February 2024 // Issue 17

www.thedatajam.org

The DataJam Download

Official Newsletter of The DataJam

Check the DataJam out on Instagram!

Meet the Data Science Professional

DataJam at the Science Olympiad

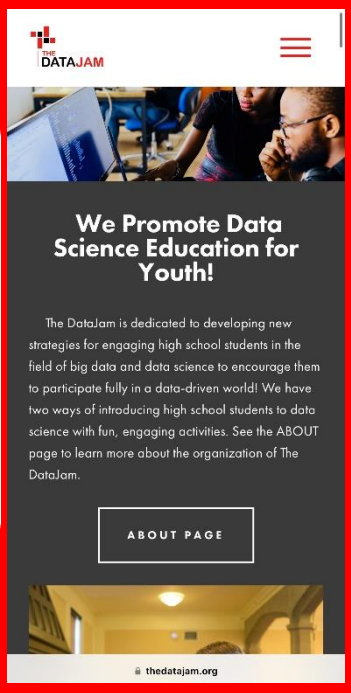
Meet the DataJam Mentors

New DataJam Resources!

DataJam Team Updates

Check The DataJam out on Instagram!

THE_DATAJAM



Join us on Instagram for a thrilling journey that inspires future data scientists. From high school competitions to fostering data fluency, we're dedicated to igniting curiosity and passion for STEM. Follow us for updates, insights, and a glimpse into the dynamic realm of Big Data and Data Science. Don't miss out! Explore #DataJam and join the conversation today!

Exciting news! Our website has been reformatted for mobile devices, making it easier than ever to access valuable resources and stay connected on the go. Check it out at thedatajam.org!

Meet the Data Science Professional

Hello! I'm Dr. Raja Sooriamurthi, a Teaching Professor in the Information Systems Program at Carnegie Mellon University and the Inaugural Program Director of the Decision Analytics and Systems minor. In various capacities, I've been involved with DataJam since its inception in Fall 2013 and currently serve on its advisory board. The motivating goal that led to the creation of DataJam 10 years ago, and which continues to be DataJam's primary goal, is to provide an opportunity and support for high school students to explore and experience the role data will play in their future. Not everyone will pursue a career as a data scientist. But no matter what major one studies in college, be it History, English, Biology or …, and what career one eventually chooses, data is going to play a vital role. Providing high school students, a learning environment to explore the power, potential, and peril of harnessing data is what DataJam has been about and will continue to be about

My own personal journey has been amazing to date and has evolved in ways I could not have predicted in high school. I did my undergraduate degree in computer science and engineering back home in India and came to the US to do my masters (in computer science) and my PhD (in AI and machine learning). I've been a professor all my career and started off in computer science. Then I moved to the Kelley School of Business at Indiana University. That was an eye-opening experience. I grew up as an engineer thinking that every problem had a technological solution to it. When I moved to a business school, I soon realized that technology is only a means to an end. Information Systems takes that strong value adding perspective to technology and I moved to the IS Program at Carnegie Mellon University in 2007.

I mention my varied professional experiences to highlight a few thoughts for your consideration. (1) Given the same problem the way an engineer looks at it, the way a computer scientist looks at it, the way a businessperson looks at it, the way an IS person looks at it is slightly different. Each perspective brings different nuances to the discussion, and it is valuable to leverage diverse perspectives. (2) In the years to come, chances are that your own career will not be a straight line. In dealing with the future and things to come there are two "P's" in your toolbelt — one is planning and the other is being prepared. Of these two I've personally found being prepared much more valuable than planning. Being prepared means doing what you are doing now, well. When we plan, we are planning for the future. The one certainty about the future is that it is unpredictable. The moment the situation changes, the plans we have developed may become ineffective. As Dwight Eisenhower (34th President of the United States and Supreme Commander of the Allied Forces during World War II) said "In preparing for battle I have always found that plans are useless, but planning is indispensable."

As a professor, I teach a range of courses dealing with the broad theme of "data": database systems, data science, machine learning, big data. Along with hardware, software, and communication data is the fourth pillar of our modern world. We've barely started tapping the potential of data. DataJam is a wonderful place to start exploring and experiencing that potential.

