EDX Capstone Project

Mark145

Fall 2019

Contents

Introduction

Weather balloons are deployed world wide every day. The data collected from each balloon is stored by the National Oceanic and Atmospheric Administration (NOAA), an American scientific agency within the United States Department of Commerce.

NOAA produces the raw and derieved (scientific calculations base on raw) data for public use and makes the data availiabe within NOAA's Integrated Global Radiosonde Archive (IGRA). The IGRA website is https://www.ncdc.noaa.gov/data-access/weather-balloon/integrated-global-radiosonde-archive

Overview

The model contained within this report seeks to predict temperature for any given altitude at any given day of a year. The data is sourced from radioscondes and available for public use via NOAA's website. The training and test data will focus on all the data produced by the UAE at Abu Dhabi International Airport.

Root Mean Square Error (RMSE) will be used to report the error this algorithm produces. Our base altitude will contain data for 1000m +/- 100m. First, error will be calculated by guessing the average temperature for the base altitude given a day. Secondly, error will be calculated by incorporating the previous day's temperature for 1000m +/- 100m and regressing to the mean for the predicted day.

Executive Summary

The model generated below, computes the average temperature for a given altitude block and associates that average to unique Month Day. It then applies that average to the predicted Month Day and computes the RMSE. The RMSE generated by this model is ~3.2*C.

Methods/Analysis

To run the code for this model, please ensure that your R is capable of accepting the following packages for installations. -tidyverse -caret -data.table -sf -rgeos -naturalearth -naturalearthdata -ggspatial

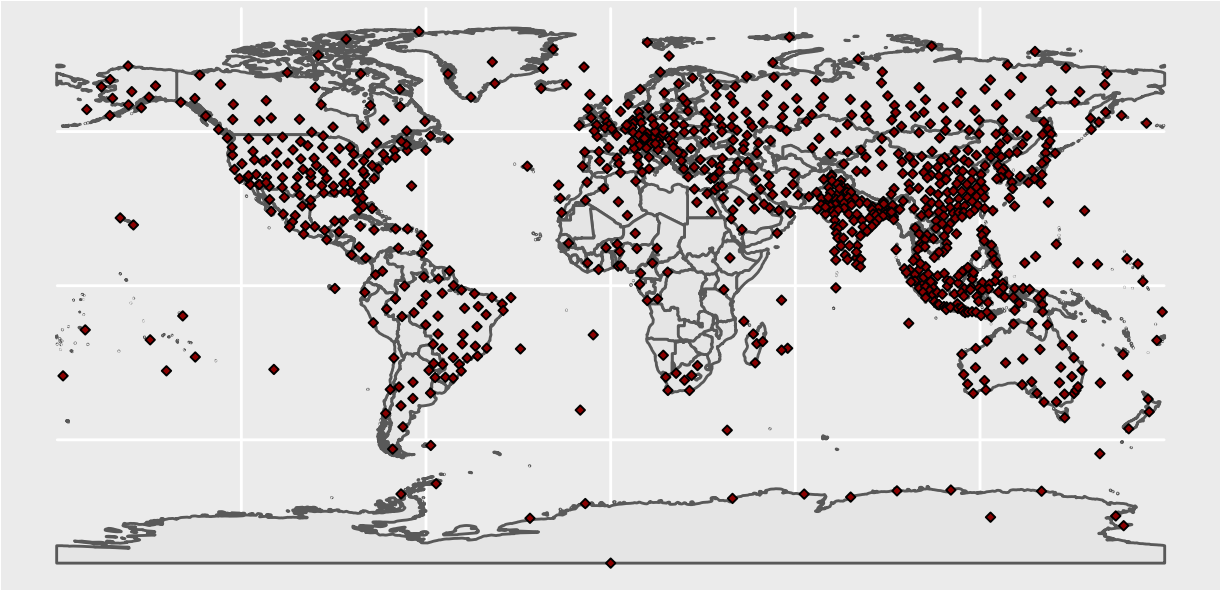
I will attempt to install the packages above with if(!require()) however, some of the mapping packages might require additional software.

