Assessing the Explanatory and Predictive Power of Google’s Community Mobility Report for COVID-19 Forecasting on a County-Level

Vincent Nguyen

**Authors**

* Vincent Nguyen

**Author affiliations**

1. College of Public Health, University of Georgia, Athens, GA, USA.

Corresponding author: vln27447@uga.edu

# 1. Summary/Abstract

**Background**: Human mobility patterns shifted significantly throughout the COVID-19 pandemic due to public health policies and individual risk perception. These changes, recorded in Google’s Community Mobility Report, may hold value in understanding transmission and forecasting cases.

**Objective:** This analysis evaluated the additive explanatory and predictive performance mobility dynamics may have on modeling daily COVID-19 case incidence at the county level in Georgia.

**Methods:** Daily case data from John Hopkins University and mobility predictors from Google’s Community Mobility Report (CMR) were used to develop machine learning models with the objective of explaining case incidence across 44 Georgia counties between March 14th 2020 and March 14th 2022. Three modeling approaches, LASSO regression, Random Forest, and XGBoost, were applied to investigate the association. Models were trained on three feature sets, a baseline set (lagged case counts and population density), a set including the mobility predictors, and a set including 7-day lagged versions of the mobility predictors. Models were evaluated using RMSE, R², and MAE metrics on a 30 day test set, with rolling window cross-validation used during training.

**Results:** In LASSO and Random Forest models, the inclusion of mobility and lagged mobility predictors resulted in significant reductions of prediction error metrics (MAE and RMSE). LASSO with mobility predictors achieved the highest R² (0.843) and identified mobility trends as key drivers of case incidence. Explanatory power, R², did not improve significantly in Random Forest and XGBoost, suggesting that mobility alone does not fully explain variation in daily new case incidence.

**Conclusion:** This analysis suggests that the incorporation of time-lagged mobility data can improve the predictive performance of some machine learning models for COVID-19 case incidence, particularly in reducing prediction error metrics like MAE or RMSE. While the improvement in explanatory power was limited, the results highlight the potential in utilizing behavioral data in forecasting disease transmission dynamics and the integration of mobility metrics in public health surveillance.

# 2. Introduction

## 2.1 COVID-19 Pandemic

The emergence of the novel corona-virus SARS-CoV-2 brought about the onset of the one of the most significant global health crises in modern history. First identified in late 2019, the virus rapidly spread across the world due to its high transmissibility and global interconnectedness, leading the World Health Organization to declare it a pandemic on March 11, 2020. Two days later, on March 13th, 2020, the United States declared the COVID-19 pandemic a national emergency, marking the biggest virus outbreak since the 1916 influenza pandemic. In response, public health agencies implemented measures to curb the virus’ spread including travel restrictions, social distancing, and lock-down procedures (Centers for Disease Control and Prevention, 2024).

## 2.2 Mobility Dynamics

Human mobility, a key driver of respiratory disease transmission, shifted significantly during the pandemic in response to public health policies and disease risk perception (Paltra et al., 2024). Uniquely, the COVID-19 pandemic utilized a new form of physical distancing measures which was known as stay-at-home orders and colloquially, lock-downs. Lock-downs involved stringent stay-at-home orders, closure of non-essential businesses, and restrictions on public gatherings. Beyond these measures which lasted only a few weeks in Georgia, individual risk perception played into the compliance of other preventative behaviors (masking, social distancing, etc.). The variability of risk individual perception has led to complex mobility patterns during the pandemic, for example, surges in mobility amidst large case outbreaks (Cipolletta et al., 2022). Previous analysis has suggested that mobility patterns are correlated with decreases in COVID-19 case growth rates (Badr et al., 2020). In order to further explore and quantify this relationship, machine learning models were adapted.

## 2.3 Machine Learning

Machine learning refers to a class of data-driven algorithms that aim to analyze associations and relationships found in data. These techniques can be supervised or unsupervised; in supervised ML, models are given labeled inputs/features to derive and predict an output. These models aim to also approximate the relationship between outputs and inputs, quantify predictions, or approximate classification tasks. Beyond regular statistical analysis, machine learning methods offer powerful tools for forecasting infectious disease trends by identifying complex, nonlinear relationships between predictors and outcomes (Rashidi et al., 2023). Some common models in disease forecasting include ARMA, ARIMA, LASSO, XGBoost, and various neural network techniques (Alfred & Obit, 2021). Several authors have utilized case counts, estimations, demographics, and more to forecast disease trends (Ogunjo et al., 2022).

## 2.4 Previous Literature

The Google CMR have been widely utilized in machine learning research. As noted by Irini et al. (2021), Google’s CMR have been integrated into diverse machine learning algorithms to predict a range of outcomes, such as case incidence, outcomes, drug development, and more. Studies have demonstrated associations between mobility (as measured by Google) and case incidence in areas such as Jakarta, as well as COVID-19 death in Ireland [Irini et al. (2021)]. Paez (2020) also employed multivariate regression to investigate the report’s potential in COVID-19 incidence. Overall, the CMR data has been validated across several international contexts and in various regression tasks. However, to our knowledge, this data has yet to be applied in a county-level American context and while utilizing modeling methods like Random Forest and XGBoost.

In addition to the predictors found in the CMR, lagged mobility predictors will also be created, given their previously demonstrated relevance (Sulyok & Walker, 2020).

# 3. Data

## 3.1 Google’s Community Mobility Report

During the pandemic, Google began to collect aggregated, anonymized data from users utilizing Google products (apps, phones, etc.) to track changes in mobility (Google, 2020a). Within the data, Google measures mobility as a percent change difference from baseline measurements; for example, a -45% change in retail and recreation mobility indicates a 45% reduction in movement to those categorical locations. These measurements are stratified by county, however, with the implication of technology use, are limited to counties with enough Google users. Additionally, two mobility metrics, transit stations and recreational parks, were omitted in this analysis due to incompleteness. According to Google, recreational parks are specifically intended to mean official national parks (Google, 2020b). Georgia, specifically, only has 11 national parks (National Park Service, 2017) .

## 3.2 John Hopkins University COVID-19 Case Data

The COVID-19 Data Repository by the Center for Systems Science and Engineering at Johns Hopkins University is a comprehensive data set that tracks global COVID-19 cases, recoveries, and deaths. This data was recorded daily for several years across every county in the US (CSSEGISandData, 2022).

## 3.3 Final Data-set

After data wrangling and cleaning for completeness, 44 counties were included in the analysis. An 80% completeness of data per county was required for inclusion into analysis, resulting in the removal of 115 counties. Many of Georgia’s counties are rural and as such, lacked many mobility metrics. The data was filtered from March 14th 2020 to March 14th 2022. This was selected specifically as March 14th marks the date of Governor Brian Kemp’s announcement of Georgia’s Public Health State of Emergency (Hart, 2025). Imputation strategies were not considered as data was not missing at random; missing data tended to come from specific counties rather than being randomly distributed across the data set. Imputing these values could obscure important patterns or misrepresent the complexitiy of the pandemic. Exclusion of these counties was chosen instead.

# 4. Methods

## 4.1 Model Selection

As informed by Alfred & Obit (2021), machine learning models have wide application in forecasting outbreaks and disease incidence. In their review, they outline several common applications of regression and classification models. Their review, along with class content and data structure, informed model selection. Previous iterations of the analysis indicated that the data was non-normal and non-stationary which violated assumptions in common time series models like ARIMA. LASSO regression was chosen for its previously demonstrated relevance (Nanda et al., 2022) Random Forest and XGBoost were chosen for their lack of assumptions required for analysis.

**LASSO Regression**

Also known as Least Absolute Shrinkage and Selection Operator, Lasso is a regularization technique used in regression modeling to prevent overfitting and improve model interpretability. Lasso adds a penalty term equal to the absolute value of the magnitude of coefficients to the loss function (IBM, 2024). This penalty can shrink coefficients to zero, effectively performing variable selection. Lasso was applied to linear regression models in the analysis.

**Random Forest**

A ensemble machine learning method that builds multiple decision trees and aggregates their predictions to improve accuracy and reduce overfitting. A decision tree is a modeling technique that partitions  data into subsets based on the input features; it aims to minimize variance of the outcome variable in each subgroup. These splits form a tree-like structure where each internal node represents a decision based on a feature, each branch corresponds to an outcome of the decision, and each leaf node represents a predicted value. Decision trees are powerful and can capture nonlinear complex relationships, however, can overfit training data.  Random forests are an average of these trees to produce a final prediction and can produce more stable/accurate models as a result (Breiman, 2001).

**XGBoost**

Also known as Extreme Gradient Boosting, XGBoost is a machine learning algorithm based on decision tress, however, instead of averaging trees, it tries to correct errors made by previous ones. This technique is known as boosting. In simple terms, XGBoost makes predictions with a single tree and looks at the errors and creates a second tree focused on minimizing the errors (Chen & Guestrin, 2016).

**Feature Set**

Three sets of predictors were chosen to assess the effect mobility changes have on case incidence. The first set, the baseline predictors, included population density and 3 spaced out lagged variables for case counts (1 day, 7 days, and 14 days). These were chosen because of  their known influence on disease transmission (Ogunjo et al., 2022). The second set, the mobility predictors, included population density, the 3 spaced out lagged case counts, and mobility indicators from Google’s report (e.g., workplace, grocery, retail, transit). Lastly, the third set included 7-day lagged versions of the mobility predictors, population density, and the three spaced out lagged case counts.

## 4.2 Cross Validation

To evaluate model performance while accounting for the temporal structure of the data, a rolling origin cross-validation technique with non-cumulative, time-blocked splits was utilized. The last 30 days of the data set were set aside and used for a test set for final model evaluation. Model training and cross-validation were conducted utilizing the remaining data set.

Cross-validation using a rolling window was implemented. Each fold contained a training window of four months (120 days per county) followed by a validation window of approximately one month (30 days). The 120 days of training and 30 days of validation is intended to mimic the 4:1 ratio of general machine learning splits (Sivakumar et al., 2024). The window was advanced forward in 15-day increments which creates multiple sequential train-validation splits. Non-cumulative windows were enforced to ensure that training sets did not grow over time. This design was intended to reflect a realistic forecasting scenario in which models are periodically re-trained using a fixed window of recent data to predict outcomes in a short future period. Each validation period occurs strictly after its corresponding training window to prevent information leakage.  
  
After tuning and validation, the final model was re-trained on the full training set and evaluated on the 30 day test set to assess its predictive performance.

