# Background information

## Understanding the variables

* Where is the variable breakdown document?
* What are the inspection types?
* What are the inspection scopes?
* What is violation type?
* What is hazcat?
* Mens versus women’s facilities?

## Why was this facility examined?

* Was there a union present?
* Check inspection type.

## How many prisons are not in the data?

* Are there specific types that are missing? (Probably make this CA specific).
* Are the missing ones known for having issues but that is not reflected in our data? This would be hard to answer without domain knowledge on prisons. Compare to the Hifeld data.
* Do we have other facilities in here like sheriffs depts that we want to remove?

# Additional Ideas:

* Think about adding additional variables to the dataset. Both from OSHA and potentially linking to census data.
* Look into whether or not prison employees present. Do the standards cover them even though they receive lower wages?
* Regional analysis?

# Cleaning steps

* Cleaning up the facility names. See the issue on this. Need to clean text in R.
  + Related: need to decide how to handle facility with the same address but with an attached #. There are some examples of this in the data.
  + DONE
* Clean up the r code files in github to be more seamless.

# How do we examine toxicity?

## Look at current penalty

* Summary stats
* Look at biggest penalties - where are they?
* How much are missing info?
* How many had initial penalties but no current?
* How many had no initial penalty?

## Initial penalty versus current penalty

* How much of a “discount” did activity numbers get?
* How many started with a penalty but did not pay anything?
* How much are missing data?
* How many contested? Is there a relationship between contesting and not paying a penalty?

## Number exposed

* Distribution
* Which facilities had the biggest hazard based on number exposed?
* How much are missing data?

## How many instances per facility?

* This could be total (if the facility was examined multiple times or it could be per activity nr).

## How many were safety or health issues?

* There is a variable for this.
* How is this difference determined?

## What types of hazards were present?

* Check the standard variable. This could be a very long process to try to understand all of the standards. We should see how hard this will be to do.
* Also look at hazsub 1-5.
  + <https://github.com/Carceral-Ecologies/Caceral-OSHA-Data/issues/6>

## Gravity

* How serious were the violations?
* Which facilities are the most serious?

# 

# How will we report or display this information?

## Ideas:

* Start with CA from the last 10 years.
  + DONE
* Create a report detailing descriptive stats for the above conceptualizations of toxicity.
  + What are the most common violations?
* Create a scale of toxicity using the above variables to create an overall score.
  + Lots of issues with missingness for this.
* Create a report of only the most serious prisons we know about from the data we have.
* Integrate information about penalties, gravity, number exposed, type of hazard into the proximity map to provide more information about hazards within prisons and to provide more context for this project because we know we do not have all prisons in our dataset.
* Create an interactive table where people can look at individual prisons in CA and look at all of the metrics of toxicity at once.
  + Done once we publish the data.

# Where will we keep track of findings?

* Do we want to use the wiki?
* Google doc?
* Spreadsheet?

Steps to Finish up Spring Quarter

**Goal:**

**What are the products that would be beneficial for our careers?**

* A report that is publicly available that demonstrates our writing and data analysis/visualization skills.
* Showing we can work as a team.
  + Organizing a team
  + Letters of recommendation.
* What are some skills that you would like to be able to demonstrate to employers, faculty, or grad schools?

# To write a report on our findings.

* + **Think about who is our audience?** 
    - **White paper to appear on the Hack4CA website?** (full project can appear here)
      * Think about the format. Should it be an r markdown? R notebook? A github.io page? Ask Lindsay about this.
      * Perhaps an interactive table
    - **UC Davis datalab blog post** (shorter version of our white paper, see if we can get Nick and Lindsay to review our work and state that it has been reviewed.)
    - UC Davis Poverty Center (may be a bit tangential)
    - Scholars Strategy Network (not beneficial to undergrads so not ideal)
    - A prison activist group (we could send them our blog post)
    - OSHA/DOL or Cal/OSHA? (we could also send them our blog post, perhaps plus an additional paper on what is wrong with their data)
    - News organizations like Calmatters? (This would need to be super streamlined). We could submit a short version of the blog post. Could talk to Nick about this.
  + **Think about how long it should be (related to above)**
  + **Think about what order to list authors in (and Lindsay and Nick as well)**
  + **Think about the topics to cover in the write up:**
    - What is toxicity? How do we measure it?
    - Prisons has been built on toxic land. Cite Nick and Lindsay.
    - What can the OSHA data bring?
      * What did we do to create the dataset?
      * What are some relevant variables?
    - What are some of the things we learned?
      * Include data visualizations for each variable
      * What prisons did we identify as repeat offenders
    - What are some problems with the data?
      * Problems with the OSHA data explorer tool
      * Issue with facility names
      * Issue with missingness
      * Difficulty extrapolating to more states
    - What can’t the data tell us?
      * Prisons that were never inspected
      * Not clear which prisons have workers (in the data but we might be able to find out ourselves)

