Addressing Inequality Through Modeling*

Updating Public Defense Funding Models In Washington State

Simon Ahn[†] Taylor Blair[‡] Robin Hardwick[§] Maxwell J.D. VanLandschoot[¶]

20 December, 2021

Abstract

This report explores and offeres alternatives to the current model used to calculate county public defense funding in Washington State. The Office of Public Defense (OPD) has identified multiple shortcomings in their current methodology including: missing variables, disequitable funding allocation, difficulties with interpretability, and abritrary or unfounded model coeficients. In this report, we consider and evaluate eight distinct multilinear and decision tree models to attempt to address the considerations raised by the OPD. Our least absolute shrinkage and selection operator (LASSO) model, in particular, is extremely promising as a candidate for further reseach and refinement.

Contents

*The authors are thankful to the Washington State Office of Public Defense –specifically Katrin Johnson for meeting to discuss RCW 10.101 and for providing data– and to the University of Washington's Center for Real Estate Research for their 2019 housing market reports.

[†] Reed College, ahns@reed.edu

[‡] Reed College, tblair@reed.edu

[§] Reed College, robhardwi@reed.edu

[¶] Reed College, maxvanlan@reed.edu

1 Introduction

The goal of this report is to investigate the RCW 10.101.060 funding model that serves as the current basis for the Washington's Office of Public Defense (henceforth referred to as OPD) to disburse funds to county offices. We identify, through this investigation, trends in spending over time, pitfalls in the current model, as well as identify how the funding scheme can be improved. While the 10.101.060 funding model is sufficient in ensuring counties receive equitable funding by population and the cases filed –an approximation for how busy public defenders are in the county– it lacks consideration for other socioeconomic factors. With the two variables considered in RCW 10.101, for instance, the percentage of county residents who live below the poverty line, household incomes, and housing costs are not taken into account. The considering of a county's caseload and population are undoubtedly important variables, but if a county has to fund more resources for its residents, they simply have less capacity than a comparatively richer county to dedicate funds to public defense. Thus, we push for the importance of beginning a conversation to reform the RCW funding model, wherein such socioeconomic complexities are considered("RCW 10.101.060" 2005).

2 Methods

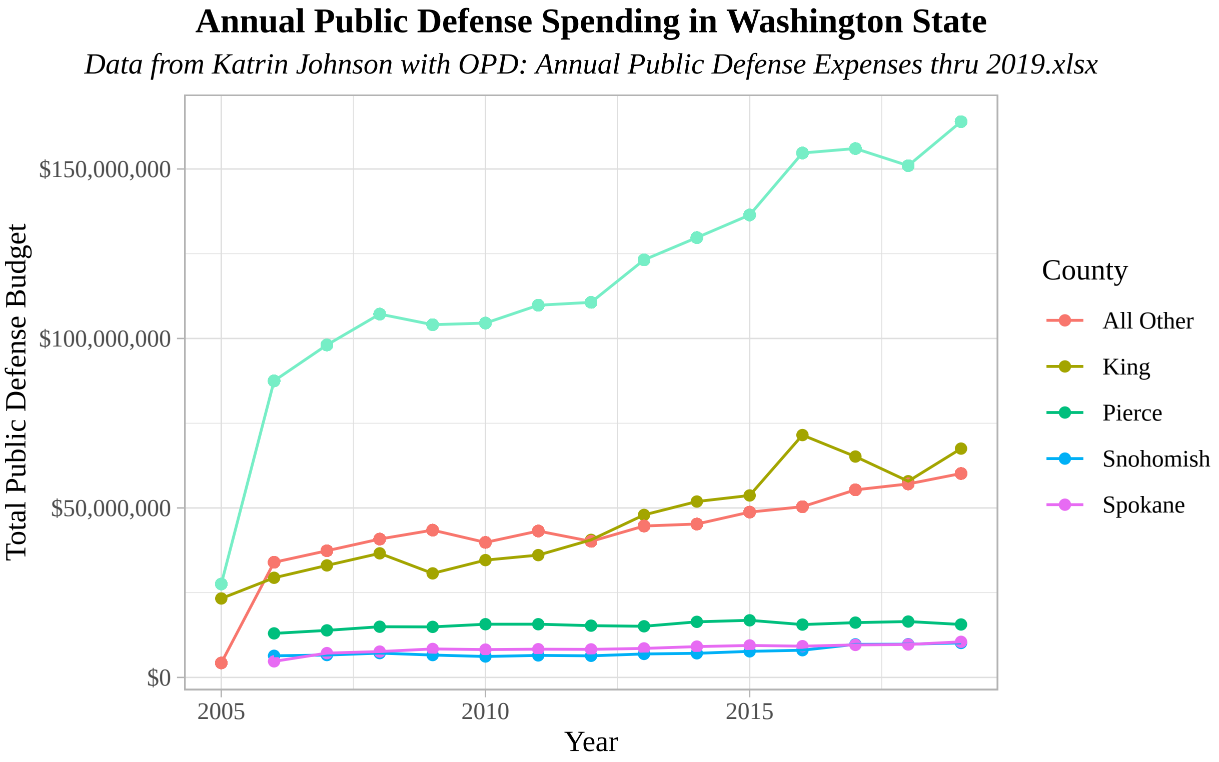
Our exploratory analysis utilizes six data files, centrally focusing on `OPD 10.101 Funding Over Time.csv` and `calculating_10_101_2021.csv` in an exploration of how the funds themselves are being distributed by county. From here, we explore how different variables are taken into account when determining how much funding an individual county receives from the OPD. As expected, those values included in the 10.101 budget allocation model (caseload and population) are greatly reflected in the funding distributions by county. Other variables, however, including housing costs, median poverty levels, median income, etc. are *not* accounted for in this model.

Every unit of `Annual Public Defense Spending by County.csv` refers to the amount of money each county spent in a unique year on public defense (including, but not limited to, the 10.101 state funds distributed to them). `OPD 10.101 Funding Over Time.csv` is much the same, except it *only* includes the funds each county spent on public defense received from the state through the 10.101 application program. It is using this data set that we calculated the *proportion of that year's state budget* allocated to each county. `calculating_10_101_2021.csv` shows the process of the state calculating the amount of money each county should receive using their current budget allocation formula. `Caseload Resources and Capacity Measured in 2018.csv` includes relevant information from 2018 about what specific resources and cases each counties are taking on. Each observation refers to one of the 39 counties, and the columns different attributes about their resources. `County Home Prices.csv` is information about median housing costs in Washington State over the last few quarters, where each row is also a county. Finally, we also use `County Statistics from 2020.csv`, which is very similar to `Caseload Resources and Capacity Measured in 2018.csv`, except containing information from 2020. Each row is one of the 39 counties. For additional information into the wrangling and data processing, reference Code Appendix 6.2.

3 Exploratory Data Analysis

3.1 Public Defense Spending Over Time

We begin our exploratory analysis generally, looking at the general trends of public defense spending from 2005-2019. As seen in the graph below, statewide funding (the top, light blue line) and funding for King County (the state's most populous which encompass much off the Seattle-Metropolitan Area) has increased over time. The state's 2nd, 3rd, and 4th most populous counties, Pierce, Snohomish, and Spokane respectively, have seen virtually no increase in funding over time. Noteworthy, too, is the fact that between 2005 and 2006, over 20 counties began reporting on funding data. Though beyond the scope fo this report, this graph begs the question of what spending on public defense looked like prior to 2005, before data was able to be collected by a centralized entity. Were indigent defendants receiving adequate legal care, supervision, and resources?



Annual Public Defense Spending in Washington State

Data from Katrin Johnson with OPD: Annual Public Defense Expenses thru 2019.xlsx

**29 counties did not report data in 2005. Douglas County is missing data in 2005...2008 and 2012*

3.1.1 County Statistics

The graph and table below add *all* of a county's reported spending to OPD since 2005. We see that the large counties, King, Pierce, etc, which overwhelmingly lead nominal funding statistics, are relatively average on a per-captia funding basis. In terms of nominal funding, counties rank roughly in order of population, with some notable standouts being Yakima and Whatcom. Other counties have such negligible differences between one another, raising the question, should this be the case, given the diversity in socioeconomic factors between them? We will return to this question in the 10.101 model section.

Public Defense Spending Per Capita in 2019

Data from the Washington State Office of Public Defense

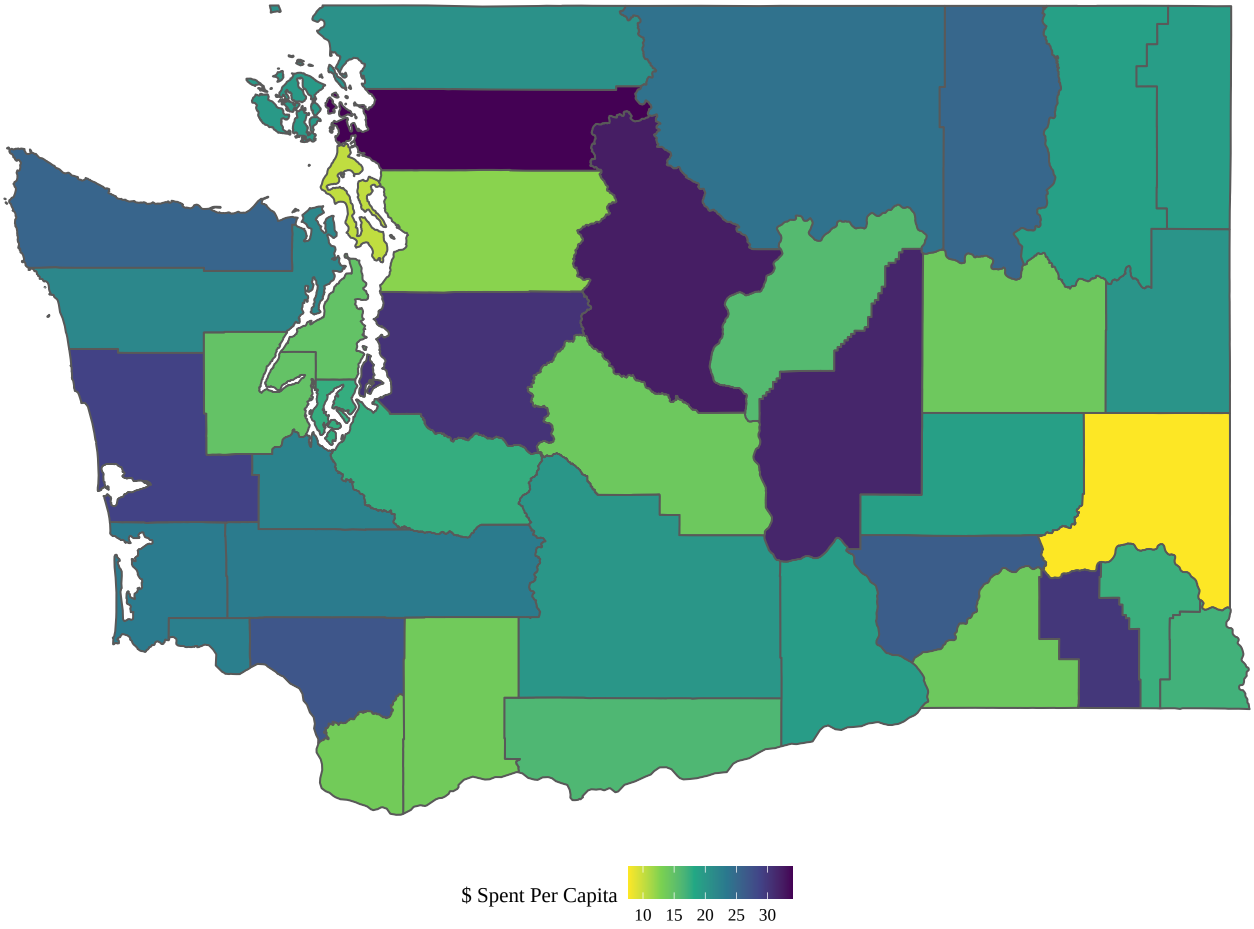
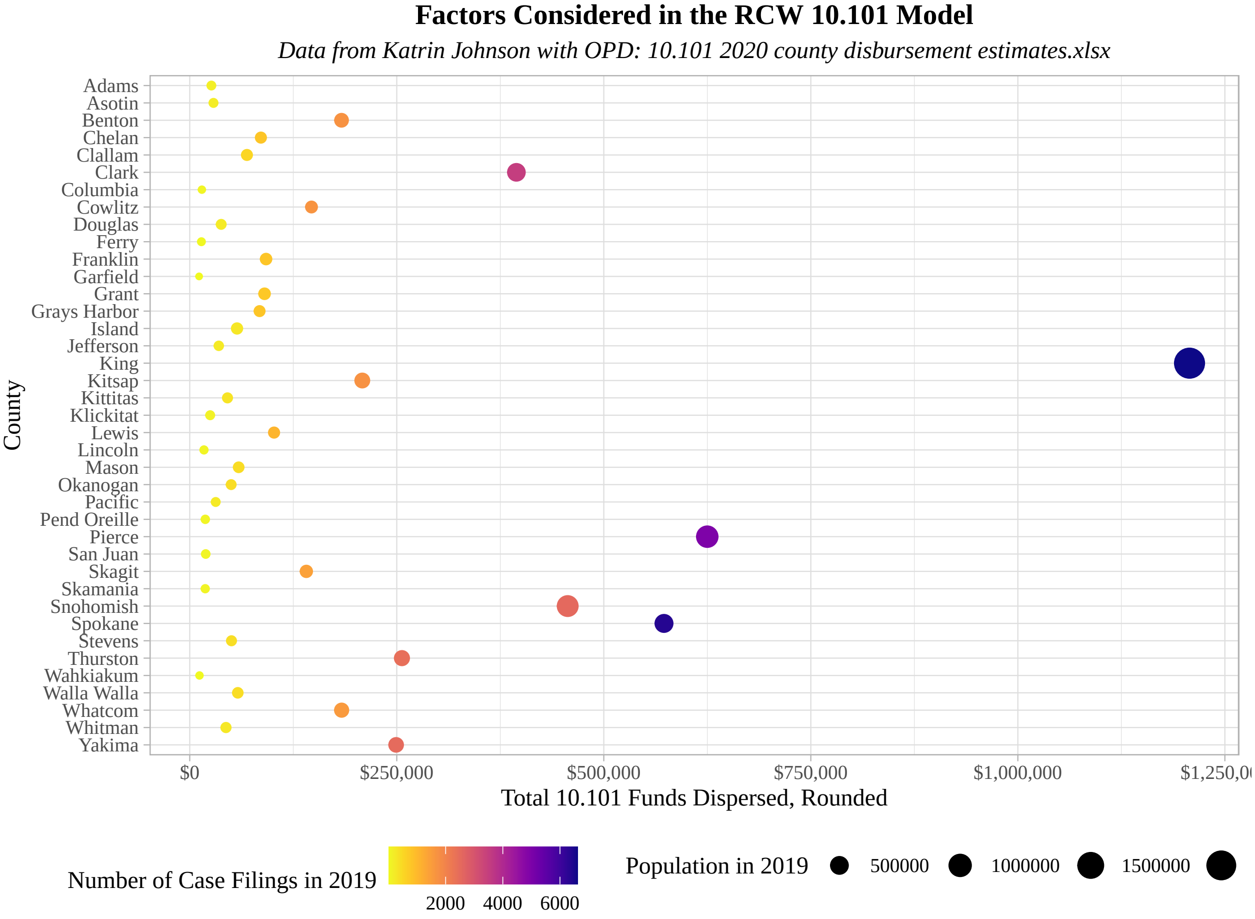


Table 1: County Budget Statistics

County	Average Budget 2005-2019	Budget Per Capita 2019
Adams	384054.53	19.11
Asotin	302813.33	17.00
Benton	2952892.21	19.48
Chelan	1814301.64	32.23
Clallam	1331168.60	25.42
Clark	5157256.36	13.62
Columbia	138556.79	30.00
Cowlitz	2171080.64	26.89
Douglas*	561369.33	15.86
Ferry	158235.43	25.17
Franklin	993369.93	26.26
Garfield	35708.43	17.31
Grant	2733708.40	31.53
Grays Harbor	1002318.07	28.96
Island	790034.21	10.30
Jefferson	524320.47	21.76
King	45326994.67	30.37
Kitsap	3154066.07	14.87
Kittitas	554425.71	14.09
Klickitat	260172.73	16.25
Lewis	1480355.43	23.09
Lincoln	130446.43	14.02
Mason	728952.14	14.80
Okanogan	945511.71	24.09
Pacific	350673.53	23.15
Pend Oreille	246891.57	19.37
Pierce	15402073.79	17.59
San Juan	236196.07	19.80
Skagit	2647859.36	33.98
Skamania	120799.33	13.76
Snohomish	7522540.57	12.42
Spokane	8492490.57	20.39
Stevens	608990.93	19.01
Thurston	3521770.71	22.31
Wahkiakum	72816.13	22.63
Walla Walla	698817.27	14.10
Whatcom	3985603.07	20.73
Whitman	321284.14	7.67
Yakima	4650878.71	20.11

3.2 The 10.101 Model

In the following section we explore of the specific facets of the 10.101 model, as well as visualize the types of variables excluded and their impact on specific counties. As outlined in RCW 10.101.070, funding is dispersed as follows: 6% of total state funding is divided equally amongst counties; 47% shall be dispersed to counties on the basis of population proportion; and 47% shall be disbursed based on the "annual number of criminal cases filed in the county superior court." RWC 10.101.070 continues in more detail, but is generally captured by the three listed metrics.



The plot above helpfully visualizes how the 10.101 formula disperses money to counties on the basis of population and case filings.

