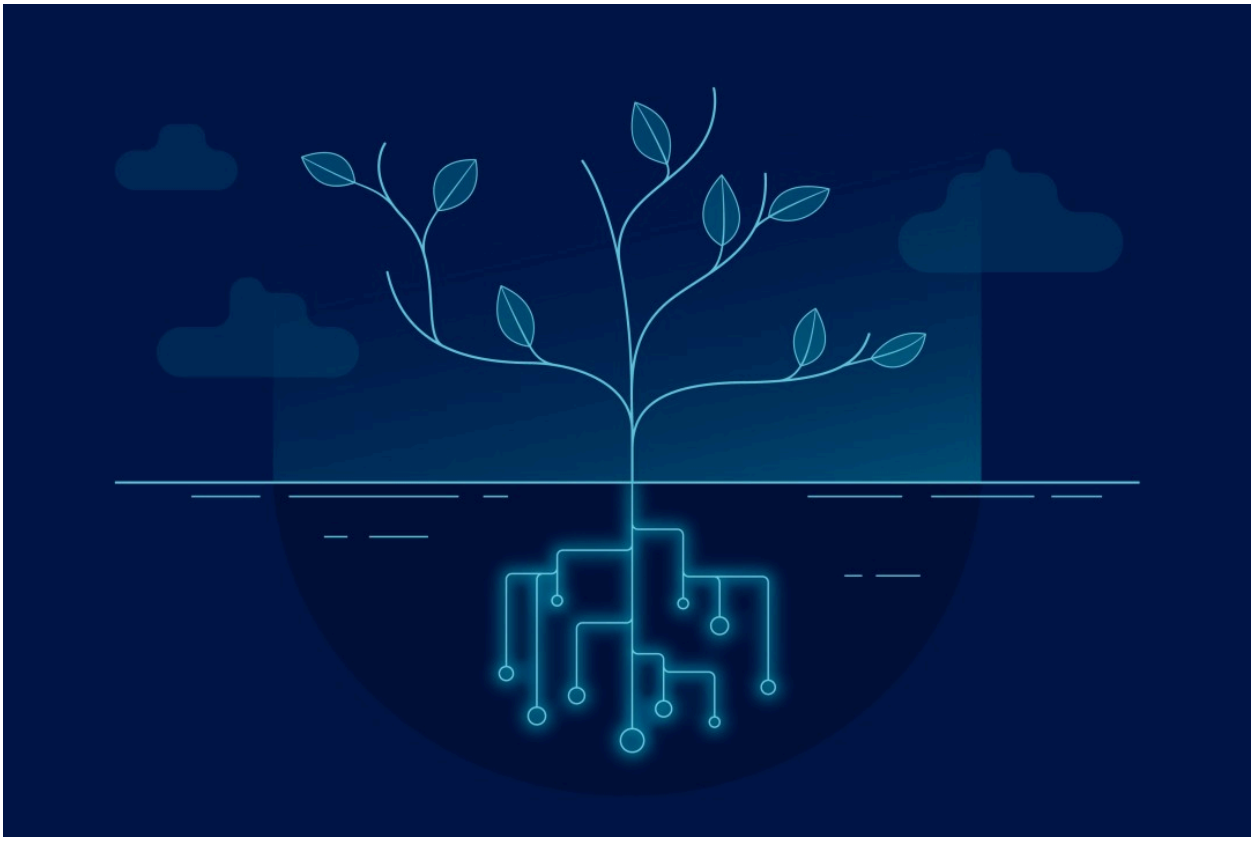
Climate Insights: Predicting Temperatures with CO_2 and GDP

Can a city's socio-economic factors help predict its temperature?