The first step taken was to reach into NOAA's IGRA file space and pull the data stored from the United Arab Emraites' Abu Dhabi International Airport station as well as Quatar's Hamad International Airport for the data to be used during exploration. Also used for exploration, is the station list data generated by NOAA. The data is then parsed and formatted for each record to represent one recorded reading of the weather balloon sensor.

Exploration

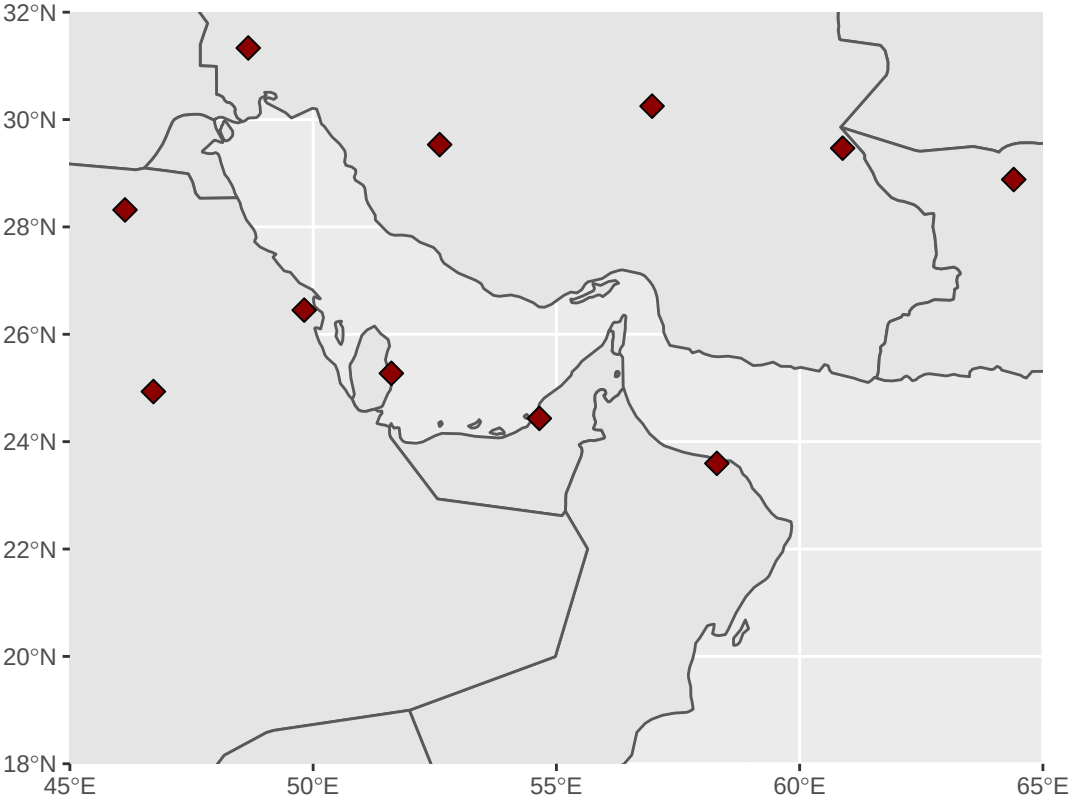
Looking at a chart of all the current readiosconde deployments worldwide. We can see which regions of the world are active in their reporting to NOAA and also know that NOAA has made that region's data avaliable.