DataJam at the Science Olympiad

Founded in 1984, Science Olympiad is the premier and largest team STEM competition in the nation. Divided between two divisions, Division B ranges from grades 6 to 9, while Division C ranges from grades 9 to 12. Across all 50 states, it tests around 6,000 teams at over 425 competitions across 23 different events from various fields of science. Events differ in style, ranging from written tests, engineering projects, to laboratory experiments.

On January 20, 2024, the University of Pittsburgh hosted a Division B invitational competition which over 200 middle school students attended, and Carnegie Mellon University (CMU) hosted a Division C invitational competition which over 300 high school students attended. Two DataJam mentors, Jatin Singh and Daniel Hufnagle, gave presentations about the DataJam-sponsored Middle School Data Science Day at the Pitt event and about the DataJam at the CMU event.

We hope that several of the more than 500 youth attending will participate in future DataJam events!

DataJam Presentation at Division B Science Olympiad on January 20, 2024. Photo by Bhaskar Chakarbarti

DataJam Presentation at Division C Science Olympiad on January 20, 2024.

Meet the DataJam Mentors

Hi! I'm Bhaskar Chakrabarti and I'm a sophomore studying Neuroscience with minors in Chemistry, Applied Statistics, and a Certificate in Conceptual Foundations of Medicine at the University of Pittsburgh. I'm so excited to serve as a DataJam mentor this year! The importance of statistics and data science cannot be overstated, as it's used in almost any field you can think of.

Originally from New York City, in high school I was heavily involved in Science Olympiad, a student competition in the sciences which include numerous categories or "events" such as Sounds of Music, where teams build a functional musical instrument and complete a test on the physics of sound and music theory fundamentals, and Experimental Design, where teams must develop and run a short experiment on a broad topic in less than 50 minutes. Science Olympiad is a collaborative effort, and teams travel to a variety of competitions at the regional, state, and national levels, as well as "invitationals" for practice.

I also conducted neuroscience research at the Albert Einstein College of Medicine which allowed me to get early experience in a laboratory setting and master the fundamentals of the scientific method, including research and data literacy. I was also fortunate enough to present my research at a number of science fairs and

professional conferences, gaining experience in communicating science to broad audiences. It was both the Science Olympiad and my research experience which sparked my interest in scientific inquiry and led me to pursue further studies in the sciences.

At Pitt, I conduct research on Multiple Sclerosis, where in addition to clinical tasks, I perform longitudinal data analysis on various patient data points. I also work as a student ambassador for the Dietrich School of Arts and Sciences, where you can find me giving tours and representing the school at events. In addition, to give back further to the Science Olympiad community, I am the Director of Logistics in Science Olympiad at Pitt, where just last month we hosted middle school teams at our student-run competition and were able to spread the word about DataJam!

In my free time, I enjoy photography and going on long bike rides. I can't wait to help you use data science to answer any question you come up with!

Hi, I'm Eeshaan Jadhav! I'm currently a Junior, statistics major at Pitt with a focus on data science. Over the past couple of years, I've had the opportunity to use data science to optimize a predictive genetics program, and as a technical writer for industrial modeling software. I'm really interested in technical optimization and finding effective ways to communicate results. Outside of statistics, I love hiking, getting involved on campus through my fraternity, and finding new restaurants around Pittsburgh.

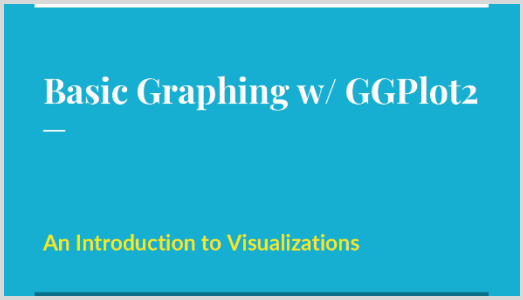
One of the reasons I love data science is because of how applicable it is. Every field can benefit from the use of data analysis whether it be by learning more about their field or by optimizing various processes. This will be my first year with the DataJam so I'm excited to work with and mentor a group of students as they learn more about data science. I hope the DataJam gets all of you as excited about data science as I am!

As I've gone through the statistics major at Pitt, I have had the opportunity to get very familiar with R through my classes. One of my favorite aspects of R is the ability to use such a wide variety of plugins that simplify and improve the data manipulation and visualization functions. One plugin that I've become familiar with is GGPlot. The variety of different graphs GGPlot offers is excellent, but because there are so many options it can be a little intimidating to use. When I started using GGPlot, I constantly had to look up the code for these graphs so I thought creating a reference that provides example code and some context about when to use these different graphs would be beneficial for students learning to use R. I hope you find my guide to "Basic Graphing with GGPlot2" to be helpful!

