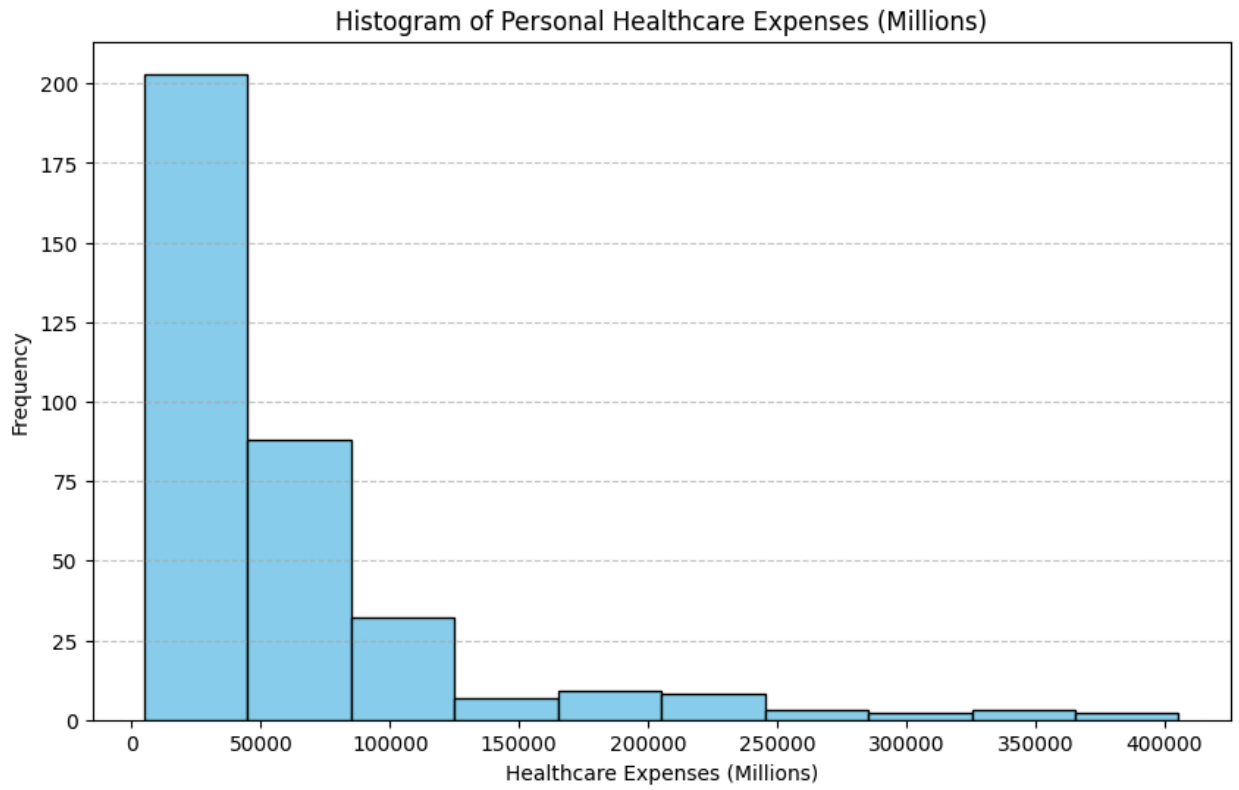
Personal Healthcare Dataset

The unit of observation for this set was a state, with the addition of the District of Columbia, totaling 51 observations with 8 columns. This dataset was modified from the Centers for Medicare & Medicaid Services, and the relevant variables taken were 'Year' and 'Healthcare Expenses.' The second variable is a quantitative measure of the net total of personal healthcare expenditures per state from the years 2014 to 2020, inclusive. There was also an additional column depicting the average expense over the range 2014-2020, which was dropped for the purposes of this experiment. Below are the basic statistics and distributions of these expenses, in millions of dollars.

