QSIDE Data Analysis

By: Ritik Jain, Ritin Kalindindi, Gabriella Stickney, Naamna Modi

Image from: https://qsideinstitute.org/

Overview

The data science capstone course is an opportunity for students to apply their knowledge to real-world problems while developing technical and professional skills. The course emphasizes teamwork, communication, and collaboration with industry and community partners. The Act-NOW: Community Health Index Project, developed in collaboration with the QSIDE Institute, aims to address social tensions between communities of color and law enforcement by leveraging quantitative methods. The project focuses on key societal issues, including gun-related violence, opioid addiction, and truancy, with the main goal of improving public safety and community well-being. The project aims to deliver actionable insights by creating a Community Health Index, a standardized tool for evaluating health and safety indicators across various communities.

Communities of color often face disproportionate challenges related to crime, law enforcement interactions, and public health disparities. These challenges strain relationships between residents and authorities, deepening cycles of poverty and inequality. By analyzing open-source data, this project aims to uncover underlying patterns that influence these dynamics, guiding the development of effective policies and targeted community interventions. The research focuses on identifying key factors affecting community health and safety, specifically in communities of color, exploring how the findings can be leveraged to provide actionable insights for improving law enforcement practices and community support initiatives to create a standardized Community Health Index reflecting social, economic, and public health conditions.

The project's primary goal is to produce a data-driven Community Health Index that helps stakeholders, including policymakers, social justice advocates, and law enforcement agencies, better understand and address community-specific challenges. The project already has a data map with approximately 60-70% of these indicators sourced from the Census Community Survey, which will require cleaning and refinement. The team has a goal to explore ways to capture the remaining 30-40% of the metrics to try and provide a complete picture of community health. By the end of the project, the team will have

developed a comprehensive dataset, analytical models, and visualization tools that contribute to a greater understanding of community health disparities.

Technical Details

Our project will utilize datasets in CSV or XLSX format, which we will read and process primarily using Python with libraries such as Pandas and NumPy. These tools will allow us to efficiently clean, manipulate, and analyze the data. Alternatively, we may consider using R if the team finds that its statistical capabilities or visualization tools better suit our needs. The choice between Python and R remains an open loop, which we will resolve based on early data exploration and team preference.

Once the data is loaded, we will conduct exploratory data analysis to assess its structure, identify missing values, and determine necessary preprocessing steps. This process will help us refine our dataset selection and ensure data quality. Since we aim to develop a health index for at-risk communities, we will explore different statistical models and machine learning techniques to weigh various factors, such as crime rates and socio-economic indicators. Determining the most effective methodology for calculating this index is another open loop that will be addressed through research and collaboration with community partners.

For modeling and analysis, the team anticipate using techniques such as regression analysis, clustering algorithms, or decision trees, depending on what best fits our data and project goals. Our group will leverage libraries like scikit-learn for machine learning, and if geospatial analysis is required, we may incorporate Geopandas and Folium for mapping crime patterns and community risk levels. The specific algorithms and tools chosen will be refined as we better understand the data and its predictive potential.

The final output will likely include visualizations, reports, and predictive models presented in an accessible format for our stakeholders. These could take the form of Jupyter Notebooks, PDFs, or interactive dashboards using Plotly or Dash. The exact format remains an open loop and will be finalized based on discussions with our sponsors and community partners.

For computation, we plan to use personal computers and university-provided computing resources. The success of our project will be measured by the accuracy and interpretability of our models, as well as the usefulness of insights provided to stakeholders. We will track performance using standard metrics like accuracy, F1-score, or Mean Squared Error (MSE), depending on the nature of our final model. Our findings and code will be shared with sponsors through GitHub, shared drives, or formal presentations, ensuring transparency and accessibility.

This technical approach will evolve as we analyze the data, ensuring our methods and tools align with project goals and stakeholder needs.

Project Goals

The success of the Act-NOW: Community Health Index project depends on addressing several key challenges while ensuring our methods produce meaningful, actionable insights. One of our primary objectives is to build upon the work completed by previous teams. While these teams have made progress, we need to review their methodologies and identify areas for improvement.

