Writeup

Writeup

Project 1

Both the pdfs "Arizona_Department_of_Correction_Inmate_Fact_Sheet_2020.pdf" and "Arizona_Department_of_Corrections_Cost_Report_2020.pdf" which can be found in ./data/Raw_Data are reports released by the Arizona Department of Corrections,Rehabilitation& Reentry.

In order to extract tables from these pdf reports into csv files, I first used Adobe Acrobat to turn the pdf files into xlsx files. The xlsx versions of these files can also be found in ./data/Raw_Data. Then, I used Excel to inspect these files and identify which tables I wanted to extract. Each table was extracted into its own csv file.

Cost Report

The Cost Report could not easily be saved as a csv file because it had images and text that was not in a table. Thus, saving it as a csv would cause information to be lost. Thus, I decided to extract the individual tables that were relevant to me from the xlsx file. The R script "Cost_Report_Data_Cleaning.R" extracts these tables into csv files. The focus of my project is the difference in expenditures between private and public prison units. Thus, although the report has lots of interesting information, I focused on tables directly related to this.

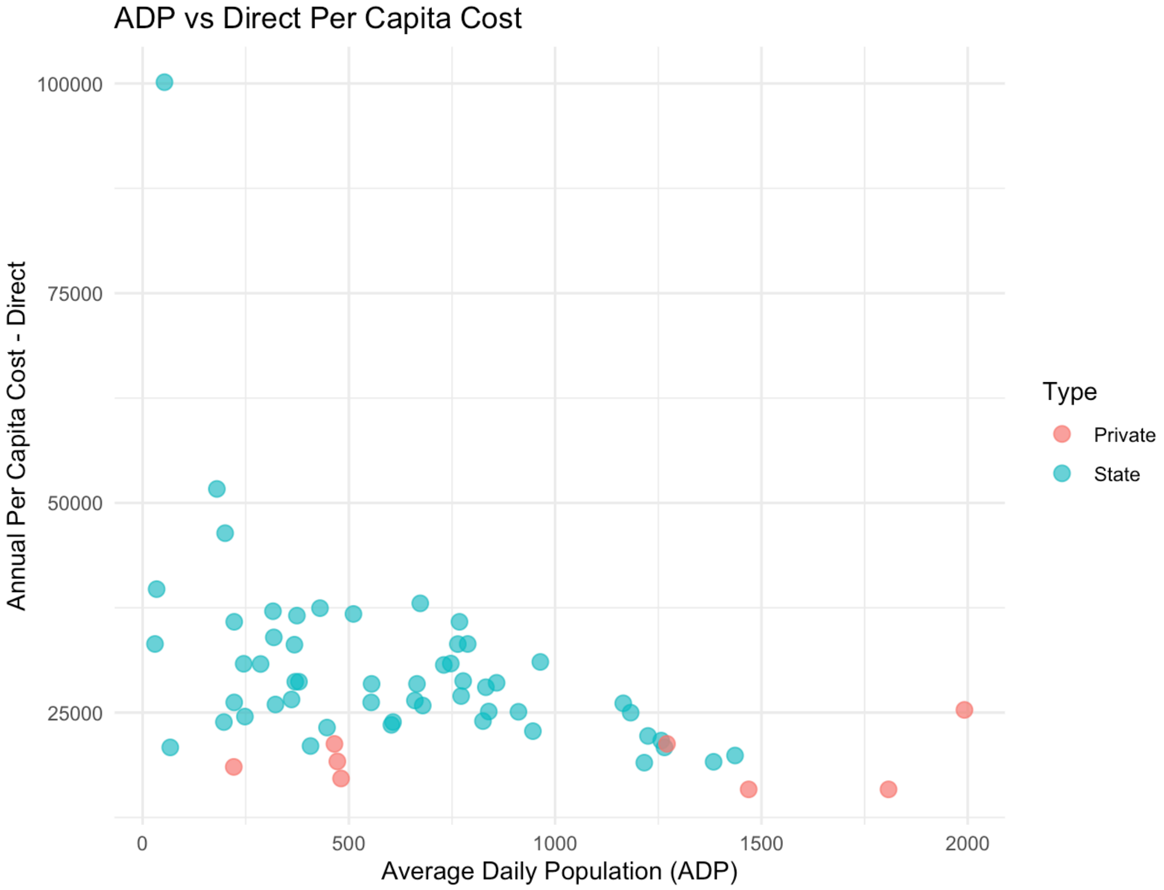
I extracted the Expenditure Summary table on page 8 of the pdf. In this table, I only extracted the information from the Private Prisons. The information on private prisons was organized by unit. I extracted the entire table from pages 10 and 11 of the pdf, which contains information on State Prison Expenditure by Prison Unit. Since both the extracted data for Private and Public prisons was then organized by unit, I could compare the data.

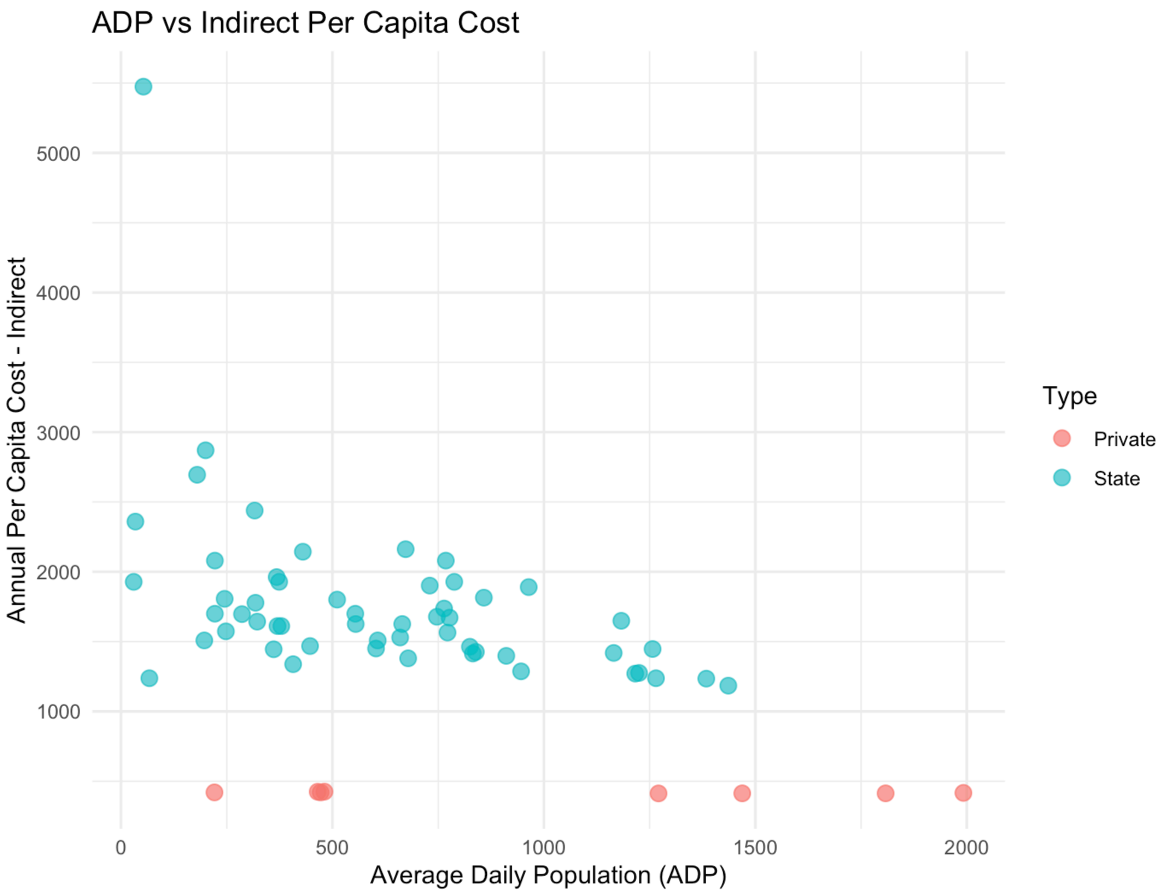
These extracted csv files are saved in ./data/Processed_Data as follows:

- Private_Prison_Revenue.csv
- State_Prison_Revenue.csv

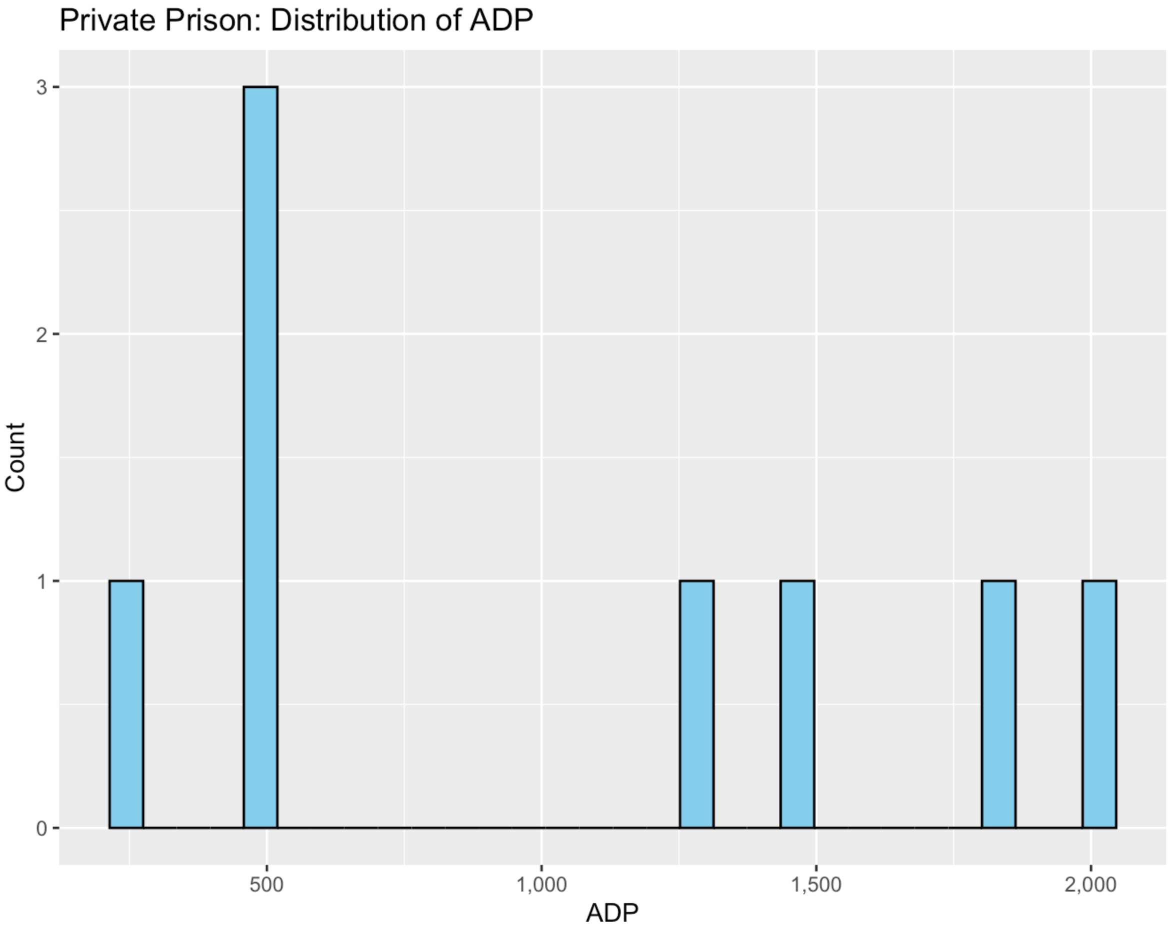
Multivariate Plot:

