

LAKE MICHIGAN INFLUENCES

BLAKE WALLACE

MAY 17, 2019

DATA SCIENCE QUESTIONS

1. Existence of significant differences in temperature?
2. Can a predictive model that explains at least 80% of the variance in the precipitation differences be constructed?

DATA SOURCES

National Centers for Environmental Information

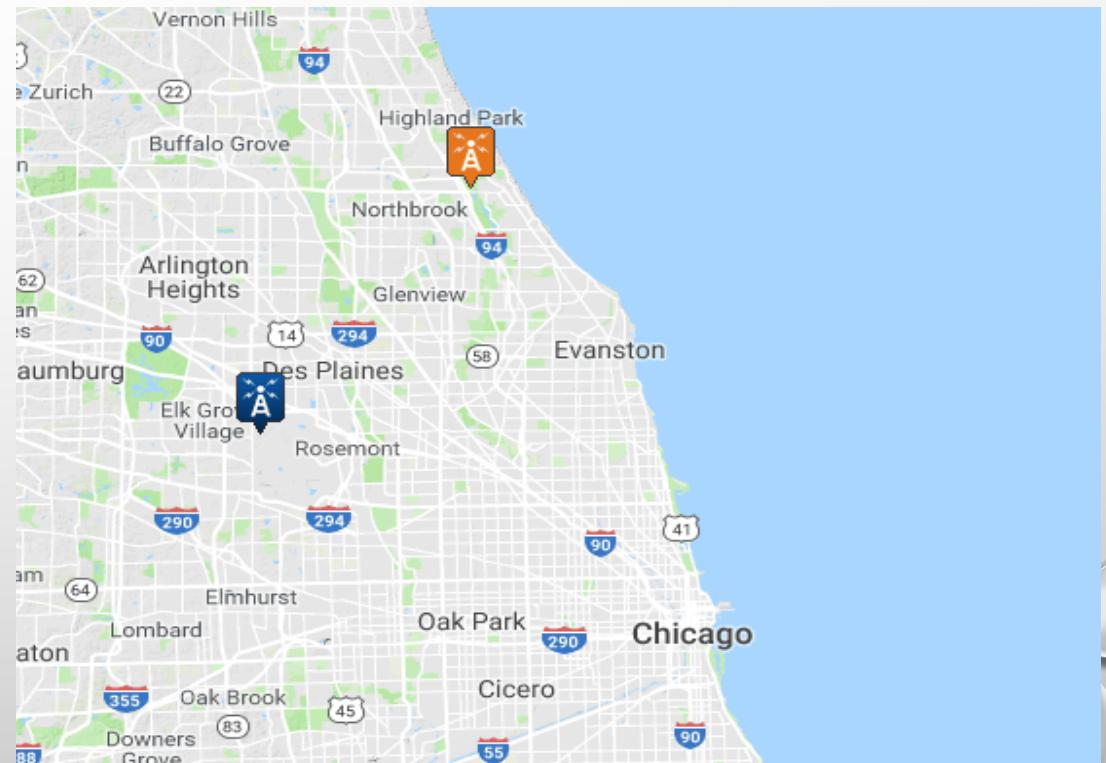
Global Historical Climatology Network (GHCN)

- a network of worldwide weather data available to the public for analysis and evaluation. - [Source](#)

- O'Hare Airport
- Chicago Botanical Gardens

Great Lakes Statistics: Average Surface Water Temperature from the Great Lakes Surface Environmental Analysis (GLSEA) - [Source](#)

- Lake Michigan Average Daily Temperatures



DATA

The Global Historical Climatological Network defines five core values

1. Precipitation
2. Snowfall
3. Snow Depth
4. Maximum Temperature
5. Minimum Temperature

Years considered:

- 1995 – 2018
- 8766 days

The Great Lakes Statistics:

