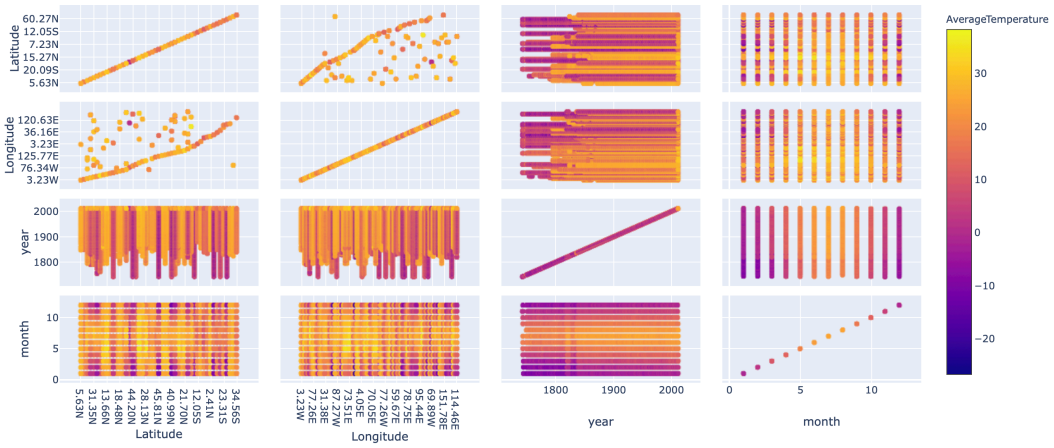
Victoria Rose | vrose9 |Data Science Final Project

Data:

- Partially Cleaned (just so it can be plotted) but mostly raw data:



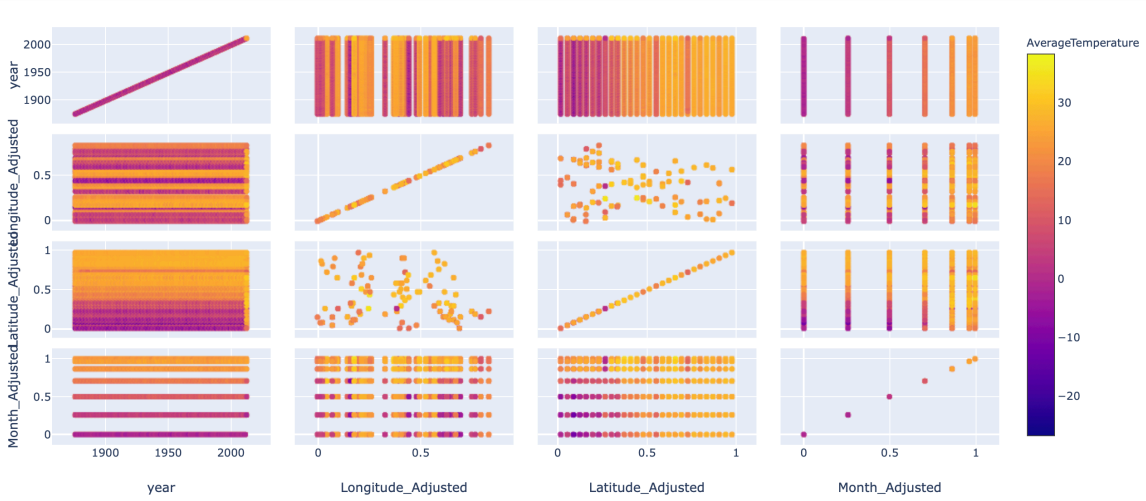
- Data Cleaning:

To clean the data I took the date (dt) column:

- I pulled month and year from the original yyyy-mm-dd format.
- For year I further cleaned it by removing all years before 1875 just so that no geographical region's temperature average was skewing the model by virtue of having sampling points from a larger range of time.
- For the months I cleaned it so that the cyclical nature of seasons could be considered by the OLS linear regression. To do this I took the months (1-12) and mapped them to a half period of a sinusoid using sin(n*pi/12) where n is the month number (1-12). This is not a linear mapping, which is a source of error within the model.

To clean the data from the Longitude and Latitude columns:

- I spliced just the numerical values.
- For Latitude I modified it to Latitude_adjusted by taking the absolute distance from the equator instead so that the equator and poles were considered uniquely but the poles were considered similarly.
- I adjusted the ranges of both of them to be between 0 and 1 for cleaner interpretation of the OLS coefficients later.



