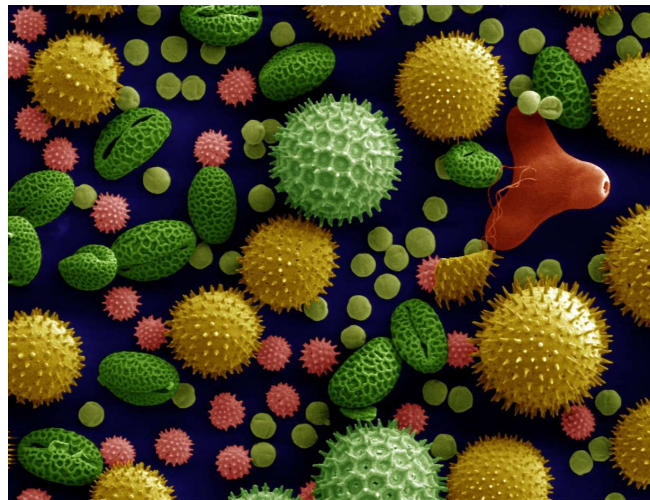


Houston, We Have an *Allergy* Problem



Predicting Pollen Counts in Texas

May, 2022

Nicholas Ross O'Keefe Kennedy



Flat Iron Data Science Graduate
www.github.com/nikennedy
nrokkennedy@gmail.com

Agenda:

1. Introduction & Business Understanding

2. Business Problem

3. Data Understanding

4. Initial Analysis

5. Modeling: *Classification & Time-Series*

6. Results

7. Species-Specific Diagnostic Tool Proposal

8. Further Considerations

9. Conclusion

10. References & Appendix

Introduction:

Allergies & Public Health After Covid-19

1. Heightened Public Awareness of Allergies
2. Seasonal Strain on Healthcare Providers



Allergies: *A Business Problem*



Preparing Medical Providers & the General Public
for Allergy Season Through Public Health Initiatives

The Data: *Pollen*

- All Pollen Data From Houston Health Dept.
- Monthly Tallies of Species-Specific Pollen Counts per m³
- Business Days from January 2013 to April 2022



The Data: *Climate*

- Daily Climate Data from NOAA
- 2 Testing Centers in Houston, TX, 1 from Shreveport, LA
- Average Daily Temp (°F),
Average Wind Speed (mph),
Precipitation (in.)



Pollen Counts & Allergies

TREE POLLEN:

- 90-1499 /m³: *Heavy*
- 1500+ /m³: *Extremely Heavy*

GRASS POLLEN:

- 20-199/m³: *Heavy*
- 200+ /m³: *Extremely Heavy*

WEED POLLEN:

- 50-499 /m³ : *Heavy*
- 500+ /m³: *Extremely Heavy*



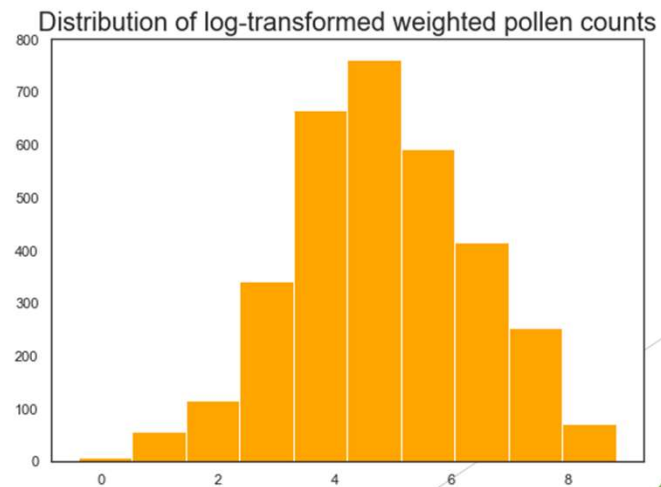
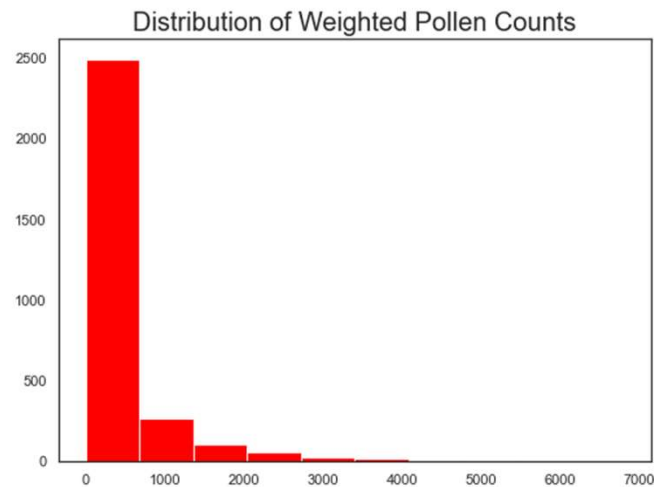
Weighted Pollen Counts

