Visualization of the Opioid Crisis in the USA
Team Big Data (114) Final Report

Introduction and Motivation, Problem Definition

The American opioid crisis has reached alarming proportions and is taking a severe toll on public health [1-3, 15]. While prior research has attempted to measure the extent of the crisis, a holistic study that includes socioeconomic factors across multiple data sources, advanced prediction techniques, and informative visualizations is lacking. This reality motivated us to develop a multidimensional analysis to understand the degree of this crisis and the factors which may explain opioid usage and mortality.

Literature Review

Two-thirds of deaths from drug overdoses are caused by opioids, and opioid abuse has doubled in prevalence since 2004 [1]. The most common explanation is over-prescription [2] [3]. Nonetheless, some authors have found that prescription rates have steadily decreased while opioid deaths have increased in the same states [4], suggesting that the issue is more complex than just excessive prescription. Indeed, several major questions are still unanswered: Which demographics are most affected? What are the characteristic factors in these populations? Which opioids are most prevalent – and most deadly? Several publications call for increased access to such analysis, including some that have raised questions around potential bias in CDC data upon which much previous work was based [5]. Data reaching across multiple secondary sources are especially lacking [8]. No study but one [19] provides a freely available visualization tool. While existing socioeconomic status indices have been shown to help predict adverse outcomes [6, 7], these indices have not been widely used in prior studies due to limited availability. Previous studies also tend to be single- or dual-factor (i.e., examining the effect of a single socioeconomic variable on opioid consumption), and typically cover effects at the aggregate level, rather than incorporating a wide array of concurrent factors and analyzing individual geographic localities [9 - 14]. Given that opioid crises are cropping up in other countries as well, we hope that this work can serve as the foundation for future analysis in the public health area [16-18].

Proposed Method

<u>Intuition</u>

Our project aims to fill several gaps in current understanding. First, opioid usage data stitched together from a wide variety of secondary sources is known to be lacking [8] -- a reality that extends both to modeling and visualization. Further, the most widely-available visualization provides only pills per county: there exists an opportunity to provide users with important or correlated sociodemographic data that helps them understand the opioid crisis much more deeply. Finally, nonparametric machine learning methods have been shown to improve tangibly upon linear regression in many arenas, providing an opportunity to advance the state of the field in terms of modeling opioid consumption* and opioid-related mortality**.

* Consumption = grams of hydrocodone or grams of oxycodone purchased in a given year in a given county

** Mortality = number of deaths involving a drug overdose related to opioids

<u>Description of Approach</u>

Data synthesis

We synthesized the 88GB ARCOS-Washington Post database of opioid transactions from 2006-2012 into groups by year, state, county, and drug type, based upon a star schema. The main dataset was 88 gigabytes in size with 179 million transaction records. We used *Dask* and *pandas* to perform grouping on each dimension which we then insert into a SQLite database using Python Jupyter notebooks; *Dask* was also used to reduce transaction entries into a *pandas* dataframe with all data and keys marshalled into usable formats. We next queried Census Bureau APIs for 1-, 3-, and 5-year ACS data to acquire data for the following sociodemographic factors by county and year: median household income, number of individuals in poverty, degree holders, number of individuals in school, non-white population, number of veterans, population with disabilities, number of housing units, median home value, number of homes with mortgages, median monthly housing costs, estimated number of bedrooms, people in labor-sector jobs. We acquired mortality data from the CDC's website and used Python to extract the deaths caused from opioids, reducing a total of 25 million rows into 55k rows. Unemployment data, workplace fatalities, and crime data were acquired from the Bureau of Labor Statistics (BLS), Occupational Safety and Health Administration (OSHA), and state governments respectively. Finally, we merged all datasets into the SQLite database based upon county FIPS code and year as a primary key. At this stage we also ran correlation analyses between each of the above variables and opioid transactions, and hosted the website framework, server code, D3 scripts, data CSV files, and SQLite database on GitHub Pages.