1. Average Daily Temperatures of Lake Michigan

FEATURE ENGINEERING

- target - absolute difference between the precipitation measurements at O'Hare and the Chicago Botanical Gardens
- garden_didrain - categorical, 1 for yes, 0 for no
- ohare_didrain - categorical, 1 for yes, 0 for no
- tmpdiff - difference between the median temperatures at ohare and the garden ($\text{ohare_medtmp} - \text{garden_medtmp}$)

DATA GATHERING

Scraping –

Lake Michigan Temperature Data

Weather Underground Weather Data (unsuccessfully)

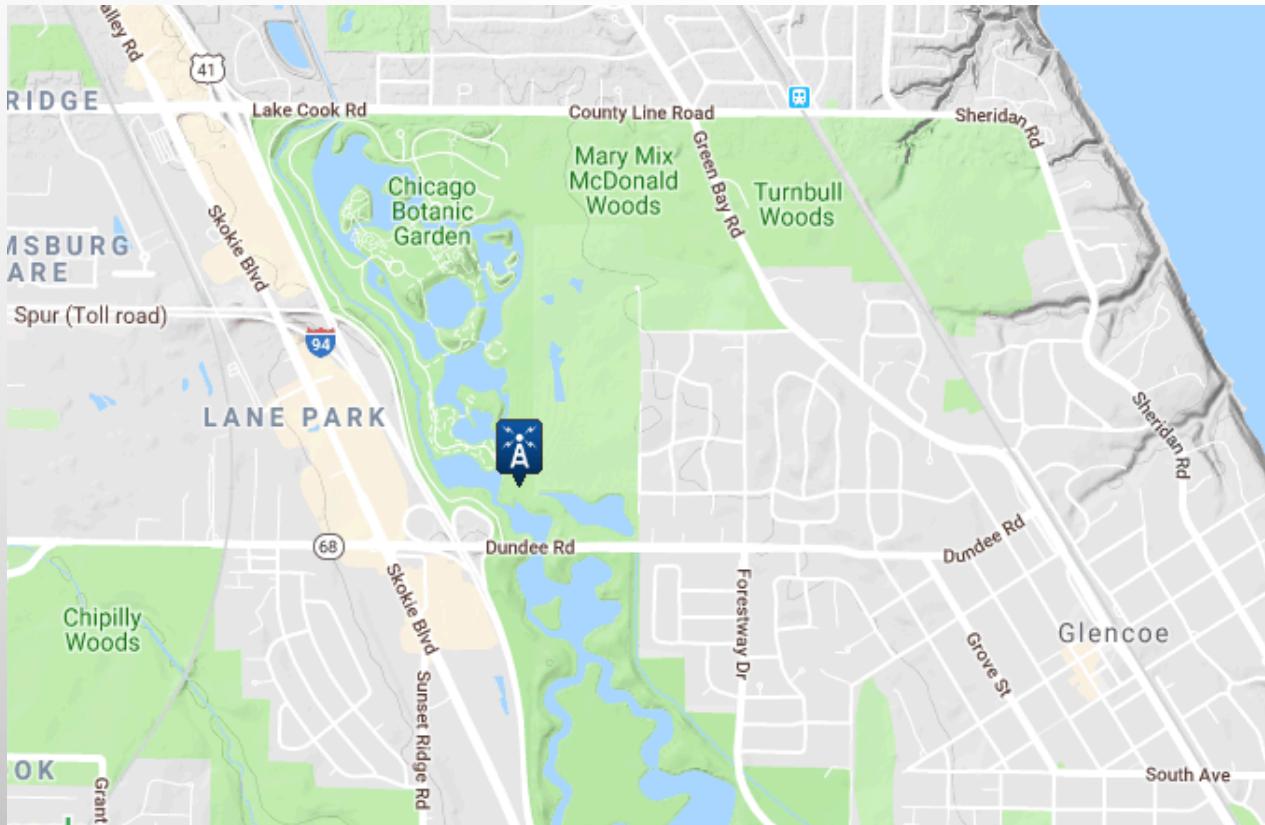
Direct Downloads –

Global Historical Climatology Network .csv files

Web API –

Global Historical Climatology Network has a direct
Application Programming Interface