Birfatehjit Grewal, Gregory Parent, Ibrahim Rehman

bsg13@sfu.ca - 301467843, gparent@sfu.ca - 301394300, irehman@sfu.ca - 301399434

06.12.2024

CMPT 353: Computational Data Science

TABLE OF CONTENTS

INTRODUCTION

Climate change is a global issue with dangerous consequences, impacting us today and threatening greater risks in the future if we fail to take action to lessen its effects. This is why we set out on a project to investigate how socio-economic factors, such as CO_2 emissions and GDP per capita, influence temperature changes. To get started, we brainstormed datasets we could investigate and believed were related to climate change. We quickly realized that we might not be able to access all the datasets we wanted to, so we chose ones that were accessible for free. After we narrowed down our choices of datasets, we thought about the techniques from class that we wanted to apply to analyze the data and relationships, and how we could make predictions using our data. This led to us forming the following questions we wished to answer in our project:

- What trends do we observe in the individual datasets?
- Are there relationships between the datasets, and if so, what kind?
- Can this data be used to predict future attributes of climate change using a machine learning model?

To start things off, we had to narrow down the scope of our project by choosing the regions we were interested in analyzing. We decided to focus on the United States and Canada, using the capital cities of each state and province. We chose capital cities because they are usually consistent, unlike urban cities which can change due to economic booms, immigration, etc., and also tend to have plentiful data. We initially assumed that the data would show an increasing trend based on our background knowledge of climate change. However, the initial data from the period we analyzed (2000-2010) revealed some unexpected patterns, which required us to thoroughly investigate the data and its trends and relationships to ensure we understood it and knew our predictive capabilities.

THE DATA

Weather Data

Weather data is integral to any climate change project, ours among them. To effectively analyze trends and relationships between climate, emissions, and GDP, we needed a robust dataset that included important weather features for the specific locations that we are using for our project.

We landed on the **Open-Meteo Archive API**, a great tool for retrieving historical weather data. This API offers cool features like daily maximum and minimum temperatures, precipitation, and wind speed. It's free and has data for all of our locations for the period that we are using for the project (which is very rare as we have learned from the rest of this project). Moving on, with the help of the capitals.json file, we wrote a script called extract_data.py to extract the data and save it locally. The script essentially fetches data for each capital in the capitals.json file using its latitude and longitude, additionally, instead of requesting data for the whole decade at once, we fetched it one year at a time for each city. At this point, we had a folder full of city-specific CSV files. This was a good start, but it wasn't optimal for our analysis. This led us to create another script to combine all the individual CSV files into one single dataset for the weather data. This was probably the easiest dataset we were able to acquire, although we did run into an issue with **API Rate Limits**. This was resolved by waiting 5 seconds between requests to ensure we didn't get blocked, however, it slowed the execution of our scripts.

Population

The Population data was needed for numerous reasons in this project. Although it was not directly used as an input feature for our machine-learning model it was essential in deriving other datasets that were used as input.

The population data can be divided into two different types. The first type of population data we gathered was the population over different years for each city in the capitals.json file. This was quite difficult because the data was not gathered as often as needed for our project. Although most of the US capital cities had annual population data available at Neilsberg, some did not. The dataset obtained from Neilsberg resulted in 47 CSV files representing different cities. The remaining US cities and all the Canadian cities did not have any datasets available for population data in different years. So, to obtain the remaining data we had to manually collect the population data for different years across multiple websites. A few capital cities in the US only had population data available with 10-year gaps. The Canadian capital cities had similar problems since the census data was collected every 5 years. This left large gaps in the population data which we used linear interpolation to fill.

The second type of population data that we collected was the population of each state and province. This data was required to transform our other datasets like emissions and GDP per capita. The Federal Reserve Bank of St. Louis has a collection of population datasets available for download which we used to get the population data for almost all the US states. The only US state that was not available was Delaware where we had to collect the population entries from the US Census Bureau. The province populations were nicely available at Statistics Canada which just needed to be extracted and transformed into the format we needed.

City Transportation CO_2 Emissions

Another possible factor we wished to examine was the amount of CO_2 emissions produced per city by transportation in the years 2000-2010. Unfortunately, we could not find this specific data in an easy-to-access format, so we had to get a little creative. We found state/province level data from the U.S. Energy Information Administration (EIA) and Canada's official greenhouse gas inventory - Canada.ca as annual estimates in megatonnes (same unit for US and Canada, luckily). Each country had different data layouts, but transforming the data and combining was done in the same way.

