Data 200 Project

Kalie Knecht

April 2021

1 Abstract

2 Introduction

2.1 Research Question

I'll be exploring countries which emit the most greenhouse gases vs which areas of the world are more affected by global warming. Can we build a model to determine temperature based on greenhouse gas emissions? Additionally I will take a look at how the nuclear energy capacity of each country correlates to their greenhouse gas output. I would also like to explore some radiation data in relation to this second question, which I will obtain from my research group's environmental monitoring stations.

I will be investigating this using the data provided by the Global Historical Climatology Network (GHCN), which contains the average daily temperature recorded at monitoring stations around the world. I will also be using the data provided by the Environmental Protection Agency (EPA) through their Greenhouse Gas Reporting System.

2.2 Previous Work

A brief survey of related work on the topic(s) of your analysis and how your project differs from or complements existing research. (TO DO)

3 Data

The data used in this project is Dataset 2A: Climate and the Environment - General Measurements and Statistics. This dataset contains some general statistics and measurements of various aspects of the climate and the environment. It includes the following reports:

- *daily_global_weather*: Daily global weather from GHCN from January to October 2020
- *greenhouse_gas_type* and *greenhouse_gas_facility*: greenhouse gas emissions data reported by EPA, detailing the specific types of gas reported by facilities and general information about the facilities themselves. The dataset is made available through EPA's GHGRP (Greenhouse Gas Reporting Program).
- *us_air_quality*: air quality on a county level from approximately 4000 monitoring stations around the United States, reported from EPA's Air Quality System (AQS)

I am also using data generated by my research group, Radwatch, from our DoseNet monitoring stations. The DoseNet stations each include a basic nuclear radiation counter which provides the dose-rate shown on our default map. Several locations now also include outside stations with an advanced gamma radiation spectrometer, a carbon dioxide monitor and particulate matter air quality sensor, and a pressure, temperature, humidity gauge.Our network begins in UC Berkeley and Lawrence Berkeley National Lab (LBNL), and stretches out to secondary schools across the Bay Area. This network also includes a growing number of satellite networks through our international partners.

Our basic dosimeters count radiation interactions that deposit energy in the device, giving the counts per minute (CPM) averaged over a 5-minute interval. The CPM is converted to a rate of radiation dose that people are being exposed to. Each dosimeter then takes this measurement and sends it to a central server, where it is displayed on RadWatch. This system allows for real-time monitoring of radiation levels.

Our monitoring stations are generally installed in local Bay Area high schools and various places around the world where we have research connections.

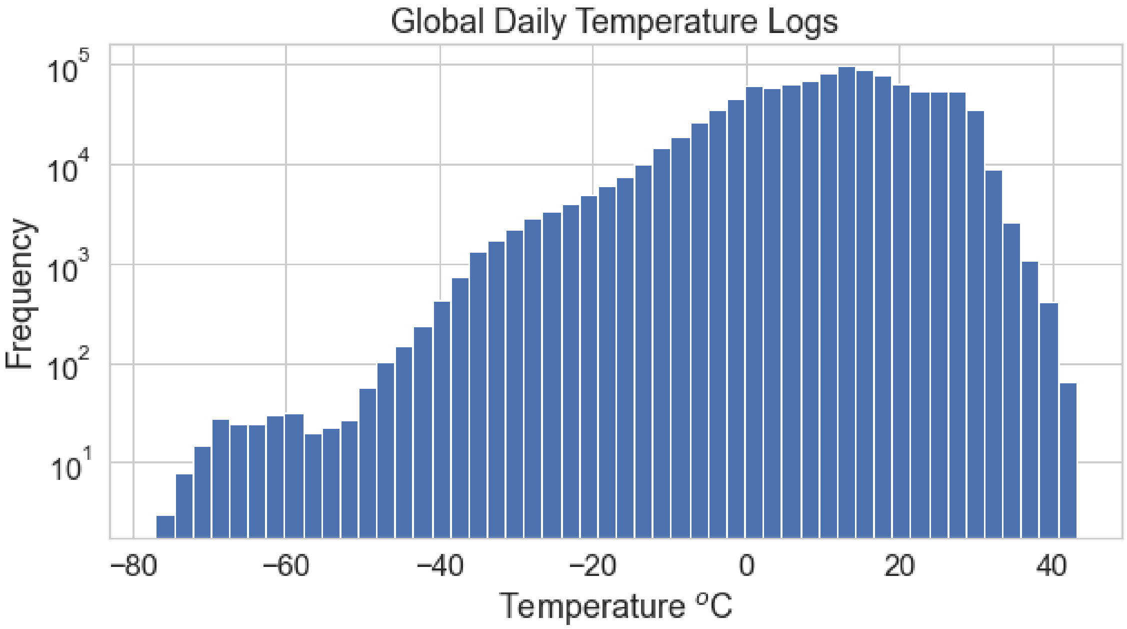

Figure 1: Histogram of daily temeprature logs for all of the stations generated in EDA section of Analysis notebook

