George Mason University

CS-504-001

Spring 2022

Climate Change: Making It Personal

Team Cowboys

Brian Monter, Kristina Longe, Ron Maxseiner, Jason Pinto, Sai Sushmitha Kadambari

Table of Contents

[**Abstract**](#_v6uug89gttfn) **4**

[**1 Introduction**](#_cm9cv52ushig) **5**

[1.1 Background and Rationale](#_qtae9yajclfz) 5

[1.2 Research](#_ju6sfsvmynwf) 6

[1.3 Project Objectives](#_e3hmmqhjsx8m) 9

[1.4 Problem Space](#_esbbqummro4l) 9

[1.5 Primary User Story](#_nyndv1do6rws) 10

[1.6 Solution Space](#_c2c7tr65j4pi) 11

[1.7 Product Vision](#_9cq9qjys04tm) 11

[1.8 Definitions](#_1a6ykit07rns) 12

[**2 Data Acquisition**](#_5timsumslkdz) **13**

[2.1 Overview](#_u8s6ih9kh8v5) 13

[2.2 Field Descriptions](#_lx5v5mx1xa2r) 13

[2.3 Data Context](#_8obfoidtwvn6) 16

[2.4 Data Conditioning](#_id8pvh5ndp18) 17

[2.6 Other Data Sources](#_l7xs7lafrari) 19

[**3 Analytics and Algorithms**](#_1w0ipukww4b5) **19**

[3.1 Algorithm Descriptions](#_l6trlehbmb5a) 19

[3.2.1 Ridge regression](#_yiy228ik8je) 20

[3.2.2 Artificial Neural Networks (ANN) - Keras](#_5pu4edgcykri) 21

[**4 Visualizations**](#_vem8sn5dpwgb) **21**

[**5 Findings**](#_etzou8u321ed) **25**

[**6 Summary**](#_yl2o6ruc6n8n) **27**

[**7 Future Work**](#_dvgo56jbwcjd) **28**

[**Appendix A - Code References**](#_ba75m3gmftcj) **29**

[**Appendix B - Risk Section**](#_y3ov07ngacep) **30**

[**Appendix C - Agile Development**](#_vhlvatcbyygc) **32**

[**Appendix D - References**](#_dgxmi1mmr7fw) **34**

# 

# 

# Abstract

When you are faced with a problem what do you do? In most cases, individuals and organizations would address the problem and figure out a way to resolve it. One of our biggest problems we currently face as a human race does not fit under that narrative. No matter how much it is talked about or how much you can actually feel the impact, some people refuse to adapt. The problem we are talking about here is Climate Change. The goal of our project is to make climate change personal to the reader so that there is a reason to act on it. While not every person is going to act, shifting the person away from opposing action towards indifference would be considered a positive step as well.

Climate change is a frequent topic of social and political discourse; it is often discussed in terms of worrisome increasing global temperature with correlations drawn between fossil fuel usage and average temperature change. Indeed, with growing popularity of electric cars and alternative energies it is worth considering how gas, coal, and oil have impacted global temperature historically and how they will impact average global temperature in the coming years. Using data from BerkelyEarth, we dissected yearly fossil fuel usage to measure its impact on global temperature and create models to predict average temperature into the year 2050. We use this data to present impactful information to users in the form of an interactive application. Popular conservationist opinion suggests that mass transportation and clean energy are quintessential to positive climate response in the form of decreasing global temperature change. Our research suggests that, while fossil fuels do impact global temperature, they should not be considered the pinnacle on which global action stands. Given the multifaceted and often unpredictable nature of climate change, it is hard to make predictions based on fossil fuels alone.

# **1 Introduction**

## **1.1 Background and Rationale**

Climate change is one of the most talked about issues in the world today, which leads us to wonder, what is climate change? Climate change is a difficult topic to discuss as there are many interpretations of the simple words, often influenced by political views and biased information. In general, it refers to long-term shifts in temperatures and weather patterns. As described by the United Nations, climate change can be a natural process where temperature, rainfall, wind, and other elements vary over decades or more. Over millions of years, our world has been both warmer and colder than it is now. But today we are experiencing unprecedented rapid warming from human activities, primarily due to burning fossil fuels that generate greenhouse gas emissions. (“Key Findings | United Nations”)

Climate change has many impacts but the primary impact is rising temperatures. According to the United Nations, Earth is now about 1.1°C warmer than it was in the 1800s. We are not on track to meet the Paris Agreement target to keep global temperature from exceeding 1.5°C above pre-industrial levels. That is considered the upper limit to avoid the worst fallout from climate change. (“Key Findings | United Nations”)

While the facts seem to be quite clear, it is difficult to get people to act in a way that would help alleviate the problem. Even though the impacts may not be that far away, it is difficult for people to grasp if this summer was just a hot summer or if that is evidence of a larger trend. Extreme weather and long-term temperature trends are part of Earth's normal behavior and this makes it difficult for individuals to determine if an event is evidence of large-scale change or a one time event.

## **1.2 Research**

Our primary data set is [Climate Change: Earth Surface Temperature Data | Kaggle](https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data?select=GlobalTemperatures.csv), which contains the following columns:

* Date: starts in 1750 for average land temperature and 1850 for max and min land temperatures and global ocean and land temperatures
* LandAverageTemperature: global average land temperature in celsius
* LandAverageTemperatureUncertainty: the 95% confidence interval around the average
* LandMaxTemperature: global average maximum land temperature in celsius
* LandMaxTemperatureUncertainty: the 95% confidence interval around the maximum land temperature
* LandMinTemperature: global average minimum land temperature in celsius
* LandMinTemperatureUncertainty: the 95% confidence interval around the minimum land temperature
* LandAndOceanAverageTemperature: global average land and ocean temperature in celsius
* LandAndOceanAverageTemperatureUncertainty: the 95% confidence interval around the global average land and ocean temperature

The data contains a total of almost 8.6 million records. Out of those records, 364,130 of the LandAverageTemperatures are NaN values. Since this only composes 4.4% of the records, it is not a significant concern as removing these rows will still leave 8.25 million records for analysis.

