# Predicting the Number of Crisis Pregnancy Centers by State

### Data Collection

What data will you collect or create?

I will collect open access datasets available from the following sources:

* percent of people uninsured by state (from US HHS/Kaggle, available here: <https://www.kaggle.com/hhs/health-insurance>)
* whether or not states have expanded Medicaid (from KFF: <https://www.kff.org/health-reform/state-indicator/state-activity-around-expanding-medicaid-under-the-affordable-care-act/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D>)
* total number of residents by state (from KFF via Census: <https://www.kff.org/other/state-indicator/total-residents/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D>)
* number of abortion providers by state (using Guttmacher Institute data: <https://data.guttmacher.org/states>)
* number of abortions provided by state (using state of occurrence, not state of residence) (from Guttmacher Institute data: <https://data.guttmacher.org/states>)
* data on abortion restrictions (whether state has any prohibition on abortion past any point in gestation, whether state requires a person to receive information on ultrasounds or undergo an ultrasound, whether state requires abortion facilities to adhere to ambulatory surgical center requirements, whether state policy protects or restrict right to abortion, whether state has a required waiting period before abortion, whether state allows use of Medicaid funds to pay for abortion, whether state requires parental consent for abortion, whether state has a so-called "partial birth" ban) - all come from KFF's Stata Data portal: <https://www.kff.org/state-category/womens-health/> using Guttmacher Institute data
* state population (from KFF dataset using Census Bureau's American Community Survey data: https://www.kff.org/other/state-indicator/total-residents/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D)
* state voting behavior in 2012 and 2016 (compiled from Federal Election Commission data: <https://transition.fec.gov/pubrec/fe2016/federalelections2016.pdf> and [https://transition.fec.gov/pubrec/fe2012/federalelections2012.pdf)](https://transition.fec.gov/pubrec/fe2016/federalelections2016.pdf)
* number of crisis pregnancy centers by state (from Reproaction: https://reproaction.org/fakeclinicdatabase/)

How will the data be collected or created?

The data is all open source, but I will use Python and pandas to clean the datasets. I will be removing empty rows and removing columns not used for analysis. From the CPC database, I isolate the state variable from their column (since city, state, and zip are all together in the same column) and then restructure the data so that each state is associated with the count of unique CPCs in that state. From there, I will merge all datasets on the state variable, and then rename the columns of the final merged dataset for clarity.

### Documentation and Metadata

What documentation and metadata will accompany the data?

The variables will be clearly labeled, using underscores (\_) to separate words in variable titles. String variables, from the original datasets, will have \_str appended to their label, while dummy variables (0 for No and 1 for Yes) will have \_yn appended to their data label.

### Ethics and Legal Compliance

How will you manage any ethical issues?

All of the data are secondary data, so all should have gone through an ethical review board process before collection. All of the data is aggregated at the state level, so it is anonymous and de-identified.

How will you manage copyright and Intellectual Property Rights (IP/IPR) issues?

All of the data are open source and available on the Internet.

### Storage and Backup

How will the data be stored and backed up during the research?

I will store the data on my computer. I backed up the data on my Google drive and sync the backup weekly.

How will you manage access and security?

I am working alone so do not need to ensure access for collaborators. My computer is password-protected.

### Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

The data are all backed up on the websites which hosted them originally, but I will keep my dataset available through my GitHub account.

What is the long-term preservation plan for the dataset?

I will retain the data on my Google Drive and my GitHub moving forward.

### Data Sharing

How will you share the data?

My data will be available on my GitHub account for anyone to use. Because my data is all open source, I do not have any proprietary ownership over the data. My code will also be available on my GitHub so that it will be replicable.

Are any restrictions on data sharing required?

no restrictions required.

### Responsibilities and Resources

Who will be responsible for data management?

I am responsible for data management and will be implementing the DMP. I will be reassessing my data management plan weekly throughout the project to make sure that I am adhering to it and to check if there are any edits I need to make.

What resources will you require to deliver your plan?

I will require GitHub and Google Drive accounts and repositories.