## 4.3 Model Evaluation and Metrics

**R²**

A statistical measure that explains the proportion of the data variation that can be explained by the model. A higher R² value generally is indicative of a better-fitting model, although could be misleading when over fitting occurs. Criteria for final model selection focused on achieving the highest R² value. R² was chosen as the final selector for best model. Previous literature has used R² as a measure of forecasting performance (Nanda et al., 2022).

**RMSE**

Root Mean Square Deviation is a statistical measure that averages the magnitude of the errors in the model’s predictions.

Arithmetically, it is the square root of the average squared differences between the predicted and actual values. A lower RMSE indicates a better predictive performance.

**MAE**

Mean Absolute Error is a statistical measure that also measures prediction error, however, utilizes the absolute difference between predicted and actual values instead of the squaring the errors. It is calculated by taking the average absolute difference between each predicted and actual value. Because of its computation, MAE is less sensitive to large errors, like outliers, than RMSE. MAE may be a more reliable metric for this data set due to the presence of large outbreaks that can inflate the errors.

Linear regression, Random Forest, and XGBoost models were trained using the baseline set of features to predict COVID-19 case incidence. These baseline models served as reference points to evaluate the added predictive value of mobility-based features and to assess the potential association between changes in population mobility and COVID-19 case incidence.

After the models were made, their hyper-parameters were tuned to improve predictive performance.

## 4.4 Software

The analysis was conducted on RStudio 4.3.2 on Windows 11. The following R packages were used: ggplot2 (Wickham, 2019), broom (Robinson et al., 2022), here (Müller & Bryan, 2020), glmnet (Friedman et al., 2021), MASS (Ripley et al., 2022), tidymodels (Kuhn et al., 2023), dplyr (Wickham et al., 2020), rsample (Frick et al., 2025), parsnip (Kuhn et al., 2025), future (Bengtsson, 2025), vip (Greenwell & Boehmke, 2023), zoo (Zeileis et al., 2021), patchwork (Pedersen, 2020), e1071 (Meyer et al., 2022), scales (Wickham et al., 2022), RColorBrewer (Neuwirth, 2022), corrplot (Wei et al., 2017), reshape2 (Wickham, 2020), tidyr (Wickham & Henry, 2020), lubridate (Spinu et al., 2022), readxl (Wickham et al., 2019), stringr (Wickham & RStudio, 2019), skimr (Waring et al., 2022), gtExtras (Mock & Sjoberg, 2023), webshot2(Chang et al., 2023), tigris (Walker, 2025), and sf (Pebesma et al., 2021)

# 5. 5 Results

## 5.1 Exploratory/Descriptive analysis

To provide context for the modeling analysis, visualizations of Georgia’s COVID-19 case incidence and mobility dynamics are presented.

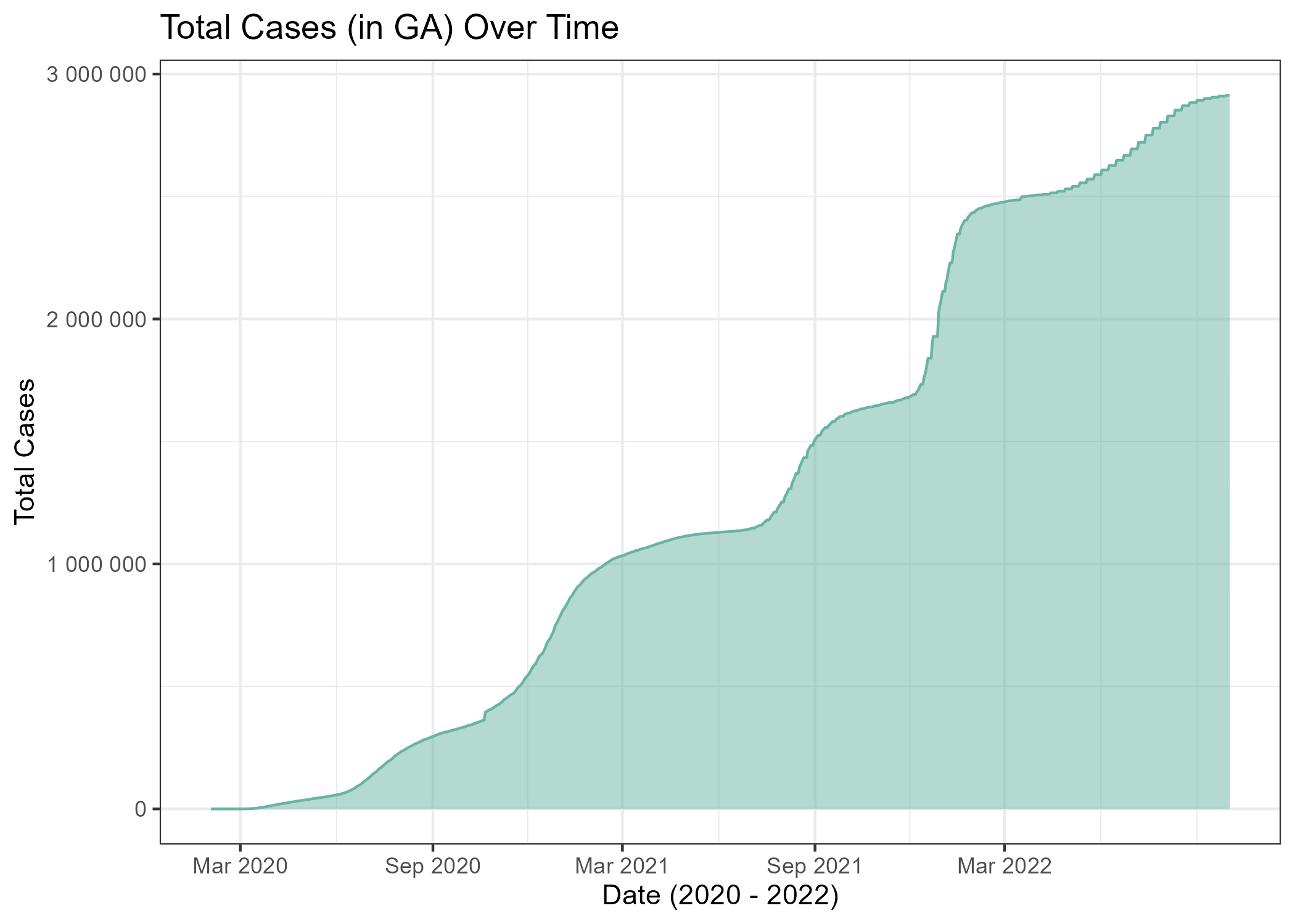


Figure 5.1: Total COVID-19 Cases in Georgia over time

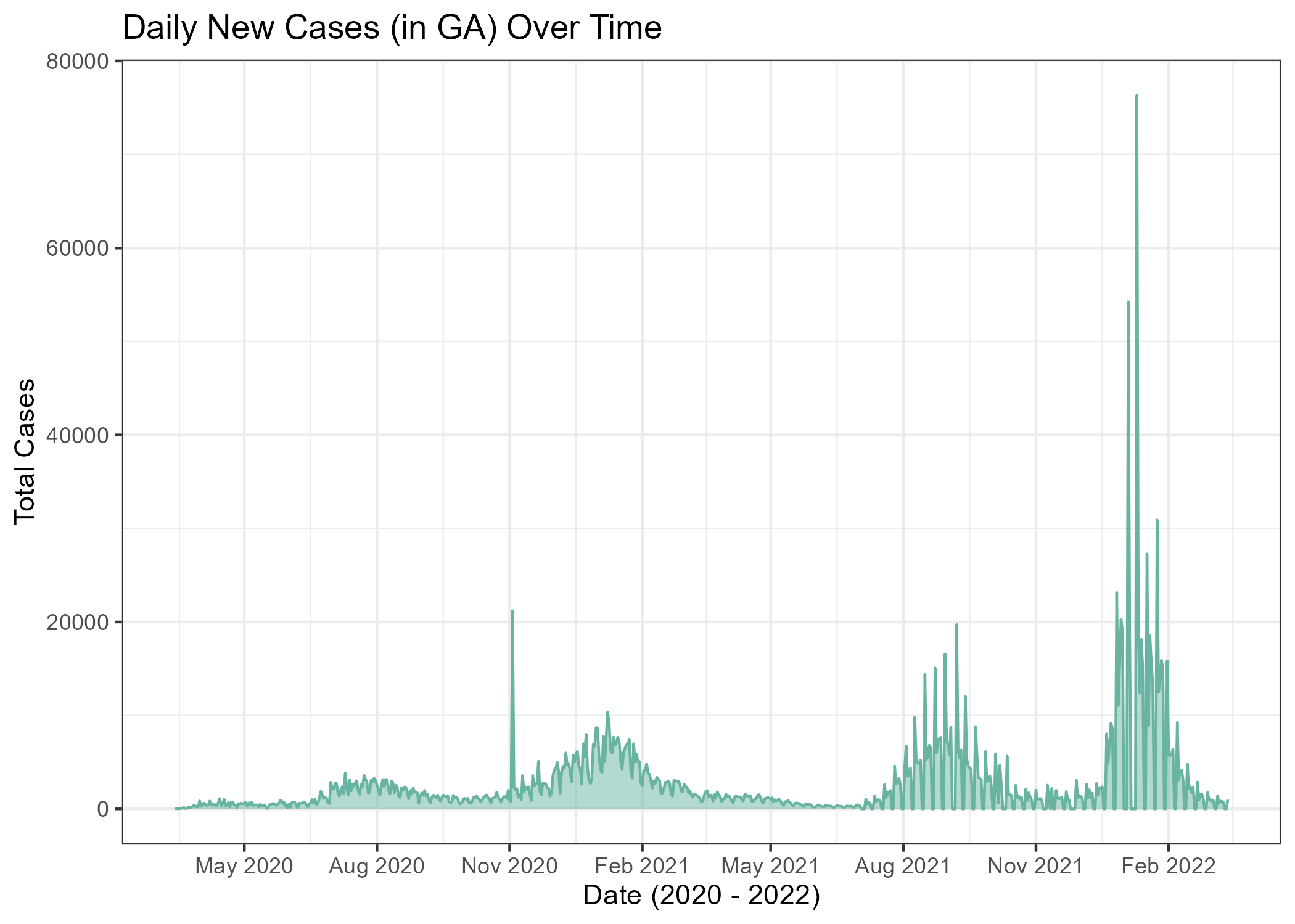


Figure 5.2: Daily New COVID-19 Cases overtime in the GA (2020-2022)

Figure 5.1 and 5.2 demonstrate distinct outbreaks of infection, with sharp peaks during the winter months and troughs in the summer, suggesting seasonal patterns and periods of increased transmission.

