**CDC Risk Behavior Assessment ETL Project Technical Report**

To find the data for this project, we went to [Kaggle](https://www.kaggle.com/raylo168/dash-yrbss-hs-2017) and explored open source data sets. Ultimately, we selected a dataset from DASH, a subset of the CDC focused on studying the health and risky behaviors of K to 12 students. The set came from [Youth Risk Behavior Surveillance System](https://chronicdata.cdc.gov/Youth-Risk-Behaviors/DASH-Youth-Risk-Behavior-Surveillance-System-YRBSS/svam-8dhg) (YRBSS), a study monitoring categories of behaviors in students. The data we used from their study focused on alcohol consumption, sexual behaviors, and tobacco use. Within each survey category, the survey questions (column headers) were the different aspects of the students’ individual behaviors as they reported in the surveys as well as general information such as survey location, sample size, year, question id, etc.)

Extraction: Excellent.

The raw data was both massive and poorly formatted. Much of the information in the set was either useless or redundant. For example, locations were given abbreviated, fully named, and in Lat/Lng. Other data had similar redundancies. Ultimately, we were able to carve each set down from 35 columns to 15. Much of that data was incomplete as well, for simplicity’s sake we made the decision to drop rows with null values. Column titles were similarly convoluted sometimes using underscores as spaces and other times just capitalizing the first letter of each new word in a given title. Columns were renamed to be formatted consistently.

Transformation: Tubular

To load the data, we created a set of empty tables in a Postgres database with columns that matched our renamed columns in our ipynb Jupyter notebook. Our original sets were without a primary key so we left it to Postgres to manage that by setting our index to false when we moved our DBs to sql with .to\_sql. We hit a bump in the road here because we had set the primary key id as int rather than serial in Postgres. We found the error, made the correction and the load went off without a hitch.

Load: Lovely

Please refer to [the repo](https://github.com/ELLowe/YRBSS_db_ksel_project) for the final database.