Prediction tasks

We built three predictive models (log-log linear regression, log-log LASSO regression, random forest) for each of two consumption-related prediction tasks (grams of hydrocodone consumed, grams of oxycodone consumed). We also built two predictive models (linear regression, random forest) for each of two mortality-related prediction tasks: mortality *without* drugs as input, and mortality *with* drugs as input. In each model, we controlled for the county's population. We assessed features of importance using standardized coefficients (regression) and feature importances (RF) (see Experiments 1 and 2). Wherever coefficients were compared to each other, we employed standardized data to guarantee consistent scales. We compared model performances by using RMSE on a held-out test set (see Experiment 3).

Visualization

We built a web application that includes the following features: pills per person, grams per person, opioid deaths, inlaid graphs of trends in the top 3 correlated factors for the selected county. It also contains a "deep dive" feature: clicking on a county generates RF model projections for future opioid transactions, projected mortality in the county, and trends in each input feature. To evaluate the effectiveness of our product compared to the existing visualization [19] (which, as noted before, provides only pill-count statistics), we contacted a panel of eight healthcare professionals and solicited their evaluations by using a simple anonymous survey (see Experiment 4). We restrict our survey respondents to healthcare professionals because of the background knowledge required to understand different opioids and their ramifications for policy.

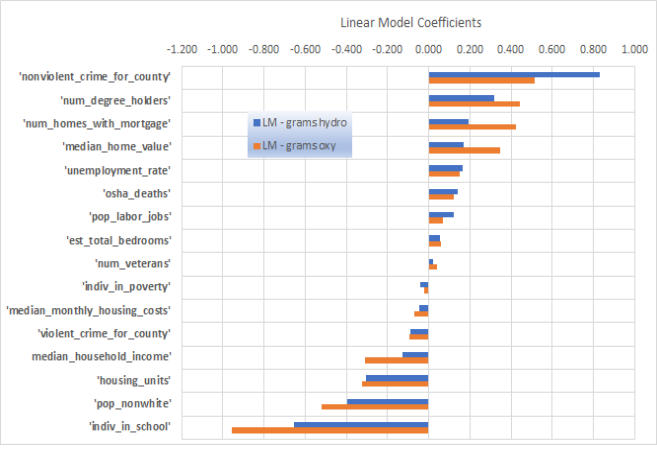
Experiments and Evaluation

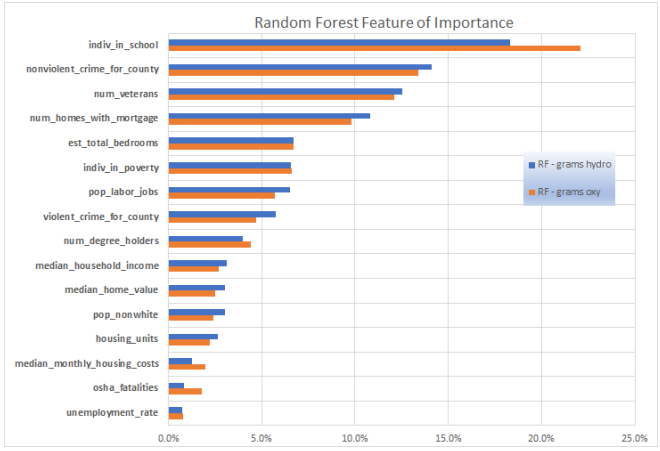
Key Question 1. What sociodemographic factors tend to correlate with, or explain the variance in, opioid transactions in the United States? What factors are most important? Do these factors indicate where

governments could increase social services? Key Question 2. To what extent are opioid consumption levels correlated with mortality rates? Put differently, does the data support the common narrative that excessive opioid prescription is the key factor in opioid-related mortality? Testbed. We employed linear regression, LASSO regression, and random forests for two prediction tasks on a per-county basis: (1) prediction of grams of opioids transacted (hydrocodone and oxycodone), and (2) prediction of opioid-related deaths (both with and without opioids included in the model). Each prediction task was conducted on our synthesized, county-level dataset spanning 2006-2012 with the variables listed above. We employed (unit-) standardized coefficients (LR, LASSO) and feature importances (RF) to indicate variables of importance. Each model was evaluated by using root mean squared error (RMSE) on a held-out test set of the data.

