*\***: Each row represents a county, and column include a year and the value of the metric (i.e. # of extreme heat events). While the values themselves were not aggregated, some of the temperature values used to calculate the number of extreme heat events were taken from a model, which aggregated data points from different locations in the county. We do not have to aggregate further, since this is already at a county-level
   3. **\*\*Scope\*\*:** The data covers years from 1979-2019. Only extreme heat events in May-Sept are included. All states except Alaska and Hawaii are included, and some counties have missing values because
   4. **\*\*Temporality\*\*:** Dates are represented in a column that describes the year of each data point. Like previous data sets, we don’t have to worry about funky time stamps.
   5. **\*\*Faithfulness\*\*:** Since this data comes from the CDC, we have confidence in its faithfulness. One could question the model (North American Land Data Assimilation System) used for some values, but since this model is used by government scientists and academics, we believe it’s trustworthy.
3. This data is used as a feature: for a given county, we’ll use extreme heat data as a feature to predict opinions on climate change.

**Data Source 5: Unemployment Data**

1. This data from the Bureau of Labor Statistics (BLS) lists average yearly unemployment rates by county for each year from 1990 to 2020. To calculate unemployment rates, BLS sends out surveys monthly that are used to calculate unemployment rate. Over the course of each year, BLS averages this data to find yearly unemployment averages.
2. **Data Integrity**
   1. **\*\*Structure\*\*:** Data is given in an XLS (CSV) file. Each year is given in its own file. Data is separated in rows for each county.
   2. **\*\*Granularity\*\*:** Each row represents a county in a given year, and all represents granularity at this county level. Inherently, data is aggregated at a county-level, which could hide sectoral trends (i.e. if unemployment among, say, women is exceptionally high). Additional aggregation is not desirable.
   3. **\*\*Scope\*\*:** This covers all states in the United States from 1979-2020, with one data point for each county in each year. We do have some missing values for
   4. **\*\*Temporality\*\*:** The only date/time factor we deal with is year. Since each data table corresponds to a certain year, we don’t have a year column, but instead identify the year using the name of each data table.
   5. **\*\*Faithfulness\*\*:** BLS is the go-to resource for labor statistics—it’s entire job is to calculate accurate unemployment statistics, so we have full confidence in this data.
3. This data is used as a feature: for a given county, we’ll use unemployment as a feature to predict opinions on climate change.

**Data Source 6: Employment by Sector data**

1. This data from the Census Bureau provides employment by sector and demographic for different states and counties (i.e. % of people in a state working in the arts). This data is collected using surveys mailed to a random sample of addresses (approximately 295,000 addresses a month). For non-respondents, ACS follows up with a personal visit to ensure completion. The responses to this data, combined with statistics from other government agencies, are combined to provide the dataset listed above.
2. **Data Integrity**
   1. **\*\*Structure\*\*:** Data is provided in a CSV download, where each year is in a separate CSV. The data is already given in rows, and columns include total workforce per state and number of people working in certain industries (i.e. mining, arts, IT, etc.)
   2. **\*\*Granularity\*\*:** Data can be accessed at both the state and county level. However, the county-level granularity does not include every county, so we opted for state-level data. This data was aggregated by the census bureau after receiving survey results from each subsection of the state. Each record represents a state.
      1. We will want to aggregate this somewhat by summing employment in “extractive” industry (i.e. people working in mining, forestry, etc) to find the proportion of workforce in each county that works in an extractive industry.
   3. **\*\*Scope\*\*:** There is data for all 50 states from at least 2005 to 2019. The project only uses data from 2014 onwards. There are no missing values for each state, however there are missing values at the county-level. Given the methodology (survey sent to random addresses), there are limits to how accurate the data can be (i.e. it can’t capture every single person). However, there are no missing years or states.
   4. **\*\*Temporality\*\*:** The only date/time factor we deal with is year. Since each data table corresponds to a certain year, we don’t have a year column, but instead identify the year using the name of each data table.
   5. **\*\*Faithfulness\*\*:** Like BLS, the Census Bureau’s job is to produce accurate statistics. While there could be some noise if people lie on their survey response, we have confidence in the source of this data. Looking at the data, we don’t see any violations of obvious dependency (i.e. number of workers being larger than population in the county) that would lead us to question this data.
3. This data is used as a feature: for a given county, we’ll use employment by certain industries. as a feature to predict opinions on climate change.

