About QSIDE’s Toy Data

This is the link to access our [toy data](https://michiganstate.sharepoint.com/:x:/r/sites/Section_SS25-CMSE-495-001-225215054-EL-32-A26-QSIDE/Shared%20Documents/QSIDE/Project_deliverables/Reproducibility_Documents/Toy_Dataset/Student_Toy_data.xlsx?d=w69dfa650517a49639e6bc556a33a0a96&csf=1&web=1&e=SZLrf6).

We followed the [Installing\_Excel.docx](https://michiganstate.sharepoint.com/:w:/r/sites/Section_SS25-CMSE-495-001-225215054-EL-32-A26-QSIDE/Shared%20Documents/QSIDE/Project_deliverables/Reproducibility_Documents/Toy_Dataset/Installing_Excel.docx?d=wdc0506b050f24cf5827ef242fee686c3&csf=1&web=1&e=TBwnZk) to create a new Excel sheet for our toy data. We created multiple sheets in order to maintain organization with all of the data that we were generating.

Our toy dataset consists of five different Excel sheets, covering student demographics, metadata, individual school trajectories, school demographics, and student school presence. It includes 25 fake students with attributes. Additionally, it incorporates data from six real Illinois schools and 80 fake excused absence letters to mimic real-world attendance records. All the variables that are included within our toy dataset were discussed prior with our community partners, listed in the [MSU - ACT NOW Data Map](https://docs.google.com/spreadsheets/d/1JXQa1UJz7rqCv2j2kmfTo1-sv1kV0tm0Cg5psaEIetI/edit?usp=sharing) under the “School Presence Data Map” sheet.



We used this outline to structure the sheets and variables for our toy dataset. Through multiple meetings with community partners, we refined and adjusted the data to better fit our needs.

With a clear framework of required variables and values, we leveraged ChatGPT to generate randomized data for most columns and sheets. A document detailing the prompts used is available [here](https://michiganstate.sharepoint.com/:w:/r/sites/Section_SS25-CMSE-495-001-225215054-EL-32-A26-QSIDE/Shared%20Documents/QSIDE/Project_deliverables/Reproducibility_Documents/Toy_Dataset/Chatgpt_Prompts.docx?d=w73865d9c968645f3b81d68fb55af4bfd&csf=1&web=1&e=1d2vAH). Student addresses and real Illinois schools were manually generated.

Once the toy data set was created, we followed the “Exporting Excel files” found within our [Installing\_Excel.docx](https://michiganstate.sharepoint.com/:w:/r/sites/Section_SS25-CMSE-495-001-225215054-EL-32-A26-QSIDE/Shared%20Documents/QSIDE/Project_deliverables/Reproducibility_Documents/Toy_Dataset/Installing_Excel.docx?d=wdc0506b050f24cf5827ef242fee686c3&csf=1&web=1&e=TBwnZk) to export and save it locally. The toy dataset Excel file was then used to extract the GeoIDs and mapping in Tableau.

File Details

Below are the 6 different Excel sheets as well as some information on the data included within each sheet. Regarding the locations included in the toy dataset, we focused on 3 cities within Illinois (Springfield, Decatur, and Peoria).

