**Colonials Project Final Report**

December 13, 2014

**Team Members**

Awilda Lopez

Faith Bradley

Bala Venkatesan

Rebecca Bilbro

**Project Trajectory**

Originally we set out to make a tool that would characterize the relationships between different sociocultural factors. Could factors like educational attainment, income levels, and housing prices be used as leading or lagging indicators for other factors like violence or occupational fatality rates? After our initial attempts to find statistical correlations failed to establish any compelling relationships, we decided to restructure the project, focusing instead on teasing out subtleties from a single, complex dataset, and individually testing different hypotheses with the shared data.

We believe this trajectory ultimately resulted in a much more meaningful learning outcome for the group members. The capstone project been a good opportunity for us to learn not only how to code in Python, collect and clean data, and tell stories, but also about effective ways of dividing large data science projects into individual components that can be coded independently and then aggregated.

**Problem Statement: A Plateau in Workplace Safety**

In the first three decades after the Occupational Safety and Health Administration was created in 1971, workplace fatalities dropped more than 65%, even as US employment doubled. But since then, the drop-off in fatalities has reached a plateau. The annual fatality count has hovered within the same narrow range for many years (4,500-5,500).

Worker deaths are down from about 38 a day in 1970 to 12 a day in 2013. OSHA would like to see that number continue to decrease. However, OSHA's budget is small and not likely to increase, so the agency can only do so much outreach, training, enforcement, and setting of new regulations. How can the agency make the most of its limited resources?

**Motivation**

Use OSHA data to build a tool that will help the agency discover more information hidden within the data and develop targeting schemas for outreach, training, enforcement, and regulation.

**Data**

The main sources of data were the OSHA website ([http://www.osha.gov](http://www.osha.gov/)) and the Department of Labor Enforcement Data Catalog (<http://ogesdw.dol.gov/>). These data were supplemented with US population data from [www.census.gov](http://www.census.gov) and unemployment rate data from the Bureau of Labor Statistics ([www.bls.gov](http://www.bls.gov)).

The primary data set is a spreadsheet that catalogs every workplace fatality and catastrophe reported to OSHA since 1984. Employers are required by law to report the workplace-related death of any employee to OSHA within 8 hours of the event. Employers are also required to report within 24 hours the hospitalization of three or more employees (termed by OSHA as a ‘catastrophe’). Catastrophes represent a small minority of the data. The original dataset includes the event timestamp, local OSHA area office code (but not state), industry code, description of the event, associated keywords, and whether the event was a fatality or a catastrophe. The data comes from an internal Oracle-based database. While the data is pre-scrubbed to exclude protected personal information (PPI), it reflects many irregularities and required several iterations of cleaning by the group.

**Theoretical underpinnings**

There is a fair amount of data analysis that has been done by academics and advocacy groups (e.g. [AFL-CIO](http://www.aflcio.org/Issues/Job-Safety/Death-on-the-Job-Report)) with publically available OSHA data, as well as many studies conducted internally by OSHA and by the Bureau of Labor Statistics. These reports and studies were used to generate the below hypotheses.

**Methodology**

Test four hypotheses that could account for or illuminate the recent plateau in the decline of occupational fatalities

1. Seasonality – If many of the most hazardous jobs are seasonal, could an increased incidence of severe weather events (more heat waves in summer, bigger snow storms in winter) be masking safety and health improvements?

2. Industrial change – As hazards decrease in some traditional industries (e.g. traditional manufacturing), are more workers being killed in new industries (e.g. green energy)?

3. Impact of the recession - Are there fewer fatalities in construction because of the housing market collapse? Has the recession depressed the number of worker fatalities, artificially diminishing the impact of hazardous workplaces?

4. Variation by state or area – The plateau could be masking significant increases in fatalities in certain parts of the country as incidents drop off in other places.

**Results**1. Seasonality – Occupational fatalities and catastrophes are seasonal, with the most occurring in the summer and the fewest occurring in winter.

2. Industrial change – Certain industries are more dangerous than others. Based on this dataset, we found that the most dangerous appear to be: roofing, power line construction, electrical work, and carpentry.

3. Impact of the recession – The recession has had a significant impact on the construction industry. We found that the unemployment rate is inversely correlated with the number of building permits issued for residential construction projects (r= -0.6).

4. Variation by state or area – Fatalities and catastrophes do vary widely by state, though generally in proportion to the population size. We also found some evidence that decreases in certain regions are correlated with increases in other regions over the last 30 years. Most notably, as the number of workplace fatalities and catastrophes has decreased in the Southeast (defined as the cluster of Alabama, Georgia, South Carolina, and Virginia), incidents seem to have increased in the Northeast (Connecticut, Massachusetts, New Hampshire, and New York), with an rof about -0.72.

**Conclusions and Next Steps: Modeling**

Our initial investigative analysis suggests several avenues for future analysis as well as predictive modeling projects.

1. Seasonality – Moving forward, the seasonality analysis could be done per year by region and/or state. With a more granular analysis, one might be able to identify if there is any correlation between workplace fatalities and extreme weather events or seasons. One event that would be a great test to run against this data would be 2012’s Hurricane Sandy or the Ice Storm in Atlanta, GA that occurred in February of 2014. A seasonality predictive model could allow OSHA to forecast future fatalities due to extreme weather by state. With this forecast the agency would be able to allocate proper budgets and plan ahead for preventative workplace training. Many industries would benefit on all future data that could be created with the data used in this project.

2. Industrial change – Moving forward it would be useful to look at how these hazardous industry rankings have changed over time. It would be interesting to look at relationships between these industries – perhaps looking for a cross correlation between and increase in fatalities in one industry and a decrease in another industry.

3. Impact of the recession – If would be interesting to expand this study to break things down industry – particularly by manufacturing and construction, and to look at the impact of previous recessions on these industries. Given the correlation that we did find, we could look at ways in which OSHA can pre-empt fatalities, with the knowledge that as unemployment rates decline, workplace fatalities are bound to increase.

4. Variation by state or area – We have enough data to break down these state incidents on a more granular level, by individual local area offices. It would be interesting to look at in-state variations, particularly over time. Even more interesting would be to incorporate county business pattern and population statistics to identify trends, and create a predictive model that anticipates fatality rates based upon industry and population growth. This kind of project could enable OSHA to better allocate resources across the regions in accordance with need, and even embark on redistricting that would ensure that each area office is grouped around a nexus of predicted hazardous activity.

**For More Information**

Read more about our project at <http://nbviewer.ipython.org/github/Colonials/capstone/blob/master/osha/index.ipynb>.

Download the files from GitHub at <https://github.com/Colonials/capstone>.