# Predicting Cherry Blossom Full Bloom Timing in Japan: Geographic Location and Initial Flowering Date Enable Two-Day Accurate Forecasts[*]

Kevin Z Shen

December 3, 2024

Cherry blossoms hold immense cultural significance in Japan, making accurate prediction of their full bloom timing valuable for both cultural celebrations and tourism planning. Using historical data spanning from 1953 to 2019 across 100 observation stations in Japan, we developed a mixed effects regression model that combines initial flowering dates, local temperature data, and geographical coordinates to predict cherry blossom full bloom timing. Our analysis shows that flowering dates and local conditions can predict full bloom timing with remarkable precision, achieving an average accuracy of two days, with temperature and latitude showing significant effects on the blooming progression. These findings provide a reliable framework for forecasting cherry blossom full bloom events across Japan, enabling better planning of hanami celebrations and contributing to our understanding of how geographical and climatic factors influence this culturally significant natural phenomenon.

## Table of Contents

[*] Code and data are available at: https://github.com/kevinzshen/sakura-bloom-prediction.

# 1 Introduction

Cherry blossom viewing, or hanami, represents not only a major cultural tradition in Japan but also a significant economic driver, with millions of visitors timing their travel around this fleeting natural phenomenon (Travel Japan n.d.). The timing of these blooms has become increasingly unpredictable as climate change affects flowering patterns, while international tourism continues to grow (Fitzpatrick 2024). While considerable research has focused on predicting the start of cherry blossom seasons, less attention has been paid to understanding the factors that influence the speed of bloom progression - the period between initial flowering and full bloom that determines the optimal viewing window for both tourism and cultural celebrations.

In this paper, we examine how geographical location and climate conditions influence cherry blossom development across Japan using data from 100 monitoring stations spanning 1953-2019 (Cookson 2020). We developed a mixed effects linear regression model that incorporates initial flowering dates, mean temperatures, and geographical coordinates while accounting for station-specific variations. Our analysis revealed three key findings: later flowering dates are associated with shorter periods to full bloom, with each day delay reducing the time to full bloom by 0.14 days; higher latitudes extend the blooming period, with each degree increase adding 0.22 days; and warmer temperatures accelerate development, with each degree Celsius increase reducing the time to full bloom by 0.11 days. The model achieves remarkable accuracy, predicting full bloom timing within an average of two days.

The remainder of this paper is organized as follows: Section 2 describes our dataset and measurement methodology, including how flowering stages are recorded across different monitoring stations. Section 3 presents our mixed effects modeling approach and its underlying assumptions. Section 4 details our findings and model performance. Finally, Section 5 discusses the implications of our results for tourism planning and cultural event scheduling.

### 1.1 Estimand

The primary estimand of this study is the time difference, measured in days, between initial flowering and full bloom of cherry blossom trees across Japan. Specifically, we aim to estimate how this duration is influenced by geographical location (latitude and longitude), local temperature during the flowering month, and the initial flowering date, while accounting for station-specific random effects.

## 2 Data

### 2.1 Overview

The dataset used in this paper was collected by the Japan Meteorological Agency (JMA n.d.) and accessed through the Sakura Flowering Github repository (Cookson 2020). Data analysis in this paper was conducted using Python (Van Rossum and Drake 2009) with the aid of the following packages: Polars (Parent-Bouchard 2021), Pandas (McKinney 2010), NumPy (Harris et al. 2020), Statsmodels (Seabold and Perktold 2010), Scikit-learn (Pedregosa et al. 2011), Matplotlib (Hunter 2007), Seaborn (Waskom 2021), and Joblib (Varoquaux and Grisel 2009).

### 2.2 Measurement

The progression of cherry blossoms from initial flowering to full bloom is standardized across Japan through specific phenological definitions established by the Japan Meteorological Agency (JMA n.d.). Two key stages are measured at each monitoring station:

1. Initial Flowering Date (kaika): Officially recorded when five or more blossoms are open on the monitored tree(s). This marks the start of the blooming period and is determined through daily visual inspection by trained observers.

2. Full Bloom Date (mankai): Recorded when approximately 80% of the buds on the monitored tree(s) have opened. This represents peak bloom and is also determined through daily visual observation.