Worldwide Radiosconde Launch Stations

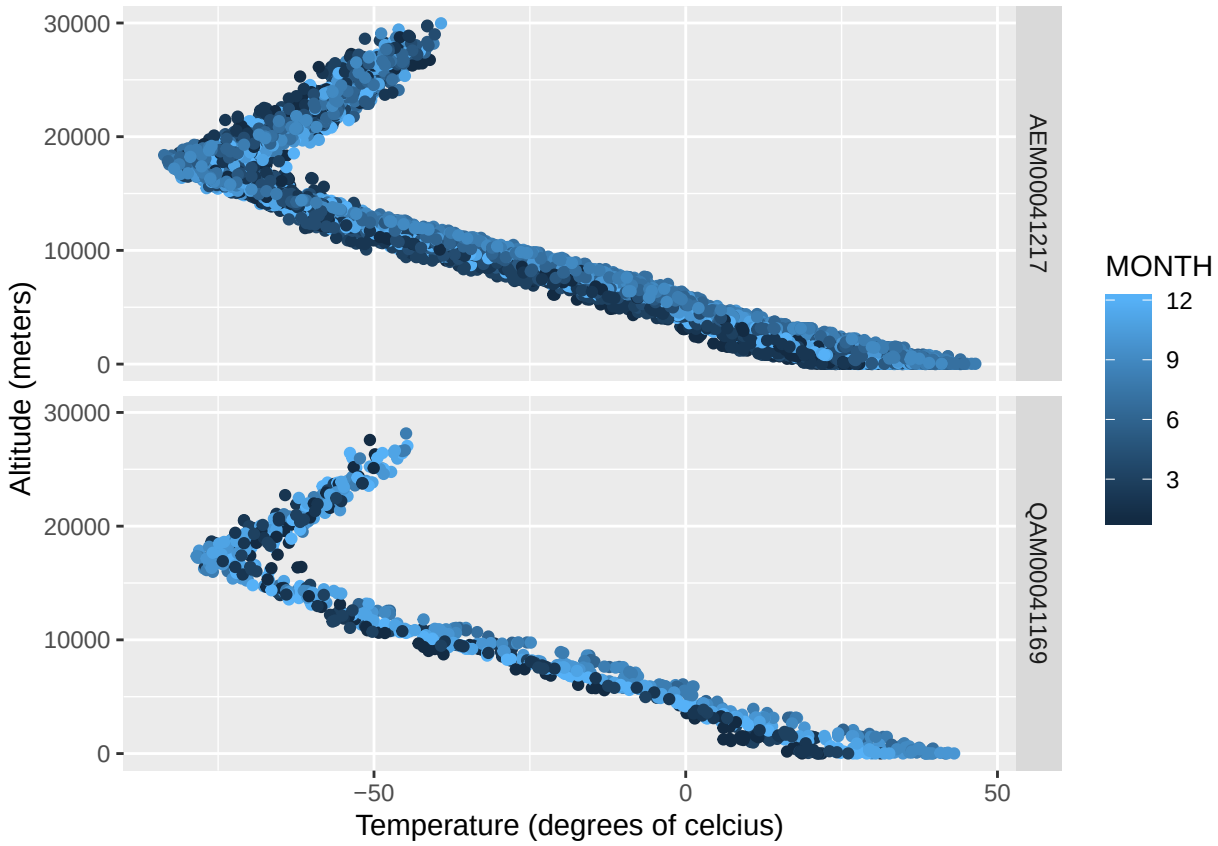


Since the region we are interested is the Arabian/Persian Gulf, lets take a look at all the stations located in the region.