Experiment 1: What factors tend to correlate with opioid consumption?

Hypothesis 1: Among socioeconomic and sociodemographic variables, variance in opioid consumption is most explained by the state of the economy. Experiment 1: We trained and evaluated important features for six models in total: a log-log linear regression model, a log-log LASSO regression model, and a random forest model to predict grams transacted of each of hydrocodone and oxycodone respectively. Findings 1: The following figures show standardized coefficients and feature importances for each prediction task. None of the methods prioritized economic factors in predicting opioid consumption, bucking the conventional narrative that economics explains the opioid crisis. We find that education levels (indiv_in_school, num_degree_holders), housing availability/affordability (num_homes_with_mortgage, est_total_bedrooms), crime rates, and veteran status are *far more important than economic factors* in explaining opioid consumption. Unemployment rate and median household income, which our team expected to find as top factors, were comparatively unimportant. Conclusions 1: Our data suggest that it would be wrongheaded to insist on fiscal explanations of the opioid crisis. Improving school systems, crime rates, and housing availability seem to be much more impactful areas for possible local government investment.

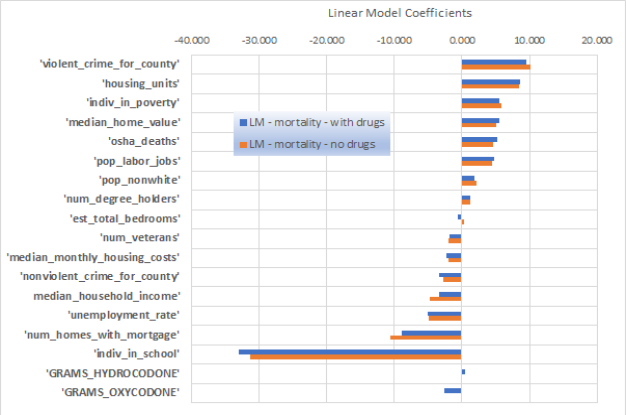


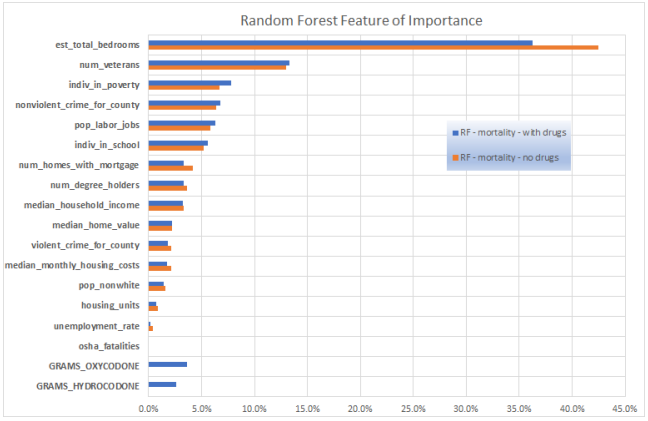


Experiment 2: What tends to correlate with opioid-related mortality? Are prescriptions most important?

Hypothesis 2: Sociodemographic factors are of greater importance than opioid consumption statistics for predicting mortality among patients who have previously taken opioids. Experiment 2: We next trained a linear regression model and random forest model to predict opioid-related mortality both *with* drugs transacted and

without it. This task specifically allowed us to check whether (1) adding opioids to a model based purely upon sociodemographic factors improved prediction accuracy, and (2) grams of hydrocodone and oxycodone proved to be the most important factors for forecasting mortality. Findings 2: The two figures below, as well as the table below, show results for this experiment. Adding opioid consumption to the mortality model showed *no benefit to predictive accuracy* in LR or RF models alike based upon hold-out set RMSE. Further, neither method prioritized hydrocodone or oxycodone consumption above key sociodemographic factors (e.g. housing, veteran status) in explaining mortality for each county. Conclusions 2: Our results suggest that sociodemographic factors, including the availability of housing and education, are paradoxically more important than opioid volume for explaining opioid-related deaths. Put simply, pill-count is not the whole story. Our findings therefore contradict the "overprescription hypothesis", which holds that an increase in prescriptions has driven the opioid crisis.





