Summary

Every state in the United States had various information about its programs and safety recorded in order to determine what may be the cause behind the different numbers and rates of injuries, illnesses, and fatalities. We also attempted to see if there are multiple different factors involved and, if so, which may prove to be more impactful. By obtaining raw data, population numbers for every state, and calculating per capita values of injuries/illnesses as well as fatalities, we were able to reach certain conclusions for questions that we were presented with. By utilizing bar charts, pie charts, and scatterplots, we could visualize these results to help show up the full picture.

The state and federal program types were compared in terms of their fatalities to see which one proved safer. When comparing using raw number of fatalities, federal programs had approximately 1000 more fatalities than the state program and accounted for ~60% of total fatalities. Using the fatality rate per 100,000 people, the federal program still accounted for ~60% of the total fatalities per 100,000 people, however the gap between the two programs shrunk to approximately 20 per 100,000 people.

When comparing the raw number of injuries/illnesses of states with the state program, California had by far the highest number with 345,400. However, when taking per 100,000 into account, California was in the bottom 25% of states; Vermont became the state with the highest number of injuries/illnesses at 1580.558 per 100,000 people.

Comparing the average number of years it took to inspect a workplace once with the number of fatalities showed a positive, extremely weak relationship. This was shown in the resulting R value of 0.12, and the resulting p-value of a regression test with a significance level of 0.05 was 0.397; this means we would reject the null hypothesis and conclude that there is no significant correlation between the two variables.

Business Problem

There were three problems we were attempting to answer with this project:

1. Which program had the highest rate of fatalities: state or federal?
2. Which state had the highest number of injuries and illnesses under a state program?
3. Is there a relationship between the rate of fatalities and the average number of years it took to inspect each workplace once? If so, what is the relationship between them?

Data

The data records the number of fatalities (in 2012), injuries/illnesses (in 2012), financial penalties (in 2013), inspectors, and years to inspect each workplace once for each state. It also calculated the rates of fatalities and injuries/illnesses as a percentage. The states are then ranked based on their number of fatalities as well as their penalties. They are also categorized by their program type: state or federal.

Methods

To compare the program with the highest rate of fatalities, the number of fatalities of each state under both programs were added up under their respective program and then compared to one another. I decided to use the “Number of Fatalities” column instead of the “Rate of Fatalities” as the latter is measured as a percentage, and since each state has a varying population, adding up the percentages does not correctly represent the data and may lead to inaccurate results; Texas and California have a might higher population compared to Hawaii and Wyoming, so it's not exactly a one-to-one comparison. A bar chart was then used to visualize the difference.

In order to determine the state under a state program with the highest number of injuries and illnesses, the states using a federal program were filtered out and the remaining states were graphed against each other using bar and pie charts via their number of injuries and illnesses.

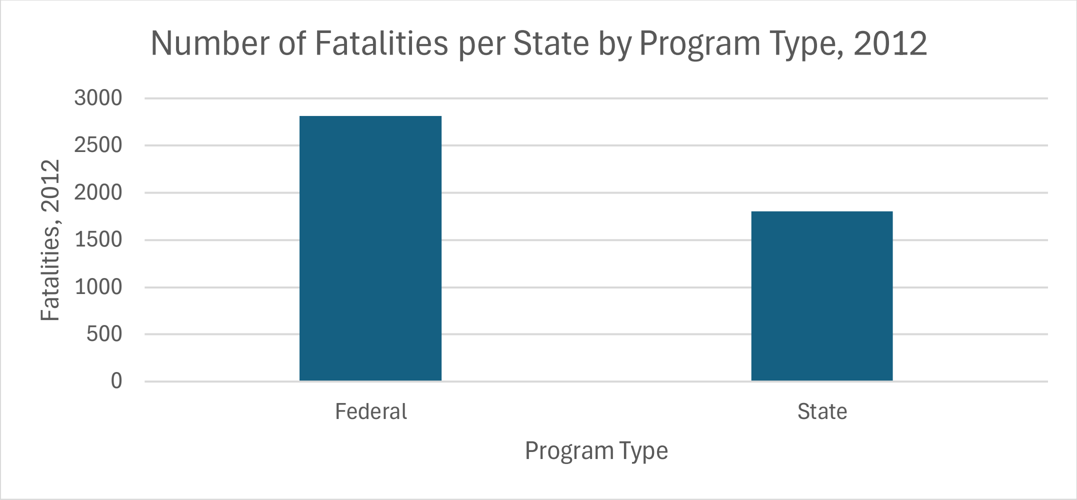
In order to determine the relationship between "Average Number of Years to Inspect Each Workplace Once" and "Rate of Fatalities," a scatterplot was created, the line of best fit was measured, and the R value was calculated. In order to ensure validity of the results, a regression test was also performed to determine if the R values matched.