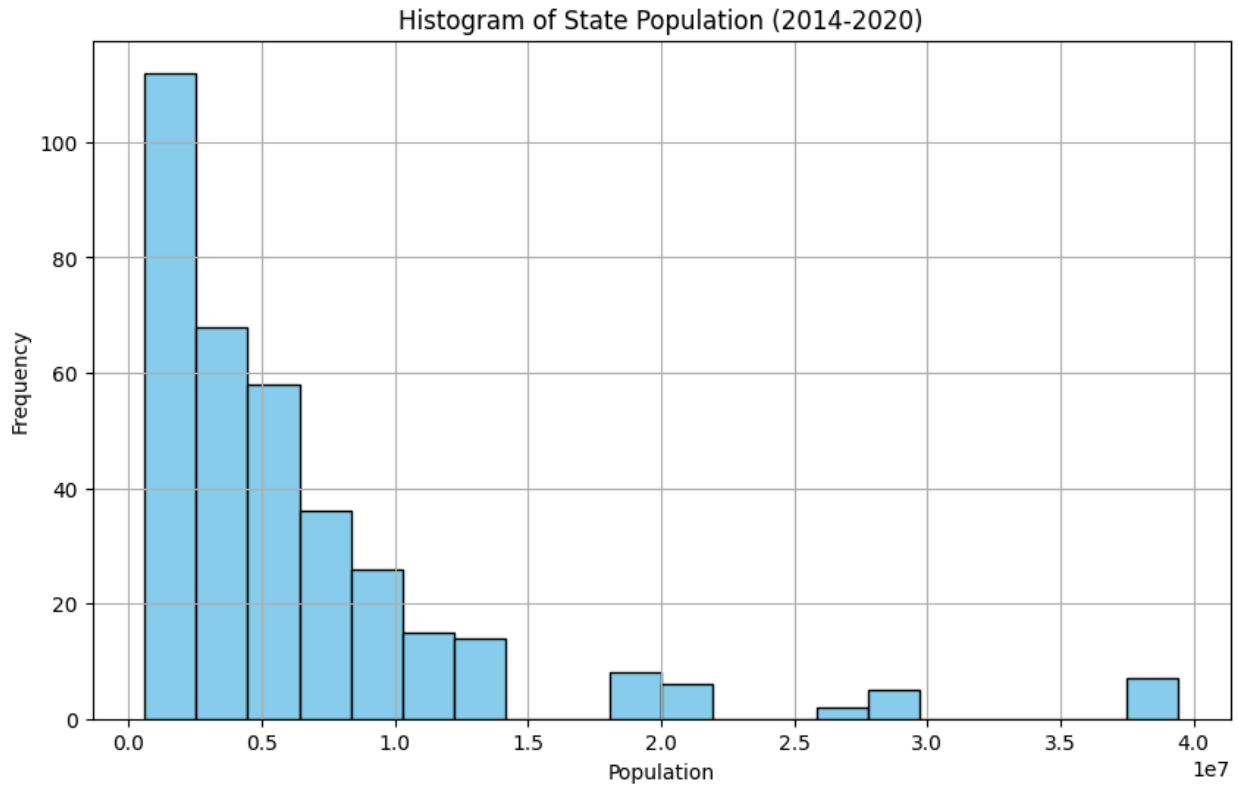
Count	357
Mean	57,301.82
Standard Deviation	64,996.34
Minimum	4822
25th Percentile	15,391
Median	38,322
75th Percentile	68,253
Maximum	405,451



State Populations Dataset

The unit of observation for this set was a state, with the addition of the District of Columbia, totaling 51 observations and 8 columns. This dataset was acquired from the United States Census Bureau and depicts the state populations as reported to the Census from the years 2014 to 2020, inclusive. Thus, the relevant variables were 'Year' and 'Population,' the quantitative variable for this dataset. Below are the descriptive statistics for the distribution of state populations over the years 2014 to 2020.