Experiment 3: Can machine learning methods offer a benefit to studies of population health?
Hypothesis 3: Advanced prediction techniques can offer an improvement in opioid-related forecasts as compared to regression. Experiment 3: As described prior, we built three predictive models (LR, LASSO, RF) for each of two prediction tasks (grams transacted, mortality). We then compared the performance of each model by evaluating its RMSE on a held-out test set unseen by the models during training. Findings 3: Results from modeling experiments are shown in the table below. LASSO models provided little-to-no improvement to baseline linear regression. Meanwhile, RF methods provide a significant benefit, sometimes reaching 50% or more improvement in hold-out set RMSE compared to linear regression. This benefit held in forecasting opioid consumption and mortality alike. Conclusions 3: Health services researchers must incorporate advanced ML methods -- such as RF -- into population health forecasting if their goal is to improve accuracy.

Prediction task	Method	R^2 / OOB score	Hold-out set RMSE	Improvement over LR
Grams hydrocodone	LR	72%	1,081,923	-
	LASSO	73%	1,173,710	-8%
	RF	88%	598,310	45%

Grams oxycodone	LR	71%	2,953,543	-
	LASSO	74%	2,523,902	14%
	RF	82%	1,342,637	55%
Mortality *without* drugs in model	LR	41%	1,222	-
	RF	93%	620	49%
Mortality *with* drugs in model	LR	40%	1,516	-
	RF	93%	503	67%

Experiment 4: Can we offer a more informative visualization tool than the current state of the field?
Hypothesis 4: A visualization application that provides correlations, local trends, and other data exploration offers significantly more benefit to policymakers than the existing visualization, which provides only pill-count. Experiment 4: We built an openly-accessible website (Visualization of the Opioid Crisis in the USA (kariato.github.io)) that uses our six-source dataset to provide background information about top correlations between opioid consumption and sociodemographic factors, model predictions for future years, and local trends. We contacted a panel of eight healthcare professionals, and solicited their anonymous feedback via a survey (Appendix B). In particular, we sought to compare our offering to the most-available existing tool, built by the El Paso Times. Findings 4: Six of eight survey respondents said our tool was a "complete improvement" over the existing visualization. All eight respondents fell between "likely" and "every time" with respect to referencing our product in the future. (Screenshots of our site are provided in Appendix C.) Conclusions 4: Healthcare professionals demand a greater degree of detail about the opioid crisis than that provided by existing tools. Important sociodemographic correlates and projections for future mortality hotspots seem to meet these desires. While eight respondents is a small sample size (largely due to the time required to find and partner with healthcare professionals during the COVID-19 pandemic), we believe these findings are representative of a broader narrative about the tools available to professionals who work with opioids.

Conclusions and Discussion

While our experiments uncover several interesting findings, understanding the opioid crisis continues to be a challenge. For instance, our opioids dataset ranges from 2006-2012, and further data collection efforts are needed in order for research to continue. Visualization tools require further improvement and validation given that our sample of survey participants is of admittedly small size. Nonetheless, our indicative findings are encouraging. Much health data is restricted (e.g. comorbid conditions) due to privacy laws, so health systems must work directly with government entities to continue exploring the opioid crisis and its correlated factors. Finally, it is plausible that much opioid consumption is driven by illegal markets. *Very little data exists* about the illegal market, and much more research effort is needed to further clarify the picture.

Distribution of Team Member Effort: All team members have contributed a similar amount of effort towards the success of this project.