One major challenge is overcoming the limitations of national datasets. While these data sources provide valuable information, they are often outdated, released in large chunks that lack real-time relevance. The most recent available data is from 2021, making it difficult to capture current trends. To address this, we aim to identify at least one or two local sites where we can collect more recent and specific data on chronic absenteeism. However, collecting local data presents its own set of challenges. Unlike national datasets, which are readily available but outdated, local data may be more current but harder to obtain due to accessibility restrictions, data privacy concerns, and inconsistencies in how different sources collect and report information. Additionally, even when access is granted, local datasets may be incomplete, lack standardization, or require significant cleaning and preprocessing before they can be used for analysis. To overcome these challenges, we will need to engage with community stakeholders to develop strategies for handling incomplete or inconsistent data. This process will require a combination of outreach, negotiation, and technical problem-solving to ensure that the data we collect is both useful and ethically sourced

Our project focuses particularly on chronic absenteeism among high school students, as truancy is often a key indicator of broader community health challenges. Chronic absenteeism, defined as missing a significant portion of the school year for any reason—excused or unexcused—has serious long-term consequences. It can hinder early literacy development, impede academic achievement, and reduce high school graduation rates. By analyzing absenteeism data, the team hopes to identify underlying factors contributing to this issue. Research indicates that students from low-income backgrounds are disproportionately affected, often due to systemic barriers such as inadequate transportation, unsafe walking routes, food insecurity, or unmet healthcare needs. Additionally, students of color and those with disabilities are more likely to experience chronic absence, further worsening educational inequalities.

One of the main uncertainties we need to resolve is determining which specific data points best reflect absenteeism trends and how these trends correlate with other community health indicators. While school attendance records provide a starting point, they do not always capture the root causes of absence. Health-related factors and mental health issues often play a significant role, as well as structural issues like housing instability or involvement in the juvenile justice system. Moreover, many schools track overall attendance rates but fail to examine patterns of chronic absence at an individual level, which can obscure the true scale of the problem.

To address this, our project will take a data-driven, diagnostic approach. By leveraging both national and local data, we aim to uncover trends that can inform targeted interventions. A key challenge will be ensuring that our findings do not reinforce harmful assumptions—such as the notion that chronically absent students or their families simply lack motivation or concern for education. Instead, we will emphasize the structural and systemic factors contributing to absenteeism and explore solutions that engage students, families, and community partners in a collaborative effort to improve school attendance.

A major deliverable for this project is the development of a Community Health Index, a tool designed to provide an accessible, data-driven assessment of community well-being. For this index to be effective, it must be both methodologically rigorous and easily interpretable by policymakers, community organizations, and the public. A key question we must address is which quantitative methods and weighting mechanisms will ensure that the index is both meaningful and actionable. Additionally, to

maximize the impact of our work, we must translate our findings into intuitive, visual tools that allow both technical and non-technical stakeholders to engage with the data. We need to determine which visualization techniques will best communicate our insights in a way that drives action while avoiding any risk of misrepresentation.

To measure the success of our project, we will use several key metrics. First, we will assess data quality and timeliness by identifying and incorporating at least one new, locally sourced data set, ensuring it is current and reliable. We will also analyze any limitations or biases within the existing data. Second, we will evaluate the analytical impact of our work by applying appropriate statistical techniques to uncover meaningful insights and validating our methodology to ensure its robustness and replicability. Third, we will track the development and usability of our Community Health Index, aiming to create a prototype that effectively synthesizes key community indicators and demonstrates its practical value through testing and feedback. Fourth, we will measure the adoption and usability of our tools by conducting at least one pilot test with community partners to gauge how easily they can apply the findings in real-world settings. Finally, we will compile our findings into a well-organized final report that clearly explains our methodology and results. By tackling these challenges and tracking success metrics, our goal is to deliver an analytically rigorous and practical solution, providing the community partners with tools to improvements in community health and safety.

Open Loops

One of the primary open loops in our project is the selection of relevant datasets. We have access to an extensive Excel sheet containing multiple datasets categorized by different themes. While we are confident that a suitable data set exists to address our project objectives, a challenge is in reviewing and filtering these datasets to identify the most relevant ones. This process involves determining which datasets align best with our research goals, ensuring data quality, and evaluating their applicability. Closing this open loop will occur once we finalize our dataset selection, allowing us to move forward with our analysis and model development.

Another open loop is the development of a health index aimed at identifying communities that are affected by high crime rates, drug trafficking, and other socio-economic stressors. While we have a conceptual understanding of the health index, we need further information to determine the appropriate models and equations required to construct it effectively. A key challenge in this process is establishing the appropriate weighting for different crime types, such as whether aggravated assaults should have a greater impact on the index compared to drug-related offenses. The closing loop for this aspect of the project will be reached once we collaborate with our community partners to finalize the model structure, define the index's parameters, and agree on the appropriate weightings for different factors.

Investigating the relationship between school absenteeism in young children and future crime rates would be another open loop for our project. Research has shown that children who miss school frequently at a young age are at a higher risk of becoming involved in criminal activity later in life. We will need to determine whether our dataset includes school attendance records and, if not, explore external sources that track absenteeism trends. Additionally, we will need to decide on the most effective statistical approach to quantify this relationship, options include correlation analysis, time-series forecasting, or classification models to predict future risks based on early education patterns.