3.3 Additional Model Factors

3.3.1 Poverty Rates

Having firmly established the parameters of the 10.101 model, we introduce the poverty levels as recorded from 2015-2019 in OPD's "County Statistics 2020" Excel sheet as a parameter for model creation. Appendix 1 contains two graphs depicting the relationship between poverty levels and funding schema. The first plot looks at each county's entire public defense budget. One should not that counties with high rates of poverty spent the least amount of money on public defense in 2020. At low funding levels, however, spending is not a perfect predictor for poverty as some counties spend minimally and maintain low rates of poverty. As funding rises above the five million dollar range, rates of poverty tend to decrease. This correlation is even more prevalent in the most highly funded counties as, once funding surpasses ten million dollar per year, counties report less than 15% of their residents living below the poverty line. The second plot in Appendix 1 looks directly at how 10.101 funding discounts the percentage of a county's residents who are below the poverty level. There are several counties who receive similar funds –due to their population sizes and case filings– but which have differing poverty rates, impacting the amount of funds and resources on hand for public defense. Adams and Asotin counties, for instance, perfectly demonstrate the funding disparity between counties of similar size with different rates of poverty.

3.3.2 Median Household Income

Counties with larger median household income, as seen in Appendix 2, generally have higher public defense budgets. We postulate this relationship exists because the higher median incomes are, the more money the county collects in taxes, and thus the more resources at hand to allocate. It may also be the case that wealthier counties place a greater emphasis on the funding of public services, but we have no way of measuring this through the data we have compiled. Looking at the 10.101 model, we find that the relationship between income and funding is not as stark, though, counties with higher household incomes generally receiving the most 10.101 funds from OPD. Counties of note are King, Clark, Snohomish, and Pierce, all of whom have household incomes above other counties yet still receive more funds than them –Spokane County is the outlier among wealthy counties.

3.3.3 Median Home Cost

Finally, we explore how the median cost of homes by county intersects with the county's public defense spending and funds received by the OPD. This section uses the same data from OPD, but also data from the University of Washington's State Housing Market Resources and Reports (specifically 2019, Quarter 4 to line up with the data obtained from 2019 from the past two sections). Appendix 3 shows that generally, the higher the median home price, the more that county allocates on public defense. There is an argument to be made that more populated counties have more expensive housing markets, and thus of course would spend more on public defense due to more citizens and cases (both of which are included in the 10.101 model). We will see if this is true by comparing the median home price to the 10.101 allocations. Interestingly, the trend of higher home prices corresponding to higher rates of defense spending is not reflected in the 10.101 model. The majority of counties –those receiving below $150,000 from the state– receive the same amount as other counties despite having different median home prices. For instance, look at Wahkiakum, Kittitas, Jefferson, Island, and Douglas in comparison to Adams, Asotin, Columbia, and Stevens. Other counties, such as Clark, King, Pierce, and Skamania have the highest median home prices yet still receive the most 10.101 funds(Moore 2018). Given how property taxes can provide local governments with more money, taking into consideration median house prices will allow counties to receive more equitable funding.

3.4 Exploratory Data Analysis Conclusions

This section uncovers how socioeconomic variables, percentage of individuals under the poverty level, median income, and median home cost, inform public defense spending by county in Washington State. When comparing overall public defense spending with these three variables, we generally find that the less poverty and the higher resources (income/home cost), the more that county spends on public resources in total. When zooming in to look at RCW 10.101 funds, however, many counties receive the same amount of funds as others despite how poverty levels, income, and home costs differ by county. This suggests the importance of improving the model from its current budget allocation formula, as it fails to consider other influencing factors on how much a county can spend on public defense—meaning some counties offer more adequate, robust public defense services than others. The current funding paradigm is not only inequitable, but may even be constitutionally questionable *Gideon vs. Wainwright* ("Gideon v. wainwright" 1963).

3.5 Evaluation Methodology

For evaluation of our model performances –the means by which we select and present evidence for the capabilities of our model– we will use the response variable *proportion of the yearly budget allocated to each county*. This measure indicates the fraction of the OPD budget recieved by one county in a given year. For example, if King County receives 30% of the funds, it will have a proportion of 0.30(Moore 2018). Our data spans 15 years and 39 counties (give or take some NA values), so we have about 585 data points with which to train and test our model. We will split these values into a 75%/25% training/test set, as well as use k-fold to determine which type of model (see below for a discussion of potential models) performs the best and most closely to the desired values yearly budget proportion. To be even more secure of our results, we will also *bootstrap* our data to have more available data points to fit to models, increasing both our model's accuracy and ability to narrow down the best fitting model. For the sake of interpretability we hope to present both methods to our desired audience, alongside an example of what bootstrapping does and the role it plays to help with modeling when working with fewer data points.

It is also important to discuss how our particular response variable and the data itself presents a unique challenge in regards to evaluation, as our goal here is not explicitly to perform inference or prediction. So, we will have to be creative, even before we construct a model, to identify a means of *validation* to determine which model yields the "best" yearly proportion per county. We propose two possible approaches to this problem. 1) We could create a full linear model with every relevant variable, regressing for the proportion of annual state funding to predict the proportion of the budget each county receives that year, and then attempting to create a simpler model with less predictors that can more reasonably estimate annual funding. 2) We could make a comparison to a model that looks at historical funding proportions and determining which variables drastically increase the operating costs of public defenders (is it caseloads, and if so, what types of cases? Other external resources? The presence of municipal funding? etc.). Both of these comparisons would serve as metrics of rMSE that we could then compare between a new, proposed model to argue for its increased ability to consider multiple variables when determining budget allocation. It will, additionally, be valuable to recreate previously used plots that show funding relative to other predictors considered in our models –such as median poverty levels– to show how our new model performs with these variables in mind.

3.6 Potential Models

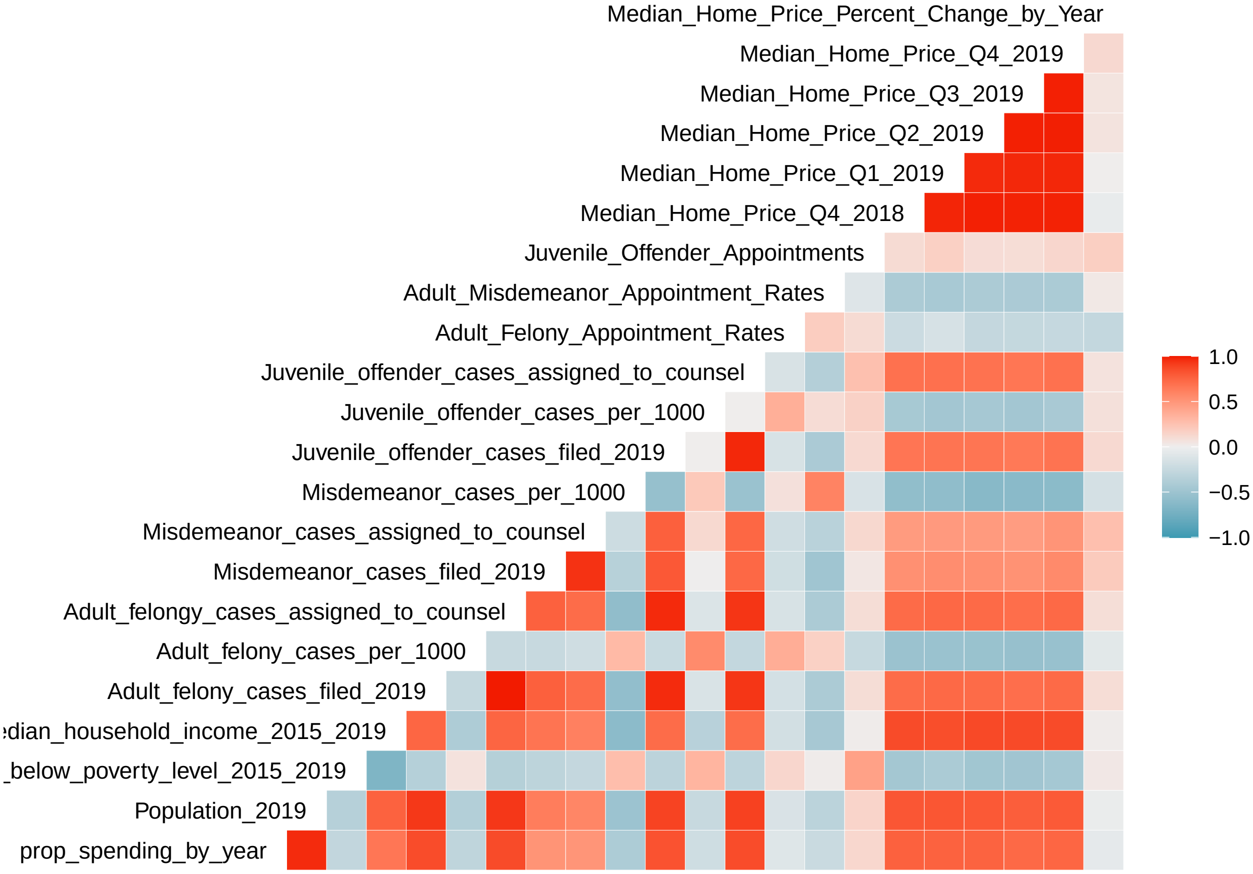
Considering the types of evaluation metrics available to us, we believe it most appropriate to use ridge regression or trees to build the "best" model. While other methods, notably LASSO, have advantages, we prefer ridge regression and trees in this context for their relative ease of understanding. This is an important consideration because the final audience for this report, government officials, may not have a robust technical background. We also do think that keeping all of our initial variables has some positive benefit as, even if a LASSO model could reduce rMSE further than a ridge regression model, it would feel inappropriate to leave out a variable like population or poverty rates. Random forests and bagging are two further techniques that would allow us to consider *every* possible explanatory value and data point from each county, while properly weighing their importance and significance on the final predicted values per county (this will also improve our predictions by allowing for us to have lower variance and bias, both issues in the context of this small data set across a very variable number of counties). We will though, of course, compare these models to ones created using other selection methods, like forward, backward, or manual selection(Moore 2018).

4 Results

In this section, we will walk through the eight models which we have fitted to predict the proportion of the RCW 10.101 public defense budget allocated to each county in 2020 using data from 2019. For each model, we present the model results and their attributes by looking at penalty sizes and feature selection. We will, further, compare these models and their accuracy, abstracting the results to inform OPD's budget questions of interest.

4.1 Full linear model

As a simple baseline, we create a *full* linear regression model, with all possible predictors included. We use all traditional linear modeling tools, such as line of least squares to calculate our coefficient values. To visualize the relationship between the elements of our full linear model, the heat map below was created. Clearly, many of our predictors are *highly* correlated with one another (over 0.5 in magnitude). This effect is especially prevalent for blocks of variables measuring similar attributes. Consider, for instance, the extremely high correlation between home price variables, as well as the high correlation between types of cases filed —-especially cases of the same category. These findings will likely pose issues in our full linear model due to the emergent multicollinearity.



Next, we used the four diagnostic plots below to assess if a full linear model is appropriate given the characteristics of the data. As shown, the model's residuals are certainly *not* normally distributed, nor are they even across residuals or even standardized residuals. There are even a few noticeable leverage points –likely King and Pierce Counties– that we should be cognizant of when making predictions. Most worrisome is the non-normally distributed residuals, which is likely occurring because of outlier counties with extremely large reletive budgets, population, government attention, and so on. Thus, most counties are similar in predictors and have lower fitted values in addition to overall lower residuals, so models have more predictive power in places that report more data. This is in contrast to the few higher fitted values, which have less observations, and thus less predictive power(James et al. 2013).

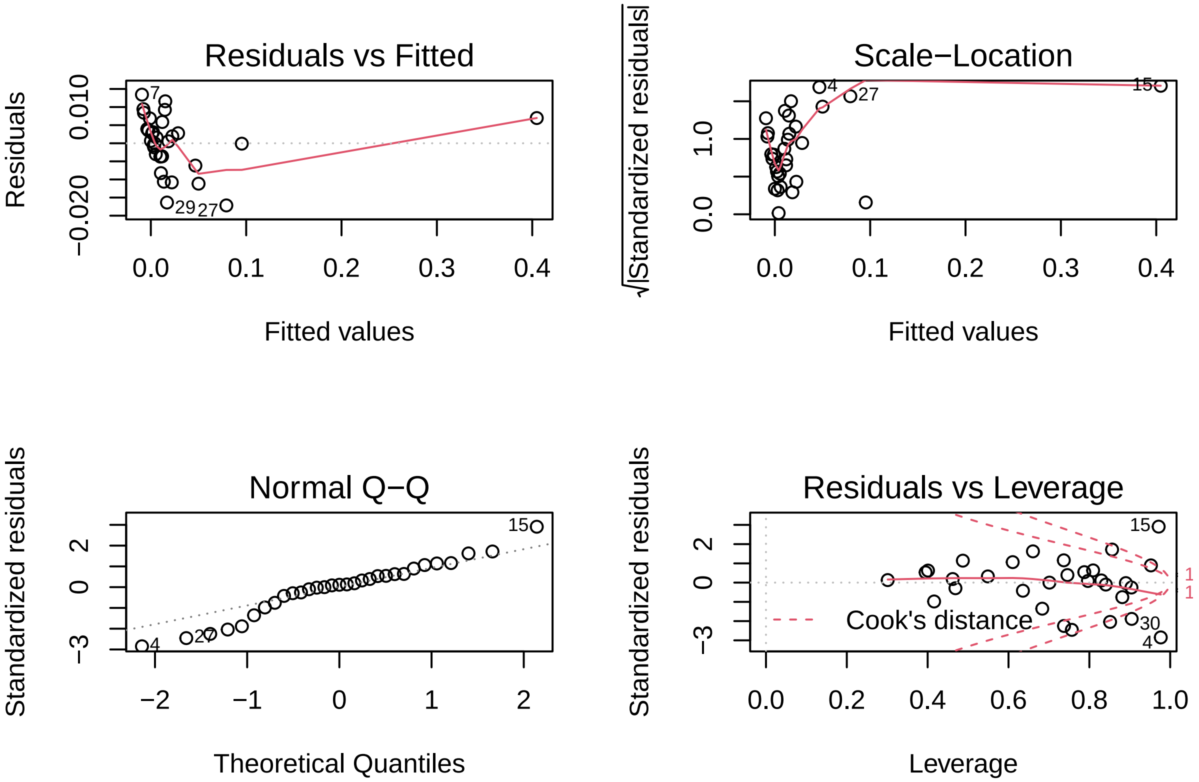


Table 2: Full Model Coefficients and Significance (Scaled by 100)

	Coefficient	p > |t|
(Intercept)	1.0938234	0.8881928
Population_2019	0.0000196	0.0002821
Percent_Individuals_below_poverty_level_2015_2019	16.9610942	0.3228544
Median_household_income_2015_2019	-0.0000872	0.3538241
Adult_felony_cases_filed_2019	-0.0125184	0.1263492
Adult_felony_cases_per_1000	0.1010382	0.5289460
Adult_felongy_cases_assigned_to_counsel	0.0124684	0.1546612
Misdemeanor_cases_filed_2019	-0.0004754	0.5165165
Misdemeanor_cases_assigned_to_counsel	0.0000698	0.9604754
Misdemeanor_cases_per_1000	0.1951371	0.0439635
Juvenile_offender_cases_filed_2019	-0.0110360	0.4659264
Juvenile_offender_cases_per_1000	-0.3268065	0.6815413
Juvenile_offender_cases_assigned_to_counsel	0.0151385	0.1748788
Adult_Felony_Appointment_Rates	2.0945733	0.5953053
Adult_Misdemeanor_Appointment_Rates	-3.2453707	0.1738182
Juvenile_Offender_Appointments	-3.8910377	0.2326657
Median_Home_Price_Q4_2018	-0.0001622	0.2507798
Median_Home_Price_Q1_2019	-0.0000194	0.4330230
Median_Home_Price_Q2_2019	0.0000506	0.1928434
Median_Home_Price_Q3_2019	0.0000249	0.4063801
Median_Home_Price_Q4_2019	0.0001063	0.3998337
Median_Home_Price_Percent_Change_by_Year	-0.2333054	0.4446537

The preceding table shows each variable's estimated coefficient as well as its statistical significance. A general note is that it is somewhat difficult to make a sweeping claim about coefficients because some of our variables are measured on quite different scales. For instance, one might be inclined to say that Adult Felony Appointment Rates have a larger impact, due to the larger coefficient, than population, but one also has to consider the fact that population measures can be in the millions for certain counties –far outshining the effect of most other variables. Population and case filings, as we expected, are significant at the 5% level. Several other variables, like Adult_Misdemeanor_Appointment_Rates and Juvenile_offender_cases_assigned_to_counsel, however, have promising levels of significance and are not considered in the 10.101 model. Performing a prediction with the reserved test data, too, returned an rMSE of 0.0264804. On first inspection, this value is promising, but as we will show, the full model has the second worst performance of any of the models we put forth in this report. A major limitation of the full linear model is the fact that there was no feature selection and variables like Misdemeanor_cases_assigned_to_counsel, with virtually no statistical significance, were allowed to remain in the model. In the next section, we will attempt to remedy these shortcomings in the full linear model by performing linear subset selection.

4.2 Linear Subsets

Linear subset selection can work in one of three ways. The first, most basic but time intensive way, is to create every possible linear combination of variables and calculate test rMSE. This method is often refered to as "best subsetting," and while it can produce the best linear models, it is incredibly inefficient. Bearing this in mind, we did not choose to perform best subset selection. The remaining two methods, backwards and forwards selection, however, we did perform. Forward selection works by starting with the single most explanatory variable and adds additional variables until the the model evaluation parameter begins to decrease. Backward selection works in reverse, starting from a full model and whittling down variables that did not have significant explanatory power. As seen in Table 3 and Table 4, these two model building devices each produce four models based on different assessment criteria: `adjr2.max`, `rss.min`, `cp.min`, and `bic.min`. The specific differences between the criteria are somewhat outside of the scope of this report, but they attempt to capture different aspects of the bias-variance trade off. For our purposes, we will continue forward with both of the `cp.min` models. Once created, all of the same rationale applies to these linear subset models as did to the full linear model. So as to not repeat ourselves, we will just note that both of these models were able to alleviate the overfitting problem of the full model and that the forward selection model performed on-par with more sophisticated models analyzed in the rest of this report.