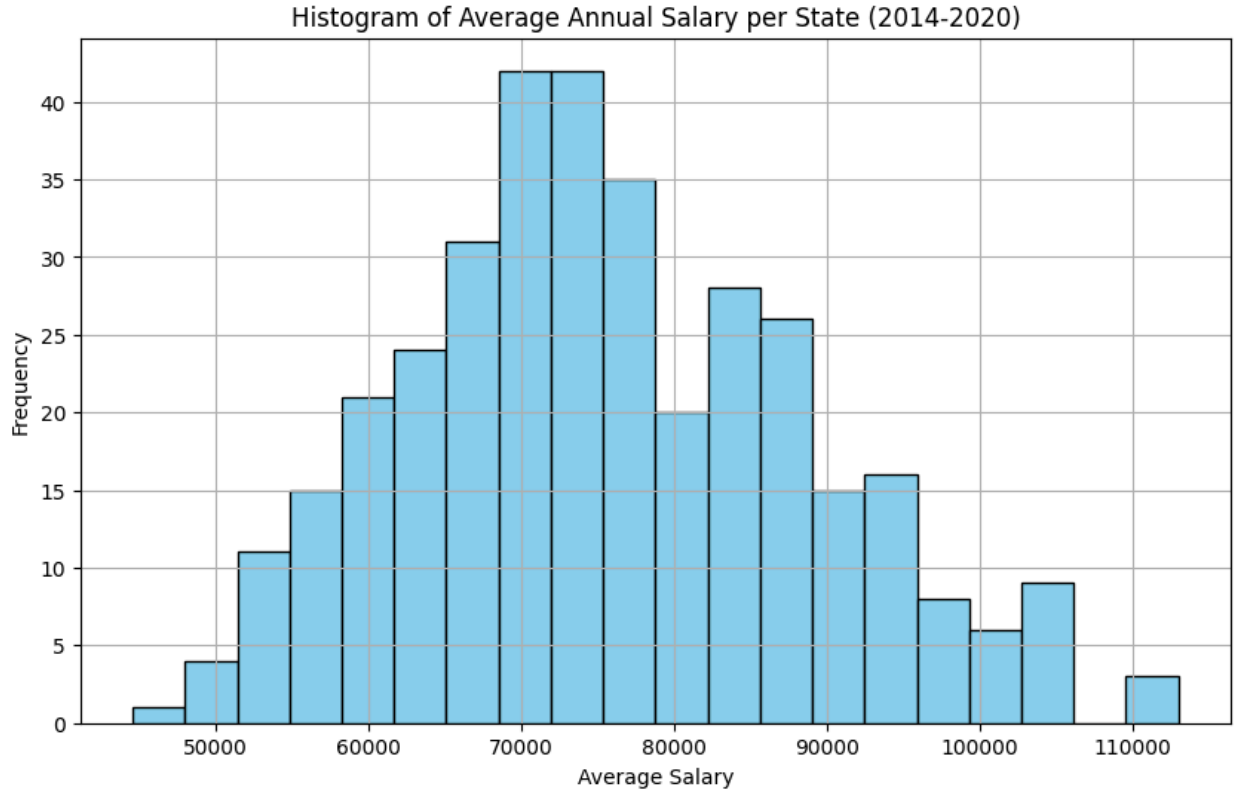
Count	357
Mean	6,364,066
Standard Deviation	7,203,519
Minimum	579,054
25th Percentile	1,784,787
Median	4,455,590
75th Percentile	7,299,961
Maximum	39,437,610



Average Annual Salaries Dataset

The unit of observation for this set was a state, with the addition of the District of Columbia, totaling 51 observations and 8 columns. Each column represents a year 2014 to 2020 inclusive, and originally had the years 2021-2023 for potential prediction in the analysis plan. However, it was dropped because the predictors would not have added much value to this project's conclusions. The relevant variables for this dataset are 'Year' and 'Average Salary,' which is a monetary average annual salary per state as reported by the Federal Reserve of Economic Data (FRED). This quantitative variable has many descriptive statistics as listed below.