One issue did present itself, however; while opting to not use the "Rate of Fatalities" for reasons of mathematical inaccuracy was valid, the use of the raw number of fatalities and injuries/illnesses for the first two questions also proved to be biased for the same reason. As mentioned previously, different states have varying population sizes, and as a result, states with a much higher population (in the tens of millions) will naturally have a higher number of injuries, illnesses, and fatalities than those with a smaller population (in the hundreds of thousands). In order to account for this, I decided to run an two additional tests to make sure the data was accurate, but this time I would be using the number of injuries/illnesses and fatalities per capita (or for more clean numbers, per 100,000 people) to create a more accurate comparison.

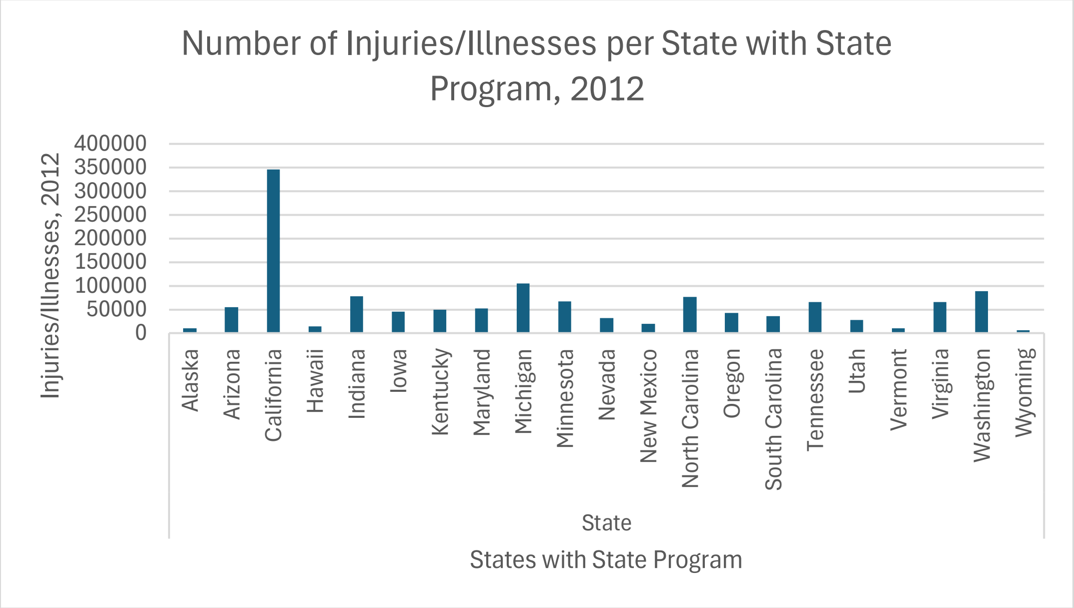
In order to do this, I first harvested data of state populations from the U.S. Census Vintage data estimates from 2012. Then, I divided that number from both the "Number of Fatalities" and "Number of Injuries/ Illnesses" in order to obtain the number of fatalities and injuries/illnesses per capita respectively. These numbers were extremely low (10^{-5} and 10^{-3} respectively), so I multiplied all the results with 100,000 to achieve cleaner numbers. I then redid the process for the first two questions, only now I am using the new "per 100,000 people" data set that I have created. The same visualizations were used as the previous examples.

