Extra Credit: Climate Change Dataset

Colin Kwiecinski

12/1/2019

Introduction

This dataset is a time series data set of average global land and ocean temperatures. It includes columns for average land temperature, max and min land temperature, and average land and ocean temperature, as well as confidence interval columns for each of these variables. For simplicity sake, we will not be utilizing the confidence intervals in this exploration of the data.

The data comprises 3192 observations, starting in 1750 and running up to the end of 2015. The data was collected by weather and meteorology organizations like NOAA's MLOST and NASA's GISTEMP, and then further organized and compiled by Berkley Earth, and uploaded to kaggle.com for public use. Berkley Earth drew from almost 2 billion temperature reports from multiple archives to create this dataset.

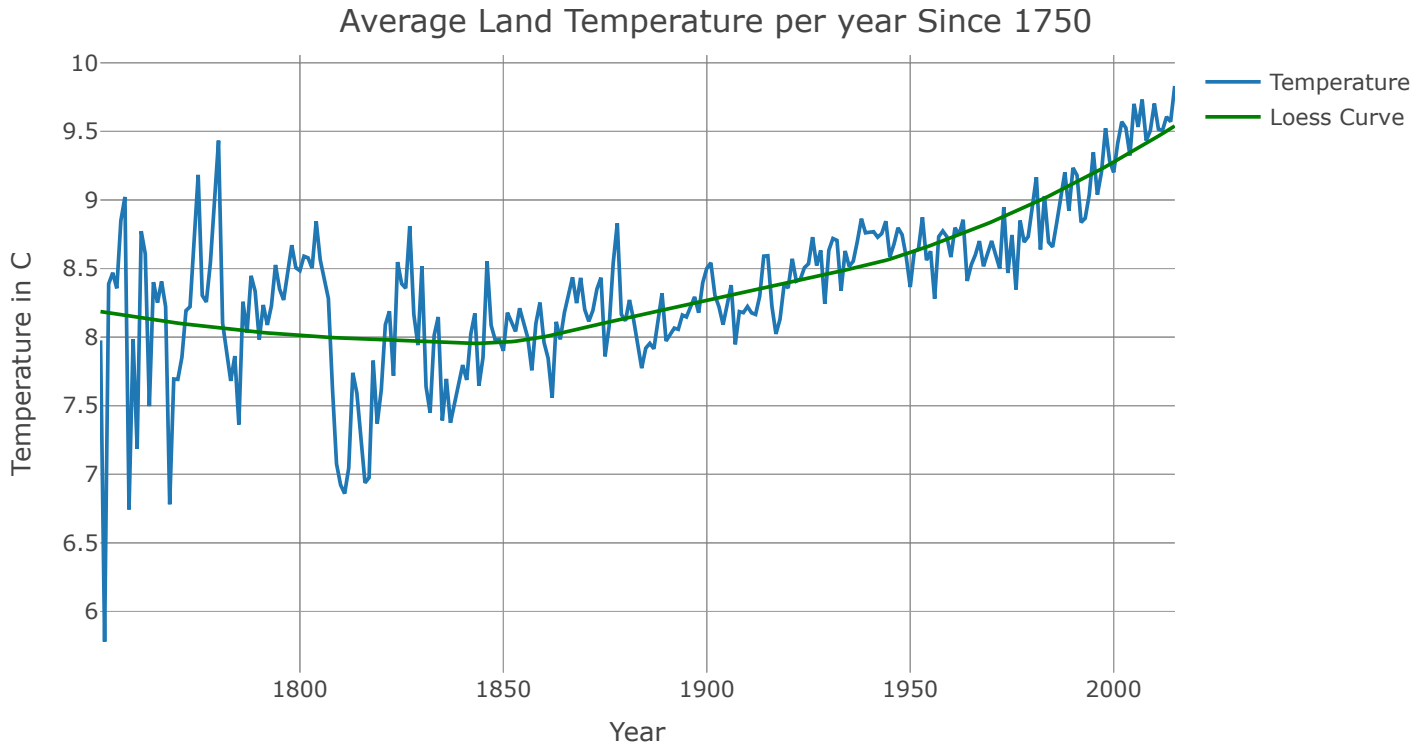
The information provided by this dataset will be very useful for examining trends in global temperature, so that we may see if global temperature has actually been rising, to confirm what others have already said about climate change.

