**NYC Planning**

**Data Engineering Data Challenge**

**March 2018**

NYC Planning’s Data Engineering team provides services to other teams throughout NYC Planning by modernizing legacy datasets, developing new data products, and improving existing data.  These datasets are then published and used within NYC Planning, by other City agencies, and by the public in a variety of analyses, reports, and research projects.

NYC Planning’s Data Engineering team receives requests for new datasets via an Enterprise Data Management (EDM) request form, one of which is appended to the end of this document.

**Challenge**: Start the request made by the Capital Planning team by beginning to develop the SCA Capital Projects data tables using the open data sources listed.

**Time:** 1.5 hours.  After 1.5 hours please stop and share what you’ve completed so far.

**Deliver**: Short answers to the questions below, any code you write, and any output generated.

1. What was the first thing you Googled (technical or contextual) in relation to this challenge?
2. What tools did you use to complete this challenge?
3. What aspect of developing the data tables did you enjoy the most and why?
4. What did you learn while working through this challenge?
5. If you did not complete building the data tables, where did you get stuck and what questions would you ask to help move you forward?

**EDM request form**

|  |  |
| --- | --- |
| **Dataset** | SCA Capital Projects database |
| **Request date** | Thursday, April 5, 2018 |
| **Business owner** | Capital Planning |
| **Request** | Build the SCA Capital Projects database using Open Data |
| **Reason for request**  **(use case)** | The School Construction Authority (SCA) Capital Projects database is a resource for planners to understand, visualize, and analyze SCA capital projects in an area of interest, so that they can determine if the future needs of a community will be met and where SCA should plan to invest in the future.  SCA publishes several separate data tables about their capital projects on on Open Data; however, the data are most useful for analyses when they are consolidated and the projects are mapped.  It is important that a planner know what the project is, how many students the project may accommodate, and where the project is taking place. |
| **Relevant datasets and location** | Capacity projects:   * [Capacity projects by schools](https://data.cityofnewyork.us/Education/Capacity-Projects-by-schools/a94k-kjys) no location info, only project ids and names * [Capacity project site location](https://data.cityofnewyork.us/Education/Capacity-Project-Site-Location/tesz-9suw)   Capital projects:   * [School based programs by borough](https://data.cityofnewyork.us/Education/School-Based-Programs-by-Borough/esmb-8zkm)   Spatial data: [School point locations](https://data.cityofnewyork.us/Education/School-Point-Locations/jfju-ynrr) |
| **Desired output** | Two data tables (Capacity projects and Capital projects) with the common fields normalized across the tables.  Join the two tables in each category together via a common ID.  Add point geometries (spatial data) onto each project using latitude / longitude, address, or existing point geometry.  Output Format: CSV, Geojson, or shapefile |
| **Desired timeframe** | * Next ~2 weeks if possible, to allow time for QA/QC and incorporation into analyses |