Results

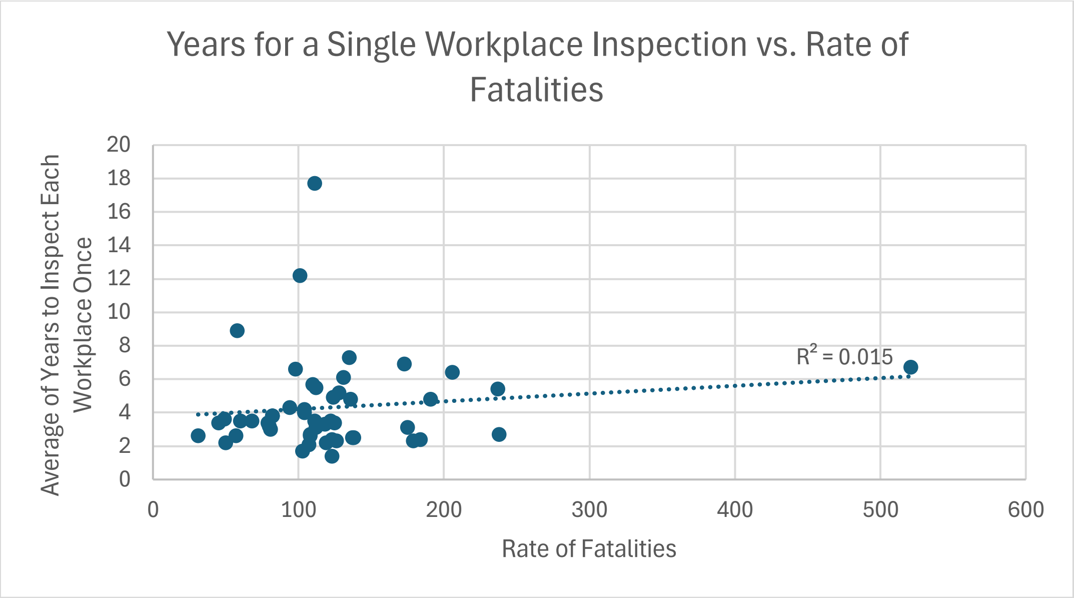
When using the raw number of fatalities, it was determined that the program with the highest rate of fatalities was the federal program; the federal program had a total of 2,814 fatalities while the state program had 1,803 total fatalities.

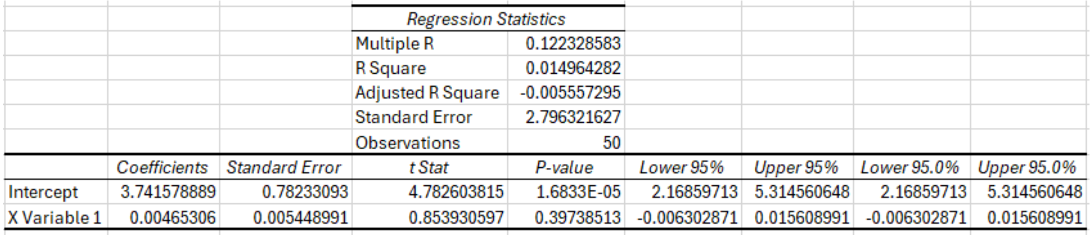


When using the raw number of injuries/illnesses, the state under a state program with the highest number of injuries and illnesses was California with a total number of 345,400.



The R value that was calculated between the “Average Numbers of Years to Inspect Each Workplace Once” and “Rate of Fatalities” was 0.12. This shows a positive, yet extremely weak correlation between the two variables. The regression test at a significance level of 0.05 also resulted in an R value of 0.12, as well as a p-value of 0.397, therefore we would reject the null hypothesis and conclude that there is no significant correlation between the two variables.