**Data Source 7: Presidential Speech Data**

1. This data set includes transcripts from official presidential speeches. Most transcripts were accessed from the Miller Center, an affiliate of the University of Virginia that does research and archival work on American politics. The Miller Center keeps records of official presidential speeches and their transcripts, and this data table compiles them.
2. Data integrity
   1. **\*\*Structure\*\*:** Table given as a CSV with predefined rows, where each row is a given speech. Columns include date, president, party, transcript and url of transcript for each speech.
   2. **\*\*Granularity\*\*:** Each row represents a county, and column include a year and the value of the metric (i.e. # of extreme heat events). While the values themselves were not aggregated, some of the temperature values used to calculate the number of extreme heat events were taken from a model, which aggregated data points from different locations in the county. We do not have to aggregate further, since this is already at a county-level
   3. **\*\*Scope\*\*:** The data table includes all official presidential speeches from 1792 to 2019. For the period of interest (2010 onwards), there are certainly no missing values or speeches.
   4. **\*\*Temporality\*\*:** \*\*Temporality\*\* represented by dates of each speech. All dates represented in a YYYY-MM-DD format, and in extracting the year from this date column we ensured that there were no strange values for years (i.e. if a certain date was written in DD-MM-YYYY, we would have identified that; although we didn’t find any such dates).
   5. **\*\*Faithfulness\*\*:** Especially for the period of interest, where there is recording and transcripts of almost everything a president does, we are confident that this dataset includes the proper transcript of all official presidential speeches. Additionally, its connection with the Miller Center gives us additional confidence in the original creator of the information that this dataset uses.
3. This data is used as a feature: for a given year (and for all counties in that year), we’ll use the number of times a keyword is mentioned in a speech as a feature to predict opinions on climate change.

**Data Source 8: County Population Totals**

1. This data from the [US Census Bureau](https://www.census.gov/programs-surveys/popest/technical-documentation/research/evaluation-estimates/2020-evaluation-estimates/2010s-counties-total.html) provides estimates of county population totals from the years 2010-2020. These values are estimates of the number of residents, based on the most recent decennial census data and factoring in births, deaths, and migrations. The Census Bureau utilizes a “vintage” method in which the release of a new year’s estimates includes a re-release of previous years’ estimates with improved methodology and input data.
2. Data integrity
   1. **\*\*Structure\*\*:** Table given as a CSV with predefined rows, where each row is a given county. Columns include state name, county name, 2010 Census population observation, and population estimates for years 2011-2020.
   2. **\*\*Granularity\*\*:** Each row represents a county, and columns include a year and the value of the population estimate. While estimates of population by age, sex, and race are created through a top-down aggreagate approach, the total population estimates at the county level are not aggregated. We do not have to aggregate further, since this is already at a county-level.
   3. **\*\*Scope\*\*:** The Census Bureau offers data dating back to 1900, but the data table we pulled includes all US counties from 2010-2020. For our period of interest (2010 onwards), there are certainly no missing values or speeches.
   4. **\*\*Temporality\*\*:** The only date/time factor we deal with is year. We identify the year of the population estimate using the name of each column.
   5. **\*\*Faithfulness\*\*:** The Census Bureau’s job is to produce accurate statistics. While there is some uncertainty in the population values given that they are estimates, we have confidence in the source of this data. Looking at the data, we don’t see any violations of obvious dependency that would lead us to question this data.
3. This data is used as a feature: for a given year (and for all counties in that year), we’ll use the population of a county as a feature to predict opinions on climate change.