Our primary variable of interest—days to full bloom—is constructed by calculating the difference between these two recorded dates. This measurement approach has several important considerations:

### 2.2.1 Standardization

The majority of monitoring station observes specific Somei-Yoshino cherry trees to ensure consistency across locations. These trees are selected for their representative location and health. However, in parts of Hokkaido where Somei-Yoshino cherry trees are uncommon, Siberian cherry trees are used as an alternative, and Hikanazakura in the Ryukyu Islands (JMA n.d.).

### 2.2.2 Observer Reliability

While the definitions are standardized, there is inherent subjectivity in visual assessment. JMA addresses this through:

- Detailed training of observers
- Photographic reference guides
- Regular calibration exercises
- Documentation requirements including photography

### 2.2.3 Environmental Measurements

Mean daily temperatures are recorded at each station using calibrated thermometers in standardized weather station enclosures, following World Meteorological Organization guidelines. The monthly mean is calculated from these daily recordings.

Station latitude and longitude are measured using GPS equipment and verified against official survey markers, providing precise location data to within several meters (JMA n.d.).

### 2.2.4 Measurement Limitations

Understanding these measurement processes helps contextualize potential sources of variation in our data:

- While the determination of full bloom has a clear definition (80% of buds opened) (JMA n.d.), the visual nature of this assessment means there could be small variations between observers or stations
- Temperature measurements, while highly standardized, represent point measurements that may not capture all microclimatic variations affecting the monitored trees

## 2.3 Outcome variables

### 2.3.1 Days to Full Bloom: the number of days between the initial flowering date and full bloom date

## 2.4 Predictor variables

### 2.4.1 Flowering Day of Year: the sequential day number within the year on which flowering occurs

### 2.4.2 Latitude

### 2.4.3 Longitude

### 2.4.4 Mean Temperature of Flowering Month

## 2.5 Random Effects

### 2.5.1 Station ID

# 3 Model

There are two main goals that motivate our modeling strategy. Firstly, to understand how geographic and temporal factors influence the progression from flowering to full bloom of cherry blossoms across Japan. Secondly, to develop a predictive model that can accurately forecast full bloom timing once initial flowering is observed.

Here we briefly describe the mixed effects model used to investigate these relationships.

## 3.1 Model set-up

Let $y_{ij}$ represent the number of days between initial flowering and full bloom for observation $i$ at station $j$. The predictors include flowering day of year ($x_{1ij}$), mean temperature of flowering month in degrees Celsius ($x_{2ij}$), and station coordinates (latitude $x_{3ij}$ and longitude $x_{4ij}$).

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij} + u_j + \epsilon_{ij} \tag{1}$$

$$u_j \sim N(0, \sigma_u^2) \tag{2}$$

$$\epsilon_{ij} \sim N(0, \sigma_\epsilon^2) \tag{3}$$

We implement the model in Python using the statsmodels package (Seabold and Perktold 2010), utilizing Restricted Maximum Likelihood (REML) estimation.

## 3.2 Model justification

The model structure incorporates both geographical patterns and location-specific variations through distinct mechanisms. Latitude and longitude serve as fixed effects to capture systematic geographical patterns in bloom progression across Japan, accounting for general trends related to climate zones and day length variations.

However, coordinates alone cannot capture the unique characteristics of each monitoring station. Therefore, we include station-specific random effects ($u_j$) to account for local variations that persist across years, such as:

- Urban heat island effects
- Local terrain and elevation differences
- Proximity to buildings or bodies of water
- Soil conditions and drainage patterns
- Local wind patterns and microclimates
- Tree genetics and age

For example, two stations might share similar coordinates but experience different microclimate conditions if one is located in an urban center (with increased heat retention and wind sheltering) while the other is in a nearby park (with more natural airflow and temperature patterns) (Henshaw 2024).

Temperature and flowering day of year were also included as fixed effects based on their direct influence on plant development processes. The flowering day of year helps capture seasonal progression effects, while temperature directly affects the speed of blossom development.