Hi everyone! My name is Neha Dutt, and I am a senior at Pitt majoring in Computational Biology. My goal is to become a strong data scientist who can apply the commonly used technologies in various aspects and fields. Outside of school, I enjoy reading and dancing, and in fact, I am currently on a competitive dance team at my school! I'm also on the board of a club that organizes a classical dance competition in Pittsburgh! This year, for DataJam teams, I've created a guide called "Making Sense of Your Dataset". I know that finding the right dataset can be quite challenging, so I hope this guide proves to be helpful. It'll walk you through all the important questions to consider when finding a dataset to match your general topic and question, and help you finalize a dataset that is the most representative and of a size that is manageable for your team to analyze. Some of the important points the guide urges you to consider are about the relationship of your variables, the scope of your

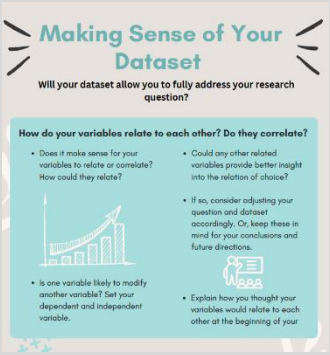
dataset, the representativeness of your dataset, the implications of your dataset, and whether your dataset truly can best answer your question. I look forward to the upcoming DataJam season, and I hope that this guide can be a valuable tool for your team and project!

New DataJam Resources



Basic Graphing with GGPlot2

This new guide is designed to introduce DataJam teams to the wide variety of graphs that can be made with GGPlot2 and simple directions for how to create these graphs. Strategies for plotting both discrete and continuous variables are included. This guide was developed by DataJam mentor Eeshaan Jadhav (featured this month in the Meet the Mentor column).



Making Sense of Your Dataset

This new guide is designed to guide DataJam teams in evaluating if a dataset that they find will actually work to help them answer the question that is the focus of their DataJam project. It helps teams think through whether there may be missing data, or if the dataset is large how to potentially filter the data to hone in on the pertinent data. This guide was developed by DataJam mentor Neha Dutt (featured this month in the Meet the Mentor column).

How to Do Supervised Learning in R?

This new guide explains what supervised learning is and how to make and use decision trees using R. The guide also covers how to perform logistic regressions in R. The guide ends with explaining a Random Forest and how to make a Random Forest and use it in R. The guide was developed by DataJam mentor Yewon Kim.

DataJam Team Updates

As announced in the January 2024 DataJam Download, we are excited this month to introduce "DataJam Team Updates"! In this section from February to May we will provide space in the newsletter for DataJam teams to write a short paragraph about the project they are working on and provide a figure if desired. Our goal is for teams across the country to have more communication about their DataJam projects with each other. If your team would like to provide a DataJam Team Update for the March issue of the DataJam Download, please email your submission to datajam@thedatajam.org.

Oakland Catholic (Pittsburgh, PA)