Link to the dataset on Kaggle.com (https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data#GlobalTemperatures.csv)

Plots

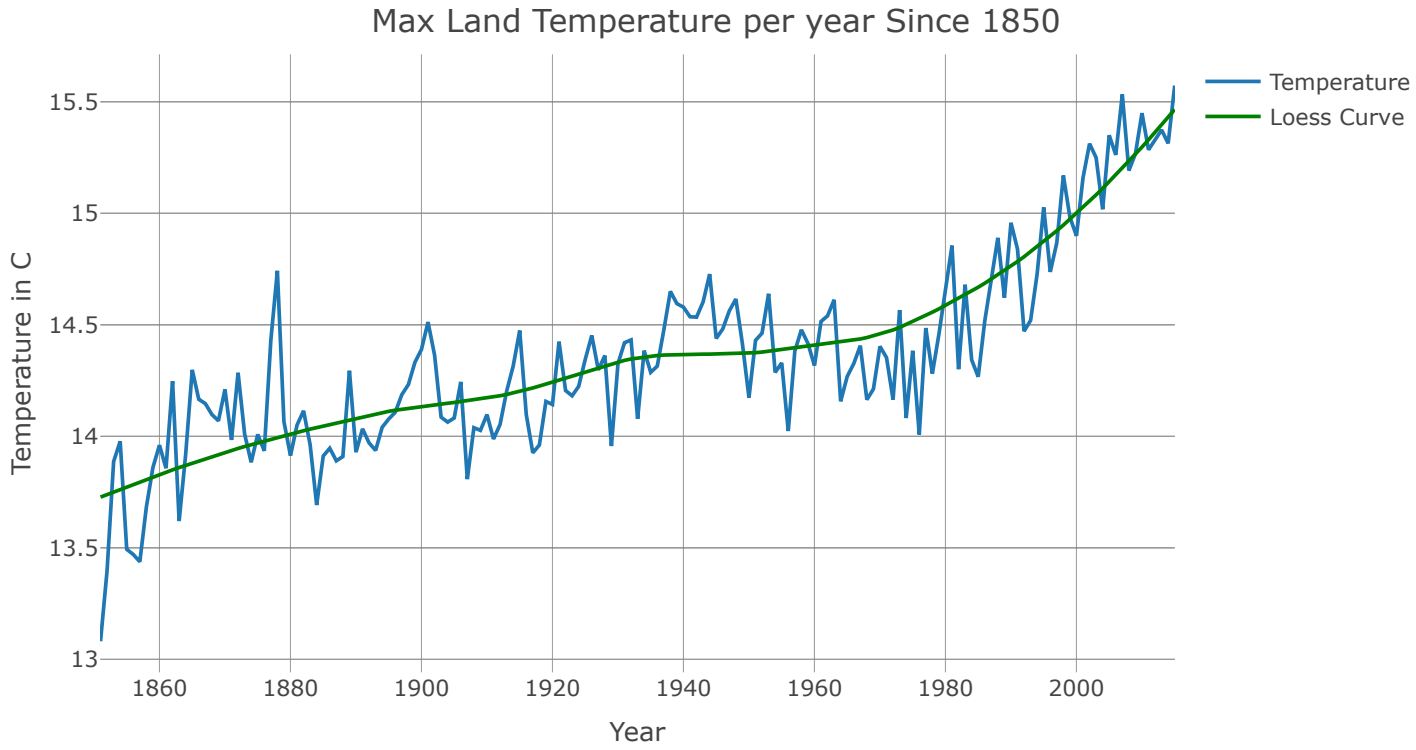
The following three plots have been creating by grouping observations into years and summarizing them to get the average for each year, and then plotting the temperature vs year and adding a loess fit curve to show the general trend of the temperature over the chosen time period.

The plots are also fully interactive using plotly, so the user may zoom in and control them as they see fit.