CLEANING THE DATA

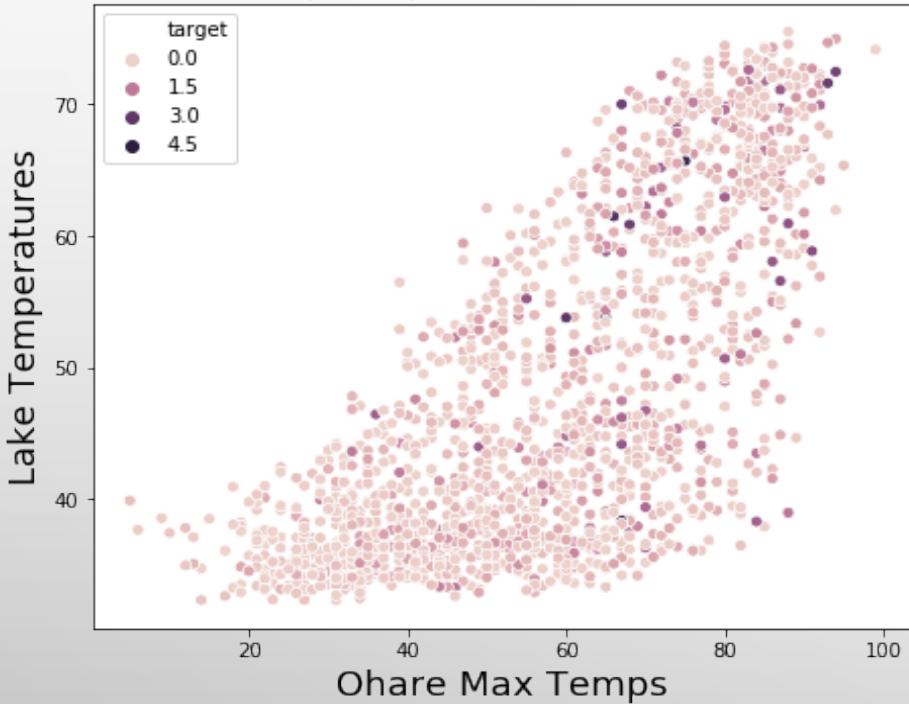


Incomplete Data
determined the choice
of the Botanical Gardens

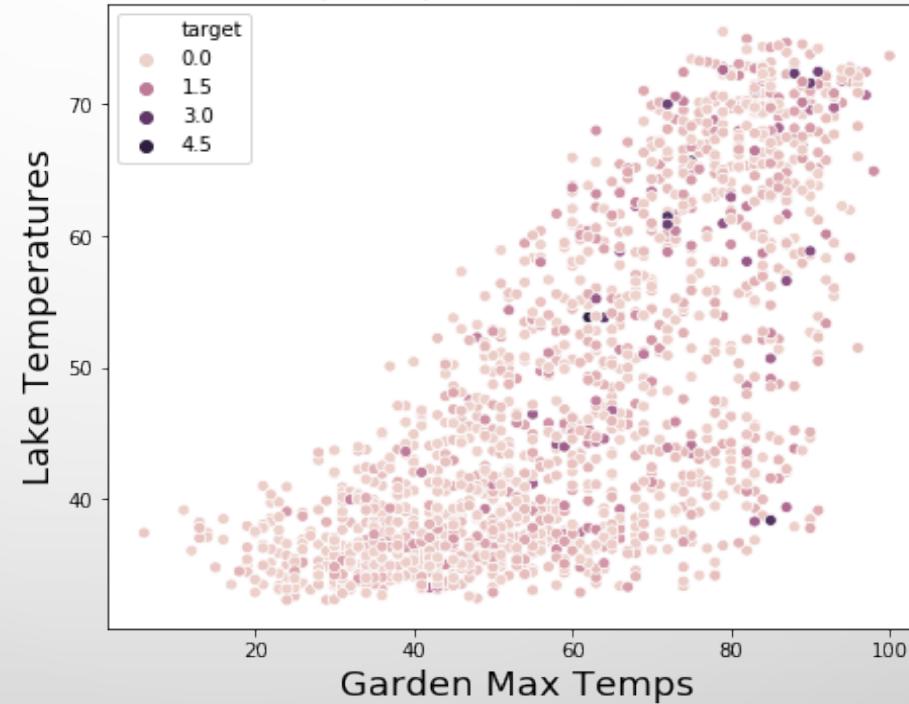