Our secondary data set is [Consumption in the world of fossil fuel | Kaggle](https://www.kaggle.com/code/gianlab/consumption-in-the-world-of-fossil-fuel), which contains 9 columns:

* Entity: State
* Code: 3-letter code that identifies the state
* Year: from 1966 to 2019
* Entity: Country
* Fossil.fuels..TWh.growth ... sub.method. : Annual change in fossil fuels
* Coal..TWh.growth ... sub.method.: Annual change in coal
* Oil..TWh.growth ... sub.method. : Annual change in oil
* Gas..TWh.growth ... sub.method. : Annual change in gas
* region: Continent
* sub.region: Area of the continent

There are no null values in this data set but it has only 3,987 records. This data is summarized by year, state and country code.

An initial review of these data sets indicates that their quality is good. To further verify the quality of the data, outlier analysis will be performed to determine if outliers exist in the data. A method of dealing with both the missing and outlier data needs to be assessed and implemented. In order to better understand the data itself, descriptive analysis will first be performed. This will include statistical measures such as mean and standard deviation and the analysis will need to be performed over time and location.

Following the descriptive analysis, modeling the data will begin. Since this is time series data, there are two significant modifications that need to be made to the data. First, time series data can have delayed impacts. For example, this month’s temperature can impact next month’s temperature. Time delayed series will be added to the source data and included in models to determine if the current month’s data or last month’s data has the highest correlation with the temperature data. Second, surface temperatures certainly are dependent on seasons. This can be removed by summarizing by year or adding a sinusoidal element to the prediction. More investigation is needed in this area to determine the best way to handle these results. Also, in the case of fossil fuel consumption, the impact on the surface temperatures may be more related to the cumulative CO2 in the atmosphere than the previous year's amount. This may require the introduction of a term that is the sum of the last X years of data where X could be anything from a few years to all the previous data.

The two largest risks that we see at this time are data volume and the complexity of the data. The data will require several additional predictor variables to determine the true relationships. With 8.3 million records, the computational needs could get quite high for some of the non-linear models. Also, this is a complicated problem and it is possible that even with all the data we have, it is only a small number of the predictors required that actually contribute to the rising surface temperatures. In this case, all our models will not be strong predictors as the data is simply not available that would enable accurate prediction.

## **1.3 Project Objectives**

Our project has three main objectives: to analyze historical surface temperature data and build a predictive model, to explore the relationship between the use of fossil fuels and surface temperature changes, and to make climate change more real and accessible to individual people.

The predictive model will be used to provide an expected distribution of temperatures for a given area which will be displayed in a graphical representation. The representation will show the user if the weather they are experiencing falls inside of a normal range of expected temperatures.

By exploring the relationship between fossil fuel usage and long-term surface temperature change, we can provide guidance on how changes that individual people can make will impact climate change.

## **1.4 Problem Space**

We will use the analysis of surface temperature data to provide predictions for distributions of future temperatures. This information will be presented to users in relation to the acute weather conditions they are experiencing to determine if they are statistically normal or on a trend toward higher temperatures.

It seems that people may understand the data, but they do not seem to be acting on it. We believe that by presenting evidence for change in their own local climate in a graphical manner, users will be more likely to engage with the data and make more climate-friendly decisions.

Some problems that we could encounter when building a predictive model based on our current dataset of average surface temperatures are computation costs given the size of our dataset, and the selection of descriptive enough explanatory datasets such as fossil fuel usage. If the trend of temperatures itself is not easily modeled, and fossil fuels do not explain much of the change we see in surface temperatures over time, we may have trouble producing an accurate model.

## **1.5 Primary User Story**

Sally exclaims that it is hot today and wonders if the current temperature is normal for this time of year. She opens the app to see that the temperature for today is normal by the historical trend that climate change has predicted. This also shows that temperatures will continue to rise over the next five years and extreme temperatures will be even higher into the future. She is also able to see how driving or purchase patterns affect temperature on a large-scale. We hope that she uses this information to change her decisions to ones that help slow down the rapid change of the climate.

## **1.6 Solution Space**

Our program will help users understand climate trends from an objective, unbiased point of view. It will provide users with a better understanding of global climate change and how their lifestyle impacts it. Research suggests that “going green” elicits positive emotions; users will be able to derive satisfaction by making actionable decisions towards environmental improvement (Venhoeven et al). Our goal is to help users understand climate change from a non-political point of view and to make informed lifestyle changes that align with personal environmental goals and understandings.

## **1.7 Product Vision**

Our product vision is to take the results and findings, and turn them into useful information for the general public. As our title alludes to, how can we make climate change personal? One question we considered while contemplating our product vision was, how can we make the information easily accessible? We considered two scenarios that would make this information to the public accessible, both of which bring the information to the general public via an application on your phone, tablet, or computer.

Our first scenario is to create an application that provides suggestions to the general public on how to reduce their own fossil fuel consumption. In addition, we would provide information on how their current use of fossil fuels is impacting surface temperature trends and how any changes made may impact the surface temperatures if these changes were to be adopted broadly amongst the general public.