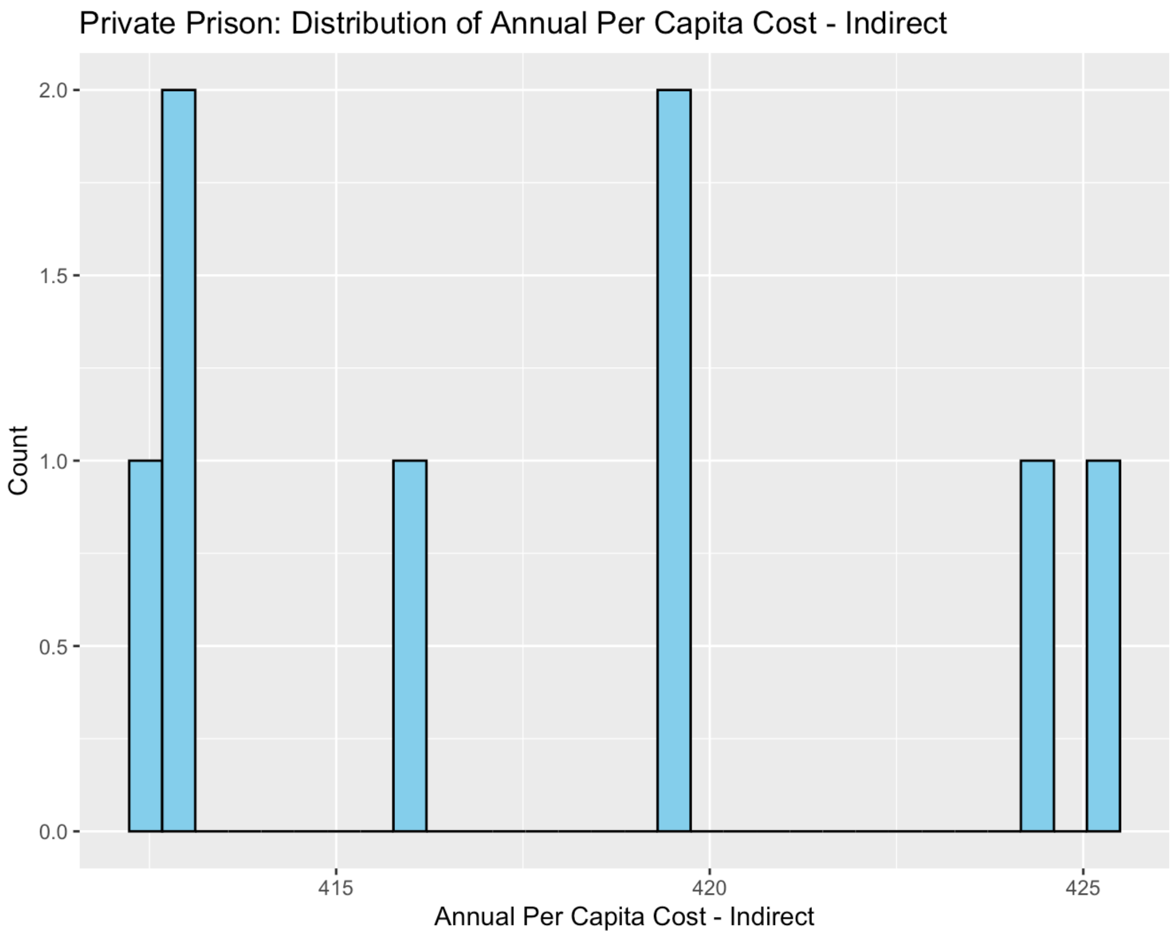
Private prisons have less Direct and Indirect Annual Per Capita Cost than State prisons. Direct Costs are tied to the operation of a specific prison facility or unit. They include salaries for staff; food, clothing, and hygiene products for inmates; utilities (electricity, water, etc.) used in the facility; and facility maintenance. Indirect costs are shared or overhead costs that support the prison system but aren't directly linked to one facility. They include central administration salaries (e.g., the state corrections department headquarters); IT systems that serve multiple facilities; legal services and oversight; state-level contracting or procurement services; and training programs for correctional officers across facilities.