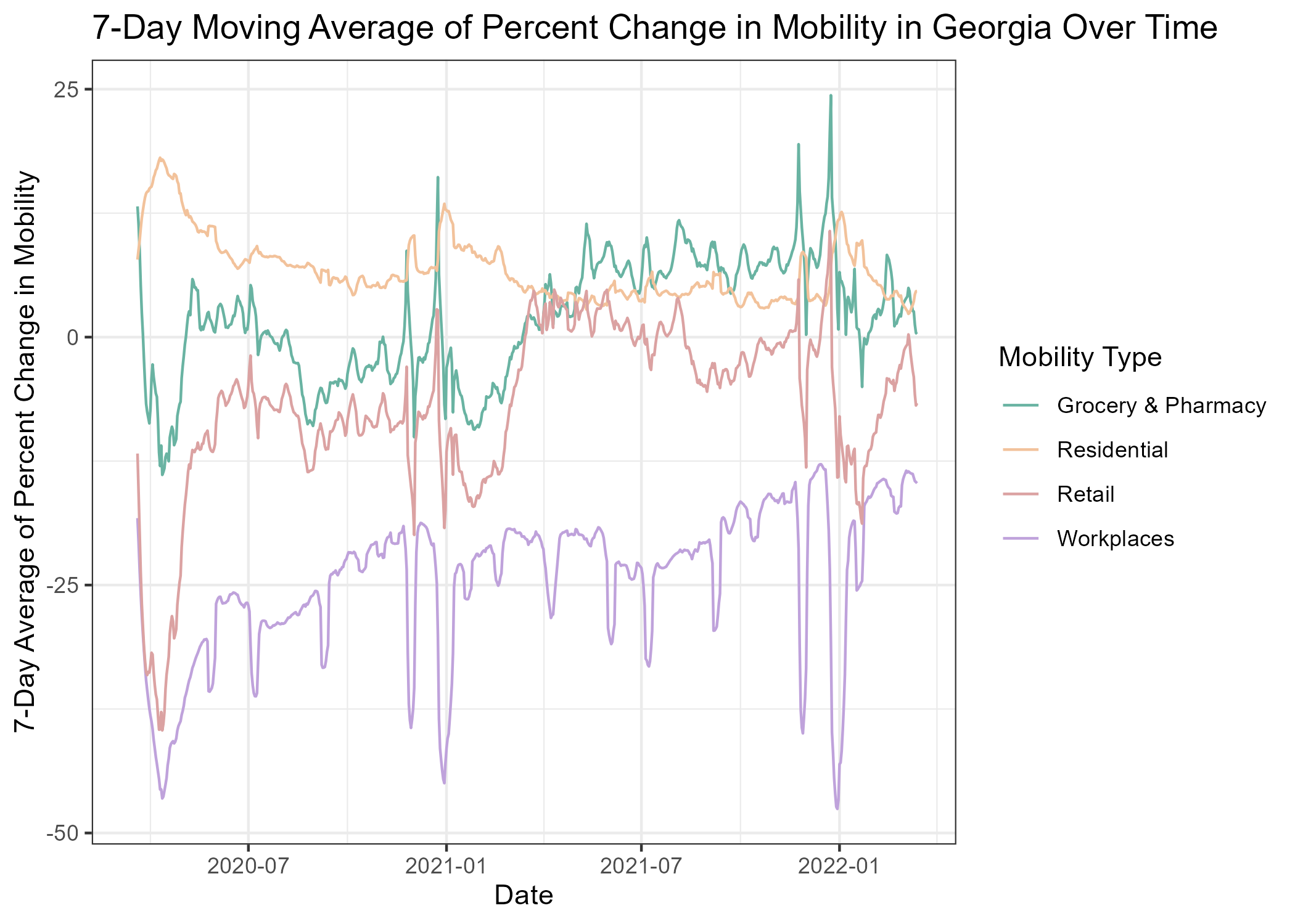


Figure 5.3: Percent Changes in Mobility per week in GA



Figure 5.4: Percent Changes in Mobility per week in GA

Figure 5.3 and 5.4 demonstrate distinct changes in mobility dynamics throughout the pandemic, with sharp drops corresponding with outbreaks and the seasons. Workplace mobility tended to remain below the baseline level while residential mobility increased during this period. There were frequent fluctuations in retail and grocery mobility, possibly a result of outbreaks, changing policies, or even holiday-related activity. Levels of mobility sharply declined and would remain below baseline until a year later. This seasonality violates the assumption of stationarity required for many time-series modeling techniques.

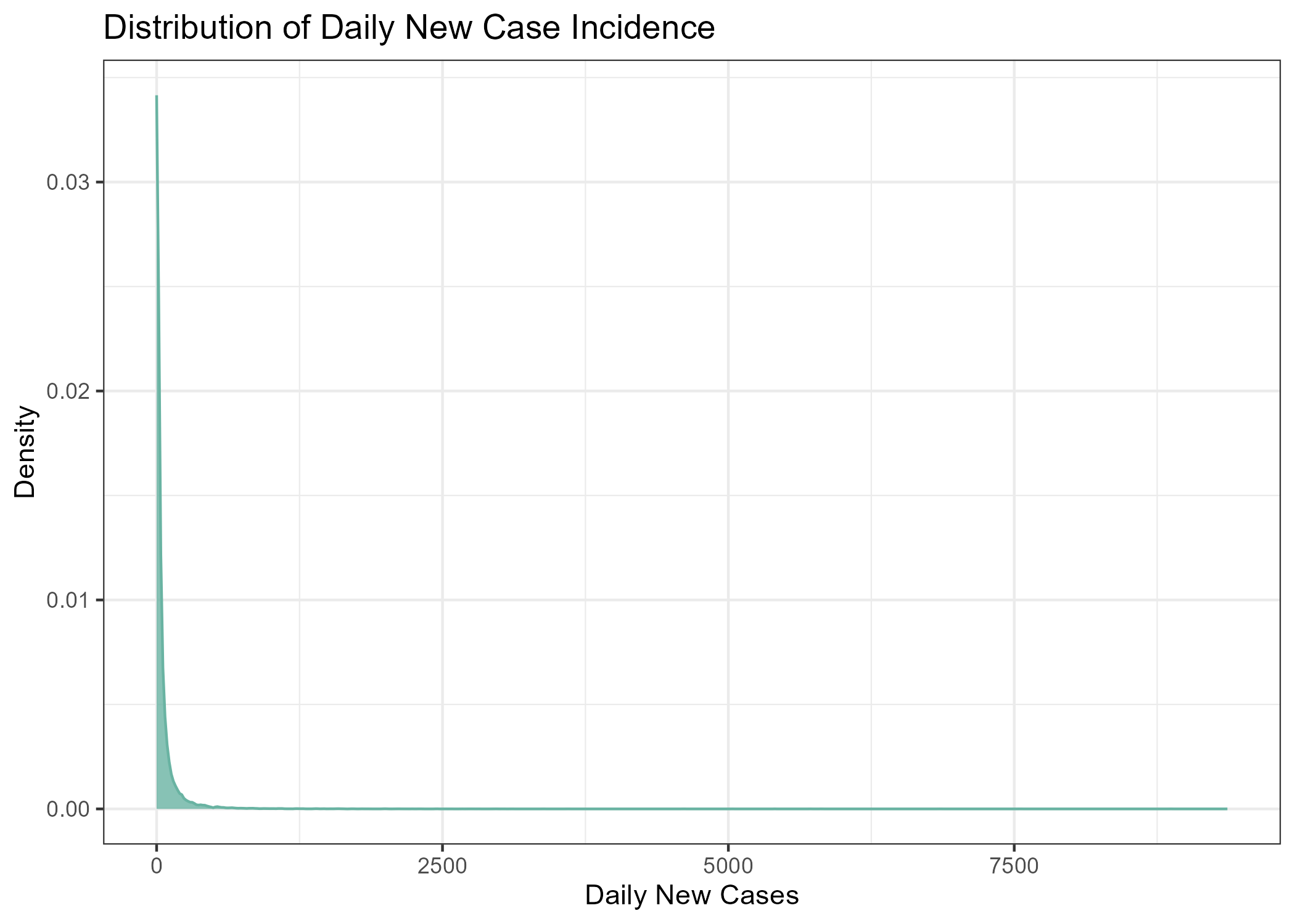


Figure 5.5: Density Plot of Daily New Cases

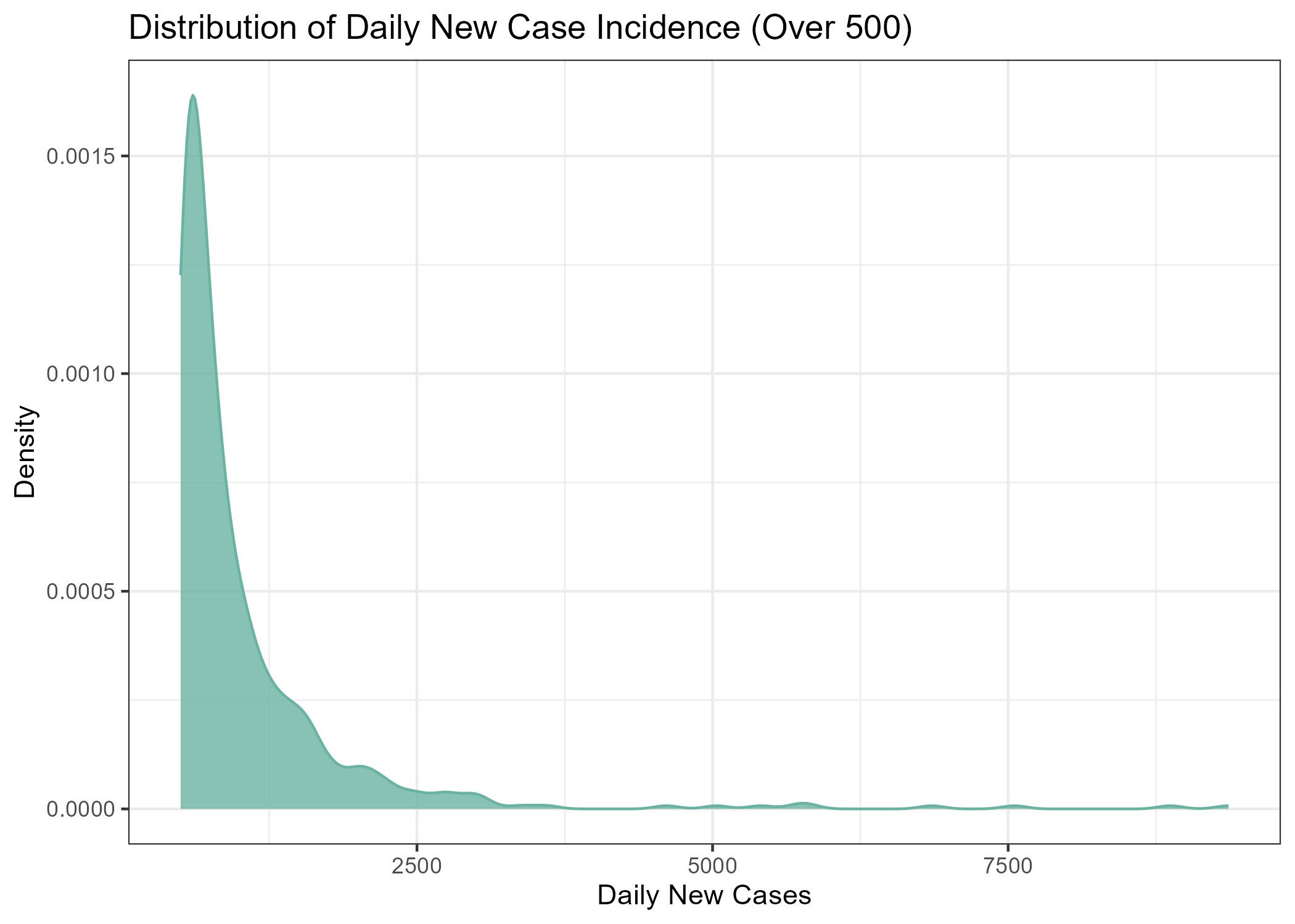


Figure 5.6: Density Plot of Daily New Cases (Over 500)