EXPLORATORY DATA ANALYSIS

As temperatures increase and there is precipitation at both locations, the difference in precipitation between locations increases.

Comparing Ohare vs Lake Temperatures
with precipitation differences

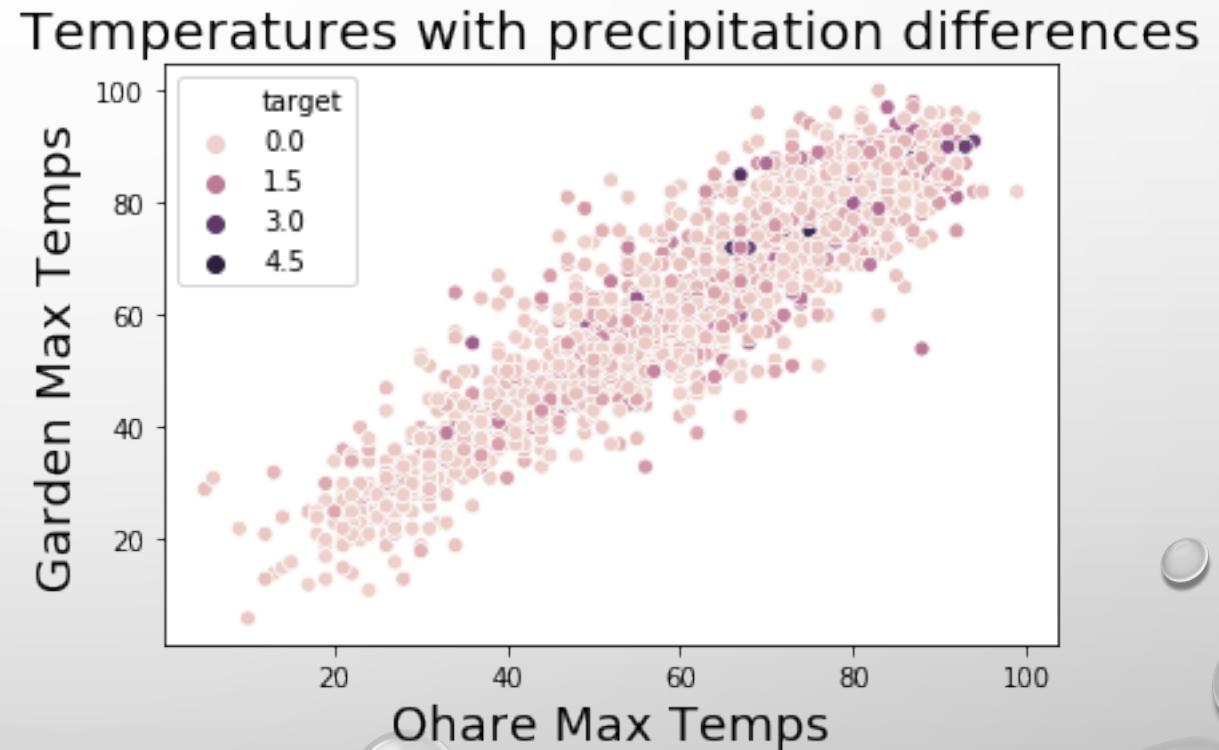
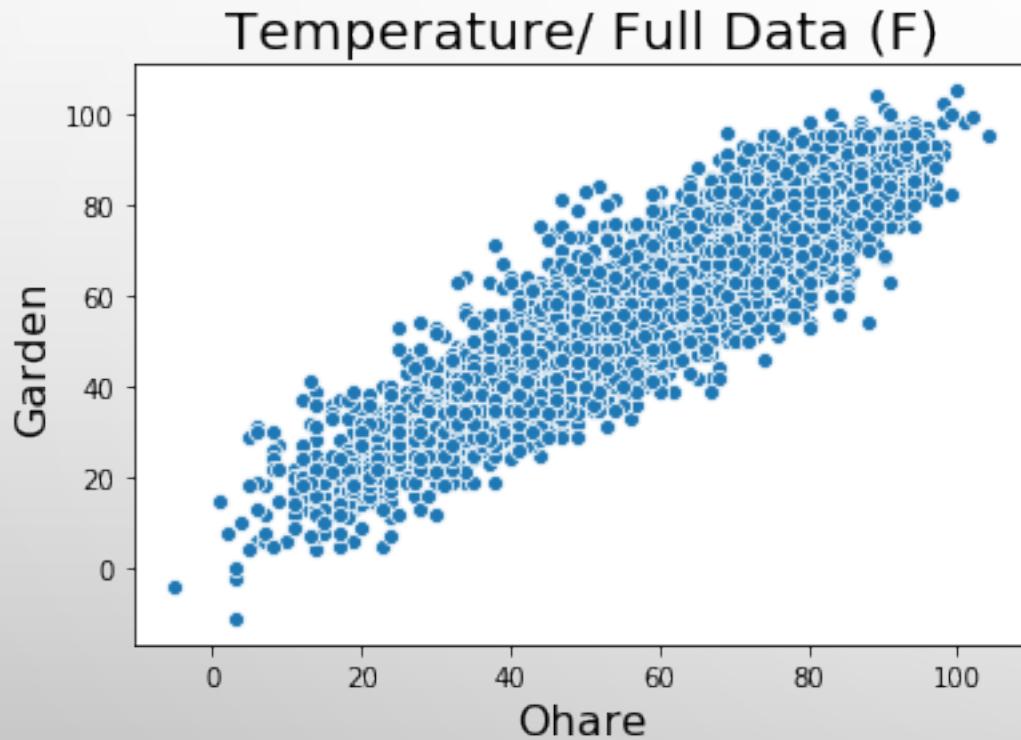