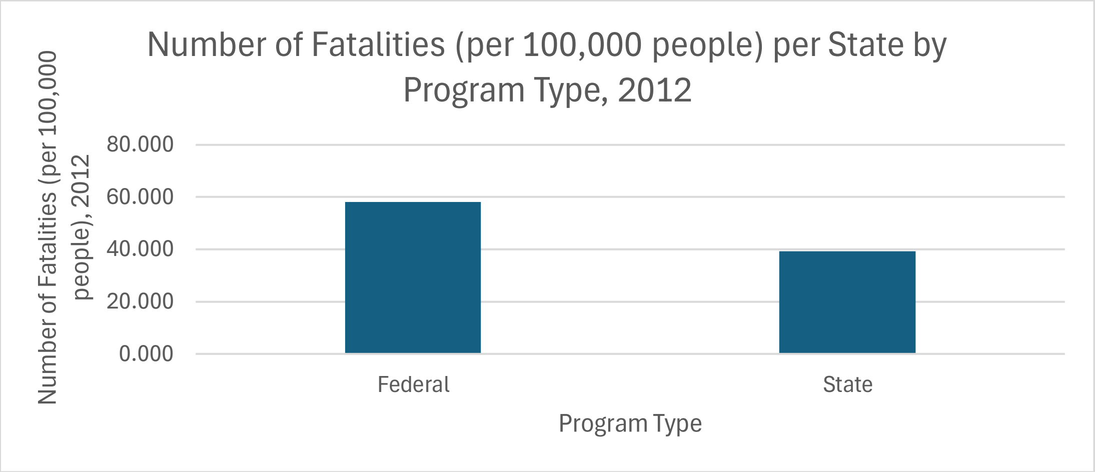




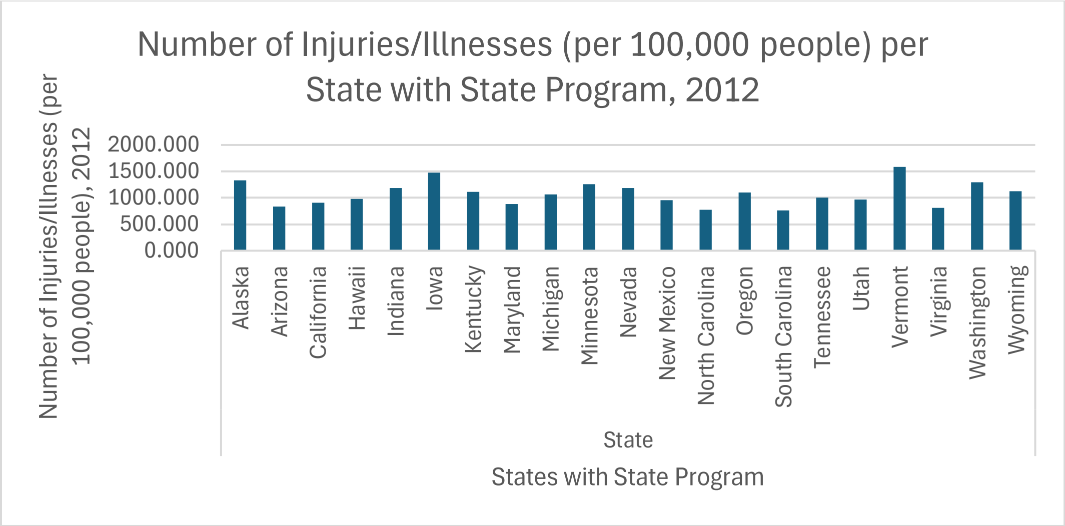
Regression Statistics	
Multiple R	0.122328583
R Square	0.014964282
Adjusted R Square	-0.005557295
Standard Error	2.796321627
Observations	50

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	3.741578889	0.78233093	4.782603815	1.6833E-05	2.16859713	5.314560648	2.16859713	5.314560648
X Variable 1	0.00465306	0.005448991	0.853930597	0.39738513	-0.006302871	0.015608991	-0.006302871	0.015608991

When looking at the program with the highest number of fatalities per 100,000 people, the federal program still had the highest value of 58.251 fatalities per 100,000 people versus the state program's 39.353 fatalities per 100,000 people.

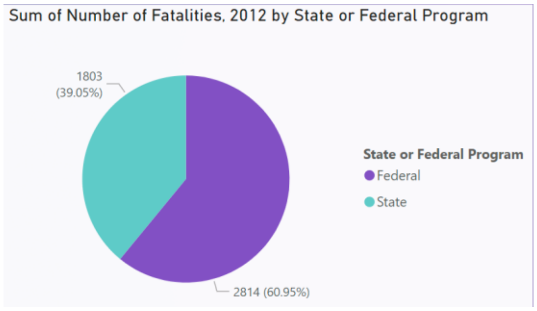


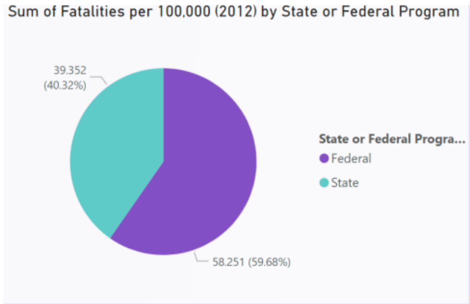
When looking at the injuries/illnesses per 100,000 people for all the states with a state program, Vermont had the highest number with 1580.558 injuries/illnesses per 100,000 people.