Radiosconde Launch Stations Around The UAE

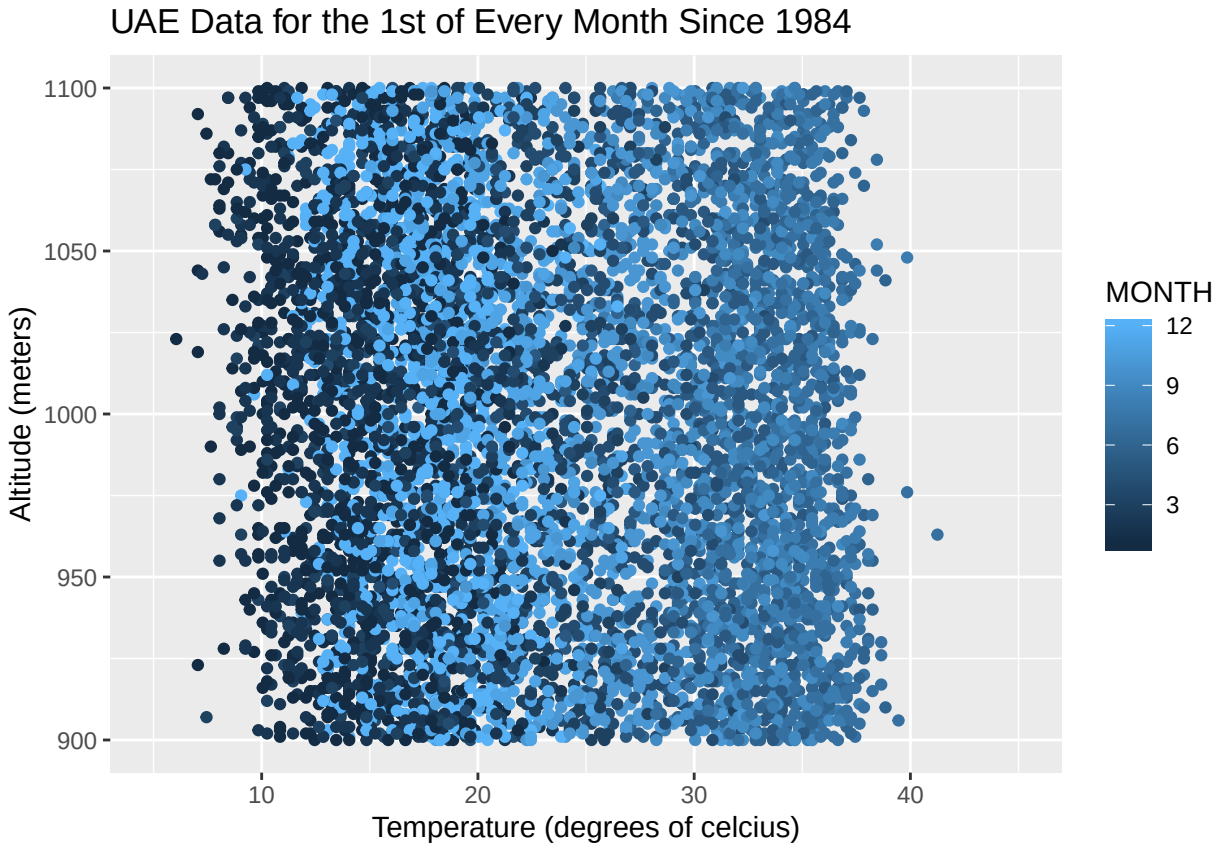


Now lets look at the temperature data for the UAE site and how it compares to the station in Quatar. The data found in both plots is derived from the first day of each month and a launch time of 1200 UTC.



Here is a plot adjustment, with all the UAE data, for the altitude we are attempting to predict. -Altitude 1000m +/- 100m

```
## Warning: Removed 837969 rows containing missing values (geom_point).
```



As depected above, the winter months (lightest and darkest blue) have lower temperatures while the summer (blended blue) have warmer temperatures.

Algorithm Development

I have generated a training and validation set by selecting the same station and altitude block (1000m +/- 100m) and separated the data based on launch time. 0 UTC launches will be used for training and 12 UTC launches will be used for validation.

The check function, how well we did will be computed by 'root mean square error' (RMSE). We will find the difference for each days temperature rating at this layer, to what was actually measured at the layer. Then square this value and average it for all days with measurements. Lastly, we will find the square root of that average and report the value. A value of 1 means 1 degree of celcius off from reality.

There are a couple of ways of determining the same time every year. One method would be to convert the date to a number 1-365 and group all of the same numbers throughout the year. Perhaps the most accurate way to identify the same time each year would be the incorporation of celiestial bodies. For the sake of this model, I will group by Month Day and calculate the average temperature.

The RMSE for predicting the average daily temperature is 2.79*C on our training data.

Modeling Results

Now the daily average for UTC(0) will be applied to the validation set UTC(12).

We can see that our RMSE increased to ~3.2*C.

Conclusion and Future Work

Our RMSE increased when we applied the derieved mean temperature to each record. This was likely due to the variance in temperature when measuring at different times during the day.

Future model development should acount for all times of the day and some sort of historisis. Perhaps the previous day's average temperature or the previous altitude block's average temperature or a look at the previous day's temperature.

It might be possible to compare the previous day's temperatures and identify a trend and rate of increase or decrease. However, when examining this data, it was noted that some days are missing, perhaps due to holidays or weekends.