Table 3: Evaluating Forward Selection

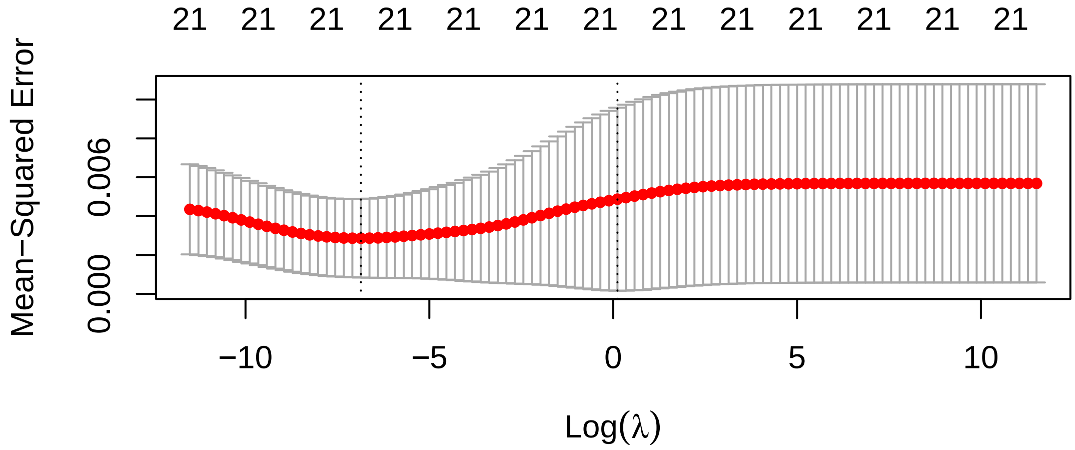
Model	Number of Predictors	rMSE
adjr2.max	14	0.0159818
rss.min	21	0.0264804
cp.min	8	0.0145060
bic.min	3	0.0129758

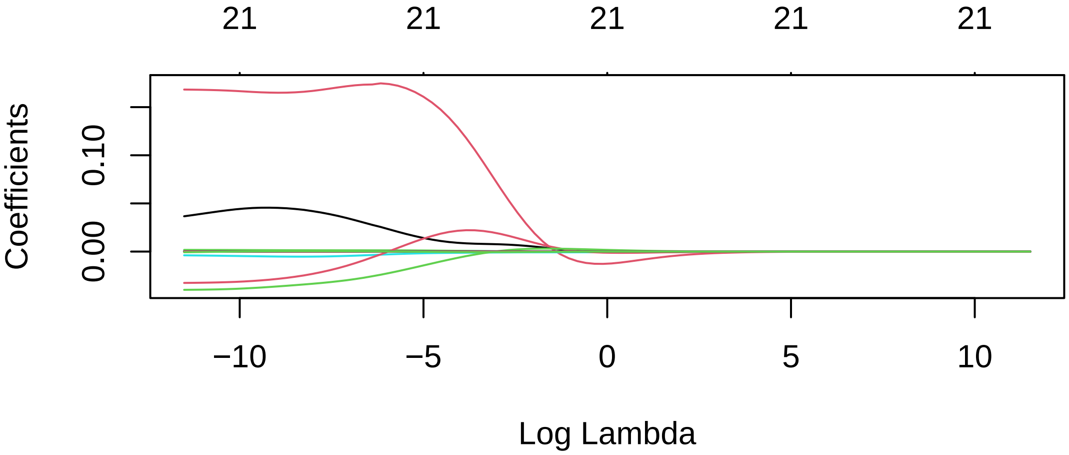
Table 4: Evaluating Backward Selection

Model	Number of Predictors	rMSE
adjr2.max	11	0.0189844
rss.min	21	0.0264804
cp.min	10	0.0186755
bic.min	11	0.0189844

4.3 Ridge Regression

Ridge Regression models are similar to the models explored previously, but help to mitigate violated assumptions. Through the usage of a shrinkage penalty λ, ridge regression will, through cross-validation tuned selection, shrink coefficients by a specified value to increase bias at the cost of drastically limiting variance(James et al. 2013).



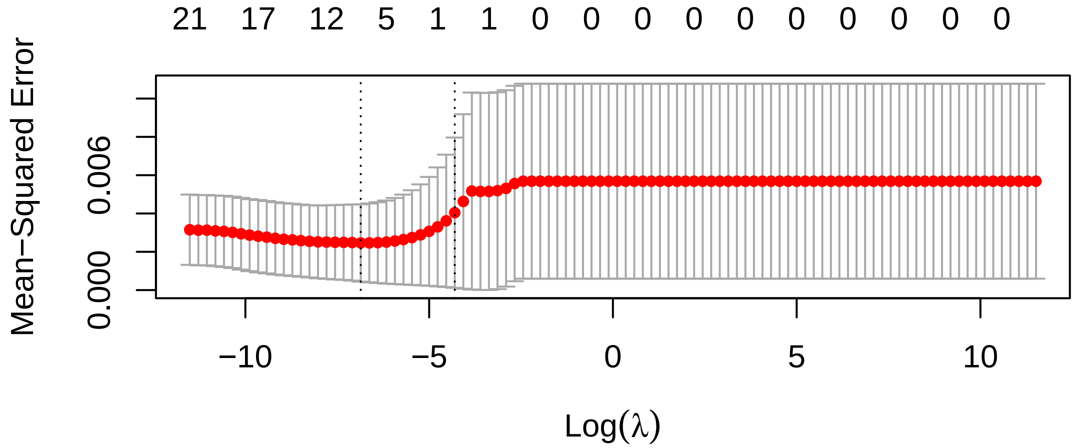


As the above two graphs show, the standard error and deviation at each lambda value is very large. The error seems to be the lowest is within the region from log(-7) to log(-1). It seems, from this, that a relatively small penalty is warranted in this instance. Furthermore, the graph of coefficient size is telling. Most predictors are zero at all λ values, though there are four that are selected to impact in the model before the penalty is large enough to reduce all coefficients.

Ridge Regression's best (minimum) lambda value is 0.0123297, suggesting that a small shrinkage penalty may be needed. Within one standard deviation it becomes more sizable, but not greatly so. Just as with the linear model, the test rMSE is extremely low:0.0159625. Accuracy is pronounced here because Ridge Regression can favor the most impact variables which are, of course, population and misdemeanor caseloads. By amplifying the effects of these two variables –which directly influence a county's proportion of 10.101.060 funds– our models become more accurate. But do the predictions help improve the model, as we explored above?

4.4 LASSO

While the Ridge Regression model was successful in reducing our rMSE and helped remedy some of the violated assumptions in our initial linear model, we also wanted to try the LASSO model, which performs feature selection, and could help us further narrow in on helpful variables. A downside to this approach, though, is that population and misdemeanor cases will likely stay in the model –as they are the two variables used to create the budget proportion– while other variables that may be more helpful to holistically expand the equity and robustness of budgets by county will be removed (James et al. 2013).



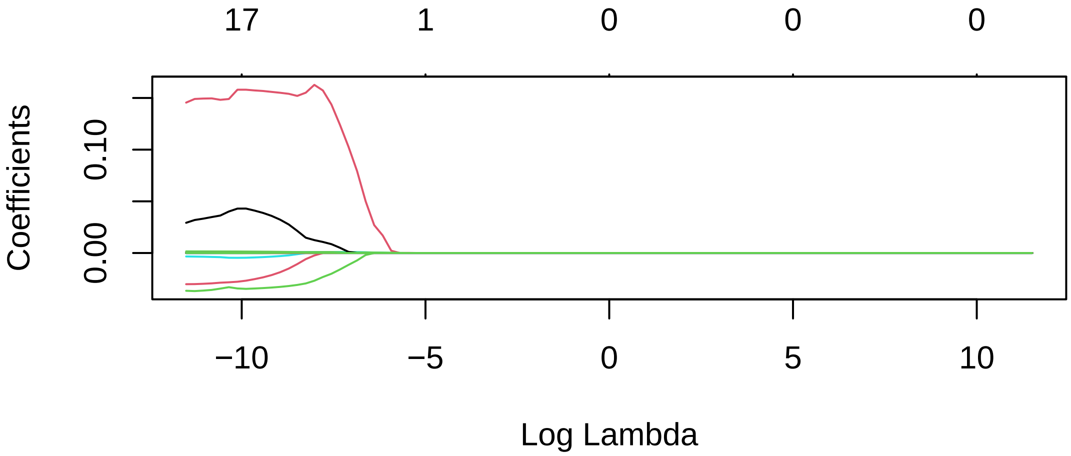


Table 5: Comparison of LASSO Tuning Parameters

Best Lambda Value	Best Lambda within 1 SE
0.0010476	0.0135305

Looking at the first LASSO plot, we observe that having a model between 5 and 1 predictor values is the optimal space for this LASSO model. It is also significant to note that with these small proportional values, the mean squared error is tiny—though relative to one another, it has extremely large standard error, as displayed graphically by the bars. A similar story is shown in the coefficient size plot, which shows that as the tuning parameter increases, the number of predictors quickly increases to one, then rapidly drops to zero. The minimum lambda value and the minimum value within 1 SE is shown in the above table. Generally, this tuning value is large relative to the magnitude of our response variable. Using our minimum value of λ in a test regression yields an rMSE of 0.0068809.

Table 6: Best LASSO Model Coefficient

Variable	Coefficient
Intercept	-0.0286993
Population_2019	0.0000002
Percent_Individuals_below_poverty_level_2015_2019	0.0500727

Variable	Coefficient
Median_household_income_2015_2019	0.0000000
Adult_felony_cases_filed_2019	0.0000000
Adult_felony_cases_per_1000	0.0006815
Adult_felongy_cases_assigned_to_counsel	0.0000000
Misdemeanor_cases_filed_2019	-0.0000036
Misdemeanor_cases_assigned_to_counsel	0.0000000
Misdemeanor_cases_per_1000	0.0005871
Juvenile_offender_cases_filed_2019	0.0000000
Juvenile_offender_cases_per_1000	0.0000894
Juvenile_offender_cases_assigned_to_counsel	0.0000128
Adult_Felony_Appointment_Rates	0.0000000
Adult_Misdemeanor_Appointment_Rates	0.0000000
Juvenile_Offender_Appointments	-0.0019388
Median_Home_Price_Q4_2018	0.0000000
Median_Home_Price_Q1_2019	0.0000000
Median_Home_Price_Q2_2019	0.0000000
Median_Home_Price_Q3_2019	0.0000000
Median_Home_Price_Q4_2019	0.0000000
Median_Home_Price_Percent_Change_by_Year	0.0000000

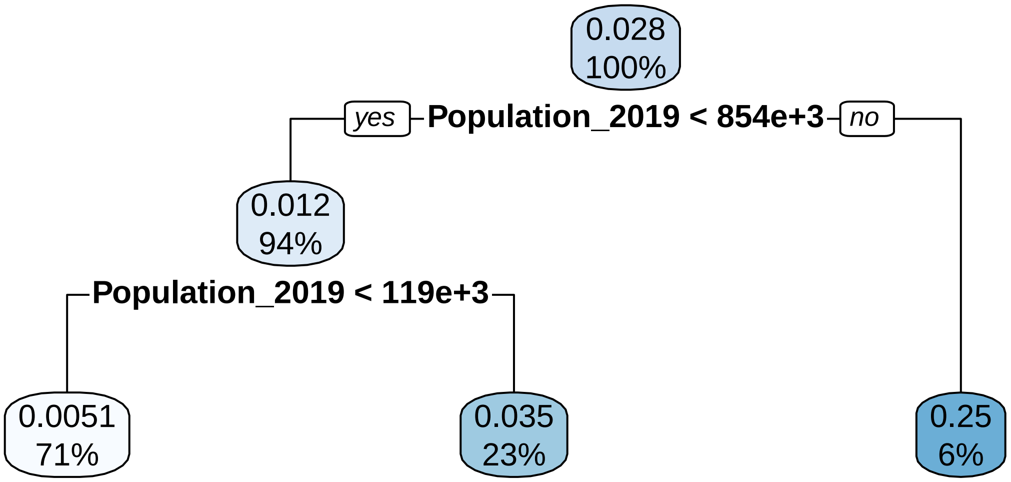
Table 7: Table 1: Non-Zero LASSO Model Coefficients

Variable	Coefficient
Intercept	-0.0286993
Population_2019	0.0000002
Percent_Individuals_below_poverty_level_2015_2019	0.0500727
Adult_felony_cases_per_1000	0.0006815
Misdemeanor_cases_filed_2019	-0.0000036
Misdemeanor_cases_per_1000	0.0005871
Juvenile_offender_cases_per_1000	0.0000894
Juvenile_offender_cases_assigned_to_counsel	0.0000128
Juvenile_Offender_Appointments	-0.0019388
Median_Home_Price_Q2_2019	0.0000000

Above, we look at the specific output and which variables were included in our LASSO model. The model, notably, had 9 non-zero predictors, which were: 2019 population, individuals below the poverty line, adult felony cases per 1000 residents, misdemeanor cases filed in 2019, misdemeanor cases per 1000 residents, juvenile offender cases per 1000 residents, juvenile offender cases assigned to counsel, juvenile offender appointments, and median home price in quarter 2 of 2019. This insight is particularly helpful for our purposes, as we do see our variables of misdemeanor cases and population as important. It, however, suggests that there are other uncorrelated variables that could enhance how Washington state allocates budgeting to its counties. These impact how the model makes predictions and, accordingly, supports the argument of including more variables in the OPD funding model.

4.5 Pruned Tree

In these last models, we look at various forms of decision trees, starting with the simplest: a pruned tree. The pruned tree model, shown below, has the most interpretable results to a non-statistical audience, though the random forests and bagging have higher accuracy and are built on a more interpretable logic from which we can attempt to extrapolate to a non-statistical audience.



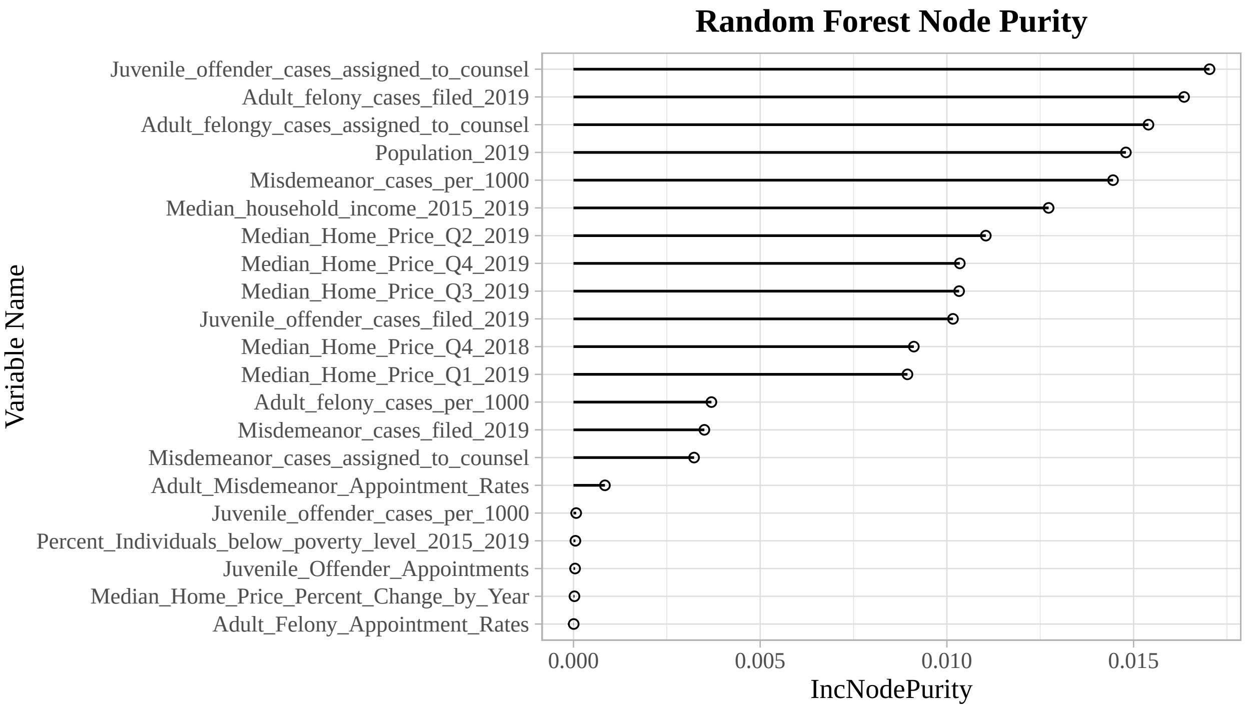
Visualized above is our pruned decision tree, with two branches (spots of yes/no questions), with three potential predictions as a result. Most counties were assigned the value of 0.0051 (71%), with the other two nodes grabbing the last 23% and 6%, respectively. A tree with three splits is favored over one with fewer splits (though all are within one standard deviation of each other). Given our small sample size, it is difficult to have more than three splits in our tree. Calculating the test rMSE for our best pruned tree yields 0.0114375, meaning it is only currently outperformed by the LASSO model. Only considering two nodes, this model is at a disadvantage compared to the others as it leaves out the other variable (municipal cases) used to calculate proportion of the state's public defense budget should be allocated to that county.

4.6 Random Forest

Next, to increase our model's diversity in predictions, we created a random forest model. For our purposes, this means creating an *ensemble model* with multiple trees of varying size being considered and factored in, on various combinations of predictors.