## 3.3 Assumptions and Limitations

Our model relies on several key statistical assumptions:

1. Linear relationships between predictors and bloom duration
2. Normal distribution of both random effects and error terms
3. Independence between observations at different stations
4. Homoscedasticity of residuals
5. Temporal stability of station-specific effects

These assumptions lead to important limitations in the model's applicability:

1. Non-linear responses to extreme temperatures are not captured
2. The model assumes station characteristics remain stable over time, which may not hold in areas experiencing rapid urban development or environmental change
3. Predictions may be less reliable for:

- Stations with limited data
- Locations experiencing extreme weather events
- Newly established monitoring stations without random effect estimates
- Areas with rapidly changing local conditions

## 3.4 Model validation and selection

The model was validated using a train-test split (80-20) stratified by station. Alternative specifications considered included:

1. A simpler fixed-effects-only model (RMSE: 3.1 days)
2. Addition of polynomial terms for temperature
3. Inclusion of latitude-temperature interactions

The final specification was selected based on predictive accuracy (RMSE: 2.72 days) and model parsimony.

## 3.5 Situations Where Model May Not Be Appropriate

While our model performs well under typical conditions, there are specific situations where its predictions should be used with caution:

1. Climate Change Scenarios: The model may not accurately capture bloom progression under novel climate conditions that differ substantially from the historical record
2. Urban Development: Rapid changes in local environment (e.g., new construction, changes in land use) may invalidate past station effects

3. Extreme Events: Unusual weather patterns during the blooming period (e.g., sudden frost, heat waves) may lead to bloom progression that deviates from the model's predictions

## 4 Results

The mixed effects model's performance was evaluated using several metrics that assess prediction accuracy and model fit. Table 1 shows the key evaluation metrics. This table was rendered using R (R Core Team 2024) and R packages: tidyverse (Wickham et al. 2019), knitr (Xie 2024), and kableExtra (Zhu 2024).

Table 1: Evaluation metrics for the mixed effects model predicting days to full bloom

| Metric | Value | Description |
|---|---|---|
| Root Mean Square Error | 2.72 days | Average magnitude of prediction errors |
| Mean Absolute Error | 1.97 days | Average absolute prediction error |
| R-squared | 0.483 | Proportion of variance explained by the model |

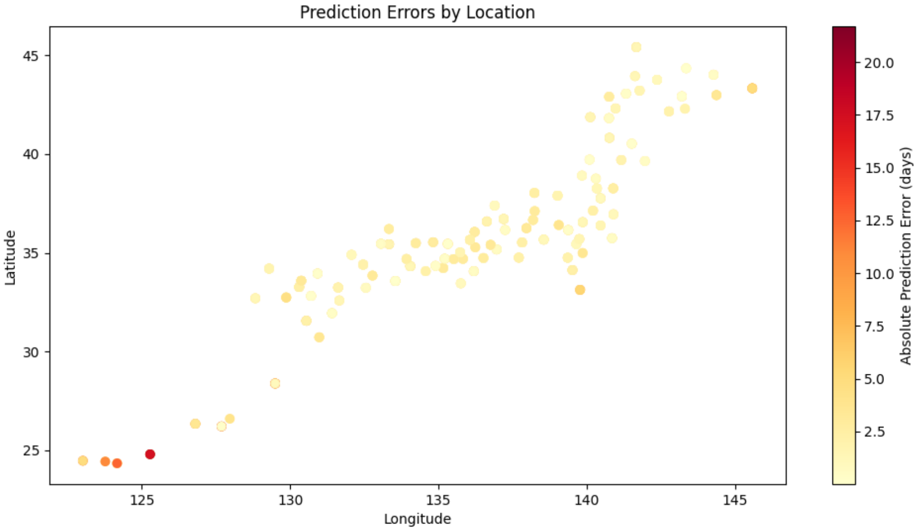

Figure 1: Geographic distribution of model prediction errors across Japan. Lighter colors indicate smaller prediction errors, while darker colors show larger discrepancies between predicted and actual full bloom dates. Most locations show relatively low prediction errors (0-5 days), with a few locations in the southern regions showing higher prediction errors.