4 Methods

Methodology: carefully describe the methods you use and why they are appropriate for answering your search questions. It must include

4.1 Casual Inference

a brief overview of causal inference, which should be written in a way such that another student in Data 100 who has never been exposed to the concept can carry out the analyses involving the datasets in your project.

4.2 Exploratory Data Analysis

4.3 Modeling

a detailed description of how modeling is done in your project, including inference or prediction methods used, feature engineering and regularization if applicable, and cross-validation or test data as appropriate for model selection and evaluation.

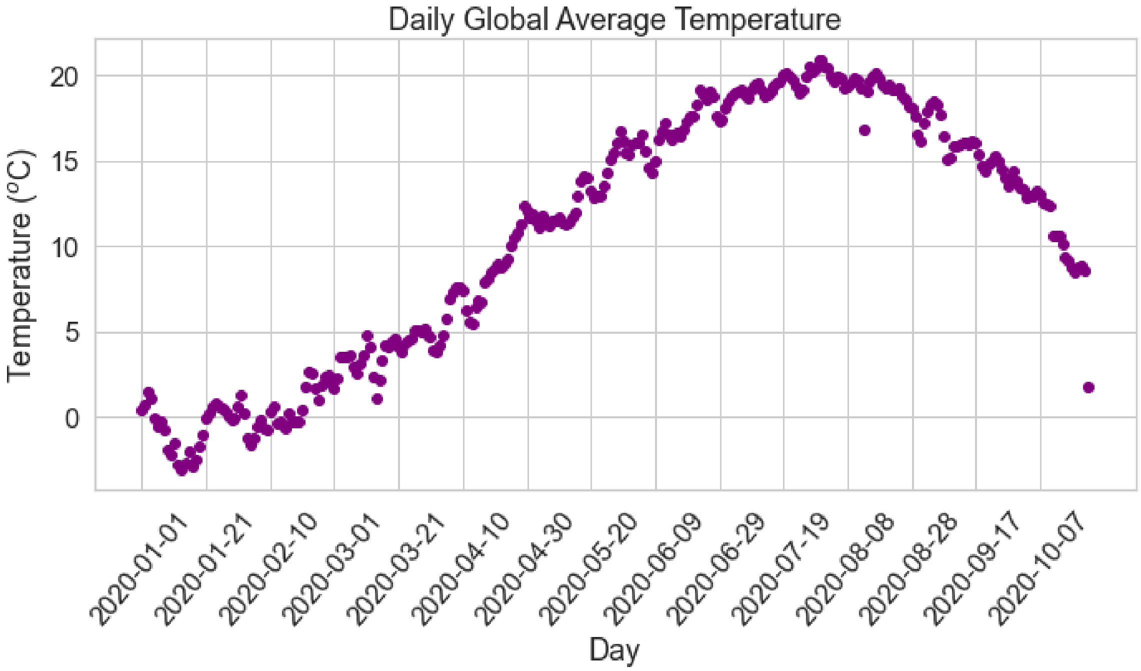

Figure 2: Daily Average Temperature generated in EDA section of Analysis notebook

5 Results

Interesting findings* about each dataset when analyzed individually. Include visualizations and descriptions of data cleaning and data transformation necessary to perform the analysis that led to your findings. Interesting findings* involving your datasets. Include visualizations and descriptions of data cleaning and data transformation necessary to perform the analysis that led to your findings.

* Examples of interesting findings: interesting data distributions and trends, correlations between different features, the relationship between the data distribution for the general population and specific datasets (e.g., the gender distribution in the census dataset vs. in the mental health dataset), specific features that are notably effective/ineffective for prediction.

6 Conclusions

Analysis of your findings to answer your research question(s). Include visualizations and specific results. If your research questions contain a modeling component, you must compare the results using different inference or prediction methods (e.g., linear regression, logistic regression, or classification and regression trees). Can you explain why some methods performed better than others? An evaluation of your approach and discuss any limitations of the methods you used. Describe any surprising discoveries that you made and future work.