Figure 5.5 and 5.6 shows the distribution of daily new COVID-19 cases in Georgia. The distribution is heavily right skewed, with most days clustered around lower incidence leels and fewer days where extremely high case transmission occurs. This skew suggests surges were relatively infrequent compared to more moderate levels of daily incidence. This kurtosis can affect the performance of linear regression models and ARIMA where normality is an assumption whereas Random Forest and XGBoost are more robust to this skewing.

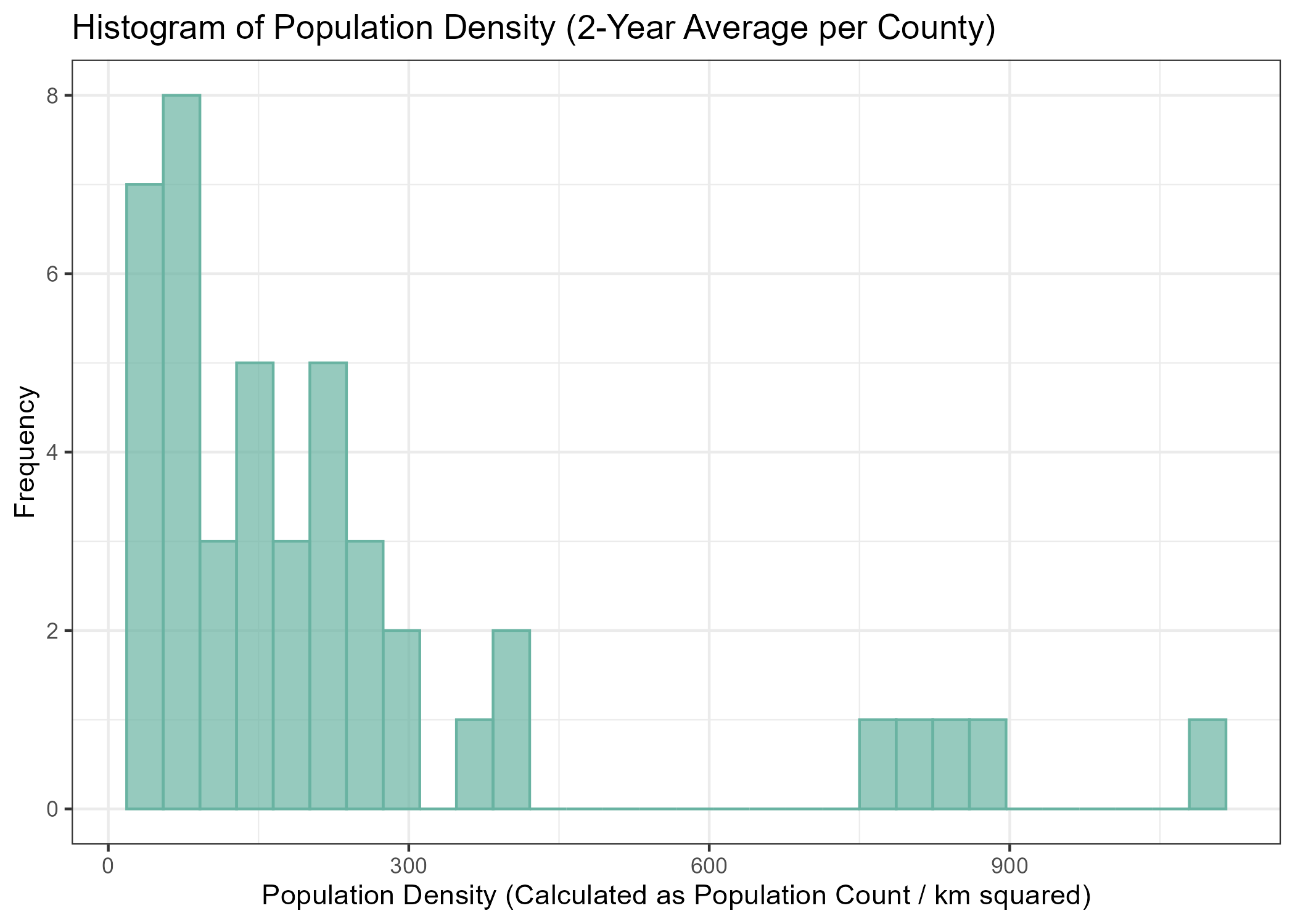


Figure 5.7: Distribution of Population Density among Counties included in Data

Population density across the included Georgia counties is also heavily right-skewed, with the majority of counties having lower densities. This imbalance may influence transmission dynamics and was included for modeling.

## 5.2 Correlation Analysis

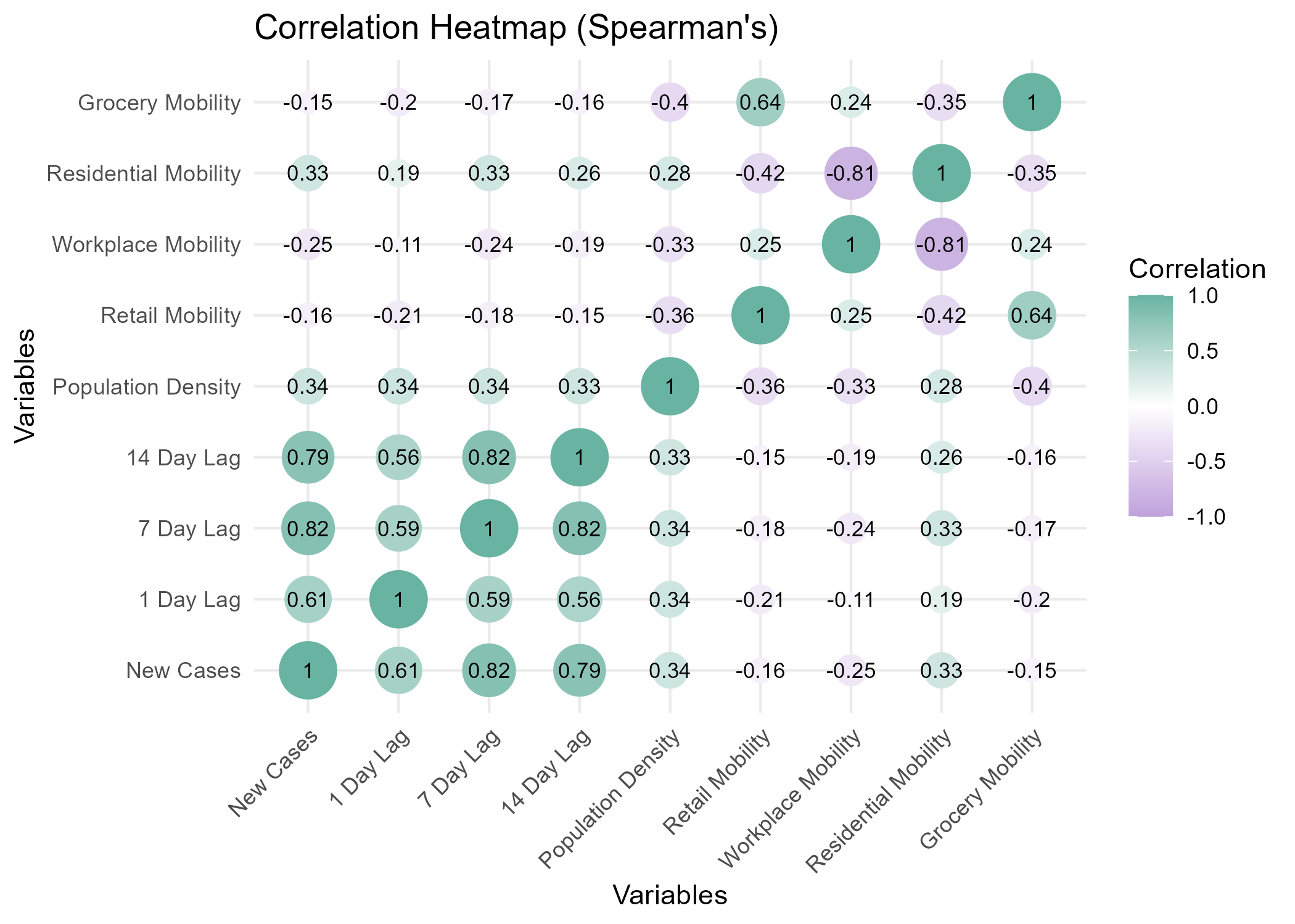


Figure 5.8: Correlation Matrix of Predictors using Spearman’s

In addition to visual analysis, correlation analysis was done. The non-parametric method, Spearman’s rank correlation coefficient, was used to assess correlation among predictors. Analysis reveals that lagged case counts were strongly correlated with daily incidence while in contrast mobility variables demonstrated weak relationships.

## 5.3 Model Analysis

To evaluate the explanatory and predictive value of mobility trends on new COVID-19 case incidence, three modeling techniques were utilized, LASSO, Random Forest, and XGBoost. For each model, three variants were created based on the different set of predictors: baseline predictors, mobility predictors, and lagged mobility predictors. This allowed for analysis of how mobility and its delated effects can influence COVID-19 case predictions.

