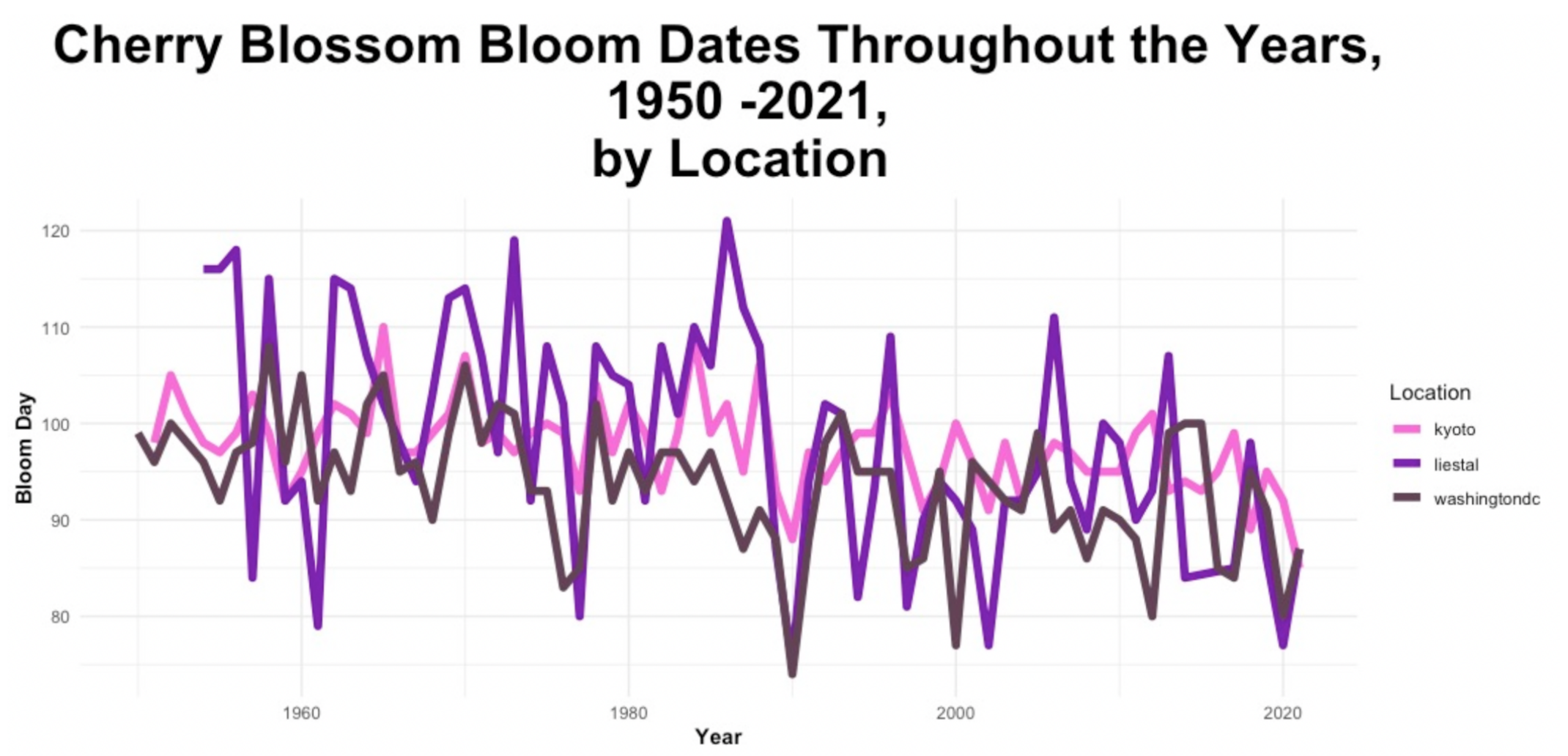
- CHERRY BLOSSOM PEAK BLOOM PREDICTION -

Stephanie Plaza, Elise Rust, Amalia Stahl, Adiam Tesfaselassie

I. Introduction

Cherry blossoms are an icon of spring and are celebrated in many cultures and cities. The peak Cherry Blossom bloom date is defined as the day when 70 percent of the blossoms are in full bloom (the United States Environmental Protection Agency, 2021). Peak bloom in Washington, DC, Kyoto, Japan, and Liestal, Switzerland, occurs typically between mid-March to mid-April. The entire blooming period can last up to 14 days, including the days leading up to peak bloom. Since 1921, peak bloom dates have shifted by approximately six days earlier, and peak bloom dates for cherry trees continue to occur earlier than they did in the past (United States Environmental Protection Agency, 2021).

