DATASPACE

USING DATA & STORIES TO BRING DISCIPLINES TOGETHER

AMATYC WEBINAR
JANUARY 20, 2022
SHONDA KUIPER, GRINNELL COLLEGE
Outline

• Briefly introduce goals

• Dataspace game example

• Dataspace story example

• Describe motivation for this work
Goals for DataSpace

Integrate examples that are “real to the students” (Gould, 2010)
• Find patterns that matter (tell a story with your data)
• Deeper meaning and insights so that better decisions can be made.

Critical Thinking: Address common misunderstandings (GAISE guidelines)
• bias, conditional probabilities, assumptions, abstraction

Transition from small/carefully vetted data to large/messy data (2014 Curriculum Guidelines)

Active Learning (Laursen and Rasmussen(2019))
• Deep Engagement, peer-to-peer learning
Goals for DataSpace Games

Students take ownership for their decisions

• Train students to experiment with the data, find their own patterns, and ask their own questions.
  Students learn to handle larger/messier datasets.
  Students have input on what questions are asked even with a common dataset

• Challenge students to think carefully about data and the models they choose to build.
  Interpretation of the numbers is just as important as the calculations!

*Learning is essentially hard; it happens best when one is deeply engaged in hard and challenging activities*
- PApet
Greenhouse Game

• An online environment simulates a greenhouse.
  If players want to make money, they must be careful about the amount of water, fertilizer, and sunlight each crop receives. While players can choose any options they like, they are more likely to profit if they use the interactive graphs and statistical models to improve their strategies.

• Each season players
  • buy seed
  • plant crops
  • harvest crops
  • sell crops

• Initially the only variable to consider is water.
  • Students add different amounts of water to each plot
  • The total amount of water will influence the harvest
  • What amount of water is best?
Here are some key terms each player should know:

- **Yield** is the amount of crops produced. It is measured by the number of bushels produced for each plot. In this game, yields reflect actual crop growth in Iowa farmland.

- A **bushel** originally was a large basket used to carry crops. For example, a bushel of corn typically weighs about 56 pounds.

- **Water** is the amount of water (in inches) that you assign to each plot.

- A **plot** is the area of ground where you are planning your seeds. In this Greenhouse game, each test plot represents 1/10th of an acre of land.

- The game will ask you for a **PlayerID**. You can use any name you like, but this will be on the internet, so do not use a name that will easily identify you.

- If you are playing this game for a course, your instructor will provide a **GroupID** so you will be able to compete with everyone in your class.
Greenhouse Level 1

• Estimate Corn yield based upon the amount of water.
• Enter a **PlayerID**: You can use any name you like, but this will be on the internet, so do not use a name that will easily identify you.
• Enter the **GroupID**: AMATYC1
• Use the following settings:
After you select Randomized Play, you will be able to:

- Move your cursor over each plot to identify the crop type and the amount of water the plot has received.

- Click the Harvest button on the right. Then click Sell All to sell your crops. The game will display how much profit you made.

- After you have harvested and sold your crops, enter your data using the Check Your Understanding form below.
Online visualizations and auto graded questions ensure students understand the core ideas.
Discussion questions and competitions encourage students to take ownership of their decisions.

Grades are not based upon the competition!
Let’s Play!

https://dataspace.sites.grinnell.edu/greenhouse1.html

If you don’t have access to a computer, watch the video here:
https://www.youtube.com/watch?v=F_pdfB-y5sU
Greenhouse Level 1

After growing the corn, the farmer sold the corn for $4 a bushel. Thus Revenue = 4*(Yield). If we use Revenue instead of profit, the correlation coefficient will:

- Increase
- Decrease
- Stay the same

The correlation = 0.914, does this mean we can trust this model to determine the best amount of water to use?
Greenhouse Level 1

Modify the X Axis Limits in the Greenhouse Models app. When is a linear regression model a good fit for this data?

- When $X \leq 25$
- When $X \leq 30$
- When $X \leq 35$
Questions???
Greenhouse Level 1

- Conduct a hypothesis test to determine if we can show that yield is better when water is 35 instead of when water is 25.
  - How many samples do we need?
  - Is it possible to observe a plot where yield is higher when water = 25?
  - What would a large p-value mean?
  - What are the challenges of testing water = 0 versus water = 10?

- Regression estimation, prediction, and extrapolation
  - Estimate your yield when Water = 10, 20, 30, and 50
  - When are your estimates accurate?

- Confidence intervals
  - If everyone in the class played the same game and used the same amount of water, why is there variation?