Table 8: Variables Split on in at Least One Random Forest Iteration

	Split Variables
1	Adult_felongy_cases_assigned_to_counsel
2	Juvenile_offender_cases_filed_2019
4	Juvenile_offender_cases_assigned_to_counsel
6	Misdemeanor_cases_filed_2019
7	Population_2019
8	Misdemeanor_cases_per_1000
11	Median_Home_Price_Q3_2019
17	Adult_felony_cases_filed_2019



As shown in the table above, there are eight variables that split the data in various tree models. These line up with the node purity, or *importance* values assigned to each predictor. As expected, misdemeanor cases per 1,000 was deemed the most important, as were other caseload information, population, median household income, home prices, and so on. Notably, poverty was not an important variable in this model

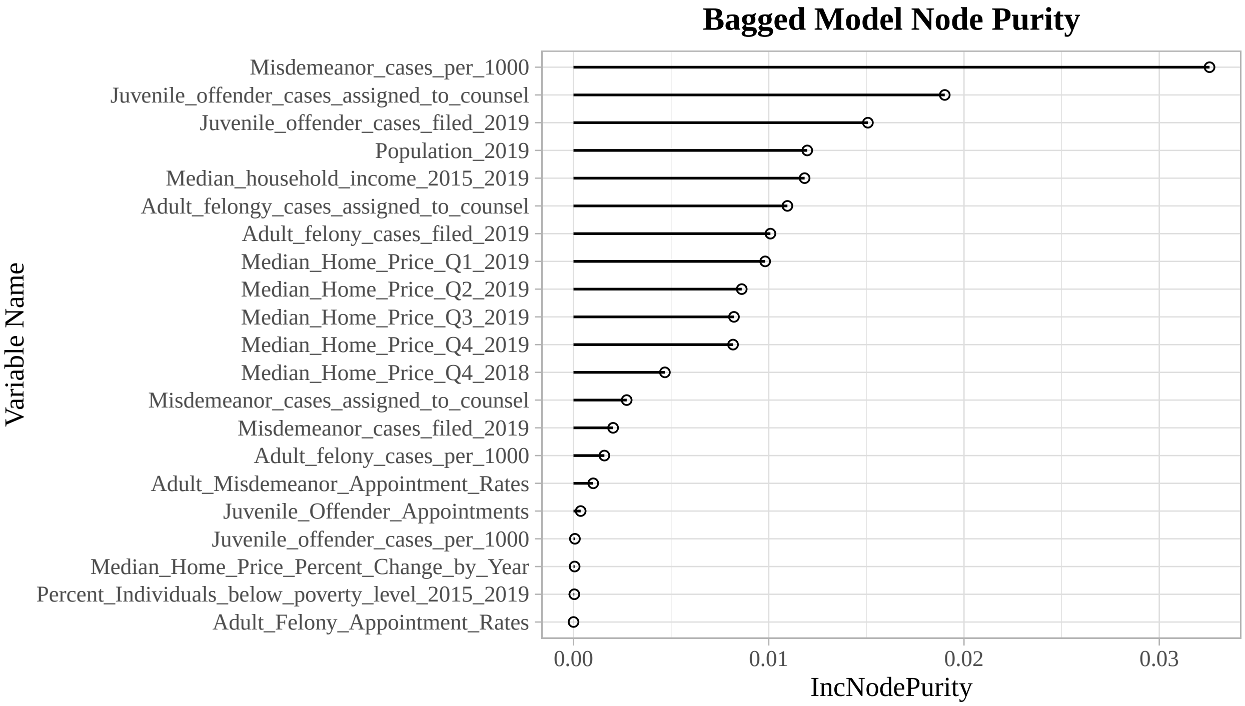
Though appearing promising, calculating the test rMSE on this random forest model, 0.0296918, shows that this is, in fact, our worst performing model –worse even than the full linear model.

4.7 Bagging

Lastly, we look at a bagged model. Bagging takes different subsets of the data, fits a tree model on that, then combines all of the resulting trees into a larger ensemble model. For our purposes, this could be quite useful if it helps consider a wider variety of predictors when making splits.

Table 9: Variables Split on in Bagged Model

	Split Variables
1	Adult_Misdemeanor_Appointment_Rates
3	Adult_felongy_cases_assigned_to_counsel
4	Population_2019
5	Juvenile_offender_cases_filed_2019
6	Misdemeanor_cases_assigned_to_counsel
7	Adult_felony_cases_per_1000
15	Juvenile_Offender_Appointments



Our results, however, do not differ much from the random forest model. Similar variables emerge as important and others as less important, with minimal shuffling. It is likely we will run into similar issues with rMSE being large and not considering the importance of poverty for budget allocation.

While lower than the random forest RMSE, the bagged RMSE, 0.0209888 is still quite large in comparison to the pruned tree and penalized regression models (James et al. 2013).

4.8 Model Comparisons and Accuracy

Table 10: Model rMSE Comparison

Model	rMSE
LASSO	0.0068809
Pruned Tree	0.0114375
Forward Selection	0.0129758
Ridge Regression	0.0159625
Backward Selection	0.0186755
Bagged Tree	0.0209888
Full Linear	0.0264804
Random Forest	0.0296918

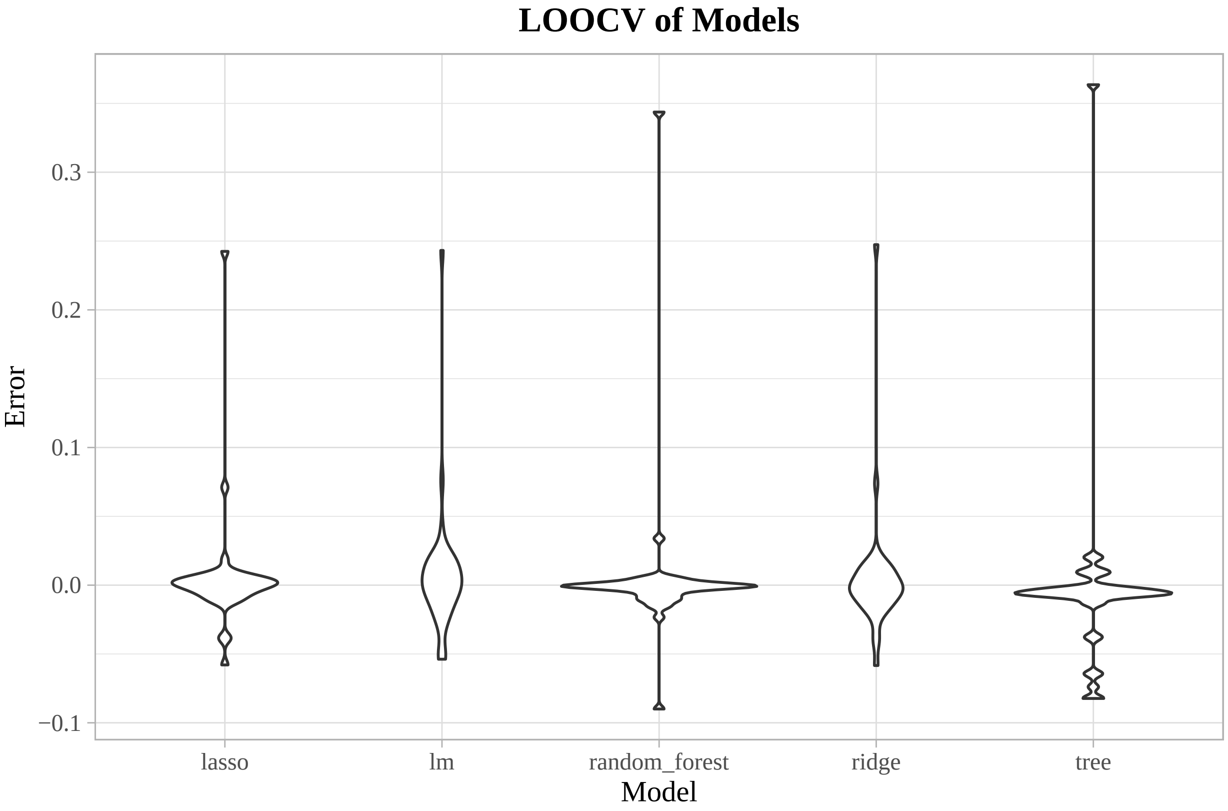
Having walked through all of the proposed models, we are able to compare them through their rMSE values. It is significant to note here that the linear models did relatively well in comparison to some of the more opaque and sophisticated methods. Of course, however, the best overall model, with an incredibly small rMSE of 0.0068809, was the LASSO model. The combination of penalized regression and feature selection far outperformed any other types of models we created (James et al. 2013). Additionally, the LASSO model is supremely useful in reporting back to the OPD what data they should collect from counties as, due to limited resources, knowing what to ask with their limited collection ability is key. While not as directly transparent as, say, a pruned tree or full linear model, we believe that the LASSO model is relatively easy to explain premise and should be interpenetrate to a non-statistical audience following very minor outside research. If the interperatablility of the LASSO model is a further concern, then the linear model resulting from forward selection, with an rMSE of 0.0129758, might be optimal.

The core impotence for creating this report was to make a model, surpassing the OPD's, that successfully incorporated socioeconomic factors. Hopefully, as seen through some of our models, this aim has been achieved. We were not able to comment on every socioeconomic variable found in our models, but an in depth look at capturing poverty rates can be found in Appendix 5. In Appendix 5, crucially, we show that our LASSO model is better able to encompass and equitable disburse funds on the basis of poverty rates.

It is notable, too, that the pruned tree model, using just one variable and two splits outperformed ridge regression, which included every variables. However, when looking at the graphs of poverty for both the ridge and tree model (in Appendix 5), it was clear that ridge regression is better able to take into account the ramifications of a county's poverty level on how much funding they should receive from the state. Given our intentions to build a more equitable model, I would say that ridge and LASSO clearly emerge as the winners for this method, while still getting proportions that are not too terribly far off from what they used to be. A dramatic change in state funding could drastically destabilize county's public defense systems, and would also require more political capital for the state to pass. Tweaking the model to produce slightly different results, but with a bent of equitable distribution by county when considering more variables, would be a possible with our models. Alternatively, if the state is unable to collect all of the above data about counties one year (say, for instance, COVID-19 or other budgetary constraints got in the way), using a tree model and just a county's population, we can efficiently calculate how much funding each county should receive. Therefore, no matter the data OPD has and their interests in revising the funding model, one of our models should be an appropriate fit.

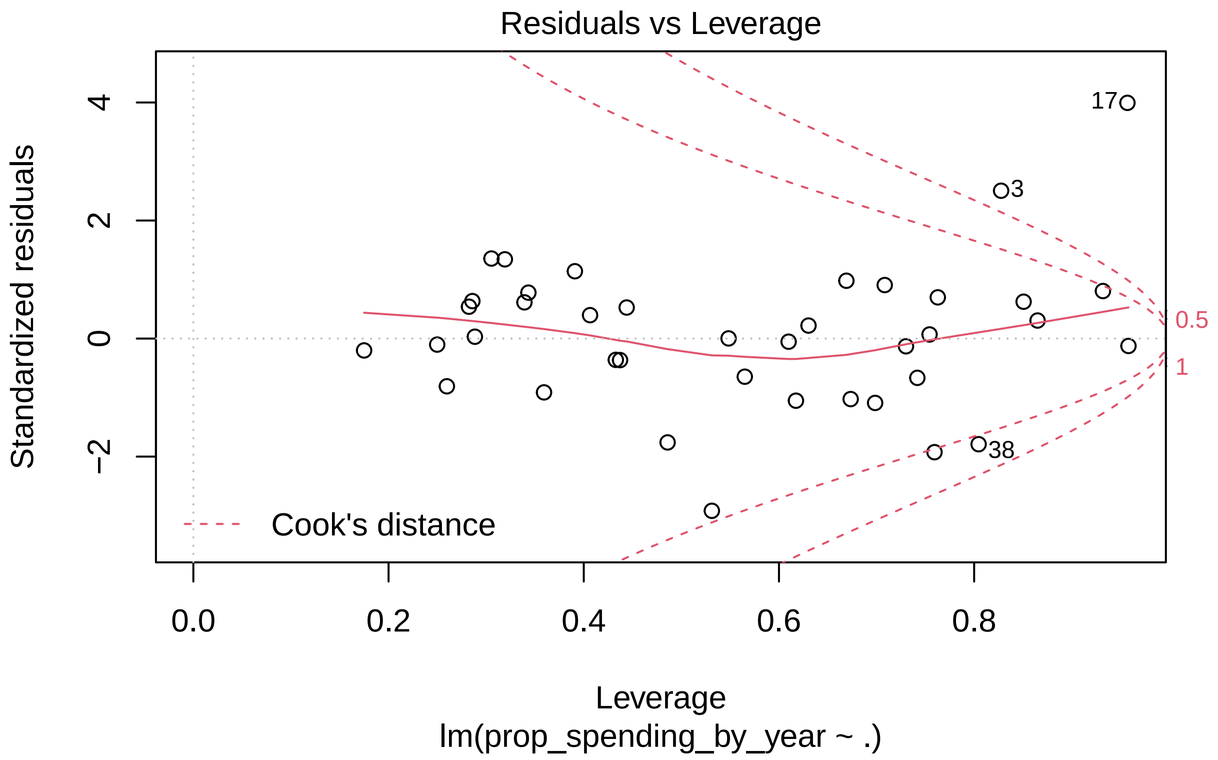
5 Discussion

5.1 Modeling with Sparse Data: Looking at our Models with LOOCV



Outling point is King County

The previous analysis outlines the issues that specific models have with the general data. This section introduces another confounding issue for consideration, the impact of leverage points in sparse data modeling. The plot above is a violin graph of several models and the residuals they produce when one county is left out of the model. The largest outlying point is King County, if it is not present in the train dataset then it overstates the error of the model in the test error.



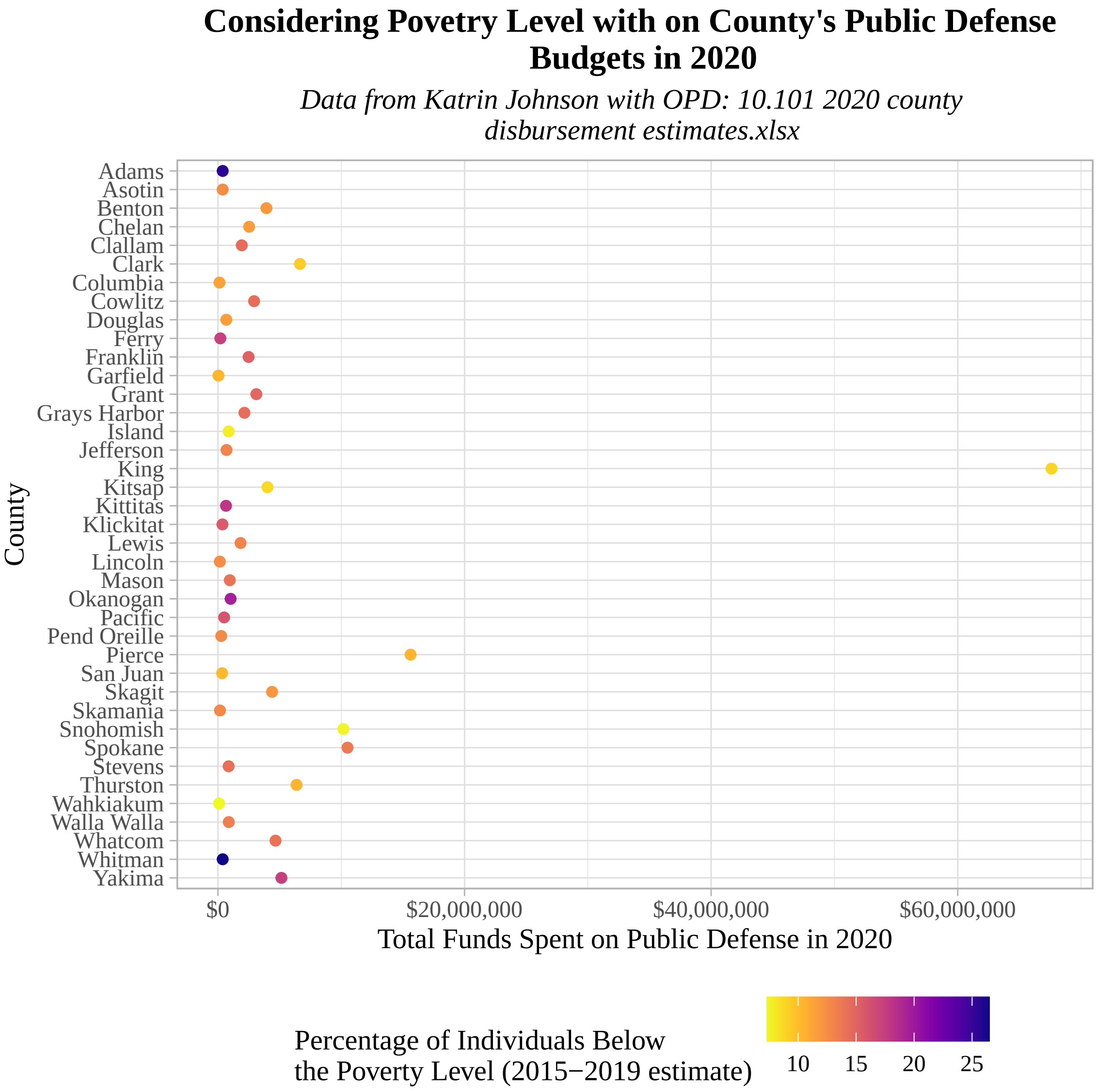
When present in the train dataset King County leverages the model. This can be seen in the full linear model residual vs leverage graph. Of note are the handful of points that have a cooks distance greater than one. Once again the furthest point is King County. We can fix some of the issues above with further data wrangling normalizing data, transforming data, removing extraneous variable, etc. But these can act as Texas sharpshooter fallacy (painting targets on outputted values).