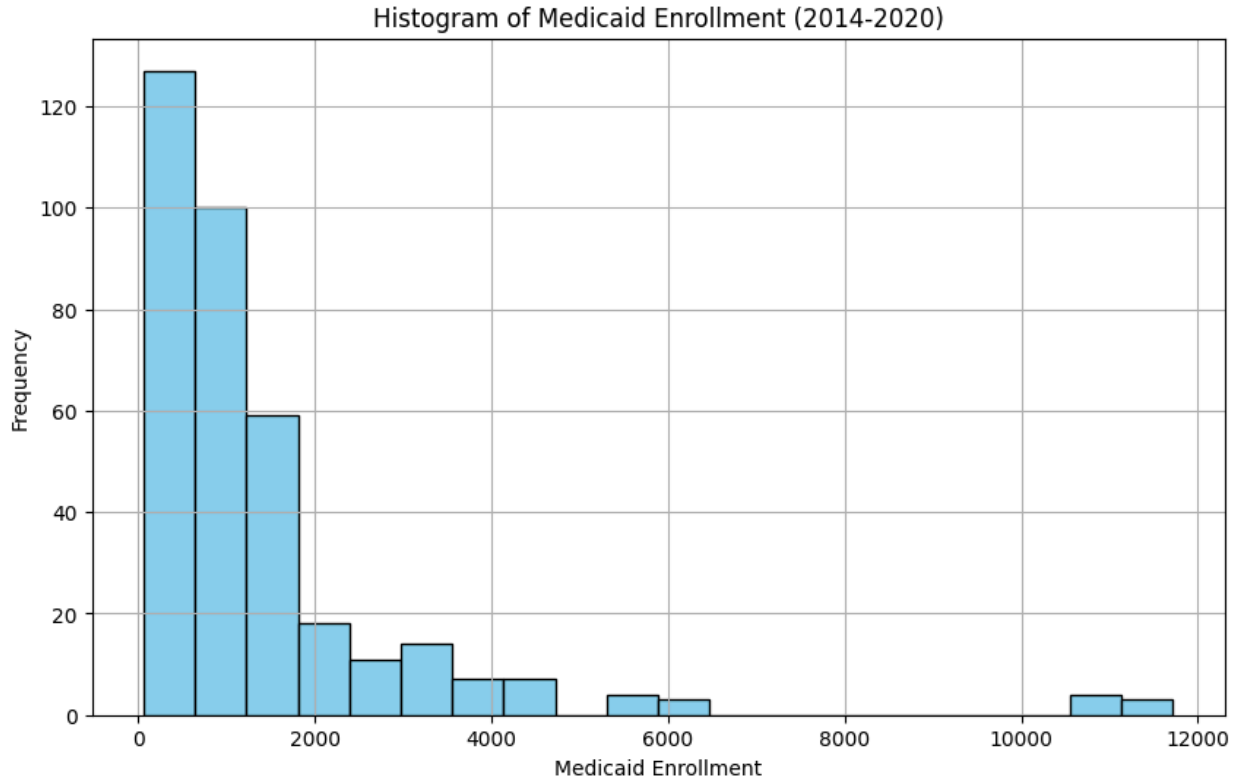
Count	357
Mean	$75,698.77
Standard Deviation	$13,020.94
Minimum	$44,590.00
25th Percentile	$66,650.00
Median	$74,370.00
75th Percentile	$84,900.00
Maximum	$113,000.00



Medicaid Enrollment Dataset

The unit of observation for this set was a state, with the addition of the District of Columbia, with 51 observations and 8 columns for years 2014 to 2020, inclusive. The relevant variables for this dataset are 'Year' and 'Medicaid Enrollment,' which was reported directly from the Centers for Medicare & Medicaid Services. The enrollment variable is a quantitative one, whose descriptive statistics and distribution information is listed below.

Count	357
Mean	1421
Standard Deviation	1812
Minimum	57
25th Percentile	301
Median	1034
75th Percentile	1616
Maximum	11,718



Results Dataset

The unit of observation for this set was a state, with the addition of the District of Columbia, which adds up to 51 observations and 8 columns. This is the output dataset from the ARIMA model that was used in the analysis plan. It uses all four input datasets as listed above to ensure a robust conclusion can be made about the impact of Medicaid expansion policy adoption on personal healthcare expenses. A new dataset was made, `cost_over_salary`, with a formula as follows:

$$Cost = \frac{Personal\ healthcare\ expense\ per\ state * 1{,}000{,}000}{State\ population\ for\ that\ respective\ year} \div State's\ annual\ salary * 100$$