Comparing Gardens vs Lake Temperatures
with precipitation differences



EXPLORATORY DATA ANALYSIS

Again, as the temperatures increase, the difference in precipitation increases.



SIGNIFICANCE TESTING

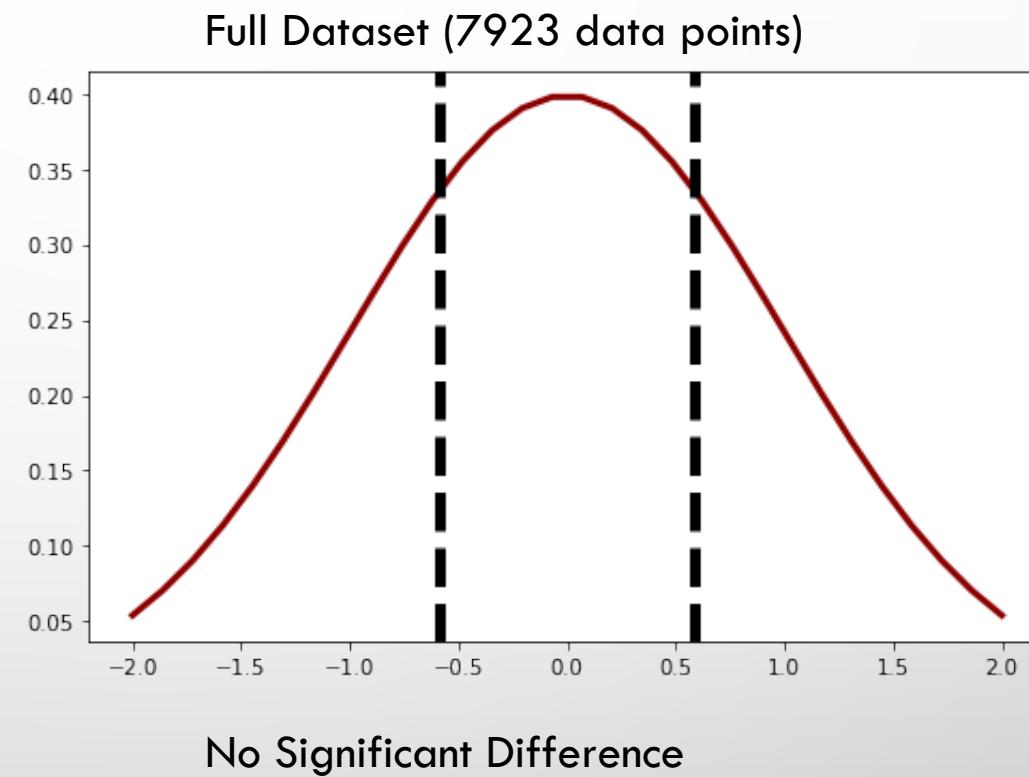
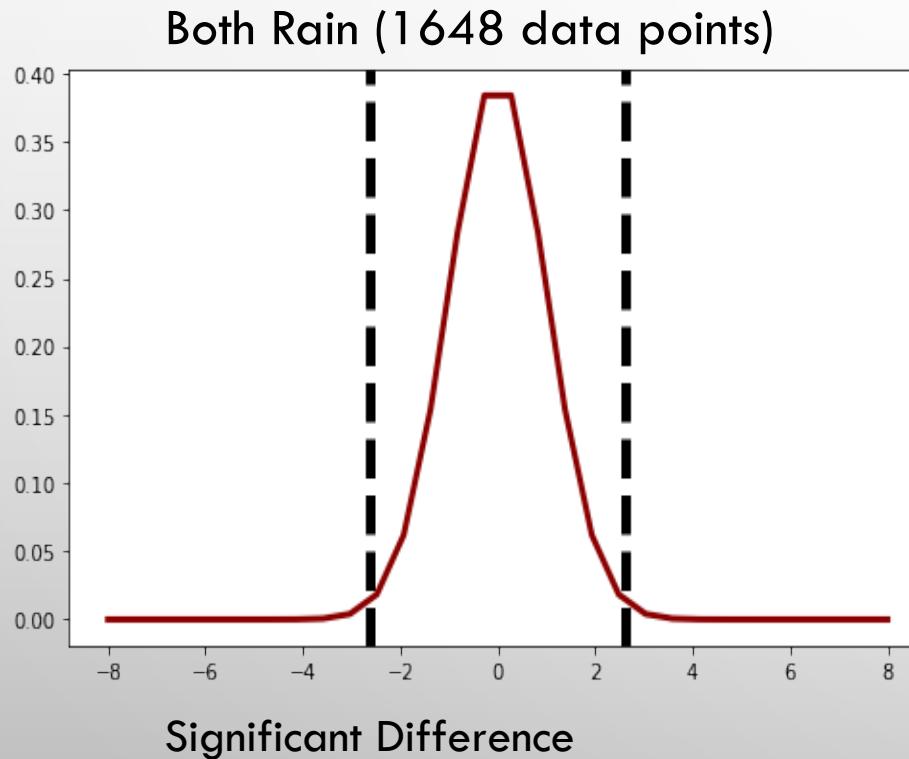
Testing to see if there is any significant difference between the temperatures at O'Hare and the Chicago Botanical Gardens.