Figure 1: Time Series



Cherry blossoms' peak bloom dates can be used as indicators of climate change. Chung, JI, and SH (2011) state that due to cherry trees' sensitivity to winter and early spring temperatures, the timing of cherry blossoms is an ideal indicator of the impacts of climate change on tree phenology. Similarly, JH et al. (2015) found that in the Korean Peninsula, rapid temperature hikes in late spring likely brought forward the peak bloom dates of cherry blossom. Shi et al. (2017) also found that rising winter low temperatures delay the first flowering time.

Figure 2: Scatterplot

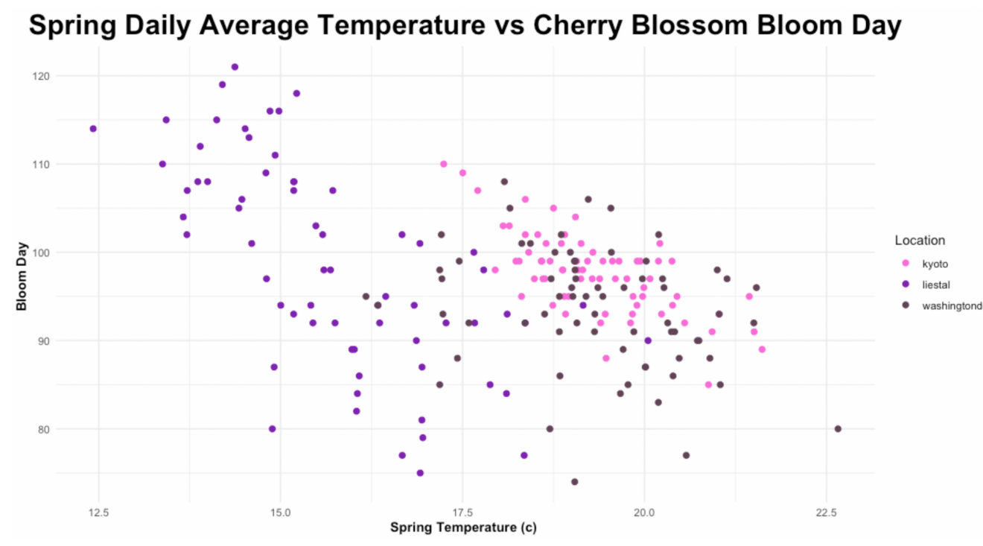
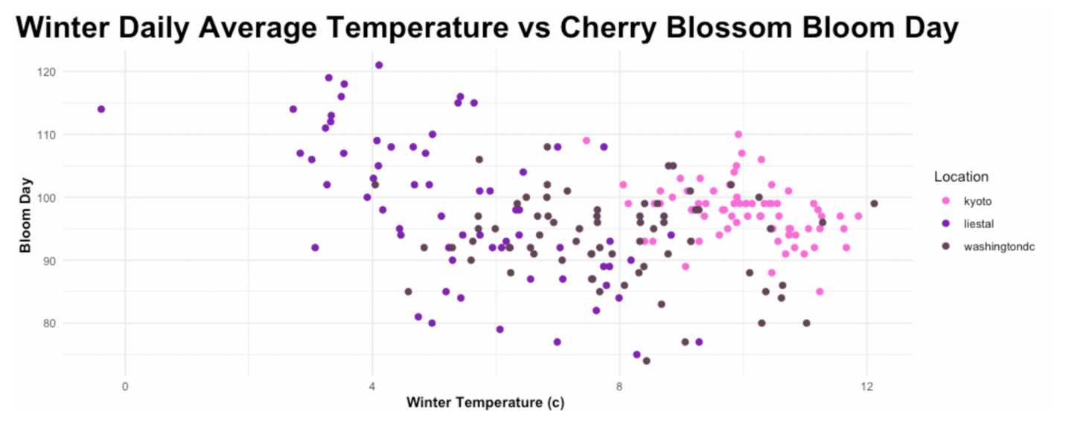


Figure 3: Scatterplot



Goals

The goal of this project is to predict the peak bloom day in 3 day different locations.

Research Questions

I. What at variables are the best predictors of cherry blossom peak dates?
II. Using regression, decision tree, random forest, and gradient boosting which location has the best prediction results?
III. With given variables, between regression, decision tree, random forest, and gradient boosting, which model is the highest predictive power?

