Algonquin College

CST 2213_300: Business Intelligence Programming 2: Advanced Concepts

Final Project

Prof. Sanket Ganatra

August 09, 2025

Submitted by

Tanush Anand Ilanchezhian

041127144

# Table of Contents

## Problem Definition

The primary objective of this analysis is to investigate the relationship between global mean temperatures, atmospheric $CO_2$ levels, and socioeconomic factors (income groups). Specifically, we aim to:

1. Check if mean temperature can be reasonably predicted using year and $CO_2$ levels.
2. Determine whether mean temperatures differ significantly across income groups.

This is not intended as a comprehensive climate model but as a statistical exercise to satisfy predictive analysis requirements for the project.

## Methodology

### Data Sources

- Global Land Temperatures by Country – Kaggle (Mean annual land temperatures by country)
- World Bank Country & Lending Groups – Kaggle (World Bank income group categories)
- CO2 Emissions – Kaggle (Annual $CO_2$ emissions (kilotonnes) by country)

The datasets were merged to align country, year, mean temperature, $CO_2$ emissions, and income group.

### Data Processing

- Filtered data by selected countries and years via a Streamlit dashboard.
- Applied log transformation to $CO_2$ emissions for certain plots to handle skewness.
- Grouped data by income category when number of selected countries exceeded display limits.

## Predictive Check

### Model

$$\text{MeanTemp}=a\times\text{Year}+b\times\log(CO_2)+c$$

### Approach

- Applied simple linear regression using year and log-transformed $CO_2$ emissions as predictors.
- Calculated in-sample $R^2$ and cross-validation $R^2$ (CV $R^2$).

## Findings

A simple linear regression model was built using CO2_kt to predict MeanTemp.

**Predictive Check Results**

- In-sample $R^2$: 0.627 — The model explains ~62.7% of the variance in mean temperature.
- CV $R^2$: -3.535 — Cross-validation indicates poor generalization, suggesting the model may be overfitting or too simple for unseen data.

## Recommendations

- Model Refinement - Include additional predictors such as geographic location, urbanization rates, and land use changes.
- Statistical Testing – Use post-hoc pairwise tests (e.g., Tukey HSD) to identify specific group differences.
- Statistical Testing – Use post-hoc pairwise tests (e.g., Tukey HSD) to identify specific group differences.
- Visualization – Use clearer legends and annotated plots for improved interpretability.

## Conclusion

This analysis confirms a statistically significant association between $CO_2$ emissions and mean temperature, as well as notable differences in temperature between income groups. However, the predictive model's poor cross-validation performance suggests that $CO_2$ and year alone are insufficient to fully explain temperature variations.

The results highlight the complexity of climate systems and the necessity of incorporating multiple interacting factors in predictive models. Additionally, the disparities across income groups indicate that climate change adaptation and mitigation strategies should be tailored to specific regional and economic contexts.

## References

*Global land temperatures by country. (2021b, July 1). Kaggle.* *https://www.kaggle.com/datasets/vijayvvenkitesh/global-land-temperatures-by-country*

*World Bank Country and lending groups*. (2019, November 17). Kaggle. https://www.kaggle.com/datasets/taniaj/world-bank-country-and-lending-groups

*CO2 emissions*. (2023, February 28). Kaggle. https://www.kaggle.com/datasets/ulrikthygepedersen/co2-emissions-by-country