All three models were tuned and utilized rolling origin cross-validation to respect the temporal structure of the data and prevent data leakage from the future. The best-performing hyperparameters were selected based on R². Predictions were made on the test data set and final model performance was evaluated using RMSE, R², and MAE.

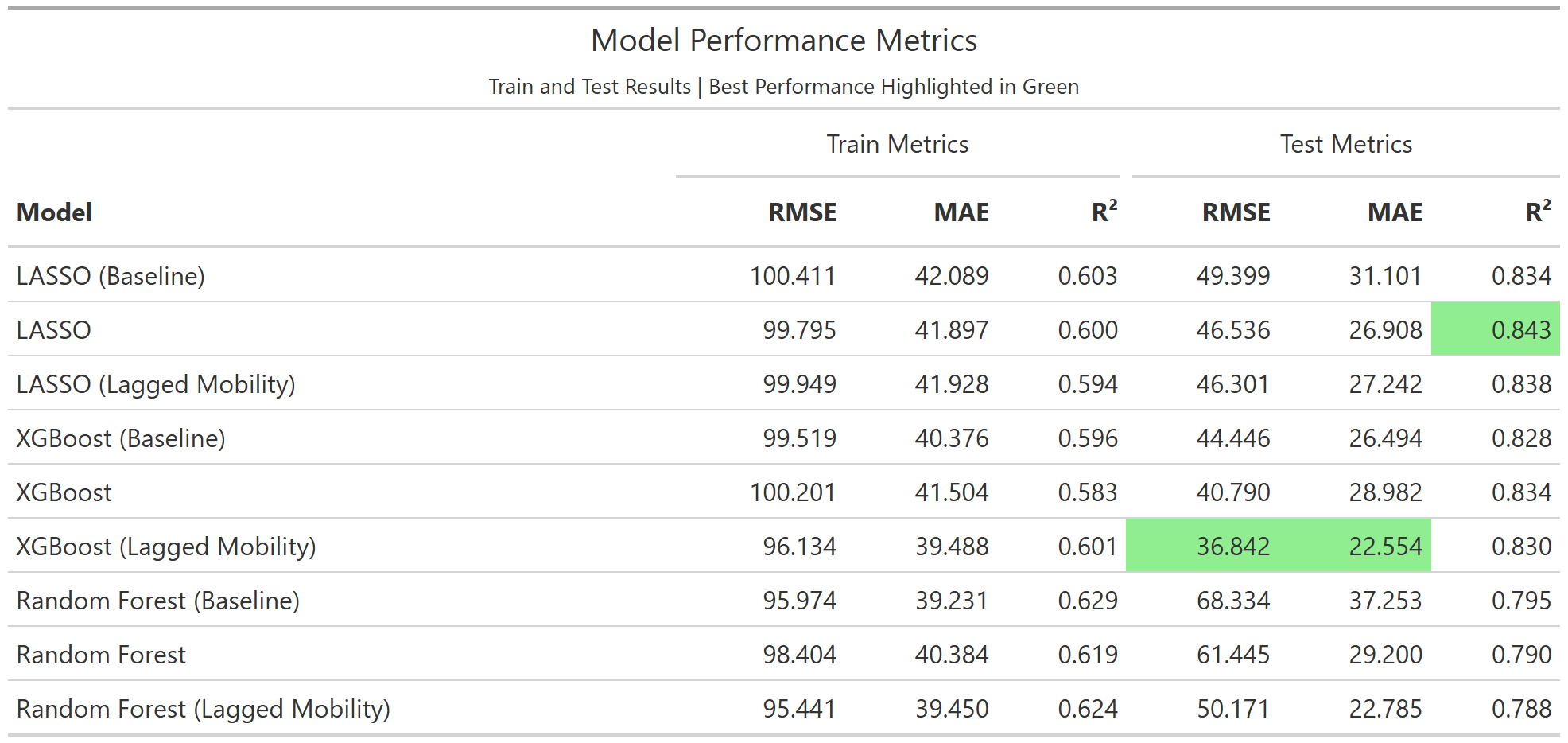


Table 5.1: Model Performance

The LASSO regression model with non-lagged mobility predictors had the strongest R² while XGBoost’s inclusion of lagged mobility predictors had the lowest MAE and RMSE.

Actual vs. predicted plots were created to visually assess model fit and diagnostic accuracy. These plots can help identify under- or over- prediction patterns.