Conclusions

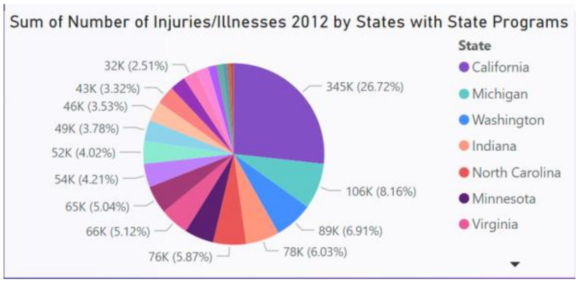
In regards to the first question, 7 of the 8 largest states in the U.S. by population – Texas, New York, Florida, Illinois, Pennsylvania, Ohio, and Georgia – are using the federal program. As a result, that program would naturally have a higher number of fatalities as the amount of people in these states alone (combined total of 112,067,021) make up 35.7% of the entire country's population (313,877,662) and almost match the entire population of the states with a state program (total population of 130,597,060). For this reason, measuring and comparing the programs in regards to fatalities yields similar results whether using just raw numbers or numbers per 100,000 people. The split between federal and state programs when in regards to fatalities was 60.95% to 39.05% when using raw numbers, while it was 59.68% to 40.32% when using numbers per 100,000 people.

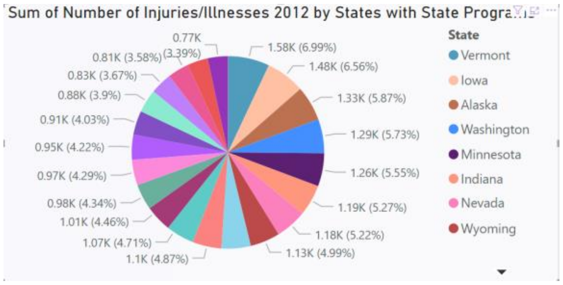




However, it is important to note that the difference in values between the federal and the state programs is also very important. Without the added context of population density, an approximate 1000 increase in fatalities for the federal program seems extremely high, and might convince the one to opt for the state program. With the knowledge that the federal program accounts for two-thirds of the entire country's population versus the state program's one-third market share, that value begins to make more sense as to why it is seemingly so high. Finally, when looking at the fatalities per 100,000 people, that number shrinks drastically from approximately 1000 fatalities to only about 20, making the federal program seem much more appealing than it once did.

The second question truly demonstrates how important the difference between using raw numbers and per capita numbers is. Similar to the first question, the only state of the eight highest populated states with a state program would naturally also be the state with the highest number of injuries and illnesses compared to other states with the state program. California is also unique as it is the state with the highest population in the entirety of the U.S. California's number of injuries/illnesses was more than triple that of second place, Michigan, who had a total number of 105,500. This makes sense when you consider that California's population (37,944,551) is about quadruple that of Michigan's (9,898,289). Digging even deeper, it becomes apparent that California's number of injuries and illnesses is almost as much as the entire population of Wyoming (576,656). However, both states had the same rate of injuries and illnesses in 2012 (3.5%), showing that raw numbers alone do not account for population density. When taking per 100,000 numbers into account, California falls from the extremely uncontested number one spot to 16th out of 21 states. One who looks at raw numbers might think California is insanely dangerous, especially considering how much higher its results were even compared to the second place spot, but when looking at per 100,000 numbers, California doesn't even register as something important to consider. Additionally, the results for the per 100,000 analysis have a might tighter range, as opposed to the obvious outlier that the raw numbers had shown.


Sum of Number of Injuries/Illnesses 2012 by States with State Programs


Sum of Number of Injuries/Illnesses 2012 by States with State Progra...

Future improvements of the project could see its scope expanded to encapsulate the availability, accessibility, and affordability of healthcare in the states in order to get a more complete understanding of the causes behind a state's injuries, illnesses, and fatalities. Certain states were massive outliers when it came to their rates of fatalities, particularly North Dakota (17.7%) and Wyoming (12.2%). Healthcare obviously plays a vital role when it comes to health, including injuries, illnesses, and more vitally, fatalities. Seeing if, and to what extent, healthcare plays a role in the fatalities in each state may show that the programs it adopts is not as important to the overall health of its population.

Lastly, for future improvements of the project, I would make sure to record data for the number of (and as a result, the rate of) injuries and illnesses for certain states that currently have no information available. All these missing information involves states under the federal program, meaning we cannot analyze the injuries and illnesses of that program. In turn, it is impossible to compare it to the injuries and illnesses of the state program, meaning that vital question cannot be answered. Collecting this data in the future should be prioritized so that that avenue of study can be explored.