Methodology

Dat
a

Cherry blossom peak bloom prediction competition data

Linear Regression

+ simple
- overfitting

Decision Tree

+ easy to interpret
- overfitting

Random Forest

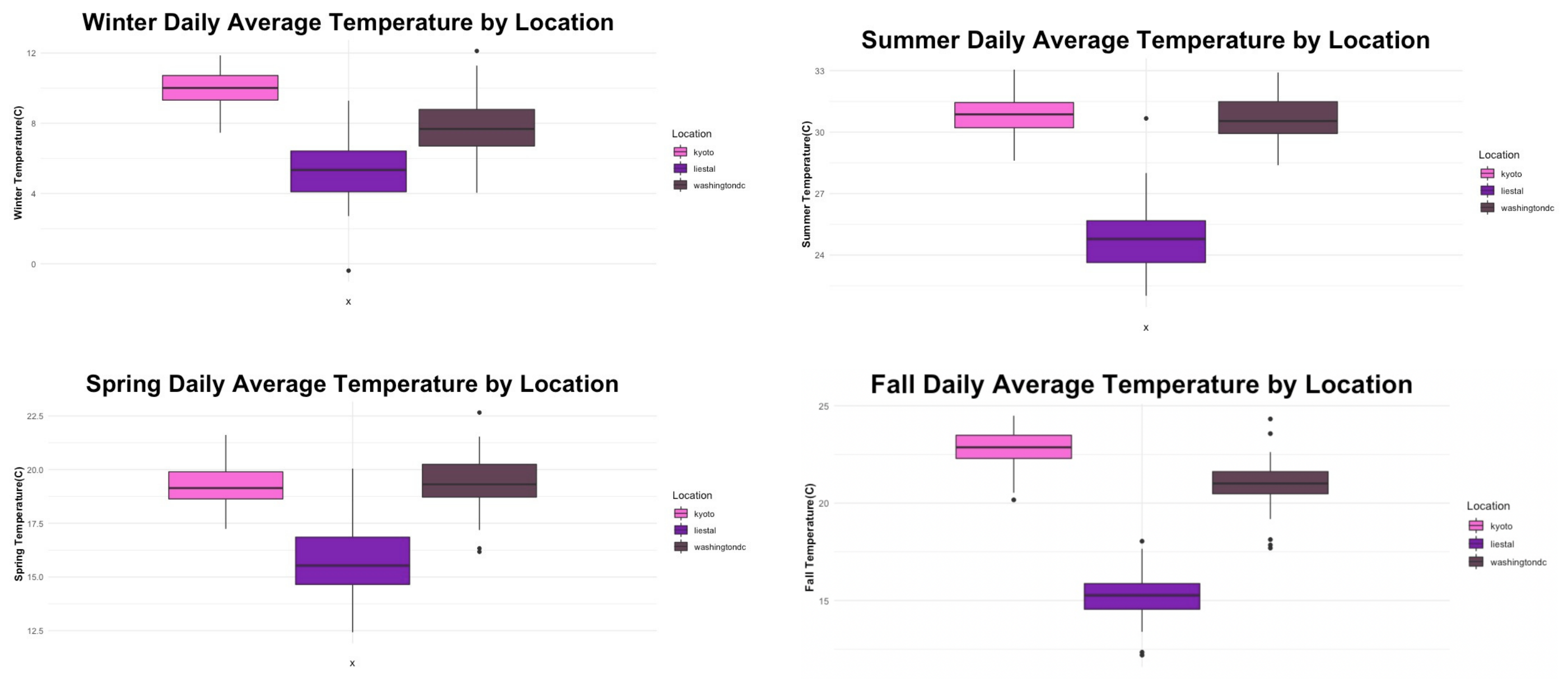
+ strong prediction
- biased towards categorical variables

Gradient Boosting

+ decrease bias error
- overfitting

Exploratory Data Analysis

Figure 4: Boxplot EDA



II. Methods

Linear Regression

Given it's simplicity, Linear regression was run first in order to predict peak bloom date. 3 regressions were run in order to account for the different peak bloom dates of each city and best model subset selection was used in order to determine the best model for each city. The results of both the regressions and the model selections are included in the tables down below.

Figure 5: Linear Regression Results

City	Number of Variables	Estimate Coef.	Year Coef.	Winter Temp Coef.	Spring Temp Coef.	Summer Temp Coef.	Regression Formula	Adj. R^2	RMSE
Kyoto	4	232.73	-0.04	-1.07	-3.59	0.72	y = 232.72836 + (-0.03894*year) + (-1.07249)*winter + (-3.58927*spring) + (0.71605*summer temp)	0.27	3.73
Liestal	2	192.60	X	-2.42	-5.20	X	y = 192.6047 + (-2.4208* winter) + (-5.2039* spring)	0.43	8.96
Washington DC	2	593.18	-0.24	X	-1.70	X	y = 593.18407 + (-0.23576* year) + (-1.69682 *spring)	0.28	7.42

