Data Science Final Project Pre Proposal

- **Who your group is:** Marianne Aubin Le Quere, Meryl Charleston, Emma Friesner, Callum Nelson
- **Vision:** (Callum)
    - What is your “big idea”?
        - Our big idea is to use data on schools - performance, location, type etc. - to allow us to look at the relationship between type of school, school performance, and location of school. We initially thought to look at Detroit because of the reputation of their charter schools for being underperforming, particularly in underserved neighborhoods. Our goal is to get enough data of this sort to come up with a scheme for predicting what type of school will open in a specific zip code, for example, based on the certain traits of the area.
    - What might you find at the end of your project?
        - We’re hoping to find correlations between school types, performance and demographics of the neighborhoods surrounding the schools. We’re also hoping to have a means for predicting the likelihood of certain kinds of schools in certain areas. These are obviously very high level ideas and really depend on the kinds of data available to us.
- **Data:** (Emma)
    - What dataset do you plan on using?
        - We plan on finding public datasets with information about public ,private and charter schools in different cities. There are some datasets for all schools across America and some specific to cities. We will also use datasets with other characteristics about the city (crime rates, household income, etc) and location within the city to gain insights into backgrounds of families with students attending each school.
        - Ex: CSV files where each line contains info about a different school
            - Public school location Detroit - https://data.detroitmi.gov/Education/Detroit-Public-School-Locations/h8ij-tu87

        - Public schools performance 2008 - 2011 https://www.ed.gov/news/press-releases/us-department-education-releases-school-level-assessment-data-reading-and-math-a
    - How big is it?
        - Datasets specific to one city would most likely have < 1000 entries, while datasets containing information for all American schools could have 10,000 entries
    - How do you plan to collect it?
        - Search online for public records of school performance, location, etc, mostly listed on government data websites.
        - Ex: data.gov, data.detroitmi,gov
    - How do you plan to clean it?
        - We can programmatically parse and clean CSV files
- **Methodology**: (Meryl)
    - What do you plan on doing with your data?
        - We plan on using our data on schools to learn whether there is a correlation between a school's location, type (public/private/charter) and performance. To do this, we will organize the data we have so that we can rank schools within locations, types, and then overall. We would also like to be able to predict where certain types of schools might open and how well they might perform.
    - What techniques do you think you will use to analyze the data?
        - We will have our data in a database with three tables: one with school to unique ID, one with ID and performance, one with ID and type and one with ID and location. We will be able to query our database and order/organize our results by these parameters.
    - How might you visualize your results?
        - One idea we had was a map of the schools. On hovering over a school, the stats would show up. We were also thinking of color coding by type of school/performance, and then having a slider to change the time and then incorporate our predictions.
- **Task List:** (Marianne)

- *What do you plan to accomplish by the first TA check-in?* At a minimum, we would like you to have collected and cleaned your data, but hope that you will make the effort to also integrate and organize it.
    - By the first check-in, we will have decided exactly which cities we want to focus on, and exactly which data we want to gather for them, e.g. crime rate, school achievement measures, etc.
    - We will have picked our exact datasets for each city
    - We will have cleanly mined the data from those different datasets into one large database to use for our final project
- *Likewise, what will you plan to accomplish by the midterm report?* We hope that by this point you will be solidified in your project idea and have begun working on analysis and possibly visualization.
    - We will have finalised our project idea, and which data we want to track/show
    - We will have begun our analysis of the data
    - We will have a plan for the full visualisation
    - We may implement some basic elements of our visualisations, but no animations