- Model assumptions and residual plots
  - What should we do with outliers or unusual observations?
  - Is a linear model appropriate?
  - What happens if your quadratic model predicts negative yields?
Advanced Options

• Each plot in the game is based upon actual yields for 1/10 acre of land.

• **Multiple variables**: amount water, type of crop, amount of fertilizer, amount of pesticides, cost of seed and sale prices
  
  • Can you build a model the will help you optimize yield?

  • Which crop will optimize profits? Profits depend on yield, the amount of water and the amount of fertilizer

Coefficients:

| Term                        | Estimate | Std. Error | t value | Pr(>|t|) |
|-----------------------------|----------|------------|---------|---------|
| (Intercept)                 | -6.808e+00 | 1.196e+00 | -5.691  | 1.82e-08 *** |
| WaterB                      | 1.029e+00  | 5.218e-02 | 19.712  | < 2e-16 *** |
| WaterSquaredB               | -2.873e-02 | 1.448e-03 | -19.843 | < 2e-16 *** |
| NitrateLevel                | -1.518e-02 | 2.732e-03 | -5.556  | 3.86e-08 *** |
| SqrtNitrate                 | 6.029e-01  | 1.032e-01 | 5.841   | 7.79e-09 *** |
| Insects                     | -1.404e+00 | 1.665e-01 | -8.431  | < 2e-16 *** |
| WaterSquaredB:NitrateLevel  | 4.140e-07  | 1.828e-06 | 0.226   | 0.821   |

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.036 on 737 degrees of freedom
Multiple R-squared:  0.6289, Adjusted R-squared:  0.6259
F-statistic: 208.2 on 6 and 737 DF,  p-value: < 2.2e-16
Goals for DataSpace Stories

• Incorporating data journalism into courses in order to provide project-based materials that emphasize real-world applications and conceptual understanding.

• Create materials that are designed to ease the workload of faculty while still incorporating research-like experiences into their own classes.
Structure

• **Introductory Article:** We will start with an easy-to-read article

• **Interactive Apps to Investigate Claims:** Throughout the online article, we will provide interactive data visualizations, data tables and/or statistical models to explore claims made in the article.

• **Additional Questions to Investigate:** We will provide additional lists of questions for readers to explore by modifying each of the data visualizations, data tables or models.
Exploring Racial Disparities in New York City’s Stop-and-Frisk Policies

By Shonda Kuiper and Yusen He

Topics: Data Visualizations, Bias, Multivariate Data

Many individual cases have been found to show evidence of discrimination by police officers. However, showing evidence of systematic patterns (e.g., clear evidence of bias across an entire city) is more difficult. It is essential to consider the type of data available, how the data was collected, and how the data was organized. In this example, we use several New York City crime and education datasets to describe how easy it is to draw incorrect conclusions from these types of data. The emphasis is not on conducting a single statistical test that provides one undisputable answer. Instead, exploratory data analysis, visualization and modeling are used to make decisions with complex and messy data.
NYPD Stop and Arrests

- **CLAIM 1**: There is evidence of racial discrimination in the NYPD stops and arrests.

- Every year, the New York City Police Department (NYPD) stops individuals for suspected criminal involvement.

- “This is a proven law enforcement tactic to fight and deter crime, one that is authorized by criminal procedure law (Long 2009).”

- In recent years, the NYPD had been accused of being racially discriminatory in their stops and arrests.
Figure 1A: Bar graph representing the total number of people arrested in New York by race of the suspect from 2006 to 2016.
Figure 1A: Bar graph representing the total number of people arrested in New York by race of the suspect from 2006 to 2016. Figure 1B: Bar graph of the percentage of arrests (Total Arrested/Total Stopped) for each race between 2006 and 2016.
NYPD Stop and Arrests

- **KEY IDEA**: Whenever we are shown a percentage, we should always ask the question, “Percentage of what?”

- Figure 1A shows that about 50% of all arrests made in New York involve a black suspect.

- Figure 1B shows that just over 6% of Asians, Hispanics, and whites are arrested after they are stopped.

- The core issue is that the denominator chosen can be mathematically accurate, but lead to very different conclusions.
NYPD Stop and Arrests

• **CLAIM 2**: After a 2013 court ruling, there has been a reduction in the racial disparity in police stops.

• Figure 2 provides shows a significant decrease in the number of police stops after 2012.

