ANLY-501
Project Assignment 1
October 7, 2019
Project Team: Ryan Callahan (rpc29), Ruchikaa Kanar (rrk47), Chris Fiaschetti (cf807), Masha Gubenko (mg1865)

Data Science Problem

Temperatures are increasing across the globe, and emissions from various sources are contributing to this increase. In this study, we are trying to determine whether increases in certain “global warming factors” are most indicative of, or at least most positively correlated with increases in temperature. To do so, we are comparing increasing in temperature and “factors” over time across different locations. Previous research has compared temperature changes across locations over time, as well as relative changes in “global warming factors” across locations over time, but only separately. This research has tended not to use each location as a single sample/trial and look simultaneously at the change of “global warming factors” and temperature across these locations over time to determine factors correlated especially tightly to temperature increases.

Analyses

“Climate Change Factors” - In this study, we look at the location-specific change over time in the quantity of elements linked to “global warming factors”. These factors include transportation emissions, forest coverage, population, cost of living, etc… Some of these statistics are direct, such as forest acreage; others are second-order, such as vehicle-registration data (i.e. increased vehicle registration tends to indicate more miles driven) and population (i.e. population increases lead to more public transport, POV, and HVAC usage).

Datasets

1. *State-level vehicle registration data by type (auto, bus, truck, motorcycle) -- 1995-2014*
    a. We can analyze the impact of number of vehicles on increasing temperatures by comparing across multiple states
    b. File: Vehicle_Registration_By_State_By_Year.csv
2. *Population by state -- 2000-2019*
    a. We can use this data set to better analyze the effects of “global warming factors” by accounting for population growth
    b. File: Population_by_State_by_Year_2000_2010.csv & Population_by_State_County_by_Year_2014_2018.csv
3. *Population by city -- 2000-2019*

a. We can use this data set to better analyze the effects of “global warming factors” by accounting for population growth
b. File: Population_by_City_by_Year.csv

4. *Transit ridership/usage by cities -- 2001-2017*
   a. We can study the effects of public transit on increasing temperatures by comparing areas with different amount of transit ridership
   b. File: Transit_Ridership_By_City_By_Year.csv
5. *U.S electric vehicle registration data -- 1999-2015*
   a. We can analyze the impact of growing number of electric vehicles on increasing temperatures
   b. File: EV_Registrations_by_Type_US_by_Year.csv
6. *Commuter Volume (includes categories by modes of transportation data by state)*
   a. We can use these data to further analyze the effects of public transportation on increasing temperatures
   b. File: Commuter Dataset.csv
7. *General carbon consumption*
   a. We will analyze the environmental impact of carbon consumption
   b. File: Emissions.csv
8. *Location-specific Forest Coverage*
   a. We will use these data to study the effects of “global warming factors” on forest acreage over time for various regions
   b. Files: Forest_Data.csv & Live_Trees_in_Timberland.csv & Timberland_Planted.csv & Urban_Land(2000-2010).csv & US_Fires.csv
9. *U.S. Ecological Footprint*
   a. Dataset consisting of US ecological footprint as the country overall from 1961-2016
   b. File: US_Ecological_Footprints.csv
10. *Cost of living in various U.S. urban locations*
    a. We can use this dataset to account for socio-economic factors when comparing environmental impacts of different urban areas within the U.S.
    b. File: CPI.csv
11. *Monthly temperature averages at 1000s of U.S. locations -- 1961-2019*
    a. We can look at the extent of the correlation between these temperatures and each of the “global warming factors” above over time in each location
    b. File: Temps.csv
12. *Alternative Fuel Stations (includes the location of a nuclear power plant by region and city)*
    a. We can analyze how many states have accommodations for hybrid/electric cars
    b. *File: AlternativeFuelStationLocation.csv*
13. *Airport Locations (includes airport size, type, name, longitude, latitude, city and state information)*
    a. We can use these data to account for nearby airports when analyzing environmental factors for specific geographic regions
    b. FIle: AirportLocation.csv

14. *Facility Pollution (includes information that lists every facility in the state and greenhouse gas emissions) -- 2010-2014*
    a. We can use this dataset to account for the largest pollution factors across the U.S
    b. File: FacilityPollution.csv
15. *Nuclear Power Plant Locations (includes the location of a nuclear power plant by region and city)*
    a. We can use this dataset account for the environmental impact of the nuclear plants
    b. File: NuclearLocation.csv
16. *Vehicle Data (includes make, model, mpg data, among other attributes. Links with Emissions.csv via ID)*
    a. We can use this dataset, along with commuter volume and registration data to understand the overall environmental impact vehicles are having
    b. File: VehiclesC.csv

Data Issues

- The primary issue with this question is that emissions-induced global warming functions on a more global scale than other atmospheric issues such as air pollution. Because CO2 disperses much more broadly than smog, it is unclear whether disproportionate increases in emissions in specific locations actually lead to disproportionate increases in temperature in those locations.
- The other issue with the scoping of the question is that global warming occurs over a very long time period, and the datasets that we are using primarily only address the 1960-2019 timeframe.
- For datasets that were in the CSV format, there were many missing values, some columns included comments such as “not reported”, and many columns were deleted because the titles were not self explanatory.
- When it comes to transit and transportation data, most of it is collected by different state agencies and aggregated outside of those organizations. This leads to discrepancies in reporting and combining of datasets, which could, in turn, lead to incomparable data and skewed results.

Data Cleaning and Cleanliness Metrics

In the temperature dataset, the clean() function cleans all 12 columns of average monthly temperatures. This function converts numerals improperly saved as strings, and then attempts to approximate invalid values (strings, null values, etc…) using the previous and following years’ values. If any cell in the row is not able to be approximated, that row is discarded from the dataset. The cleanliness() function outputs overall statistics regarding the percentage of data points that were clean upon arrival, able to be approximated to a clean value, and not able to be approximated.