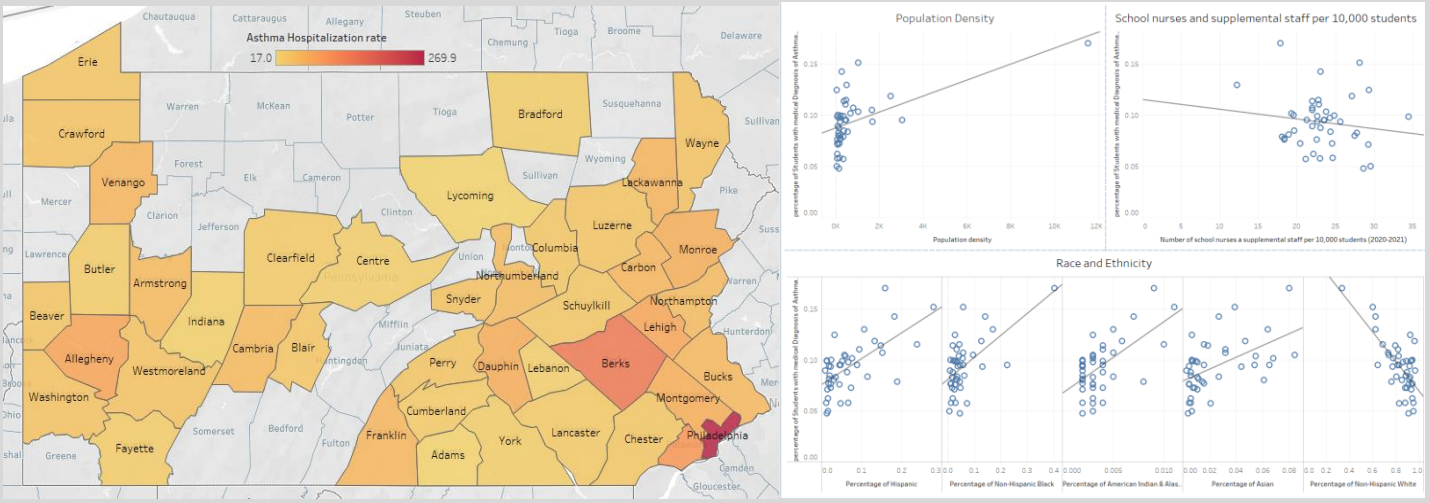
Our Data Jam project this year is about the relationship between the cleanliness of Pittsburgh parks and the crime rates in the surrounding area. First, we are looking to understand the relationship between park location and location of crimes. Then, we will seek to understand the relationship between the cleanliness of a given park (based upon how long it has been since the last clean-up) and crimes committed nearby. We are not very far along yet, but we are working towards creating graphs to interpret soon. We are really excited to present our findings in April!

North Allegheny (Pittsburgh, PA)

Asthma is a chronic condition that impacts 9.9% of children in PA; that's almost one in every 10 children! Understanding factors that impact asthma hospitalization rates and diagnosis most significantly in children can

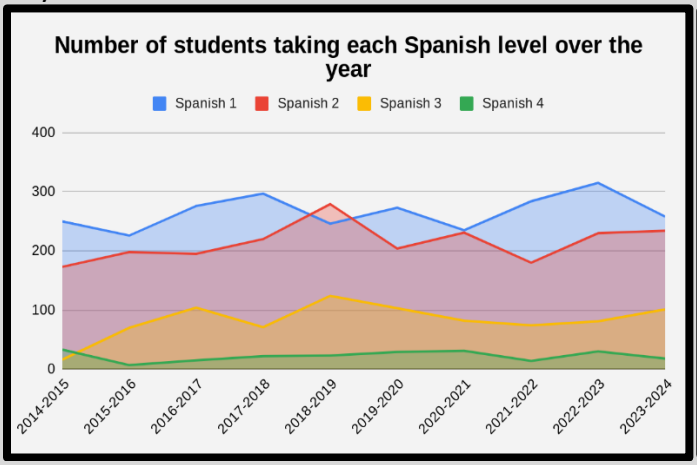
help millions worldwide. Asthma may be affected by a wide variety of things: genetic history, race, pollution, poverty, behaviors, population density, accessibility to healthcare, etc. We started by hypothesizing that population density would be the factor with the most drastic impact. We searched for datasets and our main data set was “Asthma School Health Data Report by Count Asthma Control Program”. This included information about many of the factors listed above. Additionally, we also used data from other sources to get information on population density, air pollution, lead poisoning, tobacco retail, etc.

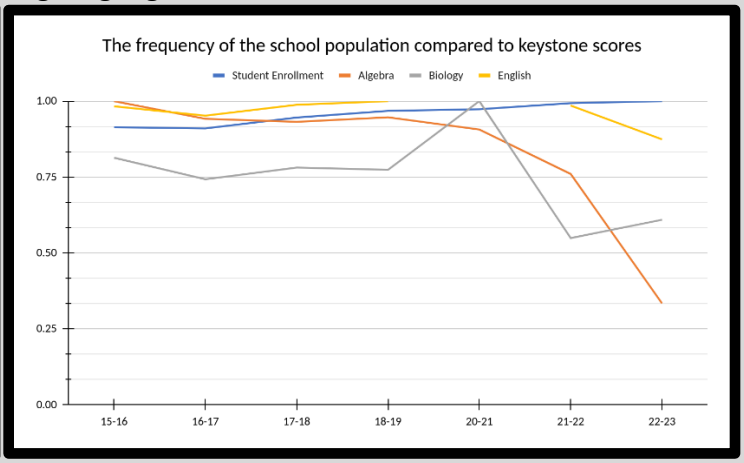
To narrow down the number of variable factors, we first ran a linear regression model to find each factor’s p-value. A p-value less than .05 made it significant. Before running a further multilinear regression model, we determined the correlation between factors to remove any multi-collinearity. With the significant factors, we then ran a multilinear regression to calculate their impact on hospitalization rates. This process was then repeated for its impact on the diagnosis rate. Through research and all these tests, simply put, we can conclude that population density, and the percentage of Hispanic population were the two main impacting factors, with air pollution, lead poisoning, and tobacco smoke also having significant impacts.