Closing this loop will involve finalizing our data sources, selecting an appropriate predictive model, and collaborating with community partners or education departments to ensure the insights are actionable.

The final open loop involves reviewing and understanding the code from previous student teams who have contributed to this project. The existing code provides a foundation for our work, but it requires examination to comprehend its structure, functionality, and potential areas for improvement. This process may take some time as we analyze the logic and determine how to build upon or refine the existing framework. Once we have fully grasped the previous implementation and integrated any necessary modifications, this open loop will be closed, allowing us to continue with the project.

By systematically addressing these open loops and refining our technical approach, we aim to develop an effective solution for identifying at-risk communities and providing valuable insights to our stakeholders.

Success Metrics

To evaluate the success of the Act-NOW: Community Health Index project, we will measure several key factors that reflect both the technical and practical impact of our work. First, we will assess the quality and timeliness of the data, ensuring it meets our internal standards as well as the expectations of our community partners for accuracy and reliability, and that we incorporate up-to-date datasets, particularly by sourcing at least one local dataset per key indicator. We will also focus on data completeness, aiming to reduce missing values and ensure that the necessary metrics are covered.

For model performance, we will track metrics such as F1-score, Mean Squared Error (MSE), and AUC-ROC to measure the accuracy and predictive power of our models. A key goal of ours is to balance model precision with interpretability, ensuring stakeholders can easily understand and apply the results.

The Community Health Index will be validated by comparing it to expert evaluations and by measuring how effectively it can guide decision-making, aiming for a user adoption rate of at least 50% from community stakeholders. Impact and adoption rate will be tracked through the usability of our visualizations, with success defined by a score of at least 4 out of 5 on usability tests and through the application of our findings in real-world community health interventions. We will gather feedback from stakeholders to ensure the tools are practical and actionable, aiming for 70% of users to report the tools as valuable in improving community health and safety.

Finally, we will measure the overall completion and quality of deliverables, ensuring that all key components of the project, including the index, reports, and visualizations, are finalized and well-documented. The success of the project will be reflected in the effectiveness of the Community Health Index in supporting community-based interventions and policy improvements.

Division of Labor

Our team is made up of individuals with diverse strengths and experiences, each contributing unique skills to ensure the success of the project. Gabriella has experience in Python, R, and data analysis, and she is particularly strong in organization. This allows her to connect data insights to real-world applications. She plans to contribute with her written skills, organizational abilities, and by providing analytical insights on the data. Gabriella’s personal goal is to improve her ability to explain insights and

the actions to be taken based on the analysis to stakeholders and others, as well as to gain a better understanding of what it's like to work on a full project from start to finish.

Naamna brings her experience with Python, R, and various machine learning models to the team. She is skilled in notetaking and time management, and her contributions will be focused on applying these technical skills to the project's data analysis. Naamna's goal for the project is to strengthen her proficiency in machine learning models and further hone her time management and collaborative skills.

Ritik has a strong background in data analysis with Python and R, and he is familiar with machine learning models such as random forests and decision trees. He also has expertise in data visualization techniques and Excel. Ritik's main contribution will be in analyzing the data, creating machine learning models, and visualizing key trends. His goal for the project is to deepen his understanding of data science and analysis in a real-world context, as well as to become more proficient in Power BI and Tableau.

Ritin is proficient in Python, R, C++, SQL, and Excel, and has experience in data analysis, data visualization tools, and machine learning algorithms. His contributions will focus on applying these skills to analyze and visualize community health data, while also helping to implement machine learning techniques. His personal goal is to further develop his skills by applying them in a real-world setting, learning more about factors affecting community health, and utilizing Power BI to enhance the project's outputs.

Together, each team member will leverage their strengths to support the project's development, while working toward their individual goals and building valuable experience in data science.

Communication

Our team communicates regularly using a combination of Microsoft Teams, a group text chat, and in-class discussions. We make it a point to stay connected daily or every other day, ensuring that everyone is up to date on ongoing tasks, updates, and progress. Microsoft Teams serves as our central hub for sharing documents, meeting notes, and other important information, while the group text chat is used as a quicker, more informal way for us to coordinate in real-time.

In terms of engagement with our community partners, we have planned weekly meetings with them, which we have tentatively scheduled for Wednesdays at 4 p.m. These meetings will be an opportunity for us to discuss project updates, share insights, and gather feedback with our community partners. In addition, we will maintain communication with our community partners through emails with any questions or important updates regarding the project. This ensures that we are consistently aligned with our partners and can address any concerns or changes in project direction.