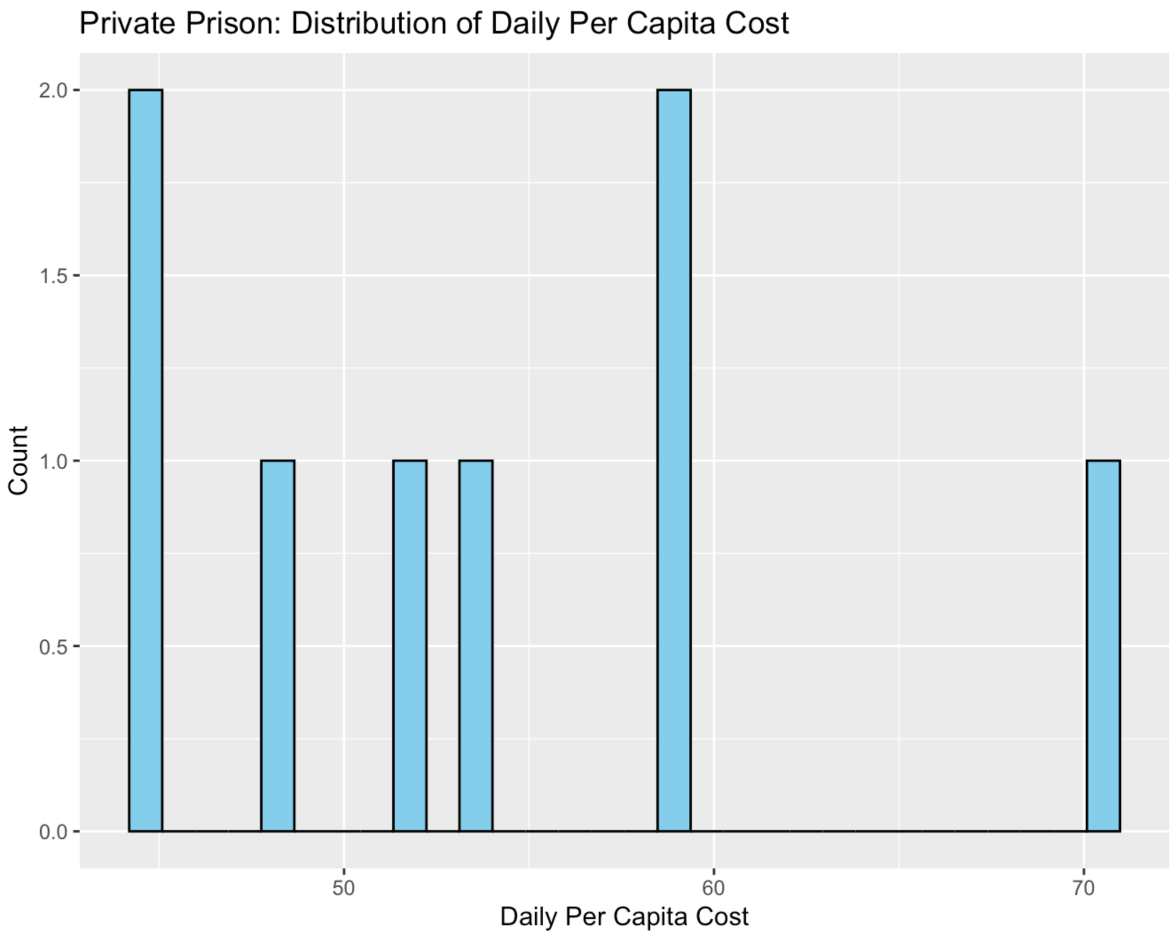


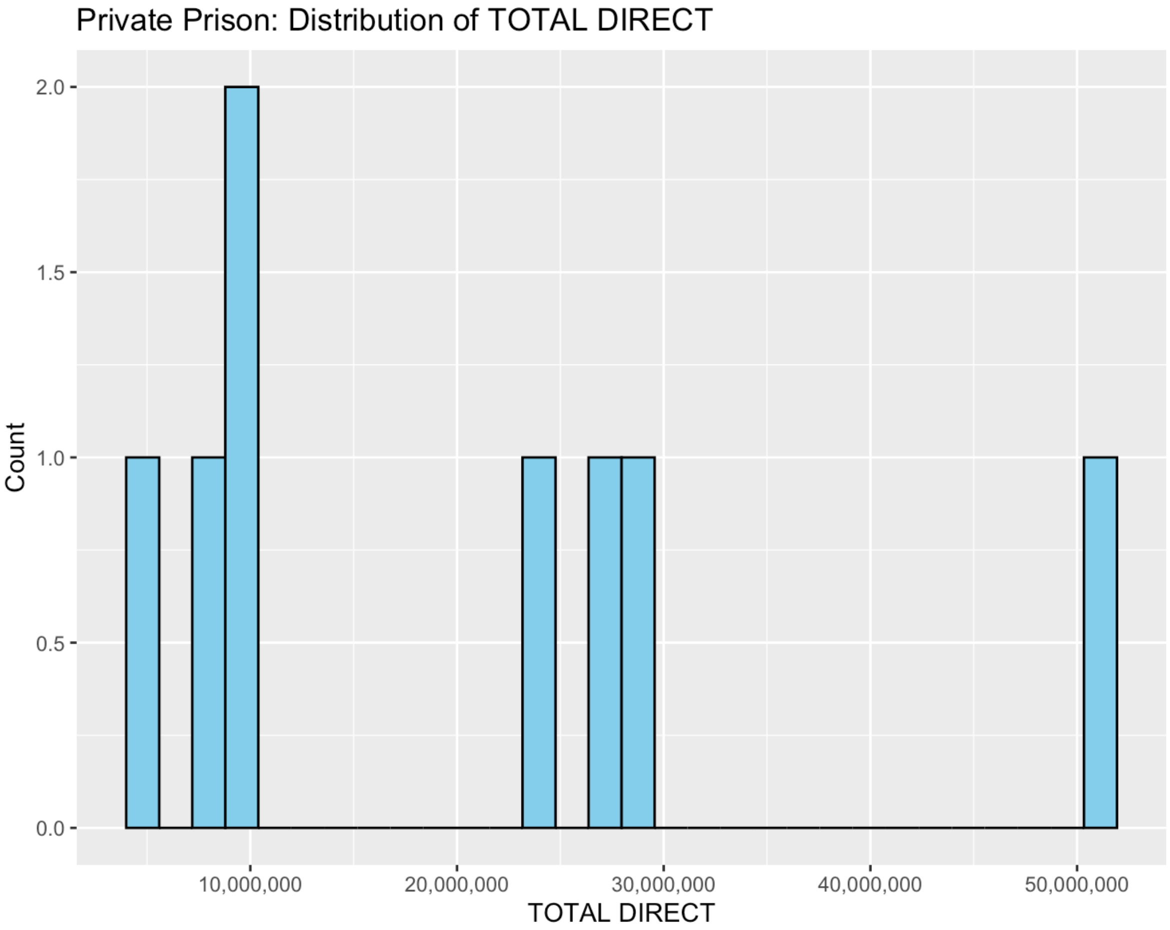


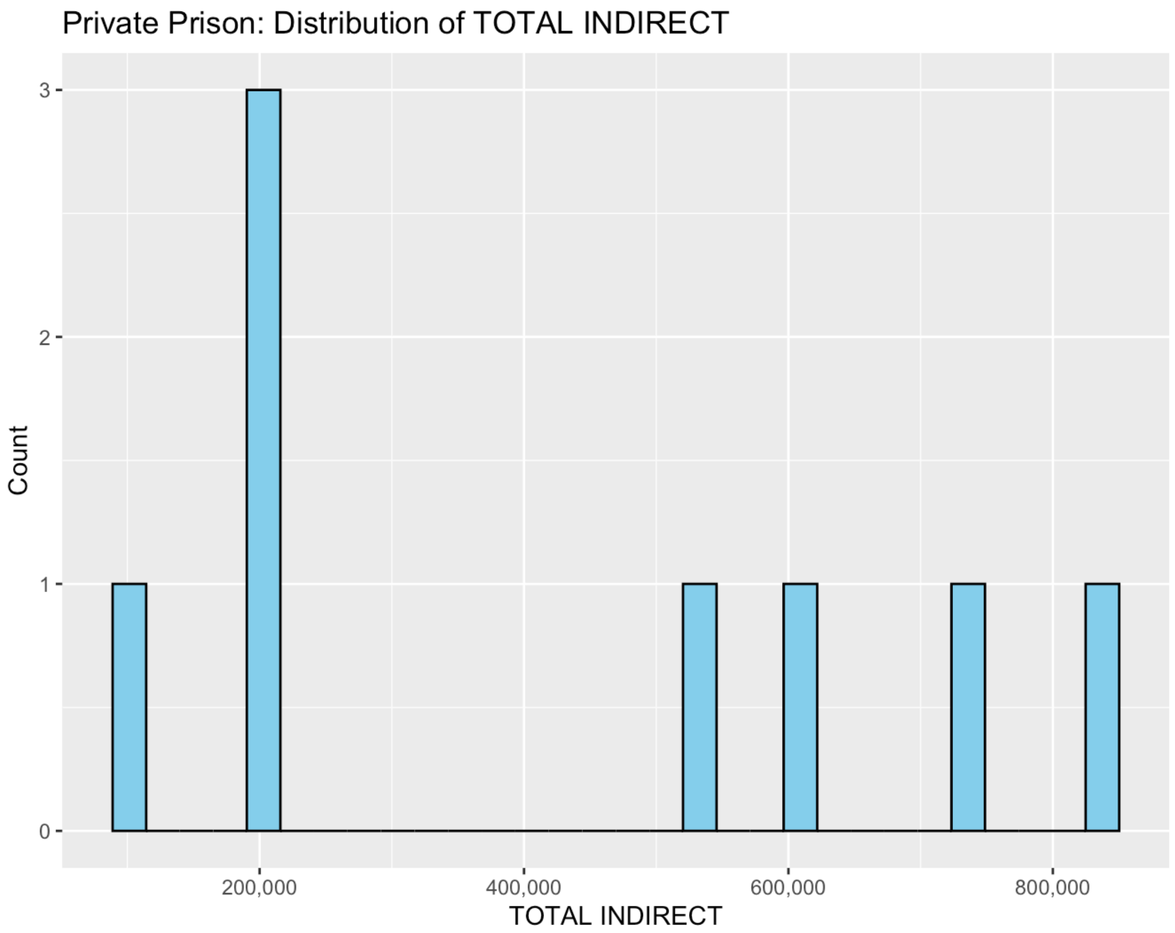
Private Prison Distributions:


Private Prison: Distribution of ADP
Count
3
2
1
0
500
1,000
1,500
2,000
ADP

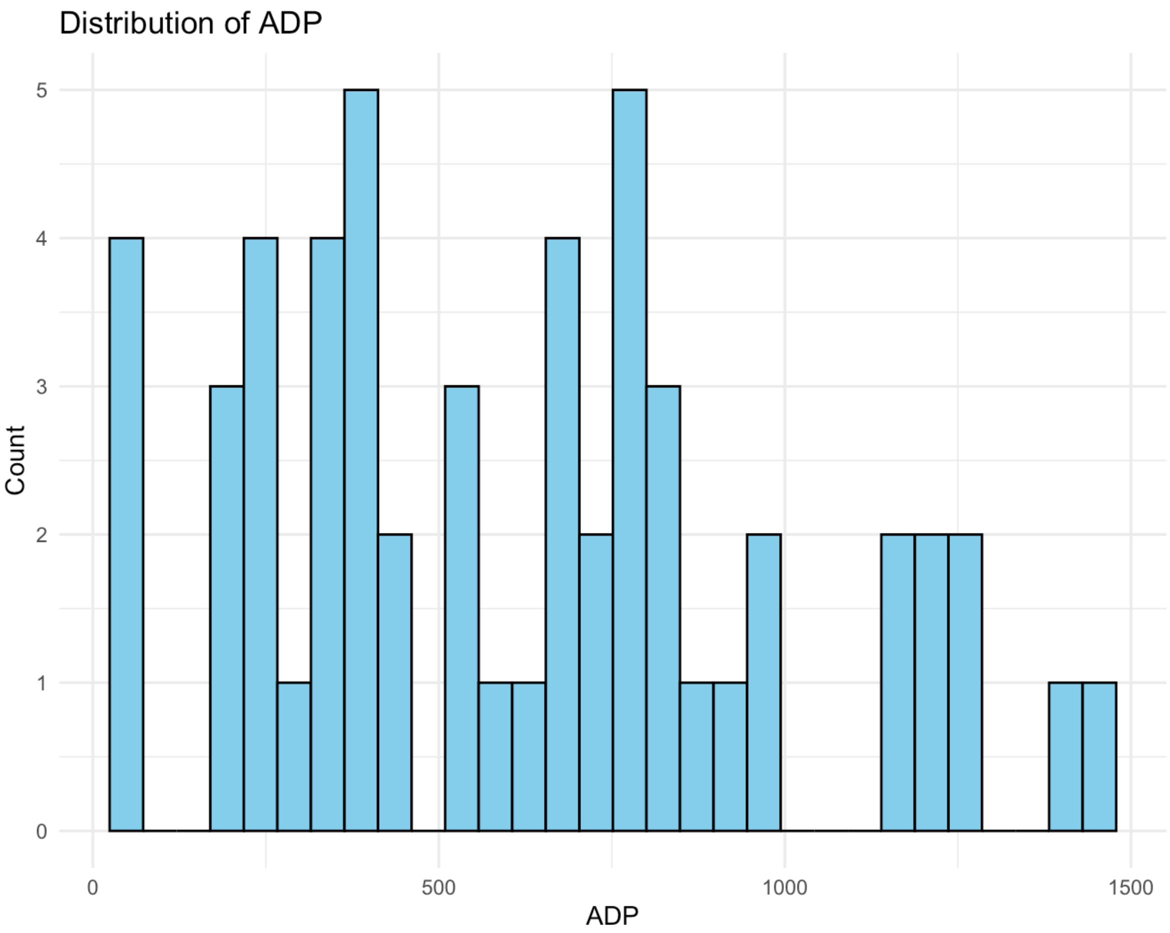

Private Prison: Distribution of Annual Per Capita Cost - Indirect
Count
2.0
1.5
1.0
0.5
0.0
415
420
425
Annual Per Capita Cost - Indirect

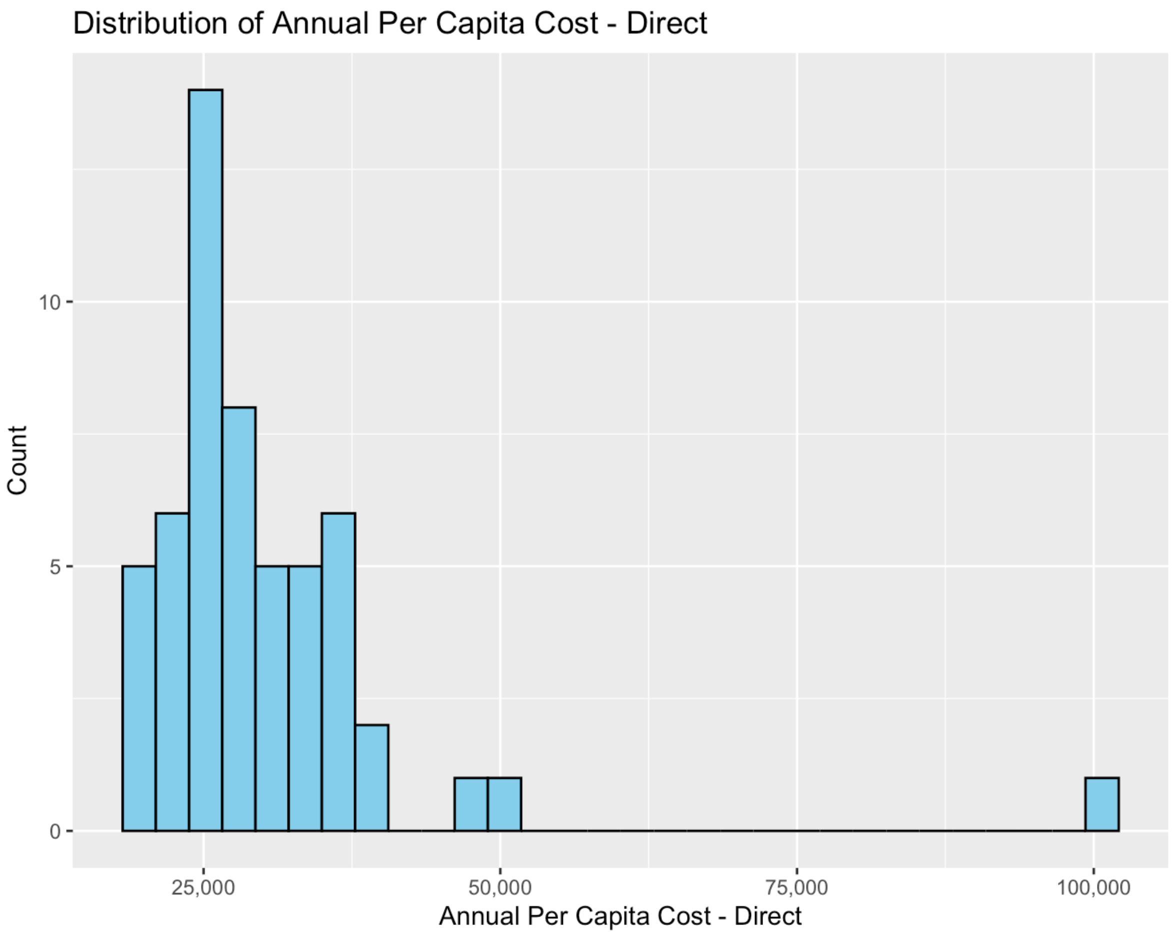

Private Prison: Distribution of Daily Per Capita Cost
Count
2.0
1.5
1.0
0.5
0.0
50
60
70
Daily Per Capita Cost