Our second scenario for our product vision is to add an interaction to a weather application that shows how localized temperature measurements compare to previous values in the same area. From there, we would want to create a graphical representation for the public to visualize the distribution of expected temperatures.

## **1.8 Definitions**

* Climate Change - As described by the United Nations, Climate Change can be a natural process where temperature, rainfall, wind, and other elements vary over decades or more. Refers to long-term shifts in temperatures and weather patterns ([What Is Climate Change? | United Nations](https://www.un.org/en/climatechange/what-is-climate-change))
* Surface Temperature - According to NASA, surface temperature is how hot the “surface” of the Earth would feel to the touch in a particular location ([Land Surface Temperature (nasa.gov)](https://earthobservatory.nasa.gov/global-maps/MOD_LSTD_M)).
* Fossil Fuels - According to National Geographic, fossil fuels are made from decomposing plants and animals. These fuels are found in the Earth’s crust and contain carbon and hydrogen, which can be burned for energy ([Fossil Fuels | National Geographic Society](https://www.nationalgeographic.org/encyclopedia/fossil-fuels/)).
* Greenhouse Gas - According to the EPA, greenhouse gases are gases that trap heat in the atmosphere. Examples include carbon dioxide (CO2), methane (CH4), nitrous oxide (N20), and fluorinated gases ([Overview of Greenhouse Gases | US EPA](https://www.epa.gov/ghgemissions/overview-greenhouse-gases)).

# 

# **2 Data Acquisition**

## **2.1 Overview**

The two main datasets we will use for this project are *Climate Change: Earth Surface Temperature Data* and *Consumption in the World of Fossil Fuel*; both data packages are from Kaggle.com. The *Climate Change* files are repackaged datasets from over 16 archives collected and synthesized by Berkeley Earth. It includes 1.6 billion observations from NOAA’s MLOST, NASA’s GISTEMP, and the UK’s HadCrut ([Climate Change: Earth Surface Temperature Data | Kaggle](https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data?select=GlobalTemperatures.csv)). The main file we will use for temperature change analysis, *GlobalLandTemperaturesByMajorCity.csv*, ranges from the years 1750 to 2015 and contains observations from 3,448 different cities. The *Fossil Fuel* data set is sourced from BP’s Statistical Review of World Energy; it ranges from 2019 and contains measured changes in coal, oil, and gas usage from 79 countries ([Annual Change in Fossil Fuel Consumption | OurWorldInData.com](https://ourworldindata.org/grapher/annual-change-fossil-fuels)).

## **2.2 Field Descriptions**

*Climate Change: Earth Surface Temperature Data*

* Date (Type: datetime) – The year, month, and day of record observations as parsed by Berkeley Earth. It is in YYYY-MM-DD format, with the day for all observations being “01”. This field is recorded as an object datatype by Berkeley Earth. To better process the data, it is parsed as a datetime object upon import to Python for this project. This field should not be null. The data set contains 239,177 observations.
* AverageTemperature (Type: float32) – The global average land temperature in celsius. This field may be null; this dataset contains 11,002 null values.
* AverageTemperatureUncertainty (Type: float32) – the 95% confidence interval around the average. This field may be null; this dataset contains 11,002 null values.
* City (Type: object) – The city in which the observation was recorded. Cities are sourced from data obtained from NASA, NOAA, and HadCrut as procured by Berkeley Earth. This field cannot be null.
* Country (Type: object) – The country in which the observation was recorded. This field cannot be null.
* Latitude (Type: int32) – A major city’s location north or south of the equator in degrees. This cannot be null.
* Longitude (Type: int32) – A major city’s location west or east of the prime meridian in degrees. This cannot be null.

*Consumption in the World of Fossil Fuel*

* Entity (Type: object) – The state or country in which an observation was made. It cannot be null.
* Code (Type: object) – 3-letter code that identifies the state or country.
* Year (Type: int64) – Recorded as an object from the Kaggle data source. This field is converted to a date data type after importing to Python for analysis purposes. The data ranges from 1966 to 2019.
* Fossil.fuels..TWh.growth ... sub.method. (Type: float64) – Annual change in all fossil fuels as recorded by BP. Measured in tonnes.
* Coal..TWh.growth ... sub.method. (Type: float64) – Annual change in coal, measured in terawatt hours, as recorded by BP.
* Oil..TWh.growth ... sub.method. (Type: float64) – Annual change in oil measured in terawatt hours, as recorded by BP.
* Gas..TWh.growth ... sub.method. (Type: float64) – Annual change in gas measured in terawatt hours, as recorded by BP.
* region: Continent (Type: object) – The continent on which the observation was recorded. The five regions in this data set are: 'Asia', 'Americas', 'Oceania', 'Europe', and 'Africa'.
* sub.region: Area of the continent (Type: object) – A geographically specific location of the observed continent. BP defines the 14 sub-regions as: 'Western Asia', 'Latin America and the Caribbean', 'Australia and New Zealand', 'Western Europe', 'Southern Asia', 'Eastern Europe', 'Northern America', 'Eastern Asia', 'Northern Europe', 'Northern Africa', 'Southern Europe', 'South-eastern Asia', 'Central Asia', and 'Sub-Saharan Africa'.