References

[1] Vadivelu, N., Kai, A.M., Kodumudi, V. et al. The Opioid Crisis: a Comprehensive Overview. Curr Pain Headache Rep 22, 16 (2018). https://doi.org/10.1007/s11916-018-0670-z

[2] Weiner, Scott G. MD, MPH; Malek, Sayeed K. MD; Price, Christin N. MD The Opioid Crisis and Its Consequences, Transplantation, April 2017, 101(4): 678-68. https://doi.org/10.1097/TP.0000000000001671

[3] Teresa A. Rummans, M. Caroline Burton, Nancy L. Dawson. "How Good Intentions Contributed to Bad Outcomes: The Opioid Crisis". Mayo Clinic Proceedings, Volume 93, Issue 3, 2018, Pages 344-350, ISSN 0025-6196, https://doi.org/10.1016/j.mayocp.2017.12.020.

[4] Wickramatilake S, Zur J, Mulvaney-Day N, Klimo MC von, Selmi E, Harwood H. How States Are Tackling the Opioid Crisis. Public Health Reports. 2017; 132(2):171-179. https://doi.org/10.1177/0033354916688206

[5] Schatman, M. E., & Ziegler, S. J. (2017). Pain management, prescription opioid mortality, and the CDC: is the devil in the data?. Journal of pain research, 10, 2489–2495. https://doi.org/10.2147/JPR.S153322

[6] Bhavsar NA, Gao A, Phelan M, Pagidipati NJ, Goldstein BA. Value of Neighborhood Socioeconomic Status in Predicting Risk of Outcomes in Studies That Use Electronic Health Record Data. JAMA Netw Open. 2018;1(5):e182716. https://doi.org/10.1001/jamanetworkopen.2018.2716

[7] Kind AJH, Jencks S, Brock J, et al. Neighborhood socioeconomic disadvantage and 30-day rehospitalizations: an analysis of Medicare data. Ann Intern Med 2014;161(11):765-74.

[8] Rosanna Smart, Courtney A. Kase, Erin A. Taylor, Susan Lumsden, Scott R. Smith, Bradley D. Stein, Strengths and weaknesses of existing data sources to support research to address the opioids crisis, Preventive Medicine Reports, Volume 17, 2020, 101015, ISSN 2211-3355, https://doi.org/10.1016/j.pmedr.2019.101015.

[9] Altekruse SF, Cosgrove CM, Altekruse WC, Jenkins RA, Blanco C (2020) Socioeconomic risk factors for fatal opioid overdoses in the United States: Findings from the Mortality Disparities in American Communities Study (MDAC). PLoS ONE 15(1): e0227966. https://doi.org/10.1371/journal.pone.0227966.

[10] Shannon M. Monnat, David J. Peters, Mark T. Berg, Andrew Hochstetler, "Using Census Data to Understand County-Level Differences in Overall Drug Mortality and Opioid-Related Mortality by Opioid Type", American Journal of Public Health 109, no. 8 (August 1, 2019): pp. 1084-1091.

[11] Alex Hollingsworth, Christopher J. Ruhm, Kosali Simon, Macroeconomic conditions and opioid abuse, Journal of Health Economics, Volume 56, 2017, Pages 222-233, ISSN 0167-6296, https://doi.org/10.1016/j.jhealeco.2017.07.009.

[12] Sarah Gebauer, Joanne Salas, Jeffrey F. Scherrer, Neighborhood Socioeconomic Status and Receipt of Opioid Medication for New Back Pain Diagnosis, The Journal of the American Board of Family Medicine Nov 2017, 30 (6) 775-783; https://doi.org/10.3122/jabfm.2017.06.170061.

[13] Currie, J., Jin, J. and Schnell, M. (2019), "US Employment and Opioids: Is There a Connection?", Health and Labor Markets (Research in Labor Economics, Vol. 47), Emerald Publishing Limited, pp. 253-280. https://doi.org/10.1108/S0147-912120190000047009

[14] Hammersley, R., Forsyth, A., Morrison, V. and Davies, J.B. (1989), The Relationship Between Crime and Opioid Use. British Journal of Addiction, 84: 1029-1043. https://doi.org/10.1111/j.1360-0443.1989.tb00786.x