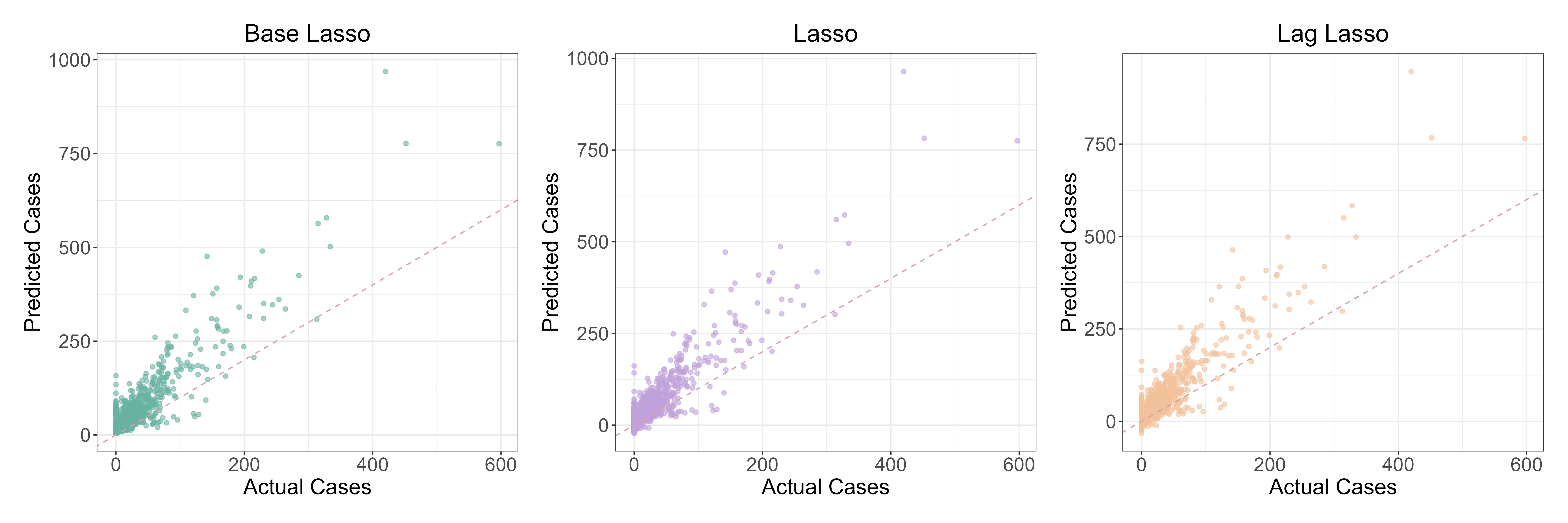


Figure 5.9: Plot of Actual vs Predicted Cases in Lasso Regression Models

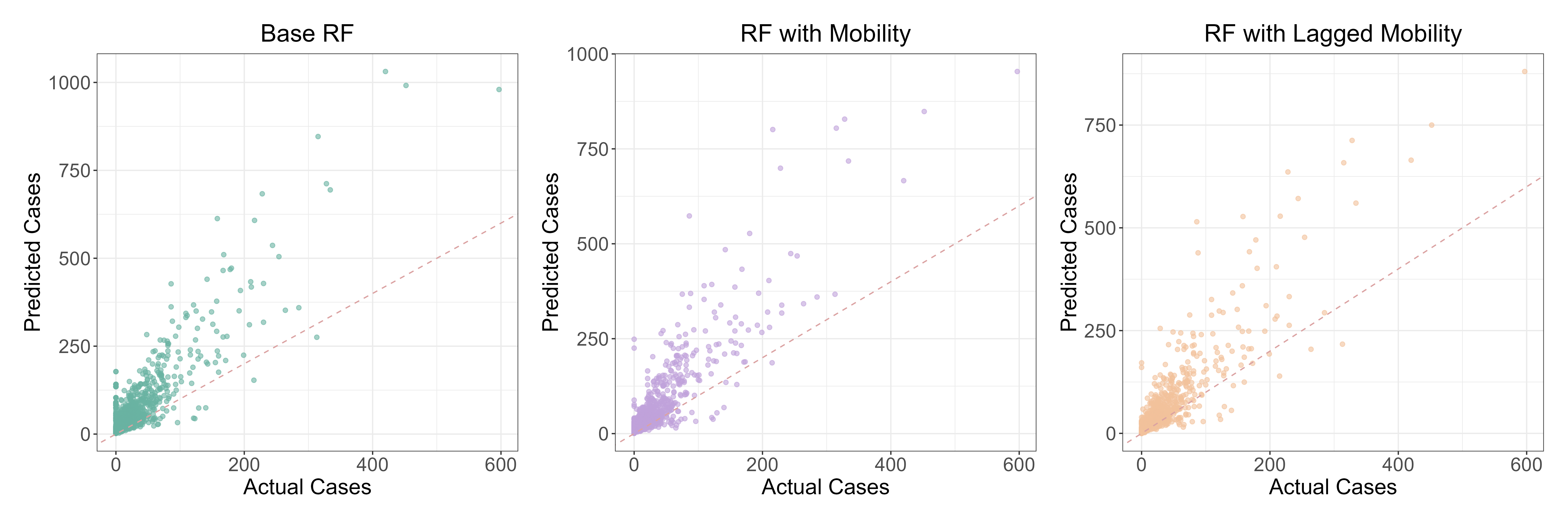


Figure 5.10: Plot of Actual vs Predicted Cases in Random Forest Models

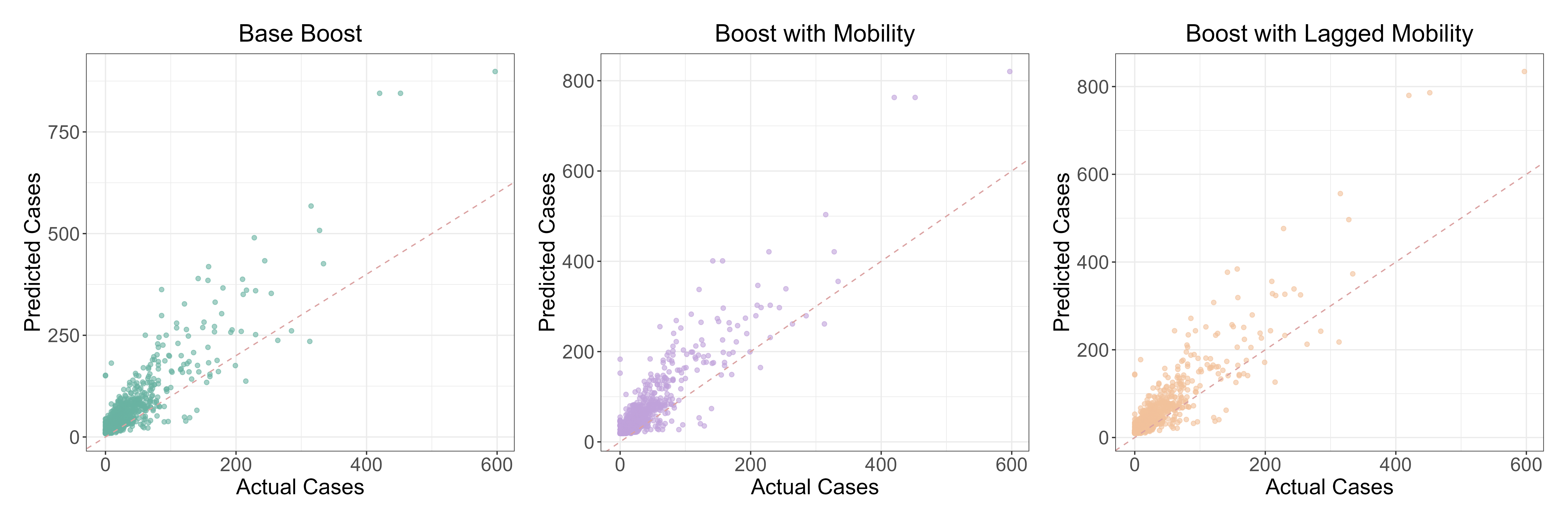


Figure 5.11: Plot of Actual vs Predicted Cases in XGBoost Models

In all three figures, the largest residuals are associated with over-predictions by the models. This is most likely due to the large prevalence of low case count days.

To assess the importance of the predictors, variable importance plots were generated for each model type. These plots can provide insight into how mobility predictors contributed to the predictive power of each model.