## 

## 

## 

## **2.3 Data Context**

*Climate Change: Earth Surface Temperature Data*

This dataset is a synthesized repackaging of over 1.6 billion temperature observations from 16 different archives. Data is collected from temperature observation programs operated by National Oceanic and Atmospheric Administration (NOAA), National Aeronautics and Space Administration (NASA), and HadCRUT (Hadley Centre/Climatic Research Unit Temperature) in the UK ([Climate Change: Earth Surface Temperature Data | Kaggle](https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data?select=GlobalTemperatures.csv)). Because the data has a span of almost 300 years, we can expect changes in technology to play a factor in observations. We see a decrease in confidence intervals as time progresses; we also notice some countries lack temperature observations, perhaps due to late establishment, lack of historic records, technological changes, or weather center relocation.

*Consumption in the World of Fossil Fuel*

This data set contains observations of coal, gas, and oil from 5 different regions, dating from 1966 to 2019. Sourced from Kaggle, the primary data source comes from BP’s annual energy review, which was first published in 1965. BP gathers its data from dozens of sources and states its data is from government sources and published materials ([Statistical Review of World Energy | bp.com](https://www.bp.com/en/global/corporate/energy-economics/statistical-review-of-world-energy/using-the-review/links-to-the-contributors.html)). While data exists before these dates (notably from the onset of the Industrial Revolution when we begin to see fossil fuels become the primary source of energy), the latter half of the 20th century sees a sharp increase in global use of fossil fuels ([Global Fossil Fuel Consumption | Our World In Data](https://ourworldindata.org/fossil-fuels)). This is where BP becomes the primary source for fossil fuel data acquisition, and where we focus our analysis; the contemporary shift from coal to gas and oil will allow for more relevant energy suggestions to users.

## **2.4 Data Conditioning**

All data sets are downloaded from the sources and uploaded into a shared Google Drive. From there, they are imported into Python and read as .csv files.

**2.5 Data Quality Assessment**

*Global Land Temperatures*

* Completeness: There are approximately 4.8% null values in the data. This occurs in the AverageTemperature and AverageTemperatureUncertainy fields. This notably occurs in large time gaps within a few countries. Berkeley Earth notes that missing data could be a result of several factors, including technological changes or weather station relocation.
* Uniqueness: This data contains no duplicate observations.
* Accuracy: The accuracy of the data increases over time, which is evident as the confidence intervals decrease when technology changes and improves. Berkeley Earth notes: “Early data was collected by technicians using mercury thermometers, where any variation in the visit time impacted measurements… In the 1980s, there was a move to electronic thermometers that are said to have a cooling bias” ([Climate Change: Earth Surface Temperature Data | Kaggle](https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data?select=GlobalTemperatures.csv)).
* Atomicity: Given the nature of the data, we do not feel there are issues with atomicity.
* Conformity: The data is taken using objective, standardized measurements. It conforms to industry standards.
* Overall Quality: While there are missing observations, the data is of very good quality overall. We believe the missing data can be accounted for in a way that will not adversely affect predictions and we are currently gathering more information to appropriately process the missing data.

*Fossil Fuel Changes*

* Completeness: This data set has no missing values. While the data does not encompass trends from 1800 to the 1960s, it is capable of reflecting current trends given that coal will likely not reemerge as a primary energy source ([Coal Demand Has Seen its Biggest Drop | World Economic Forum](https://www.weforum.org/agenda/2021/01/coal-demand-asia-decarbonize-emissions/#:~:text=A%20forecasted%20rebound%20in%202021,to%20remain%20stable%20till%202025.)).
* Uniqueness: This data set does not have duplicate observations.
* Accuracy: This dataset is sourced from BP, which gathers its statistics from government sources and published data ([Statistical Review of World Energy | bp.com](https://www.bp.com/en/global/corporate/energy-economics/statistical-review-of-world-energy/using-the-review/links-to-the-contributors.html)). It can be considered reviewed and accurate.
* Atomicity: Given the nature of the data, we do not feel there are issues with atomicity.
* Conformity: BP has subject matter expertise in measurements and conversions of energy measurements. The measurements of the observed fossil fuels conform to industry standards.
* Overall Quality: The quality of this data can be considered excellent. It meets all assessment points.

## **2.6 Other Data Sources**

Other data sets that were considered from Berkeley Earth, include:

* Global Average Land Temperature by Country (GlobalLandTemperaturesByCountry.csv)
* Global Average Land Temperature by State (GlobalLandTemperaturesByState.csv)
* Global Land Temperatures By Major City (GlobalLandTemperaturesByMajorCity.csv)

These data sets were not used because our current data set provides a more granular look at the data and will allow our project to make more localized projections and suggestions.

# **3 Analytics and Algorithms**

## **3.1 Algorithm Descriptions**

There are two problems that are being addressed with the algorithms. The first problem (Problem 1) is to determine the relationship between fossil fuel use and average temperature. The second problem (Problem 2) is to predict the average temperature for a given location. The fossil fuel data for the first problem does not have monthly data and therefore we are limited to finding the temperature data on a yearly basis. The fossil fuel data set suffers from highly correlated predictor variables as population, coal, oil, and gas are all increasing steadily over time. This relationship appears to be linear. The predictors for the second problem (month, year, latitude, and longitude) likely have a nonlinear relationship to average temperature and therefore it is likely that a non-linear model will provide the best results.

For both problems, we tested a number of algorithms and focused on the algorithm that provided the best initial results. For Problem 1, this was Ridge Regression; for Problem 2, this was an Artificial Neural Network.

## **3.2.1 Ridge regression**

Ridge regression is an implementation of linear regression used to resolve multicollinearity in feature variables. It does this by imposing a penalty on beta coefficients; this is called the L2 penalty. The L2 penalty minimizes coefficient sizes while also preventing the values from becoming zero (which would result in them being removed by the model). A hyperparameter, lambda, “controls the weighting of the penalty to the loss function” [(“How to Develop Ridge Regression Models in Python” | John Brownlee)](https://machinelearningmastery.com/ridge-regression-with-python/). This allows stability in responses between feature and target variables.