Model Fitting:

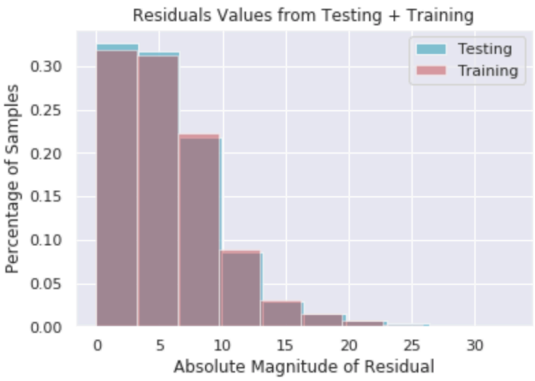
- **Overfitting:**

To prevent overfitting the data:

- Kept the number of features small relative to the number of samples (~10k+ samples/feature)
- Trimmed down the range of years so that all geographical locations were represented by a similar amount of points and that the data was not overfit to single locations.
- 25 % of the data was also withheld for testing--as it was done in class--to allow for validation of the model

- **Multivariable Linear Regression:**

```
                  Results: Ordinary least squares
==================================================================
Model:              OLS                Adj. R-squared:     0.464
Dependent Variable: AverageTemperature AIC:                834547.0287
Date:               2019-10-07 12:57   BIC:                834595.6525
No. Observations:   123595             Log-Likelihood:     -4.1727e+05
Df Model:           4                  F-statistic:        2.673e+04
Df Residuals:       123590             Prob (F-statistic): 0.00
R-squared:          0.464              Scale:              50.117
------------------------------------------------------------------
                    Coef.   Std.Err.     t      P>|t|    [0.025   0.975]
------------------------------------------------------------------
Intercept         -13.6210   0.9899 -13.7606 0.0000 -15.5611 -11.6809
Month_Adjusted     11.6649   0.0639 182.4141 0.0000  11.5396  11.7902
year                0.0085   0.0005  16.6632 0.0000   0.0075   0.0095
Longitude_Adjusted  0.2559   0.0933   2.7417 0.0061   0.0730   0.4389
Latitude_Adjusted  19.9880   0.0755 264.8912 0.0000  19.8401  20.1359
------------------------------------------------------------------
Omnibus:               7642.299       Durbin-Watson:           0.345
Prob(Omnibus):         0.000          Jarque-Bera (JB):        9131.715
Skew:                  -0.651         Prob(JB):                0.000
Kurtosis:              3.276          Condition No.:           95612
==================================================================
```




```
The average absolute residual of the training data was 5.707128
and the average absolute residual for the testing data was 5.724044.
These represent the average error this model had in determining
average monthly temperatures
```

Model fitting 4. If the fit is linear or logistic regression, did they interpret the covariates?

The covariates represent the difference in predicted temperature in degrees celcius for every unit increase in each of the respective features. For each of these variables, as they have been largely normalized or adjusted, these transformations and mappings must be considered when interpreting the varaibles:

- for months, this coefficient represents the predicted difference in the average monthly temperature in degrees celsius for every increase in ordinal month number, with january indexing at n = 1 and december at n = 12, and all points inbetween being mapped to the range 0 to 1 via the function sin((n+1)*(pi/12)). This might seem convoluted and there are probably better, more linear ways to achieve this, but it was most critical to ensure that the mapping conveyed the cyclical nature of the months and their influence on temperature.
- for latitude this represents one forty-fifth of the anticipated average temperature in degrees celsius for every degree away from the equator (north or south agnostic)
- for longitude this represents the one one hundred and eightieth of the anticipated average temperature change in degrees celsius for each degree increase in longitude
- for years this maps intuitively, this is the change in the predicted average temperature in degrees celsius for every additional year

Model fitting 3. Did the student describe how they evaluated the performance of the model?

I validated the performance of the model by comparing both the average residual and distribution of residuals for the training and testing data sets and comparing them. Their similarity in both distribution and average encourages the belief that the model did not overfit to the training data.

Interpretation:

1. Did the student summarize and interpret their findings in non tehcnical terms?

From this project I studied the relationship between the variables of latitude, longitude, year, and month and the average monthly temperature. Anecdotally we can all describe our personal understanding of the relationship between geographical location and time of the year, and I thought that it would be fun to use a model to quantitatively describe it. This model randomly considers 75 % of the original data (leaving the rest to be used later to check how good the model is at describing the larger group behavior) in order to understand the larger trends and relationships between these variables of geography and time on temperature. The findings of this project confirm what many of us intuitively might guess: distance from the equator and time of year are the two greatest predictors of what the average monthly temperature will be. Locations closest to the poles at the end and beginning of the year have lower anticipated temperature averages than their counterparts closer to the equator during the middle of the year.

2. Did the student summarize performance metrics and make suggestions on improving model performance in the terms of other models to fit or other data to collect?

What this project did well & discovered:

- Confirmed intutive understandings of how time and geography influence temperature via the covariate magnitudes.
- Created a model to approximate average monthly temperature of a location given geographical and temporal location within about 5.7 degrees C on average, both inside and outside the training set.
- The standard deviation of the raw data is about 9 degrees C, and so the average error of just under 6 is an improvement from that.

How this project to expand and improve:

- The weight of *year* as a varaible in estimating the average monthly temperature was slightly lower than I think I was anticipating after reading articles about climate chnange and so considering carbon emissions or greenhouse gas levels, or something in that league, might give a more focused measure on how temporal location with respect to year influences estimated averages.
- I also think that following our intuitive understanding of Earth's orbit and it's influence on the seasons, a variable that captures the influence of the Earth's axis via a relationship between latitude+longitude.
- Because the data set is so large, I think that adding just these 2 or 3 extra features the model could become more precise without overfitting.