The data for the US was spread out across fifty Excel spreadsheets, one for each state, and so we had to read each file, extract the rows we wanted, and combine it all into a single file. The Canada data was just a CSV file with entries for each province, year, total emissions, and sector, so we only had to filter it down to what we wanted. Once the data was extracted, we needed to transform the state-level data into city-level data. To do this, we used the gathered data on the populations for each state and province and calculated ratios of city population divided by state (or province) population and multiplied to get our estimated City Transportation CO_2 Emissions annual data. To match the other data, we also interpolated the data between the years to estimate the monthly trends, however, this came with the limitation of **not considering within-year variations like seasonal effects**.

GDP Per Capita

We chose to include the GDP per capita data since we believed that there might be a connection between the production in a city and the temperature due to the carbon emissions in the process of product production. Additionally, since we plan on predicting the temperature using a machine learning model we needed additional input features that could help increase the prediction score. The GDP per capita data is obtained from Federal Reserve Economic Data which provides the GDP per capita for most US cities. Some cities did not have GDP per capita data available so we had to make do with the available data like the GDP per capita of the county the city is in and sometimes even had to rely on the GDP per capita of the state. Additionally, Canada does not have GDP per capita data available for the cities so the closest we were able to get was the GDP of the province which we transformed with the province population dataset to estimate the GDP per capita at the city level.

ANALYSIS

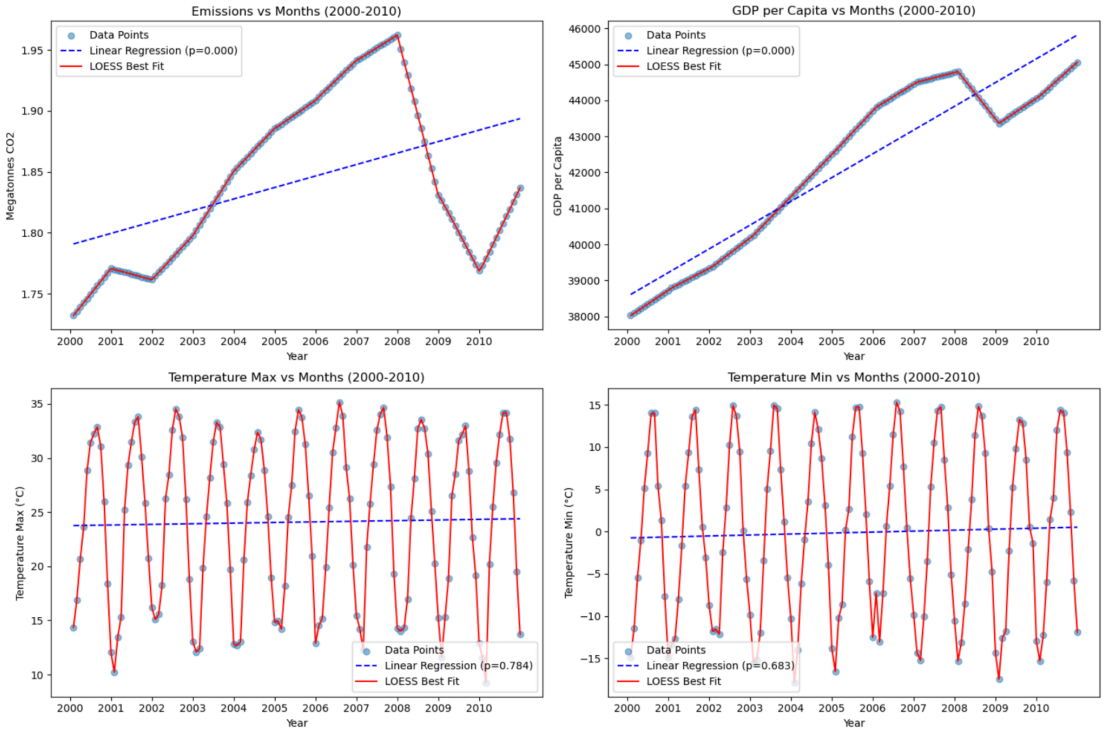
We finally have all our data together combined in a single CSV file, however, before directly jumping into building a machine learning model to predict monthly minimum and maximum temperatures using GDP and emissions data, we wanted to take a step back and understand the dataset. A few questions we thought were worth answering:

- How are these variables changing over time?
- Do they correlate in ways we expect, or are there any surprises?