# Data work to do to finish up

* + - We need to decide if we want to do this by facility or by violation
    - Number of total violations
      * 797 violations out of 1093 inspections (296 no violation, only inspection)
      * Average number of violations per facility
    - Number of total facilities present.
      * Some idea of what types of facilities are missing or how many. Need to compare to the HIFLD.
        + 149 DONE - Savannah
    - Correlation matrix for all key toxicity variables? - Selam
    - Descriptive stats for each toxicity variable (in one table)- Ambar
      * Mean (if applicable)
      * Median (if applicable)
      * Mode
      * Frequency (if applicable - no mean or median)
      * Min
      * Max
      * SD
      * Number and percent missing data for that variable
        + Identify if this is missing because it had no violation (it was a record from the inspection dataset only) versus true missingness. - Savannah **DONE**
    - Simple plot showing the distribution of each toxicity variable. (this might not be as easy as it sounds and perhaps not necessary considering the table above but may be more visually appealing.)
      * We should decide on a standard look of the plots.
      * Grouped histogram for variables most compelling.
    - A second plot showing a relationship between the variable and another key variable of interest. - Selam
      * This is a place to be creative I think. After doing the above steps.
    - Create a table of all facility names and addresses and perhaps their mean or total value on each toxicity variable. With dates as well.
      * This could be interactive.
      * How would we deal with this considering the variables will be in wide format. So we would have to add up current penalty for example across multiple variables. Add we will have to include FTA penalties - example: https://www.osha.gov/pls/imis/establishment.inspection\_detail?id=312663560
    - Identify top “toxic” prisons from this work.
    - Perhaps think about a way to create a scale of toxicity from the OSHA data.

# What are all of the key variables:

* + - **Important background variables**
      * Safety\_hlth
      * Insp\_type
      * Insp\_scope
      * Union\_status
      * Nr\_in\_estab
      * Why\_no\_insp
      * Rec
      * Owner\_type
    - **Toxicity related (**[**Data Dictionary**](https://drive.google.com/file/d/1njaHqWazjpi4iFmuyHThHU3aHfkhC4Ba/view?usp=sharing)**,** [**Inspection Detail Definitions**](https://www.osha.gov/data/inspection-detail-definitions)**)**
      * [**See this document for summary info**](https://docs.google.com/document/d/1mDduWKWnT_mX0dWebAx9eSxdG1Y01Ok86Km5vLtVXO4/edit)**)**
      * Standard - add to blog post top 5 or so and then also a note that there is a lot of valuable information here for someone with the expertise to understand all of the violations.
        + Define that
        + We need to clean the text here. Remove spaces. Remove (), and trimws(). Then find table of most common. We could also separate out the numeric from the letter (put them in different columns) so we can just see the overarching standard that was violated.
      * Viol\_type - O, R, S - look up what this means
        + DONE - Savannah
      * Current\_penalty
      * Initial\_penalty
        + Think about providing a total or amount change
      * FTA
        + DONE
      * Nr\_instances
        + DONE
      * Nr\_exposed (could make percentage compared to nr in establishment)
        + DONE
      * Gravity
        + DONE
      * Emphasis? - define that - all NA’s in the dataset. I believe this has something to do with OSHA special enforcement programs.
      * Hazcat? - define that - all NA’s in the datset
      * Rec? - define that
        + DONE - Savannah
      * Fta (there are a few - failure to abate - could be people in danger longer)
        + There are also FTA penalties. Should we add this to the current penalty?
      * Hazsub - make a table - simple frequencies

1. **Publish the data with a DOI**
   * Drop columns in the code for variables that are completely NA - nevermind. I am going to leave them so people can see what is “available”
     + DONE
   * Re-organize columns so related activity is at the end because mostly NAs
     + DONE
   * Find out from Lindsay and Jared the process for this and where we might want to submit it.
   * Add acknowledgements to the github page to thank all the contributors.