5.2 Final Discussion

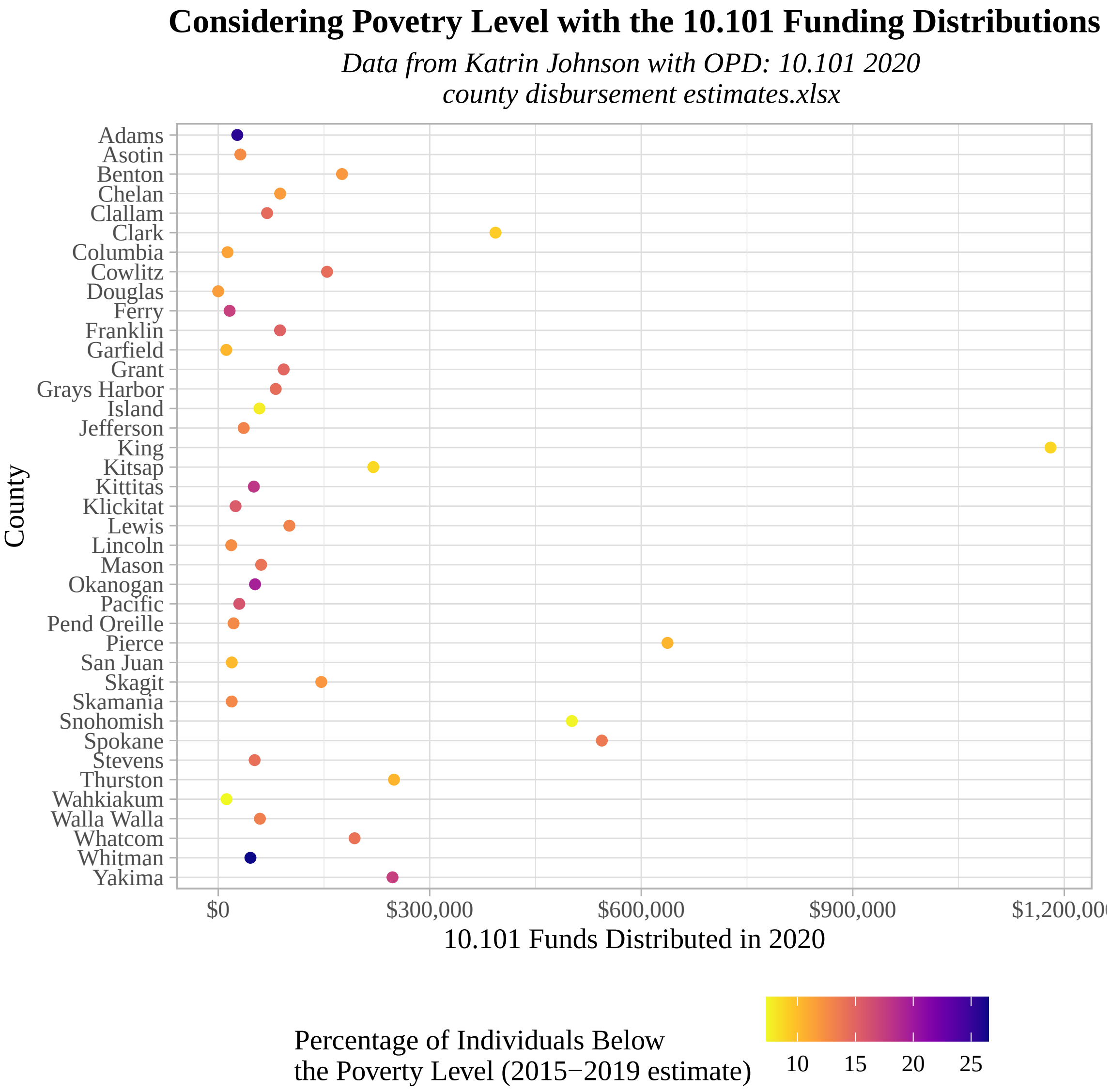
Our models were created and trained on limited data, and it would be inappropriate to claim that we have created here the best possible models for disbursing OPD funds in Washington. We can, however, assert that we created improved, statistically grounded models with the available resources that successfully increased model transparency and interpretability, and accounted for a greater variety of socioeconomic factors. A clear avenue to improve our work, naturally, is to collect and analyze additional data from a wider range of sources and years. Another more difficult, but potentially transformative update to our work would be to secure outside funding distribution models with which to compare to ours. We hope, in closing, that our work can be used –in some small part– to help update and modernize the criminal justice system to best support the needs of all people.

6 Appendices

6.1 Appendix 1 Considering Poverty Levels


Considering Povetry Level with on County's Public Defense Budgets in 2020

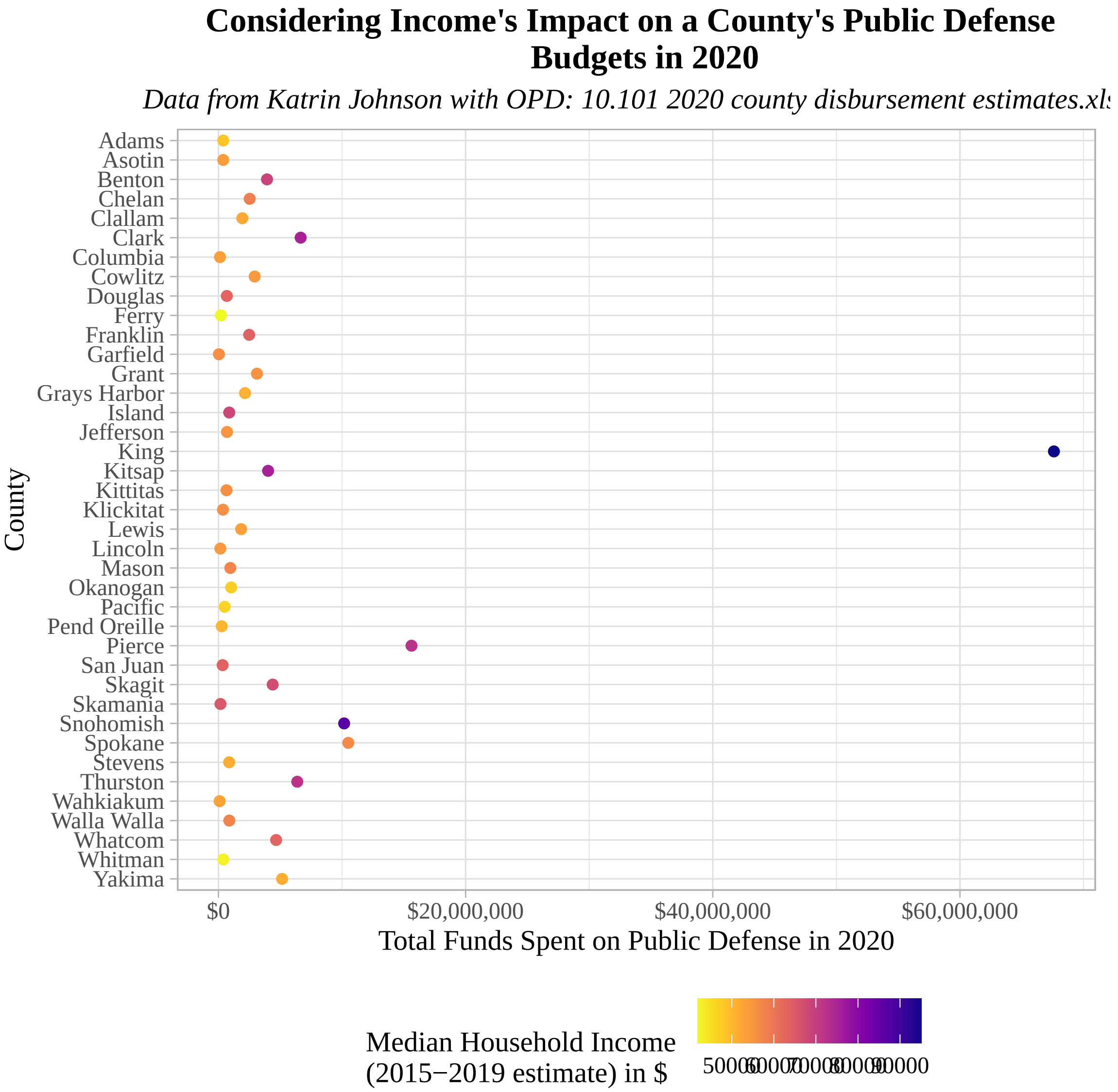
Data from Katrin Johnson with OPD: 10.101 2020 county disbursement estimates.xlsx



Considering Povetry Level with the 10.101 Funding Distributions

Data from Katrin Johnson with OPD: 10.101 2020 county disbursement estimates.xlsx

6.2 Appendix 2 Considering Income

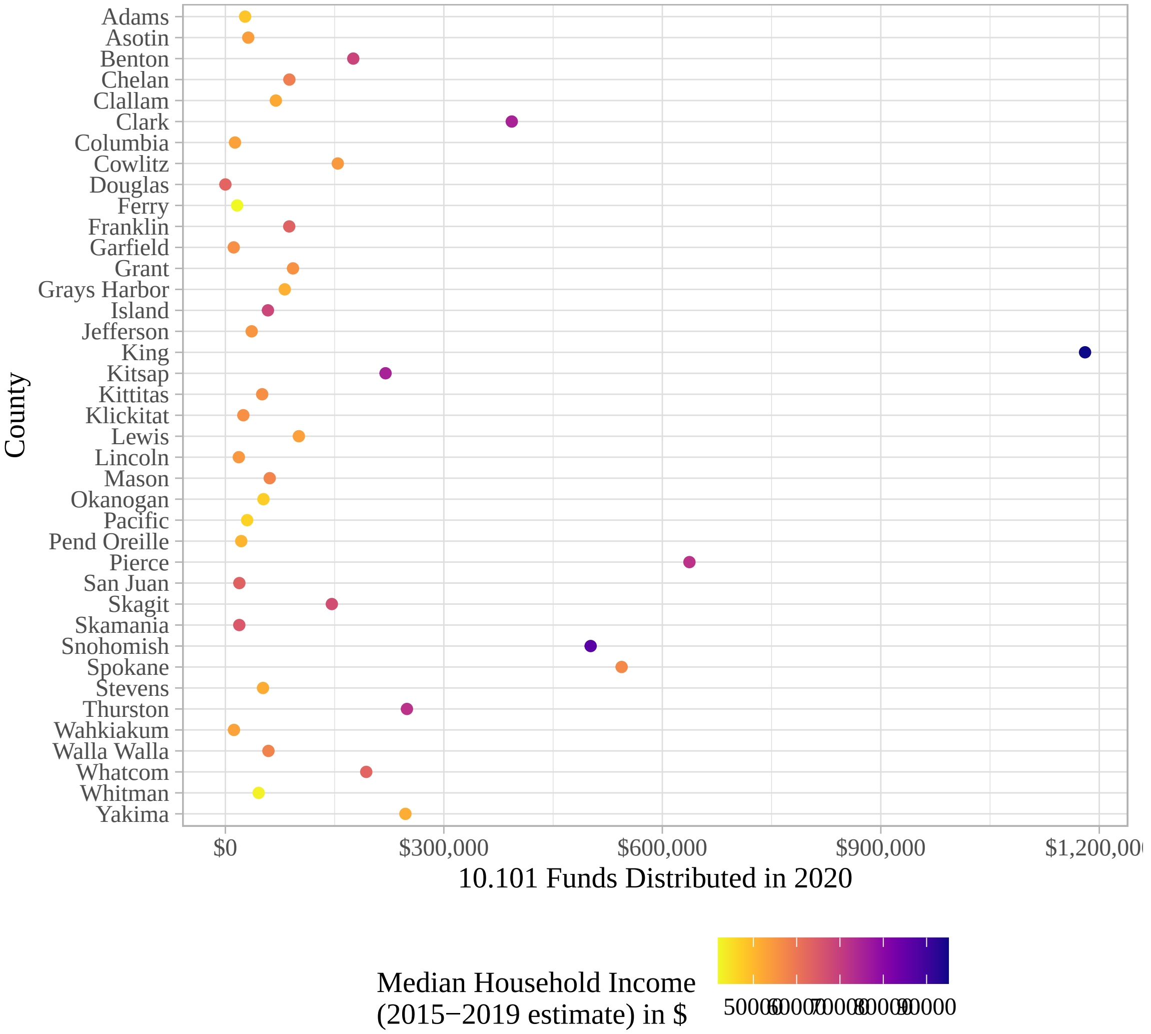


Considering Income's Impact on a County's Public Defense Budgets in 2020

Data from Katrin Johnson with OPD: 10.101 2020 county disbursement estimates.xls

Considering Income's Impact on the 10.101 Funding Distributions

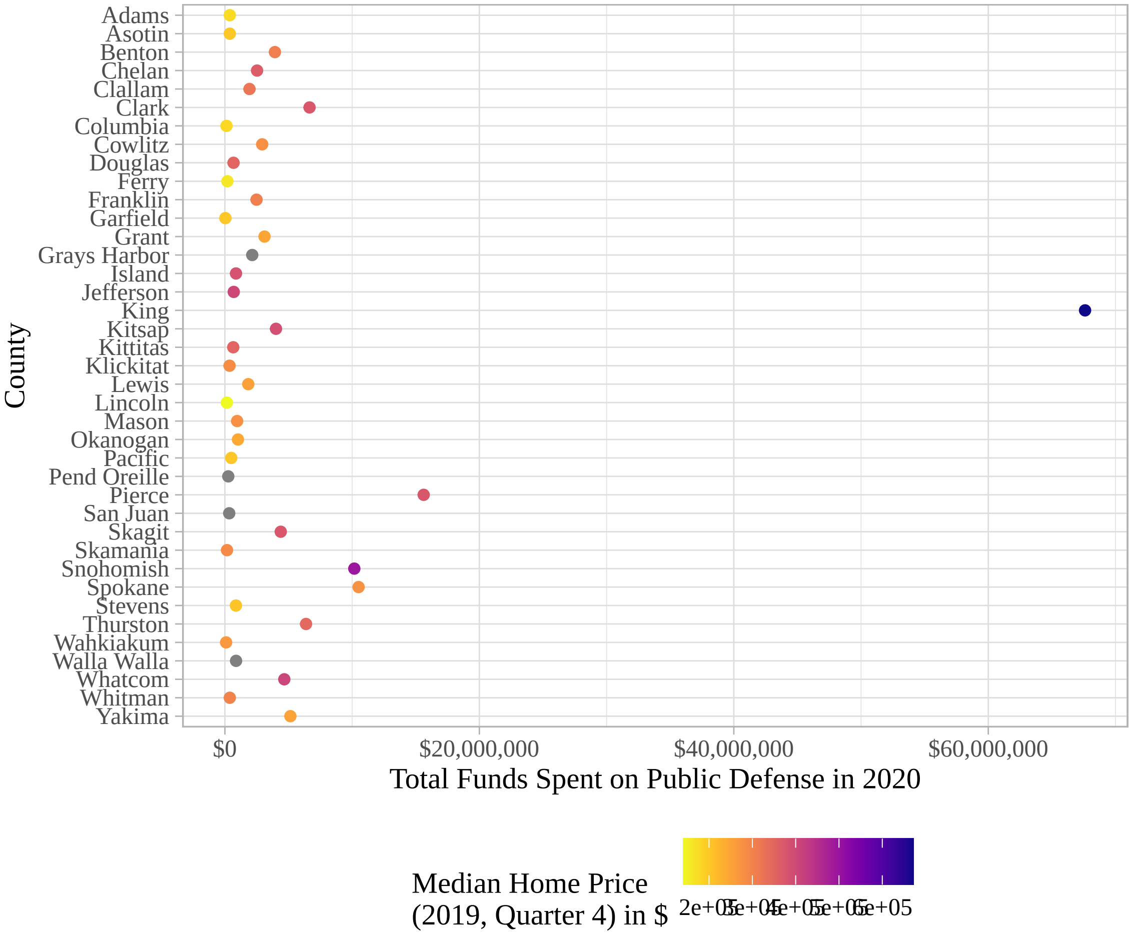
Data from Katrin Johnson with OPD: 10.101 2020 county disbursement estimates.xl



