One Page Proposal

Delightful Data

**We are planning to determine if there is a correlation between Superfund Site and socio- economic status for various cities across the US.**

**Superfund sites** are polluted locations requiring a long-term response to clean up hazardous material contaminations. Environmental Protection Agency (EPA) creates a list of such locations, which are placed on the National Priorities List (NPL)

**Data Collection**

We gathered the data regarding NPL Superfund sites from the EPA resources provided below:

* <https://www.epa.gov/superfund/superfund-data-and-reports>

We will gather the data from the census using the census module found in Python

**Goals and Ideas**

We are trying to:

* Trend of the number of sites in a region. (Stacked bar chart, state/county for each of the year)
* Changes in number of sites by state and year.
* No. of sites added/deleted on a yearly basis state-wise.
* We will be comparing average values of certain parameters from census data for National Average against the averages of areas around superfund sites.

Simple T test –population of zip codes of Superfund sites versus mean population all zip codes

Use ttest equal to compare superfund sites versus deleted superfund sites

Use chi test to compare population stats of nation vs superfund

Chisquare test – is there a causal relationship? Race to socio-economic status – won’t tell you what the significance is, only if there is a significance

ANOVA – superfund/total w Median household income, population, home value