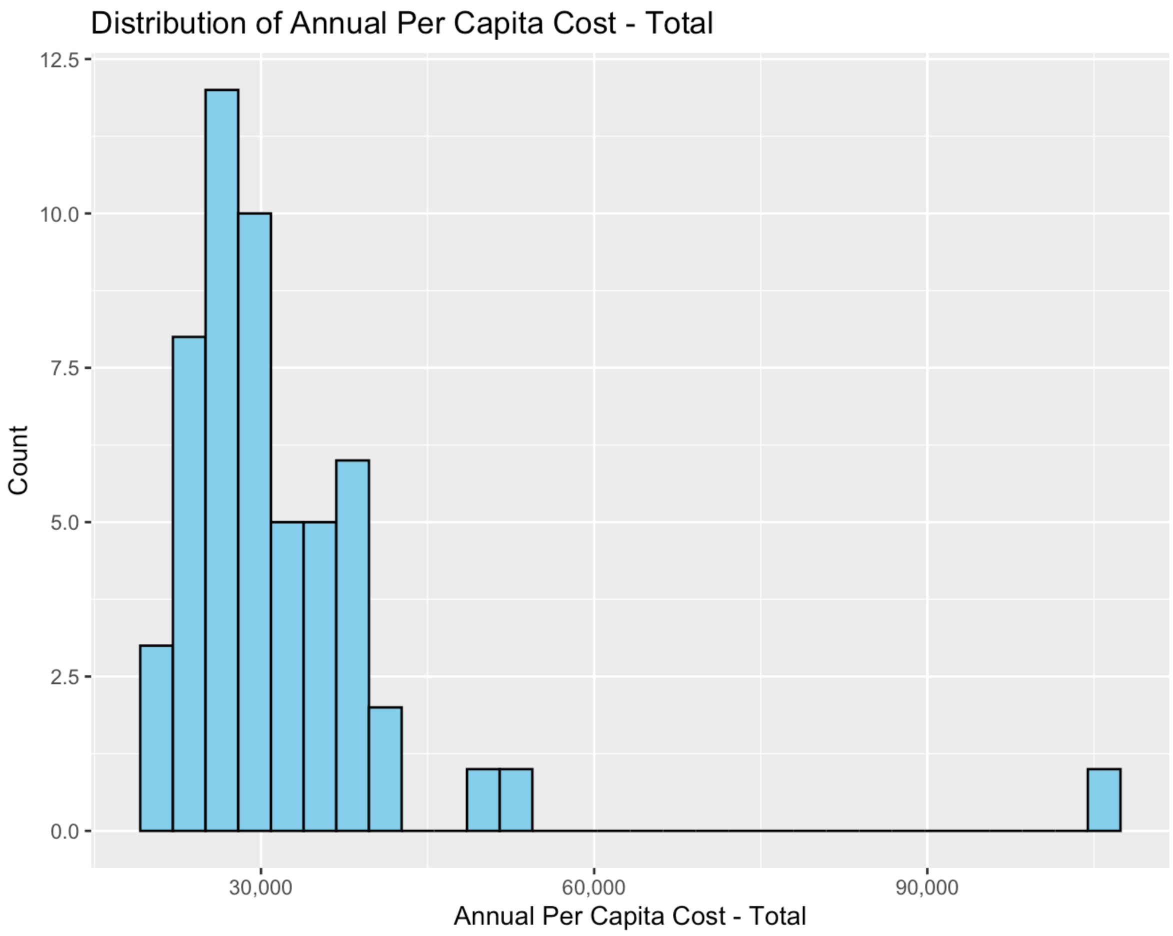

Private Prison: Distribution of TOTAL DIRECT
Count
2.0
1.5
1.0
0.5
0.0
10,000,000
20,000,000
30,000,000
40,000,000
50,000,000
TOTAL DIRECT

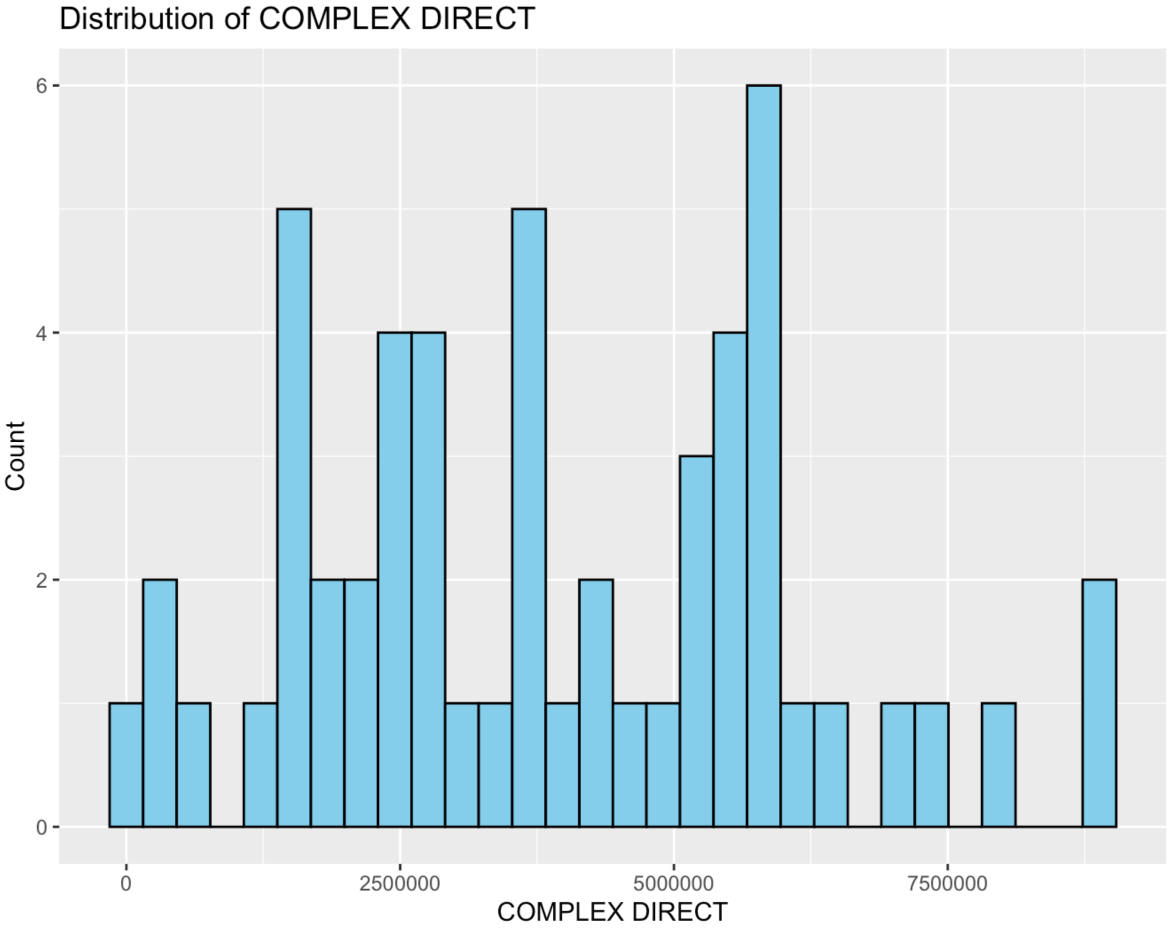


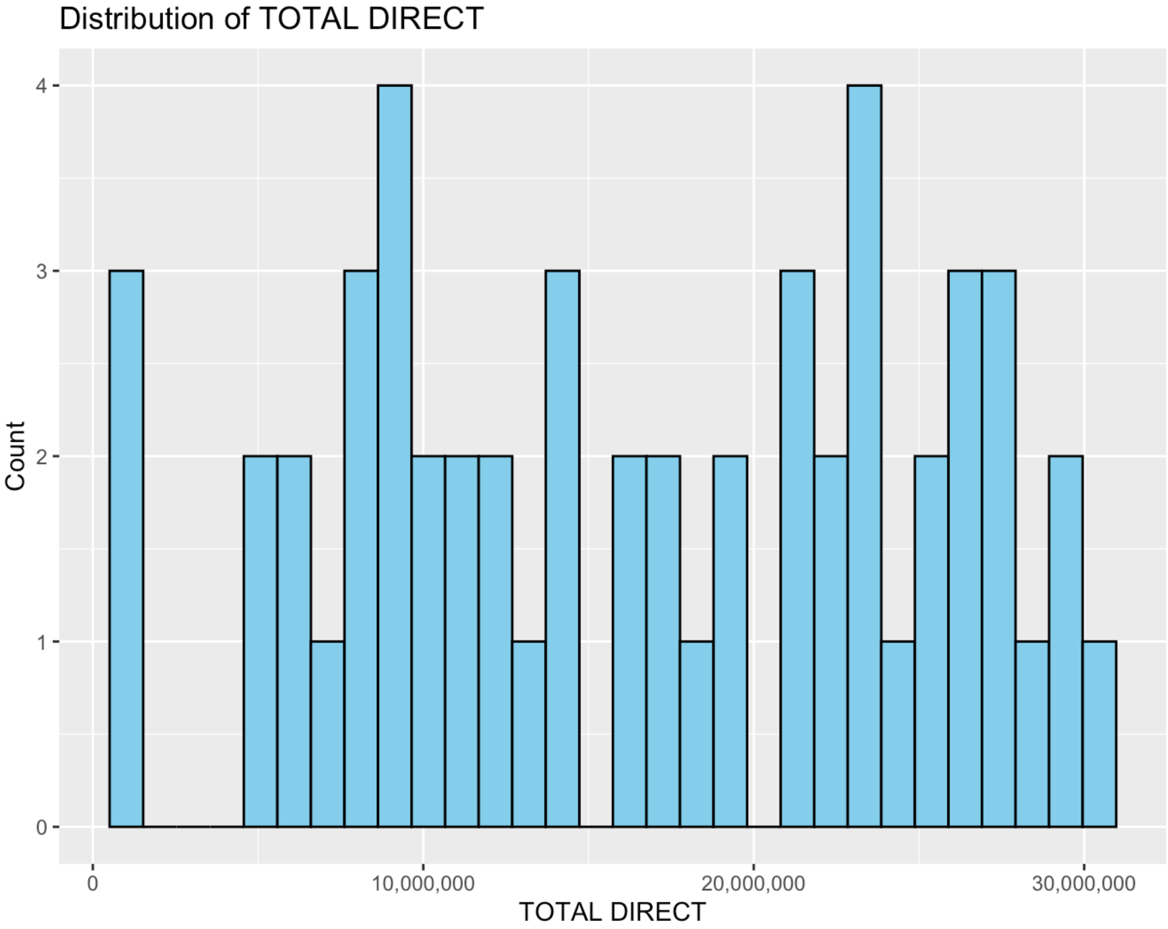
State Prison Distributions:

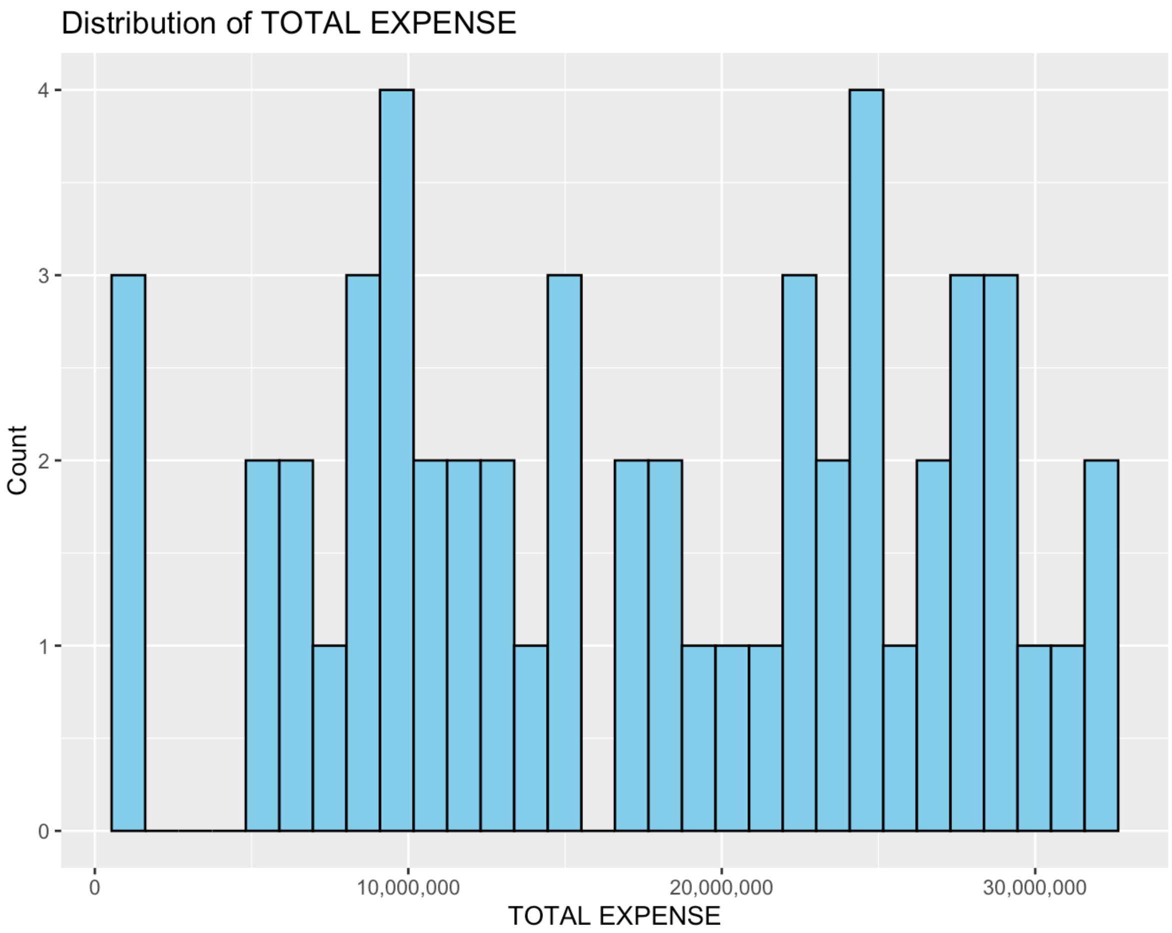

Distribution of ADP
Count
5
4
3
2
1
0
0
500
1000
1500
ADP

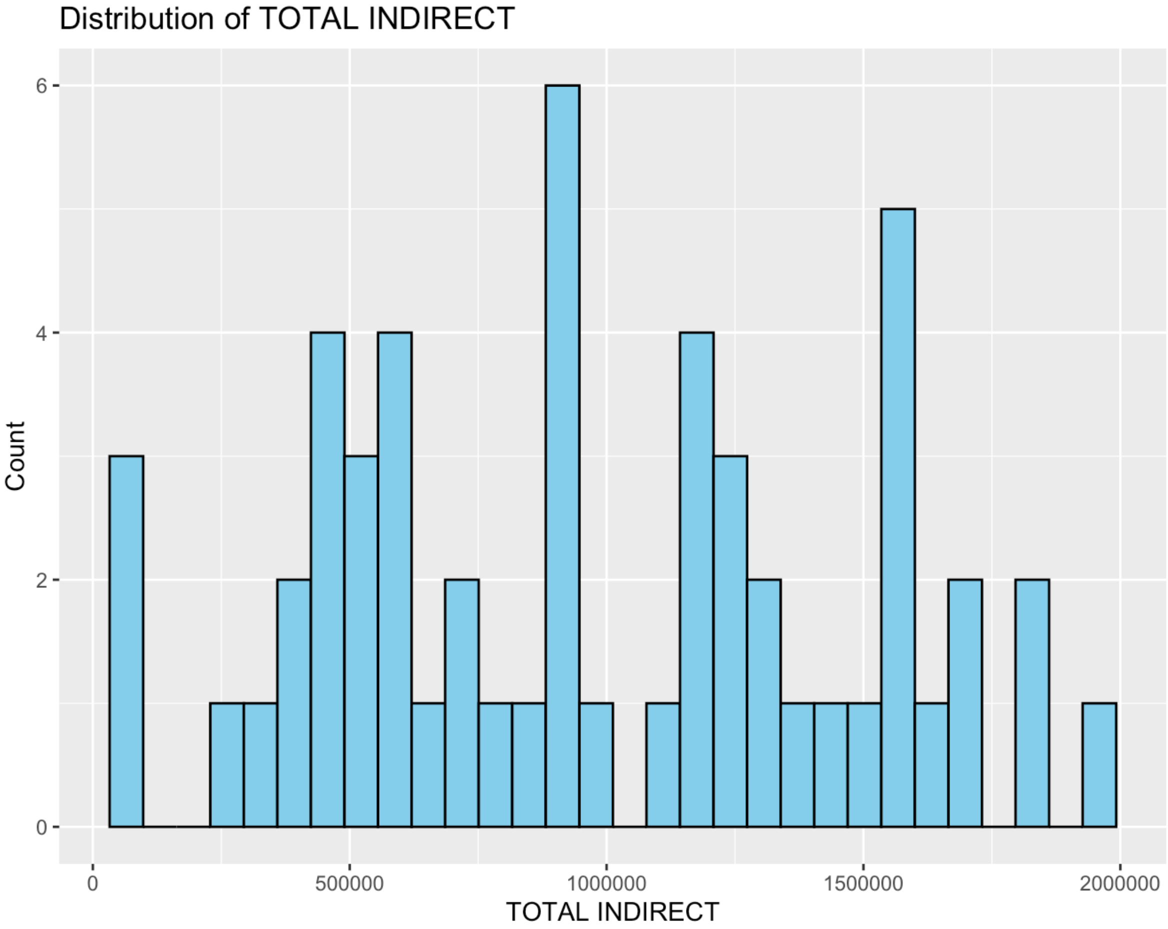

Distribution of Annual Per Capita Cost - Direct
Count
0
5
10
25,000
50,000
75,000
100,000
Annual Per Capita Cost - Direct

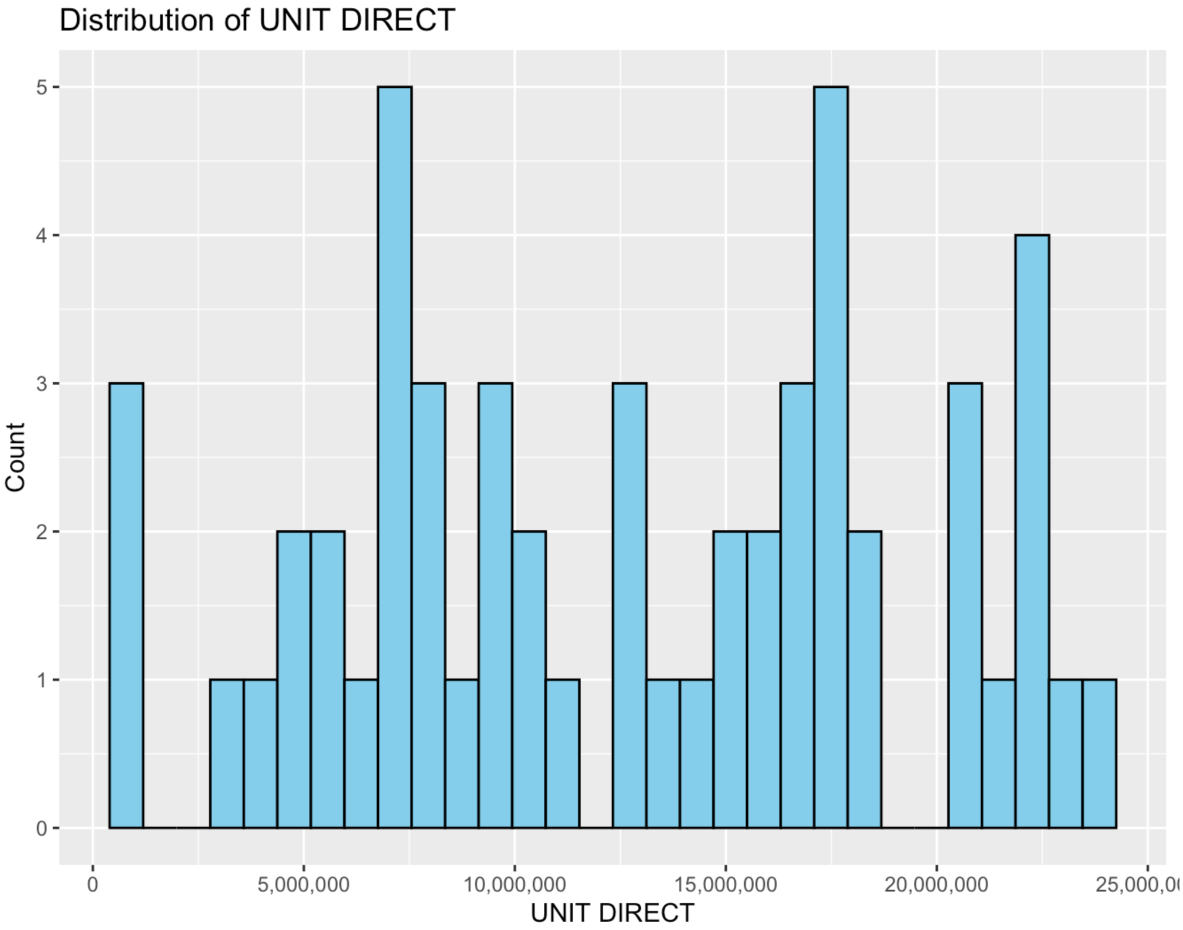

Distribution of Annual Per Capita Cost - Total
Count
12.5
10.0
7.5
5.0
2.5
0.0
30,000
60,000
90,000
Annual Per Capita Cost - Total


Distribution of COMPLEX DIRECT
Count
6
4
2
0
0
2500000
5000000
7500000
COMPLEX DIRECT


Distribution of TOTAL DIRECT
Count
4
3
2
1
0
0
10,000,000
20,000,000
30,000,000
TOTAL DIRECT


Distribution of TOTAL EXPENSE
Count
4
3
2
1
0
0
10,000,000
20,000,000
30,000,000
TOTAL EXPENSE


Distribution of TOTAL INDIRECT
Count
0
2
4
6
0
500000
1000000
1500000
2000000
TOTAL INDIRECT



Project 3

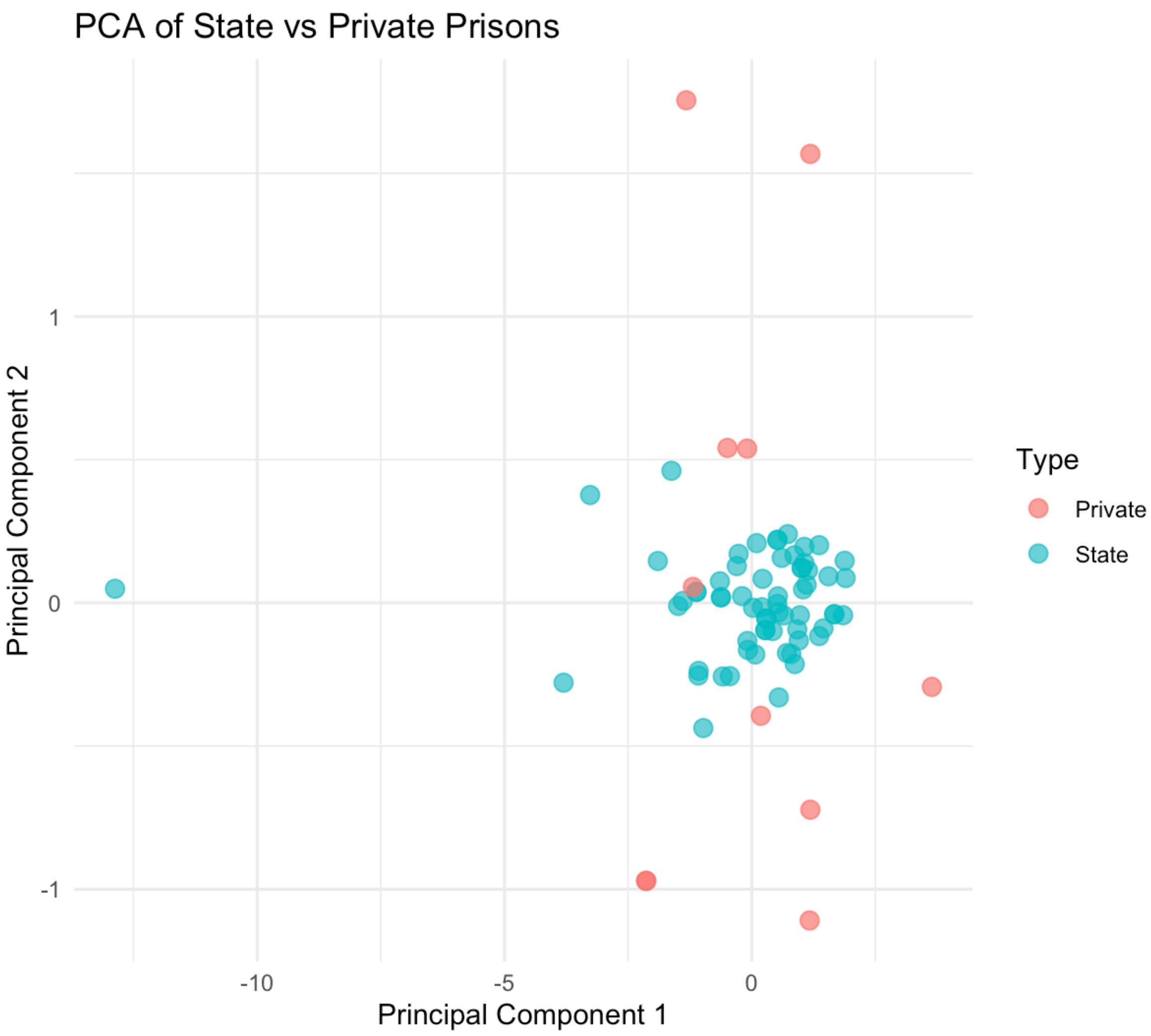
Model

In order to analyze the difference between private and public prisons, I used Principle Component Analysis. After reducing dimensionality with PCA, I created a plot in which private and public prisons are colored differently.