*Algorithm Application*

Ridge regression is available in Python’s sci-kitlearn and deployed via the Ridge class. Our utilization of ridge regression results in a code size of approximately 100 lines and is relatively easy to implement. Github file sharing is used for version control between the team members.

## **3.2.2 Artificial Neural Networks (ANN) - Keras**

Keras is an open-source API built by Google. It is built on TensorFlow and used as an interface for creating artificial neural networks ([About Keras | keras.io](https://keras.io/about/)). ANN is a process by which all data is stored in a network (TensorFlow, in our case) and patterns observed by the network are stored as vectors. Essentially a process of elimination, vectors are passed through layered nodes (which can be described generally as individual linear regression models). If the output coefficients meet the criteria of the model, the data is passed to the next node and so on until the defined number of nodes is reached and the output is generated.

*Algorithm Application*

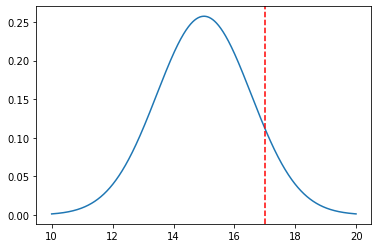
Keras is available via pip install. It requires additional installation and importation of SciPy and a TensorFlow. Our implementation of Keras results in a code size of approximately 100 lines and can be considered mid-level difficulty; it requires some understanding of neural networks and the use of a variety of libraries to support the Keras functions. Github file sharing is used for version control between the team members.

# **4 Visualizations**

*Probability Distribution Visualization*

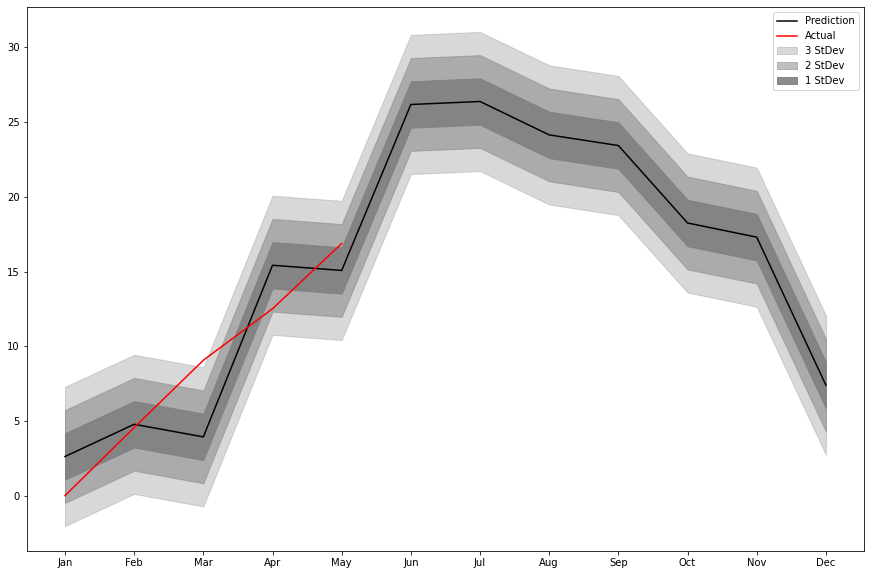
The probability distribution function allows for a visualization of the distribution of predicted data based on historical data. We apply this to our program by using historical temperatures to predict the average temperature. This graph enables a better visual understanding of what could be considered a “normal” temperature on any given day.

The probability distribution visual will be implemented using Python. This visualization requires approximately 20 lines of code and is of medium complexity. It will be tested by various team members and version control will be maintained using Github.