Figure 6: Model Selection

City	Model 1	Model 2	Model 3	Model 4	Model 5	Adjusted R^2	Final Model
Kyoto	Spring temp	Winter temp + Spring temp	Winter temp + Spring temp + Summer temp	Winter temp + Spring temp + Summer temp + Year	Winter temp + Spring temp + Summer temp + Year + Fall Temp	1. 0.4509852 2. 0.4892900 3. 0.4943406 4. 0.5009383 5. 0.4965139	4
Liestal	Spring temp	Winter temp + Spring temp	Winter temp + Spring temp + Fall Temp	Winter temp + Spring temp + Fall Temp + Summer Temp	Winter temp + Spring temp + Fall Temp + Summer Temp + Year	1. 0.4583858 2. 0.5488743 3. 0.5482505 4. 0.5446729 5. 0.5433808	2
Washington DC	Year	Year + Spring temp	Year + Spring temp + Winter temp	Year + Spring temp + Winter temp + Summer temp	Year + Spring temp + Winter temp + Summer temp + Fall temp	1. 0.1985136 2. 0.2938646 3. 0.2915135 4. 0.2903028 5. 0.2755351	2

Decision Trees

Decision Tree regression was used to predict Day of Peak Bloom using all season's temperatures, as well as location data. Cross Validation was used to optimize the tree complexity and select the number of terminal nodes that minimize classification error rate.

Results Summary:

1. MSE = **50.89**
2. RMSE = **7.134**
3. Terminal Nodes after Pruning = **8 nodes**

Figure 7: Pruning Cross-Validation

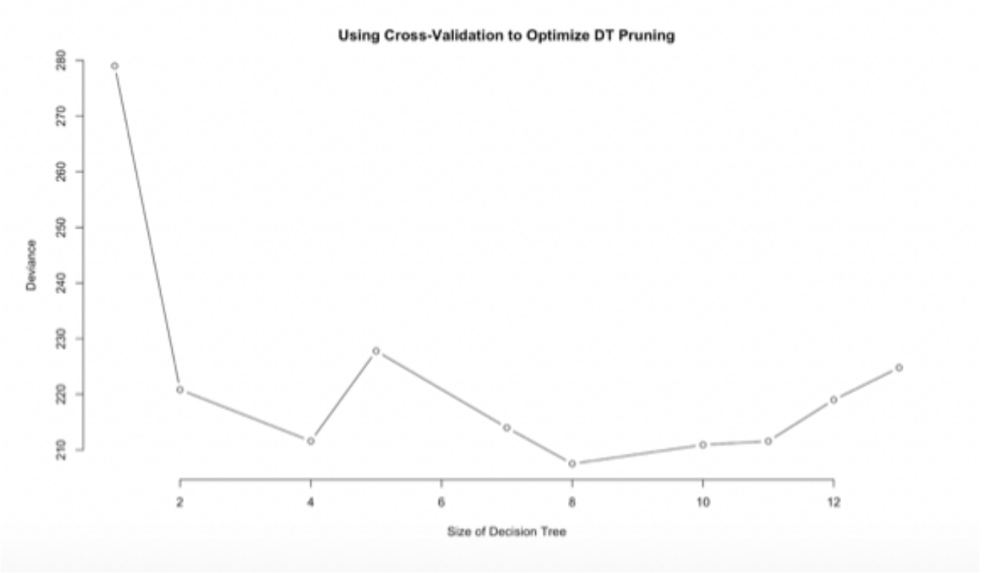
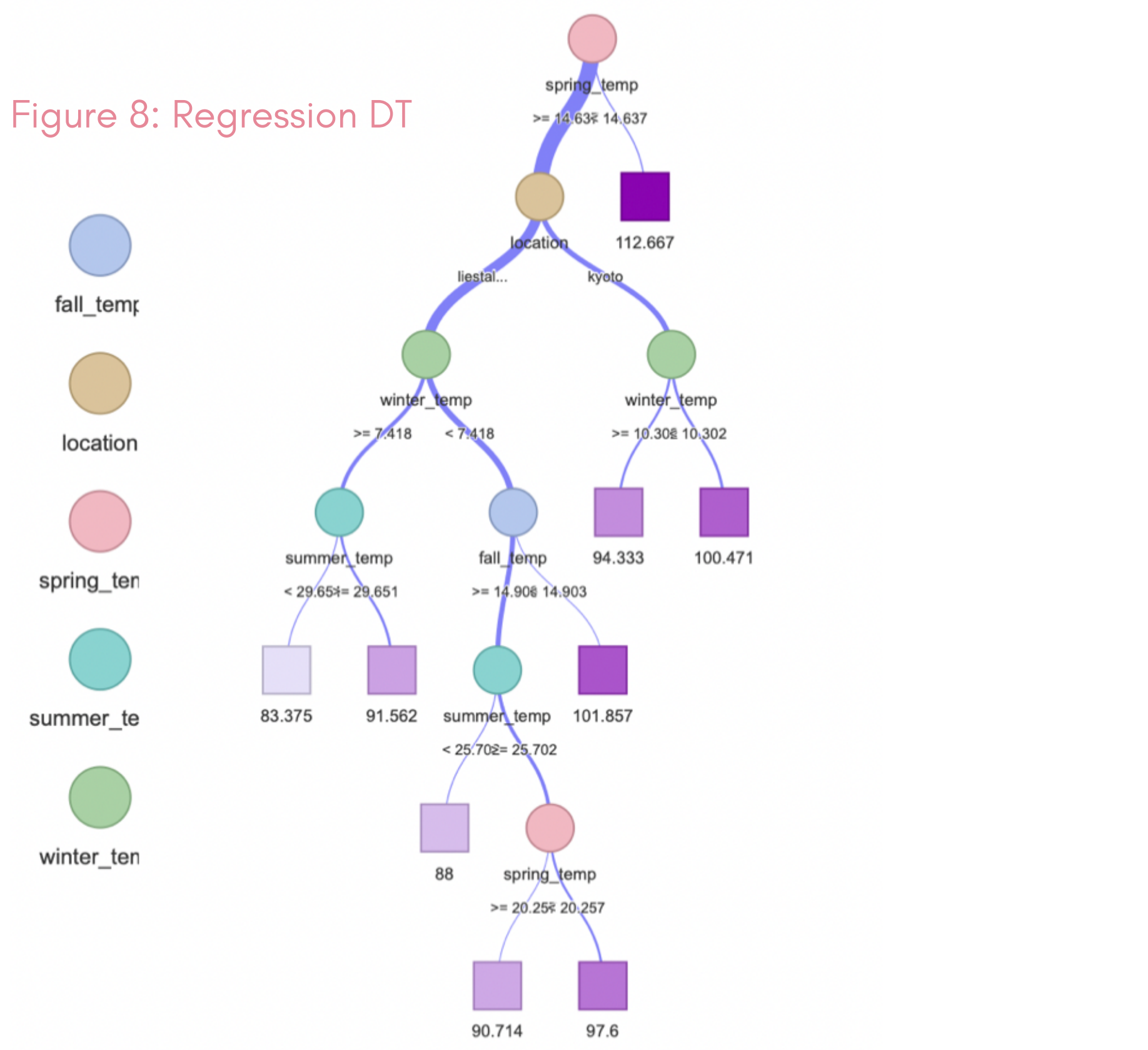