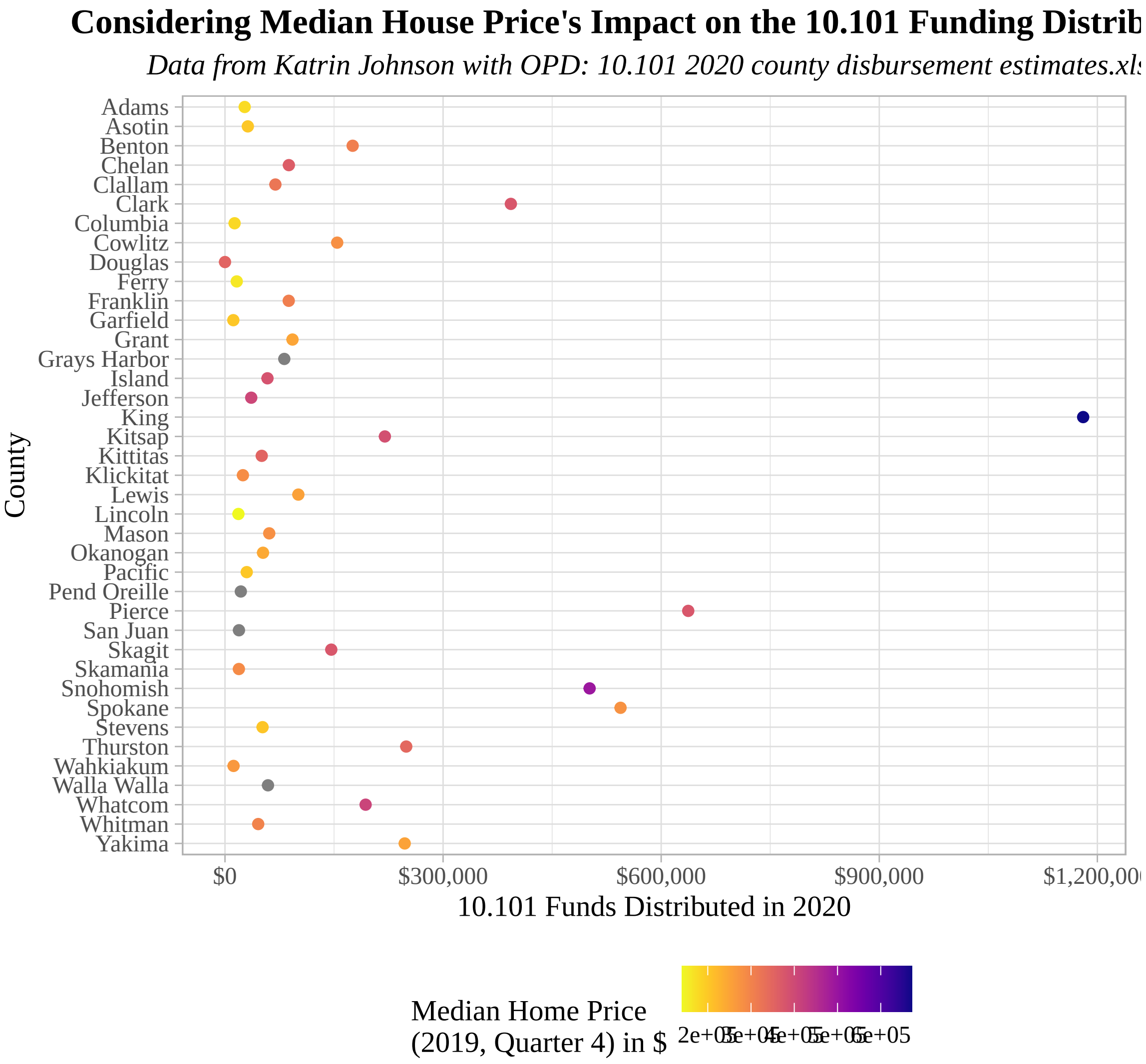
6.3 Appendix 3 Considering Median Home Prices

Considering Median House Price's Impact on a County's Public Defe
Budgets in 2020

Data from Katrin Johnson with OPD: 10.101 2020 county disbursement estimates.xl



** Gray points indicate an NA value.*



Considering Median House Price's Impact on the 10.101 Funding Distrib

Data from Katrin Johnson with OPD: 10.101 2020 county disbursement estimates.xl

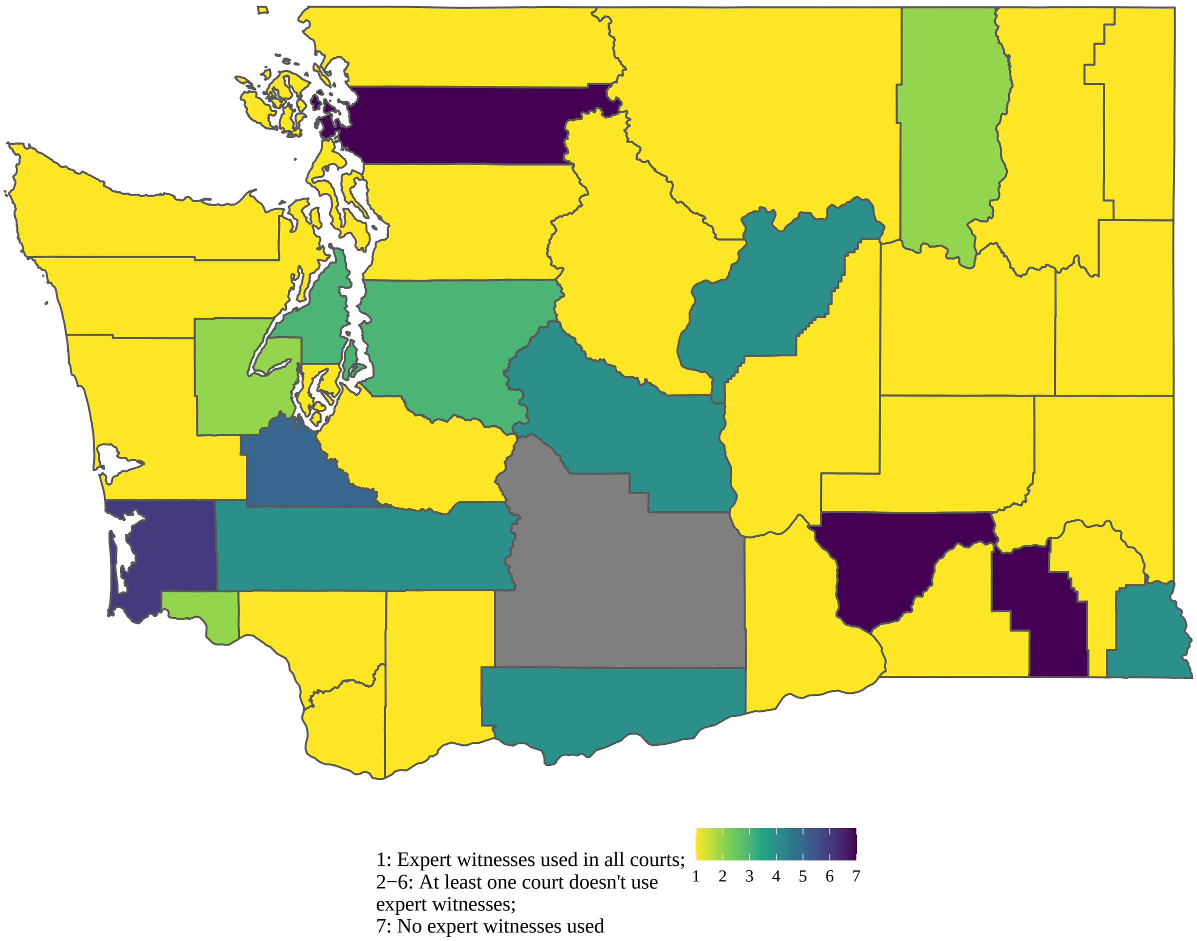
** Gray points indicate an NA value.*

6.4 Appendix 4 Additional Data Visualizations and Considerations

6.4.1 County’s Usage of Expert Witnesses

Expert Witness Resources by County (2018)

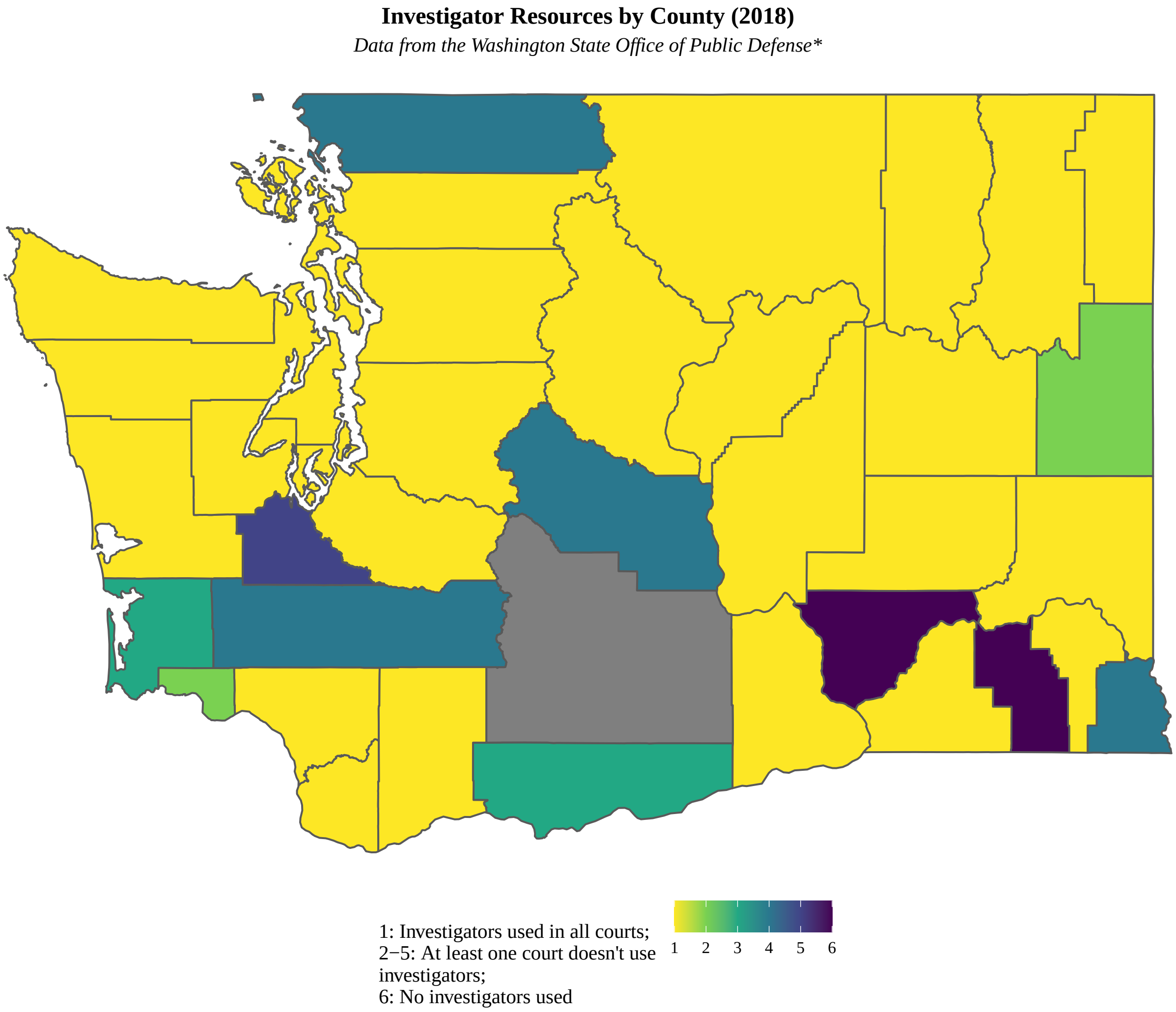
*Data from the Washington State Office of Public Defense**


1: Expert witnesses used in all courts;
2−6: At least one court doesn't use expert witnesses;
7: No expert witnesses used
1
2
3
4
5
6
7

**Douglas County did not report their data to OPD in 2018.*

Courts Expert Witnessed Used In	Legend Value
Yes: all courts	1
No: District; Yes: Superior, Juvenile	2
No: Juvenile; Yes: Superior, District	3
No: District, Juvenile; Yes: Superior	4
No: Superior, Juvenile; Unanswered: District	5
No: Superior, Juvenile; Yes: District	6
No: all courts	7
NA	NA

6.4.2 County's Usage of Investigators



Investigator Resources by County (2018)

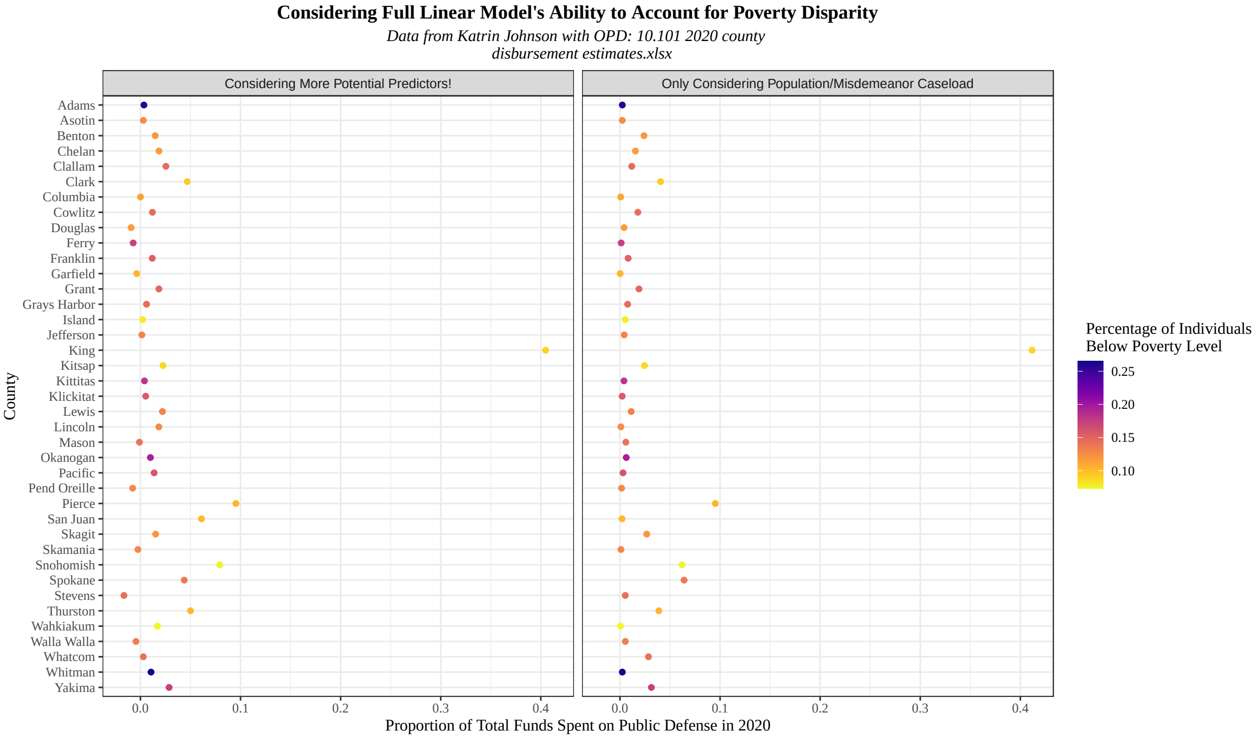
*Data from the Washington State Office of Public Defense**

1: Investigators used in all courts;
2−5: At least one court doesn't use investigators;
6: No investigators used
1
2
3
4
5
6

**Douglas County did not report their data in OPD in 2018.*

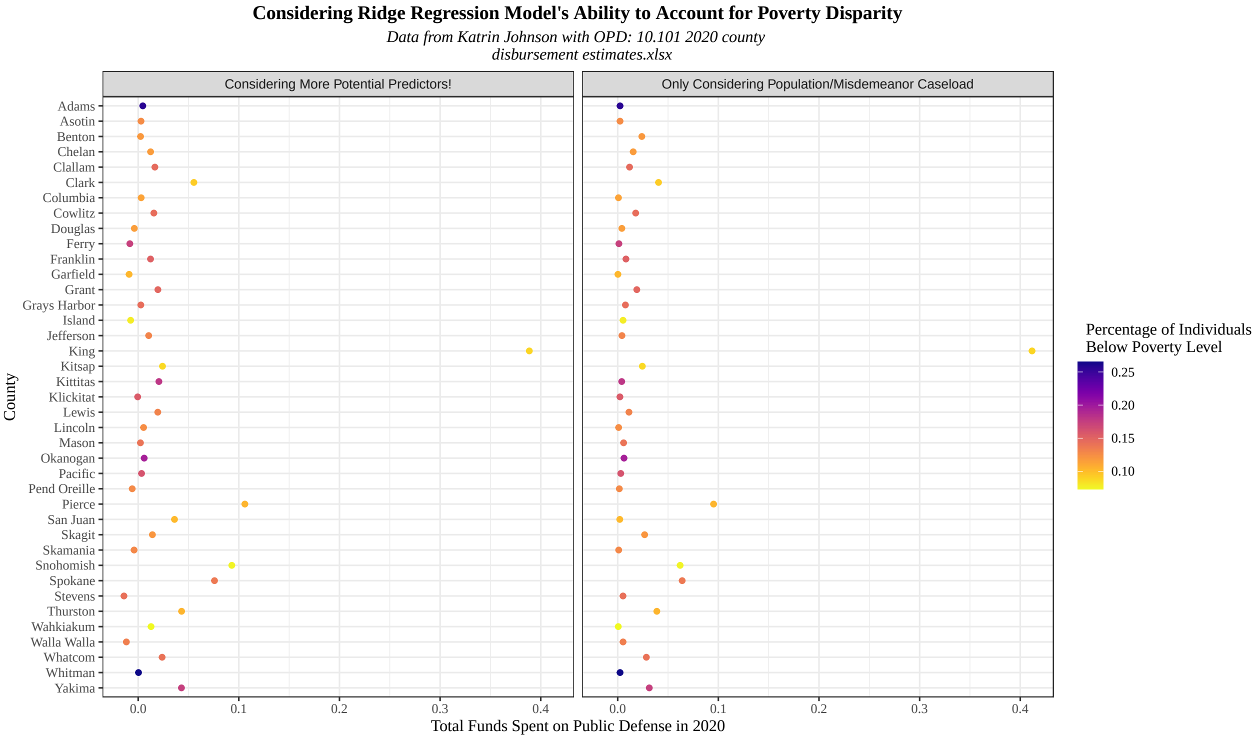
Courts Investigators Used In	Legend Value
Yes: all courts	1
No: District; Yes: Superior, Juvenile	2
No: Juvenile; Yes: District, Superior	3
No: District, Juvenile; Yes: Superior	4
No: Superior, Juvenile; Unanswered: District	5
No: all courts	6
NA	NA

6.5 Appendix 5 Comparing Models Selecting for Poverty Rates



Here, to evaluate the strength of the new linear model when taking into account new, important predictors for budget consideration (such as poverty), we look at the percentage of poverty by county, and the predicted and actual proportion of funds for public defense. This variable is especially significant here because employees from the Washington Office of Public Defense specifically expressed worry about this factor not being taken into account—as counties with more poverty tend to have less tax money coming in, and have higher budgeting constraints due to need for funds in other areas. Therefore, this report considers how well our new models improves upon considering the importance of poverty.