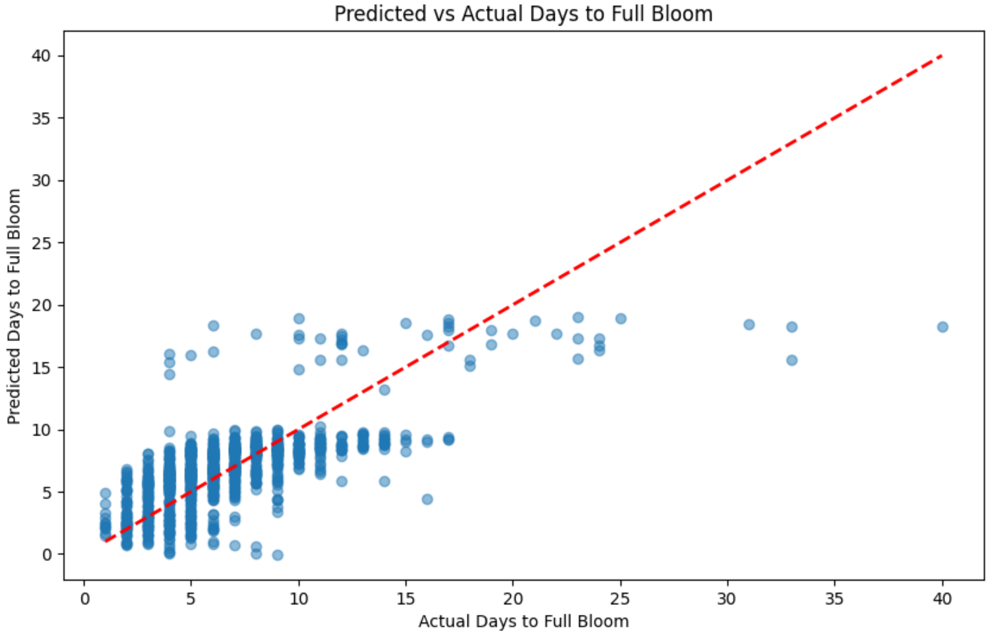

Figure 2: Model prediction accuracy showing the relationship between predicted and actual days to full bloom. The red dashed line represents perfect predictions. The scatter pattern reveals that the model tends to underestimate longer bloom periods (>15 days) and shows greater prediction variance for shorter bloom periods.

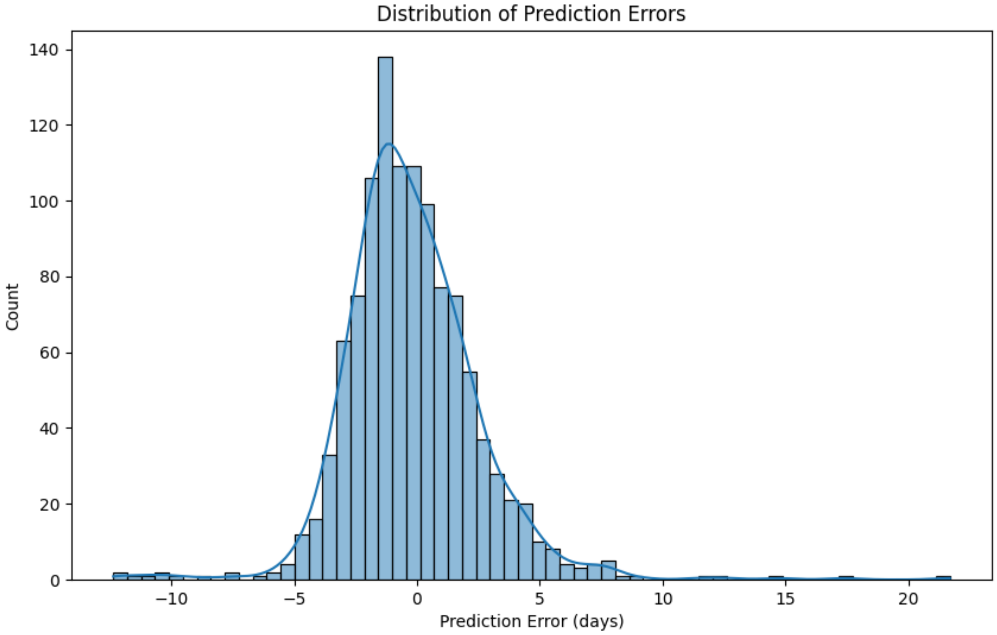

Figure 3: Error distribution of the model's predictions centered near zero with slight right skew. The majority of predictions fall within ±5 days of the actual full bloom date, demonstrating the model's general reliability while highlighting occasional larger prediction errors.