**Data Source 9: Educational Attainment**

1. This data from the [US Census Bureau](https://data.census.gov/cedsci/table?t=Educational%20Attainment&g=0100000US%240500000&tid=ACSST5Y2014.S1501) provides estimates of various measures of educational attainment based on the American Community Survey. The particular values of interest are proportion of individuals with only a high school diploma or equivalent and proportion of individuals with a bachelor’s degree or higher in the subpopulations of 18-24 year olds and 25 year olds and older.
2. Data integrity
   1. **\*\*Structure\*\*:** Table given as a CSV with predefined rows, where each row is a given county. Columns include geographic area name (county and state), estimates of various quantifications of educational attainment, and their respective margins of error.
   2. **\*\*Granularity\*\*:** Each row represents a county, and columns include a proportion value. Because the underlying observations are a survey and not a census, the Census Bureau aggregates responses and statistically transforms observations to come up with estimates at the county level. We do not have to aggregate further, since this is already at a county-level.
   3. **\*\*Scope\*\*:** The Census Bureau offers data from every 5-year ACS period as well as 1-year estimates. Because the educational attainment values of interest to us (proportions) are omitted from the 1-year estimates, the scope is limited to every 5-year period, particularly 2014 and 2019.
   4. **\*\*Temporality\*\*:** The only date/time factor we deal with is year. We identify the year of the educational estimate based on the year of the table, since each year’s data is contained in a separate file.
   5. **\*\*Faithfulness\*\*:** The Census Bureau’s job is to produce accurate statistics. While there is some uncertainty in the educational attainment proportions given that they are estimates, we have confidence in the source of this data. Looking at the data, we don’t see any violations of obvious dependency that would lead us to question this data.
3. This data is used as a feature: for a given year (and for all counties in that year), we’ll use the educational attainment of a county as a feature to predict opinions on climate change.

**Data Source 10: Number of Fossil Fuel Power Plants**

1. This data from the [EIA](https://www.eia.gov/electricity/data/eia860/) provides generator-level specific information about existing and planned generators and associated environmental equipment at electric power plants with 1 megawatt or greater of combined nameplate capacity. Because the data is clearly delineated into currently operating plants and also what type of generation each plant uses for each year, the data can be transformed to represent the number of fossil fuel power plants per county in the US.
2. Data integrity
   1. **\*\*Structure\*\*:** Table given as a CSV with predefined rows, where each row is one generator facility. Columns include county name, state abbreviation, name of facility, type of generation, nameplate capacity, and more.
   2. **\*\*Granularity\*\*:** Each row represents one generation plant, with supplemental column information about what county it resides in. This means we have to aggregate further, grouping by county name to get a value that represents the number of plants at the county-level.
   3. **\*\*Scope\*\*:** The EIA offers data for every year dating back to 1990. The scope is limited to the years from 2014 to 2019. Geographically, the data contains information from power plants across the entirety of the US.
   4. **\*\*Temporality\*\*:** The only date/time factor we deal with is year. We identify the year of the power plant information based on the year of the table, since each year’s data is contained in a separate file.
   5. **\*\*Faithfulness\*\*:** The EIA’s job is to produce accurate information. While there is some uncertainty in how exactly the data for each power plant is collected with this survey form, we have confidence in the source of this data. Looking at the data, we don’t see any violations of obvious dependency that would lead us to question this data.
3. This data is used as a feature: for a given year (and for all counties in that year), we’ll use the number of fossil fueled power plants in a county as a feature to predict opinions on climate change.

**Data Source 11: Median Household Income and Poverty Estimates**

1. This data from the [US Census Bureau](https://www.census.gov/data/datasets/2019/demo/saipe/2019-state-and-county.html) provides estimated demographic data related to income and poverty as part of the Small Area Income and Poverty Estimates (SAIPE) Program.
2. Data integrity
   1. **\*\*Structure\*\*:** Table given as an XLSX file with predefined rows, where each row is one geographic area (USA, state, and county). Columns include area name, estimates for number of people living below the poverty line across different age ranges, statistical information about the estimate, and median household income estimate in dollars.
   2. **\*\*Granularity\*\*:** Different rows represent different geographic scales. The first row represents the entire US, while subsequent rows represent each state followed by its respective counties. This means we have to drop national and state rows to retain only county-level data.
   3. **\*\*Scope\*\*:** The US Census Bureau offers annual estimates as part of the SAIPE Program. The scope is limited to the years from 2014 to 2019, specifically the years 2015, 2017, and 2019.
   4. **\*\*Temporality\*\*:** The only date/time factor we deal with is year. We identify the year of the income and poverty information based on the year of the table, since each year’s data is contained in a separate file.
   5. **\*\*Faithfulness\*\*:** The Census Bureau’s job is to produce accurate information. While there is some statistical uncertainty in the estimates, we have confidence in the source of this data especially considering the confidence intervals provided for each estimate. Looking at the data, we don’t see any violations of obvious dependency that would lead us to question this data.
3. This data is used as 2 features: for a given year (and for all counties in that year), we’ll use the proportion of population living under the poverty line and median household income as features to predict opinions on climate change.