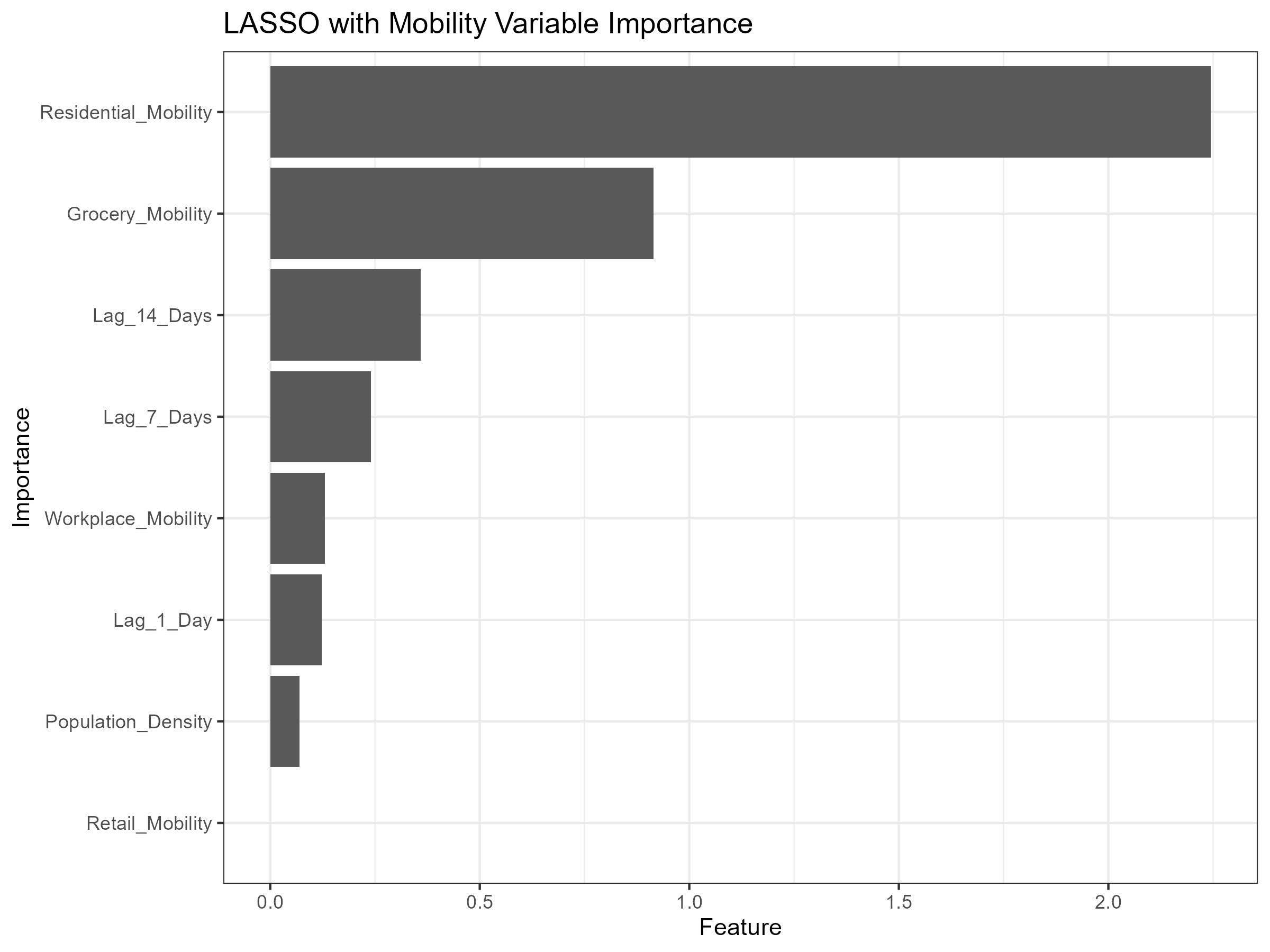


Figure 5.12: Variable Importance Plot of Mobility Included LASSO Model

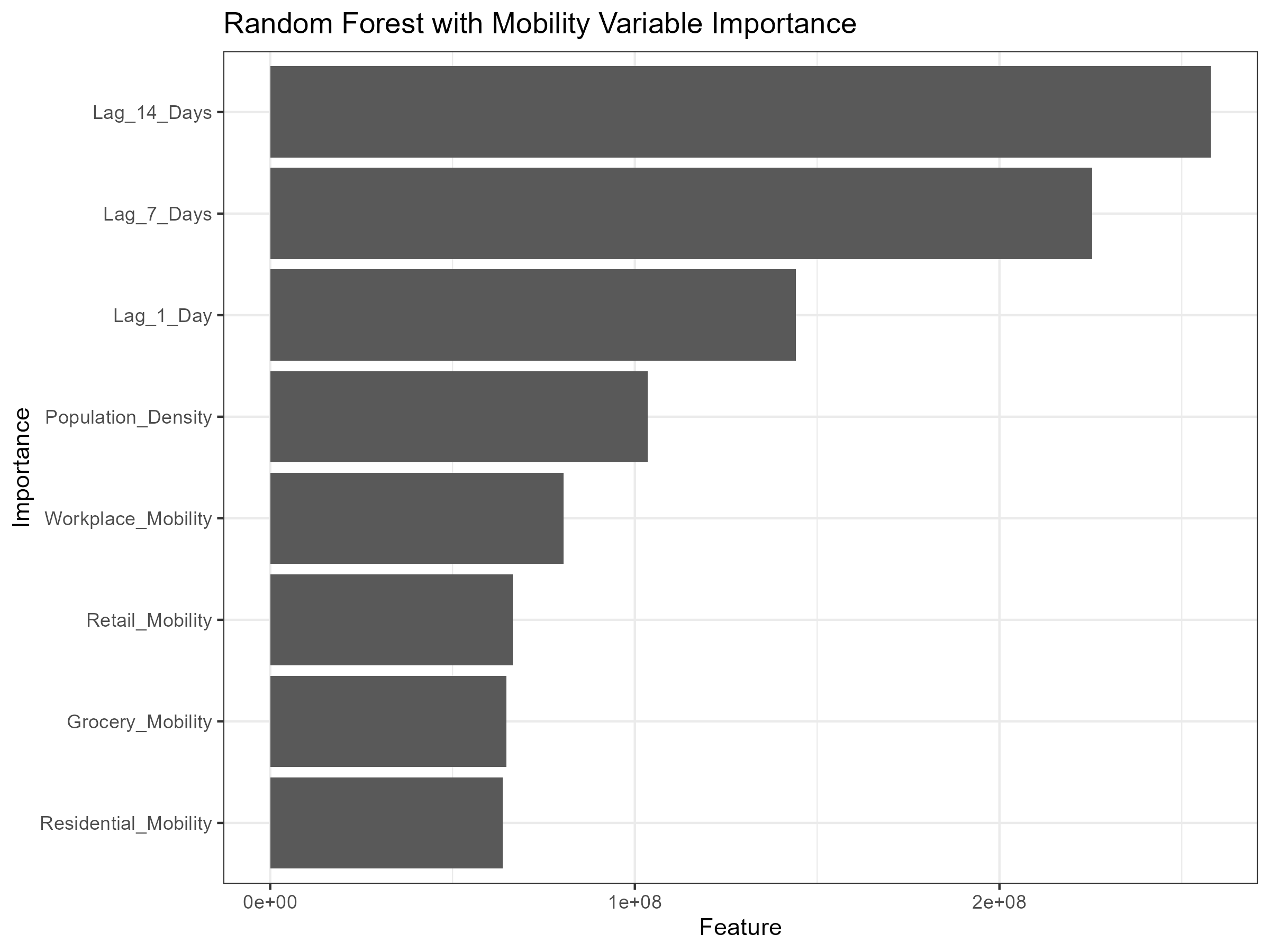


Figure 5.13: Variable Importance Plot of Mobility Included Random Forest Model

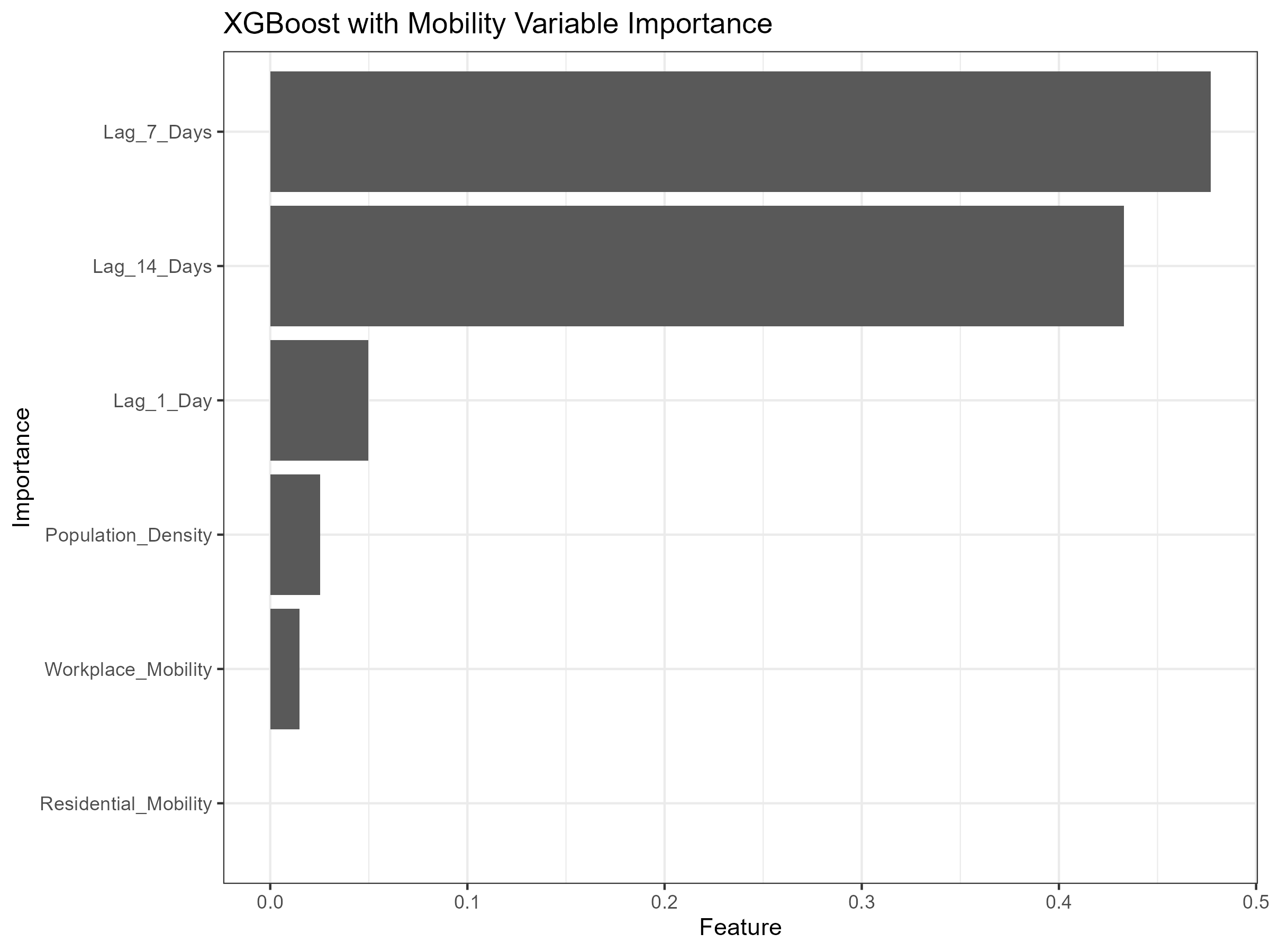


Figure 5.14: Variable Importance Plot of Mobility Included XGBoost Model

Across the three models, LASSO regression identified the mobility predictors as highly important, indicating a strong relationship between the mobility dynamics and case incidence. Random Forest indicated moderate importance to these predictors while XGBoost found the predictors to be insignificant.

Additional VIP plots for the other predictor sets can be found in the supplementary materials for the analysis.

# 6. Discussion

## 6.1 Model Performance

This study aimed to investigate the extent to which mobility dynamics predict variation in COVID-19 case incidence at the county level in Georgia using three different machine learning methods, LASSO, Random Forest, and XGBoost. To assess practical predictive performance, RMSE and MAE were also reported

Among all models, LASSO with mobility predictors achieved the best overall performance, with an RMSE of 46.5, an R² of 0.843, and MAE of 26.9. The inclusion of mobility predictors were able to improve the RMSE, MAE, and R² from the baseline model. In addition to improvements among the baseline LASSO model, this model outperformed all models in R² and had great relative performance in RMSE and MAE, suggesting that mobility predictors contributed meaningfully to the model’s capacity to explain variance in COVID-19 cases while also improving predictive accuracy. LASSO identified mobility variables as key predictors, suggesting a linear relationship between mobility trends and COVID-19 case incidence. This is consistent with previous literature where linear regression models incorporating mobility data have been effective in case forecasting.

Considerable improvement in MAE and RMSE were observed in mobility included models for Random Forest. Compared to baseline, the mobility included predictors had a MAE reduction of 21.72% and 10% reduction in RMSE. The lagged predictor set performed even better, with a 38.87% reduction in MAE and 26.5% reduction in RMSE. These predictors were assigned moderate importance, indicating that that both non-lagged and lagged mobility predictors contributed useful predictive value. However, the inclusion of mobility predictors resulted in only a slight improvement in R², suggesting that while these features reduced prediction errors, they did not substantially enhance the model’s ability to explain the overall variance in COVID-19 case incidence. This may imply that the model’s explanatory power mostly comes from the other variables present rather than mobility. Additionally, unlike in LASSO, mobility predictors were not dominant drivers of case trends.

In contrast, XGBoost had strong baseline predictive performance but saw modest or negative improvement gains with the inclusion of mobility predictors. Additionally, investigating variable importance plots further, XGBoost tended to rate the mobility predictors as unimportant. XGBoost’s use of “greedy” optimization can cause the modeling technique to overlook weak or redudant predictors, like mobility, especially if their effect is context-dependent or overshadowed by stronger features. The noisiness or redundancy found in the mobility predictors may lead to its low importance.

While mostly exploratory, preliminary, and limited to this study’s context, the results suggest that the inclusion of lagged mobility predictors can significantly reduce prediction error metrics such as MAE and RMSE, specifically in LASSO and Random Forest models. These reductions indicate improved short-term predictive accuracy, even when overall explanatory power does not significantly improve. With this increase in short term forecasting accuracy, public health agencies can track mobility shifts to anticipate potential case surges and take more proactive approaches for prevention, such as medical resource allocation or adjusting guidelines of interventions. However, the inclusion of mobility predictors seemingly did not significantly improve the R² of models, suggesting that these predictors do not improve the explanatory power of models. While mobility trends can be a useful input for predictive surveillance systems, they are not sufficient alone to explain complex patterns of COVID-19 transmission, reaffirming the need to include additional structural or behavioral predictors to achieve robust model performance.

## 6.2 Mobility Dynamics and Individualism

The CMR’s explanatory capacity is seemingly greater in non-American contexts. For example, in Jakarta, the CMR was able to explain 52% of the variation in case counts with just the mobility predictors (Nanda et al., 2022). This could be the result of complex social and political contexts persisting in the US. Mobility trends may lack explanatory power in the US if they do not consistently follow or correlate with transmission. In other words, inconsistency in American behavior could limit the CMR’s explanatory power.

It is posited that the US has a “cultural default” of being more focused on individualism in which a person is prioritized over the needs of the group while some cultures, like those found in East Asia, value collectivism (Markus et al., 2024). Bazzi et al. (2021), has demonstrated that the unique sense of individualism, coined “rugged individualism” is a prominent feature of American culture that has led to hampered pandemic response, specifically, worse adherence to social distancing guidelines, mask usage, and a weaker local government effort to control the virus spread. This sense of individualism may have led to noisy and complex mobility patterns independent of COVID-19 case incidence. For instance, large spikes in cases may not affect mobility patterns due to this disregard of the severity of the pandemic.

Investment into other data collection strategies in the midset of a pandemic may prove to be more effective in forecasting disease incidence. Additionally, the demonstration of American individualism during the pandemic may be an indication for public health agencies to begin focusing on developing prevention control methods that cater to U.S. culture.

## 6.3 Strengths and Limitations

This analysis has several limitations. First, the mobility data was incomplete, leading to the filtering of majority of Georgia counties out of the data set. This could be due to the methodology used for measuring which would skew the data towards more suburban and urban areas where Google product usage is more common. This can also mask behaviors related to communities on a more individual scale. Additionally, this study did not incorporate many variables relating to policy changes or other events that could precede and explain changes in COVID-19 incidence. Lastly, machine learning models can be highly sensitive to parameter tuning and data pre-processing. In some instances, models like XGBoost and Random Forest are at risk of overfitting (Jha, 2024).

A strength of this analysis is the inclusion of open source data, along with basic feature engineering, which allows for reproduction and cost-efficiency.

## 6.4 Conclusions

This research explores the potential of incorporating mobility dynamics into public health surveillance, specifically, for COVID-19 case forecasting. The improvement of models, in the metrics of interest, suggest the added predictive value of mobility predictors, indicating that these variables can improve forecasting accuracy for COVID-19 case incidence. Limited improvements in explanatory power cautions against over-reliance on mobility predictors and demonstrate that mobility alone is insufficient to fully account for the complex dynamics of disease spread. Future research should explore the integration of additional behavioral predictors, examine these findings in different geographic and sociopolitical contexts, and assess how mobility patterns interact with other public health interventions to influence disease transmission dynamics.