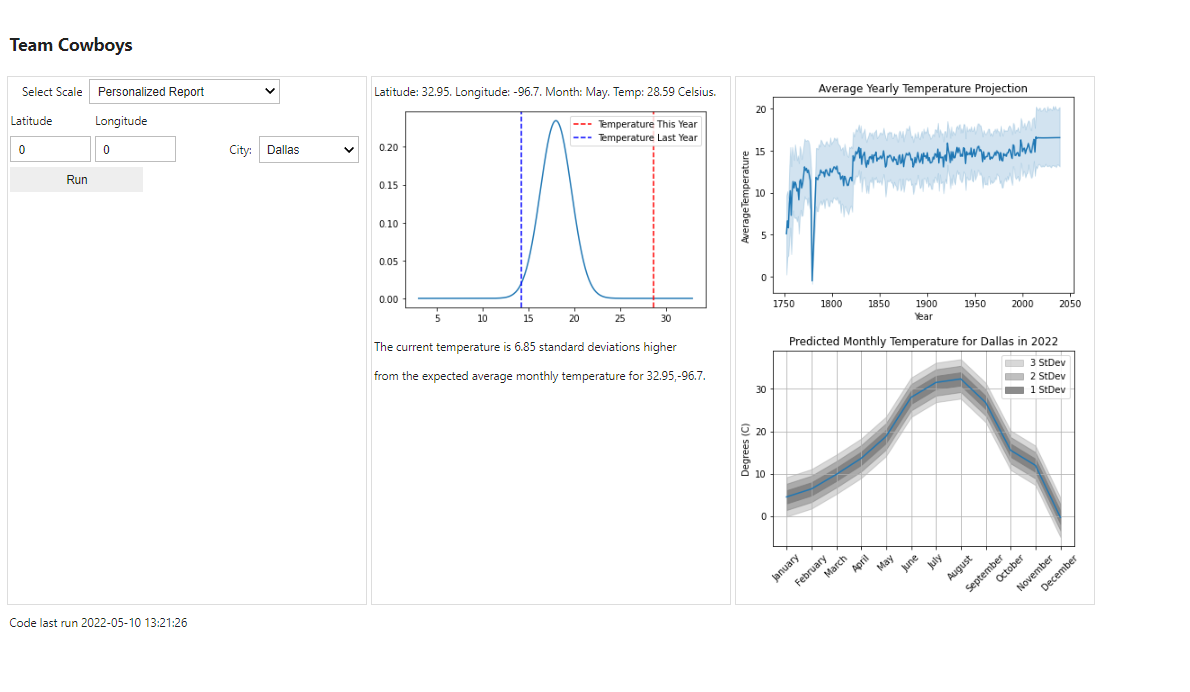
*Line Graph*

The probability distribution function allows for a visualization of the distribution of predicted data based on historical data. The prediction uses the Artificial Neural Network described in the previous section. After training the model and testing on a section of the data, the test error rate was used to estimate the standard deviation of the error. The visualization is created by predicting the average temperature for each month of 2022. The location for Fairfax, VA was used as the example. The central line for the average data was plotted using this data from the trained neural network. The banding was created by adding one, two and three standard deviations from the central line. The actual data was plotted over this in a red line. This data was acquired from the Weather Underground historical search capabilities. The first five months of this year were manually requested and entered into the program. This visual allows for a better visual understanding of what could be considered a “normal” temperature on any given day. It was implemented using Python using Matplotlib and it requires approximately 30 lines of code of low complexity. It will be tested by various team members and version control will be maintained using Github.

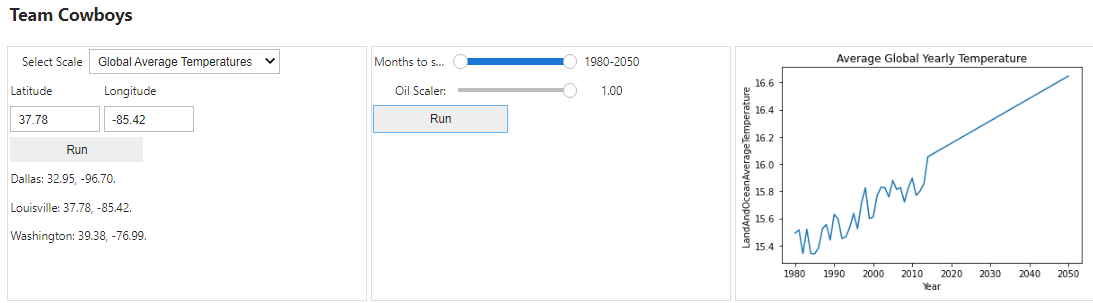


*Application Visuals*

The application uses four visualizations grouped into two categories: Personalized Report and Global Average. The Personalized report provides results to the user for their location. Once a city is selected, three graphs are provided for that location. The expected temperature chart provides a normal distribution of predicted temperatures for a location. The average temperatures for this year and last year are also provided, placing them in relation to the normal distribution of temperatures.



The second chart Average Year Temperature Prediction provides the historical temperature predictions and future prediction for location. These predictions are based on location and may not be one the locations in the data provided. The third chart, Predicted Monthly Temperatures in 2022, shows the predicted temperatures, bands that indicate one, two and three standard deviations and the actual temperatures for the first five months.



Each of these charts is intended to show current temperatures in relation to the predicted temperatures and allow users to investigate data on their own. The Global Average group only contains a single graph. The sliders to the left of the graph allow the user to modify the graph and investigate the relationship between average temperature and oil consumption. The months to show slider allows a user to select the years they want to view to zoom in on particular years. The oil scaler slider allows the user to scale the oil consumption estimates and observe the impact on the average temperature. Once again, the intention is to allow the user to interact with the data and predictions.

# **5 Findings**

The Ridge Regression model was a moderately strong predictor of Average Temperature using fossil fuel data for oil, coal and gas and population as the predictor. The R Squared value was 0.653 and the MSE was 0.038. There were two difficulties encountered with this data. First is that population, oil consumption, coal consumption and gas consumption were all collinear. While this does not diminish the predictive value of the model, the coefficients are not reliable. Second, while population, oil consumption, coal consumption and gas consumption account for about 65% of the variance in Average Temperature, other data is needed to improve the predictive ability of the model.

In order to predict future global temperatures, the population and fossil fuels needed to be estimated into the future. We used a linear model to project the data into the future and this resulted in a linear projection of the temperature. When comparing the model results to other predictions, the Ridge Regression model predicted values that were between the high and low predictions by the Center for Science Education. (“Predictions of Future Global Climate | Center for Science Education”). These models likely take into account a rising fossil fuel consumption rate and a decreasing fossil fuel consumption rate placing these predictions above and below our predictions.

The Artificial Neural Network model was a strong predictor using year, latitude, longitude and month to predict Average Temperature. There were a large number of observations with 5,376,962 used for training and 2,648,355 test values. Using the model to predict the test values the R Squared value was 0.977 and the MSE was 2.33. The MSE seems high but for temperature that we all know can vary greatly day to day this is quite reasonable. The model seemed to be accurate for locations that were outside the cities included in the original dataset.