How are these variables changing over time?

We began by examining ten-year trends for temperature, emissions, and GDP to identify clear patterns. After all, the way these variables evolve over the years could tell us a lot about their relationships. Like any good data scientist (student version), our first goal was to plot the data. Hence, we grouped the dataset by year and month, calculating monthly averages for min and max temperatures, GDP per capita and emissions for all the cities where we sampled the data. We then proceeded to plot the data using a scatter plot, we also applied linear regression to see if these variables showed statistically significant trends over the decade and used LOESS lines to fit the data to see the overall pattern.

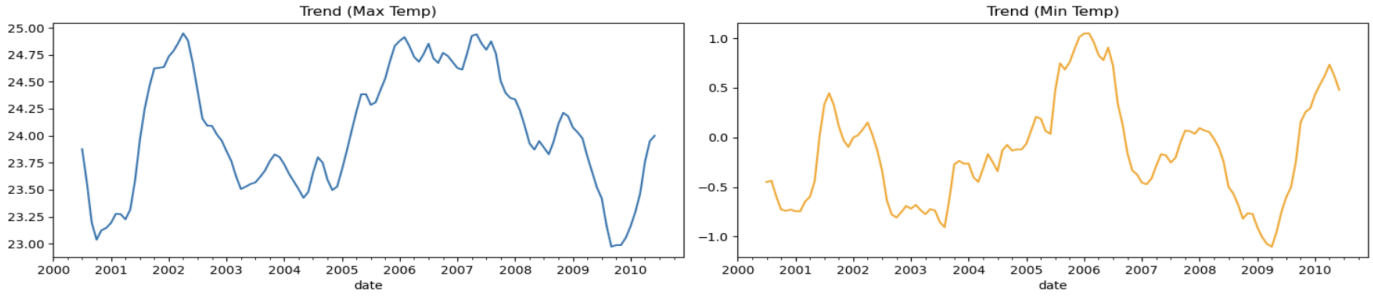
Figure 1:



As seen in Figure 1, the emissions and GDP per capita datasets turned out to be much more interesting than we had thought. First, we were able to successfully prove that the slopes for both are indeed non-zero with the linear regression p-value. Second, we can see from the plot that right around 2007, the emissions and GDP data suddenly started to trend downward until 2009, after which they started recovering. While on the other hand for the min and max temp we are unable to prove that they are changing over time with just the linear regression p-value. Additionally, the plot is also not super helpful, because of the temperature fluctuations the trend is hard to decipher and may need a closer look.

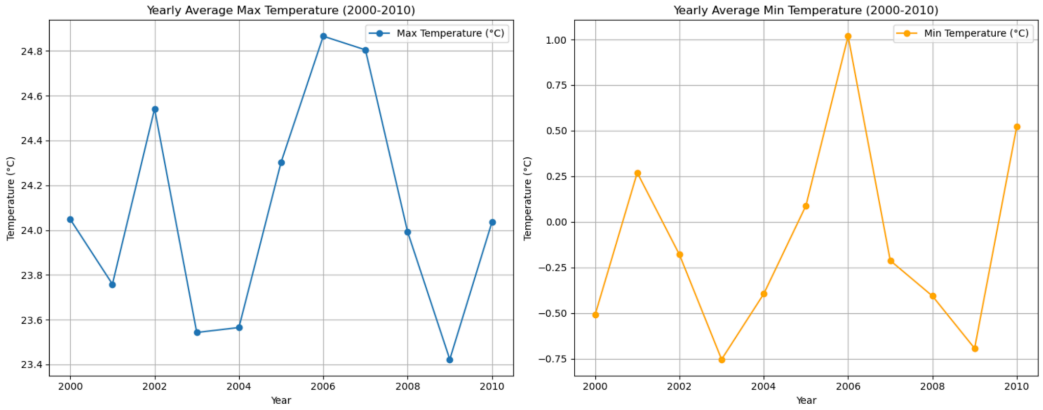
To make sense of the min and max temperature datasets we thought of two possible paths, aggregating the temperature data yearly to account for the monthly fluctuations and doing a time series analysis. The time series analysis gave us some very interesting results, as seen by the trend plots produced by it.