- Weighted using Different Severities of Tree, Grass and Weed
- “high” pollen set at 100+ particles per m³ per HHD
- ~50% of days ‘high’ pollen for Houston, TX

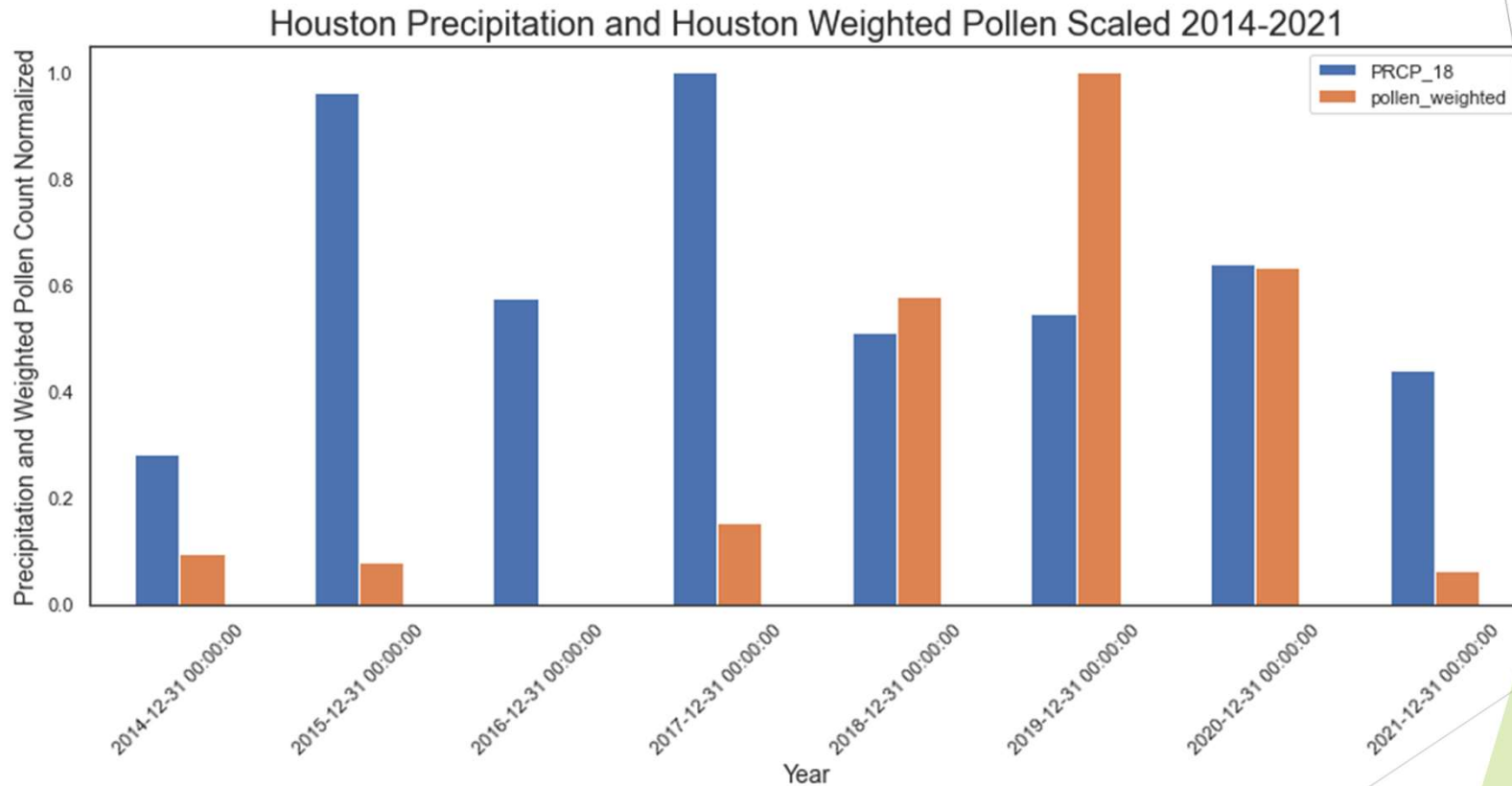


Exploratory Findings

- Many Days had pollen counts of zero especially in summertime
- Max Pollen Count of over 9,000 per m³ occurred in March, 2019
- Pollen's Logarithmic Distribution