While the data to train this model was freely available, the location based temperature data could only be provided by paid services. An educational license was able to be used to gather the data for prediction. Normally this would cost $170 per month with a limit of 40,000 requests. Creating a public application would easily exceed the 40,000 requests and likely be far more expensive to operate.

The most important finding is that making data relevant and accessible is possible. While more work is needed on predictions and the user interface, adding this information to a commonly used application such as weather would allow everyday users to use and interact with the data. The underlying data and models can be investigated and tested by those that feel the need to verify the veracity of the visualizations.

# **6 Summary**

Overall, our data confirms the multidimensional nature of temperature change; we conclude that more research is needed on the direct impacts of technological and social changes in fossil fuel usage. Our model is able to account for and predict temperature change using global oil, gas, and coal usage with relative accuracy. When comparing our model to UCAR’s study that uses both positive- and negative- outlook predictive models, our model falls in the middle of both lines, sitting more towards the positive outlook prediction ([“Predictions of Future Global Climate” | UCAR](https://scied.ucar.edu/learning-zone/climate-change-impacts/predictions-future-global-climate)) It is simple to predict the average temperature using oil alone (as it is the most impactful). But the collinear nature of fossil fuel usage and population, or other variables, makes predicting the distant future of temperature change an uncertain task when accounting for all the variables that *do* impact temperature change. That being said, our data demonstrates that fossil fuel does have a linear relationship with data; therefore, change even at a granular level, can aid in a positive impact on climate change in terms of temperature. It is our view that individuals can start by making small changes that might ultimately foster larger changes that will have globally beneficial outcomes.

# **7 Future Work**

There is a lot of future work that can be done in regards to our project and climate change in general. With it being such a popular topic in the world today, there is always more you can dive into. For our project, we came up with these ideas for future work:

* Consider other variables that contribute to climate change such as CO2 production, other greenhouse gasses, reflectivity or absorption of the Sun’s energy, changes in the Earth’s orbit and rotation, volcanic activity, and so on ([Causes of Climate Change | US EPA](https://www.epa.gov/climatechange-science/causes-climate-change)).
* Utilize the real time data more.
* Working or refining the coefficients for our model. While our model is sound, there is a lot of collinear data and would need more time to understand how all the variables impact each other.
* Get a full set of cities/locations to add to our App that users can use to look up information in specific areas, just as a weather app has the ability to show you the weather of any city you could probably think of.

# 

# **Appendix A - Code References**

<https://github.com/rmaxseiner/TeamCowboys>

Please see readme file for more information.

# 

# **Appendix B - Risk Section**

*3.2.1 Risks Associated with the Algorithm - Ridge Regression*

Because our dataset features multicollinearity (linear correlation between multiple independent variables), small changes in one regression coefficient can result in large changes in another. Multicollinearity produces the challenge of detecting statistical significance. This results in less accurate predictions ([“Understanding Multicollinearity and How to Detect it in Python” | Terence Shin](https://towardsdatascience.com/everything-you-need-to-know-about-multicollinearity-2f21f082d6dc)). We are mitigating the risks of multicollinearity by using ridge regression and customizing lambda in our model. This strategy is currently implemented in our codebase.

*3.2.2 Risks Associated with the Algorithm - Keras*

Because of the nature of ANNs, the algorithm and its predictions are generally unexplainable. After training the model and reviewing the output, there is a general lack of information as to *how* the output arrived. There is also some trial and error to determine the ideal number of nodes and the appropriate architecture to procure the most accurate predictions. There is a high risk of general obscurity when it comes to a deeper understanding of exactly what is going on in the hidden layer of the ANN. In order to ensure the Neural Network is statistically significant absent the ability to interpret the results, the data has been split into train and test data. Given our large number of records, both the train and test data sets have sufficient data. Measuring the effectiveness of the Neural Network allows us to ensure that the model is not overfitting and has good predictive capabilities. An additional mitigation strategy includes implementing the Keras Tuner library; this library helps select the optimal set of hyperparameters for the program ([“Introduction to the Keras Tuner” | TensorFlow.org](https://www.tensorflow.org/tutorials/keras/keras_tuner#:~:text=The%20Keras%20Tuner%20is%20a,called%20hyperparameter%20tuning%20or%20hypertuning.)) and provides more insight into the output. While the ANN may require additional tuning, the manual tuning performed to this point has provided an R Squared value of over 0.97 and an MSE of 2.3. These are both reasonable given our dataset. Additional tuning seeks to improve those even further.

*4 Risks Associated with the Visualizations*

The statistical nature of the visuals puts the program at risk for a lack of understanding by some users. Terms like “one standard deviation” will hold little meaning for users unfamiliar with statistical jargon. There is a moderate risk of this occurring. In the event that a user does not understand the output, the result is likely confusion and limited usefulness of the data to the user. One way to mitigate this is to supplement the visualizations with understandable notes and relevant information; wording standard deviations in terms of percentages will minimize confusion of temperature distributions and make the data more impactful to a broad range of users and.

Our data is also subject to paywall, putting it at risk for limited access thereby impacting predictive visuals. Temperature data from 2013 and earlier was obtained from openweathermap.org, which requires a monthly $175 subscription to view historic data beyond five days. Our team was able to secure a free subscription based on student status; however, this free tier is limited to 40,000 requests and will end May 2023.