Figure 2



We can see that temp data definitely changes from year to year (Figure 2), however, it just so happens that the monthly min and max temperatures in 2000 and 2010 are about the same, hence we were unable to prove that the temperature is changing over time with just a simple linear regression test.

Figure 3



On the other hand, by aggregating the temperature data yearly and plotting it. We can also clearly see that the temperature has indeed changed in the ten-year span, and we can also see that there are most likely some factors that are making it fluctuate in an odd manner, this leads us to our next question.

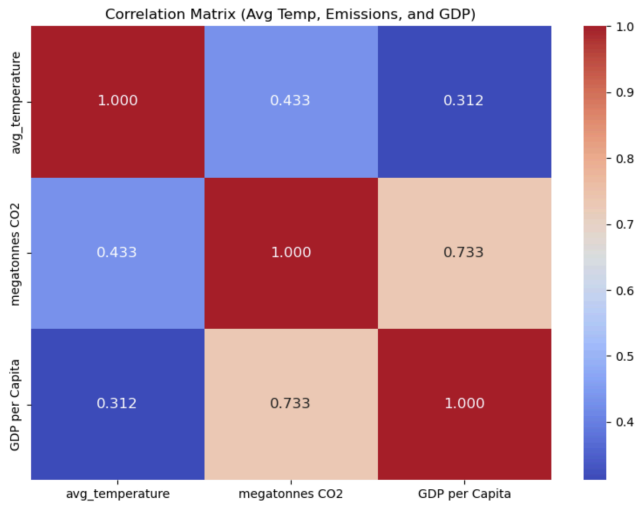


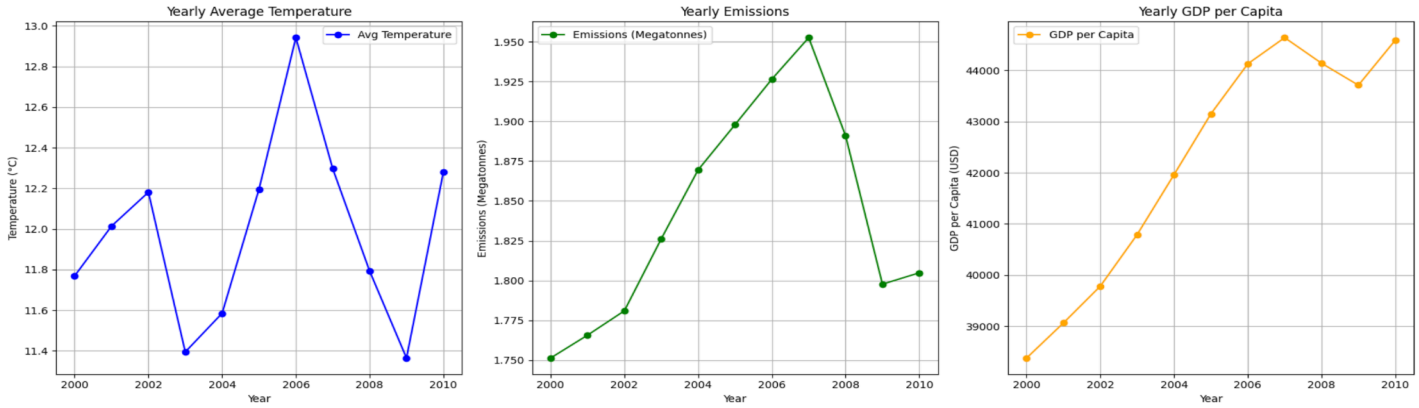
Figure 4

Do the datasets correlate in ways we expect?