Analysis

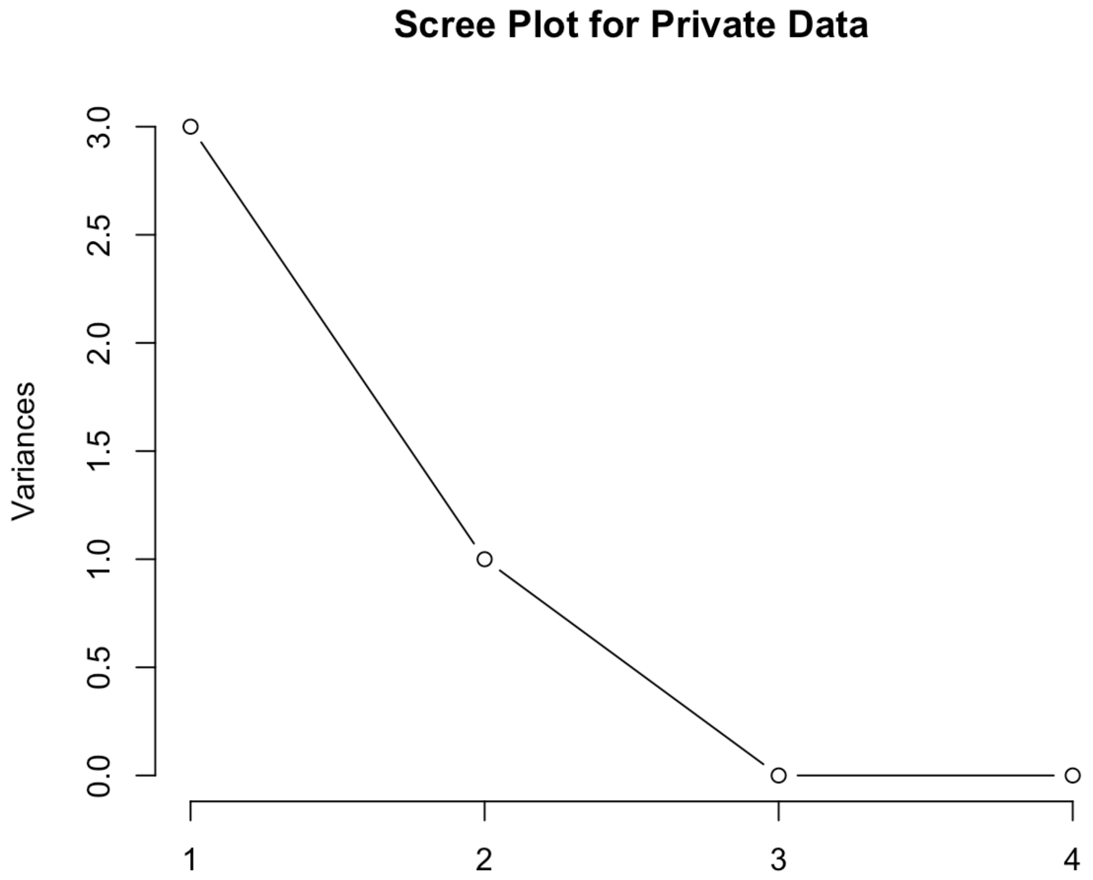
PCA

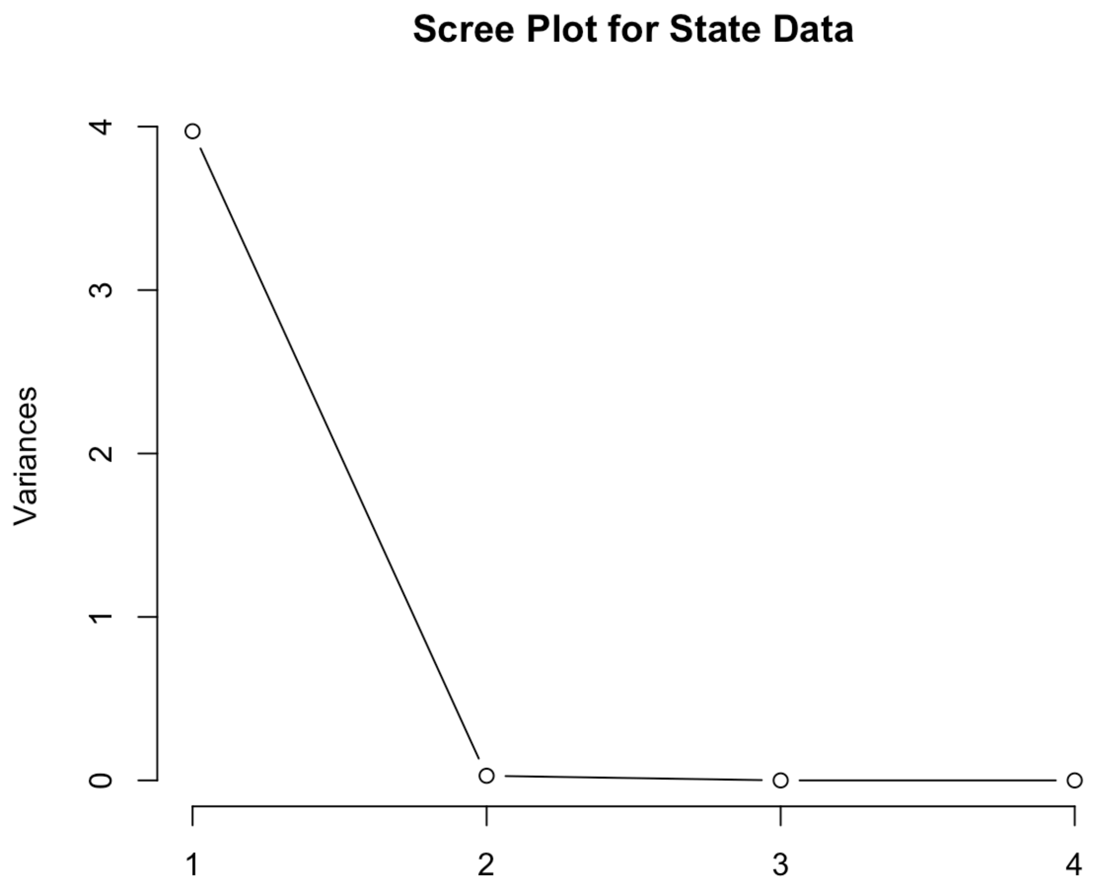
The PCA Graph shows that State prisons are more consistent in how they operate per inmate and Private prisons are more varied and some diverge significantly from the state cluster. The difference is not significant enough to determine that all private prisons are different from all state prisons.



The scree plot for private prisons shows that PCA1 and PCA2 have non-zero variances. This implies PCA found two patterns in the data.

The scree plot for state prisons shows that only PCA1 has a non-zero variance. This implies PCA found only one pattern in the data.


Scree Plot for Private Data
Variances
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1
2
3
4



These are the features in the PCA components:

PC1 (State) and PC1 (Private):

Annual Per Capita Cost - Total

Daily Per Capita Cost

Annual Per Capita Cost - Direct

Annual Per Capita Cost - Indirect

PC2 (Private):

Annual Per Capita Cost - Indirect

Annual Per Capita Cost - Direct

Daily Per Capita Cost

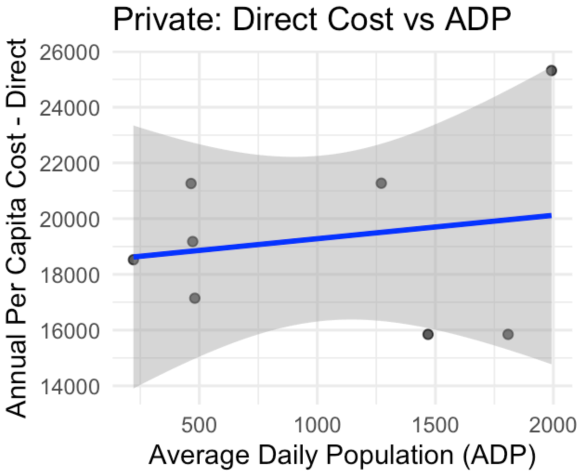
Annual Per Capita Cost - Total

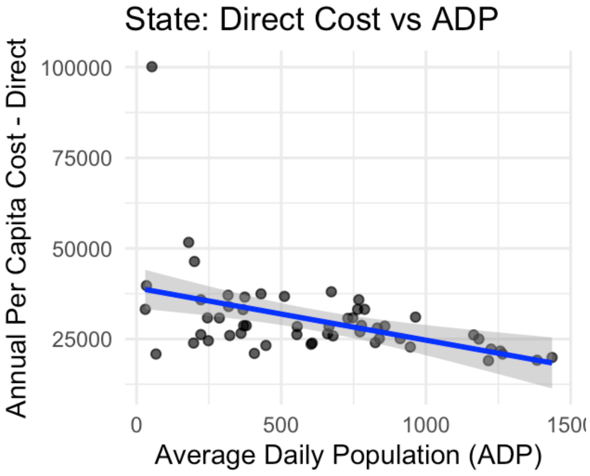
This suggests that Annual Per Capita Cost - Indirect is relevant to Private Prisons but not State Prisons.

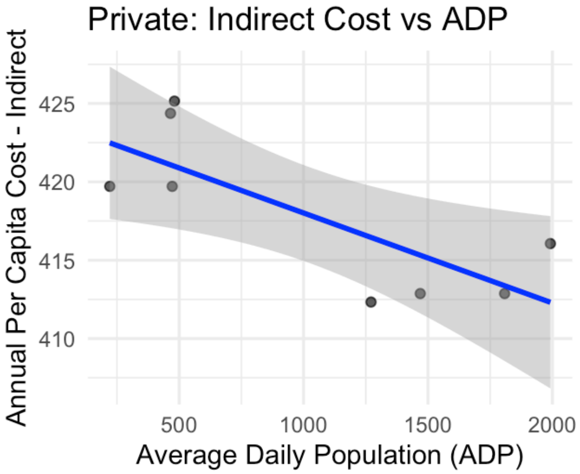
Individual Regression

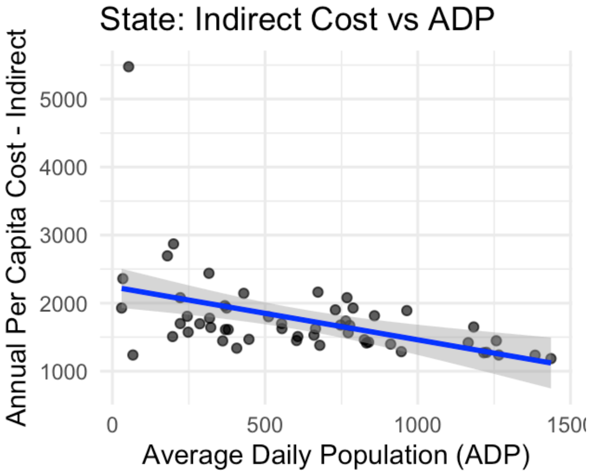
I conducted regression to quantify how prison type and ADP effects cost metrics. I controlled for number of prisoners by only looking at the per capita measures.