# 7. References

Alfred, R., & Obit, J. H. (2021). The roles of machine learning methods in limiting the spread of deadly diseases: A systematic review. *Heliyon*, *7*(6), e07371. <https://doi.org/10.1016/j.heliyon.2021.e07371>

Badr, H. S., Du, H., Marshall, M., Dong, E., Squire, M. M., & Gardner, L. M. (2020). Association between mobility patterns and COVID-19 transmission in the USA: A mathematical modelling study. *The Lancet Infectious Diseases*. <https://doi.org/10.1016/s1473-3099(20)30553-3>

Bazzi, S., Fiszbein, M., & Gebresilasse, M. (2021). “Rugged individualism” and collective (In)action during the COVID-19 pandemic. *Journal of Public Economics*, *195*, 104357. <https://doi.org/10.1016/j.jpubeco.2020.104357>

Bengtsson, H. (2025). Future: Unified Parallel and Distributed Processing in R for Everyone [R package future version 1.40.0]. *R-Project.org*. <https://cran.r-project.org/web/packages/future/index.html>

Breiman, L. (2001). *Random Forests*. Statistics Department University of California Berkeley. <https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf>

Centers for Disease Control and Prevention. (2024). *COVID-19 Timeline*. Centers for Disease Control; Prevention; CDC. <https://www.cdc.gov/museum/timeline/covid19.html>

Chang, W., Schloerke, B., Posit Software, & PBC. (2023). Take Screenshots of Web Pages [R package Webshot2 version 0.1.1]. *R-Project.org*. <https://cran.r-project.org/web/packages/webshot2/index.html>

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’16*, *1*(1), 785–794. <https://doi.org/10.1145/2939672.2939785>

Cipolletta, S., Andreghetti, G. R., & Mioni, G. (2022). Risk Perception towards COVID-19: A Systematic Review and Qualitative Synthesis. *International Journal of Environmental Research and Public Health*, *19*(8), 4649. <https://doi.org/10.3390/ijerph19084649>

CSSEGISandData. (2022). *COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University*. GitHub. <https://github.com/CSSEGISandData/COVID-19>

Frick, H., Chow, F., Khun, M., Mahoney, M., Silge, J., Wickham, H., Posit Software, & PBC. (2025). Rsample: General Resampling Infrastructure [R package rsample version 1.3.0]. *R-Project.org*. <https://cran.r-project.org/web/packages/rsample/index.html>

Friedman, J., Hastie, T., Tibshirani, R., Narasimhan, B., Tay, K., Simon, N., & Qian, J. (2021). *Glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models*. R-Packages. <https://cran.r-project.org/web/packages/glmnet/index.html>

Google. (2020a). *COVID-19 Community Mobility Report*. COVID-19 Community Mobility Report. <https://www.google.com/covid19/mobility/>

Google. (2020b). *Overview - Community Mobility Reports Help*. Google.com. <https://support.google.com/covid19-mobility/answer/9824897?hl=en>

Greenwell, B., & Boehmke, B. (2023). Vip: Variable Importance Plots [R package vip version 0.4.1]. *R-Project.org*. <https://cran.r-project.org/web/packages/vip/index.html>

Hart, A. (2025). *The COVID-19 pandemic: A Georgia timeline*. ajc; The Atlanta Journal-Constitution. <https://www.ajc.com/news/health-news/the-covid-19-pandemic-a-georgia-timeline/KW7BULU2NREIDKWN5PD5FY53PU/>

IBM. (2024). *Lasso Regression*. Ibm.com. <https://www.ibm.com/think/topics/lasso-regression>

Irini, F., Kia, A. N., Shannon, D., Jannusch, T., Murphy, F., & Sheehan, B. (2021). Associations between mobility patterns and COVID-19 deaths during the pandemic: A network structure and rank propagation modelling approach. *Array*, *11*, 100075. <https://doi.org/10.1016/j.array.2021.100075>

Jha, G. (2024). *Popular Machine Learning Models Prone to Overfitting and Why It Happens*. Medium. <https://medium.com/@post.gourang/popular-machine-learning-models-prone-to-overfitting-and-why-it-happens-8050e9c3a944>

Kuhn, M., Vaughan, D., Hvitfeldt, E., Posit Software, & PBC. (2025). Parsnip: A Common API to Modeling and Analysis Functions [R package parsnip version 1.3.1]. *R-Project.org*. <https://cran.r-project.org/web/packages/parsnip/index.html>

Kuhn, M., Wickham, H., Posit Software, & PBC. (2023). *Tidymodels: Easily Install and Load the ’Tidymodels’ Packages*. R-Packages. <https://CRAN.R-project.org/package=tidymodels>

Markus, H. R., Tsai, J. L., Uchida, Y., Yang, A. M., & Maitreyi, A. (2024). Cultural Defaults in the Time of COVID: Lessons for the Future. *Psychological Science in the Public Interest*, *25*(2), 41–91. <https://doi.org/10.1177/15291006241277810>

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.-C., & Lin, C.-C. (2022). *E1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R-Packages. <https://cran.r-project.org/web/packages/e1071/index.html>

Mock, T., & Sjoberg, D. (2023). Extending ’gt’ for Beautiful HTML Tables [R package gtExtras version 0.5.0]. *R-Project.org*. <https://cran.r-project.org/web/packages/gtExtras/index.html>

Müller, K., & Bryan, J. (2020). *Here: A Simpler Way to Find Your Files*. R-Packages. <https://cran.r-project.org/web/packages/here/index.html>

Nanda, R. O., Nursetyo, A. A., Ramadona, A. L., Imron, M. A., Fuad, A., Setyawan, A., & Ahmad, R. A. (2022). Community Mobility and COVID-19 Dynamics in Jakarta, Indonesia. *International Journal of Environmental Research and Public Health*, *19*(11), 6671. <https://doi.org/10.3390/ijerph19116671>

National Park Service. (2017). *Georgia (U.S. National Park Service)*. Nps.gov. <https://www.nps.gov/state/ga/index.htm>

Neuwirth, E. (2022). *RColorBrewer: ColorBrewer Palettes*. R-Packages. <https://cran.r-project.org/web/packages/RColorBrewer/index.html>

Ogunjo, S. T., Fuwape, I. A., & Rabiu, A. B. (2022). Predicting COVID‐19 Cases From Atmospheric Parameters Using Machine Learning Approach. *GeoHealth*, *6*(4). <https://doi.org/10.1029/2021gh000509>

Paez, A. (2020). Using Google Community Mobility Reports to investigate the incidence of COVID-19 in the United States. *Transport Findings*. <https://doi.org/10.32866/001c.12976>

Paltra, S., Bostanci, I., & Nagel, K. (2024). The effect of mobility reductions on infection growth is quadratic in many cases. *Scientific Reports*, *14*(1). <https://doi.org/10.1038/s41598-024-64230-1>

Pebesma, E., Bivand, R., Racine, E., Summer, M., Cook, I., Keitt, T., Lovelace, R., Wickham, H., Ooms, J., Muller, K., Pedersen, T. L., Baston, D., & Dunnington, D. (2021). *Sf: Simple Features for R [R package sf version 0.9-8]*. cran.r-project.org. <https://cran.r-project.org/web/packages/sf/index.html>

Pedersen, T. L. (2020). *Patchwork: The Composer of Plots*. R-Packages. <https://cran.r-project.org/web/packages/patchwork/index.html>

Rashidi, H. H., Albahra, S., Robertson, S., Tran, N. K., & Hu, B. (2023). Common statistical concepts in the supervised Machine Learning arena. *Frontiers in Oncology*, *13*. <https://doi.org/10.3389/fonc.2023.1130229>

Ripley, B., Venables, B., Bates, D. M., port 1998), K. H. (partial., port 1998), A. G. (partial., & Firth, D. (2022). *MASS: Support Functions and Datasets for Venables and Ripley’s MASS*. R-Packages. <https://cran.r-project.org/web/packages/MASS/index.html>

Robinson, D., Hayes, A., Couch, S., Patil, I., Chiu, D., Gomez, M., Demeshev, B., Menne, D., Nutter, B., Johnston, L., Bolker, B., Briatte, F., Arnold, J., Gabry, J., Selzer, L., Simpson, G., Preussner, J., Hesselberth, J., Wickham, H., … Sjoberg, D. D. (2022). *Broom: Convert Statistical Objects into Tidy Tibbles*. R-Packages. <https://cran.r-project.org/web/packages/broom/index.html>

Sivakumar, M., Parthasarathy, S., & Padmapriya, T. (2024). Trade-off between training and testing ratio in machine learning for medical image processing. *PeerJ Computer Science*, *10*, e2245–e2245. <https://doi.org/10.7717/peerj-cs.2245>

Spinu, V., Grolemund, G., Wickham, H., Vaughan, D., Lyttle, I., Costigan, I., Law, J., Mitarotonda, D., Larmarange, J., Boiser, J., & Lee, C. H. (2022). *Lubridate: Make Dealing with Dates a Little Easier*. R-Packages. <https://cran.r-project.org/web/packages/lubridate/index.html>

Sulyok, M., & Walker, M. (2020). Community movement and COVID-19: A global study using Google’s Community Mobility Reports. *Epidemiology and Infection*, *148*. <https://doi.org/10.1017/s0950268820002757>

Walker, K. (2025). Tigris: Load Census TIGER/Line Shapefiles [R package tigris version 2.2.1]. *R-Project.org*. <https://cran.r-project.org/web/packages/tigris/index.html>

Waring, E., Quinn, M., McNamara, A., Rubia, E. A. de la, Zhu, H., Lowndes, J., Ellis, S., McLeod, H., Wickham, H., Müller, K., RStudio, Kirkpatrick, C., Brenstuhl, S., Schratz, P., Korpela, M., Thompson, J., McGehee, H., Roepke, M., Kennedy, P., … Stewart, H. M. (2022). *Skimr: Compact and Flexible Summaries of Data*. R-Packages. <https://CRAN.R-project.org/package=skimr>

Wei, T., Simko, V., Levy, M., Xie, Y., Jin, Y., & Zemla, J. (2017). *Corrplot: Visualization of a Correlation Matrix*. R-Packages. <https://cran.r-project.org/web/packages/corrplot/index.html>

Wickham, H. (2019). Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. *R-Packages*. <https://cran.r-project.org/web/packages/ggplot2/index.html>

Wickham, H. (2020). *Reshape2: Flexibly Reshape Data: A Reboot of the Reshape Package*. R-Packages. <https://cran.r-project.org/web/packages/reshape2/index.html>

Wickham, H., Bryan, J., RStudio, Kalicinski, M., Valery, K., Leitienne, C., Colbert, B., Hoerl, D., & Miller, E. (2019). *Readxl: Read Excel Files*. R-Packages. <https://cran.r-project.org/web/packages/readxl/index.html>

Wickham, H., François, R., Henry, L., Müller, K., & RStudio. (2020). *Dplyr: A Grammar of Data Manipulation*. R-Packages. <https://cran.r-project.org/web/packages/dplyr/index.html>

Wickham, H., & Henry, L. (2020). *Tidyr: Tidy Messy Data*. R-Packages. <https://cran.r-project.org/web/packages/tidyr/index.html>

Wickham, H., & RStudio. (2019). *Stringr: Simple, Consistent Wrappers for Common String Operations*. R-Packages. <https://cran.r-project.org/web/packages/stringr/index.html>

Wickham, H., Seidel, D., & RStudio. (2022). *Scales: Scale Functions for Visualization*. R-Packages. <https://cran.r-project.org/web/packages/scales/index.html>

Zeileis, A., Grothendieck, G., Ryan, J. A., Ulrich, J. M., & Andrews, F. (2021). *Zoo: S3 Infrastructure for Regular and Irregular Time Series (Z’s Ordered Observations)*. R-Packages. <https://cran.r-project.org/web/packages/zoo/index.html>