In case communication were to break down for any reason, we would first attempt to reach out directly via alternative channels, such as phone or email. If we are still unable to get in touch, we would inform our instructor and ask if he can help facilitate communication with the community partner from his end, leveraging his existing relationship and access to ensure that the project remains on track. We aim to be proactive in maintaining open lines of communication throughout the project to minimize any disruptions, stay on track with our projected timelines, and keep everyone involved informed and engaged.

TimeLine

This is all subject to change, however our project timeline is structured in a week-by-week manner to ensure steady progress and avoid last-minute work. During the first week, we will set up meetings with community partners and define our project scope, laying the groundwork for the semester. In the second week, we will develop a video storyboard for our project sponsor and acquire the dataset to begin familiarizing ourselves with the data. By this time our team will be in the closing steps in understanding which datasets will be included in our project. In the third week, we produce a video proposal for community partners-a 3-5-minute video-along with initial visualizations in exploring different types of correlations among various variables, also during this week a model will be selected to close the loop.

In week four, a blog post about introducing the project and community partners is published as we are developing a feasible model framework and start doing the initial development of models. During the fifth week, we should develop a user instructional guide on how to use such software and carry out demo sessions of such software. Also, during the sixth week, we will work on a detailed storyboarding for the minimum viable demoing of projects, keeping in mind that now we will have at least some prototype in working condition.

In week seven, we will be producing a second 3-5 minute video of the minimum viable project. In week eight, we will be writing detailed instructions of how one can reproduce the figures in our project and further refine and enhance our model. The week of April 6 will be a finalization of the model, inclusive of adjustments, cleaning up the Git repository structure, and adding comprehensive documentation.

As we approach the final weeks, we will storyboard and script our final presentation video. In the last week, we will make any final adjustments to the project, write and submit the final report, submit the final project video, and clean and organize all project files for submission. This structured timeline will help us stay on track and complete each phase efficiently while ensuring a high-quality final deliverable.

Anticipating Challenges

One of the challenges we anticipate is issues with data preprocessing. There are still a lot of unknowns with the type of data we will be working with, as well as the format, but we will work to remove N/A values and get rid of any duplicates.

Another challenge we may have will be working on creating a standard for the Community Health Index. We will need to do quite a bit of research to figure out which attributes will become parameters in our model. Along with this, we will need to figure out what exactly our project deliverables will be and set tentative deadlines for them.

We anticipate that there may be challenges with adding to the project as there is content from a previous group working on it. Processing the previous materials and understanding where the code left off may be a tedious process. We anticipate working with the community partners to get a good background of what the previous team completed and what is to be added.

We also anticipate challenges with how much data we will have, and if that will be enough to complete the project well. As a backup, in the case that we do not have enough data, we will find a relevant appropriate public dataset to use as additional data.

We anticipate illness during the course of this semester. If one of our group members gets sick, the rest of the group members will temporarily fill in for meetings and other project commitments that the ill member is not able to. Then, once the group member has recovered, they will make up the time by temporarily taking on more project commitments.

Lastly, we may anticipate timing delays from both sides. In the event that there is a significant timing delay in terms of a virtual meeting, the delayed party may give a heads up via email to the other party. If the delay is significant enough, the meeting should be rescheduled to a later time.

Appendices

A data dictionary (or schema) will be one of the key appendices. This will include a detailed description of all the data fields being used in the project. For each dataset, we will document the type of data, the possible range of values, and the meaning of each field, as well as any potential data errors or inconsistencies. We will also describe how the data was collected, including any limitations or biases in the data collection process, and provide information on the source of the data (e.g., government agencies, community organizations, open-source platforms). This will allow stakeholders to better understand the context of the data and its relevance to the Community Health Index.

In addition to the data dictionary, we will include metadata that provides insights into the data's creation and usage. This will include documentation on how the data was cleaned, any transformations applied, and details on missing values or imputed data. For example, we will clarify the methods used to handle missing data, such as imputation techniques or exclusion criteria.

This section will also outline any known issues with the data, such as inconsistencies across datasets, missing or incomplete records, or discrepancies between local and national data sources.

Another potential addition to the appendices could be visualizations of the data itself, providing example charts, graphs, or maps that help explain the patterns or trends observed. These visual aids can help stakeholders better understand the insights derived from the data and see how they contribute to the development of the Community Health Index.

Finally, references to all the tools, libraries, and methods used in the project will be included to ensure transparency in the approach. This will also include any relevant academic research or articles that support the methodologies and assumptions used in the analysis, helping stakeholders understand the theoretical foundations behind the project.