# Ideas for relationships to explore:

Potential relationships to examine:

[See here for toxicity variable info](https://docs.google.com/document/d/1mDduWKWnT_mX0dWebAx9eSxdG1Y01Ok86Km5vLtVXO4/edit?usp=sharing)

[See here for the data dictionary](https://drive.google.com/file/d/1njaHqWazjpi4iFmuyHThHU3aHfkhC4Ba/view?usp=sharing)

|  |  |  |
| --- | --- | --- |
| **“independent variable”** | **“dependent variable”** | **Rationale** |
| **Union status** | **Toxicity variables**  · Noviol (True – means no violation present, False – means violation present)  · Current penalty (how serious was it)  · Initial penalty (how serious was it)  · Gravity (how serious was it)  · Viol\_type (how serious was it)  · How long was the violation open? (this might be a bit complicated but you could use the open\_date and the close\_case\_date to see if there is a relationship to how long it takes for an issue to get resolved)  **Other variables**  · Insp\_type (what type of inspection was it? Are they mostly complaints?)  · Also look at Rec (similar to insp\_type)  · Insp\_scope (are unionized places more likely to have complete inspections compared to partial?)  · Why\_no\_insp (why no inspection) | Research suggests that unions make workplaces safer. When accidents happen and unions are present, on average, issues get resolved sooner.    Are there more unionized workplaces in the dataset than none? (if true, this might not be an indication that unionized workplaces are less safe but actually than unions have the resources to be able to file complaints and get issues resolved with OSHA better than non-unionized workplaces)    Are there more violations associated with unionized workplaces?    See papers by David Weil |
| **Owner\_type** | **Toxicity variables**  · Noviol (True – means no violation present, False – means violation present)  · Current penalty (how serious was it)  · Initial penalty (how serious was it)  · Gravity (how serious was it)  · Viol\_type (how serious was it)  · Nr\_exposed  · Nr\_instances  · Standards (see slack chat for cleaning code; are different facility types prone to certain types of standards violations?)  · How long was the violation open? (this might be a bit complicated but you could use the open\_date and the close\_case\_date to see if there is a relationship to how long it takes for an issue to get resolved. You could also look at open\_date and abate\_date to see how long it took for the issue to be abated, it could be different from the close date I don’t know).  · Abate\_complete (did they fix it?)  **Other variables**  · Insp\_type (is there a relationship between owner type and inspection type?)  · Also look at Rec (similar to insp\_type)  · Insp\_scope (relationship between owner type and scope?)  · Why\_no\_insp (why no inspection) | This was something Selam noticed. Federal governments have no penalties associated with them despite exposing a lot of people.    I don’t know the literature well on this but there may be differences in the experiences of prisoners in federal facilities versus private (for-profit) prisons.  What is the distribution of owner type in the dataset? (are there more private facilities present than federal? etc) Is there a relationship between owner type and these variables? |
| **Gravity** | * Current penalty * Initial penalty | According to the OSHA field manual higher gravity should mean a higher penalty. Is that actually true? Is there a relationship here?  Here are some of the specifics from the OSHA manual.  According to manual - Serious Violation & GBP (gravity based penalty). a. The gravity of a violation is defined by the GBP: A high gravity violation is one with a GBP of $12,471. A moderate gravity violation is one with a GBP ranging from $7,126 to $10,689. A low gravity violation is one with a GBP of $5,345.  [Link](https://www.osha.gov/sites/default/files/enforcement/directives/CPL_02-00-160.pdf)  1 - low gravity  5 - moderate  10 - high gravity  There is a 3 coded which we don’t know what it is. |

# Resources:

* Making nice tables - <https://www.littlemissdata.com/blog/prettytables>
* Lindsay’s resources on tables - DT table and GT table seem ideal [How to Make Beautiful Tables in R](https://rfortherestofus.com/2019/11/how-to-make-beautiful-tables-in-r/)

# Interesting findings:

* Federal governments have paid nothing in fines yet have violations that exposed people (one at least 300 people)
* USP victorville has been investigated/cited 80 times!
* CALIFORNIA INSTITUTION FOR MEN (CIM) has been investigated/cited over 50 times!