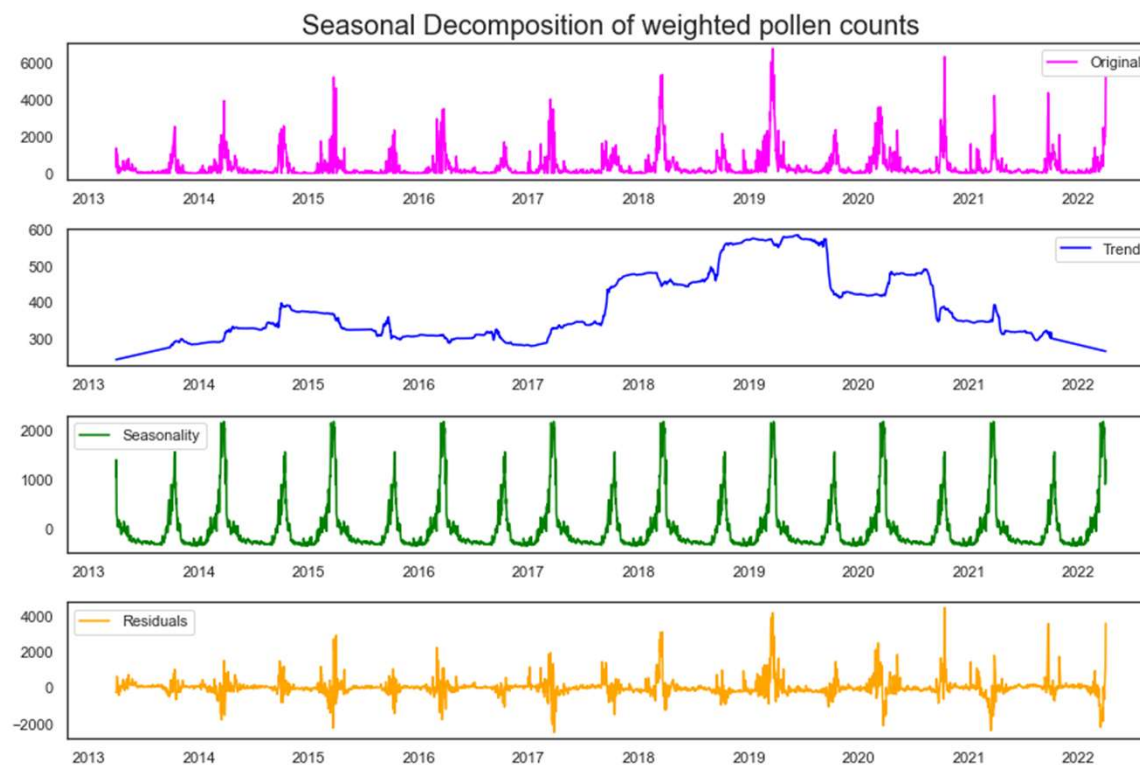
Exploratory Findings



No Clear Correlation Between Local Precipitation* and Pollen Counts for Current or Prior Year