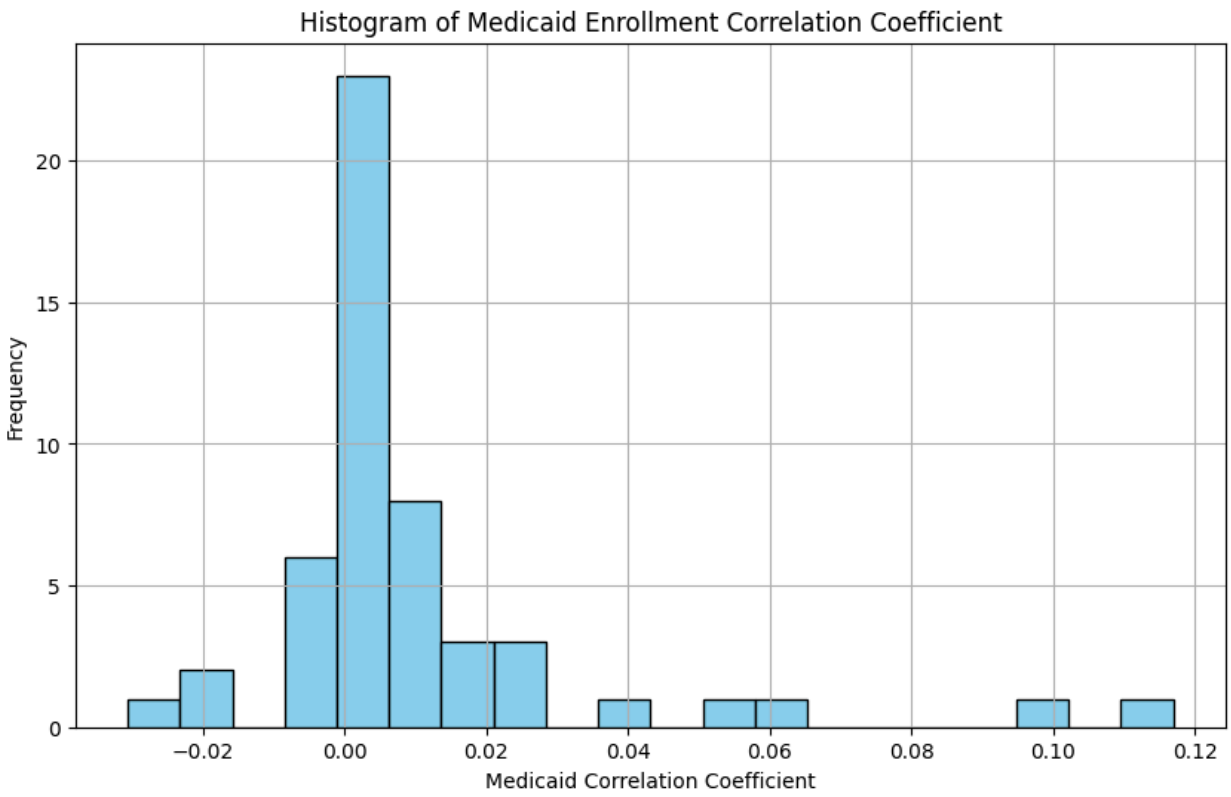
This formula calculates the personal healthcare cost per person as a percentage of salary by taking the net total personal healthcare expenses per state (in millions) divided by the state population, divided again by the state's average salary for that year. This also operates under the assumption that the personal healthcare expenses are equal for every person in the state, which, in reality, is not the case but is sufficient for the purposes of this exploration.

For the ARIMA model implementation, a three-year moving average was first conducted to 'smooth out' the percentages of healthcare costs to reduce noise. Then, the ARIMA model, created with ARIMA(endog=[per capita healthcare], order=(1,1,1), exog=[Medicaid enrollment]) was used to fit the respective data, which output seven coefficients and values. The most significant of this was the Medicaid coefficient, for which the descriptive statistics are listed below.

Count	51
Mean	0.0098
Standard Deviation	0.0257
Minimum	-0.0306
25th Percentile	-0.0004
Median	0.0017
75th Percentile	0.0098
Maximum	0.1169

It also consists of AR (auto-regressive) and MA (moving average) coefficients and p-values, as well as the Akaike Information Criterion (AIC), which was used to determine systematic impact of enrollment on healthcare costs. Because these coefficients and values were not external predictors like the Medicaid correlation coefficient, they were not used in the final analysis and in conclusions for this experiment.

It is important to note that a positive coefficient denotes that if enrollment numbers increase, healthcare costs per person will increase; otherwise, higher Medicaid enrollment is associated with lower healthcare costs per person.


Histogram of Medicaid Enrollment Correlation Coefficient
Frequency
0
5
10
15
20
−0.02
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Medicaid Correlation Coefficient