* StudentDemographics: This sheet gives information regarding the student’s personal data. There are 25 unique students listed within this Excel sheet.
  + StudentID
    - This is currently a stand in for the unique identifier for each of the students
  + StudentFirstName
    - This is the first name of the student
  + StudentLastName
    - This is the last name of the student
  + Race/Ethnicity
    - This is the race/ethnicity of each student
    - The values within this column consist of African American, Asian, Caucasian, Hispanic
  + GenderAssignedatBirth
    - This is the gender assigned at birth of each student
    - The values within this column are Male and Female
  + GenderIdendity
    - This is the gender identity that the student most recently identifies as
    - The values within this column consist of Agender, Agender Neutrois, Cisgender Female, Cisgender Male, Gender Neutral, Genderqueer Androgynous, Genderqueer Femme, Non-Binary, Non-Binary Femme, Non-Binary Masc, Pangender, Pangender Fluid, Third Gender, Third Gender Androgynous, Third Gender Fluid, Transfeminine, Transgender Female, Transgender Male, Transgender Non-Binary, Transmasculine, Two-Spirit Feminine
    - \*We did our best to have varying gender identities to best reflect the numerous gender identities. We used this [source](https://teentalk.ca/learn-about/gender-identity/#:~:text=There%20are%20many%20different%20gender,or%20a%20combination%20of%20these.) as a guide for all the gender identities included in the toy dataset:
  + Address
    - This is the student’s current address \*All of the addresses are real Illinois homes. These addresses were found through [Zillow,](https://www.zillow.com/homes/54522_rid/) inputting the wanted city of each student.
  + City
    - These are the cities that each student’s address is in (Springfield, Decatur, or Peoria)
  + State
    - Because we are only focusing on the state of Illinois, all the states are “IL”
  + ZipCode
    - The five-digit postal code corresponding to the student’s address
  + Birthdate
    - This is the birthdate of each student. This variable will be mainly used to look at the ages of each student for further analysis
  + NumofSiblings
    - This is the number of siblings that each registered student has
  + SES (High/Middle/Low)
    - This is the Socioeconomic status of the student. These are categorized into (High, Middle, or Low) categories. We used [Wikipedia](http://en.wikipedia.org/wiki/Socioeconomic_status#:~:text=Socioeconomic%20status%20is%20typically%20broken,and%20occupation)%20can%20be%20assessed.) to determine these categories. These categories were also discussed and agreed upon with the community partners.
  + StableLivingSituation (Scale 1-5: 1 Not stable - 5 Very Stable)
    - This is a scale of the living situation of the student. 1 is considered not stable while 5 is considered very stable.
  + FoodInsecure (Rating 1-5: 1 Not Insecure- 5 Very Insecure)
    - This is a scale of the food security of the student. 1 is considered not insecure while 5 is considered very insecure.
* ID\_MetaData: This sheet contains a preliminary list of potential values for the StudentID column in the StudentDemographics sheet. The StudentID must be a unique, unchanging identifier assigned to each student, ensuring there is no risk of duplication or confusion, regardless of the school they attend or their grade level.
* IndividualSchoolTrajectory: This sheet gives information regarding the educational path that each of the 25 students is currently on.
  + StudentID
    - This is currently a stand in for the unique identifier for each of the students
    - It references the StudentID that is assigned within the StudentDemographics sheet
  + ExpectedGraduationDate
    - This is the expected graduation date for each student
    - All the graduation dates are between 2024 and 2027
  + CourseofStudy
    - This is the course of study that each student is currently on.
    - Consists of Science, Engineering, Business, Arts, Math, Literature, and Technology
  + GPA
    - On a scale of a 4.0
  + GradeDistribution(A-F)
    - This column has how many A/B/C/D grades that each student has in the gradebook
    - Each student has a total of 6 grades distributed amongst the 4 letter grades
  + PhysicalDisabilityStatus
    - This column details the student's physical disability status
    - Listed disabilities include Visual impairment, Hearing impairment, and Physical Disability
  + Neurodiverse/Learning Disability
    - This column details the student's Neurodiverse/Learning Disability status
    - Listed disabilities include ADHD, Dyslexia, and Autism
  + IEP(Y/N)
    - This value is whether the student is on an Individual Educational Plan
* SchoolDemographics: This sheet gives information about the 6 unique, real Ilinois schools. \*While all of the school names and addresses are real, the remaining values are artificially generated and do not reflect the actual status of each of the schools.
  + SchoolID
    - A unique numerical identifier assigned to each school. This ensures that schools can be easily referenced without ambiguity
  + SchoolAddress
    - The full street address of the school, including building numbers and street names
  + City
    - The city where the school is located (Springfield, Decatur, or Peoria)
  + State
    - The state abbreviation (e.g., IL for Illinois) where the school operates
    - Because we are only focusing on the state of Illinois, all the states are “IL”
  + ZipCode
    - The five-digit postal code corresponding to the school's address
  + SchoolDistrict
    - The numeric identifier of the public school district the school belongs to. \*Private or charter schools may not be directly associated with a public district
  + SchoolType
    - The classification of the school, such as Public District, Private, Charter, or Magnet.
    - This indicates whether the school is government-funded, independently run, or has specialized programs
  + EducationLevel
    - The grade levels offered at the school (PreK-8, 6-8, 9-12)
  + SchoolName
    - The full name of the school
  + TotalSchoolDays
    - The total number of school days in a given academic year
  + SchoolSize
    - The total number of enrolled students at the school
  + Student:TeacherRatio
    - The ratio of students to teachers in the school, displayed in the format X:01 (e.g., 18:01 for 18 students per teacher)
    - Lower ratios generally indicate smaller class sizes
  + YearOpened
    - The year the school was originally established. This helps provide historical context about the institution's longevity
  + Four-YearGraduationRate(Overall | By Race/Gender)
    - The percentage of students who graduate within four years, both as an overall percentage and broken down by race/ethnicity and gender
  + Five-YearGraduationRate(Overall | By Race/Gender)
    - The percentage of students who graduate within five years, also segmented by race/ethnicity and gender. This can highlight differences in graduation timelines among student demographics
  + RankedinDistrict
    - The school’s ranking within its respective school district
  + RankedinState
    - The school’s ranking within the state based on standardized assessment scores or overall performance compared to other schools
  + ACTAvgerage
    - The average ACT score of students at the school
  + SATAvgerage
    - The average SAT score of students at the school
* StudentSchoolPresence: This sheet gives information regarding the attendance, specifically the absences and tardies of each of the students at their recorded school. It also includes made up absence notes for the number of ExcusedAbsences (totaling 80 fake absence letters).
  + StudentID
    - This is currently a stand in for the unique identifier for each of the students
    - It references the StudentID that is assigned within the StudentDemographics sheet
  + SchoolID
    - A numeric identifier linking the student to their school
    - It references the SchoolID that is assigned within the SchoolDemographics sheet
  + DaysPresent
    - The total number of school days the student attended
  + ExcusedAbsences
    - The number of days the student was absent for a valid, documented reason
  + UnexcusedAbsences
    - The number of days the student was absent without an acceptable excuse
  + Tardies
    - The number of times the student arrived late to school
  + DatesAbsent
    - A semicolon-separated list of dates when the student was absent (formatted as MM/DD/YYYY)
  + ReasonGiven (Y/N)
    - Indicates whether a reason was provided for absences
  + AbsenceNote
    - A detailed text description of why the student was absent
    - There is a different, unique absence note for the number of excused absences
    - These were AI generated
  + ReasonCategory
    - A semicolon-separated list categorizing the absence reasons
    - Reasons include Medical, Athletics, Family, Appointment, Mental health, Transportation
* StudentSchoolPresenceUpdated: This sheet is a modified version of the StudentSchoolPresence sheet, with the "ReasonCategory" column split into separate columns. This change was made to simplify future coding tasks. Although we have not yet implemented any code using "ReasonCategory", pre-splitting the values eliminates the need for additional processing to separate them later, as the original format used a ";" as a delimiter.