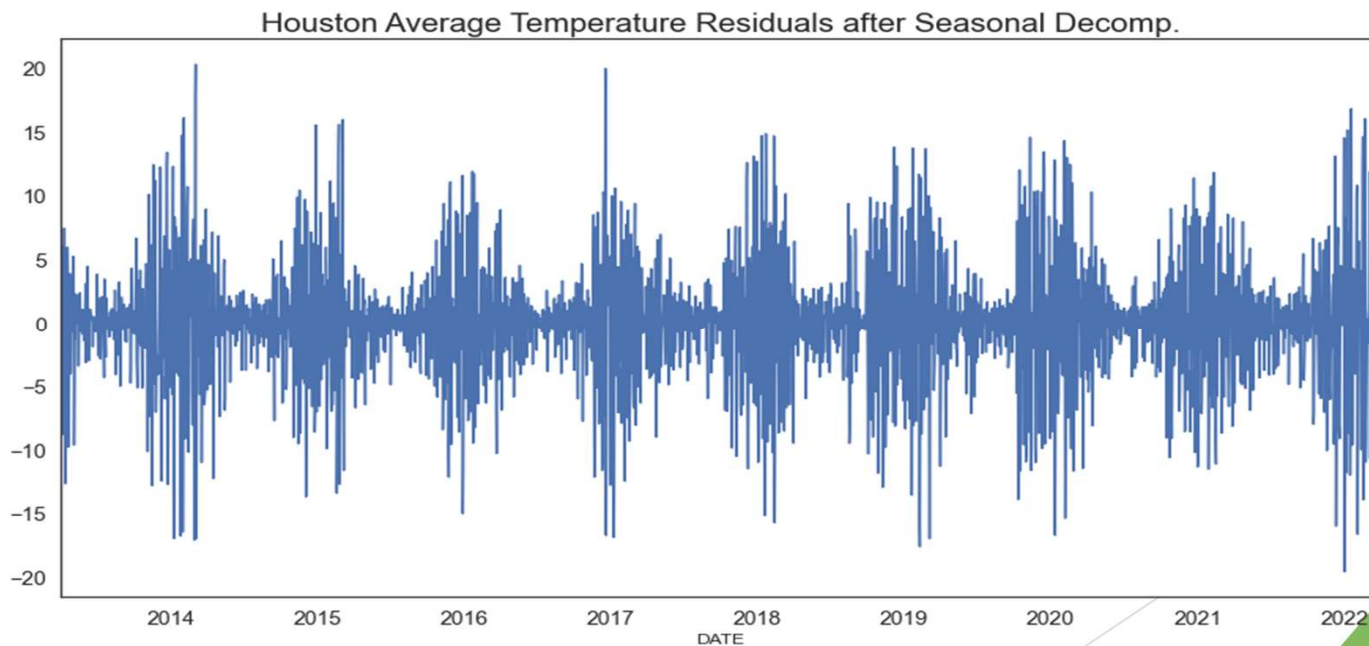
* Also True for Shreveport, & other variables (see Appendix)

Classification Modeling: *The Seasonality Problem*



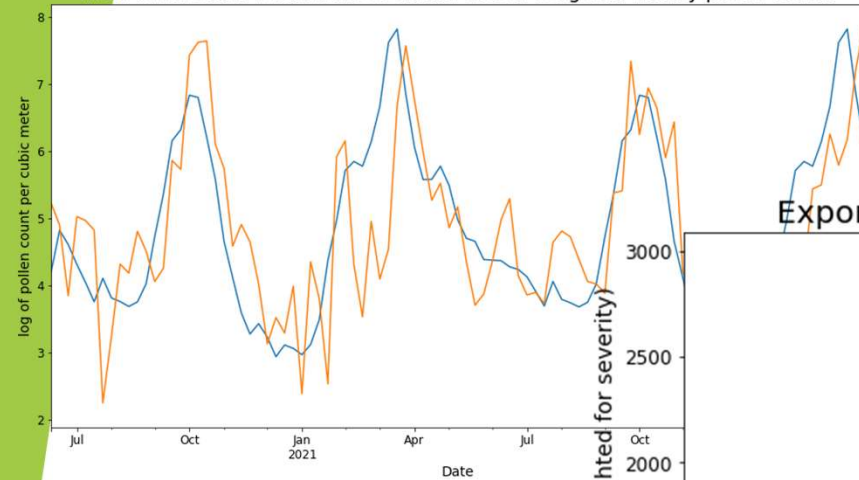
Classification Modeling

- Decision Tree: Local Temp = Most Important Feature
- Random Forest Classifier Achieved 60% Accuracy
- Seasonality persists after decomposition?