Figure 8: Regression DT



Random Forest

Random Forest was employed to address some of the pitfalls of DT - namely how coarse the splits were for just one tree. Grid Search was employed to identify the optimal number of trees to grow and the optimal number of variables randomly sampled as candidates at each split that minimize MSE. Then, quantified variable importance was derived by looking at how each variable improved node purity, as well as how they influenced MSE.

Results Summary:

1. MSE = **34.664**
2. RMSE = **5.887**
3. # of Trees = **25**
4. Candidate Variables = **2**

Figure 9: Variable Importance

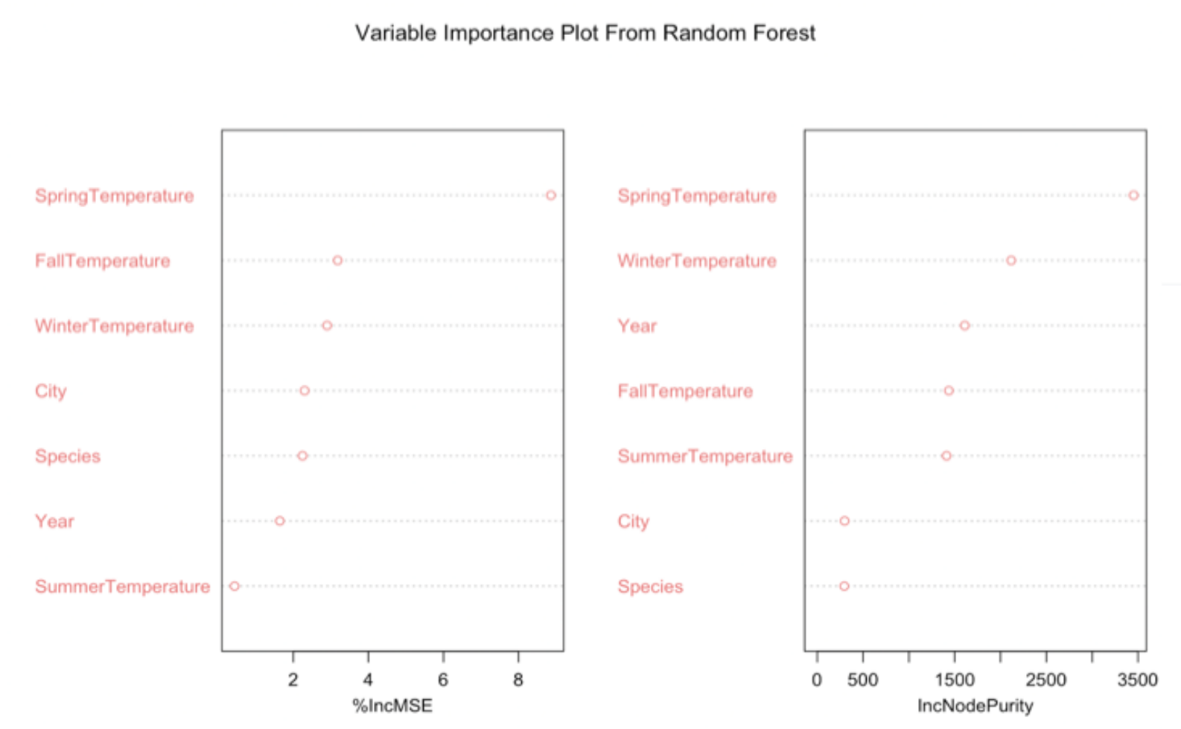
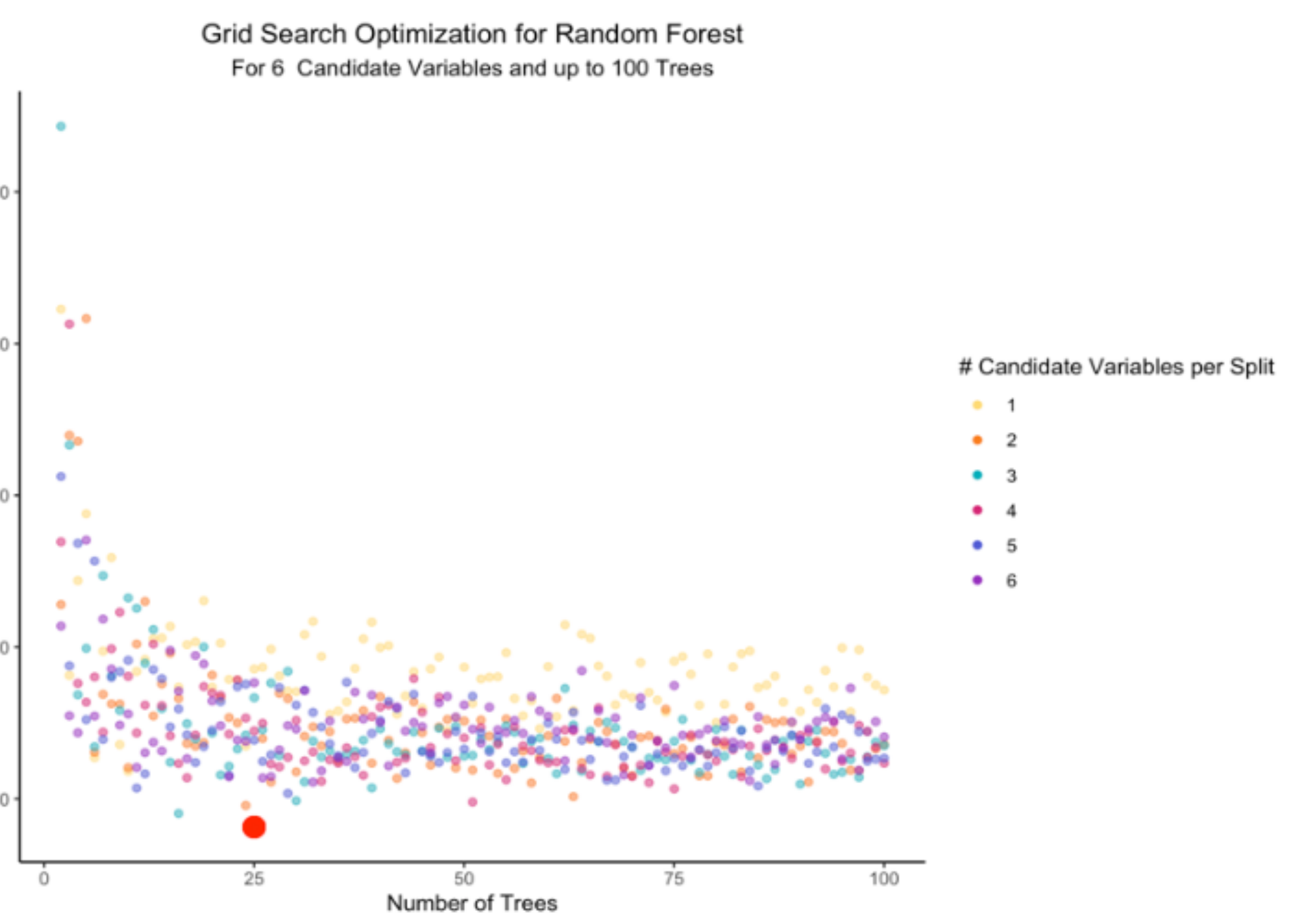


