**ETL - PROJECT**

The goal of this project was to prep data using ETL methodology and gain valuable insights about Chicago Public Schools. We reviewed datasets from data.cityofchicago.org and Scribd. We began the process by searching for various data from the City of Chicago website and then choosing a topic that was interesting to us. We then found multiple data sets and reviewed them thoroughly.

Below was our specific process and steps:

**Extract/Data Utilized**

The key data we extracted was from the City of Chicago website in the form of CSV, Text, and PDF files. The files were the Chicago Public Schools Progress Report – SY1819 and District Representative List. The third data set was from Scribd.com and was called The Accessibility List. These sites provide data in multiple formats (10) and we chose to go through three different formats for transformation.

**Transformation**

The first data set was a csv file from the City of Chicago website and this the largest and most robust data set of our selection. It contained over 150 columns across 650 schools, but we just chose a few key columns for better data readability. We used pandas to create a data frame and then SQLAlchemy to load into SQL database.

The Accessibility Report required an extensive process to transform. This report was loaded as a text file, split, stripped, cleaned and loaded into a pandas data frame.

The third data set was District Representative report which was in a PDF file which we converted to a csv file through a conversion website called pdftables.com. We removed columns that weren’t needed or had missing data. Through pandas, we then created a data frame from this data set.

**Load**

The first step in the loading process was creating tables in SQL using pgadmin. We exported the three data frames from Jupyter Notebooks into tables in pgadmin. We then joined both District Representative data with data from the Report Card data frame on the column of “School ID”.

We then did a second join of Report Card data and Accessibility Data on School\_Name aiming to get the information of availability of accessibility in schools and district representatives and to see the outliers from this.

**In Summary**

This particular set of data is a real-world example useful to us as parents. It could be used to come up with a comprehensive list of the district representatives and their e-mail addresses that would be used for information regarding insufficient handicap accessibility across all 650 CPS schools.