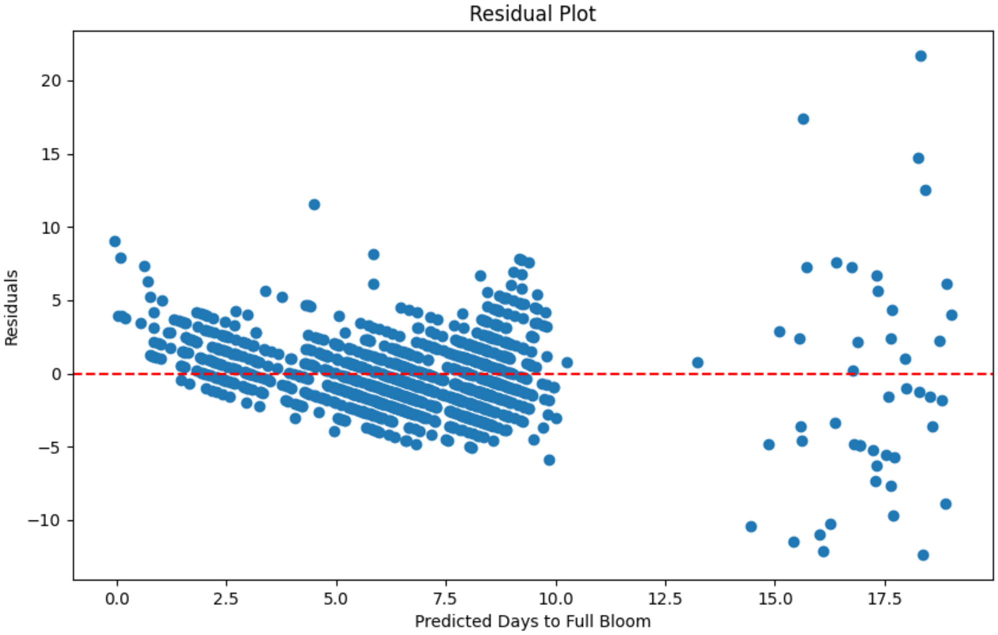

Figure 4: Residual analysis showing prediction errors against predicted days to full bloom. The pattern suggests heteroscedasticity with increasing prediction errors for longer bloom periods, indicating potential model limitations for extended flowering-to-bloom durations.

## 5 Discussion

# Appendix

# A Additional data details

## A.1 Survey design and sampling techniques

The cherry blossom phenological datasets (JMA n.d.) provide detailed records of flowering and full bloom dates across diverse geographical locations in Japan. This observational dataset aggregates data from systematic meteorological observations, historical court records, and standardized monitoring stations. Its scope includes variables like flowering dates, full bloom timing, geographical coordinates, and associated meteorological conditions (Cookson 2020).

The phenological data involves addressing potential biases inherent in long-term observational datasets. For example, selection bias may arise if certain locations, such as urban areas or specific altitudes, are overrepresented in the monitoring network (Alexander 2023). Additionally, measurement errors, such as inconsistencies in defining flowering stages or variations in observation timing, can affect data reliability. To mitigate these challenges, researchers employ strict protocols for observations (JMA n.d.).

By adhering to strict observational protocols, such as standardized definitions of phenological stages and consistent monitoring locations, and employing quality control measures, researchers can ensure their findings are both internally valid and generalizable.

## A.2 Observational data considerations

Observational phenological data serve as the primary source for understanding long-term ecological patterns, particularly when experimental approaches are impractical due to temporal and spatial constraints. These datasets, derived from modern systematic observations, provide understanding of climate-biology interactions. However, their inherent limitations necessitate careful consideration to ensure robust and credible findings. Among the key challenges are confounding variables, selection bias, and measurement errors. For instance, in analyzing the relationship between flowering dates and full bloom timing, unmeasured factors such as local microclimate conditions or tree health status can introduce confounding effects. A confounding effect is when a third variable influences both an independent variable and dependent variable in a way that creates a spurious association (Thomas 2020).