For Per Capita Cost - Direct, there is no clear relationship with Private Data. There is a negative relationship with variance for Per Capita Cost - Indirect with Private Data. State Data shows a clear negative relationship for both Per Capita Cost - Direct and Per Capita Cost - Indirect.

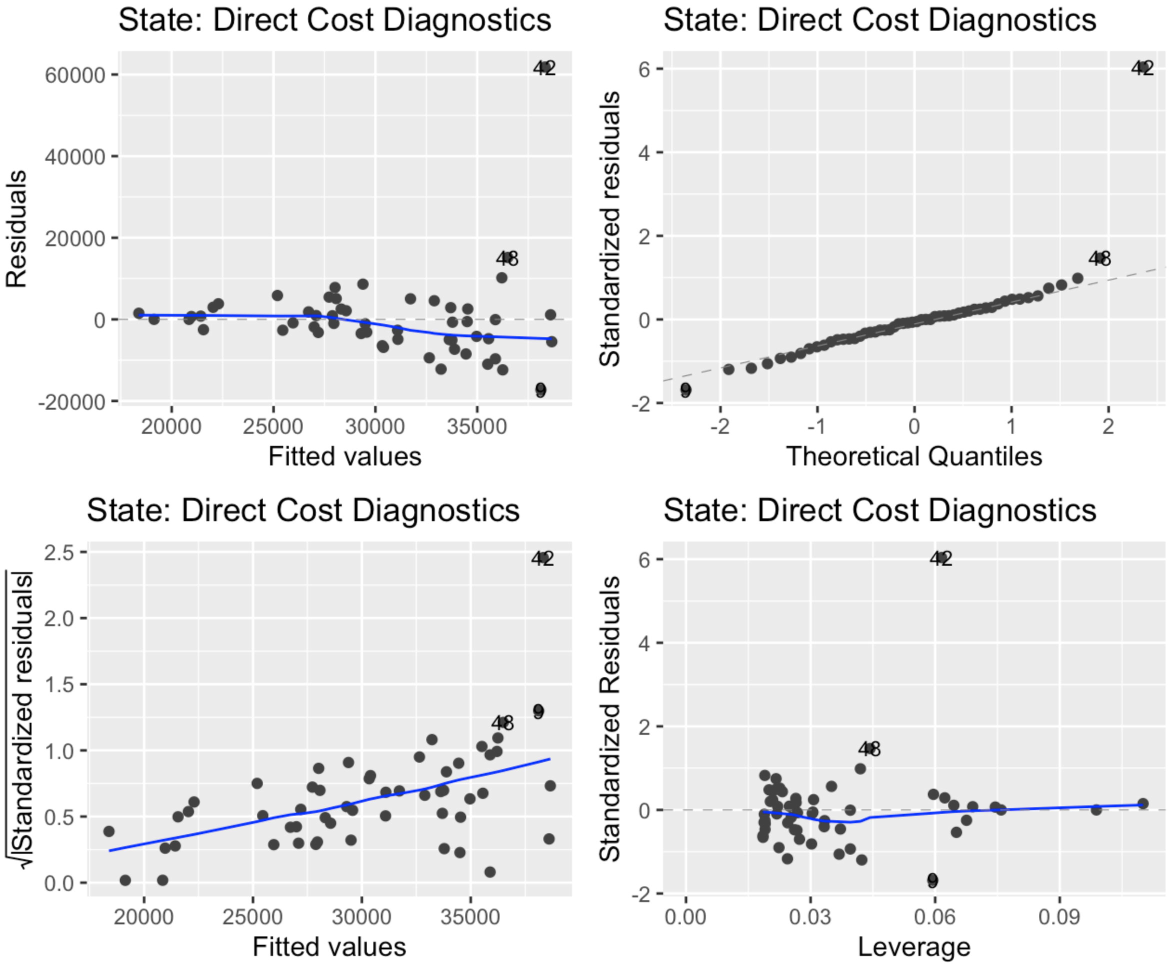




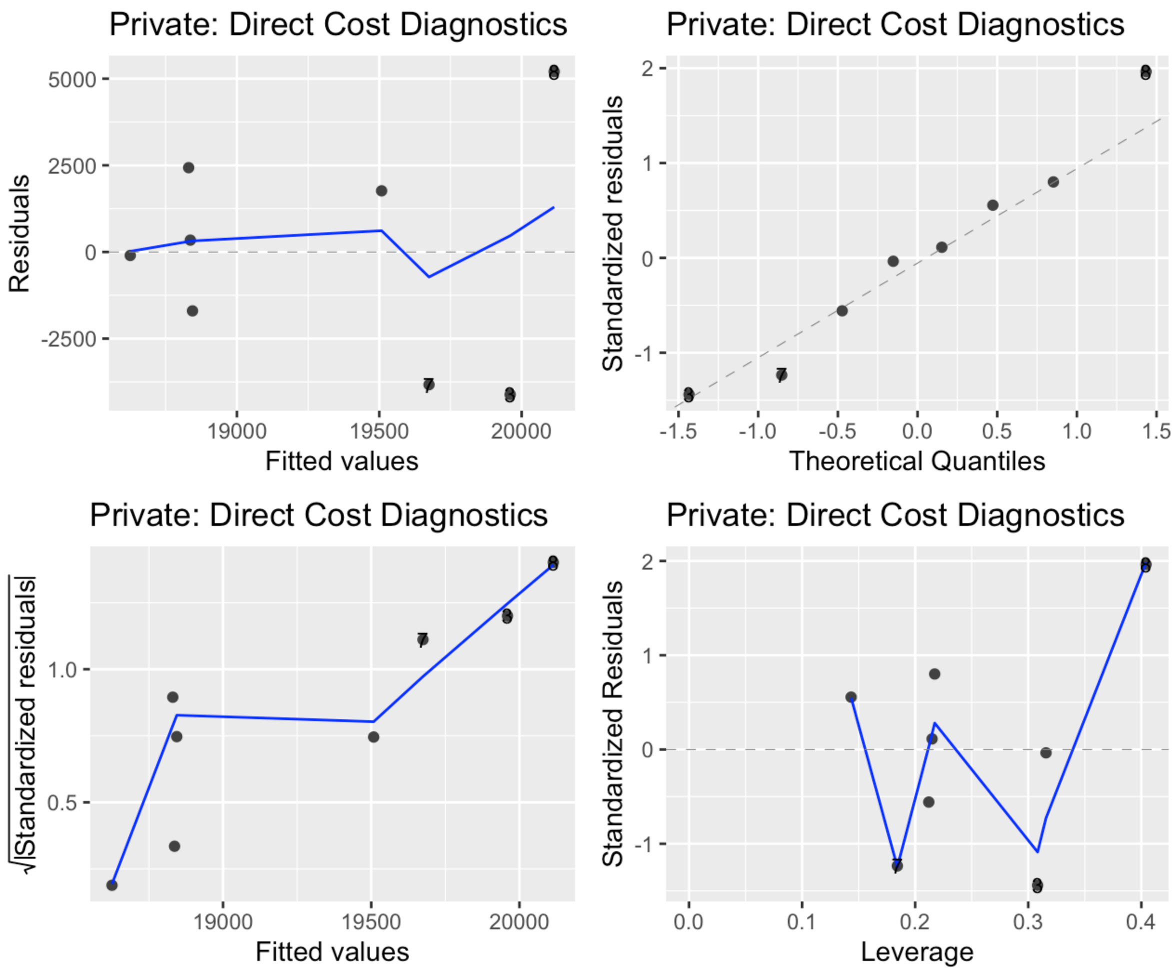




The following plots show how regression for State Data performed. For the **Residuals plot (top left), n**on-randomness suggests a linear relationship may not accurately capture relationship. For the **Q-Q Plot (top right), the d**iagonal suggests a normal distribution which is desired. For the **Scale-Location Plot (bottom left), the relatively f**lat line suggests constant variance which is desired. For the **Residuals vs. Leverage (bottom right), the** inflection points suggest one point has high influence.



The following plots show how regression for Private Data performed. Since there was few points, the plots show that regression may not accurately capture the relationship. For the **Residuals (top left), a** non-randomness plot suggests heteroscedasticity. For the **Q-Q Plot (top right),** a diagonal means normal distribution. For the **Scale-Location Plot (bottom left), a** non-flat plot suggests heteroscedasticity. For the **Residuals vs. Leverage (bottom right),** inflection points suggest one point has high influence.



The coefficients of regression were significant in all of the tested cases:

State: Annual Per Capita Cost - Direct ~ ADP

Coefficients:	Estimate	Std.	Error	z value	Pr(>|z|)
(Intercept)	39088.833	2796.711	13.977	< 2e-16	*** 0.001
ADP	-14.417	3.897	-3.699	0.000522	*** 0.001

State: Annual Per Capita Cost - Indirect ~ ADP

Coefficients:	Estimate	Std.	Error	z value	Pr(>|z|)
(Intercept)	2239.7219	148.8212	15.050	< 2e-16	*** 0.001
ADP	-0.7793	0.2074	-3.758	0.000435	*** 0.001

Private: Annual Per Capita Cost - Direct ~ ADP

Coefficients:	Estimate	Std.	Error	z value	Pr(>|z|)
(Intercept)	1.844e+04	2.265e+03	8.140	0.000185	*** 0.001
ADP	8.407e-01	1.871e+00	0.449	0.668905	1

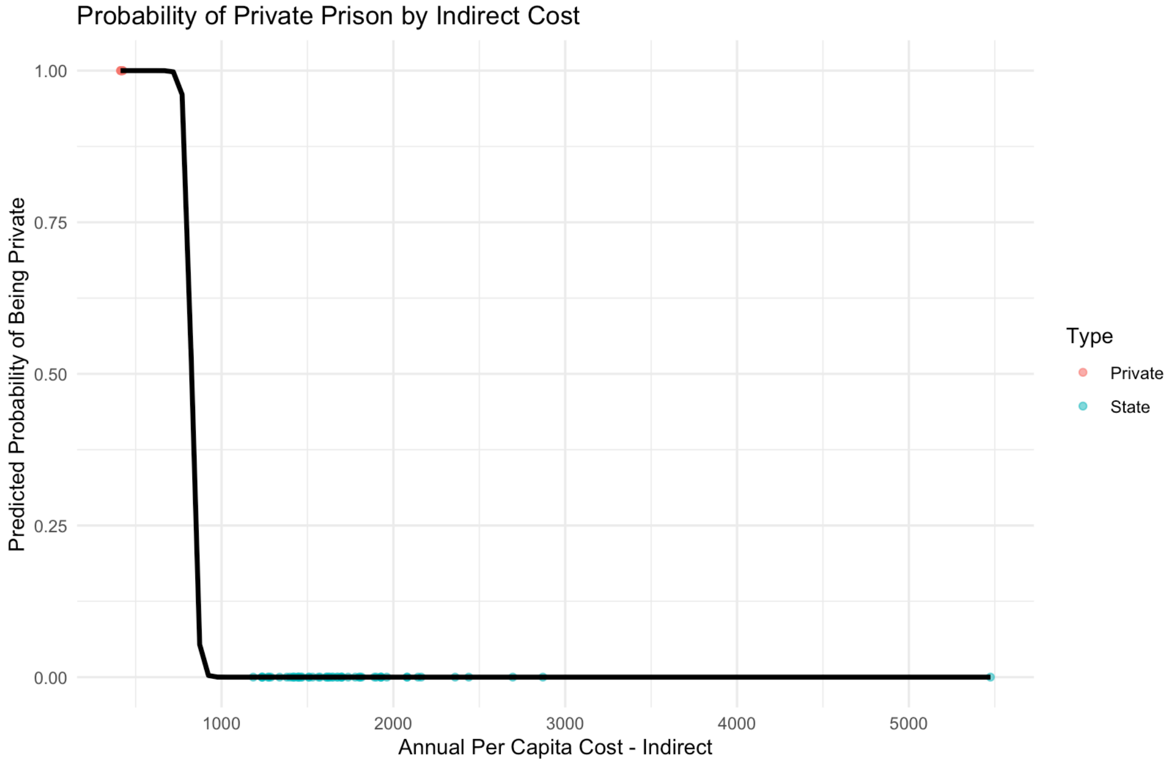
Private: Annual Per Capita Cost - Indirect ~ ADP