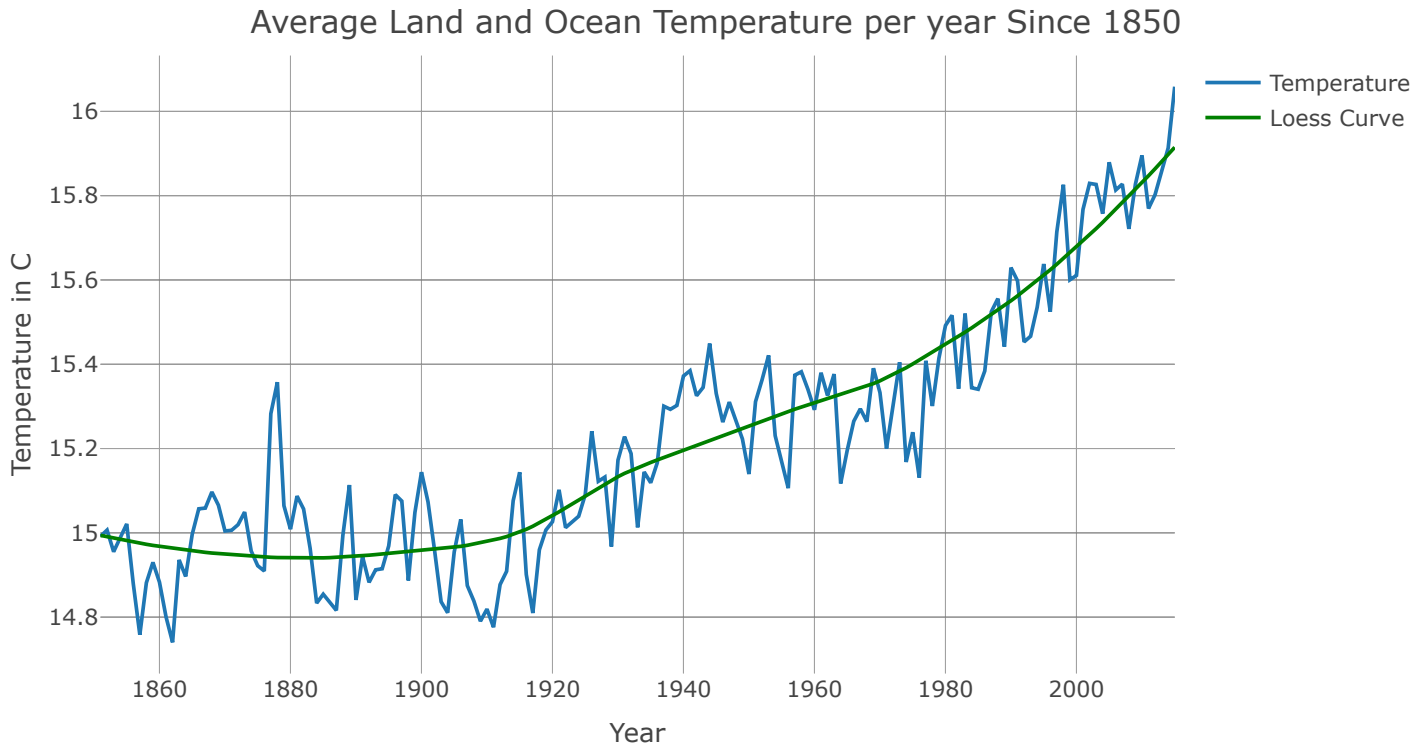


This plot shows the average land temperature recorded from 1750, and the data shows an upwards trend, meaning that the

average land temperatuer is hotter than it was in the past.



This plot shows the maximum recorded temperature on average per year since 1850. This data is tracked from 1850 instead of 1750 because this variable did not have sufficient data to be considered until 1850, at which point the accuracy and quantity of available data increased. Again, the trend shows an increase in temperature over time.



This plot shows the combined average of land and water temperature for each year starting in 1850. This data also starts in 1850 for the same reason as above, having insufficient data prior to 1850. This plot also depicts a general increase in temperature over time.

Discussion

Since this is a time series dataset, it allows us to look into the past to examine trends in temperature. We can then use this data to either confirm current observations and hypothesis, or to extrapolate and predict future trends. An example that fits both of these use cases would be to argue that climate change is happening and will continue to happen, and at an increased rate of increase.

Some stakeholders that may use this data would be cilmatologists who are trying to prove that global warming is real, and perhaps politicians (especially those who don't believe in climate change). Some indirect stakeholders would be the general population, as climate change affects everyone in the world, but they do not have policy making authority on what is done about it.

I conclude from my analysis of this data that over the past 200 years, the average temperature of the Earth has increased, and is increasing at a higher rate than in the past.

Reflection

From my analysis of this time series data I learned how to create visualizations that show a general trend over time, and how to add lines of fit that help to demonstrate the trend of the data. Even if the scatterplot version of the data looks messy and "scattered" everywhere, combining that data into a fit line can help make the trend of the data more apparent. After working with this data set, a further question I have is that I want to learn how to extrapolate using the data I have to make a prediction about what the future data may look like.

Code

Below is a sample of the code used in the analysis. After the data is initially read, the date variable is converted into a Date datatype, and the year column is added to allow for grouping by year. Then each plot is created by summarizing each year of data for the chosen variable.

To see the full code and dataset, head to the github page found here: Link to Github (https://github.com/ColinKwiecinski/climate_change_viz)

```r
library(dplyr)
library(plotly)
library(lintr)
library(lubridate)
library(knitr)

# Import data and reformat date information to be more usable.
main_df <- read.csv("GlobalTemperatures.csv", stringsAsFactors = FALSE)
colnames(main_df)[1] <- "date"
main_df$date <- as.Date(main_df$date, "%Y-%m-%d")
main_df <- main_df %>%
  mutate(year = year(date)) %>%
  group_by(year)

# Creates a plot of land temp vs time. Option to select certain start year
plot_landavgtemp <- function(df, startYear) {
  df <- df %>%
    summarise(temp = mean(LandAverageTemperature, na.rm = TRUE)) %>%
    filter(year > startYear)

  p <- plot_ly(
    data = df,
    x = df$year,
    y = df$temp,
    type = "scatter",
    mode = "lines",
    name = "Temperature"
  ) %>%
    add_lines(
      y = ~ fitted(loess(df$temp ~ df$year)),
      line = list(color = "green"),
      name = "Loess Curve"
    ) %>%
    layout(
      title = paste("Average Land Temperature per year Since", startYear),
      xaxis = list(title = "Year"),
      yaxis = list(title = "Temperature in C")
    )
  return(p)
}
avg_landtemp_line <- plot_landavgtemp(main_df, 1750)
```