Data	Quantity of Data	t-score	p-value	significance	Gardens Avg (F)	Ohare Avg (F)
All Data	7923	0.5876	0.5568	None	59.24	59.43
No Rain	4022	3.285	0.001	Yes	58.99	60.57
Both Rain	1648	-2.629	0.0086	Yes	59.48	57.7
ohareRain	1193	-1.9557	0.0506	None	59.06	57.43
gardensRain	1060	0.0904	0.928	None	59.99	60.07

Statistically significant difference when there is similar precipitation at both locations!

SIGNIFICANCE TESTING

Baseline T-distributions



MODELING

Can a predictive model that explains at least 80% of the variance in the precipitation differences be constructed?

During the modeling process we trained more than 25 models!

The best scores were obtained using a direct Random Forest.

- Full dataset (7923 data points)
- Train/test split (75/25)
- Squared polynomial features
 - Training set shape – (5942, 120)
- `n_estimator = 100`
- `max_features = 'sqrt'`

Sadly, NO.

Not with the given data.

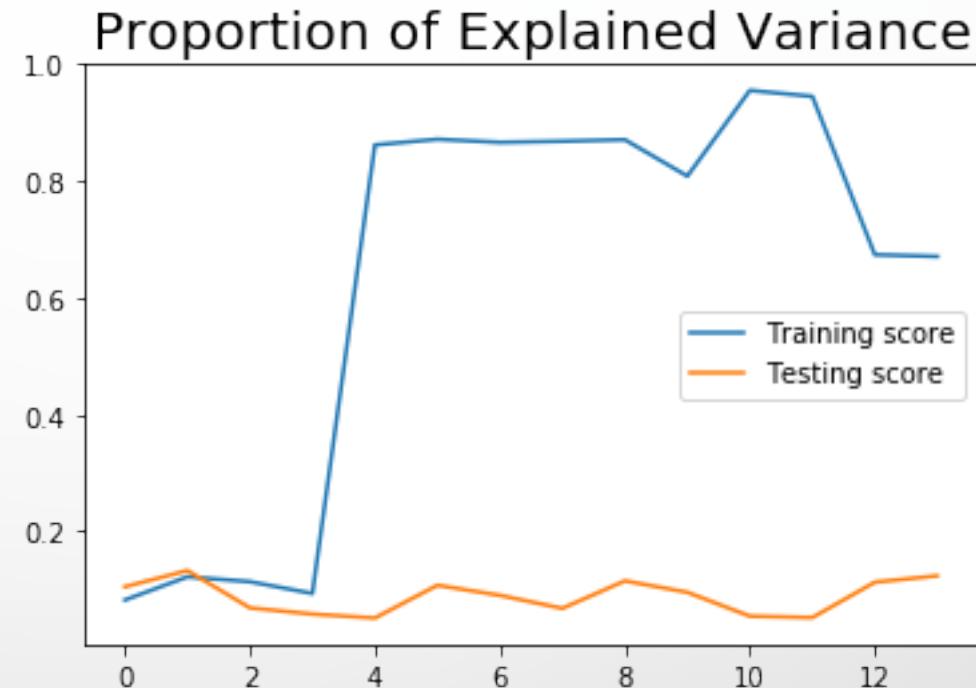
Best Scores:

- Training – 0.8704
- Testing – 0.1153
- Cross Validation – 0.0767

MODELING

Other models used:

- Linear no Poly
- Linear Grid Searches
- Regularized
- Decision Tree
- Boosted Decision Tree
- Boosted Random Forest
- Bagged Random Forest
- Feed Forward Neural Networks



CONCLUSIONS

1. Existence of significant differences in temperature?

- Even though there is a relatively close proximity between the two locations in the Greater Chicagoland area, there are indications that temperature differences do exist under certain weather conditions.

