Kenny Waite Term Project: Milestone 1

Data Sources:

1. **School District Location(Website)**:
   - This site contains all of the school districts in North Carolina, as well as the education district region they reside within.
   - Link to data:
     - https://stateboard.ncpublicschools.gov/about-sbe/education-districts

2. **School Grades(CSV/Excel)**:
   - This file contains the school grades for every school in North Carolina for the 2018-2019 school year.
   - Link to data:
     - https://files.nc.gov/dpi/documents/accountability/reporting/spg-report2019_october.xlsx

3. **School Database (API)**
   - This API contains location information about each school within North Carolina
   - Linkt to data:
     - https://live-durhamnc.opendata.arcgis.com/datasets/all-schools-pre-k-through-12/data

Data Relationship:

The three data sources have a simple relationship between them. The **School Database** contains the *school name, school ID,* and *municipality* for each school in NC. I will create a lookup on the **School District Location** data to determine the state region location of the school. I will create a new column with this information called *school district region.* I will link these two data sources on *school district name* from the **School District Location** data and the *municipality* from the **School Database**. The **School Grades Data** will connect to the **School Database** on the *school code number* from the **School Grades Data** and the *facility id* from the **School Database**

Overview:

To order to accomplish all five milestones for this project, I will have to perform different types of data wrangling and transformations on all three data sets. I believe the most challenging data set I will have to work with is the **School District Location Data.** This data is not in an easy to read tabular format. This data is written into chunks on a website page with separate headings for each district location. I will need to use a web page scraper to extract this information into a single table identifying the geographical region for each school district. I will then have to create a join with this table and the **School District Location Data** and create a calculated field to look up each school district and provide the district region for each school. This portion will be the most difficult because I do not know how to do either the web scraping or the database lookup/join.

Finally, I will have to join the **School Grades Data** with the **School Database** on the *school code number* from the **School Grades Data** and the *facility id* from the **School Database.** These are the unique keys for each school that I can link between the two datasets. This part should not be as difficult as there is a one to one relationship for school each school in each database. The School Grades Data and the School Database are relatively clean data sources. However, I will have to do some analysis to review for duplicates, errors in the data, and identifying any additional inconsistencies in the data. Also, I will have to ensure every school received a grade. If the school did not receive a grade, I would remove those schools. I will investigate the reason for a school not having a grade further to ensure the removal will not throw off the analysis, and the schools did not receive a grade for a specific reason.

To me, I would summarize each data set to contain the following information:

- The **School Database** school location information, including the *school name, school ID,* and *school district* for each school in North Carolina. The **School Database** will be the primary dataset I will use to connect with the other two data sources.
- The **School District Location** contains all of the school districts in North Carolina and state region locations. I will create a new column with this information called *school district location.*
- The **School Grades Data** contains all of the state assessed school grades based on the school's performance on the North Carolina state testing held the previous year. I will link to the **School Database** on the *school code number* from the **School Grades Data** and the *facility id* from the **School Database.**