Central Dauphin RAWMAN (Harrisburg, PA)

The RAWMAN DataJam team at Central Dauphin High School is located just outside of Harrisburg, Pennsylvania. We are researching why there is such a drastic decline in the number of students in the higher levels of language classes. To do this, we plan to survey students at all levels of every language class to get their reasons on why they did or did not continue into higher levels of their chosen language. While researching, we came across a data set relating the increase in total enrollment with the decrease in the percentage of students passing the state standardized tests. This could explain or show how the number of students per class can affect an individual's learning experience. In the end, we hope to understand better why students at our school and others are not continuing language education.





Carlynton (Pittsburgh, PA)

The Carlynton DataJam team decided early on that we wanted to focus on education and the pandemic, but that is too broad a topic. We turned to our own school, examining, and discussing issues that both students and teachers see emerging. Our mentor helped us narrow our proposal down to special education, funding, and how the pandemic affected the two variables. Thankfully, there's an abundance of data that is available to the public regarding public schooling across PA, such as the datasets in pvaas.sas.com. We've found multiple sites with data, and sites that have all that data compiled for easy access. We hope to find a trend in special education and its funding and analyze how the pandemic changed the trend - for better or worse.

Brooke (Boston, MA)

We set out to figure out what replicable aspects of students' daily lives have the greatest impact on their GPAs. This can help teachers learn to improve student performance so that students can learn methods to improve their GPAs. We thought this was important because school is too stressful and finding factors to make school less stressful would be good.

First, the data science club announced during our daily community meetings at school our goal of determining how mood and GPA correlated with each other. To gather volunteers, we sent out an email form for students to express their interest. Then, one of our mentors sent out an email every 15th 30th / 31st of each month since November, and every Tuesday we came together to discuss the data set. We determined collectively different questions that might bring us closer to our overall goal. Also, in every form, there was a feedback section for the volunteers to leave us a comment, so we could improve our forms. Every feedback that was left was implemented into the next survey to ensure it would continue to be effective throughout the process.

Currently, we hypothesize GPA will go up and down based on sleep levels, studying, and phone usage. Based on the data we collected so far in our 3 surveys, physical health, studying, and homework completion help GPA the most. Despite these having a lot of impact so far, we want to continue to understand the data by testing how much these factors impact GPA independently. We plan to give our volunteers different amounts of time to study, sleep, or do other daily things to see how it affects their GPA. Other than changing the way we get data; we plan on continuing to use r^2 to see the correlation and scatter plots.

DataJam Timeline for 2023-2024

On the DataJam page of the website the new 2024 DataJam Timeline has been posted. Click here to see the Timeline.

- **Posters will be due Fri., March 29, 2024**
- **2024 DataJam Finale will be Thur., April 25, 2024**

January to Early April

Work on DataJam Projects

Teams will be able to work on their DataJam projects. Send an email to DataJam@pghdataworks.org to arrange meetings with DataJam mentors, who are available to help with all aspects of the DataJam projects. Mentors can also be reached directly on the DataJam 2024 Slack workspace.

Friday, March 29, 2024

DataJam Posters Due

Teams should email their DataJam poster to DataJam@pghdataworks.org. Instructions for the poster are in the DataJam Guidebook. Posters should be 24"x36" in size and submitted as a PDF.

The instructions for writing the DataJam Proposal are on page 5 of the DataJam 2024 Guide Book, and a template for the one page DataJam Proposal is on page 6. The guidebook can be downloaded from the DataJam page of the Pittsburgh DataWorks website.

PITTSBURGH DATAWORKS

DataJam Guidebook - 2024

We are looking forward to DataJam 2024!
We Hope You Are Too!

Email us at datajam@thedatajam.org when you are ready to start working with a DataJam Mentor!