* **Requirements:**
  + Well-articulated problem statement with "specific aim" and hypothesis, based on your lightning talk.
  + An outline of any potential methods and models.
  + Detailed explanation of the available data. (i.e., build a data dictionary or link to pre-built data dictionaries)
  + Describe any outstanding questions, assumptions, risks, and caveats.
  + Demonstrate domain knowledge, including specific features or relevant benchmarks from similar projects.
  + Define your goals and criteria, in order to explain what success looks like.
* **Bonus:**
  + Consider alternative hypotheses: if your project is a regression problem, is it possible to rewrite it as a classification problem?
  + "Convert" your goal metric from a statistical one (like Mean Squared Error) and tie it to something non-data people can understand, like a cost/benefit analysis, etc.

**Project Problem and Hypothesis**

* **What's the project about? What problem are you solving?**

This project is about police involved fatalities. I am not necessarily trying to solve a problem, while I want the solution to be a predictive model I am doing so to get a better idea about what features out of: Cause of Death, State, Gender, Age, Race, Year of Death, whether or not the deceased was armed, and whether or not the deceased had priors might be highly correlated with whether or not the police involved in the fatality gets fired or not.

* **Where does this seem to reside as a machine learning problem? Are you predicting some continuous number, or predicting a binary value?**

Based on the response variables I will be making a prediction about two binary values: Clear or unclear whether or not the police office was fired and whether or not the police was fired or not. This based solely on the face that the three possible responses for “Was the Police officer fired or suspended?” are “Unclear, Fired, Not-Fired.

* **What kind of impact do you think it could have?**

I’m not really sure if it could have much impact aside from informative.

* **What do you think will have the most impact in predicting the value you are interested in solving for?**

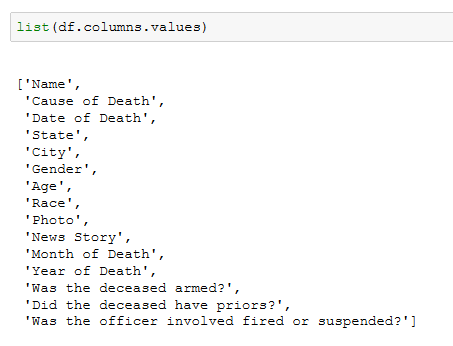
Right now I am inclined to think race will have the biggest impact but I think that is just a perception I have because of recent events. I think there is a good chance that state could have a bigger impact because stringency of laws could be variable. => which leads me to believe it might be interesting, with more data to try and accurately model whether or the police involved was fired by state.

**Datasets**

* **Description of data set available, at the field level. (see table)**

Data was acquired from the Data for Everyone part of the CrowdFlower site.





* **If from an API, include a sample return. (this is usually included in API documentation!) (if doing this in markdown, use the JavaScript code tag)**

Not applicable

**Domain knowledge**

* **What experience do you already have around this area?**

I don’t have much experience in this area, I don’t work in social justice or anything but do keep up with current events and am particularly interested in race-police relations presently.

* **Does it relate or help inform the project in any way?**

I’m not sure it helps relate or inform the project all that much, I think it makes me more biased than anything because I have preconceptions about what the data will indicate about police-race relations. However, it also helps that there are not that many potentially contributing features, as such very little pre-knowledge of the domain is necessary. That leaves me optimistic about modeling and finding correlations.

* **What other research efforts exist?**
  + Use a quick Google search to see what approaches others have made, or talk with your colleagues if it is work related about previous attempts at similar problems.
  + This could even just be something like "the marketing team put together a forecast in excel that doesn't do well."
  + Include a benchmark, how other models have performed, even if you are unsure what the metric means.

**Project Concerns**

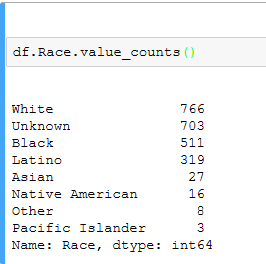
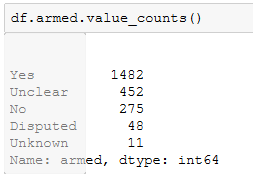
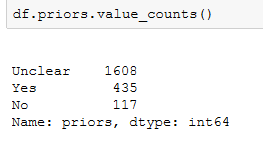
* **What questions do you have about your project? What are you not sure you quite yet understand? (The more honest you are about this, the easier your instructors can help)**

The major question I have about this project is about all the ambiguity in the data. A large number of the samples have a response variable of unclear. Does it make sense to do a predictive model for a binary that is clear and unclear? Or is this not value added? I see it as value added and still interesting but in general this doesn’t seem ideal.

I thought about removing all of the “unclear” values but if I remove them my data set is only roughly 800 samples and I am not sure if that would be enough. Another alternative is finding the data for years prior to 2013 and after 2015 but given my timeline this doesn’t seem entirely necessary or plausible but I will do some cursory searching to see if it’s available somewhere in a similar format.

Additionally, I am worried about how much data will be left when I start removing NaN values from the columns. I am worried it will make my data set unworkable.

Overall there are just unfortunately a lot of unknown values in the data set. For instance, below are three of the variables that I assumed would be most important in predicting whether or not the officer involved was fired. However, there are a ton of unknown values, which kind of makes me feel like I need more data.

* **What are the assumptions and caveats to the problem?**
  + What data do you not have access to but wish you had?

I want the same data for police involved fatalities from before 2013 and for 2016. I want a bigger data set to work with for this predictive model. I did fine more data from January 2015 here <https://github.com/washingtonpost/data-police-shootings/blob/master/fatal-police-shootings-data.csv>. Do you think it would be worthwhile to combine them? The data referenced here is just for shootings and my current data set is for all police involved fatalities. So maybe I should limit my predictive model to just shootings?

* + What is already implied about the observations in your data set? For example, if your primary data set is twitter data, it may not be representative of the whole sample. (say, predicting who would win an election)

I’m not sure I understand this question.

* **What are the risks to the project?**
  + What's the cost of your model being wrong? (What's the benefit of your model being right?)

There is no real cost to the model being right or wrong as it is mostly informative. However, if the model is right, or even just there are correlations strong enough to make a model then I will feel like have some concrete evidence to help make sense or recent events and their outcome. I will have a better idea of how things were handled historically and why.

* + Is any of the data incorrect? Could it be incorrect?

Definitely some of the data could be incorrect. I think I kind of presumed that was a possibility in any data set though, and I trust the source.

**Outcomes**

* **What do you expect the output to look like?**

Right now I just anticipate that the outcome will be some kind of logistic regression model that based upon the values of particularly important feature variables will make a probabilistic determination about first, whether or not it is clear or not that the officer got fired and then whether not the officer will get fired or not.

* **What does your target audience expect the output to look like?**

I am the only target audience at this point so, see above.

* **What gain do you expect from your most important feature on its own?**

I’m not sure, I am not even sure now which feature will be the most important.

* **How complicated does your model have to be?**

I am not sure right now, I could incorporate a lot of feature variables or just a few, the complexity will be determined by the level of correlation between feature variables the response variable.

* **How successful does your project have to be in order to be considered a "success"?**

I’m not sure.

* **What will you do if the project is a bust (this happens! but it shouldn't here)?**

I will probably try to find more data about police involved fatalities and keep trying.