• In 2014, the New York Civil Liberties Union stated that New York City Mayor Bill de Blasio had "... made stop-and-frisk reform a central issue in his campaign, and shortly after his election he moved aggressively to honor his campaign promises."
NYPD Stop and Arrests

- **CLAIM 2**: After a 2013 court ruling, there has been a reduction in the racial disparity in police stops.

*Figure 2: The total number of police stops by year and race in New York City*
NYPD Stop and Arrests

- **CLAIM 2**: After a 2013 court ruling, there has been a reduction in the racial disparity in police stops.
NYPD Stop and Arrests

• **KEY IDEA**: Data can be easily manipulated to support a particular preconceived notion.

• Summarizing a complex dataset with only one graph (or one hypothesis test) can easily misrepresent the true patterns within the data.

• Objectively look at the entire dataset before drawing conclusions.
Let’s Go!!

https://dataspace.sites.grinnell.edu/nypd1.html

In 2018, what suspected crime type had the most frisks?

- Assault
- Theft
- Possession of an illegal substance
- Possession of an illegal weapon
Questions???
Let’s Go!!!

• Has the pattern in the percentage of arrests changed over the past ten years?
• Has the amount or type of force used in a stop changed over time?
• Are there disparities in the stop or arrest data related to gender?
• What patterns occur when the data is restricted to a particular type of force, such as restricting the data to only stops where firearms were used?
• What percentage of arrests involved cases where the police drew a weapon (handgun, Taser, pepper spray or baton)?
• Are there any relationships between the types of force used and the suspected crime type? For example, are firearms used more often when the suspected crime is a felony instead of a misdemeanor?
The Growing Power of Data

• Data is easily available
• Growing interest in data analysis
• Technology has changed the discipline of statistics
• Making decisions with data in an essential life skill
The Growing Power of Data

- Data is easily available
- Growing interest in data analysis
- Technology has changed the discipline of statistics
- Making decisions with data is an essential life skill

McKinsey & Company (Manyika et al., 2011) has predicted shortfalls of 150,000 data analysts and 1.5 million managers who are knowledgeable about data and their relevance.
Challenges in adapting to the age of big data

Calculating a p-value, does not ensure correct decisions will be made.

• Students who take only an intro course are no longer equipped to apply the more relevant statistical methods in their own work\(^1\)

• The standard approach in teaching, of stressing the formal definition of a p-value while warning against its misinterpretation, has simply been an abysmal failure.\(^2\)

Help students focus on interpretation (abstract ideas) as much as the calculations (concrete rules)

---

Students who take only an intro course are no longer equipped to apply the more relevant statistical methods in their own work.

The standard approach in teaching, of stressing the formal definition of a $p$-value while warning against its misinterpretation, has simply been an abysmal failure.


Challenges in adapting to the age of big data


StatsLife

Psychology journal bans $P$ values

Test for reliability of results ‘too easy to pass’, say editors.
Math is Music Math is Music – Stats is Literature
- Dick De Veaux

The math part (Concrete Ideas, Follow the Rules)
– Give them the formula, they can get the answer with some training

The hard part (Requires More Abstraction, Guidelines)
– Putting it all together
  • Real world
  • Experience
  • Methods
Tyranny of technique

It is relatively easy to teach students a set of rules and calculations

• Students dislike uncertainty
• Students … want to look smart even if it means not learning a thing in the process. —Dweck

It is challenging to coach students to think creatively and critically

• Give space to explore questions and unique approaches
• Articulate uncertainty in context when making decisions
• Communicate these decisions with their peers

Growth Mindset-Carol Dweck, a professor of developmental psychology at Stanford University and author of “Mindset: The New Psychology of Success.”
We Need Class Testers!!!

Email kuipers@grinnell.edu

First Day of Classes (or as soon as possible) Have students complete the Online Assignment (10 minutes)
- https://grinnell.co1.qualtrics.com/jfe/form/SV_6G7Qfd5EoEdpY8K

When you are ready to use the GH Lab in your course:
Day 1: (10 minutes at the end of class)
- Show students the website: https://dataspace.sites.grinnell.edu/greenhouse1.html
- Demo the GreenHouse Game, you may want to use the “Tutorial” button.
- Assign Step 1A and Step 1B (individually or in groups)

Day 2: {typically a full class period on the following day}
- Allow students to ask questions about “Check your Understanding” questions.
- Have students work on Step 1C: Additional Questions to Consider individually or in groups
- At the end of class, assign the GreenHouse Challenge as Extra Credit (at the end of page 3).

Within a week of finishing the activity: Have students complete the second Online Assignment (10 minutes)