Figure 10: Grid Search



Most Important: **Spring Temperature**

- Increase in Node Purity = **3453**
- % Increase in MSE = **8.86**

Followed closely by Fall Temp, Winter Temp, Year

Gradient Boosting

There are mainly two types of error, bias error and variance error. Gradient boost algorithm helps us minimize bias error of the model. This algorithm starts by building a decision stump and then assigning equal weights to all the data points. Then it increases the weights for all the points which are misclassified and lowers the weight for those that are easy to classify or are correctly classified. A new decision stump is made for these weighted data points. The idea behind this is to improve the predictions made by the first stump. Upon creating this model, we can see that winter and spring temperature are the most important factors in determining the bloom day of year. However, the line plots below show us that the correlation is not strong (-0.30 and -0.44, respectively). Spring temperature has a slightly stronger negative correlation to bloom day of year.

Figures 11 and 12: Relationship Dependence

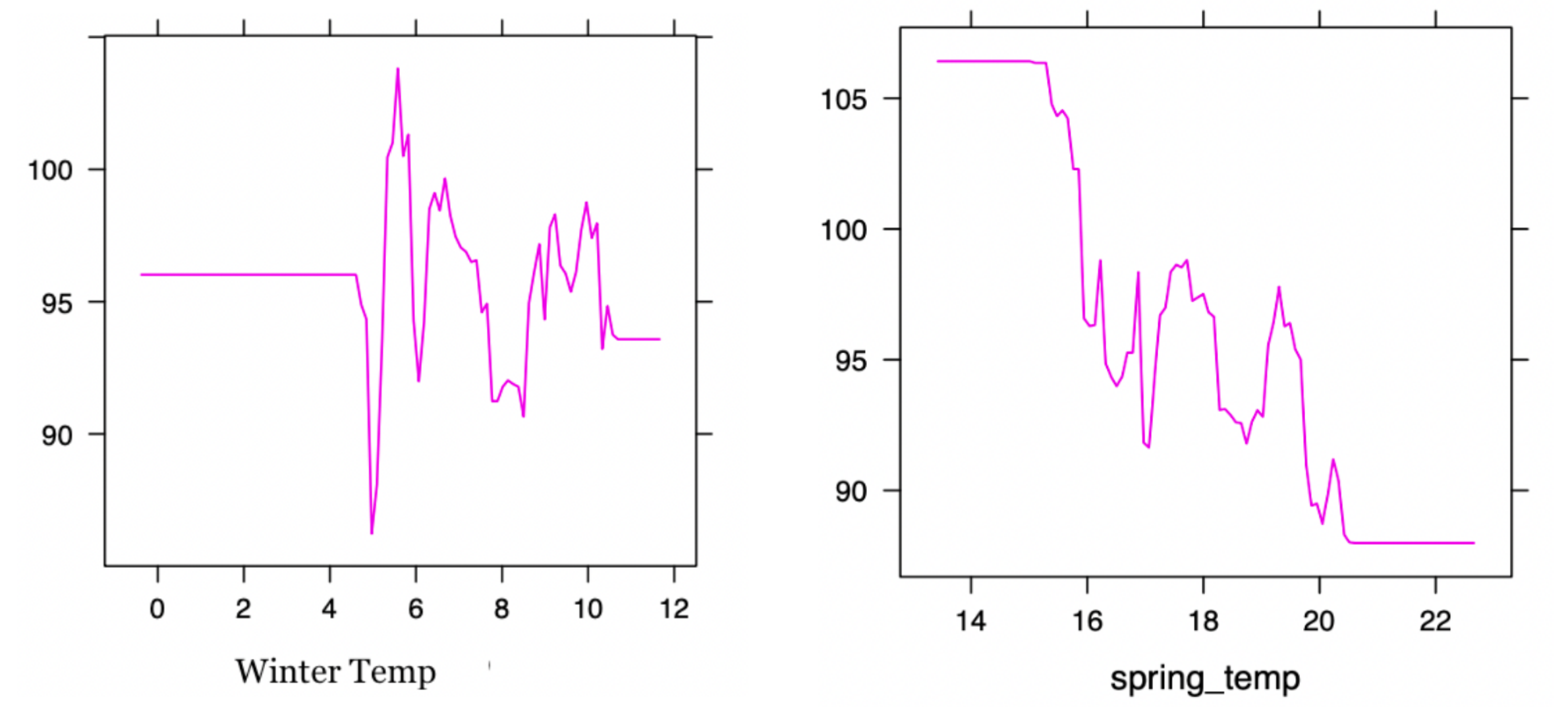
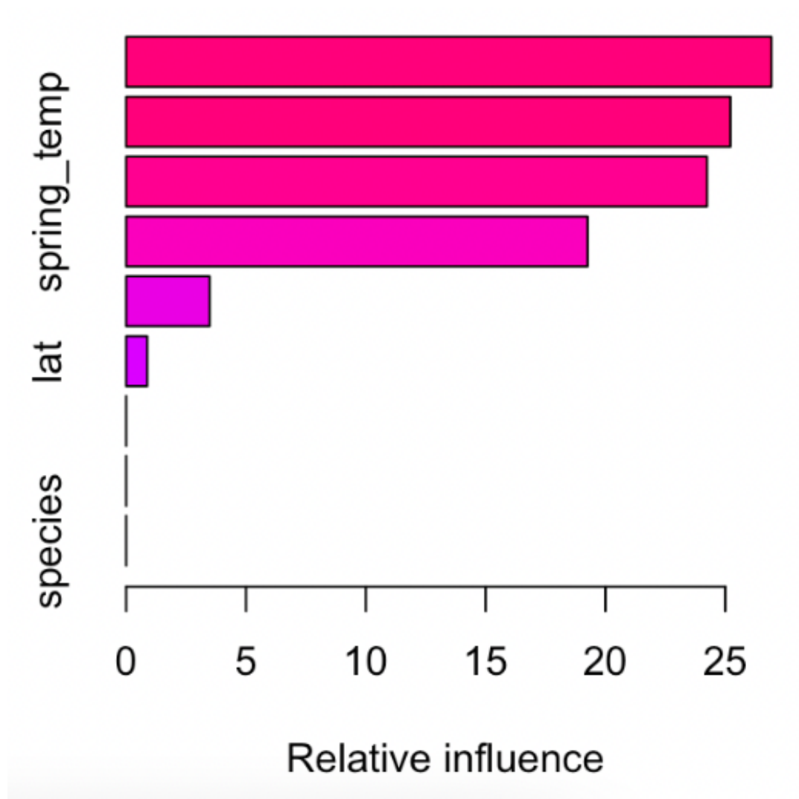