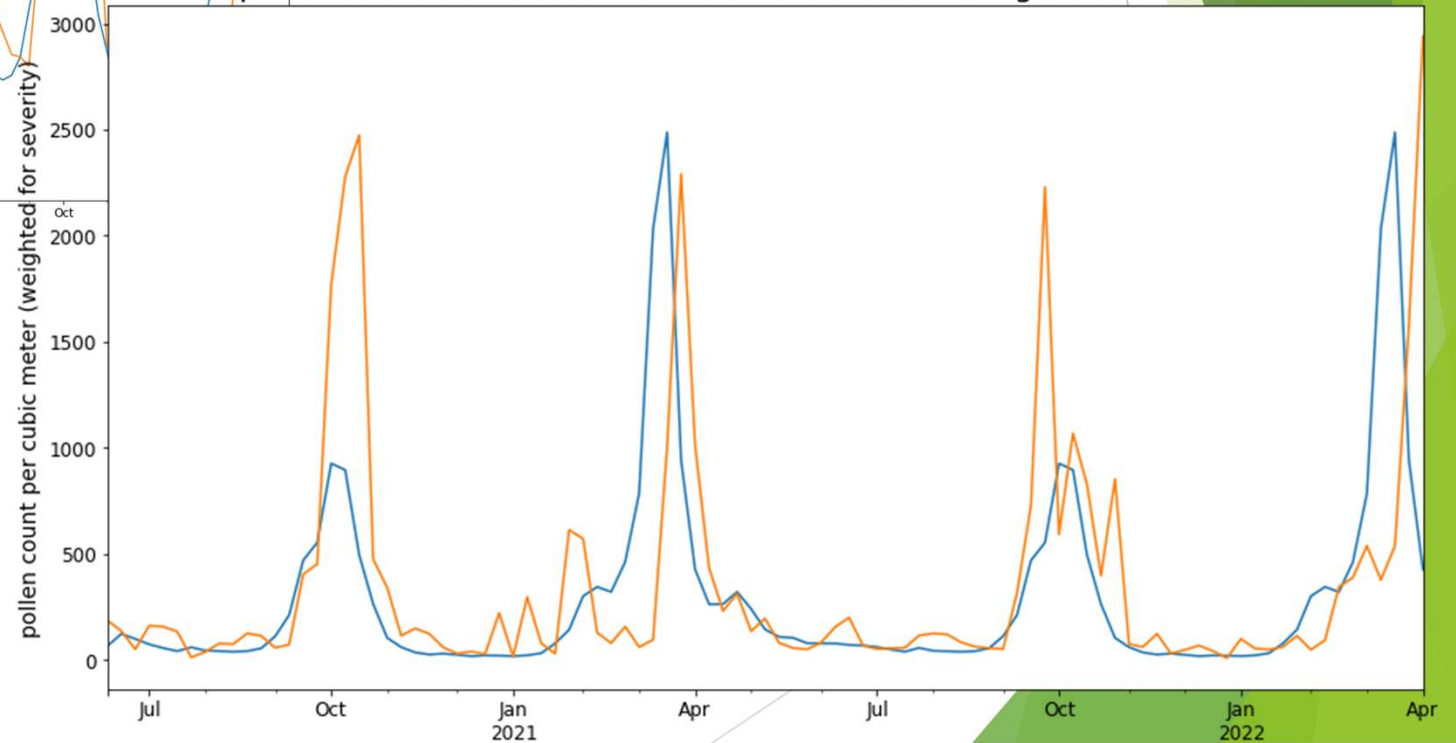


Time-Series Modeling

First Arima Model Predictions vs Actual LOG of weighted weekly pollen counts



Exponentiated First Arima Predictions vs Actual Weighted Pollen Counts



Results

Classification:

- Best Model was Random Forest 60% Accuracy, 59% Precision

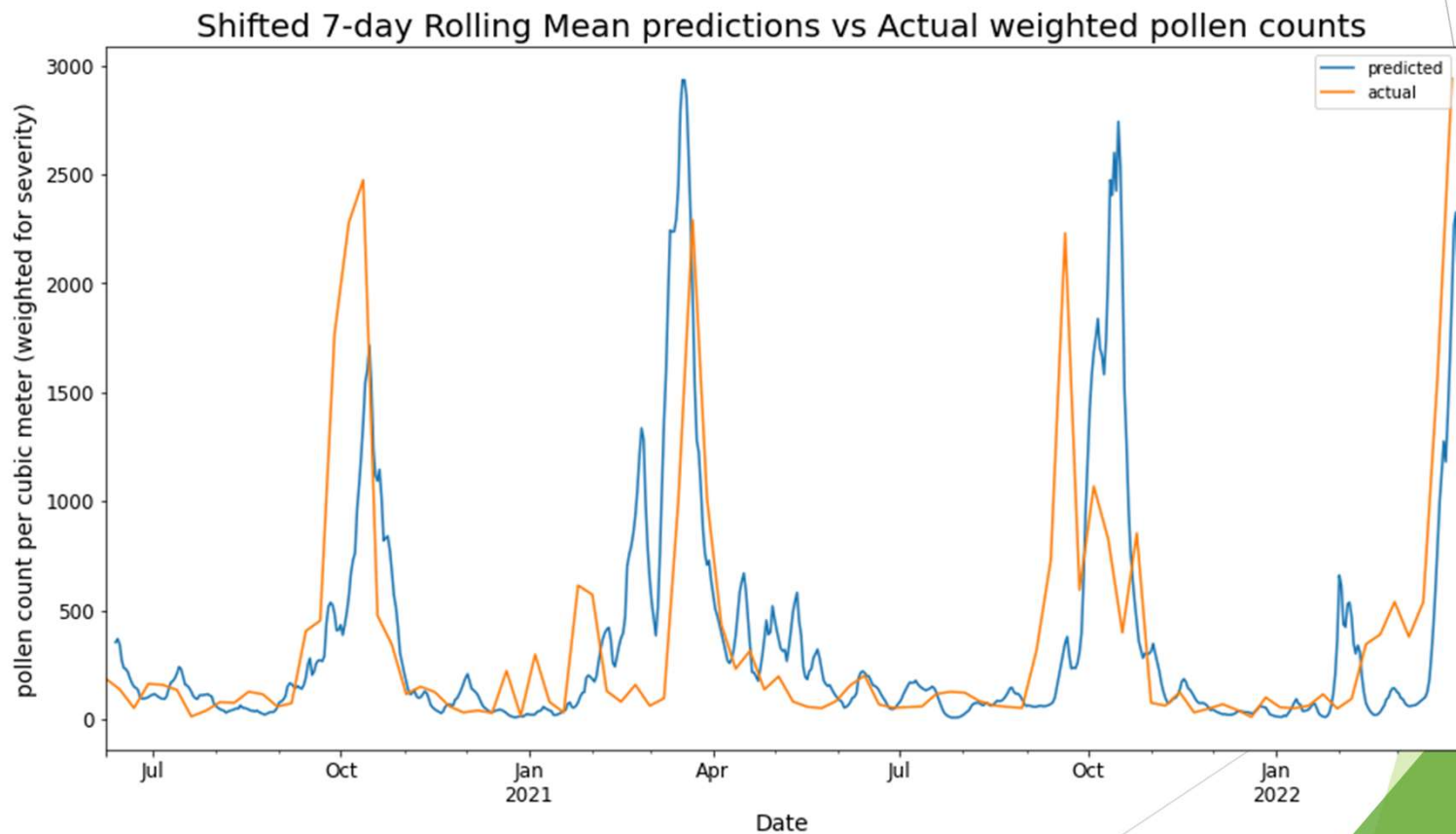
Time Series:

- Best Predictor was Prior Year's Pollen Counts
- Off by ~ 425 particles per m³ on average