Our next step to learn more about the datasets we have acquired is to see if the datasets are related to each other, for this, we can find the correlation coefficient between the datasets and create a correlation matrix. The correlation coefficient gives us the following results:
Avg Temp vs Emissions: **0.433** (moderate correlation)
Avg Temp vs GDP: **0.312** (weak correlation)
Emissions vs GDP: **0.733** (strong correlation)
Additionally, we can also see from the correlation matrix that emissions and GDP datasets are clearly related. However, average temperature only has a weak correlation with GDP and only a moderate correlation with emissions.

To further investigate, we can plot all three datasets, all being yearly aggregated this time, as monthly aggregation did not make the relation very clear to us earlier. Figure 5 (below) clearly shows us the relationship between the datasets (we averaged out the min and max together since that made more sense). We can see that all three datasets peaked around the 2006-2007 mark and then rapidly dropped until 2009, after which the data started trending upward in all three datasets.

Figure 5



Machine Learning

With the previous correlations found, we started examining linear regression models to help take advantage of the linear correlation between GDP per capita and CO_2 emissions to try and predict the temperature values in the 2000-2010 data. However, we quickly learned they would not suffice due to low training scores around 8%. This perplexed us since it did not match our earlier analysis, furthermore, we noticed non-linear models gave better results. Our best explanation for this is that there are other patterns in the data that the linear models failed to see, so we chose to continue with a non-linear regression model.

In particular, we found that the non-linear models KNeighborsRegressor, RandomForestRegressor and GradientBoostingRegressor all performed decently on their own, and so we decided to use them all in a StackingRegressor which let us use their strengths to create a stronger model by “stacking” their outputs and using a final regressor for the prediction. To further optimize our model, we used a grid search method which tested multiple combinations of hyperparameters for each specific regressor over many iterations. This process also uses 5-fold cross-validation which uses 5 random training/validation splits for every iteration and scores them then takes the average to prevent overfitting. This ensured our model had the best possible performance and gave us the following results:

Model inputs	**Training (2000-2010) score (R^2)**	**Future (2011-2013) score (R^2)**
Only Year and Month	0.677897	0.641443
Only CO_2 emissions & GDP per capita	0.912400	-0.197607
Year, Month, CO_2 & GDP per capita	0.975878	0.791278

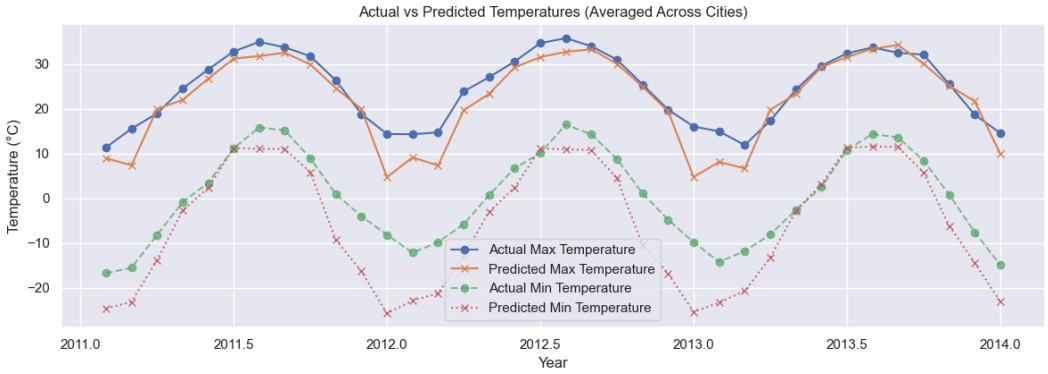
Each model was trained on an 80% training, 20% validation split on the 2000-2010 data, and tested to try and predict future temperatures that we obtained for the next three years, 2011-2013. The scoring metric we used was the coefficient of determination (R^2) and a higher score means the model explains more of the target data. From the results in the table which shows our model performances on the entire training and future data, we see that using just the year and month as inputs did perform okay with about 67% on the training data and 64% on the future data which backs up our earlier analysis on not being able to find a definitive linear relation in the temperature data over time because we used a non-linear model. On the other hand, the model that used only the CO_2 city emissions data and GDP per capita data performed well on the training data but got a negative score on the future data. This is a clear sign of overfitting and that something was missing. For our final model, we used the year, month, CO_2 data and GDP per capita data all as inputs, and it scored the highest on both the training and future data, around 97.5% and 79% respectively! This suggests that data on CO2 emissions and GDP per capita is useful in predicting future temperatures, providing insight into their contribution to climate change.