[15] Johnson KF, Worth A, Brookover D. Families Facing the Opioid Crisis: Content and Frame Analysis of YouTube Videos. The Family Journal. 2019; 27(2): 209-220. https://doi.org/10.1177/1066480719832507
[16] Katia M C Verhamme, Arthur M Bohnen, Are we facing an opioid crisis in Europe?, The Lancet Public Health, Volume 4, Issue 10, 2019, Pages e483-e484, ISSN 2468-2667, https://doi.org/10.1016/S2468-2667(19)30156-2.
[17] van Amsterdam J, van den Brink W. The Misuse of Prescription Opioids: A Threat for Europe? Curr Drug Abuse Rev. 2015;8(1):3-14. https://doi.org/10.2174/1874473708011506111842 18.
[18] Kurth AE, Cherutich P, Conover R, Chhun N, Bruce RD, Lambdin BH. The Opioid Epidemic in Africa And Its Impact. Curr Addict Rep. 2018 Dec;5(4):428-453. https://doi.org/10.1007/s40429-018-0232-9.
[19] El Paso Times. (2019). Opioid Epidemic: See how millions of pills moved through your area. Retrieved from https://data.elpasotimes.com/pain-pills/.

Appendix A: Sociodemographic factors (extracted per county per year)

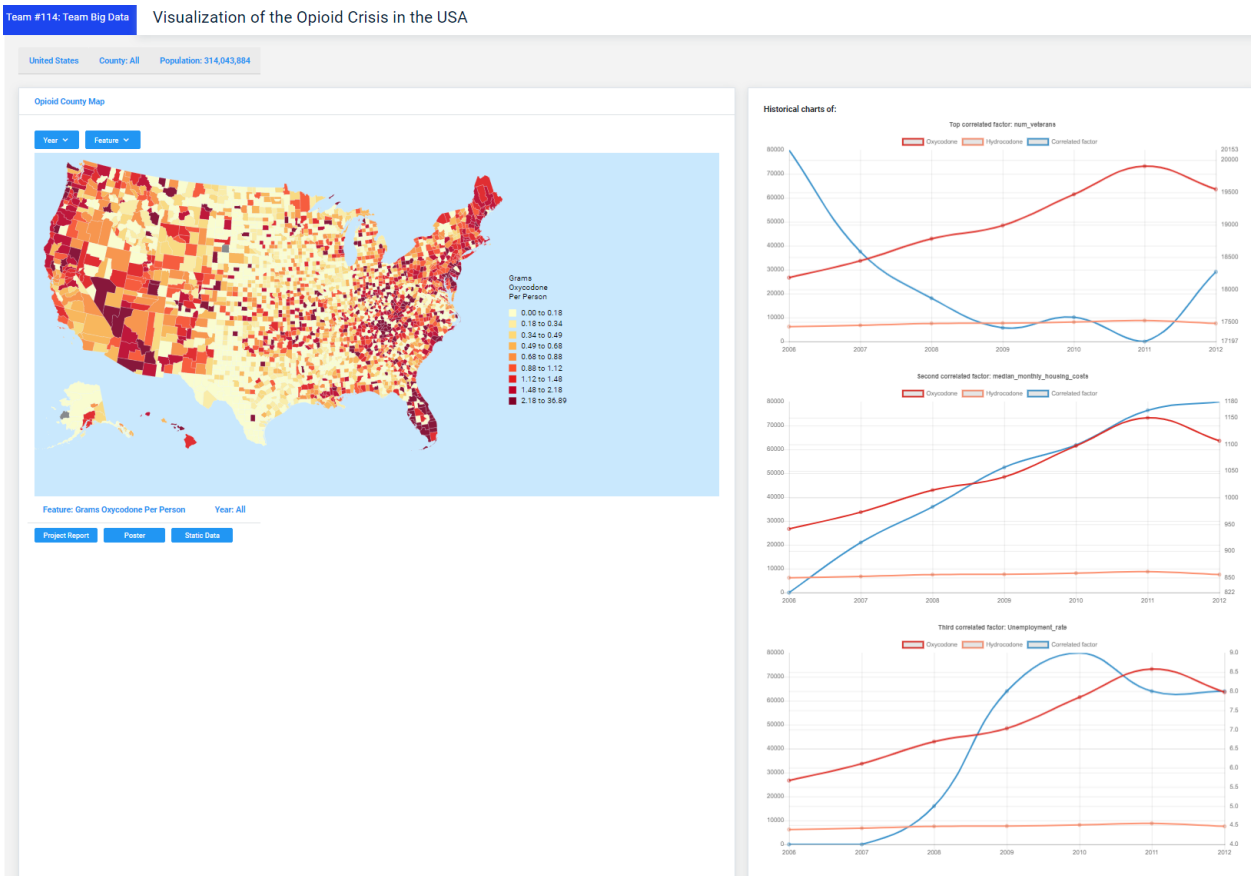
- Economy and employment
    - Unemployment rate
    - Median household income
    - Individuals in poverty
    - Population in labor-sector jobs
- Housing affordability and availability
    - Number of housing units
    - Median monthly housing costs
    - Median home value
    - Number of homes with mortgage(s)
    - Estimated total bedrooms in county
- Education system
    - Individuals enrolled in school
    - Number of degree holders
- Other demographics
    - Non-white race
    - Number of veterans
    - Nonviolent crime rate
    - Violent crime rate
    - Number of OSHA-reportable workplace fatalities