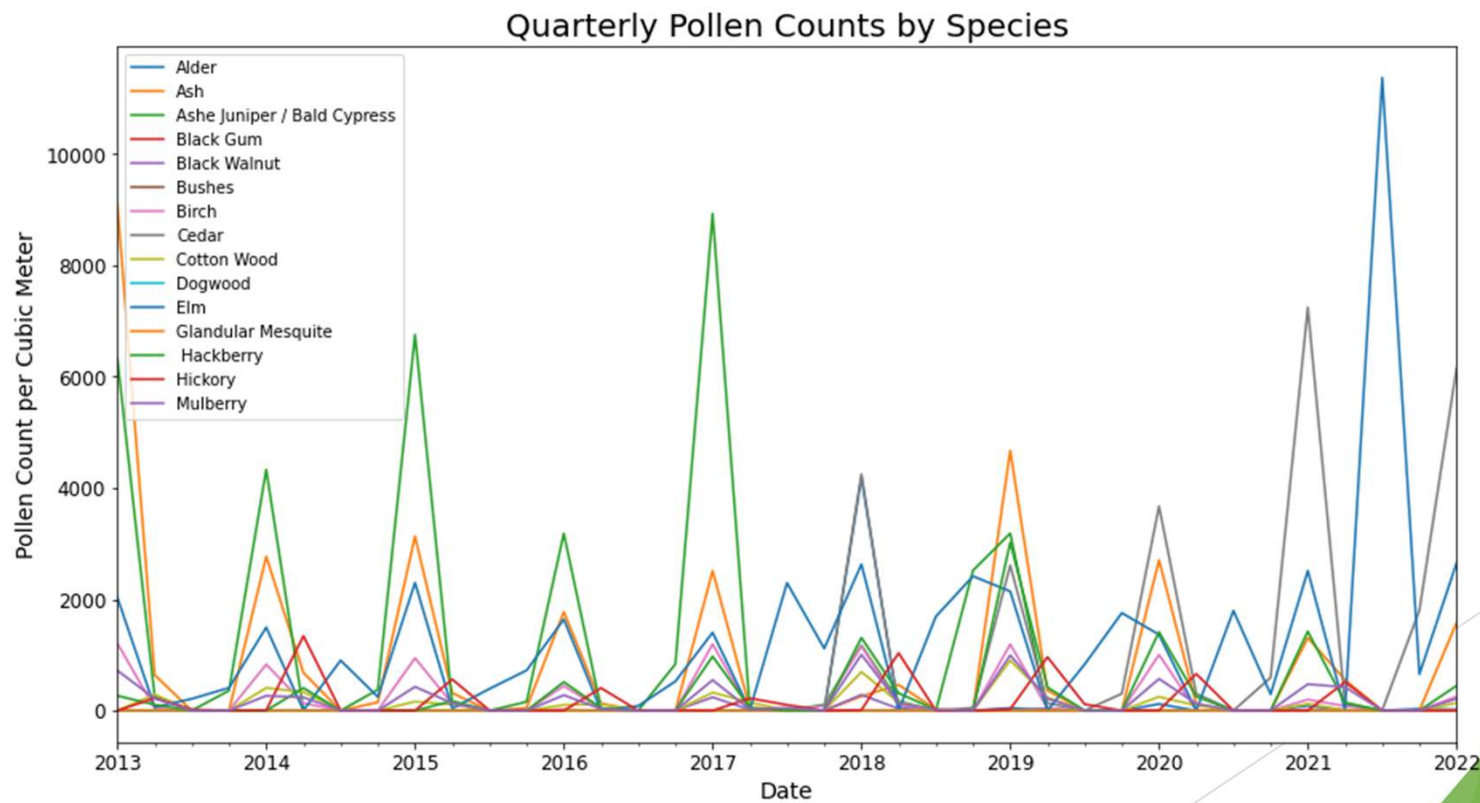


Results:

Best Predictions => Last Year's Data

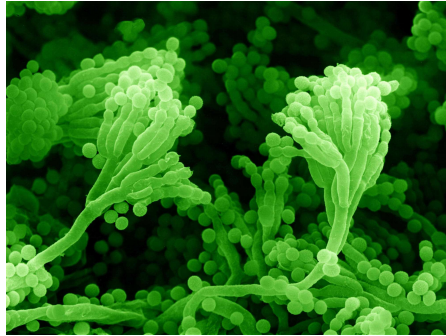


Proposal for Next Steps: *Species-Specific Diagnostic Tool*



Further Considerations:

1. Mold Spores



2. More Local and Statewide Data



Conclusion

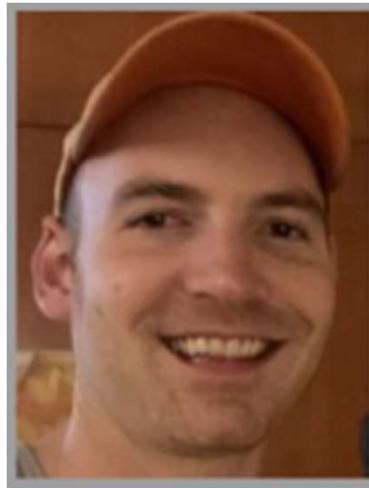
Seasonality is King

- **The local seasonality of pollen production is the best predictor of pollen counts.**



Thanks!

Questions and Comments



Flat Iron Data Science Graduate
www.github.com/nikennedy
nroknedy@gmail.com

Additional References

1. <https://www.chpa.org/sites/default/files/media/docs/2020-10/Assessing-Consumer-Benefits-of-Allergy-Rx-OTC-Switches-03012017.pdf>
2. <https://www.ochsner.org/services/allergy-asthma-and-immunology/pollen-mold>

Appendix