Figure 13: Variable Importance



III. Results

In order to best compare the different models, a summary table was created comparing predicted peak bloom for each model and location along with the residual mean square error.

Figure 14: Summary Table

City	Actual Peak Bloom	Predicted Peak Bloom	Model	# Days Off	RMSE
Kyoto	April 3rd	April 5th	Linear Regression	2	3.70
Liestal	April 3rd	April 6th		3	8.96
Washington	March 21st	March 25th		4	7.42
Kyoto	April 3rd	April 7th	Decision Tree	4	7.134
Liestal	April 3rd	April 7th		4	7.134
Washington	March 21st	April 7th		18	7.134
Kyoto	April 3rd	April 6th	Random Forest	3	5.887
Liestal	April 3rd	April 2nd		1	5.887
Washington	March 21st	March 31st		10	5.887
Kyoto	April 3rd	April 9th	Gradient Boosting	6	83.10
Liestal	April 3rd	April 9th		6	84.83
Washington	March 21st	April 3rd		13	79.23

IV. Conclusions

The task of predicting cherry blossom peak bloom is notoriously difficult given how dependent it is on the 10 days of spring temperatures preceding the bloom. However, our goal was to employ a variety of statistical models to assess how accurately they could predict bloom based on historical data and compare the effectiveness of each.

From this methodology, we concluded that linear regression analysis and random forest models were the most appropriate for the task, and were able to derive peak bloom date predictions across three cities (Washington D.C., Kyoto, Liestal) within 2-4 days of actual 2022 dates.

Ranking the models

1 Linear Regression

2 Random Forest

3 Decision Tree & Gradient Boosting

Important Variables

Winter Temperature

Spring Temperature

Year

Key Takeaway

Thus, our key takeaway is that, while difficult, predicting the date of peak cherry blossom bloom is not impossible! Once key temperature data has been collected it is possible for tourists looking to visit these cities, DC students seeking adventure in their city, or business owners preparing for increased tourist activity to ascertain a date of peak bloom, give or take a few days, by employing statistical models!