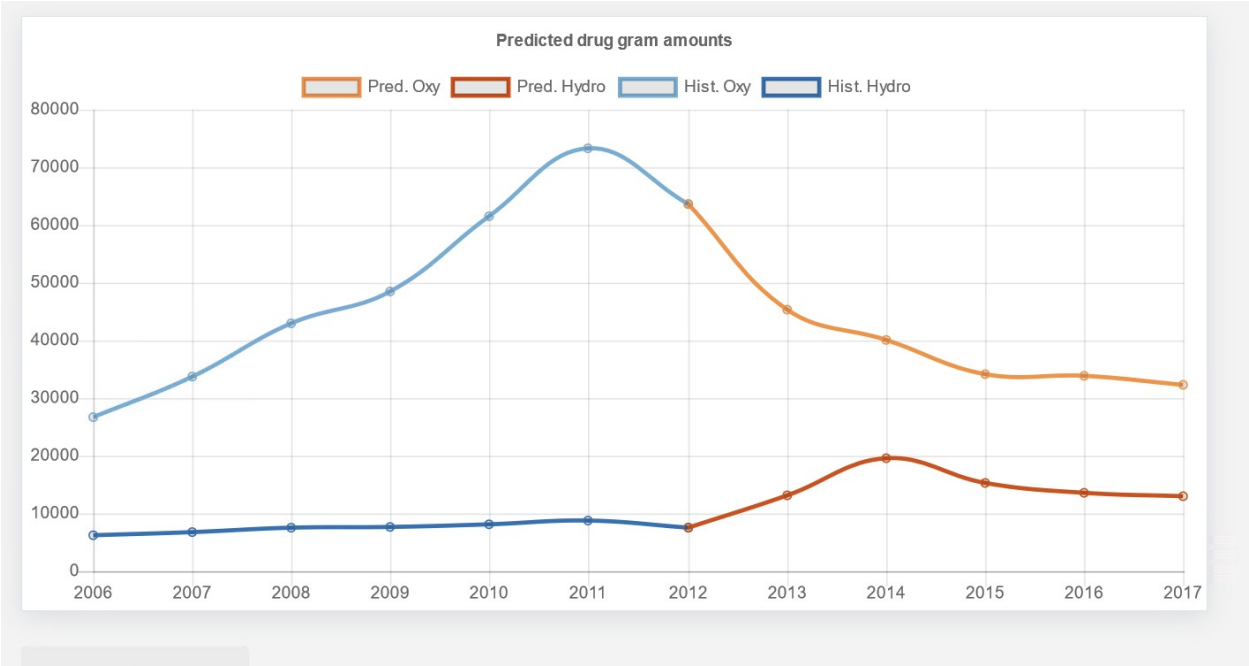
Appendix B: Survey questions

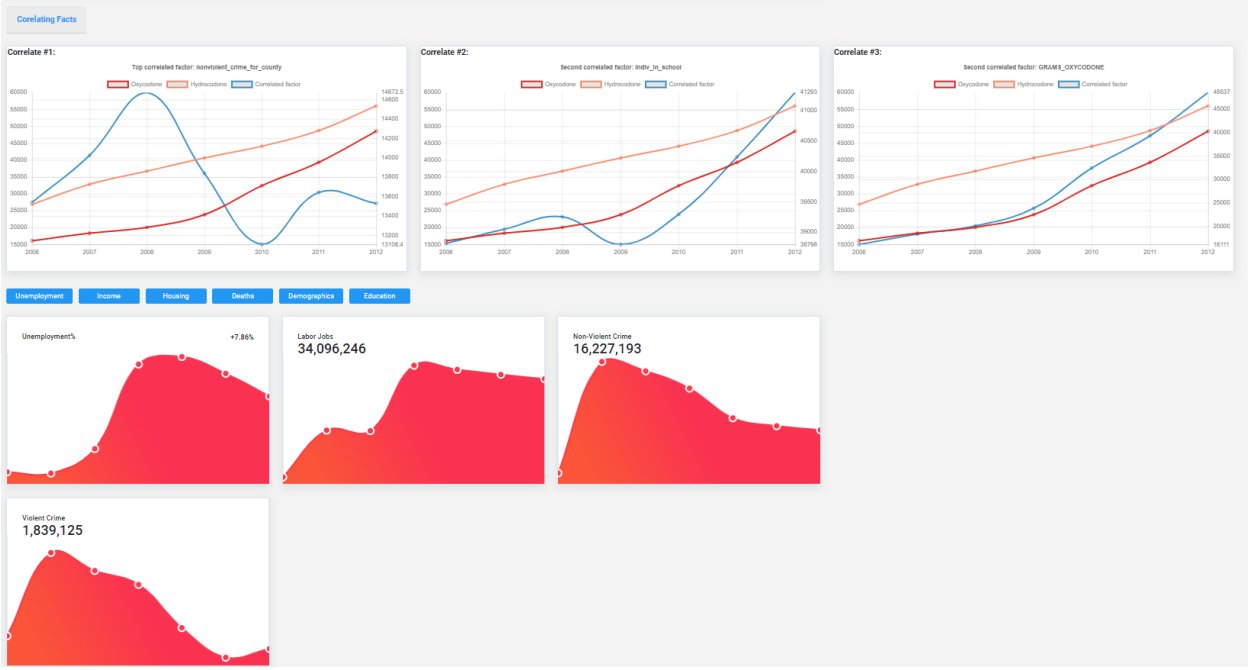
- To what extent does our tool help you understand the sociodemographic factors that correlate with the opioid crisis? (1-5, 1 = no understanding, 5 = great understanding)
- To what extent does our tool *improve upon* the existing visualization tool? (1-5, 1 = no improvement, 5 = complete improvement)

- To what extent do the visual layout, tooltip (pop-up when mousing over the map), and right-hand-side charts help you understand the opioid crisis? (1 = I learned nothing new, 5 = I learned lots of new information)
- If you hoped to understand more about opioids in the course of your work, how likely are you to reference our tool? (1 = will not reference, 5 = will reference every time)
- Which feature was most important to improving your understanding (if at all)? (text entry)

Appendix C: Website screenshots




Predicted drug gram amounts
Pred. Oxy
Pred. Hydro
Hist. Oxy
Hist. Hydro
80000
70000
60000
50000
40000
30000
20000
10000
0
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017


Corelating Facts
Correlate #1:
Correlate #2:
Correlate #3:
Unemployment
Income
Housing
Deaths
Demographics
Education
Unemployment%
+7.86%
Labor Jobs
34,096,246
Non-Violent Crime
16,227,193
Violent Crime
1,839,125