## Summary

[Brief overview of your project]

## Business Problem

[Description of the business problem your project addresses]

## Data

[Description of the dataset used in your project]

## Methods

[Description of the methods and techniques used in your project]

## Results

[Summary of the results obtained from your project]

## Conclusions

[Conclusions drawn from your project and recommendations for future work]

**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.

**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.

**Results**

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.

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 also resulted in an R value of 0.12.

**Conclusions**

While our data seems very conclusive, there is one fatal flaw when it comes to the results for our first two questions: the use of raw numbers alone, as opposed to per capita values, skews the data heavily. 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 regards to the first question, 7 of the 8 largest states in the U.S. by population are using the federal program. As a result, that program would naturally have a higher number of fatalities than the total number of fatalities under the state program. Interestingly, with a much greater number of individuals under the federal program, the approximately 1,000 higher number of fatalities suddenly seems much smaller than when looked at without the added context.

Similarly, in regards to the second 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 (~39 million) is almost quadruple that of Michigan’s (~10 million).

In addition to this, California’s number of injuries and illnesses is almost as much as Wyoming’s entire population (~581,000). Wyoming ranked as the state with a state program and the least number of injuries and illnesses (6,500). However, both states had the same rate of injuries and illnesses in 2012 (3.5%). This is why additional columns in the data set should be added for both injuries/illnesses and fatalities per capita, as that gives a more accurate representation of those values in comparison to the population without having to deal with the mathematical issues that are presented with rates and percentages.

In addition to this, I would like to expand the scope of the project 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 may prove to have a massive impact on a state’s ability to reduce the number of injuries, illnesses, and fatalities, and it is a factor that is definitely worth studying.

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. Since all those states without data fall under the federal program, it is currently not possible tell a full story of the injuries and illnesses of the federal program and, as a result, we cannot compare it to the states with a state program. This missing information may help explain a crucial aspect of the results we are looking for that we do not currently have.