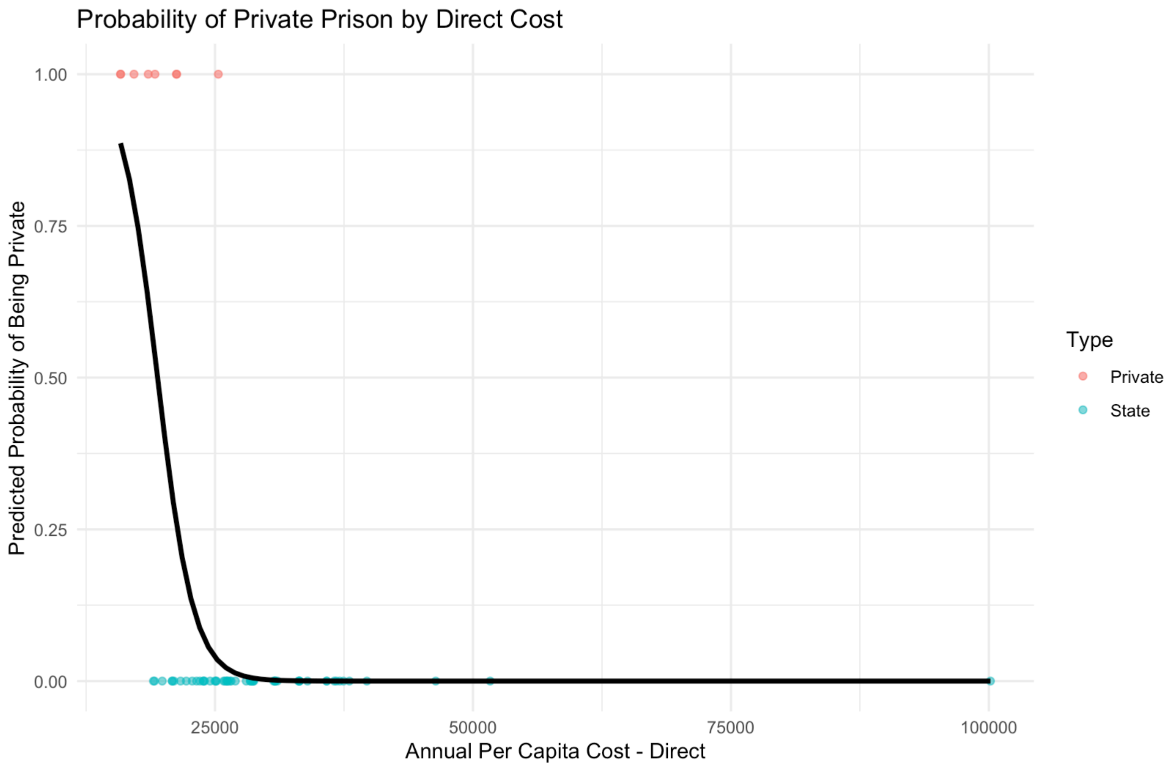
Coefficients:	Estimate	Std.	Error	z value	Pr(>|z|)
(Intercept)	423.756847	2.330667	181.818	1.87e-12	*** 0.001
ADP	-0.005745	0.001925	-2.985	0.0245	* 0.05

Combined Regression

I also combined both datasets and added a column that differentiates between private and state data. The logistic models attempted to classify a prison as Private or State by Annual Per Capita Cost - Direct and Annual Per Capita Cost - Indirect.

These graphs show the fit of the logistic regression models. The model likely overfit due to small amount of data. However, it picked up on general trend, higher per capita costs are incurred by state prisons.

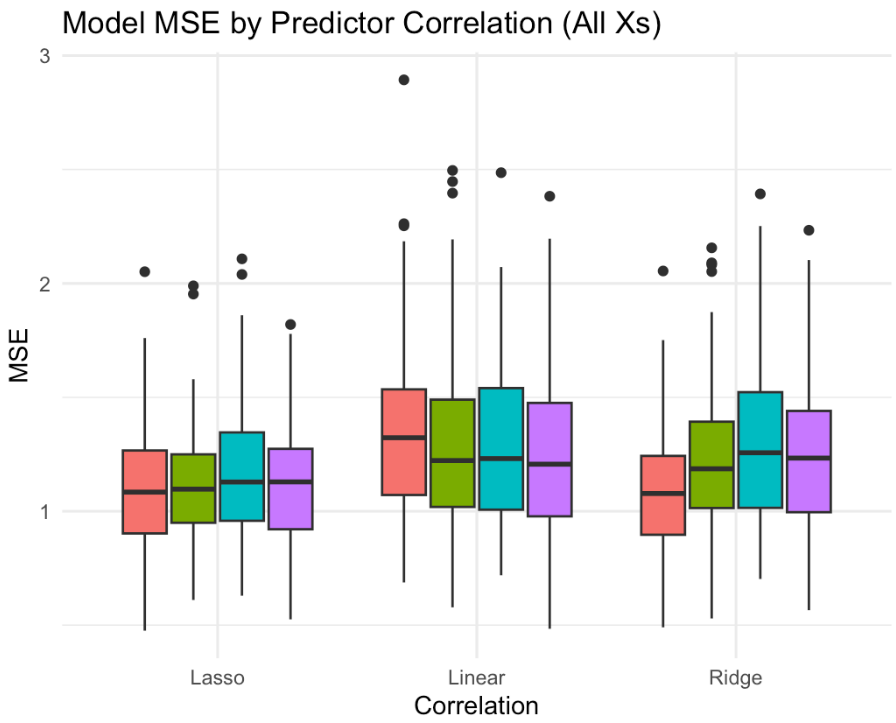


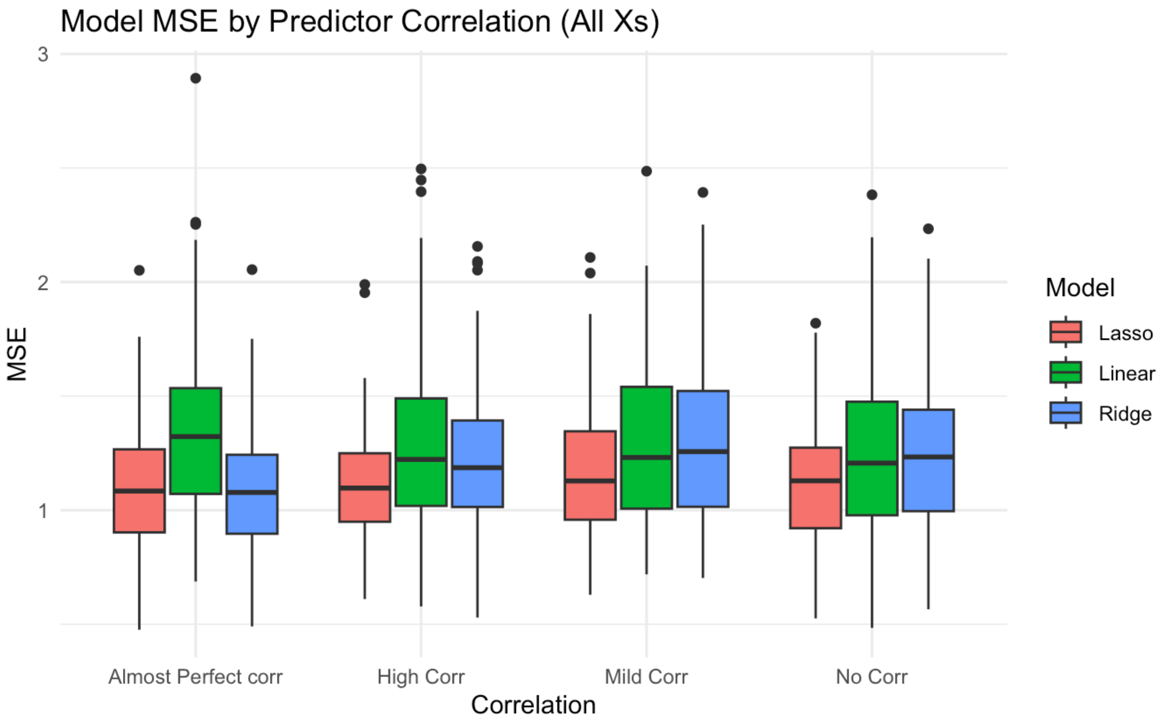


Results were not significant in both cases.

Coefficients:	Estimate	Std.	Error	z value	Pr(>|z|)
(Intercept)	11.1330960	4.1302820	2.695	0.00703	** 0.01
Annual Per Capita Cost Direct	-0.0005730	4.1302820	-2.918	0.00352	** 0.01

Coefficients:	Estimate	Std.	Error	z value	Pr(>|z|)
(Intercept)	4.886e+01	5.668e+04	0.001	0.999	1
Annual Per Capita Cost Indirect	-5.928e-02	5.601e+01	-0.001	0.999	1


Model MSE by Predictor Correlation (All Xs)
3
2
1
MSE
Lasso
Linear
Ridge
Correlation


Model MSE by Predictor Correlation (All Xs)
3
2
1
MSE
Almost Perfect corr
High Corr
Mild Corr
No Corr
Correlation
Model
Lasso
Linear
Ridge

I want to identify if there are significant differences between Linear Regression, Ridge Regression (alpha = 0), and Lasso Regression (alpha = 1) for data that is not correlated, mildly correlated, highly correlated, and almost perfectly correlated.

15 x variables are generated and 1 y variable is generated. The x variables are randomly generated from a multivariate normal distribution with equal pairwise correlation values, 0.10 for no correlation, 0.50 for mild correlation, 0.95 for high correlation, or 0.99 for almost perfect correlation.

y = 1 * x2 + 2 *x3 + 3 * x4 + 4 * x5 + noise

I am estimating if y can be predicted from all 15x's by linear, ridge, and lasso regression. It is replicated 100 times for each correlation and each model. Each model's performance is measured by MSE.

This first plot shows a statistically significant difference between Linear Regression as compared to Lasso Regression and between Linear Regression as compared to Ridge Regression in the Almost Perfect Correlation case. This difference is illustrated in the table below as well. This is the only statistically significant difference in the results. More plots can be created in Project4/project4_simulation_Annapureddy_Githika.qmd

```
    Pairwise comparisons using Wilcoxon rank sum test with continuity correction

High Corr 0.762               -         -
Mild Corr 0.037               0.464     -
No Corr   0.762               0.905     0.464
```

```
    Pairwise comparisons using Wilcoxon rank sum test with continuity correction

       Lasso   Linear
Linear 2.8e-11 -
Ridge  0.00047 0.00175
```

Model	label	Mean_MSE	SD_MSE
Lasso	Almost Perfect corr	1.091134	0.2694847
Lasso	High Corr	1.104686	0.2580040
Lasso	Mild Corr	1.176031	0.3088236
Lasso	No Corr	1.129455	0.2755578
Linear	Almost Perfect corr	1.359378	0.3654802
Linear	High Corr	1.288295	0.3775425
Linear	Mild Corr	1.296692	0.3445425
Linear	No Corr	1.246999	0.3609552

Model	label	Mean_MSE	SD_MSE
Ridge	Almost Perfect corr	1.079490	0.2656926
Ridge	High Corr	1.232065	0.3131289