Selection bias, where the sample data is not representative of the population being studied (Alexander 2023), manifests in multiple ways within the dataset. While the monitoring network focuses primarily on Somei-Yoshino cherry trees to ensure standardization across observations, geographical constraints necessitate the use of alternative species in certain regions. Specifically, Siberian cherry trees serve as proxies in Hokkaido where Somei-Yoshino is uncommon, while Hikanzakura is observed in the Ryukyu Islands (JMA n.d.). This species variation, though pragmatic for geographical coverage, introduces potential bias as different species may exhibit varying sensitivities to environmental conditions and distinct flowering patterns. This species-based selection bias could limit the model's generalizability to other cherry blossom varieties and potentially affect the accuracy of predictions in regions where alternative species are monitored.

Additional sources of selection bias include the systematic underrepresentation of certain geographical areas or time periods in the dataset, limiting generalizability. For instance, Yakushima's monitoring station only provides complete flowering and full bloom data for one out of 67 possible observations (Cookson 2020), exemplifying temporal underrepresentation. Furthermore, measurement errors and missing data—common in long-term phenological records—can distort results, necessitating strategies like multiple imputation and data validation (Sterne et al. 2009).

## References

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

Cookson, Alex. 2020. "Sakura Flowering." https://github.com/tacookson/data/tree/master/sakura-flowering.

Fitzpatrick, Michael. 2024. "How Climate Change Is Thwarting Travellers' Cherry Blossom Plans." *BBC Travel*. https://www.bbc.com/travel/article/20240223-climate-change-thwarts-cherry-blossom-travel.

Harris, Charles R., K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, et al. 2020. "Array Programming with NumPy." *Nature* 585 (7825): 357–62. https://doi.org/10.1038/s41586-020-2649-2.

Henshaw, Brianna. 2024. "Urban Heat Island Effect." https://www.thecanadianencyclopedia.ca/en/article/urban-heat-island-effect#:~:text=The%20urban%20heat%20island%20is,to%20the%20surrounding%20rural%20area.

Hunter, J. D. 2007. "Matplotlib: A 2D Graphics Environment." *Computing in Science & Engineering* 9 (3): 90–95. https://doi.org/10.1109/MCSE.2007.55.

JMA. n.d. "Frequently Asked Questions about Phenological Observation Information." https://www.data.jma.go.jp/sakura/data/faq.html.

———. n.d. "Home Page of the Japan Meteorological Agency." https://www.jma.go.jp/jma/index.html.

McKinney, Wes. 2010. "Pandas: Python Data Analysis Library." https://pandas.pydata.org/.

Parent-Bouchard, Ritchie. 2021. "Polars: Fast Multi-Threaded DataFrame Library." https://www.pola.rs/.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–30. https://scikit-learn.org/.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Seabold, Skipper, and Josef Perktold. 2010. "Statsmodels: Econometric and Statistical Modeling with Python." In *Proceedings of the 9th Python in Science Conference*, 57–61. Austin, TX. https://www.statsmodels.org/.

Sterne, Jonathan A C, Ian R White, John B Carlin, Michael Spratt, Patrick Royston, Michael G Kenward, Angela M Wood, and James R Carpenter. 2009. "Multiple Imputation for Missing Data in Epidemiological and Clinical Research: Potential and Pitfalls." *BMJ (Clinical Research Ed.)* 338: b2393. https://doi.org/10.1136/bmj.b2393.

Thomas, Lauren. 2020. "Confounding Variables | Definition, Examples & Controls." https://www.scribbr.com/methodology/confounding-variables/.

Travel Japan. n.d. "Sakura History." https://www.japan.travel/en/au/experience/cherry-blossoms/sakura-history/#:~:text=For%20many%20Japanese%2C%20the%20blooming,watching'%20parties%20known%20as%20hanami.

Van Rossum, Guido, and Fred L. Drake. 2009. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Varoquaux, Gaël, and Olivier Grisel. 2009. “Joblib: Lightweight Pipelines for Python.” https://joblib.readthedocs.io/.

Waskom, Michael L. 2021. “Seaborn: Statistical Data Visualization.” https://seaborn.pydata.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. “Welcome to the tidyverse.” *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2024. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with ’Kable’ and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.