2. Can a predictive model that explains at least 80% of the variance in the precipitation differences be constructed?

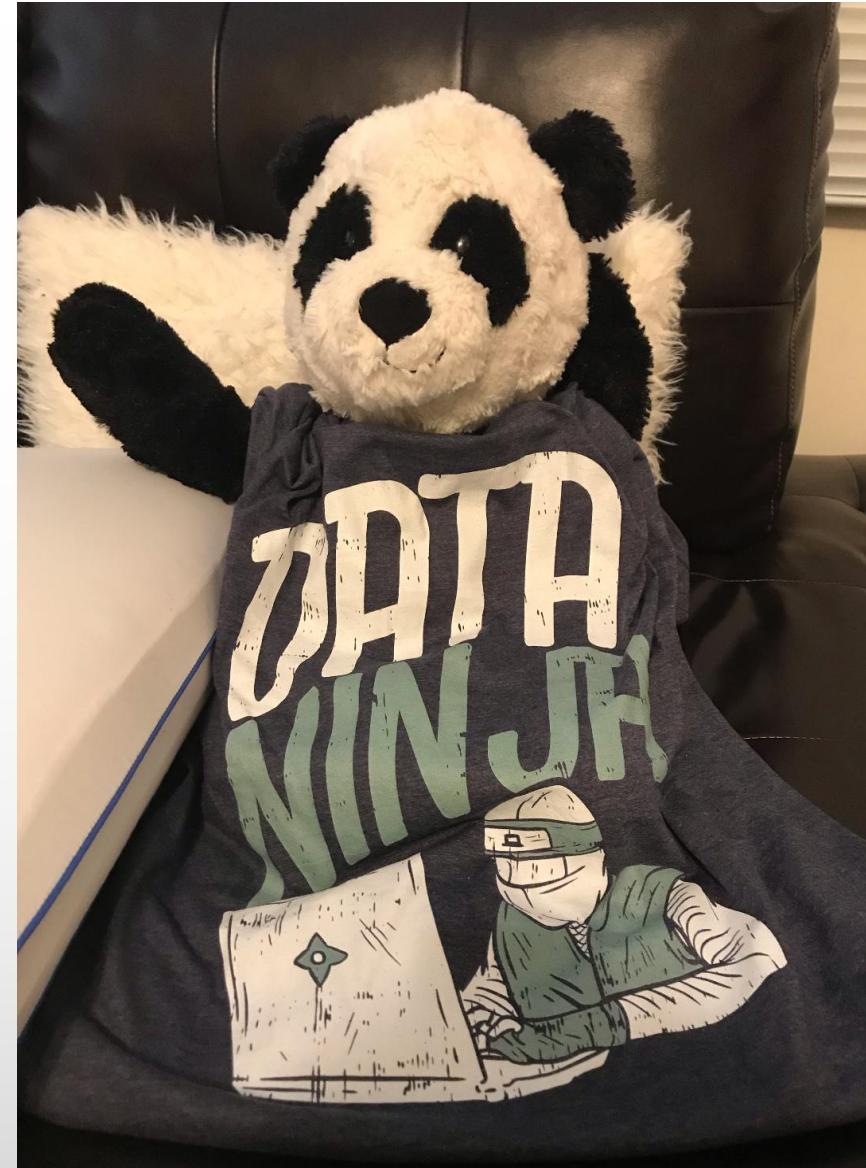
- Not when comparing only the Chicago Botanical Gardens and O'Hare airport over years 1995-2018.

FUTURE ITERATIONS

- The significant differences in temperature when there is similar precipitation activity at both the Gardens and O'Hare should be investigated further. It is possible that there is a seasonal component.
- A similar comparison between other locations close Lake Michigan and the airport should also be performed to determine whether this difference is truly a large scale trend.
- Include more data for modeling. There are hundreds of weather towers, 192 in the Chicagoland area. The next model constructions will be spatio-temporal in nature, utilizing a much wider berth of data to build temperature fields that capture trends over a larger area.

TUX SAYS, THANKS!

For preparing him to go into the
wild world of data!



QUESTIONS