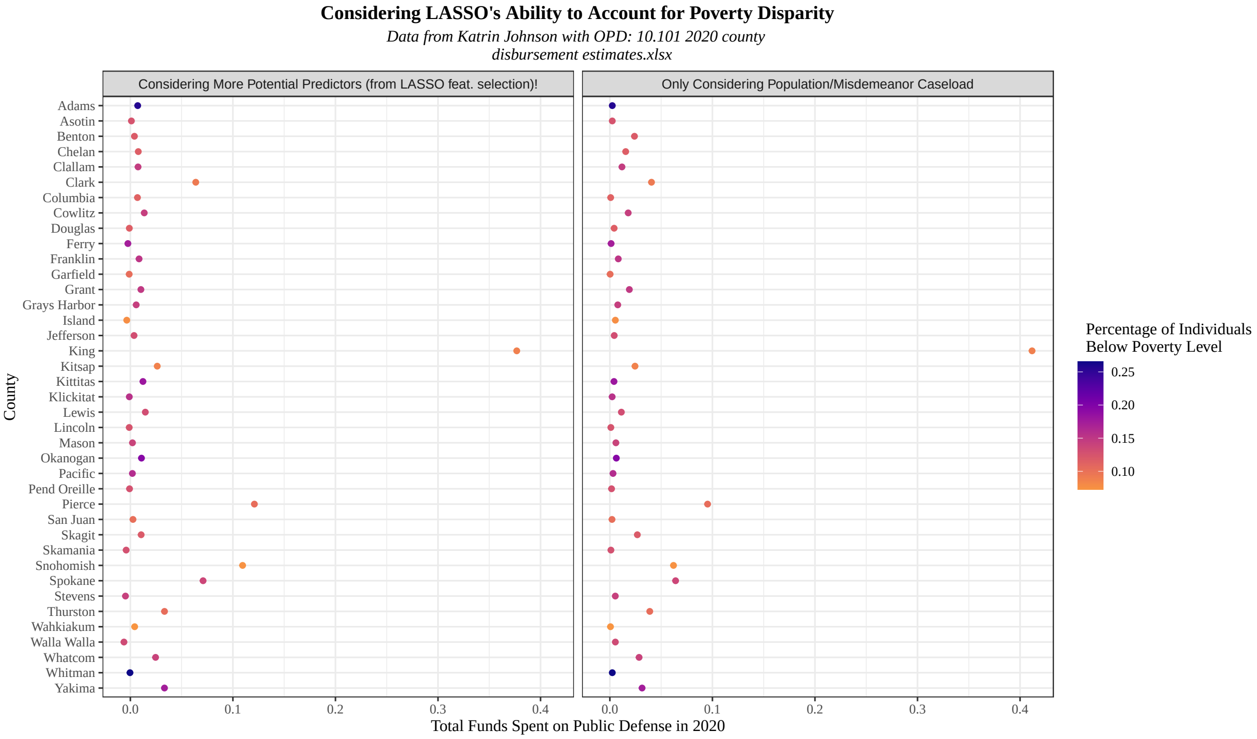
The full linear model here does a good job accounting for new variables, such as percentage of individuals below poverty level. For instance, compare Whitman and Walla Walla. Here, Whitman has a higher percentage of its population under the poverty level. Under the initial model's conditions (the right-hand panel), it receives a smaller portion of the budget than Yakima, a county with a lower poverty level. However, in the new model, these new variables help *predict proportions of the WA state budgets*. It succeeds in narrowing the gap between Whitman and Yakima (along with some other places, including Clallum and Clark, Snohomish and Spokane, etc.), often giving the advantage to the city with fewer resources. Thus, while the full linear model violates many assumptions (despite being exceedingly correct due to us possessing the exact variables it was built upon), it does offer some important insights to our model about other predictors to add.



Considering Ridge Regression Model's Ability to Account for Poverty Disparity

Data from Katrin Johnson with OPD: 10.101 2020 county disbursement estimates.xlsx

Here, we have a chance to see how the Ridge Regression model takes into account poverty levels and funding. On the right we see actual predictions, which result from only looking at population and caseloads. Unlike with the linear model, there are no drastic differences. From Franklin to Jefferson counties, there does tend to be a trend toward slightly more funding for the counties with more poverty, though this is not a systematic, universal change.



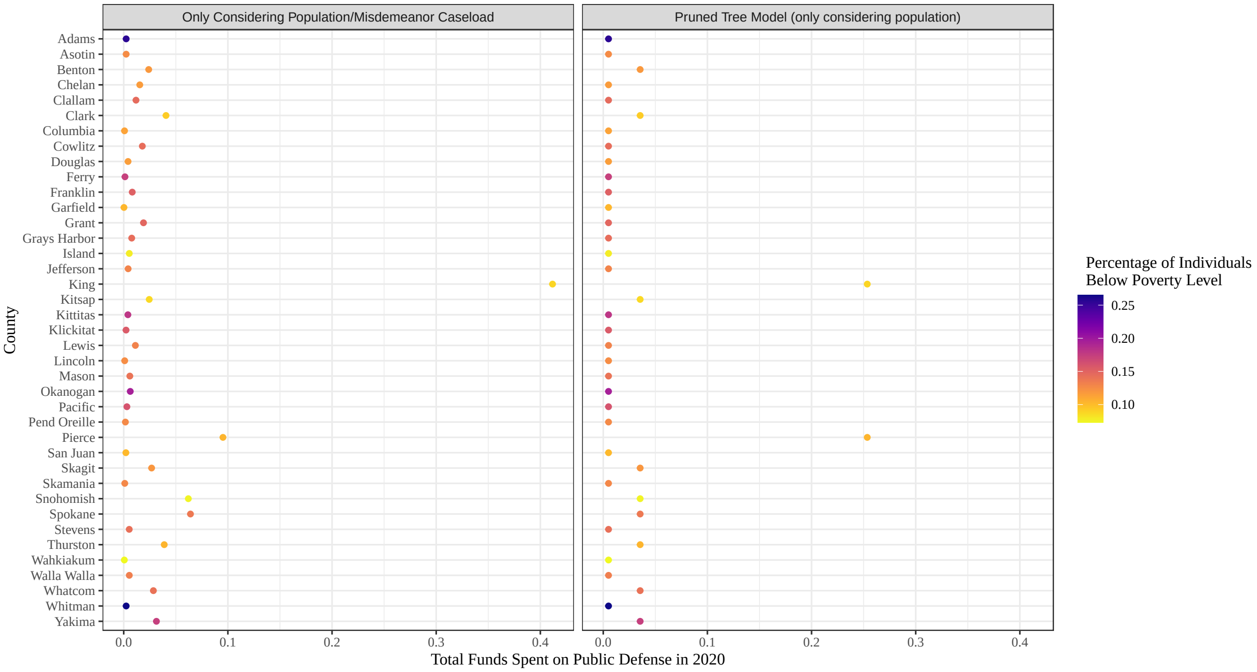
Considering LASSO's Ability to Account for Poverty Disparity

Data from Katrin Johnson with OPD: 10.101 2020 county disbursement estimates.xlsx

When considering more variables, the model on the left does in fact seem to more fairly distribute funds to counties with higher poverty levels. Take Oknogan and Pacific, or Kittitas and Klickitat. Both Okanogan and Kittitas have more poverty, though previously received approximately the same proportion of the public defense budget. In the LASSO model, however, we see more equity in the distribution as Okanogan and Kittitas receive more funds than their less impoverished counterparts.

Considering the Pruned Tree's Ability to Account for Poverty Disparity

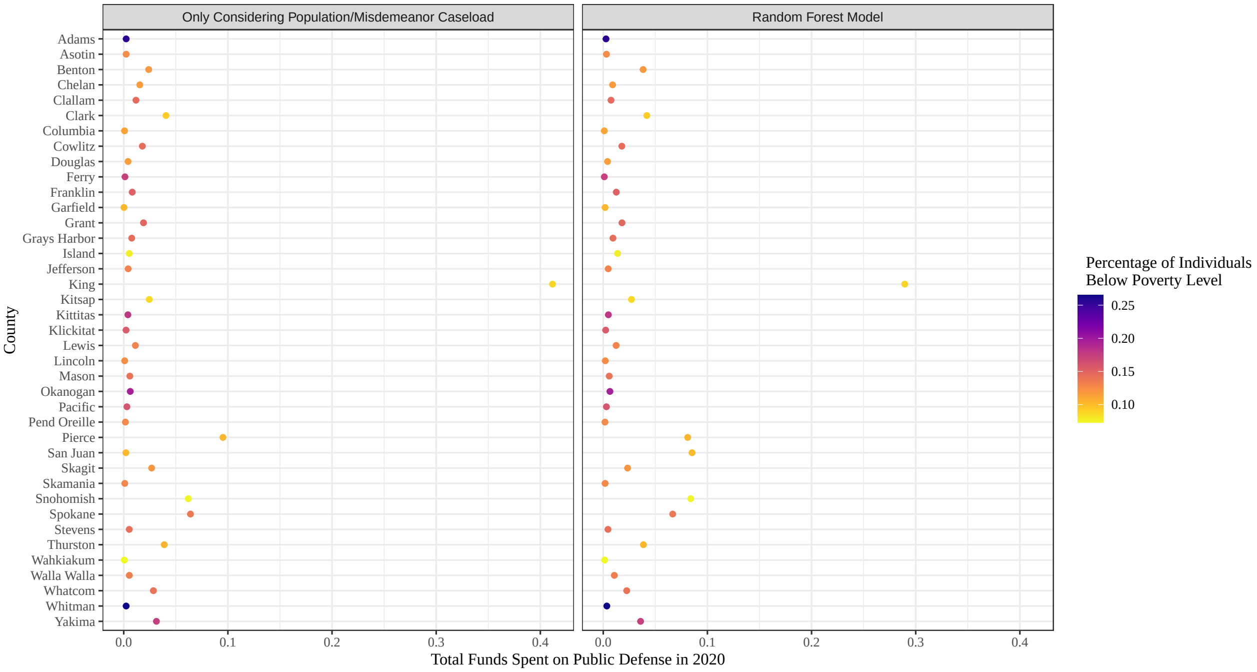
Data from Katrin Johnson with OPD: 10.101 2020 county disbursement estimates.xlsx



As our tree model only has three predictions, the right-hand facet of the above graph reflects just three funding options. Notably, when considering poverty, these predictions do *a very poor job* of ensuring counties with higher poverty levels receive more funds than other counties. While there are, of course, a myriad of other factors that may determine way counties with higher poverty shouldn't receive funds, the tree model casts many counties with *vastly* different poverty levels to receive the same proportion of the budget. This is a large concern.

Considering the Random Forest Model's Ability to Account for Poverty Disparity

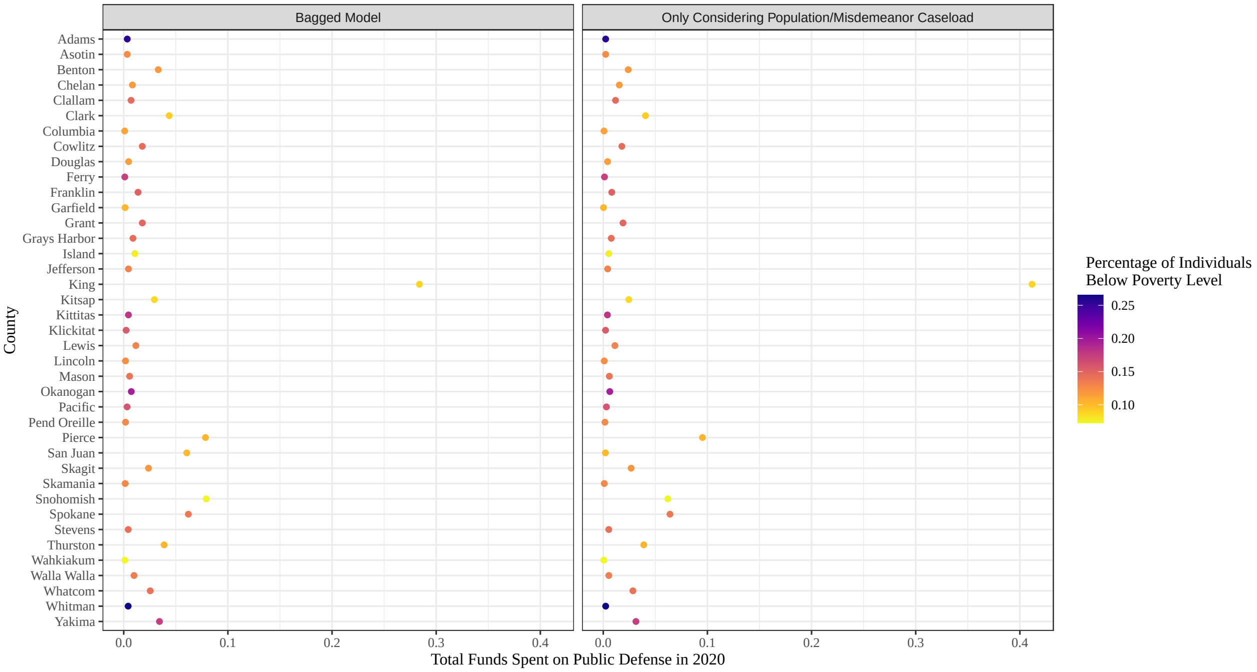
Data from Katrin Johnson with OPD: 10.101 2020 county disbursement estimates.xlsx



As shown above, the random forest plot does not do much in helping counties with more poverty receive more funding. This is likely a result of how poverty was ranked in terms of node purity/importance, meaning that it does not have an influence on the model. Considering its importance for budgeting, this is a serious limitation of the random forest model.

Considering the Bagged Model's Ability to Account for Poverty Disparity

Data from Katrin Johnson with OPD: 10.101 2020 county disbursement estimates.xlsx



Unfortunately, no overall trends emerge in the above graph where the bagged model improves upon the original model in terms of the percentage of poverty present in a county. While the predictions are more varied and thus slightly more accurate overall (likely a result of considering misdemeanor cases and population so heavily), they do not tend to help counties with more poverty generally receive more funds.

7 Code Appendix

7.1 Libraries

```r
library(readr)
library(tidyverse)
library(gt)
library(paletteer)
library(tidycensus)
library(viridis)
library(sf)
library(leaflet)
library(tidytext)
library(lsr)
library(lubridate)
library(lemon)
library(knitr)
library(jtools)
library(leaps)
library(caret)
library(glmnet)
library(rpart)
library(rpart.plot)
library(randomForest)
library(yardstick)
```

