STAT444: PROJECT PROPOSAL

BY ZAFAR ERKINBOEV[1,a] ADITYA JAYANTI[2,b] TOM SHU[3,c] BRYAN ZANG[4,d]

[1]*Department of Statistics, University of Waterloo,* [a]*zerkinbo@uwaterloo.ca*

[2]*Department of Statistics, University of Waterloo,* [b]*ajayanti@uwaterloo.ca*

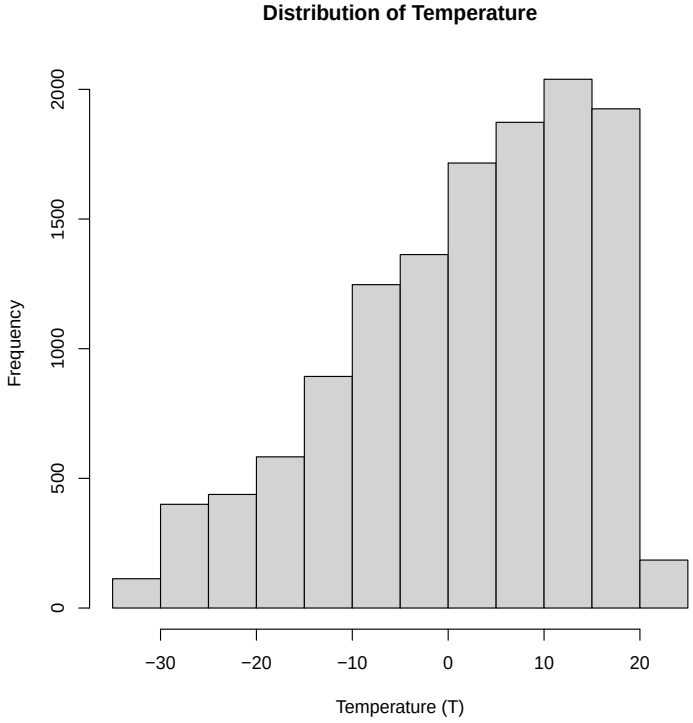
[3]*Department of Statistics, University of Waterloo,* [c]*t4shu@uwaterloo.ca*

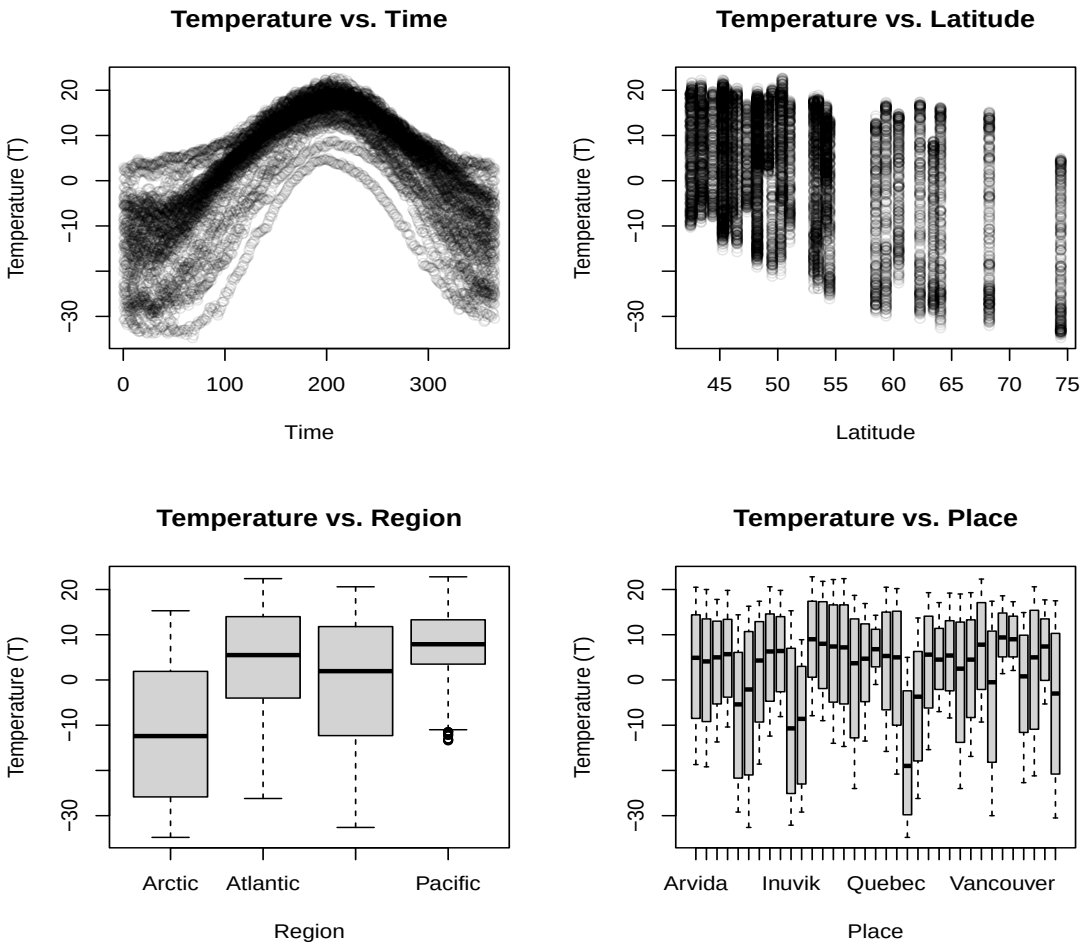
[4]*Department of Statistics, University of Waterloo,* [d]*bszang@uwaterloo.ca*

Final project proposal for spring 2023 offering of STAT444.

1. Dataset.

The dataset we chose to work with comes from the package `gamair`, it is the `CanWeather` dataset which is a dataset describing Canadian temperature (averaged over the years 1960-1994) throughout a year at 35 different locations. There is a column of integers denoting the day in the year (1-365), a numerical column for the mean temperature for that day in Celsius, a column of categorical data denoting the general location (Arctic, Atlantic, Continental, or Pacific), a numerical column denoting the location latitude, and another categorical column of location names. We conducted some brief exploratory data analysis to better understand the data and temperature arises as the response variable of interest, and covariates such as time and region appear to be interesting in the sense that we are able to partition data into portions local to some variable (i.e., the temperature change in x days or the overall temperature model in the Atlantic region).





2. Research Question.

It is worth noting that the Canadian weather dataset is originates from the `fda` package in which the data is sourced from *Functional Data Analysis 2nd ed.*, by Ramsay and Silverman (2006) and that the book is does not state why/how the data was obtained. But in general, temperatures are recorded at weather stations to monitor and study the conditions of the atmosphere and other related environmental factors. Then, the research question we hope to address in our analysis is that of the comparison of temperature change over each region and how well we can model these changes using various statistical techniques.

3. Analysis Plan.

To actually analyze the data, we would first partition the dataset by region into subsets, and then we would want to model each of the data subsets using various techniques so we can compare and contrast how well we can model the temperature changes. These techniques can include simple linear regression, polynomial regression, penalized spline regression, and etc. In general we would like to find a model that both fits the data sufficiently well enough and not require a heavy computational cost. Once we have a suitable model for each region, we can then compare the effectiveness of modelling the data over each subset — i.e., we look into questions such as "does the usage of spline functions increase how well our model fits the data?".