Weather Project

Brett Keller

Final Project

08/12/21

Intro

This project was to explore relationships in a dataset collected from a total of 161 land-based weather stations from 1940 to 1945. The dataset includes dates, coordinates (latitude, longitude, and elevation), temperatures, and precipitation data. Various relationships, tests, and models were used to interrogate the data.

Results: Relationships

Summer months are more likely to have a warmer daily average temperature. There is much more variability in temperature during the winter months than during the summer months.

It appears that there is a strong, positive, linear relationship between the daily maximum and daily minimum temperature. Pearson's r = 0.88.

There is a weak, negative correlation between mean temperature and snowfall; there is generally more snow for colder temperatures. However, it looks as though the data may better be described with a non-linear model; there seems to be a sweet spot in temperature that produces the most amount of snow.

A model was also constructed in an attempt to predict the average temperature of a given location and time of the year. Only 32% of the variation in mean temperature could be predicted from this information.

Results: April Showers

The idea that April is considered a particularly rainy month has become cliche, repeated in aphorisms and platitudes. I wanted to put this to the test.

Looking at the mean precipitation for each month, it was actually September that had the highest average rainfall- 4.1 mm. Even looking at the number of days it rains per month, August and September were tied for the rainiest months. Over a third of the days of those months have at least some rain.

A T-test confirms that the rainfall is greater in September than April, with a p-value < 0.001. We can reject the hypothesis that there is no difference, and in fact the difference is opposite of what folk wisdom tells us.

Limitations

Only a handful of the different variables measured among the stations were collected by all stations. Wind speed, thunderstorms, and other interesting measures were just too sparsely available to provide any useful insight. Future research could not only examine other variables, but look to expand the range of weather stations worldwide over a longer breadth of time.

Certain decisions had to be made to analyze the data. Trace amounts of rainfall were counted as no rainfall, a few stations missing elevation data that was supplemented by the average elevation of their location’s city, and some assumptions were made regarding precipitation as excluding snowfall.

More advanced analysis could be used to analyze the data. Time-series analysis could be used to forecast the weather, non-linear relationships could be examined, similar locations could be examined through cluster analysis, and the data could be divided to better train and test the regression model.