# **Appendix C - Agile Development**

The agile development process has different steps depending on who you talk to: meet, plan, design, develop, test, and evaluate. For the most part, we followed this process for each of our Sprints to ensure we stayed up to date on each task. The below screenshot shows the process we took when executing our weekly tasks. Each week we met once or twice, typically on Monday and Thursday. During our Monday meetings we would discuss what is needed to be done that week and who will work on particular parts of the project, this would be included in the meet and plan steps of agile development. In between our Monday and Thursday meetings, we would each work on specific tasks. This part of our week would be considered the design, develop, and test steps. During our Thursday meetings we discussed what each member had worked on and prepared our powerpoint for the upcoming presentation on Saturday. Our Thursday meetings would be considered the evaluation phase of agile development which gave us time between our final meeting and our presentation to finalize and refine the work done that week. During the two week sprints, we typically added in a third meeting with them being on both Monday’s of the two week sprint and the Thursday before we presented. After our presentations on Saturdays, we discussed on Slack what needed to be finished for our submission Sunday evening.

# 

# 

# **Appendix D - References**

United Nations. (n.d.). *Climate Action Fast Facts.* [Key Findings | United Nations](https://www.un.org/en/climatechange/science/key-findings#physical-science.).

Venhoeven, Leonie A., et al. “Why Going Green Feels Good.” *Journal of Environmental*

*Psychology*, vol. 71, 2020, p. 1. [Why going green feels good - ScienceDirect](https://www.sciencedirect.com/science/article/pii/S0272494418307801?via%3Dihub)

Berkley Earth. (n.d.). *Climate Change: Earth Surface Temperature.* [Climate Change: Earth Surface Temperature Data | Kaggle](https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data?select=GlobalTemperatures.csv)

Kaggle Contributor. (n.d.). *Consumption in the World of Fossil Fuel.* [Consumption in the world of fossil fuel | Kaggle](https://www.kaggle.com/code/gianlab/consumption-in-the-world-of-fossil-fuel)

National Geographic. (n.d.). *Resource Library Encyclopedic Entry: Fossil Fuels.* [Fossil Fuels | National Geographic Society](https://www.nationalgeographic.org/encyclopedia/fossil-fuels/)

National Aeronautics and Space Administration (NASA). (n.d.). *Land Surface Temperature.* [Land Surface Temperature (nasa.gov)](https://earthobservatory.nasa.gov/global-maps/MOD_LSTD_M)

United Nations. (n.d.). *What is Climate Change?.* [What Is Climate Change? | United Nations](https://www.un.org/en/climatechange/what-is-climate-change)

Shin, Terence. *Understanding Multicollinearity and How to Detect It in Python*. Towards Data Science, 4 Dec. 2021, [Understanding Multicollinearity and How to Detect it in Python | by Terence Shin | Towards Data Science](https://towardsdatascience.com/everything-you-need-to-know-about-multicollinearity-2f21f082d6dc)

“Predictions of Future Global Climate Change .” *Center for Science Education*, UCAR, [Predictions of Future Global Climate | Center for Science Education (ucar.edu)](https://scied.ucar.edu/learning-zone/climate-change-impacts/predictions-future-global-climate)

“Links to the Contributors: Energy Economics.” *Bp Global*, [Links to the contributors | Energy economics | Home (bp.com)](https://www.bp.com/en/global/corporate/energy-economics/statistical-review-of-world-energy/using-the-review/links-to-the-contributors.html)

Wood, Johnny. “Coal Demand Has Seen Its Biggest Drop since World War II. But It's Not All Good News.” *World Economic Forum*, 6 Jan. 2021, [Coal use is declining in the US and Europe - but not in Asia | World Economic Forum (weforum.org)](https://www.weforum.org/agenda/2021/01/coal-demand-asia-decarbonize-emissions/#:~:text=A%20forecasted%20rebound%20in%202021,to%20remain%20stable%20till%202025.)

Ritchie, Hannah, and Max Roser. “Fossil Fuels.” *Our World in Data*, 28 Nov. 2020, [Fossil Fuels - Our World in Data](https://ourworldindata.org/fossil-fuels)

“Annual Change in Fossil Fuel Consumption 2020.” *Our World in Data*, BP Statistical Review of World Energy, [Annual change in fossil fuel consumption, 2020 (ourworldindata.org)](https://ourworldindata.org/grapher/annual-change-fossil-fuels)

“Overview of Greenhouse Gases.” *EPA*, Environmental Protection Agency, [Overview of Greenhouse Gases | US EPA](https://www.epa.gov/ghgemissions/overview-greenhouse-gases)

“Land Surface Temperature.” *NASA*, NASA, [Land Surface Temperature (nasa.gov)](https://earthobservatory.nasa.gov/global-maps/MOD_LSTD_M)

“Causes of Climate Change.” *EPA*, Environmental Protection Agency, [Causes of Climate Change | US EPA](https://www.epa.gov/climatechange-science/causes-climate-change)

Max Kuhn (2021). caret: Classification and Regression Training. R package version 6.0-90. [CRAN - Package caret (r-project.org)](https://cran.r-project.org/web/packages/caret/index.html)

“Introduction to the Keras Tuner .” *TensorFlow*, [Introduction to the Keras Tuner | TensorFlow Core](https://www.tensorflow.org/tutorials/keras/keras_tuner#:~:text=The%20Keras%20Tuner%20is%20a,called%20hyperparameter%20tuning%20or%20hypertuning)

Team, Keras. “Keras Documentation: About Keras.” *Keras*, [About Keras](https://keras.io/about/)

Brownlee, Jason. “How to Develop Ridge Regression Models in Python.” *Machine Learning Mastery*, 9 Oct. 2020, [How to Develop Ridge Regression Models in Python (machinelearningmastery.com)](https://machinelearningmastery.com/ridge-regression-with-python/)