RESULTS/CONCLUSION

The goal of this project was to find out whether socio-economic factors, such as CO_2 emissions from transportation and GDP per capita, could be used to predict temperature trends over time. To answer the question at hand, we collected data from the capital cities of each US state and Canadian province (excluding territories). We then analyzed if and how the datasets are related to each other and finally trained a machine learning model to predict the temperature min and max data given the Month, Year, CO_2 emissions and GDP per capita. However, did this actually answer the question?

Model Results

Figure 6



Our model performed well, achieving a testing accuracy of **79% R^2** on future data (2011–2013). Taking a look at Figure 6 which plots the min and max temperatures in the 2011-2013 data and the predictions from our best model over the same 3-year timespan, we can see that our predicted values follow the trends closely, but the values can sometimes deviate a bit, especially for the min temperature prediction. However, despite these deviations, our model seems to make decent predictions, allowing us to partially achieve our goal.

Why partially? While our results answer the question of whether socio-economic factors like GDP and CO_2 emissions can predict temperature trends, can this really be used to predict climate change? Climate is a highly complex concept influenced by numerous variables, and our model reflects only a small part of a much larger picture. Our approach had many limitations and revealed many aspects that we could have done differently.

Limitations and what we could have done differently

Our dataset focused solely on the capital cities of US states and Canadian provinces, this limited the relevance of our findings because they can not be generalized to other regions or reflect global climate patterns, especially since we also excluded Canadian territories. Sampling more cities, territories, and other countries would have provided a broader, more diverse and reliable dataset.

We used GDP per capita and transportation-related CO_2 emissions as our primary socio-economic features. While these were effective starting points, climate can be influenced by countless factors, including industrial waste, renewable energy programs, land use, and natural phenomena. Having incorporated more diverse and climate-specific features would have likely improved our predictions.

Finally, many datasets we used, such as population and emissions, had significant gaps. To address this, we had to do a lot of manual extrapolation and interpolation on the data, for example, we estimated city-level emissions using population ratios and filled census gaps with linear interpolation. While this process was necessary, it also introduced potential biases, which ultimately affected the reliability of our findings.

Final Thoughts

Despite all the limitations, our project demonstrates the potential for using socio-economic factors like GDP and CO_2 emissions to predict temperature changes over time, providing a strong foundation for future research to build on.

PROJECT EXPERIENCE SUMMARY

Birfatehjit Grewal

- Collected comprehensive datasets on population and GDP per capita for 60 regions, including cities, provinces, and states, ensuring accurate and diverse data for model training and analysis.
- Combined and transformed datasets using the pandas library in Python to format and calculate to input features for model training and statistical analysis.
- Documented the testing process by developing clear, user-friendly guidelines on project dependencies and code execution procedures, significantly decreasing the ambiguity on how to run the code.
- Co-authored a report on the relationship between GDP per capita, CO_2 Emissions, and Climate change

Gregory Parent

- Estimated CO_2 emissions for 60 cities across the United States and Canada using state-level data and population metrics to support machine learning predictions and analysis.
- Constructed a machine-learning model using Scikit-Learn, Python and the gathered data to predict minimum and maximum temperatures using year, month, CO_2 emissions and GDP per capita data.
- Contributed to the report by providing details on how we constructed the machine learning model and documented its accuracy and precision relative to actual data.

Ibrahim Rehman

- Created custom scripts to extract and consolidate historical weather data (2000–2010) for 60 North American cities using the Open-Meteo API.
- Conducted statistical testing to analyze relationships between climate variables, leveraging linear regression and time series analysis to support data-driven insights into climate change factors.
- Visualized prediction accuracy by plotting error graphs to compare predicted and actual temperatures for 2011–2013, highlighting model precision.
- Co-authored a detailed report on socio-economic and environmental factors influencing climate change.