7.2 Data Loading and Wrangling

```r
set.seed(1234)

###############
# Loads CSV's #
###############

#Washington state county data

OPD_Funding_Over_Time <- read_csv("data/OPD 10.101 Funding Over Time.csv")

Annual_Public_Defense_Spending_by_County <- read_csv(
  "data/Annual Public Defense Spending by County.csv")

Caseload_Resources_2018 <- read_csv(
  "data/Caseload Resources and Capacity Measured in 2018.csv")

County_Statistics_2020 <- read_csv("data/County Statistics from 2020.csv")

County_Home_Prices <- read_csv("data/County Home Prices.csv")

Preds_10_101_2021 <- read_csv("data/calculating_10_101_2021.csv")

County_Home_Prices_revised <- read_csv("data/County_Home_Prices_revised.csv")
```

```
#######################
# Feature Engineering #
#######################

Caseload_Resources_2018 <- Caseload_Resources_2018 %>%
  mutate(Expert_Witnesses_Used_2018_Factor = case_when(
   Expert_Witnesses_Used_2018 ==
     "Yes: all courts" ~ 1,
   Expert_Witnesses_Used_2018 ==
     "No: District, Juvenile; Yes: Superior"
   ~ 4,
   Expert_Witnesses_Used_2018 ==
     "No: District; Yes: Superior, Juvenile"
   ~ 2,
   Expert_Witnesses_Used_2018 ==
     "No: Juvenile; Yes: Superior, District"
   ~ 3,
   Expert_Witnesses_Used_2018 ==
     "No: Superior, Juvenile; Unanswered: District" ~ 5,
   Expert_Witnesses_Used_2018 ==
     "No: Superior, Juvenile; Yes: District"
   ~ 6,
   Expert_Witnesses_Used_2018 ==
     "No: all courts" ~ 7)) %>%
  mutate(Investigators_Use_2018_Factor = case_when(
    Investigators_Use_2018 ==
   "Yes: all courts" ~ 1,
 Investigators_Use_2018 ==
   "No: District, Juvenile; Yes: Superior" ~ 4,
 Investigators_Use_2018 ==
   "No: District; Yes: Superior, Juvenile" ~ 2,
 Investigators_Use_2018 ==
   "No: Juvenile; Yes: District, Superior" ~ 3,
 Investigators_Use_2018 ==
   "No: Superior, Juvenile; Unanswered: District" ~ 5,
 Investigators_Use_2018 ==
   "No: all courts" ~ 6)) %>%
  rowid_to_column("id") %>%
  mutate(Municipal_Cases_Included = gsub("N/A","NA",as.character(Municipal_Cases_Included)),
         Amt_Spent_Experts = gsub("Included in invest amount","0",(Amt_Spent_Experts)),
         Total_PD_Spent_2018 <- gsub("Contract Amount   ","0",(Amt_Spent_Experts)))

Annual_Public_Defense_Spending_by_County <- Annual_Public_Defense_Spending_by_County %>%
  group_by(Year) %>%
  summarise(County = County,
            Total_Public_Defense_Budget,
    prop_spending_by_year = Total_Public_Defense_Budget/sum(Total_Public_Defense_Budget,
                                                            na.rm = T))

nodollar_crcmn2018 <- data.frame(lapply(Caseload_Resources_2018,
                                        gsub,
                                        pattern = "$",
                                        fixed = TRUE,
                                        replacement = ""))


County_Home_Prices_revised[, 2:6] <- lapply(County_Home_Prices_revised[, 2:6],
                                            parse_number)

###############
# Map Loading #
###############

api_key <- "24086c192dc2d5d94ae412de18add92dcac3f739"

wa_state_counties <- get_acs(state = "WA", geography = "county",
                             variables = "B25064_001", geometry = FALSE,
                             key = api_key)
wa_state_counties <- wa_state_counties %>%
  rowid_to_column("id")

pd_spend_county <- left_join(wa_state_counties,
                             Annual_Public_Defense_Spending_by_County %>%
                               drop_na(Total_Public_Defense_Budget) %>%
                               group_by(County) %>%
                               summarize(
                                 Total_Spent_Recorded = sum(Total_Public_Defense_Budget),
                                 Average_Yearly_Spending = mean(Total_Public_Defense_Budget)) %>%
                               rowid_to_column("id"), by = "id")

wa_state_counties_geom <- get_acs(state = "WA", geography = "county",
                                  variables = "B25064_001", geometry = TRUE,
                                  key = api_key)

pd_spend_geom <- merge(wa_state_counties_geom, pd_spend_county, by = "GEOID",
                       all.x = TRUE)

##################
# Robin's Themes #
##################

robins_ggplot_theme <- function(font = "Times") {
  theme_light() +
    theme(plot.title = element_text(hjust = 0.5, family = font, face = "bold"),
          plot.subtitle = element_text(hjust = 0.5, face = "italic", family = font),
          axis.title = element_text(family = font),
          axis.text = element_text(family = font),
          legend.title = element_text(family = font),
          legend.text = element_text(family = font),
          plot.caption = element_text(family = font, face = "italic", hjust = 0))
}

robins_facet_theme <- function(font = "Times") {
  theme_bw() +
    theme(plot.title = element_text(hjust = 0.5, family = font, face = "bold"),
          plot.subtitle = element_text(hjust = 0.5, face = "italic", family = font),
          axis.title = element_text(family = font),
          axis.text = element_text(family = font),
          legend.title = element_text(family = font),
          legend.text = element_text(family = font),
          plot.caption = element_text(family = font, face = "italic", hjust = 0))
}

################
# Training Set #
################

model_data <- County_Statistics_2020 %>%
  select(County,
         `2019_Population`,
         `Percent_Individuals_below_poverty_level_2015-2019`,
         `Median_household_income_2015-2019`,
         Adult_felony_cases_filed_2019,
         `Adult_felony_cases_per_1000`,
         Adult_felongy_cases_assigned_to_counsel,
         Misdemeanor_cases_filed_2019,
         Misdemeanor_cases_assigned_to_counsel,
         Misdemeanor_cases_per_1000,
         Juvenile_offender_cases_filed_2019,
         Juvenile_offender_cases_per_1000,
         Juvenile_offender_cases_assigned_to_counsel,
         Adult_Felony_Appointment_Rates,
         Adult_Misdemeanor_Appointment_Rates,
         Juvenile_Offender_Appointments
         ) %>%
  rename(
    Population_2019 = `2019_Population`,
    Percent_Individuals_below_poverty_level_2015_2019 =
      `Percent_Individuals_below_poverty_level_2015-2019`,
    Median_household_income_2015_2019 = `Median_household_income_2015-2019`) %>%
  mutate(
   Adult_Felony_Appointment_Rates =
     parse_number(Adult_Felony_Appointment_Rates)/100,
   Adult_Misdemeanor_Appointment_Rates =
     parse_number(Adult_Misdemeanor_Appointment_Rates)/100,
   Juvenile_Offender_Appointments =
     parse_number(Juvenile_Offender_Appointments)/100,
   Percent_Individuals_below_poverty_level_2015_2019 =
     parse_number(Percent_Individuals_below_poverty_level_2015_2019)/100,
   Median_household_income_2015_2019 =
     parse_number(Median_household_income_2015_2019)) %>%
  inner_join(Annual_Public_Defense_Spending_by_County %>%
               filter(Year==2019) %>%
               mutate(County = str_replace_all(County, "\\*", "")),
             by = "County") %>%
  select(-Year,
         -Total_Public_Defense_Budget) %>% # Removed County
  left_join(County_Home_Prices_revised,
            by="County") %>%
  relocate(prop_spending_by_year, .after = County) %>%
  select(-County)

model_data_splits <- loo_cv(model_data)
```

7.3 Data Partitioning

```
set.seed(1)
index <- createDataPartition(model_data$prop_spending_by_year, p = 0.75, list = FALSE)
train <- model_data[index, ]
test <- model_data[-index, ]
```

7.4 Full Linear Model

```
model_lin<- lm(prop_spending_by_year ~., data = train)
pred_lm <- predict(model_lin, test)

library(jtools)
options("jtools-digits" = 5)
export_summs(model_lin, scale = T,
             model.names = c("Model 1: Full Linear Model"),
             coefs = c("Population in 2019" =
                         "Population_2019",
                       "% Individuals Below Poverty Level 2015-2019" =
                         "Percent_Individuals_below_poverty_level_2015_2019",
                       "Median Household Income 2015-2019" =
                         "Median_household_income_2015_2019",
                       "Adult Felony Cases Filed in 2019" =
                         "Adult_felony_cases_filed_2019",
                       "Adult Felony Cases per 1000" =
                         "Adult_felony_cases_per_1000",
                       "Adult Felony Cases Assigned to Counsel" =
                         "Adult_felongy_cases_assigned_to_counsel",
                       "Misdemeanor Cases Filed 2019" =
                         "Misdemeanor_cases_filed_2019",
                       "Misdemeanor Cases Assigned to Counsel" =
                         "Misdemeanor_cases_assigned_to_counsel",
                       "Misdemeanor Cases per 1000" =
                         "Misdemeanor_cases_per_1000",
                       "Juvenile Offender Cases Filed in 2019" =
                         "Juvenile_offender_cases_filed_2019",
                       "Juvenile Offender Cases per 1000" =
                         "Juvenile_offender_cases_per_1000",
                       "Juvenile Offender Cases Assigned to Counsel" =
                         "Juvenile_offender_cases_assigned_to_counsel",
                       "Adult Felony Appointment Rates" =
                         "Adult_Felony_Appointment_Rates",
                       "Adult Misdemeanor Appointment Rates" =
                         "Adult_Misdemeanor_Appointment_Rates",
                       "Juvenile Offender Appointments" =
                         "Juvenile_Offender_Appointments",
                       "Median Home Price Quarter 4 2018" = "Median_Home_Price_Q4_2018",
                       "Median Home Price Quarter 1 2019" = "Median_Home_Price_Q1_2019",
                       "Median Home Price Quarter 2 2019" = "Median_Home_Price_Q2_2019",
                       "Median Home Price Quarter 3 2019" = "Median_Home_Price_Q3_2019",
                       "Median Home Price Quarter 4 2019" = "Median_Home_Price_Q4_2019",
                       "Median Home Price Change (% by Year)" =
                         "Median_Home_Price_Percent_Change_by_Year"))

lm_dat <- data.frame(truth = test$prop_spending_by_year, estimate = pred_lm)
lm_rmse <- (mean((lm_dat$truth - lm_dat$estimate)^2))^.5
```

7.5 Linear Subsets

```
feature_selection <- function(dat_train, dat_test, sub_method){
  subset <- regsubsets(prop_spending_by_year ~., data = dat_train, method = sub_method,
                       nvmax = ncol(dat_train))
  adjr2.max <- which.max(summary(subset)$adjr2)
  rss.min <- which.min(summary(subset)$rss)
  cp.min <- which.min(summary(subset)$cp)
  bic.min <- which.min(summary(subset)$bic)
  data.frame(adjr2.max, rss.min, cp.min, bic.min)

  coefs <- coef(subset, adjr2.max)
  coef_names <- names(coefs)
  formula_adjr2 <- reformulate(coef_names[-1], response = "prop_spending_by_year")
  mod_adjr2 <- lm(formula_adjr2, data = dat_train)
  mod_adjr2_preds <- predict(mod_adjr2, dat_test)
  mod_adjr2_rmse <- (mean((dat_test$prop_spending_by_year - mod_adjr2_preds)^2))^.5

  coefs <- coef(subset, rss.min)
  coef_names <- names(coefs)
  formula_rss <- reformulate(coef_names[-1], response = "prop_spending_by_year")
  mod_rss <- lm(formula_rss, data = dat_train)
  mod_rss_preds <- predict(mod_rss, dat_test)
  mod_rss_rmse <- (mean((dat_test$prop_spending_by_year - mod_rss_preds)^2))^.5

  coefs <- coef(subset, cp.min)
  coef_names <- names(coefs)
  formula_cp <- reformulate(coef_names[-1], response = "prop_spending_by_year")
  mod_cp <- lm(formula_cp, data = dat_train)
  mod_cp_preds <- predict(mod_cp, dat_test)
  mod_cp_rmse <- (mean((dat_test$prop_spending_by_year - mod_cp_preds)^2))^.5

  coefs <- coef(subset, bic.min)
  coef_names <- names(coefs)
  formula_bic <- reformulate(coef_names[-1], response = "prop_spending_by_year")
  mod_bic <- lm(formula_bic, data = dat_train)
  mod_bic_preds <- predict(mod_bic, dat_test)
  mod_bic_rmse <- (mean((dat_test$prop_spending_by_year - mod_bic_preds)^2))^.5

  data.frame(model = c("adjr2.max", "rss.min", "cp.min", "bic.min"),
             model_n = c(adjr2.max, rss.min, cp.min, bic.min),
             rMSE = c(mod_adjr2_rmse, mod_rss_rmse, mod_cp_rmse, mod_bic_rmse),
             included_vars = c(as.character(formula_adjr2[3]),
                      as.character(formula_rss[3]),
                      as.character(formula_cp[3]),
                      as.character(formula_bic[3])))
  }

feature_selection(dat_train = train, dat_test = test, sub_method = "forward") %>%
  select(-included_vars) %>%
  kable(format = "simple", caption = "Evaluating Forward Selection", col.names =
        c("Evaluation", "Predictors", "rMSE"))

feature_selection(dat_train = train, dat_test = test, sub_method = "backward") %>%
  select(-included_vars) %>%
  kable(format = "simple", caption = "Evaluating Backward Selection", col.names =
        c("Evaluation", "Predictors", "rMSE"))
```

7.6 Ridge Regression

```
grid = 10^(seq( -5, 5, length = 100))
x<-model.matrix(prop_spending_by_year ~., data = train)[,-1]
y<-train$prop_spending_by_year
ridge_mod <- glmnet(x, y, alpha = 0, lambda = grid)

ridge_mod_cv <- cv.glmnet(x, y, alpha = 0, lambda = grid, nfolds = 10 )
plot(ridge_mod_cv, xvar = "lambda")

plot(ridge_mod, xvar = "lambda")

best_L <- ridge_mod_cv$lambda.min
reg_L <- ridge_mod_cv$lambda.1se

x_tst <- model.matrix(prop_spending_by_year ~., data = test)[,-1]
pred_ridge <- predict(ridge_mod, s = best_L, newx = x_tst)

ridge_dat <- data.frame(truth = test$prop_spending_by_year, estimate = pred_ridge)

ridge_rmse <- (mean((ridge_dat$truth - ridge_dat$s1)^2))^.5
```

7.7 Lasso

```
grid = 10^(seq( -5, 5, length = 100))
x<-model.matrix(prop_spending_by_year ~., data = train)[,-1]
y<-train$prop_spending_by_year
lasso_mod <- glmnet(x, y, alpha = 1, lambda = grid)

set.seed(21)
lasso_mod_cv <- cv.glmnet(x, y, alpha = 1, lambda = grid, nfolds = 10 )

plot(lasso_mod_cv, xvar = "lambda")

plot(lasso_mod, xvar = "lambda")

best_L <- lasso_mod_cv$lambda.min
```

```
reg_L <- lasso_mod_cv$lambda.1se
d <- data.frame(best_L, reg_L)
kable(d, format = "simple", col.names = c("Best Lambda Value", "Best Lambda within 1 SE"),
      caption = "Comparison of LASSO Tuning Parameters")

x_tst <- model.matrix(prop_spending_by_year ~., data = test)[,-1]
pred_lasso <- predict(lasso_mod, s = best_L, newx = x_tst)

lasso_dat <- data.frame(truth = test$prop_spending_by_year, estimate = pred_lasso) %>%
  rename(estimate = s1)

lasso_rmse <- (mean((lasso_dat$truth - lasso_dat$estimate)^2))^.5
```

7.8 Pruned Tree

```
set.seed(13334)
c_dt <- rpart(prop_spending_by_year ~., control = rpart.control(minsplit = 6,
                                                                xval = 10,
                                                                maxdepth = 8),
              data = train)

cp <- prune(c_dt, cp = 0.017)

rpart.plot(cp)

pred_prune <- predict(cp, test)

tree_dat <- data.frame(truth = test$prop_spending_by_year, estimate = pred_prune)
tree_rmse <- (mean((tree_dat$truth - tree_dat$estimate)^2))^.5
```

7.9 Random Forest

```
rfmodel <- randomForest(prop_spending_by_year ~., data = train, mtry = (ncol(test) - 1)/3)

df <- getTree(rfmodel, labelVar = TRUE) %>%
  dplyr::select(`split var`) %>% drop_na() %>% distinct()
kable(df, format = "simple", col.names = c("Split Variables"),
      caption = "Variables Split on in at Least One Random Forest Iteration")

# Had to rework vapImpPlot() to make this work
varImpPlot <- function(x, sort=TRUE,
                       n.var=min(30, nrow(x$importance)),
                       type=NULL, class=NULL, scale=TRUE,
                       main=deparse(substitute(x)), ...) {
    if (!inherits(x, "randomForest"))
        stop("This function only works for objects of class `randomForest'")
    imp <- importance(x, class=class, scale=scale, type=type, ...)
    ## If there are more than two columns, just use the last two columns.
    if (ncol(imp) > 2) imp <- imp[, -(1:(ncol(imp) - 2))]
    nmeas <- ncol(imp)
    if (nmeas > 1) {
        op <- par(mfrow=c(1, 2), mar=c(4, 5, 4, 1), mgp=c(2, .8, 0),
                  oma=c(0, 0, 2, 0), no.readonly=TRUE)
        on.exit(par(op))
    }
    for (i in 1:nmeas) {
        ord <- if (sort) rev(order(imp[,i],
                                   decreasing=TRUE)[1:n.var]) else 1:n.var
        xmin <- if (colnames(imp)[i] %in%
                    c("IncNodePurity", "MeanDecreaseGini")) 0 else min(imp[ord, i])
    }
    if (nmeas > 1) mtext(outer=TRUE, side=3, text=main, cex=1.2)
    invisible(imp)
}

rf_var_imp <- as.data.frame(varImpPlot(rfmodel))
rf_var_imp$varnames <- rownames(rf_var_imp)
rownames(rf_var_imp) <- NULL

ggplot(rf_var_imp, aes(x = reorder(varnames, IncNodePurity), y = IncNodePurity)) +
  geom_point(shape = 1) +
  geom_segment(aes(x = varnames,xend = varnames,y = 0,yend = IncNodePurity)) +
  ylab("IncNodePurity") +
  xlab("Variable Name") +
  ggtitle("Random Forest Node Purity") +
  coord_flip() +
  robins_ggplot_theme()

rf_dat <- data.frame(truth = test$prop_spending_by_year, estimate = predict(rfmodel,
                                                                            test))
rf_rmse <- (mean((rf_dat$truth - rf_dat$estimate)^2))^.5
```

7.10 Bagging

```
bag_model <- randomForest(prop_spending_by_year ~., data = train, mtry = (ncol(test) - 1))

df <- getTree(bag_model, labelVar = TRUE) %>%
  dplyr::select(`split var`) %>% drop_na() %>% distinct()
kable(df, format = "simple", col.names = c("Split Variables"),
      caption = "Variables Split on in Bagged Model")

var_imp <- as.data.frame(varImpPlot(bag_model))
var_imp$varnames <- rownames(var_imp)
rownames(rf_var_imp) <- NULL

ggplot(var_imp, aes(x = reorder(varnames, IncNodePurity), y = IncNodePurity)) +
  geom_point(shape = 1) +
  geom_segment(aes(x = varnames,xend = varnames,y = 0,yend = IncNodePurity)) +
  ylab("IncNodePurity") +
  xlab("Variable Name") +
  ggtitle("Bagged Model Node Purity") +
  coord_flip() +
  robins_ggplot_theme()

bag_dat <- data.frame(truth = test$prop_spending_by_year, estimate = predict(bag_model,
                                                                              test))
bag_rmse <- (mean((bag_dat$truth - bag_dat$estimate)^2))^.5
```

Works Cited

"Argersinger v. hamlin" (1972), Supreme Court.

Carney, K., and Morales, A. (2020), "Washington counties by population," *Washington Demographics*, United States Census Bureau.

Coplen, D. (2012a), "Washington state office of public defense," *Washington State Office of Public Defense*, Washington State Office of Public Defense.

Coplen, D. (2012b), "Public defense improvement program," *Washington State Office of Public Defense*, Washington State Office of Public Defense.

Fabelo, T. (2004), "What policy-makers need to know to improve indigent defense systems," *Nyu rev. l. & soC. ChaNge*, HeinOnline, 29, 135.

"Gideon v. wainwright" (1963), Supreme Court.

Hardwick, R. (2021a), *OPD funding and graphs*, Washington Defender Association.

Hardwick, R. (2021b), "OPD funding and graphs," *RPubs*, Washington Defender Association.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013), *An introduction to statistical learning*, Springer.

Legislature, 59th (2005), "RCW 10.101.070 county moneys." https://app.leg.wa.gov/RCW/default.aspx?cite=10.101.070.

Moore, J. I. (2018), *Annual fiscal report*, Washington State Office of Public Defense.

Ogletree, C. J. (1995), "An essay on the new public defender for the 21st century," *Law and Contemporary Problems*, JSTOR, 58, 81–93.

"RCW 10.101.060" (2005), *RCW 10.101.060: County moneys.*, Washington State Legislature.

"Understanding gideon's impact, part 2: The birth of the public defender movement" (2021), Sixth Amendment Center, Inc.